2011

Bayesian Hierarchical Generalized Linear Models For Detecting (Rare) Haplotype-Haplotype And Haplotype-Environment Interactions In Genetic Association Analysis

Jun Li

University of Alabama at Birmingham

BAYESIAN HIERARCHICAL GENERALIZED LINEAR MODELS FOR DETECTING (RARE) HAPLOTYPE-HAPLOTYPE AND HAPLOTYPE-ENVIRONMENT INTERACTIONS IN GENETIC ASSOCIATION ANALYSIS

by

JUN LI

NENGJUN YI, COMMITTEE CHAIR
NIANJUN LIU
UPENDER MANNE
BORIS C. PASCHE
KUI ZHANG

A DISSERTATION

Submitted to the graduate faculty of The University of Alabama at Birmingham,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

BIRMINGHAM, ALABAMA

2011

BAYESIAN HIERARCHICAL GENERALIZED LINEAR MODELS FOR DETECTING (RARE) HAPLOTYPE-HAPLOTYPE AND HAPLOTYPE-ENVIRONMENT INTERACTIONS IN GENETIC ASSOCIATION ANALYSIS

JUN LI

BIOSTATISTICS

ABSTRACT

This dissertation research focuses on genetic association analysis based on haplotypes in the context of both population-based and family-based studies. Haplotype-based association analysis is powerful in the discovery and characterization of the genetic basis of complex human diseases. However, statistical models that fit haplotype-haplotype and haplotype-environment interactions have not yet been fully developed. Furthermore, statistical methods for detecting the association between rare haplotypes and disease have not kept pace with their counterpart of common haplotypes. For both population-based and family-based association analyses, we herein propose two efficient and robust methods to separately tackle these problems based on Bayesian hierarchical generalized linear models. Our models simultaneously fit environmental effects, main effects of numerous common and rare haplotypes, and haplotype-haplotype and haplotype-environment interactions. The key to the approaches is the use of a continuous prior distribution on coefficients that favors sparsity in the fitted model and facilitates computation. We develop a fast expectation-maximization (EM) algorithm to fit models by estimating posterior modes of coefficients. We incorporate our algorithm into the iteratively weighted least squares for classical generalized linear models as implemented in the R package `glm`. We evaluate the proposed methods and compare their statistical properties to existing approaches on extensive simulated data. The results show that the proposed methods perform well under all situations and are more powerful than the competitors.

Keywords: Bayesian methods, Generalized linear models, Association studies, Haplotype, Interactions, Rare variants

DEDICATION

This dissertation is especially dedicated to Ying, my beloved wife, and Yunbo, my cherubic little son.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my mentor, Dr. Nengjun Yi. He introduced me to the field of Bayesian inference and inspired me to study Statistical Genetics. I have benefited from his wisdom, sharpness, preciseness, and efficiency. Without his guidance and support, my doctoral studies could not have been so fruitful.

Next, I would like to thank Dr. David Allison for providing me financial support during my dissertation work. The Section on Statistical Genetics headed by him has provided me with excellent academic environment and many learning opportunities. I would not have been able to complete my dissertation without his support.

I thank my committee members, Drs. Kui Zhang, Nianjun Liu, Boris Pasche, and Upender Manne, for all the constructive comments and suggestions on my work.

Many thanks to the Department of Biostatistics, a place nurtured my scientific thought and spurred my academic growth.

I want to express my deepest gratitude to my wife for her love, understanding, and support all the way, which make this dissertation possible. I also want to thank my little son, Yunbo, who has brought unbelievable joy to my life. They make all of my efforts meaningful. I am also grateful to my family in my hometown for their support.

Finally, I thank my friends, Lang Chen, Xuehua Chen, and Mei Huang, and my fellow students, Jiatao Ye, Thomas Birkner, Nathan Wineinger, Guo-Bo Chen, Jihua Wu, and Milind Phadnis, for their invaluable assistance on academic and non-academic issues.

TABLE OF CONTENTS

Page

CHAPTER

LIST OF TABLES

Table *Page*

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Over the past few decades complex human diseases such as cancer, diabetes, obesity, and cardiovascular diseases have constituted enormous health burden around the world and therefore become a particularly great concern to both the public and health professionals (e.g., King *et al*., 1998; Lopez *et al*., 2006; Boyle and Levin, 2008; Finkelstein *et al*., 2008; Ramahi, 2010).

To gain great insight into the mechanisms by which such diseases are developed, a considerable effort and expense have been put forth during the same time period (Altmuller *et al*., 2001). The first step toward this particular goal is to discover which genes, or more precisely, which genetic polymorphisms or variants, are involved in the diseases. Until recently hundreds of genetic variants contributing to complex human diseases have been identified (Hindorff *et al*., 2009; Hindorff *et al*., 2010). Furthermore, the identification of such genetic variants is partly revolutionizing the field of medicine, providing more effective prevention, earlier diagnosis, and more-targeted, personalized therapies (e.g., Risch, 2000; Collins *et al*., 2003; Shastry, 2006; van't Veer and Bernards, 2008).

The most promising approach for identifying genetic variants that are related to complex human diseases is generally accepted to be genetic association analysis, provided that the frequencies of disease-susceptibility variants are not too low (Risch and

Merikangas, 1996; Zondervan and Cardon, 2004). The basic idea underlying genetic association analysis is to test whether the frequencies of alleles or genotypes at one locus or loci are different between comparison subpopulations, usually diseased subjects and healthy controls, by which researchers attempt to find out a genetic variant that either directly predisposes to disease or is in linkage disequilibrium (LD) with a causal variant (Cordell and Clayton, 2005). LD is known as a nonrandom alignment of alleles at different tightly linked loci in a population, and plays a fundamental role in the study of population genetics as a potentially powerful tool for the localization of genetic variants for complex human diseases (Hartl and Clark, 2006). The success of a genetic association analysis depends, in part, on the extent of LD of a disease-susceptibility locus with a genetic marker locus within a population. This means that a disease-susceptibility variant initially occurred close to a specific allele of a nearby genetic marker. As generations (and meioses) proceed over time, the disease-susceptibility variant and the marker allele remain statistically associated because their physical proximity remarkably reduces the number of recombination that occurs between them.

Genetic association analysis has been shown to be often more statistically powerful than linkage analysis because a valid association may be detected in a sample in which linkage is not detectable, particularly when the genetic variant is playing only a moderate role in disease susceptibility. This is especially likely when a marker itself is a susceptibility variant (Risch and Merikangas, 1996; Risch, 2000; Botstein and Risch, 2003). In addition, along with the rapid development of high-throughput genotyping technology and the availability of large amount of genetic markers, genetic association analysis has become increasingly popular in both genome-wide scan and fine mapping of candidate regions (Laird and Lange, 2006; Huang *et al*., 2009). Many causal variants for disease such as type 1 and type 2 diabetes, prostate cancer, breast cancer, and inflammatory bowel diseases have been identified through genetic association analysis (McCarthy *et al*., 2008). This breakthrough findings offer renewed hope for the fight against complex human diseases.

To detect a disease-susceptibility variant successfully, association analysis requires a high-density map of genetic markers because LD among variants occurs only over a short genetic distance. There are many kinds of genetic markers that can be used to construct such a map, for example, restriction fragment length polymorphism (RFLP), simple sequence repeat (SSR), and single nucleotide polymorphism (SNP). Among all the genetic markers, SNP is the most widely used one for gene mapping in complex human diseases due to its (nearly) complete coverage over the whole human genome with a high density, although all the other genetic markers are still very useful for genetic association analysis (Carlson *et al*., 2001).

SNP could be defined as the variation of deoxyribonucleic acid (DNA) sequence occurring when a single nucleotide (A, T, C, or G) in the genome differs between members of a species or paired chromosomes in an individual. Single-SNP-based methods are suitable for detecting association of genetic variants with disease, provided that LD between the disease-susceptibility variant and the allele of a genetic marker is strong. When LD decreases, however, the power of single-marker-based association analysis might suffer. The reason is that the power of single-marker-based methods in association analysis depends on the LD between the disease-susceptibility variant and the allele of a genetic marker. LD information contained in flanking markers generally is not incorporated, which can result in a reduction in power (Kaplan and Morris, 2001). When multiple SNPs are simply used in a study, some big challenges will present in analyzing hundreds of thousands of SNPs from a huge sample size, not only because of the high dimensionality of data, but also because of their complicated interrelated structure. To avoid these problems and, more important, take advantage of the linkage information from multiple SNPs together, SNP-based association analysis has been expanded to haplotype-based association analysis.

Haplotype refers to the specific combination of alleles that are in alignment on a single homolog, one of the two homologous chromosomes in humans, and that tend to be inherited together. As the unit of analysis for statistical tests in association analysis, haplotypes have long been of great interest and have drawn much attention in recent years (Clark, 2004; Davidson, 2000; Schaid, 2004; Schaid *et al*., 2002). There are several reasons behind this phenomenon. First, haplotypes are biologically relevant. There is strong evidence that several mutations within a gene may interact together (*cis*-interaction) to cause disease such as neural tube defects and prostate cancer (Tavtigian *et al*., 2001; Joosten *et al*., 2001; Fitze *et al*., 2002). Haplotype-based methods provide a natural way to capture such *cis*-interactions by examining a number of adjacent loci and accommodating the joint effects from them (Morris and Kaplan, 2002). Second, haplotype-based association methods are generally regarded as being more powerful than methods based on single markers since the former fully exploits LD information from multiple markers (Akey *et al*., 2001; Morris and Kaplan *et al*., 2002). Both simulation (Akey *et al*., 2001; Zaykin *et al*. 2002) and empirical studies also support this conclusion. Third, haplotype-based methods can be advantageous over SNP-based methods when multiple disease-susceptibility variants occur within a single gene and each of these variants originates and predisposes to disease independently of the other variants (Morris and Kaplan, 2002). In addition, driven by the international HapMap project, considerable information concerning haplotype structures and haplotype frequencies has been gained from several populations (The International HapMap Consortium, 2005). Nowadays, haplotypes have been widely accepted as a major tool for identifying disease-susceptibility variants in genetic association analysis.

However, one difficulty in applying haplotype-based association analysis is that actual haplotypes for each individual can not be easily obtained directly. Although it is possible to determine haplotypes through molecular techniques, such techniques are often expensive and too laborious to be practical in large-scale studies (Michalatos-Beloin *et al*., 1996; Eitan and Kashi, 2002). In the routine laboratory work, the polymerase chain reaction (PCR), the current standard genotyping technique, is usually used to generate marker genotypes, in which, for a normally diploid organism such as humans, only the two alleles at a single locus can be discerned for an individual, without providing any information regarding the chromosome which is associated with each allele, known as phase information. Therefore, for an individual who is heterozygous at more than one locus, say n ($n > 0$) loci, there are total of 2^{n-1} possible haplotype pairs that are consistent with the observed single locus genotypes, and haplotype phase for this individual is said to be ambiguous. For example, if we consider three observed diallelic SNPs with genotypes (A,a), (B,b), and (c,c), the first two genotypes denoting the heterozygosity and the last one denoting the homozygosity, then there are two possible pairs of haplotypes that the individual may carry: *ABc*/*abc* or *Abc*/*aBc*, where "/" is used to separate the two haplotypes within a haplotype pair, with each aligned on one of two homologous chromosomes. This ambiguity of haplotype phase complicates haplotype-based association analysis.

To overcome this difficulty, numerous methods have been proposed for inferring haplotypes through the estimation of haplotype frequencies for the study population and the resolution of haplotype pairs within individuals (e.g., Clark, 1990; Excoffier and Slatkin, 1995; Fallin and Schork, 2000; Stephens *et al*., 2001; Niu *et al*., 2002; Qin *et al*., 2002). Among these methods, the expectation-maximization (EM) algorithm is probably most frequently used (Excoffier and Slatkin, 1995; Fallin and Schork, 2000; Qin *et al*., 2002). The EM algorithm is a well-established approach for estimating unobservable parameters in the context of missing data. Details on the original algorithm and further intuition behind its inception can be found in much of literature, including Dempster *et al*. (1977), Sundberg (1974), Wu (1983), and so on. Typically, the basic idea behind the EM algorithm is that in the E-step the posterior probability of each possible haplotype pair is estimated within an individual that are consistent with the observed genotypes, and in the M-step the haplotype frequencies are updated given the current estimated posterior probabilities. Then iteration between these two steps proceeds until convergence. Given these estimated haplotype frequencies, the posterior probability that an individual with the observed genotypes has a specific haplotype pair can be computed using Bayes' rule. These posterior probabilities of haplotype pairs for each individual can be used to define haplotype variables to be included in a standard analysis such as logistic regression (Zaykin *et al*., 2002; Stram *et al*., 2003).

With respect to defining haplotype variables, a naive approach is to assign the most likely haplotype pair to an individual, and then the standard analysis is implemented as if the haplotype pair was exactly observed. The potential pitfall of this naive strategy is that ignoring the uncertainty in the haplotype assignment can introduce measurement error and further induce bias into the estimates of haplotype effects (Lin and Zeng, 2006; Lin and Huang, 2007; Kraft and Stram, 2007). The second way to handle the uncertainty in the haplotype assignment is a multiple imputation technique. Using this approach, a number of replicate datasets are generated by randomly assigning a haplotype pair to an individual that is in accordance with the individual's haplotype posterior probabilities. Then haplotype effects are estimated by taking the average of the estimates across the imputed datasets (Kraft *et al*., 2005). In addition, an innovative method has been proposed in which haplotype frequencies and haplotype risk effects can be estimated simultaneously (Schaid *et al*., 2002; Epstein and Satten, 2003; Zhao *et al*., 2003; Stram *et al*., 2003). The attractive feature of this method is that it can jointly deal with uncertainty in haplotype assignment and uncertainty in haplotype frequency estimates (Kraft *et al*., 2005). Finally, a relative simple but powerful way to handle this problem is to use the expectation-substitution method to compute the expected number of copies of a specific haplotype (estimate of haplotype dosage) for an individual using all the possible haplotype pairs that are compatible with his or her observed genotypes (Zaykin *et al*., 2002; Stram *et al*., 2003; Kraft and Stram, 2007). Although the method is so-called "single imputation", it can provide pretty good reliability and decent power for estimating haplotype risk effects (Kraft and Stram, 2007).

In the past two decades, large numbers of haplotype-based genome-wide and candidate gene association analyses have been conducted and it has been demonstrated that haplotype-based association analysis is a potentially cost effective and statistically powerful tool to unravel the genetic mechanisms that are underlying complex human diseases (Risch and Merikangas, 1996; Botstein and Risch, 2003). In haplotype-based association analysis, a lot of statistical methods have been proposed to examine the association between haplotypes and human complex diseases (e.g., Zaykin *et al*., 2002; Lake *et al*., 2003; Zhao *et al*., 2003; Cordell *et al*., 2004). Although many of these approaches have been widely used in the mapping of genes contributing to complex human diseases, the majority of them only focused on estimation of marginal effects of haplotypes and detection of association between common haplotypes and disease, while comparatively little attention has been paid so far to investigating interacting effects between haplotypes and environmental factors, especially those between haplotypes in different haplotype blocks, and exploring disease association with rare haplotypes (Becker *et al.*, 2005; Guo and Lin, 2009).

Complex human diseases are believed to be influenced by numerous different genetic and environmental factors, and the interplay of these two kinds of factors (e.g., Moore, 2005; Cordell, 2009). Consideration of interaction in analysis can potentially lead us to a better understanding of fundamental biological mechanisms and pathways in disease progression. Thus, an ideal strategy of analysis is to simultaneously consider all the genetic loci, environmental factors, and particularly their interactions. Such a joint analysis could enhance the power for detecting genetic variants that are involved in the etiology of disease mainly through an interacting effect with no marginal effect

(Chapman and Clayton, 2007), and/or ascertaining environmental factors that act primarily in genetically susceptible individuals (Thomas, 2010a). In addition, accommodating interaction in analysis can overcome the limited success in the detection of disease-predisposing genetic variants for complex human disease, or improve the explanation of heritability of most complex diseases that might be attributed to interactions or more complex pathways involving multiple genetic and environmental factors (Manolio *et al.*, 2009; Eichler *et al.*, 2010).

However, identifying interactions that are causal in complex human diseases is not an easy task (Cordell, 2009; Kooperberg *et al*., 2009; Thomas, 2010a; Yi, 2010). Primarily, the detection and characterization of interactions are limited due to the lack of powerful statistical methods and/or large sample sizes. When numerous interactions are fitted explicitly in a model, the degrees of freedom for the corresponding test statistics would grow rapidly, and, as a result, sufficient power cannot be guaranteed to detect possibly significant effects in the model, especially in a relatively small sample size (Luan *et al*., 2001; Boks *et al*., 2007; Mukherjee *et al*., 2008; Cordell, 2009; Thomas, 2010a). This issue may become more severe in haplotype-based association analysis, where haplotypes are usually inferred form SNPs as discussed before. With increasing number of SNPs, the number of possible haplotypes can become extremely large, leading to the related problems of high-dimensional data and sparse data for many of the haplotypes. The classical statistical methods such as logistic regression usually have no sufficient power and flexibility to handle these problems (Lake *et al*., 2003; Becker *et al*., 2005; Kwee *et al*., 2007; Hein *et al*., 2009). Furthermore, up to now few innovative methods have been developed to tackle such problems in haplotype-based association analysis. Therefore, a sophisticated method is desired that accommodates interactions as well as high dimensionality and sparsity. This poses a considerable challenge and serves as the motivation for our present research.

Another potential factor that could help explain more proportion of the heritability of most complex human diseases is rare variants (Manolio *et al.*, 2009; Eichler *et al.*, 2010). The rare variant is what has a relatively low minor allele frequency (MAF) or a rare homozygous genotype frequency in the population. So far there is no clear and consistent definition for the rare variant. Some researchers in the literature defined a variant with a MAF less than or equal to 0.05 or 0.01 as rare (e.g., Asimit and Zeggini, 2010, Bansal *et al.*, 2010). But most of authors used this term loosely, only to refer to variants that have less common MAF than those routinely studied. Anyway, no matter how the rare variant is defined, it has received very little attention for a long time in genetic association analysis, although it supplies valuable information on the mechanism by which disease is caused (e.g., Pritchard, 2001; Cohen *et al.*, 2004; Azzopardi *et al.*, 2008). This is understandable, because such variants with very low frequencies and individually small contributions to the overall inherited disease susceptibility cannot be detected unless the statistical method is much powerful or the sample size is unusually large (Altshuler *et al.*, 2008; Gorlov *et al.*, 2008; Li and Leal, 2008; Bodmer and Bonilla, 2008; Basu and Pan, 2011). This is the main reason why the common disease-common variant (CDCV) hypothesis prevails in the contemporary genetic studies. However, although hundreds of genetic variants associated with common diseases have been detected in the studies under the CDCV hypothesis, those variants have only a weak effect on disease risk, and hence only explain a small proportion of the heritable, genetic component of susceptibility to those diseases (Maher, 2008; Dickson *et al.*, 2010; Morris and Zeggini, 2010; Robinson, 2010). The unexplained part of heritability could be partly due to rare variants (Manolio *et al.*, 2009; Eichler *et al.*, 2010). This motivated researchers to consider the contribution of rare variants to susceptibility to common diseases, which is known as the common disease-rare variant (CDRV) hypothesis. The hypothesis postulates that disease is caused by some genetic variants with detectable strong effects, each of them being only found in a few individuals in the population (Bodmer and Bonilla, 2008; Morris and Zeggini, 2010; Robinson, 2010; Hoffmann *et al.*, 2010). The role of rare variants in complex human diseases such as hypertension, type 1 diabetes, and obesity has been identified by recent studies under the CDRV hypothesis (Ji *et al.*, 2008; Nejentsev *et al.*, 2009; Bochukova *et al.*, 2010).

Rare haplotypes, just like other genetic rare variants, could be important disease-predisposing variants and should not be ignored in investigating the genetic susceptibility to complex human diseases (Liu *et al*., 2005; Zhu *et al*., 2005; Yende *et al*., 2007; Semsei *et al*., 2008; Kitsios and Zintzaras, 2010). Rare haplotypes can be seen frequently in genetic association studies and even they might be produced by common SNPs in a population (Souverein *et al*., 2008; Guo and Lin, 2009). Regarding statistical modeling, however, rare haplotypes can result in nonidentifiability of parameters in model fit, which means the coefficients of predictors cannot be identified or estimated uniquely because of huge, even infinite standard errors (Gelman *et al*., 2003). There are several methodological and computational issues that complicate research of nonidentifiability of parameters. The big, even huge, estimate of parameter is the major obstacle. A common, but negative solution to this issue in the literature is to pool all rare haplotypes into one single group (Schaid *et al*., 2002; Zhao *et al*., 2003) or pool rare haplotypes with common ancestral haplotypes (Seltman *et al*., 2003; Durrant *et al*., 2004; Tzeng, 2005). These approaches in nature ignore rare haplotypes by lumping them together, and consequently any rare haplotype that might contribute to the risk of disease cannot be identified distinctly.

Obviously, the development of methods that can detect the rare variants and handle the nonidentifiability of parameters is a much needed area of research. Progress in this area requires introducing comprehensive, standardized, and precise approaches to capture all information arising from both common and rare haplotypes. This provides the second motivation for the research in this dissertation.

In summary, our research focuses on the two main topics: haplotype-related interactions (haplotype-haplotype and haplotype-environment interactions) and rare haplotypes in association analysis. These two topics are thoroughly investigated in both population-based and family-based association analyses in CHAPTER 2 and CHAPTER 3 of this dissertation, respectively, because population-based and family-based association analyses are the two major branches in the contemporary genetic studies, classified based on the study design and sample collection of a study. To make our proposed methods to be easily implemented and publicly available, we incorporate them into `R/BhGLM` software and briefly describe them in CHAPTER 4.

CHAPTER 2

POPULATION-BASED HAPLOTYPE-ASSOCIATION ANALYSIS IN CASE-CONTROL STUDIES

In general, population-based haplotype-association analysis aims to detect the relationship between haplotypes and disease or quantitative phenotypes, using unrelated individuals as the primary sampling units. This kind of analysis has marked a great potential for unraveling genetic mechanisms that are underlying complex human diseases (Stephens *et al*., 2001; Schaid *et al*., 2002; Botstein and Risch, 2003; Clark, 2004; Schaid, 2004).

This chapter starts with a summary of major features of population-based haplotype-association analysis. Subsequently, basic concepts of case-control studies in the context of genetic association investigations are described. Some existing statistical methods for population-based haplotype-association analysis are briefly described in Section 2.3. The chapter finishes with deriving a new Bayesian hierarchical generalized linear model that can detect (rare) haplotype-haplotype and haplotype-environment interactions in population-based association analysis.

2.1 Main Features of Population-based Haplotype-Association Analysis

2.1.1 Assumption of Independence

In population-based haplotype-association analysis, a fundamental assumption is that individuals under investigation are unrelated, which implies that most of association methods for uncorrelated data can be applicable directly. It is worth noting that while the assumption of independence often holds in population-based haplotype-association analysis, situations might arise where we would expect departure from independence, for example, when a trait on the same individual is measured repeatedly in a longitudinal study. In such situations, inference on haplotype effects can be biased for the methods with the assumption of independence. To solve this problem, some advanced statistical methods which accommodate correlated data are warranted and are essential for correctly estimating variance components (Fitzmaurice *et al*., 2004; Gelman and Hill, 2006; Song, 2007).

2.1.2 Ambiguity of Haplotype Phase

Another remarkable aspect is that the information of allelic phase for a SNP is generally not available in the context of population-based association analysis and hence the corresponding haplotypes of an individual usually cannot be acquired directly as described in CHAPTER 1. This presents a considerable challenge for analyzing haplotype-based association with traits in population-based studies. To address this challenge, we must first perform haplotype inference including estimating haplotype frequencies and reconstructing haplotype patterns for each individual based on the observed genotype data collected from unrelated individuals, and then we can conduct haplotype-based association analysis. Consequently, these studies tend to differ in data structure and statistical methods from some other genetic association studies, for example, SNP-based association studies and family-based haplotype-association studies that will be discussed in the next chapter (Thomas, 2004; Schaid, 2004).

2.1.3 Population Stratification

Human populations often exhibit a systematic difference in allele frequencies between subpopulations, which is usually referred to as population stratification or population structure. The main reason of population stratification is that random mating occurs within each of subpopulations while non-random mating, or more precisely, gene migration occurs between subpopulations. Note that there is another concept, population admixture, which often appears together with population stratification (McKeigue, 2007). Population admixture is used loosely in the scientific literature to indicate a population in which multiple subpopulations with different allelic distributions are present. In the population-based association studies, we usually focus on the issue arising from population stratification.

Population stratification may result in spurious association in genetic studies (Li, 1969; Devlin and Roeder, 1999; Pritchard *et al.*, 2000a). As an example, consider the situation in which a population consists of two subpopulations each having different allele frequencies at a locus and differing prevalences of disease. We assume that the locus is not causally associated with the disease and, for simplicity, that the first subpopulation has a higher allele frequency at the locus as well as a higher prevalence of the disease. A random sample of cases from the population will tend to have more individuals of the first subpopulation than a random sample of controls drawn from a natural population. Then an unstratified case-control study will lead to inflated estimates of the effect of the locus on the disease risk.

Several methods have been proposed to control the impact of population stratification in population-based association studies (e.g., Devlin and Roeder, 1999;

Pritchard *et al*., 2000a; Price *et al*., 2006; Epstein *et al*., 2007), among which the widely used ones broadly follow one of three concepts: genomic control, structured association, and principal components. The approach of genomic control uses random marker loci to obviate the false positive association due to population stratification (Devlin and Roeder, 1999; Reich and Goldstein, 2001). The method of structured association directly infers population structure and incorporates the estimated population structure in the test of association (Pritchard *et al*., 2000a,b; Satten *et al*., 2001; Chen *et al*., 2003; Hoggart *et al*., 2003; Purcell and Sham, 2004). The principal components analysis identifies principal components that represent the population structure based on genetic correlations among individuals (Yu *et al*., 2006; Malosetti *et al*., 2007; Zhao *et al*., 2007). However, all these methods have had only limited success in controlling the false association signals due to population stratification. It is important to recognize that these methods can only minimize the potential impact of population stratification or, to put it another way, they cannot completely remove the spurious association resulting from population stratification because the hidden population structure is usually unknown and cannot be corrected for at the time of statistical analysis (e.g., Lange *et al*., 2008; Zhang *et al*., 2008; Price, *et al*., 2010).

Although population-based haplotype-association analysis has potential to lead false positive findings attributable to population stratification, they have some advantages over family-based haplotype-association studies that will be discussed in the next chapter. For example, in population-based haplotype-association analysis, it is relatively easy to recruit subjects, and each individual contributes one observation to the statistical test. Haplotype-association studies are usually more efficient in terms of time, money, and logistics. Moreover, for late-onset diseases, it is impossible to collect parents of the affected subjects (Scott *et al*., 1997). Therefore, there is a great need for population-based haplotype-association analysis.

2.2 Case-Control Studies

In this section, basic concepts and properties of case-control studies are briefly described, followed by a discussion of how population-based haplotype-association analysis fits within case-control studies. Further discussions of case-control studies in the context of genetic association investigation can be found in Thomas (2004), Schaid (2004), Clayton (2007), and Ziegler and Koenig (2007).

2.2.1 Concept and Principles of Case-Control Studies

In a case-control study, two groups are sampled and compared with respect to their potential exposure of risk or protective factors; one group consists of affected subjects referred to as cases, and the other consists of unaffected subjects referred to as controls. The basic assumption of such studies is that the two groups of subjects may be employed to provide unbiased estimates of the corresponding distributions of the cases and controls. Based on this assumption, some statistical methods are used to determine whether there is a difference of past exposure to the suspected risk or protective factors between the cases and controls. If the exposure and the disease do not occur independently from each other, an association between the exposure and the disease is said to exist. The strength of association is usually assessed by a measure, odds ratio (OR), which is generally the ratio of the odds of an event occurring in one group to the odds of it occurring in another group.

2.2.2 Advantages and Disadvantages of Case-Control Studies

Since the exposure of interest is collected after the development of the disease in question (the true order is that subjects have to be exposed before a disease is developed), case-control studies can be called a retrospective study. This is a prominent characteristic of case-control studies and it offers some advantages and disadvantage over another frequently-used class of studies, cohord studies, in which subjects with different exposures to the suspected risk or protective factors are recruited and followed over time for the occurrence of disease, and then the occurrence rates of the disease are measured and compared between the two groups. Case-control studies are a relatively cheap, quick, and reliable approach of establishing evidence of an association between exposure to risk or protective factors and disease.

Case-control studies have proved particularly useful in studying rare and late-onset diseases. However, their retrospective nature limits the strength of their conclusions because the mechanism of disease cannot be studied and a proof of causation cannot be established. In addition, several biases such as selection bias and information bias can be introduced into case-control studies in the process of identifying study population, measuring information on exposure or disease, and so on. Bias is defined as any systematic error in a study that results in an incorrect estimate of the association between exposure and disease, and it should be avoided by an appropriate design and careful data collection. Another problem in case-control studies is that some risk factors might act as confounders. A confounder is a third (extraneous) factor which is related to both exposure and disease but not an intermediate step between the exposure and disease, and it can lead to an overestimation or underestimation of the true relation between the exposure and disease. Confounding is not an error in a study, but rather is a true phenomenon that exists in nature in a study and must be identified, understood, and interpreted in study design and/or analysis of data. In the design, some techniques could be used to restrict for potential confounders, and in the analysis, stratification and/or multivariable (adjusted) analysis could be used (Rothman *et al*. 2008a,b).

Nonetheless, the value of case-control studies in rapid and inexpensive assessment of a new or serious disease has been proved beyond doubt (e.g., Breslow and Day, 1980; Rothman *et al*., 2008a,b,c).

2.2.3 Population-Based Genetic Association Analysis in Case-Control Studies

Case-control studies are very popular in the context of population-based genetic association investigations of complex human diseases. In this area, genetic risk factors are used as exposure to investigate association with the status of case and control which is commonly termed phenotype. There are kinds of genetic risk factors, but we limit our discussion to the observed genetic sequence information, or more precisely, the combination of alleles located on homologous chromosomes, which is defined as genotype. In case-control studies, a genetic locus that is supposed to be investigated is genotyped for cases and controls, and the frequency of an allele or genotype of the genetic locus is compared between the cases and controls. If there is a difference in the frequency of the allele or genotype under test between the two groups, an association is said to exist between the genetic locus and the disease, which means that the genetic locus may increase the risk of the disease (itself is causal), or be in linkage disequilibrium with a causal locus which does. If several genetic loci are genotyped they can either be tested separately or jointly for association with disease. Since the joint analysis employs to the greatest possible advantage of LD information from multiple loci, it has been considered in the most of genetic studies. Haplotype-based studies are an excellent instance of jointly analyzing multiple loci. But for haplotypes-based association studies, since the haplotypic phase is generally unobservable in population-based association studies of unrelated individuals, a special consideration for analysis is required as described in detail in CHAPTER 1.

In case-control studies, false positive results caused by population stratification may also occur when we use population-based haplotype data. The issue of population stratification is briefly described in the previous section. Such a spurious effect should be eliminated by an appropriate design and careful data collection because the population stratification is usually unknown and cannot be corrected for in statistical analysis.

2.3 Existing Statistical Methods for Population-Based Haplotype-Association Analysis

A variety of statistical methods have been developed to detect haplotype-disease association through use of population-based data from case-control studies. Early attempts to such methods were made over ten years ago by simply comparing the estimated haplotype frequencies between cases and controls (Zhao *et al*., 2000; Fallin *et al*., 2001). These approaches perform global tests of haplotype association with disease and can be implemented easily in the routine statistical analysis. However, they do not provide estimates for individual haplotypes due to the nature of omnibus test. Moreover, the estimated haplotype effects cannot be adjusted for environmental factors.

Schaid *et al*. (2002), Zaykin *et al*. (2002), Stram *et al*. (2003), and Zhao *et al*. (2003) developed separate methods to deal with these problems. All these methods treat haplotypes as explanatory variables in a regression model, and thus they can estimate the effects of individual haplotypes and adjust for environmental factors. The method of Schaid *et al*. (2002) is build upon the conditional probability distributions of subjects' possible haplotype pairs given the observed genotype data and inferred haplotype frequencies. To take into account ambiguity of inferred haplotypes, it uses an EM algorithm to compute the posterior probabilities of haplotype pairs for each subject. Moreover, to account for uncertainty for haplotype assignment for each subject, it calculates expected haplotype score. Similarly, by using an EM algorithm, Zaykin *et al*. (2002) computed the expected counts based on the posterior probabilities of haplotype pairs given the observed genotype data, and then fitted the counts to the disease using a regression model. Stram *et al*. (2003) constructed a joint likelihood of disease and genetic and environmental covariates, from which they obtained the maximum likelihood estimates of individual haplotype effects. Zhao *et al*. (2003) applied a similar joint likelihood method as those of Stram *et al*. (2003) but assumed Hardy-Weinberg equilibrium (HWE) of haplotype frequencies within the sample of controls.

Although each of these methods has its attractive features, particularly in estimating individual haplotype effects directly and accounting for non-genetic covariates, all of them only focus on estimation of marginal effects of haplotypes, and no attention was paid to investigate interacting effects between haplotypes and environmental factors, especially those between haplotypes in different haplotype blocks. However, increasing evidence suggests that gene-gene and gene-environment interactions play an important role in susceptibility to complex human diseases (Cheverund and Routman, 1995; Wolf *et al.*, 2000; Moore, 2003; Carlborg and Haley, 2004; Moore, 2005). Investigating such interactions may provide great insight into disease etiology and ultimately inform new strategies for treatment and prevention.

As an earlier attempt to explore the interaction between haplotypes and environmental factors, Lake *et al*. (2003) proposed a likelihood-based method in the generalized linear model framework, which has been widely used in haplotype-based association studies because it is available free and easy to implement with its R package. This approach, however, is limited by ignoring interacting effects between haplotype blocks. Subsequently, several methods have been developed to study haplotype-related interactions but these methods do not consider all potential haplotypes and interactions simultaneously (Lin *et al*., 2005; Spinka *et al*., 2005; Lin and Zeng, 2006; Kwee *et al*., 2007; Chen *et al*., 2008). Recently, Guo and Lin (2009) proposed a generalized linear model with regularization to detect interacting haplotype effects. However, their method applies a global test and consequently does not provide inference on the effects of individual haplotypes and their interactions.

In our literature review, we also found that little attention has been paid so far to developing statistical methods for exploring disease association with rare haplotypes. However, it has been argued that rare haplotypes may account for a substantial fraction of the multifactorial inheritance of common diseases (Liu *et al*., 2005; Zhu *et al*., 2005; Yende *et al*., 2007; Semsei *et al*., 2008; Kitsios and Zintzaras, 2010). Guo and Lin (2009) adopted a least absolute shrinkage and selection operator (LASSO) penalty in their model which allows assessment of the effects of rare haplotypes by shrinking the coefficients of unassociated haplotypes to zeros so that the associated ones, particularly those that are rare, can stand out. It is a quite attractive approach for precisely estimating the effects of rare haplotypes. However, since the distribution of LASSO estimators is non-standard, the pairwise comparisons between the tested haplotype and the reference haplotype are likely to suffer efficiency losses.

Detecting interacting haplotypes and rare haplotypes associated with disease is a big challenge to population-based association analysis in case-control studies. How to address these issues is the main topic of our research as described in CHAPTER 1, which motivates us to develop a new method as in the following section.

2.4 Bayesian Hierarchical Generalized Linear Model for Population-based Haplotype-association Analysis

2.4.1 Brief Description

We propose a new approach to investigate the association between haplotypes and human diseases based on the hierarchical generalized linear model. The proposed method is built upon a Bayesian framework with weakly informative priors on the coefficients. Although our method can be applied to continuous, binary, or ordinal traits, we herein describe it only for binary disease status in case-control studies. It can simultaneously fit a large number of effects, including main effects of numerous common and rare haplotypes, main effects of environmental factors, haplotype-haplotype interactions, and haplotype-environment interactions. We fit our Bayesian generalized linear models by incorporating an EM algorithm into the usual iteratively weighted least squares as implemented in the R package `glm`. This strategy leads to stable and flexible computational tools and allows us to apply any generalized linear model to haplotype-based association studies. We investigate the statistical properties and performance of the proposed method and compare it with three existing methods, the classical generalized linear model, the method of Lake *et al*. (2003), and the method of Guo and Lin (2009), through extensive simulation studies.

2.4.2 Methods

Generalized linear models of interacting haplotypes

Suppose that a population-based association study consists of *n* unrelated individuals, phenotyped for a disease trait, genotyped for multiple genetic variants (e.g., SNPs) in multiple genomic regions or haplotype blocks, and recorded for some non-genetic exposures, referred to as environmental factors. Although our method can deal with various phenotypes, we demonstrate its performance with a binary disease trait as measured in case-control studies. That is, let y_i denote the disease status of individual *i*, with $y_i = 1$ representing a case and $y_i = 0$ representing a control.

We use generalized linear models to relate disease status to haplotypes and environmental factors. A generalized linear model consists of three components: the linear predictor, the link function, and the distribution of the outcome variable (McCullagh and Nelder, 1989; Gelman *et al*., 2003). We simultaneously fit main effects of environmental (*E*) factors, main effects of haplotypes (*H*), haplotype-haplotype (*H*×*H*) and haplotype-environment (*H*×*E*) interactions. Therefore, the generalized linear model is expressed as

$$h(\Pr(y_i = 1)) = (\beta_0 + \boldsymbol{X}_E\boldsymbol{\beta}_E + \boldsymbol{X}_H\boldsymbol{\beta}_H + \boldsymbol{X}_{HH}\boldsymbol{\beta}_{HH} + \boldsymbol{X}_{HE}\boldsymbol{\beta}_{HE})_i\ @\boldsymbol{X}_i\boldsymbol{\beta},\ \ i = 1,\ ...,\ n, \qquad (2.1)$$

where *h* is a link function or transformation which relates the linear predictor $\boldsymbol{X}_i\boldsymbol{\beta}$ to the disease probability Pr(y_i = 1), β_0 is the intercept, $\boldsymbol{\beta}_E$ and $\boldsymbol{\beta}_H$ are the vectors of environmental effects and all possible haplotype main effects, respectively, $\boldsymbol{\beta}_{HH}$ is the vector of all possible haplotype-haplotype interactions between different haplotype blocks, and $\boldsymbol{\beta}_{HE}$ is the vector of haplotype-environment interactions, and $\boldsymbol{X}_E$, $\boldsymbol{X}_H$, $\boldsymbol{X}_{HH}$, and $\boldsymbol{X}_{HE}$ are the corresponding design matrices of explanatory variables. We describe the construction of these design matrices in the next subsection.

Various link functions are provided in generalized linear models (McCullagh and Nelder, 1989), all of which can be adapted in our Bayesian models. Wray and Goddard (2010) recommended using logistic or probit model for multi-locus analysis of genetic risk of disease in case-control studies. The *logit* transformation defines $h(p)$ = logit(p) = log(p/(1-p)), leading to a logistic regression which is commonly used in case-control studies and considered in our study.

Construction of the design matrices

Since usually haplotypes are not directly measured, we first compute the posterior probabilities of haplotype pairs based on the observed genotype data for each subject to account for this ambiguity by using existing methods of haplotype inference (e.g., Excoffier and Slatkin, 1995; Niu *et al.*, 2002; Stephens *et al*., 2001; Zaykin *et al*., 2004). These posterior probabilities are then used to compute the estimates of haplotype dosage (Stram *et al*., 2003).

The estimate of haplotype dosage is the estimate of the number of copies of a specific haplotype for a subject. For the haplotypes that can be unambiguously resolved based on the observed genotype data, the values of haplotype dosage of a haplotype for a subject can be zero (indicating that the haplotype is not possible based on the subject's genotypes), one (indicating heterozygosity for the haplotype based on the subject's genotypes), or two (indicating homozygosity for the haplotype). But for the haplotypes that cannot be unambiguously resolved, the values of haplotype dosage of a haplotype for a subject would be non-integer, ranging form zero to two, which reflect the possibility of the haplotype based on the subject’s genotypes. For each subject, the sum of haplotype dosage across all haplotypes within a haplotype block is equal to two. After obtaining the estimates of haplotype dosage, we can use them to construct the design matrix $\boldsymbol{X}_H$. We treat the estimate of haplotype dosage as a surrogate variable for the true haplotype.

Suppose there are W_q possible haplotypes in the qth haplotype block in the population, q = 1, 2, …, Q, and let d_{iqw}, w = 1, 2, …, W_q, denote the estimate of haplotype dosage of the wth haplotype in the qth haplotype block for subject i. Therefore, we can set $(\boldsymbol{X}_H)_i = \left(d_{i11},\ ...,\ d_{i1W_1},\ \mathrm{L}\ ,\ d_{iQ1},\ ...,\ d_{iQW_Q}\right)$. For example, unphased genotype data at two SNPs was observed, and, for a subject, two haplotype pairs, $\left(h_1, h_1\right)$ and $\left(h_2, h_4\right)$, were estimated with posterior probabilities, say, 0.9 and 0.1, respectively. Then the values of $\boldsymbol{X}_H$ for this subject are (1.8, 0.1, 0.0, 0.1).

Note that, at the time of statistical analysis, we exclude one haplotype from $\boldsymbol{X}_H$ to ensure identifiability of parameters in model fit.

For the environmental factors, the raw values are transformed to have a mean of 0 and a standard deviation of 0.5, by subtracting the mean and dividing by 2 × SD (the standard deviation of the raw values) (Gelman *et al*., 2008; Yi and Banerjee, 2009). This transformation standardizes all the environmental effects to have a common scale. The matrices of interacting variables, $\boldsymbol{X}_{HE}$ and $\boldsymbol{X}_{HH}$, are set up by simply multiplying two corresponding realizations of $\boldsymbol{X}_E$ and $\boldsymbol{X}_H$.

Prior and posterior distributions

The model above can include a large number of highly correlated explanatory variables, and most of which are likely to be zero or at least negligible, leading to the problems of high dimensionality, collinearity and sparsity that preclude the use of classical maximum likelihood methods. We handle these problems by using a Bayesian approach that places appropriate prior distributions on coefficients to capture the notion that most of the components of $\boldsymbol{\beta}$ probably approach to zero or can be at least ignored; such prior distributions are known as shrinkage priors (Gelman *et al*., 2003; Yi and Banerjee, 2009). We assume independent Student-*t* priors $t_{\nu_j}(0,\ s_j^2)$ on coefficients β_j, with ν_j and s_j chosen to give each coefficient a high probability of being near zero while still allowing for occasionally large effects (Gelman *et al*., 2003; Gelman *et al*., 2008; Yi and Banerjee, 2009). We are motivated to use the *t* distribution because it allows for flexible modeling, robust inference, and easy and stable computation (Gelman *et al*., 2008; Yi and Banerjee, 2009; Yi *et al*., 2010). There is no easy way to estimate coefficients directly using the *t* densities, but it is straightforward to deal with the two-level formulation of *t* distribution (Gelman *et al*., 2003; Gelman *et al*., 2008). The distribution $t_{\nu_j}(0,\ s_j^2)$ can be expressed as a mixture of normal distributions with mean 0 and variance distributed as scaled inverse-χ^2

$$\beta_j \mid \tau_j^2 \sim N(0,\ \tau_j^2)\,,\quad \tau_j^2 \sim \text{Inv-}\chi^2(\nu_j,\ s_j^2)\,,\quad j = 0,\ 1,\ \ldots,\ J\,, \tag{2.2}$$

where J is the total number of effects in the model, and the hyperparameters $\nu_j > 0$ and $s_j > 0$ represent the degrees of freedom and the scale of the distribution, respectively.

The hyperparameters ν_j and s_j control the global amount of shrinkage in the effect estimation; larger ν_j and smaller s_j^2 induce stronger shrinkage and force more effects to be near zero. We use the method of Yi *et al*. (2010) to choose ν_j and s_j. For β_0, $\boldsymbol{\beta}_E$ and $\boldsymbol{\beta}_H$, we use the weakly informative priors recommended by Gelman *et al*. (2008), i.e., $(\nu_0,\ s_0)=(1,\ 10)$ for β_0, and $(\nu_j,\ s_j)=(1,\ 2.5)$ for $\boldsymbol{\beta}_E$ and $\boldsymbol{\beta}_H$. For haplotype-environment interactions $\boldsymbol{\beta}_{HE}$, we set $(\nu_j,\ s_j)=(1,\ 2.5\times l_H / l_{HE})$, where l_H and l_{HE} are the total numbers of main effects of haplotypes and haplotype-environment interactions, respectively. For haplotype-haplotype interactions $\boldsymbol{\beta}_{HH}$, we set $(\nu_j,\ s_j)=(1,\ 2.5\times l_H / l_{HH})$, where l_{HH} are the total number of haplotype-haplotype interactions. Because there are many more interactions than main effects, these priors apply more stringent restrictions on interactions and allow reliable estimates of main effects and interactions (Yi *et al*., 2010).

With the above prior distributions, we can express the log-posterior distribution of the parameters $(\boldsymbol{\beta},\boldsymbol{\tau}^2)$ as

$$\begin{aligned}\log p(\boldsymbol{\beta},\boldsymbol{\tau}^2 \mid \boldsymbol{y}) &\propto \sum_{i=1}^{n}\log p(y_i \mid \boldsymbol{X}_i\boldsymbol{\beta})+\sum_{j=0}^{J}\log p(\beta_j \mid \tau_j^2)+\sum_{j=0}^{J}\log p(\tau_j^2 \mid \nu_j, s_j^2)\\ &\propto \sum_{i=1}^{n}\log p(y_i \mid \boldsymbol{X}_i\boldsymbol{\beta})-\frac{1}{2}\sum_{j=0}^{J}\left(\log\tau_j^2+\frac{\beta_j^2}{\tau_j^2}\right)+\sum_{j=0}^{J}\left(\frac{\nu_j}{2}\log s_j^2-(\frac{\nu_j}{2}+1)\log\tau_j^2-\frac{\nu_j s_j^2}{2\tau_j^2}\right)\end{aligned}, \quad (2.3)$$

where $\boldsymbol{\tau}^2=(\tau_0^2,\ldots,\tau_J^2)$, and the likelihood of $p(y_i \mid \boldsymbol{X}_i\boldsymbol{\beta})$ depends on the logit link function and the linear predictor that is defined in (2.1).

EM algorithm for model fit

Our hierarchical generalized linear model can be fitted using Markov chain Monte Carlo (MCMC) algorithms that fully explore the joint posterior distribution $p(\boldsymbol{\beta}, \boldsymbol{\tau}^2 \mid \boldsymbol{y})$ by alternatively sampling each parameter from its conditional posterior distribution. However, it is desirable to have a faster computation that provides a point estimate of coefficients, e.g., the posterior mode, and standard errors (and thus *p*-values). Such an approximate calculation has been routinely applied in statistical analysis (Gelman *et al.*, 2008).

We use the EM algorithm to fit the hierarchical haplotype models with the Student-*t* priors by estimating the marginal posterior modes of the coefficients β_j (Yi and Banerjee, 2009; Yi *et al.*, 2010). We incorporate our algorithm into the iteratively weighted least squares for classical generalized linear models as implemented in the R package `glm`, for example. The standard iteratively weighted least squares algorithm approximates a generalized linear model by a normal linear model (Gelman *et al.*, 2003; Gelman *et al.*, 2008). Specifically, at each iteration, pseudo-data z_i and pseudo-variances σ_i^2 are calculated for subject *i* by

$$z_i = \hat{\eta}_i - \frac{L'(y_i \mid \hat{\eta}_i)}{L''(y_i \mid \hat{\eta}_i)}, \quad \sigma_i^2 = -\frac{1}{L''(y_i \mid \hat{\eta}_i)}, \tag{2.4}$$

where $\hat{\eta}_i = \boldsymbol{X}_i \hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}}$ is the latest estimate of $\boldsymbol{\beta}$, $L'(y_i \mid \eta_i) = d \log p(y_i \mid \eta_i) / d\eta_i$, and $L''(y_i \mid \eta_i) = d^2 \log p(y_i \mid \eta_i) / d\eta_i^2$. Then the generalized linear model likelihood $p(y_i \mid \boldsymbol{X}_i \boldsymbol{\beta})$ is approximated by a normal likelihood $N(z_i \mid \boldsymbol{X}_i \boldsymbol{\beta}, \sigma_i^2)$, and finally the parameters β_j are updated by a weighted normal linear regression.

Our EM algorithm uses the two-level expression of the *t* prior distribution and treats the unknown variances τ_j^2 as missing data. From (2.3), we can see that only the terms $1/\tau_j^2$ are linked to β_j, so we need to calculate the expectation of $1/\tau_j^2$. It can be easily shown that the conditional posterior distribution of τ_j^2 is Inv-$\chi^2\left(1+\nu_j, \frac{\nu_j s_j^2+\hat{\beta}_j^2}{1+\nu_j}\right)$, and thus the conditional expectation of $1/\tau_j^2$ is equal to $\left(\frac{\nu_j s_j^2+\hat{\beta}_j^2}{1+\nu_j}\right)^{-1}$ (e.g., Yi and Xu, 2008). Therefore, the E-step of our EM algorithm is equivalent to replacing the variances by

$$\hat{\tau}_j^2 = \frac{\nu_j s_j^2+\hat{\beta}_j^2}{1+\nu_j}. \tag{2.5}$$

Given the variances τ_j^2, the priors $\beta_j \mid \tau_j^2 \sim N(0,\tau_j^2)$ can be treated as additional "data points", added to the weighted normal regression $N(z_i \mid \boldsymbol{X}_i\boldsymbol{\beta}, \sigma_i^2)$. Now we have an augmented weighted regression

$$\boldsymbol{z}_* \sim N(\boldsymbol{X}_*\boldsymbol{\beta},\ \Sigma_*), \tag{2.6}$$

where $\boldsymbol{z}_* = \begin{pmatrix}\boldsymbol{z}\\ \boldsymbol{0}\end{pmatrix}_{(n+J)\times 1}$ is a vector of all z_i and *J* zeros of all prior means, $\boldsymbol{X}_* = \begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{I}_J\end{pmatrix}_{(n+J)\times J}$ is a matrix constructed by concatenating the design matrix $\boldsymbol{X}$ of the regression $z_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma_i^2)$ with the identity matrix $\boldsymbol{I}_J$, and Σ_* is a diagonal matrix of all pseudo-variances σ_i^2 and prior variances $\hat{\tau}_j^2$. Then we can update $\boldsymbol{\beta}$ by performing this augmented weighed regression. Obviously, with the augmented design matrix $\boldsymbol{X}_*$, this regression is identifiable even if the original data are high-dimensional and have collinearity or separation (Gelman *et al*., 2008).

Thus, in each M-step, the standard iteratively weighted least squares algorithm is applied to the augmented weighted normal regression to estimate the coefficients β_j. We implement these computations by modifying the `glm` function in R for fitting generalized linear models, inserting the steps for calculating the augmented data and updating the variances into the iterative procedure.

The EM algorithm is initialized by setting each τ_j to a small value, say, τ_j = 0.1, and β_j to the starting value provided by the standard iteratively weighted least squares for the classical generalized linear model as implemented in the R function `glm`. We repeat the E-step and the M-step until convergence. At convergence of the algorithm, we obtain all outputs from the R function `glm`, including the estimates $\hat{\beta}_j$, standard errors, and *p*-values (for testing $\beta_j = 0$). The standard errors are calculated from the inverse second derivative matrix of the log-posterior density evaluated at $\hat{\beta}_j$ (Gelman *et al*. 2008). The *p*-values are then determined by the estimates of $\hat{\beta}_j$ and their standard errors as in the classical framework.

In summary, the algorithm starts with initial values for each τ_j^2 and β_j, and then proceeds as follows:

1) Based on the current values of β_j, calculate pseudo-data z_i and pseudo-variances σ_i^2;

2) E-step: replace each variance τ_j^2 by its conditional posterior expectation;

3) M-step: perform the weighted least square regression based on the normal likelihood approximation to obtain estimates $\hat{\beta}_j$;

4) Repeat steps 1-3 until convergence.

2.4.3 Simulation Study

We carry out an extensive simulation study to evaluate the statistical properties and performance of the proposed method. We utilize TGFBR1 haplotype-tagging SNP (htSNP) data published in the genetic association study that investigated the relationship between TGFBR1 haplotypes and risk of non-cell lung cancer (Lei *et al*., 2009). The six htSNPs are partitioned into two blocks, one forming 2-SNP haplotypes and the other forming 4-SNP haplotypes, based on the estimates of Lewontin coefficient ($D^{'}$) and squared correlation coefficient (r^2). The haplotype frequencies are estimated for the 2-SNP and 4-SNP haplotypes, respectively, and are presented in Table 2.1. Given these haplotype frequencies, we generate case and control subjects, assuming HWE for the haplotype pair of each individual and a logistic regression model for the disease risk. The baseline penetrance of disease (the proportion of affected subjects with a pair of non-disease-associated haplotypes) is set at 10%. A binary variable, smoking status with the proportion of 49% as in Lei *et al*. (2009), is included in the model as a covariate and is considered in haplotype-environment interactions. The results from the proposed method (referred to as BayesGLM) were compared with those from the classical generalized linear model (referred to as GLM), the method of Lake *et al*. (2003) (referred to as ScoreGLM), and the method of Guo and Lin (2009) (referred to as rGLM). The method of Lake *et al*. (2003) has been implemented in the freely available software R/haplo.stats (http://mayoresearch.mayo.edu/mayo/research/schaid_lab/software.cfm). Guo and Lin

(2009) also created an R package to carry out their method and it is available free at the website: http://www.stat.osu.edu/~statgen/SOFTWARE/rGLM/.

Table 2.1. Haplotype Patterns and Their Frequencies

4-SNP Haplotype			2-SNP Haplotype		
Haplotype	Pattern	Frequency	Haplotype	Pattern	Frequency
haplo4.1	1111	3.27×10^{-1}	haplo2.1	11	4.28×10^{-1}
haplo4.2	1112	2.71×10^{-2}	haplo2.2	12	3.03×10^{-1}
haplo4.3	1121	7.04×10^{-3}	haplo2.3	21	3.33×10^{-2}
haplo4.4	1211	6.64×10^{-2}	haplo2.4	22	2.36×10^{-1}
haplo4.5	1212	9.50×10^{-9}			
haplo4.6	1221	1.35×10^{-1}			
haplo4.7	1222	2.06×10^{-9}			
haplo4.8	2111	2.78×10^{-3}			
haplo4.9	2121	4.82×10^{-3}			
haplo4.10	2211	1.22×10^{-2}			
haplo4.11	2212	4.14×10^{-1}			
haplo4.12	2222	3.27×10^{-3}			

Note: From “A Bayesian hierarchical model for detecting haplotype-haplotype and haplotype-environment interactions in genetic association studies” by Jun Li, Kui Zhang, and Nengjun Yi, 2011, Human Heredity, 71, p. 151. Copyright 2011 by S. Karger AG, Basel. Reprinted with permission.

Simulation settings

Five scenarios were posed to carry out our evaluation processes. To examine whether the proposed method can be applied to both common and rare haplotypes, we considered a rare haplotype, *haplo4.3*, two moderately rare haplotypes, *haplo4.2* and *haplo4.4*, and a common haplotype, *haplo4.1*, in the 4-SNP haplotype block, and a moderately rare haplotype, *haplo2.3*, in the 2-SNP haplotype block to be associated with the disease in the five scenarios.

In the first two scenarios, we considered only the main effects of haplotypes arising from the 4-SNP haplotype block. Specifically, in the first scenario, we assumed that *haplo4.1* and *haplo4.3* increased the odds of getting disease by 2 and 3 fold, respectively, and *haplo4.2* and *haplo4.4* were not associated with the disease. In the second scenario, we assumed that *haplo4.1*, *haplo4.2*, *haplo4.3*, and *haplo4.4* increased the odds of getting disease by 2, 3, 4, and 3 fold, respectively, and none of the other eight haplotypes in the 4-SNP haplotype block were associated with the disease (Table 2.2).

In the third to fifth scenarios, we considered both the main and interacting effects arising between haplotypes in the two haplotype blocks, and between the haplotypes and smoking status. We assumed the effects in a similar way as we did in the first two scenarios (Table 2.2). But note that in the last scenario, we considered all the main effects of haplotypes and smoking status, and all possible interacting effects between the two haplotype blocks and between the haplotypes and smoking status. In this scenario there are a total of eighty-one terms, including seventeen marginal and sixty-four interacting terms (Table 2.2).

Each of these five scenarios had three different sample sizes: 250, 500, and 1000, with equal numbers of cases and controls. A total of 1000 replicates were generated under each of these fifteen settings. All of the generated data were analyzed by using ScoreGLM, GLM, rGLM, and BayesGLM, respectively.

Table 2.2. Marginal and Interacting Terms and Their Effects in the Five Scenarios

Scenario 1		Scenario 2		Scenario 3		Scenario 4		Scenario 5	
Term	OR	Term	OR	Term	OR	Term	OR	Term	OR
haplo4.3	3	haplo4.3	4	haplo2.3:haplo4.1	4	haplo2.3:haplo4.3 smoke:haplo4.3	5	haplo2.3:haplo4.3 smoke:haplo4.3	5
haplo4.1	2	haplo4.2 haplo4.4	3	haplo4.3 smoke:haplo4.1	3	haplo4.3 haplo2.3:haplo4.1 smoke:haplo2.3	4	haplo4.3 haplo2.3:haplo4.1 smoke:haplo2.3	4
haplo4.2 haplo4.4	1	haplo4.1	2	haplo2.3 haplo4.1 smoke	2	haplo2.3, haplo4.2 haplo4.4 smoke:haplo4.1	3	haplo2.3, haplo4.2, haplo4.4 smoke:haplo4.1	3
		haplo4.5, haplo4.6 haplo4.7, haplo4.8 haplo4.9, haplo4.10 haplo4.11, haplo4.12	1	haplo4.2 haplo4.4	1	haplo4.1 smoke	2	haplo4.1 smoke	2
						haplo2.1, haplo2.2 haplo2.4, haplo4.5 haplo4.6, haplo4.7 haplo4.8, haplo4.9 haplo4.10, haplo4.11 haplo4.12	1	Other seventy effects	1

“:” stands for an interaction between two terms (before and after “:”).

Note: From “A Bayesian hierarchical model for detecting haplotype-haplotype and haplotype-environment interactions in genetic association studies” by Jun Li, Kui Zhang, and Nengjun Yi, 2011, Human Heredity, 71, p. 152. Copyright 2011 by S. Karger AG, Basel. Reprinted with permission.

In summary, our procedure of the data generation, statistical analysis, and results comparison proceeded as follows:

1) *Genotype data generation:* Randomly drew two haplotypes (phased haplotype pairs) for each subject from the observed haplotypes (Table 2.1).

2) *Covariate data generation:* Smoking status for each subject was determined from a Bernoulli distribution with the observed proportion of smoking.

3) *Case/control data generation:* Set up the “true” values of parameters as described in the simulation settings. Using these “true” values as well as the generated phased haplotypes and smoking status, assigned an individual to be a case or control according to the probabilities derived from a classical logistic regression model.

4) *Model fit:* The generated phased haplotypes and smoking status were used as explanatory variables to fit four kinds of models based on ScoreGLM, GLM, rGLM, and BayesGLM, respectively.

5) *Replication:* the step 1 through the step 4 were repeated for 1000 times.

6) *Statistics calculation:* (1) Calculated 68% and 95% intervals that covered the “true” values for each parameter in the model: $|b_j - \hat{b}_j| < z_\alpha se_j$, where b_j is the “true” value of the *jth* parameter, $j = 1,\ 2,\ ...,\ J$, $\hat{b}_j$ is an estimated coefficient of the *jth* parameter, z_α is an upper critical value of the standard normal distribution for a desired significance level α, se is a standard error of estimated coefficients. (2) Calculated empirical powers for each of parameters in the model: $power = 1 \big/ R \sum_{r=1}^{R} I_{\left(p_{rj} \le \alpha\right)}$, where R is the number of replicates required, p_{rj} is the *p*-value of the *jth* parameter in the *rth* replicate, α is the significance level taking three values of 0.05, 0.01, or 0.001.

2.4.4 Results

Nonidentifiability of parameters in model fit

There was one main problem, the nonidentifiability of parameters, that was encountered in the model fit using the classical methods. This problem is first pointed out here because it frequently occurred and resulted in serious problems. Specifically, we found that the standard errors of some predictors in the models were large and hence the coefficients were essentially infinite when using `haplo.glm` in R/haplo.stats based on ScoreGLM or using `glm` in R based on GLM, whereas there was no such problem when using the proposed method, BayesGLM (data not shown). We could not evaluated the nonidentifiability of parameters when using `rGLM` because, as mentioned earlier, `rGLM` can only perform an overall test based on permutation and consequently does not provide standard errors for each predictor in the model fit.

The further question that might be asked is how often and how serious the problem is. To this end, we summarized the results regarding the nonidentifiability of parameters in the model fit for all of the simulation settings in Table 2.3. We can see that, as the sample size was increased, the proportions of nonidentifiability of parameters decreased in each of the first four scenarios of ScoreGLM and GLM. Under a fixed sample size, the proportions of nonidentifiability of parameters followed the order: scenario 5 > scenario 4 > scenario 2 > scenario 3 > scenario 1 for both ScoreGLM and GLM. In the scenarios 2, 4, and 5 of ScoreGLM and GLM, all of the proportions exceeded 50% except that in the scenario 2 of ScoreGLM with a sample size of 1000 (39%). In contrast, in the scenarios 1 and 3, only the proportion in the scenario 3 of GLM with a sample size of 250 barely exceeded 50% (51%). For BayesGLM, there was no problem observed with the nonidentifiability of parameters in all of the simulation settings. Obviously, the larger the proportions of nonidentifiability of parameters, the less stable the estimated coefficients (Albert and Anderson 1984; Lesaffre and Albert 1989). Therefore, our results involving comparisons of the three methods were derived only from the replicates without the nonidentifiability of parameters in the scenarios 1 and 3, unless otherwise specified.

Table 2.3. Proportions of Nonidentifiability of Parameters for All of the Simulation Settings

Sample size	Scenario	ScoreGLM	GLM	BayesGLM
250	1	0.34	0.47	0.00
	2	0.69	0.78	0.00
	3	0.42	0.51	0.00
	4	0.79	0.88	0.00
	5	1.00	1.00	0.00
500	1	0.22	0.33	0.00
	2	0.58	0.67	0.00
	3	0.30	0.38	0.00
	4	0.68	0.73	0.00
	5	1.00	1.00	0.00
1000	1	0.09	0.16	0.00
	2	0.39	0.54	0.00
	3	0.16	0.21	0.00
	4	0.56	0.61	0.00
	5	1.00	1.00	0.00

Note: From “A Bayesian hierarchical model for detecting haplotype-haplotype and haplotype-environment interactions in genetic association studies” by Jun Li, Kui Zhang, and Nengjun Yi, 2011, Human Heredity, 71, p. 153. Copyright 2011 by S. Karger AG, Basel. Reprinted with permission.

Main effect model

In the scenarios 1, only four haplotypes in the 4-SNP haplotype block were modeled as main effects for the disease (Table 2.2). The “true” values prespecified for these four haplotypes were first compared to their corresponding estimated coefficients based on the four methods (left column of Figure 2.1). Under the sample size of 250, wider estimated 68% and 95% intervals that covered the “true” values calculated based on BayesGLM were observed for each of four haplotypes compared to those calculated based on the other three methods, with the only exception that rGLM had little wider estimated intervals than BayesGLM did for *haplo4.1* (top left corner of Figure 2.1). With the increase of sample sizes, however, the superiority of reliability of BayesGLM was faded out for all of the haplotypes except *haplo4.3* (middle left and bottom left corner of Figure 2.1), although its two coverage rates maintained a low growth rate. For all of the four methods, *haplo4.3* had lower coverage than the other haplotypes did no matter what sample sizes were considered.

In this and the following subsections, we did not consider rGLM in the evaluation of empirical power as well as Type I errors because, as mentioned before, its omnibus test does not produce *p*-values for individual effects. Therefore, the empirical powers were calculated based only on ScoreGLM, GLM, and BayesGLM for *haplo4.1* and *haplo4.3*, from which we tried to evaluate the ability of these methods to detect any disease-predisposing haplotypes. Under the sample size of 250, BayesGLM demonstrated higher probabilities for detecting genetic effects compared to both ScoreGLM and GLM (top right corner of Figure 2.1). Although the advantage of BayesGLM in the statistical validity was diminishing with the increase of sample sizes, it still persisted, especially for the rare haplotype, *haplo4.3*, and for the powers under $\alpha = 0.001$ and 0.01 (middle right and bottom right corner of Figure 2.1). For all of the three methods, a sample size of 500 was sufficient to detect a common haplotype with power of 90% approximately, and a sample size of 1000 was sufficient to identify a rare haplotype with power of 85% approximately.

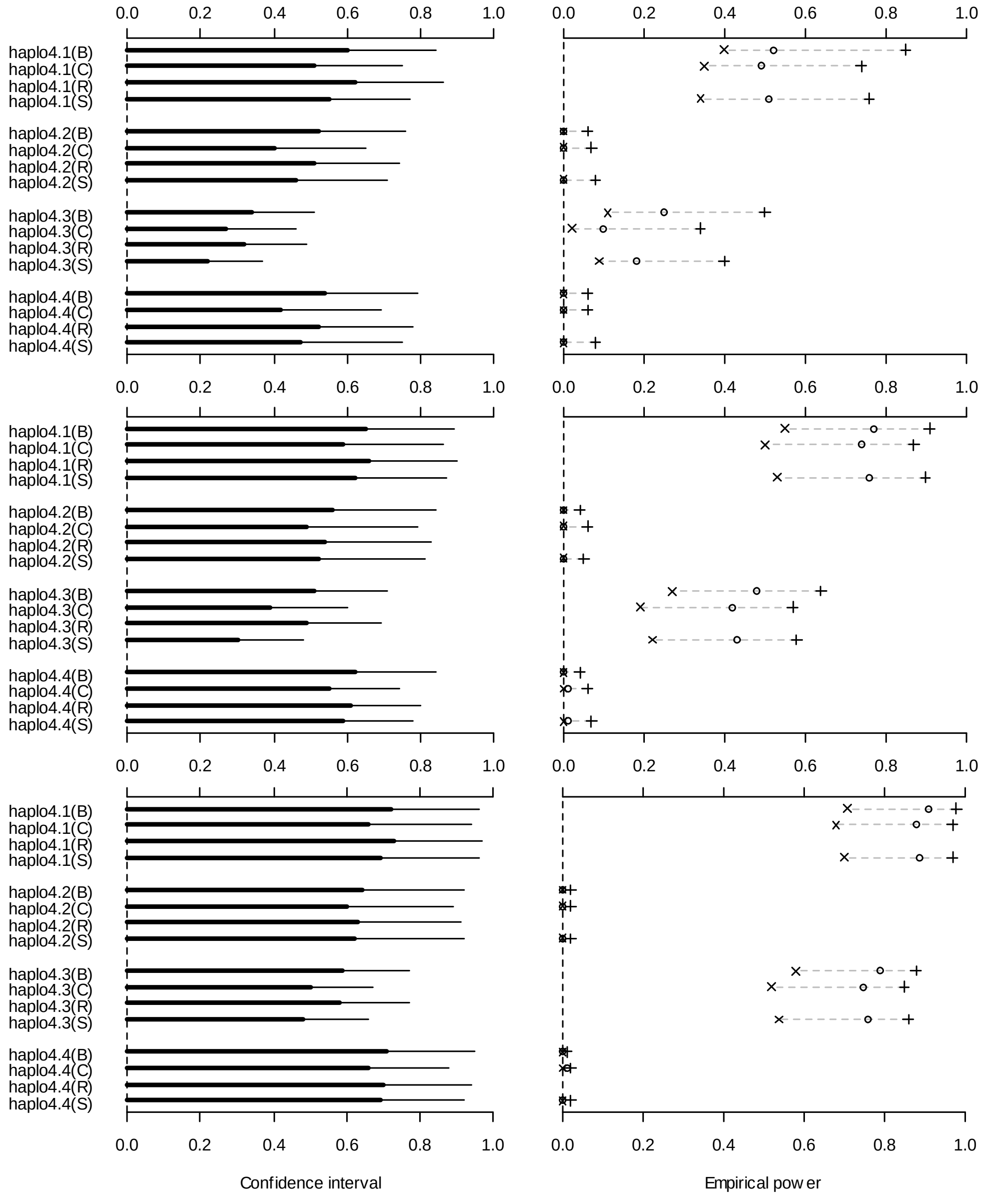


Figure 2.1. Main Effect Model. Estimated 68% and 95% coverages of the “true” values (indicated by bold and thin horizontal lines in the left column, respectively) and empirical powers or Type I error rates (× indicated the empirical powers or Type I error rates for $\alpha = 0.001$, o for $\alpha = 0.01$, and + for $\alpha = 0.05$) for each of four haplotypes based on the four methods under the sample sizes of 250 (top), 500 (middle), and 1000 (bottom). The notations, B, C, R, and S, stand for BayesGLM, GLM, rGLM, and ScoreGLM, respectively.

Note: From “A Bayesian hierarchical model for detecting haplotype-haplotype and haplotype-environment interactions in genetic association studies” by Jun Li, Kui Zhang, and Nengjun Yi, 2011, Human Heredity, 71, p. 154. Copyright 2011 by S. Karger AG, Basel. Reprinted with permission.

The empirical Type I error rates were also calculated for *haplo4.2* and *haplo4.4* based on ScoreGLM, GLM, and BayesGLM (right column of Figure 2.1). For the sample sizes of 250 and 500, BayesGLM had a little lower Type I error rates under $\alpha = 0.05$ than both ScoreGLM and GLM did. As the sample size went up to 1000, all of Type I error rates shrank to zero.

Main and interacting effect model

In the scenario 3, both the main and interacting effects arising between the two haplotype blocks and between the haplotypes and the environmental factor were jointly considered in the model fit for the four methods (Table 2.2). However, since *H*×*E* interactions cannot be fitted by using the current version of `rGLM`, interactions between *smoke* and *haplo4.1* were set only for ScoreGLM, GLM, and BayesGLM, and since *H*×*H* interactions cannot be fitted by using `haplo.glm` based on ScoreGLM, interactions between *haplo2.3* and *haplo4.1* were set only for GLM, rGLM, and BayesGLM. So there were total of eight terms as predictors included in the model with six of them assumed to be disease-associated (Figure 2.2). Under the sample size of 250, wider estimated 68% and 95% intervals that covered the “true” values calculated based on BayesGLM were found for each of eight predictors compared to those calculated based on the other three methods, with the only exception that rGLM had little wider estimated intervals than BayesGLM did for smoking status (*smoke*) and *haplo4.1* (top left corner of Figure 2.2). Although the lead of BayesGLM in the statistical reliability was narrowed with the increase of sample sizes, it continued to exist, especially for the rare haplotype, *haplo4.3*, and the interacting terms, *smoke*:*haplo4.1* and *haplo2.3*:*haplo4.1* (middle left and bottom left corner of Figure 2.2). For all of the four methods, the rare haplotype and the interacting terms had quite lower coverages than the other predictors in the model did no matter what sample sizes were considered, which was in agreement with the finding in the foregoing analysis of main effects.

The empirical powers were calculated for *smoke*, *haplo2.3*, *haplo4.1*, *haplo4.3*, *smoke*:*haplo4.1*, and *haplo2.3*:*haplo4.1* based on ScoreGLM, GLM, and BayesGLM. For *smoke*, the powers were comparable for ScoreGLM, GLM, and BayesGLM no matter what sample sizes were considered (top three lines in each of three right panels of Figure 2.2). This is reasonable because, for a common environmental factor with a decent frequency, any statistical test can achieve similar power for detecting it and the possible difference of powers among some tests can be explained by the random variability. For the predictors: *haplo2.3*, *haplo4.1*, and *haplo4.3*, the results were almost the same as those in the preceding subsection of main effects. For *smoke*:*haplo4.1*, under the sample size of 250, BayesGLM had higher power only for $\alpha = 0.05$ compared to ScoreGLM (top right corner of Figure 2.2). With the increase of sample sizes, however, the situation was soon improved and eventually turned around (middle right and bottom right corner of Figure 2.2). For *haplo2.3*:*haplo4.1*, BayesGLM demonstrated a higher probability for correctly detecting genetically interacting effects under each of three fixed Type I error rates and each of three sample sizes compared to both ScoreGLM and GLM (bottom two lines in each of three right panels of Figure 2.2).

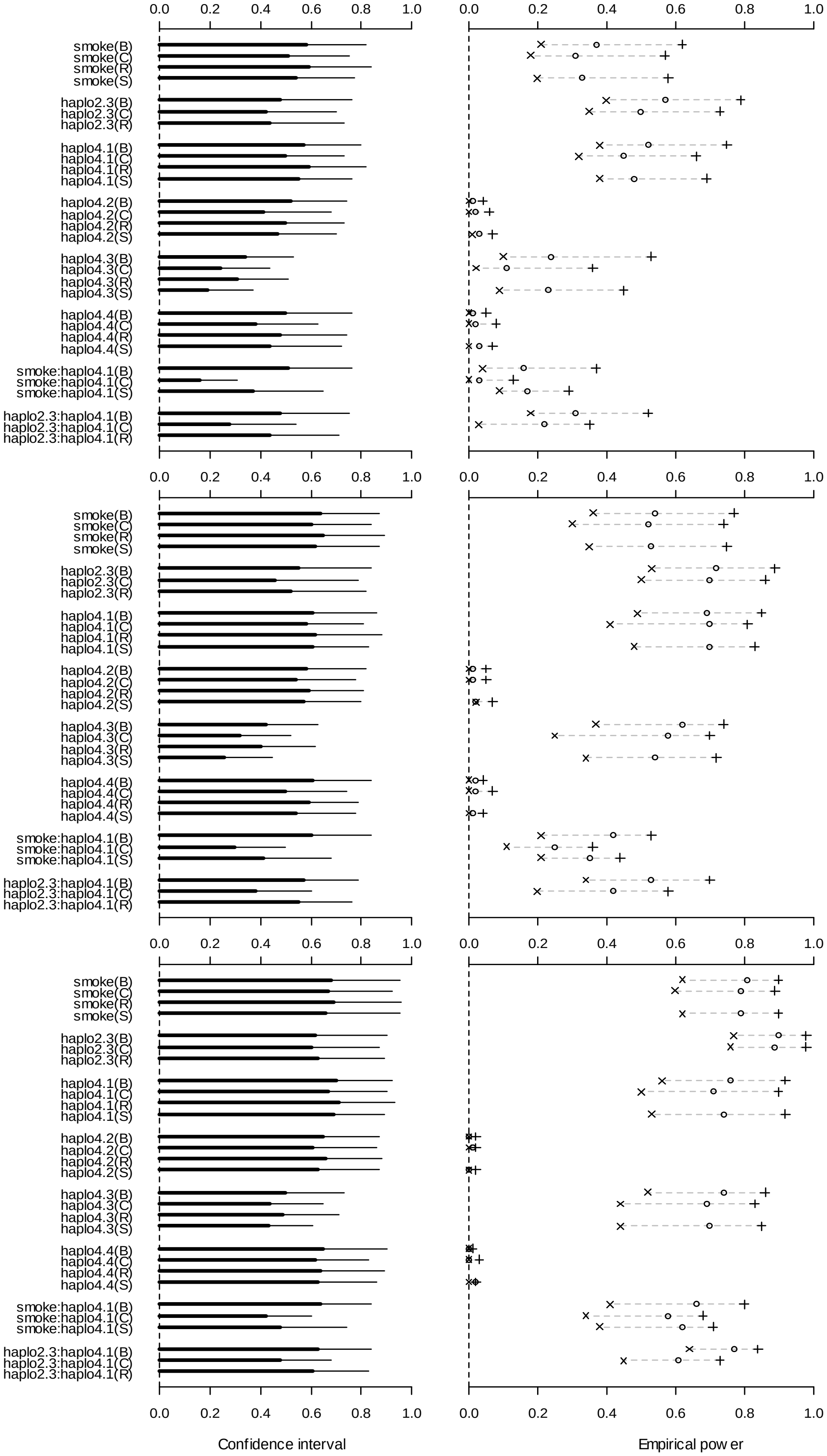


Figure 2.2. Main and Interacting Effect Model. Estimated 68% and 95% coverages of the “true” values (indicated by bold and thin horizontal lines in the left column,

respectively) and empirical powers or Type I error rates (× indicated the empirical powers or Type I error rates for $\alpha = 0.001$, o for $\alpha = 0.01$, and + for $\alpha = 0.05$) for each of four haplotypes based on the four methods under the sample sizes of 250 (top), 500 (middle), and 1000 (bottom). The notations, B, C, R, and S, stand for BayesGLM, GLM, rGLM, and ScoreGLM, respectively.

Note: From “A Bayesian hierarchical model for detecting haplotype-haplotype and haplotype-environment interactions in genetic association studies” by Jun Li, Kui Zhang, and Nengjun Yi, 2011, Human Heredity, 71, p. 155. Copyright 2011 by S. Karger AG, Basel. Reprinted with permission.

The empirical Type I error rates were also calculated for *haplo4.2* and *haplo4.4* based on ScoreGLM, GLM, and BayesGLM as in the preceding subsection of main effects, and the similar results were observed (right column of Figure 2.2).

Full model

In the scenario 5, a total of eighty-one marginal and interacting terms arising between the two haplotype blocks and between the haplotypes and the environmental factor were simultaneously considered (Table 2.2). As we have seen from Table 2.3, however, all the proportions of nonidentifiability of parameters were jumped to 1 for both ScoreGLM and GLM in the scenario 5. Consequently, the statistical estimations under these situations should be much instable and any comparison to them does not make sense. Since the current version of rGLM cannot fit *H*×*E* interactions, rGLM cannot be used to fit the full mode. Therefore, a single model based on BayesGLM was fitted to demonstrate its performance in a case where the number of predictors in a model is huge. As in the analyses of main and interacting effect models in the foregone subsections, the “true” values prespecified for all predictors in the model were first compared to their corresponding estimated coefficients for each of three sample sizes respectively (first, third, and fifth columns of Figure 2.3). From the graph we can see that, along with the increase of sample sizes, the estimated 68% and 95% intervals increased that covered the “true” values for each of eighty-one predictors. We also fund that the rare haplotypes (*haplo4.3* and *haplo2.3*) and the interactions (*smoke*:*haplo2.3*, *smoke*:*haplo4.1*, *smoke*:*haplo4.3*, *haplo2.3*:*haplo4.1*, and *haplo2.3*:*haplo4.3*) had quite lower coverages than the other predictors in the model did no matter what sample sizes were considered. All these findings were consistent with those observed in the foregoing subsections.

The empirical powers were calculated for a total of eleven disease-associated predictors in the model under each of three fixed Type I error rates ($\alpha = 0.001$, 0.01, and 0.05) (second, fourth, and sixth columns of Figure 2.3). From the graph we can see that although the power increased along with the increase of sample sizes, they started at quite low levels and maintained low growth rates. Under the sample size of 1000, eight predictors (*smoke*, *haplo2.3*, *haplo4.1*, *haplo4.3*, *smoke*:*haplo4.1*, *smoke*:*haplo4.3*, *haplo2.3*:*haplo4.1*, and *haplo2.3*:*haplo4.3*) had an 80% chance or more of being indentified under $\alpha = 0.05$, while three predictors (*haplo4.2*, *haplo4.4*, and *smoke*:*haplo2.3*) had a 60% chance or more of being indentified under $\alpha = 0.05$.

The empirical Type I error rates were also calculated for a total of seventy non-disease-associated predictors in the model. As the sample size went up to 500, almost all of the Type I error rates shrank to zero.

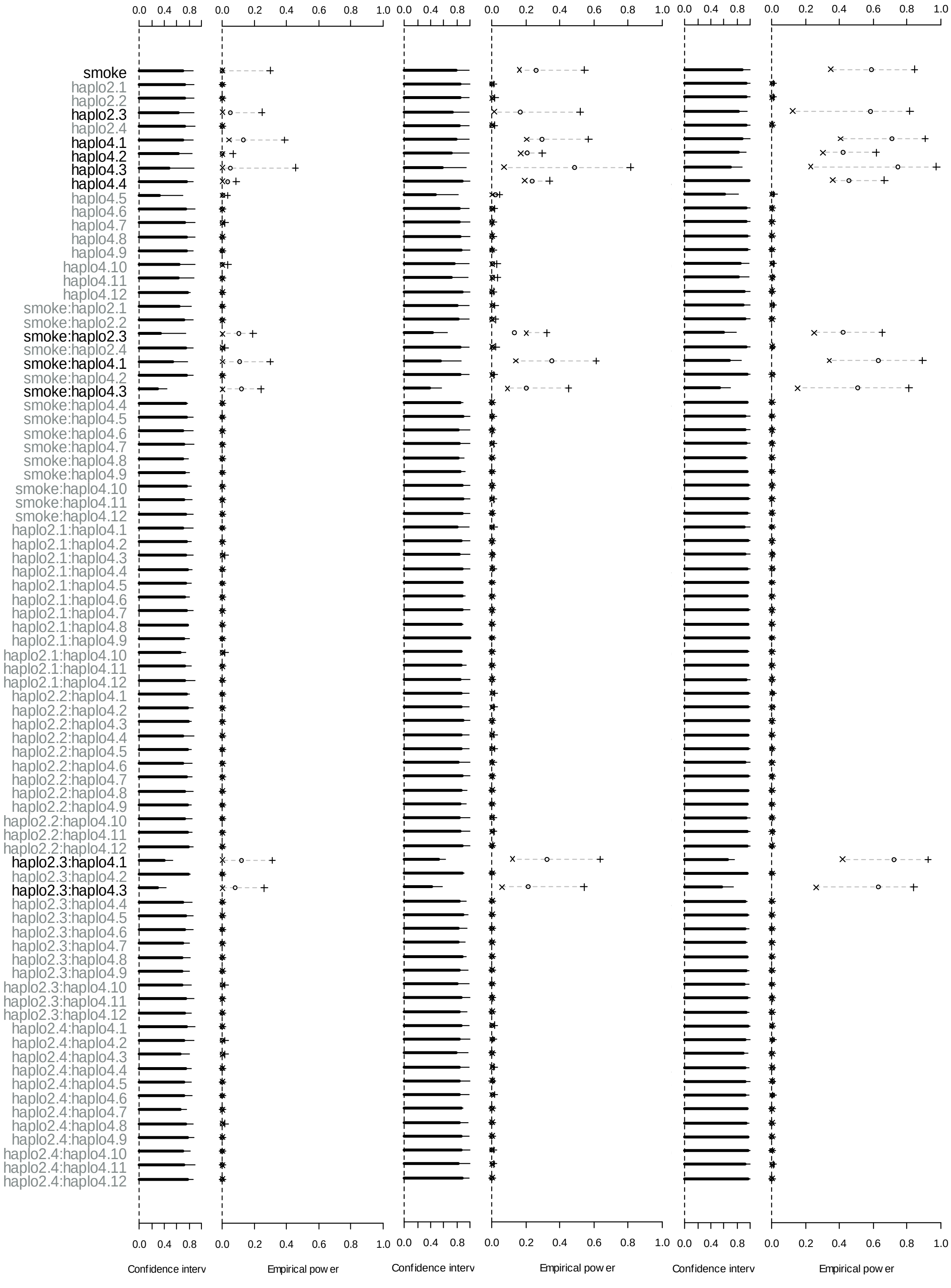


Figure 2.3. Full Model. Estimated 68% and 95% coverages of the "true" values (indicated by bold and thin horizontal lines in the first, third, and fifth columns, respectively) and empirical powers or Type I error rates (× indicated the empirical

powers or Type I error rates for $\alpha = 0.001$, o for $\alpha = 0.01$, and + for $\alpha = 0.05$) for each of eighty-one predictors based on BayesGLM under the sample sizes of 250 (first two columns), 500 (third and fourth columns), and 1000 (last two columns). The black labels on the vertical axis stand for the disease-associated predictors, while the gray labels stand for the non-disease-associated predictors.

Note: From “A Bayesian hierarchical model for detecting haplotype-haplotype and haplotype-environment interactions in genetic association studies” by Jun Li, Kui Zhang, and Nengjun Yi, 2011, Human Heredity, 71, p. 157. Copyright 2011 by S. Karger AG, Basel. Reprinted with permission.

2.4.5 Discussion

Complex human diseases are believed to be influenced by genetic and environmental factors, and their interactions. However, identifying interacting effects is challenging. In general, the identification and characterization of interactions are limited due to the lack of powerful statistical methods and/or large sample sizes. When numerous interactions are fitted explicitly in a model, the degrees of freedom for the corresponding test statistics would grow rapidly, and, as a result, sufficient power cannot be guaranteed to detect possibly significant effects in the model, especially in a relatively small sample size (Luan *et al*., 2001; Boks *et al*., 2007; Mukherjee *et al*., 2008; Cordell, 2009; Thomas, 2010). This issue is also confronted in haplotype-based association studies by classical methods, which usually has insufficient power and inflexibility to handle a large number of interactions (Lake *et al*., 2003; Becker *et al*., 2005; Kwee *et al*., 2007; Hein *et al*., 2009).

The challenges might be further aggravated when there are rare haplotypes present. Rare haplotypes can be seen frequently in genetic association studies and might be produced by common SNPs (Souverein *et al*., 2008; Guo and Lin, 2009). As already noted rare haplotypes, just like other genetic rare variants, could be important disease-predisposing variants and should not be ignored in exploring the genetic susceptibility with common diseases. Regarding statistical modeling, however, rare haplotypes can result in nonidentifiability of parameters, which means the coefficients of predictors cannot be identified or estimated uniquely because of huge, even infinite standard errors (Gelman *et al*., 2003). A commonly used approach to this issue in the literature is to pool all rare haplotypes into one single group (Schaid *et al*., 2002; Zhao *et al*., 2003) or pool rare haplotypes with common ancestral haplotypes (Seltman *et al*., 2003; Durrant *et al*., 2004; Tzeng, 2005). These approaches ignore rare haplotypes by lumping them together, and consequently any rare haplotype that might contribute to the risk of disease cannot be identified distinctly.

Statistical methods that can detect the haplotype-related interactions and handle the nonidentifiability of parameters are much needed area of research. In the present study, we propose a Bayesian hierarchical generalized linear model with weakly informative priors to simultaneously analyze a large number of effects, including main effects of common and rare haplotypes, environmental effects, and their all possible interactions. Our model fitting algorithm takes advantage of the classical generalized linear model procedure, leading to a computationally stable tool. An extensive simulation study was conducted to evaluate the statistical properties and performance of the proposed method, and the results were compared with the classical generalized linear model, the method of Lake *et al*. (2003), and the method of Guo and Lin (2009). The main reason for considering these three methods as reference is that the classical generalized linear model is a flexible and basic approach to analyze case-control data, the method of Lake *et al*. (2003) is the commonly used method for haplotype-based analysis in association studies, and the method of Guo and Lin (2009) takes account of both rare haplotypes and haplotype interactions between two haplotype blocks.

In our simulation study, the identifiability of parameters in model fit was first assessed because it is a common problem in the conventional methods. The results show that, for ScoreGLM and GLM, the estimates of coefficients were substantially nonidentifiable in most of the simulation settings, while for BayesGLM, the nonidentifiability of parameters was not observed. This demonstrates the appealing features of the proposed method in terms of robustness of parameter estimation and efficiency of statistical computation over the existing methods, especially in the case that has a large number of interactions and some rare haplotypes in the model.

With respect to the statistical properties of the proposed method, statistical power is our primary interest in the evaluation processes. The results indicate that the proposed method outperforms ScoreGLM and GLM in terms of statistical power for detecting associations, especially for rare haplotypes and interactions with the moderate sample sizes. However, with the increase of number of predictors fitted in the model, the proposed method had a relative loss of power, but still acceptable (Figure 2.3). This is reasonable because, as we already know, the high dimensionality is traded with loss of power in model fit.

The reliability of the proposed method concerning parameter estimation was examined by comparing the “true” values prespecified for the predictors in the models to their corresponding estimated coefficients. The proposed method can yield better coverage of confidence interval, especially for the interactions and the rare haplotypes, than ScoreGLM and GLM (Figure 2.1 and Figure 2.2). But, at most of time, the proposed method unsurprisingly has similar results to rGLM (Figure 2.1 and Figure 2.2). However, the proposed method provides more features than rGLM in its current implementation. Moreover, the proposed method has been implemented in our R package `BhGLM` and is available to practitioners (http://www.ssg.uab.edu/bhglm/).

CHAPTER 3

FAMILY-BASED HAPLOTYPE-ASSOCIATION ANALYSIS IN MATCHED CASE-CONTROL STUDIES

Of particular concern in genetic association analysis is the potential confounding resulting from incomparable ethnic backgrounds across subpopulations being compared; such confounding creates the motivation for family-based tests of association. Along with the advances in genomic science and the collective efforts of statistical genetics, family-based association analysis has gained in popularity for mapping disease-susceptibility genes of complex human diseases (Risch and Merikangas, 1996; Khoury and Yang, 1998; Umbach and Weinberg, 2000; Cordell *et al*., 2004; Chatterjee *et al*., 2005; Weinberg Lange *et al*., 2008). Haplotypes, as very important genetic variants, have been extensively studied in family-based association analysis, and they play a crucial role in the gene mapping due to their functional and statistical advantages over their counterparts based on SNPs (e.g., Schaid, 2004; Kraft *et al*., 2005; Levenstien *et al*., 2006).

The main purpose of this chapter is to introduce a new method that employs Bayesian hierarchical generalized linear model to detect haplotype-haplotype and haplotype-environment interactions, particularly involving rare haplotypes, using family data. We begin in Section 3.1 with a succinct description of family-based haplotype-association analysis as well as its advantage and disadvantage relative to population-based association analysis, and then in Section 3.2 we discuss the main features of matched case-control studies and its relationship with family-based case-control studies.

We also provide a short review on existing statistical methods for family-based haplotype-association analysis in Section 3.3. Finally, we present our new method in Section 3.4.

3.1 Brief Description of Family-Based Haplotype-Association Analysis

Family-based haplotype-association analysis includes a broad range of methods that aim to investigate the association of haplotypes with measures of disease progression or disease status, employing information derived from family samples.

The distinct property of family-based association analysis is that controls are selected from within the same families as cases. In family-based designs, nuclear families are most commonly considered which are composed of two parents and a number of full siblings. Sometimes subsets of nuclear families such as sib pairs or single parent are used. Extended pedigrees including, e.g., cousins may also be employed in family-based association analysis.

An advantage of choosing family members as controls is that they are ethnically matched and they can also share lifestyle, life experiences, or some socioeconomic factors. Therefore, family-based association analysis is immune to the notorious confounding due to population stratification that usually occurs in population-based association analysis (see Subsection 2.1.3 for more details). The properties of matching in both family-based and population-based designs are further discussed in Subsections 3.2.8 and 3.2.9. However, it is noteworthy that although family-based designs offer the advantage of robustness against genetic heterogeneity, this feature comes at the price of reduced statistical power when compared with population-based designs because the genetic similarity of cases and controls lessens power (Laird and Lange, 2009; Thomas, 2010b). In general, the relationship of family controls to the affected cases serves to reduce the difference of distributions of genetic variants under study between two comparison groups. Furthermore, family-based designs may be less powerful than population-based case-control designs. However, the difference between these two designs is generally small, particularly when family trio data (consisting of one affected offspring and two parents per family) is used (Witte *et al*., 1999, McGinnis *et al*., 2002).

Another advantage of selecting family members as controls is that family-based association analysis is potentially more efficient for estimating gene-environment interactions, particularly, when rare genetic variants are involved, relative to population-based association analysis (Witte *et al*., 1999; Gauderman, 2002), and more useful for detecting gene-gene interactions (MacLean *et al*., 1993; Zhao *et al*., 2006). Here the efficiency means that family-based designs generally require fewer matched sets than population-based case-control designs to achieve the same power for detecting a gene-environment interaction.

In addition, significant findings in family-based association analysis indicate both linkage and association between marker loci and disease-susceptibility loci. However, there are some disadvantages of family-based designs arising from practical matters of recruitment and cost; it is usually difficult and expensive to recruit a large number of families because, e.g., family members may not live together and be hard to reach or may refuse to participate; and it is even impossible to recruit parents of the affected subjects for late-onset diseases in which both parents may be deceased.

One should note that in the most of family-based genetic studies, we also need to infer haplotype frequencies based on the observed genotype data because the information of haplotype phase is usually unavailable that is the same as in population-based genetic studies. But haplotype inference based on family data is more reliable than that based on population data because family data can provide additional constraints that help us phase family members based on Mendelian law (Zhang and Zhao, 2006; Li and Li, 2007). However, this superiority is traded with more laboratory work to genotype additional family members. For example, for case-parent trio design, we need to genotype at least three people in a family to obtain required data.

Since family-based and population-based association analyses have different advantages and disadvantages, most contemporary genetic association studies take the view that the two designs are strongly complementary in the effort to unravel the genetic mechanisms that are underlying complex human diseases.

We herein briefly discuss the strength and weakness of family-based haplotype-association analysis with compared to population-based association analyses. For a comprehensive and in-depth discussion on these topics, we could see Gauderman *et al.*, 1999; Risch, 2000; McGinnis *et al.*, 2002; Cardon and Palmer, 2003; Laird and Lange, 2006; Dudbridge, 2007; Liu *et al.*, 2008; Zhang and Zhao, 2010.

3.2 Matched Case-Control Studies

In this section, the most important aspects of matched case-control studies are reviewed prior to discussing how family-based haplotype-association analysis fits within matched case-control studies. This overview is by no means exhaustive. There is a lot of literature devoted to the topics of matched case-control studies (e.g., Breslow and Day, 1980; Schlesselman, 1982; Costanza, 1995; Rothman *et al*., 2008a,b,c) and their application in family-based genetic association analysis (e.g., Thomas, 2004; Ziegler and Koenig, 2007; Clayton, 2007).

3.2.1 Concept and Principles of Matching in Case-Control Studies

Matching is an intuitively attractive strategy in study design for ensuring balance on one or more potential confounding factors between two comparison groups. In a case-control study, if controls are selected to match cases on some potential confounders, such a design is then called a matched case-control study. Here matching means that the controls have the same or similar values of the matching variables as the cases.

In general, we match to make sure that the two groups being compared are similar with respect to confounding factors that might distort a relationship under investigation. To fix ideas, consider a study conducted to explore the possible effect of cigarette smoking exposure on the risk of lung cancer. It is known that older age increases the risk of lung cancer and that older people are more likely to be smokers than younger ones. Age, therefore, is a probable confounding factor in the relationship between the cigarette smoking and the lung cancer. In this example, we could match the cases and controls on the similar age, e.g., to within five years, to eliminate any age difference between the cases and the controls. If, after matching in this way, we then observe an association between the cigarette smoking and the lung cancer, we would know that we could not attribute the association to the age difference.

3.2.2 Benefits of Matching in Case-Control Studies

Note that here we do not say that the process of matching itself can control confounding, but we say instead that the process of matching forces cases and controls to have similar distributions across confounding factors. This is because there has been a lot of debate within the scientific community about the purpose of matching in case-control studies (e.g., Breslow and Day, 1980; Kupper *et al*., 1981; Schlesselman, 1982; Rothman *et al*., 2008a). In the earlier publications, matching is usually described as a way to control confounding effects in case-control studies (e.g., Miettinen, 1970; Breslow *et al*., 1978). However, in the later literature, particularly very recently, the opinion seems to prevail that while matching is intended to reduce confounding effects, it cannot attain that objective in case-control studies. Matched case-control studies can only enhance the efficiency of study by balancing on some potential confounders between cases and controls (e.g., Breslow and Day, 1980; Kupper *et al*., 1981; Schlesselman, 1982; Costanza, 1995; Rothman *et al*., 2008a). Furthermore, Rothman *et al*. (2008a,b,c) said that the process of matching can introduce bias in case-control studies sometimes. In their opinion, matching in case-control studies can not prevent confounding effects directly, but it can make the stratified analysis more efficient. Occasionally, stratification and/or multivariable analysis is still necessary to control bias and confounding left after matching. It must be said, however, that they also pointed out that matching is desirable or even necessary in some situations. This is an issue we will revisit in Subsection 3.2.7.

3.2.3 Types of Controls in Matched Case-Controls

Matching can be broadly divided into the two categories of individual matching and frequency matching in case-control design. Individual matching means that one or more controls are selected for each individual case by matching variable(s). Individual matching can be implemented in various ways, including one case to one control (1:1 or pair matching), one case to two or more controls (1:*m* or triplets, quadruplets, ..., matching), or many cases to many controls (*n*:*m* matching), where *n* or *m* is a varying number of cases or controls in the matched sets, but usually the ratio of case to control is 1:5 because little statistical power is gained by further increasing this ratio (Rothman *et al*., 2008a). 1:1 matching is the most common situation in both genetic and non-genetic such as clinical researches, particularly when cases and controls cost the same.

Frequency matching is also called group or category matching, which means that controls are selected to ensure that the frequency of a matching variable is the same as found in cases, e.g., if 5% of cases are under age 35, 5% of controls are also. Since individual matching is most commonly seen in genetic association analysis, all discussion will henceforth focus on it, unless otherwise specified.

3.2.4 Selection of Matching Variables and Overmatching

Matching may be by gender, age, race, and some other established confounding variables. What a particular variable will be considered as a confounder in a study is usually determined by examining the relationship of the variable with the disease and the exposure under investigation. The methods of ascertaining whether a variable is a confounder have been well established through earlier epidemiologic studies and can be seen in the literature cited at the beginning of this section.

Variables for matching should be selected carefully, and only those that are known to be a true confounder in advance should be taken into account. If cases and controls are matched on a variable that is not a confounder, such matching can impact the efficiency or validity of study. For example, if a matching variable is associated with exposure but not associated with disease, the matching will result in a large number of exposure concordant case-control pairs, such pairs of subjects do not contribute any information to the statistical analysis (matching analysis depends only on exposure discordant case-control pairs of subjects). This matching reduces the statistical efficiency relative to an unmatched design (Kupper *et al*., 1981; Thomas and Greenland, 1983). In addition, if a matching variable is an intermediate step in the casual pathway from exposure to disease (the variable is affected by the exposure, and it in turn affects the disease), then the crude and adjusted effect estimates will be biased. In fact, a casual intermediate is not a confounder of exposure-disease association; it is part of exposure effect that we wish to study. This matching harms the validity of study (Greenland and Neutra, 1981). Moreover, controls may be selected from neighbors or friends of each case when cost and convenience are first considered. In this case, the method for recruiting controls automatically entails matching; we are in effect matching for socioeconomic status, cultural and lifestyle characteristics, or some other characteristics of a neighbor or a friend. As a result, these matched characteristics could no longer be investigated in the study. Thus this matching impacts cost efficiency (Rothman *et al*., 2008a).

Essentially, these examples above show that the factor is matched which is not a confounder of exposure-disease association. This phenomenon is usually known as overmatching in epidemiology. From these examples, we can see that overmatching can abrogates the main virtue of matching and it is irreversible. Therefore, one must use the technique of matching wisely and carefully.

3.2.5 Number of Variables for Matching

A practical concern with matching is that how many factors we should consider to match on in a study. In general, determining the number of factors for matching depends on practical consideration and the extent to which we care to establish close comparability. Basically, the more factors we choose to match on at a time the more impossible and expensive it is to find such a control. Additionally, as the number of matching factors increases, the cases and controls will become more and more similar with regards to the exposure being studied and the study may yield a false result or provide no information (Breslow and Day, 1980). The number of matching factors should therefore be reduced to as few as possible in a study.

3.2.6 Maintenance of Matching in Statistical Analysis

When matching is carried out in a case-control study, pairing as formed initially needs to be maintained throughout the study including the stage of statistical analysis. This means that we should perform relevant analysis such as stratification and/or multivariable analysis for matched data (Rothman *et al*. 2008a,b). If unmatching analysis is implemented on matched data, the analysis may drive the estimate of OR even closer towards unity (Schlesselman, 1982; Jewell, 2003; Rothman *et al*. 2008a,b). However, Breslow and Day (1980) noted that unmatching analysis could yield approximately valid results for matched data when the numbers of both cases and controls within a stratum are large and stratum-specific intercepts are included in the logistic regression model. When the numbers of both cases and controls within a stratum are small or the number of matching strata is large, the conditional analysis would be preferred. In fact, whenever possible, it would always be preferred for matched data to perform conditional analysis.

One must also keep in mind that matching usually complicates statistical analysis when he or she plans a matched case-control study (Rothman *et al*. 2008a,b).

3.2.7 Desirable Situations for Matching

Since matched case-control studies have some weakness such as complication of design and statistical analysis, cost of finding matched controls, and possible overmatching, one may well ask whether matching is ever justified. Unfortunately, the procedure of deciding whether to match in a case-control study is not always so clear. However, some studies demonstrated that the matched case-control design is essentially required in some situations. For example, when the effect of a confounder needs controlling but the confounder is not easily measured (e.g., Jablon *et al*., 1967; Costanza, 1995). In this case the best thing that we would like to do is to ensure that the distributions of cases and controls are similar on the confounder, so that the occurrence of disease is more likely to be attributable to the exposure, not to the confounder. This situation is very common in genetic association studies where we need to control some genetic characteristics such as population stratification that is usually unknown across comparison groups. This is the topic we discuss in the next subsection.

Rothman *et al*. (2008a) also mentioned another situation in which matching is extremely valuable. In the situation the information of exposure and confounding is expensive to obtain from the subjects, so the efficient way to get more information is maximizing the amount of information obtained per subject by individual matching of subjects, rather than spending the same money on recruiting more subjects.

3.2.8 Family-Based Case-Control Studies

Matched case-control studies are frequently found in the literature which investigated family-based genetic association with complex human diseases. In family-based genetic association analysis, cases are compared to their healthy relatives, typically their healthy siblings (e.g., Curtis, 1997; Spielman and Ewens, 1998), cousins (e.g., Witte *et al*., 1999), parents, spouses (e.g., Valle *et al*., 1998; Li and Boehnke, 2006), etc. Each of these controls establishes similarity with the cases on one or more characteristics. For example, sibling or parent controls share with the cases the whole or half of genetic materials and early life experiences; cousin controls share with the cases part of genetic materials and even some life experiences; spouse controls share with the cases household characteristics, lifestyle, nutrition, and some socioeconomic factors. Thus, what become clear from these similarities is that family-based genetic association analysis automatically entails matching. Consequently, family-based case-control studies can be thought of as a special case of matched case-control studies (e.g., Hsu *et al*., 2007; Chatterjee *et al*., 2005; Martin, 2006; Bernardinelli *et al*., 2007).

The most important property of family-based case-control studies is that the internal matching within a family guarantees that the cases and controls originate from the same homogeneous, including ethnically homogeneous, source population. This property offers complete robustness against population stratification and truly motivates us to use family controls in genetic association studies (e.g., Self *et al*., 1991; Ewens and Spielman, 1995; Witte *et al*., 1999).

3.2.9 Population-Based Matched Case-Control Studies

One should note that matched case-control studies can also be implemented in population-based genetic association analysis. The process of matching in population-base case-control studies is quite straightforward, i.e., when you select controls from unrelated individuals in a population, you just make them matched with cases on some variables which we may be concerned. Several methods have been proposed for population-based matched case-control studies in the context of haplotype-based association analysis (e.g., Lee, 2004; Kraft *et al.*, 2005; Zhang *et al.*, 2006; Zhang *et al.*, 2007; Chen and Rodriguez, 2007). The main reason that these investigators selected the controls from unrelated individuals not from relatives in their studies is that they aimed at avoiding the difficulty of recruiting family controls. Some, but not all, of investigators explicitly described the matching scheme in their studies and argued that the possible confounding induced by population stratification was controlled by matching the controls with the cases on some variables such as ethnicity, race, nationality, and ancestry.

Among all of the variables above, from the genetic perspective, ethnicity is the main source of the confounding (Risch, 2000). If controls are not comparable to cases with respect to ethnicity, there will be a difference of allele frequency at a locus between the two groups. Based on the self-reported information of ethnicity in a population-based study, the ethnicity may not fully specify the complex nature of fine-scale genetic structure within the population, in other words, these subjects are descended from different origins, although they asserts they have the same ethnicity (Sinha *et al.*, 2008). In addition, the information of ancestry is usually unavailable, and even when available, it may not reflect the genetic architecture of a population because of the inexplicit definition of ancestry groups (Guan *et al.*, 2009). The information of race may be even less reliable. Even within sibships, the disagreement often occurs concerning the original countries of their parents (Hahn *et al.*, 1996).

Most genetic association studies take the view that population-based matched case-control studies cannot completely eradicate the confounding effect of population stratification while family-based case-control studies can offer well protection against the spurious results. Therefore, only for robustness against population stratification, family-based case-control studies may be preferred to population-based matched case-control studies in genetic association investigation. By the way, since our primary interest lies in family-based haplotype-association analysis in the matched case-control studies, this dissertation will not address more matching in population-based genetic association studies.

3.3 Existing Statistical Methods for Family-Based Haplotype-Association Analysis

In this section we briefly review the relevant literature on family-based haplotype-association analysis, in order to demonstrate gaps in available methods, thereby creating a rationale for new methods.

In family-based association analysis, the traditional but still popular method is transmission/disequilibrium test (TDT). Originally, TDT was developed for analysis of the transmission of alleles from parents to affected offspring (Spielman *et al.*, 1993). For a biallelic marker, comparisons are made within parent-offspring trios to discern the similarities or differences between the number of heterozygous parents who transmit one allele and the number of heterozygous parents who transmit another allele to the affected offspring. TDT is a landmark in the development of family-based association analysis, and has some advantages such as simplicity for implementation and robustness to potential spurious association results caused by population stratification (Liu *et al.*, 2008; Zhang and Zhao, 2010). Nevertheless, there are many situations in which the original TDT cannot be applied directly, for instance, quantitative phenotypes, missing parents, general pedigrees, multi-allelic loci, and haplotypes with missing phase (Laird and Lange, 2006). To increase power and generalizability, various extensions have been developed based on the original TDT (Laird and Lange, 2006; Liu *et al.*, 2008; Zhang and Zhao, 2010).

One of the useful extensions is to employ a conditional likelihood function in analysis. Recalling that TDT is very close to a matching analysis that compares transmitted and non-transmitted alleles within parent-offspring trios, the standard approach to account for this ascertainment effect from epidemiology would be to carry out a matched case-control analysis by treating the transmission as a response variable and the alleles as predictors in a conditional logistic regression (Waldman *et al.*, 1999). For multi-allelic loci, however, there will be numerous parameters which affect the efficiency of analysis. For avoidance of this problem, an extension in the same direction has been further developed by Schaid (1996), Cordell and Clayton (2002) and others. In this extension, an offspring genotype is modeled as a function of parental genotypes and offspring disease status in a conditional logistic regression. As already noted, this method is equivalent to a classical matched case-control analysis in which the method is to regard the analysis not in terms of transmission from parents to offspring, but rather in terms of comparing a case (the affected offspring's genotype) to pseudocontrols. Pseudocontrols are so-called because they are formed from the other three genotypes that could have been transmitted from parents except the affected offspring's genotype (Self *et al*., 1991; Schaid, 1996).

This conditional logistic regression extension of TDT possesses all of the desirable features of the traditional TDT, and many advantages associated with the conventional regression analysis. In particular, this kind of extension can take full advantage of the well-established algorithms and software developed primarily for classical regression analysis.

Another approach for analyzing family-based association is to simply treat each family as a matched set and employ a conditional logistic regression to model the relationship between disease and some genetic and/or environmental factors as in the usual matched case-control analysis (e.g., Goldstein *et al*., 1989; Andrieu and Goldstein, 1996; Witte *et al*., 1999; Kraft and Thomas, 2000; Siegmund *et al*., 2000). Through an intensive simulation study and the evaluation of asymptotic expectation, Witte *et al*. (1999) have shown that the estimates based on a conditional likelihood function are unbiased for many types of family member controls such as sibling controls.

The conditional likelihood function for logistic regression was given in Cox (1970) and applied to matched case-control analysis in Breslow and Day (1980), Breslow (1982) and others. Conditional logistic regression has been commonly used and well-studied for estimating relative risks in matched case-control studies (Rothman *et al*., 2008a,b). In family-based case-control studies, since we introduce a family stratum with each case, we must include such stratum effects in the model. However, the number of parameters in the model increases just as fast as the total sample size. In this circumstance, if we use usual logistic regression, the asymptotic properties of likelihood inference will break down, while if use conditional logistic regression, the parameters expressing the stratum effects are eliminated from the likelihood by use of a conditional argument rather than by attempting to estimate them (Breslow and Day, 1980).

There is a significant volume of literature on various applications of conditional likelihood in family-based haplotype-association analysis (e.g., Clayton, 1999; Zhao *et al*., 2000; Dudbridge, 2003; Horvath *et al*., 2004; Cordell *et al*., 2004; Allen and Satten, 2007; Vansteelandt *et al*., 2008). Out of these applications, Clayton (1999) utilized a full likelihood function conditional on offspring's disease status to generate a TDT test and first tried to deal with the problem of phase uncertainty for multilocus haplotypes. Dudbridge (2003) applied a similar likelihood function as that of Clayton (2003) but introduced an EM algorithm to maximize the likelihood function under both the null hypothesis and the alternative hypothesis in the presence of ambiguous haplotypes. Horvath *et al*. (2004) proposed a weighted conditional approach which is an extension of the family-based association test (FBAT) originally developed by Rabinowitz and Laird (2000). This approach can examine both linkage and association between multiple loci and disease even when the haplotype phase may be ambiguous and the parental genotype data may be missing. Allen and Satten (2007) developed a method based on the projection conditional on parental haplotypes which is different from the general FBAT. The approach is robust to misspecification of the parental-genotype distribution and hence robust to population stratification. The authors also showed that their approach has improved power relative to the FBAT approach of Horvath *et al*. (2004). Given their methodological strengths and easy implementation with the help of program packages, some of these methods have been successfully employed in many studies (Liu *et al*., 2008).

However, despite the rapid improvement of family-based haplotype-association analysis to date, there are still many limitations in methodologies such as how to test and estimate haplotype-haplotype and haplotype-environment interactions, and how to identify rare haplotypes. A few researchers in the literature have tried the investigation of haplotype-related interactions in the area. For instance, Allen and Satten (2007) and Vansteelandt *et al*. (2008) presented separate methods that allow for testing interactions between haplotypes and environmental factors. Cordell *et al*. (2004) described a unified approach to analyze both haplotype-haplotype and haplotype-environment interactions in the nuclear families. Nevertheless, due to the relatively complicated structure of family data and the presence of haplotype phase uncertainty, extensions of many well-established approaches to family-based haplotype-association analysis are not straightforward and have not yet been implemented (Purcell *et al*., 2005). Furthermore, the difficulties become much greater when attempting to take into account of rare haplotypes in family-based association analysis. Though some work has been done in the context of population-based haplotype-association analysis (see CHAPTER 2 for more details), to the best of our knowledge, little is known about how to identify the effects of rare haplotypes in family-based association analysis.

Thus, the main subject of this dissertation research is to propose a new method that can accommodate both haplotype-related interactions and rare haplotypes but remain robust and computationally efficient to the high dimensionality and sparsity of data.

3.4 Bayesian Hierarchical Generalized Linear Model for Family-Based Haplotype-Association Analysis

3.4.1 Brief Description

In our previous report (Li *et al.*, 2011, and also see Section 2.4 for more details), we described a unified approach for conducting haplotype-based association analysis with quantitative traits (usually disease status) in the sample drawn from unrelated individuals. The approach is built on Bayesian hierarchical generalized linear model which allows for simultaneously analyzing the main effects of haplotypes and environmental factors as well as their interactions. An extensive simulation study shows that our approach outperforms the existing methods in terms of statistical power of indentifying disease risk factors and computational efficiency. However, it cannot be applied to family data directly. To take full advantage of family data such as immunity to population stratification (e.g., Self *et al*., 1991; Ewens and Spielman, 1995; Witte *et al*., 1999); higher power in detecting rare variants associated with a particular disease (Manolio *et al.*, 2009; Zhu *et al.*, 2010; Feng *et al.*, 2011) and superiority in efficient for estimating gene-environment and gene-gene interactions, particularly, when rare genetic variants are involved (MacLean *et al*., 1993; Witte *et al.*, 1999; Gauderman, 2002; Zhao *et al*., 2006), relative to population-based association analysis (Witte *et al*., 1999; Gauderman, 2002); and robustness to some other confounders depending on matching scheme and matched variables (see Subsection 3.2 for more details), we would like to extend our method to the context of family-based association studies.

We propose a modified conditional likelihood approach for inferring haplotype-related association with disease in family-based case-control studies, where controls are cases' relatives within a family and individually matched to the cases on some factors.

Although our method can be readily applied to continuous, binary, or ordinal traits, we herein describe it only for binary disease status in family-based case-control studies. We only assume that the genotypes of SNPs are available in our research, and hence we need to estimate haplotype frequencies and assign possible haplotype pairs to subjects that they might carry based on their observed genotypes using some existing methods (e.g., Excoffier and Slatkin, 1995; Niu *et al.*, 2002; Stephens *et al.*, 2001; Zaykin *et al.*, 2004). We utilize a logistic regression model to characterize the relationship between haplotypes and disease, and the model is fitted in a Bayesian framework with weakly informative priors on the coefficients. The model can simultaneously fit a large number of effects, including main effects of numerous common and rare haplotypes, main effects of environmental factors, haplotype-haplotype interactions, and haplotype-environment interactions. To facilitate the process of model fit in presence of high dimensionality and rare haplotypes, we create a fast and stable algorithm by incorporating an EM algorithm into the usual iteratively weighted least squares as implemented in the R package `glm`. We investigate the statistical properties and performance of the proposed method and compare it with the existing methods through an extensive simulation study.

3.4.2 Methods

Data structure, notation and model

Assume that there are a total of n case-control strata (e.g., families or sibling sets), and n_i cases and m_i controls in the ith stratum, where $n_i > 1$ and $m_i > 1$. For the jth individual in the ith stratum, we observe the vector of explanatory variables $\boldsymbol{X}_{ij}$, including haplotypes, environmental factors, haplotype-haplotype and haplotype-environment interactions. Denote the disease status by y_{ij} for the jth individual in the ith stratum, with

y_{ij} being 1 or 0 for case or control, respectively. The relationship between the disease risk and the explanatory variables can be modeled using a logistic regression of the following form:

$$\Pr(y_{ij}=1)=\text{logit}^{-1}(\alpha_i+\boldsymbol{X}_{ij}\boldsymbol{\beta}),\quad i=1,\ldots,n;\ \ j=1,\ldots,n_i+m_i, \tag{3.1}$$

where the α_i is the stratum-specific effect for the *i*th matched set, and $\boldsymbol{\beta}$ is a vector of haplotype main effects, environmental effects, and all possible interacting effects (see Subsection 2.4.2 for more details).

Construction of the design matrices

We use the same way to construct the design matrices for family-based haplotype-association analysis as those for population-based haplotype-association analysis (see Subsection 2.4.2 for more details).

Conditional likelihoods for matched case-control studies

A naïve method to estimate the parameters $\boldsymbol{\beta},\ \alpha_1,\ldots,\alpha_n$ in (3.1) is to directly use the logistic regression approach. Although simple, this method could be problematic because it does not take account of the ascertainment, i.e., each set includes at least one case and one control. The commonly used approach to this issue is the conditional logistic regression (Breslow and Day, 1980). Without loss of generality, for the *i*th matched set, we assume that the first n_i individuals are cases. The conditional likelihood for relative risk parameters $\boldsymbol{\beta}$ is

$$L(\boldsymbol{\beta},\boldsymbol{\alpha})=\prod_{i=1}^{n}\Pr(y_{i1}=\ldots=y_{in_i}=1,\ y_{i(n_i+1)}=\ldots=y_{i(n_i+m_i)}=0\,|\,\textstyle\sum_{j=1}^{n_i+m_i}y_{ij}=n_i)$$

$$=\prod_{i=1}^{n}\frac{\prod_{j=1}^{n_i}\Pr(y_{ij}=1)\prod_{j=n_i+1}^{n_i+m_i}\Pr(y_{ij}=0)}{\Pr(\sum_{j=1}^{n_i+m_i}y_{ij}=n_i)}=\prod_{i=1}^{n}\frac{\prod_{j=1}^{n_i}\Pr(y_{ij}=1)\prod_{j=n_i+1}^{n_i+m_i}\Pr(y_{ij}=0)}{\sum_{l=1}^{L}\prod_{s=1}^{n_i}\Pr\left(y_{il_s}=1\right)\prod_{s=n_i+1}^{n_i+m_i}\Pr\left(y_{il_s}=0\right)}$$

$$= \prod_{i=1}^{n} \frac{\prod_{j=1}^{n_i} e^{\alpha_i + \boldsymbol{X}_{ij}\boldsymbol{\beta}}}{\sum_{l=1}^{L} \prod_{s=1}^{n_i} e^{\alpha_i + \boldsymbol{X}_{il_s}\boldsymbol{\beta}}} = \prod_{i=1}^{n} \frac{\prod_{j=1}^{n_i} e^{\boldsymbol{X}_{ij}\boldsymbol{\beta}}}{\sum_{l=1}^{L} \prod_{s=1}^{n_i} e^{\boldsymbol{X}_{il_s}\boldsymbol{\beta}}}, \tag{3.2}$$

where the summation in the denominator is over L terms, each of which involves a permutation of n_i possible cases from any of $n_i + m_i$ individuals in the ith stratum. Note that the stratum-specific effects α_i defined in (3.1) have been eliminated in the conditional likelihood and thus will not be estimated.

The conditional likelihood is complicated to fit for case-control studies with sets consisting of more than one case. Usually cases are rare but controls are readily available. Therefore, most case-control studies consist of a single case and one or multiple controls for each or most of matched sets. Hereafter we refer such matched case-control data to as 1:*m* design. Statistical procedures for this kind of design can be much simplified. Suppose that we have *n* matched case-control sets, with 1 case and m_i controls for the *i*th stratum. The conditional likelihood is simplified to

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \Pr(y_{i1} = 1,\ y_{i2} = \cdots = y_{i(1+m_i)} = 0 \mid \sum_{j=1}^{1+m_i} y_{ij} = 1) = \prod_{i=1}^{n} \frac{e^{\boldsymbol{X}_{i1}\boldsymbol{\beta}}}{\sum_{j=1}^{1+m_i} e^{\boldsymbol{X}_{ij}\boldsymbol{\beta}}}, \tag{3.3}$$

which is equivalent to the multinomial logistic model

$$y_i = (y_{i1}, \cdots, y_{i(1+m_i)}) \sim \text{Multin}(1,\ 0,\ ...,\ 0;\ \alpha_{i1},\ \cdots,\ \alpha_{i(1+m_i)}),\ i = 1,\ \cdots,\ n \tag{3.4}$$

with $y_{i1} = 1,\ y_{i2} = \cdots = y_{i(1+m_i)} = 0$, and $\alpha_{ij} = \frac{e^{\boldsymbol{X}_{ij}\boldsymbol{\beta}}}{\sum_{l=1}^{1+m_i} e^{\boldsymbol{X}_{il}\boldsymbol{\beta}}},\quad j = 1,\ \cdots,\ 1 + m_i$.

For data with a small to moderate number of variables, the above conditional likelihood can be handled directly using multinomial logistic procedure. However, it may be more efficient to use the Poisson equivalence, often referred to as Multinomial-Poisson transformation (e.g., Baker, 1994; Gelman *et al*., 2003). As this method is useful in performing computations, we describe it here and extend it to handle high-dimensional models as in the analysis of multiple interacting genes. The conditional likelihood (3.3) can be re-expressed as

$$L(\boldsymbol{\beta}) \propto \prod_{i=1}^{n} \prod_{j=1}^{1+m_i} \left(\exp(\lambda_i + \boldsymbol{X}_{ij}\boldsymbol{\beta}) \right)^{y_{ij}} e^{-\exp(\lambda_i + \boldsymbol{X}_{ij}\boldsymbol{\beta})} \tag{3.5}$$

with $\lambda_i = -\log \sum_{l=1}^{1+m_i} e^{\boldsymbol{X}_{il}\boldsymbol{\beta}}$. This relation allows us to analyze 1:*m* case-control data using the Poisson generalized linear model

$$y_{ij} \sim \text{Poisson}(\mu_{ij}), \;\; \log \mu_{ij} = \lambda_i + \boldsymbol{X}_{ij}\boldsymbol{\beta} \cong \eta_{ij}, \; i = 1, \ldots, n; \; j = 1, \ldots, 1 + m_i \, . \tag{3.6}$$

The simplest matched case-control data consists of a single control per case for each set, which is a special case of the 1:*m* design and thus can be analyzed as above. However, a simpler analysis is to use an unconditional logistic regression, since the conditional likelihood can be expressed as

$$\prod_{i=1}^{n} \Pr(y_{i1} = 1, \, y_{i2} = 0 \,|\, y_{i1} + y_{i2} = 1) = \prod_{i=1}^{n} \frac{e^{\boldsymbol{X}_{i1}\boldsymbol{\beta}}}{e^{\boldsymbol{X}_{i1}\boldsymbol{\beta}} + e^{\boldsymbol{X}_{i2}\boldsymbol{\beta}}} = \prod_{i=1}^{n} \frac{e^{(\boldsymbol{X}_{i1} - \boldsymbol{X}_{i2})\boldsymbol{\beta}}}{1 + e^{(\boldsymbol{X}_{i1} - \boldsymbol{X}_{i2})\boldsymbol{\beta}}} \, , \tag{3.7}$$

which can be obtained using the unconditional logistic regression

$$\Pr(y_{i1} = 1) = \text{logit}^{-1}\left((\boldsymbol{X}_{i1} - \boldsymbol{X}_{i2})\boldsymbol{\beta} \right), \; i = 1, \ldots, n \tag{3.8}$$

with all responses being 1 and no intercept term in the model.

Prior and posterior distributions

Association analysis is equivalent to estimating parameters $\boldsymbol{\beta}$ in the above model. The number of parameters in the model can be large and the predictors can be highly correlated, which preclude the use of classical maximum likelihood methods. We solve this problem by placing prior distributions on $\boldsymbol{\beta}$ to capture the notion that most of the components of $\boldsymbol{\beta}$ are likely to be zero or at least negligible; such prior distributions are often referred to as shrinkage priors.

We assume independent Student-*t* priors $t_{\nu_k}(0,\ s_k^2)$ on parameters β_k, with ν_k and s_k chosen to give each parameter a high probability of being near zero while still allowing for occasionally large effects. We are motivated to use the *t* distribution since it can produce robust inference, shrinkage estimation, and easy computation (Gelman *et al*., 2008; Yi and Xu, 2008; Yi and Banerjee, 2009). There is no easy way to estimate parameters directly using the *t* densities, but it is straightforward to deal with the two-level formulation of *t* distribution (Gelman, *et al*., 2003; Gelman *et al*., 2008). The *t* distribution $t_{\nu_k}(0,\ s_k^2)$ can be expressed as a mixture of normal distributions with mean 0 and variance distributed as scaled inverse-χ^2

$$\beta_k \mid \tau_k^2 \sim N(0,\ \tau_k^2)\,,\quad \tau_k^2 \sim \text{Inv-}\chi^2(\nu_k,\ s_k^2)\,,\quad k=1,\ \ldots\ ,K\,, \tag{3.9}$$

where K is the number of the parameters, and the hyperparameters $\nu_k > 0$ and $s_k > 0$ represent the degree of freedom and the scale of the distribution, respectively.

The priors (3.9) introduce parameter-specific variances, resulting in distinct shrinkage for different parameters. A small value of τ_k^2 will force β_k close to zero. The variances τ_k^2 are not the parameters of interest, but they are useful intermediate quantities to make the computation easy and efficient. The hyperparameters ν_k and s_k affect the amount of shrinkage in the parameter estimates and should be carefully chosen. Our algorithm highlights how these hyperparameters affect the estimates of the parameters.

With the above prior distributions, we can express the log-posterior distribution of the parameters $(\boldsymbol{\beta}, \boldsymbol{\tau}^2)$ as

$$\log p(\boldsymbol{\beta}, \boldsymbol{\tau}^2 \mid \boldsymbol{y}, \boldsymbol{X}) \propto \sum_{i=1}^{n} \sum_{j=1}^{1+m_i} \log p(y_{ij} \mid \eta_{ij}) + \sum_{k=1}^{K} \log p(\beta_k \mid \tau_k^2) + \sum_{k=1}^{K} \log p(\tau_k^2 \mid \nu_k, s_k^2)$$

$$\propto \sum_{i=1}^{n} \sum_{j=1}^{1+m_i} \log p(y_{ij} \mid \eta_{ij}) - \frac{1}{2} \sum_{k=1}^{K} \left(\log \tau_k^2 + \frac{\beta_k^2}{\tau_k^2} \right) + \sum_{k=1}^{K} \left(\frac{\nu_k}{2} \log s_k^2 - (\frac{\nu_k}{2} + 1) \log \tau_k^2 - \frac{\nu_k s_k^2}{2\tau_k^2} \right), \quad (3.10)$$

where $\eta_{ij} = \lambda_i + \boldsymbol{X}_{ij}\boldsymbol{\beta}$, and $p(y_{ij} \mid \eta_{ij})$ is the Poisson likelihood function defined in (3.6).

Model fit algorithm

Before describing our model fitting algorithm, we show two ways to handle the Poisson regression (3.6). The first way takes the relation $\lambda_i = -\log \sum_{l=1}^{1+m_i} e^{\boldsymbol{X}_{il}\boldsymbol{\beta}}$ and hence only includes the parameters $\boldsymbol{\beta}$ in the model, while the second method treats λ_i as additional parameters with uniform priors and estimates them along with $\boldsymbol{\beta}$. Since the maximum likelihood estimate of $\mu_{ij} (= e^{\lambda_i + \boldsymbol{X}_{ij}\boldsymbol{\beta}})$ is y_{ij}, the conditional maximum likelihood estimate of λ_i equals $-\log \sum_{l=1}^{1+m_i} e^{\boldsymbol{X}_{il}\boldsymbol{\beta}}$. Thus these two methods could produce identical estimates. Although the second method can be directly implemented with the Poisson procedure, it is computationally intensive when there are many matched sets and thus improper to be applied to models with many variables. We develop our method for identifying interacting genes based upon the first way. Our computational idea is to treat λ_i as constants (i.e., offset in the terminology of generalized linear models) when updating $\boldsymbol{\beta}$. This method could be as fast as that for a Poisson regression without these nuisance parameters.

Estimating the posterior mode

We extend the Bayesian generalized linear models and the iterative model fitting algorithm developed by Yi and Banerjee (2009) to our Poisson model for matched case-control studies. The procedure of Yi and Banerjee (2009) fits generalized linear models with the Student-*t* priors by incorporating an EM algorithm into the standard iteratively weighted least squares (IWLS) as implemented in the R routine `glm`.

The IWLS algorithm approximates a generalized linear model by a normal likelihood and updates parameters from the weighted normal linear regression (Gelman *et al.*, 2003). At each iteration, we construct pseudo-data z_{ij} and pseudo-variances σ_{ij}^2 for each individual based on the latest estimates of $\boldsymbol{\beta}$ and λ_i as follows

$$z_{ij} = \hat{\eta}_{ij} - \frac{L'(y_{ij} \mid \hat{\eta}_{ij})}{L''(y_{ij} \mid \hat{\eta}_{ij})}, \ \ \sigma_{ij}^2 = -\frac{1}{L''(y_{ij} \mid \hat{\eta}_{ij})}, \ \ i = 1, \ \mathrm{L} \ , n; \ j = 1, \ \mathrm{L} \ , \ 1 + m_i, \qquad (3.11)$$

where $\hat{\eta}_{ij} = \hat{\lambda}_i + \boldsymbol{X}_{ij}\hat{\boldsymbol{\beta}}$, $\hat{\lambda}_i = -\log \sum_{l=1}^{1+m_i} e^{\boldsymbol{X}_{il}\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}}$ is the latest estimate of $\boldsymbol{\beta}$, $L'(y_{ij} \mid \eta_{ij}) = d \log p(y_{ij} \mid \eta_{ij}) / d\eta_{ij}$, $L''(y_{ij} \mid \eta_{ij}) = d^2 \log p(y_{ij} \mid \eta_{ij}) / d\eta_{ij}{}^2$, and $p(y_{ij} \mid \eta_{ij})$ is the Poisson likelihood defined in (3.6). The Poisson likelihood is approximated by the weighted normal likelihood

$$z_{ij} \sim N(\hat{\lambda}_i + \boldsymbol{X}_{ij}\boldsymbol{\beta}, \ \sigma_{ij}^2) \qquad (3.12)$$

so that under the classical framework (i.e., with uniform priors) $\boldsymbol{\beta}$ can be easily updated from this normal linear regression.

Under our Bayesian model, we update $\boldsymbol{\beta}$ from the model: $z_{ij} \sim N(\hat{\lambda}_i + \boldsymbol{X}_{ij}\boldsymbol{\beta}, \ \sigma_{ij}^2)$, $\beta_k \mid \hat{\tau}_k^2 \sim N(0, \ \hat{\tau}_k^2)$, conditional on the latest estimates $\hat{\tau}_k^2$ and $\hat{\lambda}_i$. By treating the *K* prior means as additional data points with residual variances $\hat{\tau}_k^2$, this two-level model can be re-expressed as an augmented weighted regression

$$\boldsymbol{z}_* \sim N(\boldsymbol{X}_*\boldsymbol{\beta}, \ \boldsymbol{\Sigma}_*), \qquad (3.13)$$

where $\boldsymbol{z}_* = \begin{pmatrix} \boldsymbol{z} - \hat{\boldsymbol{\lambda}} \\ \boldsymbol{0} \end{pmatrix}$ is the vector of all $z_{ij} - \hat{\lambda}_i$ and K prior means 0, $\boldsymbol{X}_* = \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{I}_K \end{pmatrix}$ is constructed by the design matrix $\boldsymbol{X}$ of the regression $z_{ij} - \hat{\lambda}_i \sim N(\boldsymbol{X}_{ij}\boldsymbol{\beta},\ \sigma_{ij}^2)$ and the $K \times K$ identity matrix $\boldsymbol{I}_K$, and $\boldsymbol{\Sigma}_*$ is the diagonal matrix of all pseudo-variances σ_{ij}^2 and K prior variances $\hat{\tau}_k^2$. Thus, we can update $\boldsymbol{\beta}$ by performing this augmented weighted regression.

As in Yi and Banerjee (2009), we treat the unknown variances $\boldsymbol{\tau}^2 = (\tau_1^2,\ \ldots,\ \tau_K^2)$ as missing data and average over them by replacing the terms involving both $\boldsymbol{\tau}^2$ and $\boldsymbol{\beta}$ in the posterior distribution (3.10) by their expected values conditional on the latest estimate $\hat{\boldsymbol{\beta}}$. Since the conditional posterior distributions of τ_k^2 is $\text{Inv-}\chi^2\left(1+\nu_k,\ \dfrac{\nu_k s_k^2 + \hat{\beta}_k^2}{1+\nu_k}\right)$, the conditional expectations of $1/\tau_k^2$ equals $\left(\dfrac{\nu_k s_k^2 + \hat{\beta}_k^2}{1+\nu_k}\right)^{-1}$. Therefore, we update the variances by

$$\hat{\tau}_k^2 = \frac{\nu_k s_k^2 + \hat{\beta}_k^2}{1+\nu_k}\ ,\quad k = 1,\ \ldots,\ K\ . \tag{3.14}$$

We initialize the algorithm by setting each τ_k to a small value, say, τ_k = 0.1, and β_k to the starting value provided by the `glm` function. At each step of our EM algorithm, we average over the variances $\boldsymbol{\tau}^2 = (\tau_1^2,\ \ldots,\ \tau_K^2)$ and then update $\boldsymbol{\beta}$ by maximizing the posterior density (3.10). In summary, our algorithm proceeds as follows

1) Based on the current value of $\boldsymbol{\beta}$, set each λ_i to be $-\log\sum_{l=1}^{1+m_i} e^{\boldsymbol{X}_{il}\boldsymbol{\beta}}$;

2) Calculate pseudo-data z_{ij} and pseudo-variances σ_{ij}^2 using (3.11);

3) E-step: replace each variance τ_k^2 by its conditional expectation using (3.14);

4) M-step: determine the augmented weighted normal linear model (3.13) and run this regression to obtain the estimate $\hat{\boldsymbol{\beta}}$;

5) Repeat steps 1 - 4 until convergence.

We apply the criterion in the `glm` function to assess convergence. we obtain all of the outputs produced by the `glm` function, including the latest estimate $\hat{\boldsymbol{\beta}}$, their standard errors and *p*-values (for testing $\hat{\beta}_k = 0$), and some additional values (e.g., for the variances).

Standard error correction

The standard errors for the parameter estimates $\boldsymbol{\beta}$ are underestimated because the fitting algorithm treats λ_i as known in the last iteration of the estimation of $\boldsymbol{\beta}$. We consider three ways to correct the standard errors and we shall compare their accuracy and efficiency in the following simulation study so that we can find a proper one as a default approach for computing the standard error in routine applied work.

We first propose a simple, yet ingenious, approach to correct the standard error. The basic idea is based upon the equivalence between the first and the second algorithms described above. As mentioned, if we treat λ_i as additional parameters, we can directly use `glm` or `bglm` to estimate the parameters λ_i and $\boldsymbol{\beta}$ and of course the standard errors for the estimates. Therefore, we use the second algorithm to obtain the estimates of $\boldsymbol{\beta}$, and then using these estimates as initial values we run the first algorithm just one iteration via `glm` or `bglm` to obtain the correct standard errors and the *p*-values. For convenience, we refer to this method as one-more-step correction of the standard errors (OMSC).

Second, we use the multivariate Delta Method to obtain the standard errors (Baker 1994). Let $\pi(\hat{\boldsymbol{\beta}})$ denote the estimates of parameters, we have

$$\mathrm{Var}(\pi(\hat{\boldsymbol{\beta}})) = \sum_{i}^{n}\sum_{j=1}^{1+m_i}\left(\frac{\partial\pi(\hat{\boldsymbol{\beta}})}{\partial E(y_{ij})}\right)^T \mathrm{Var}(y_{ij})\left(\frac{\partial\pi(\hat{\boldsymbol{\beta}})}{\partial E(y_{ij})}\right). \quad (3.15)$$

Third, we apply Bootstrap technique to derive the estimates of the standard errors (Efron and Tibshirani, 1993). We first draw a number of resamples from the empirical distribution of the observed data with equal sample size to the observed data, each of which is obtained by randomly sampling with replacement from the original dataset. Note that the sample unit is family in our present study. Next, we run the second algorithm to obtain the estimates of $\boldsymbol{\beta}$ for each of Bootstrap samples. Thus we get an estimate of the distribution of $\hat{\boldsymbol{\beta}}$, and we can then compute the variance of $\hat{\boldsymbol{\beta}}$ by

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\boldsymbol{\beta}}_b - \bar{\hat{\boldsymbol{\beta}}}\right)^2, \quad (3.16)$$

where b is the number of Bootstrap samples, b = 1, …, B, and $\bar{\hat{\boldsymbol{\beta}}}$ is the mean over all of the $\hat{\boldsymbol{\beta}}_b$.

3.4.3 Simulation Study

To evaluate the statistical properties and performance of the proposed method, we considered numerical evidence in the form of simulation studies. We simulated nuclear families consisting of two parents and two to four offspring with one sibling affected by a disease. For simplicity of exposition, herein we just demonstrated the performance of our proposed method in the situation where only one family member has a disease, although it can be applied in the situation where several diseased individuals may exist in a family.

We considered haplotypes formed from a number of biallelic polymorphisms as genetic factors in this simulation study. We assumed that the family members were accurately genotyped for all of the genetic loci and the phase information of the haplotypes was obtained through some molecular techniques. We also assumed that the haplotype frequencies were estimated at the population level and they were distributed uniformly between 0.001 and 0.4. Given these assumptions, the haplotype data were generated as follows. As a first step, we randomly drew two haplotypes (phased haplotype pairs) for each pedigree founder within a family with replacement from the haplotype pool according to the estimated haplotype frequencies by using a multinomial distribution. We then assigned haplotypes to the offspring from their parental haplotypes according to the Mendelian transmission. In this procedure, none inter-locus recombination and allele mutation were assumed as the haplotypes were dropped through pedigrees.

We assumed that the risk of disease was influenced by environmental factors apart from genetic factors. So, for the simplicity in our exposition, we also generated one binary environmental exposure for individuals in a family. For this purpose, we first draw several correlated random variables according to varying family sizes as described above from a multivariate normal distribution with marginal means 0, marginal variances 1 and a correlation parameter that was fixed at 0.4 so that it represented only a modest correlation between the environmental exposures for individuals in a family. We then converted each of these variables into a 0/1 scaled variable in order that the marginal probability of exposure to the suspected risk of disease for the underlying population is 0.3, which reflected a typical exposure to common environmental factors for family members.

We determined the disease status of individuals in a family using a disease risk model as follows

$$\frac{exp\left(\alpha_i + x_E\beta_E + \sum_{w=1}^{W} c\left(h_1, h_2\right)\beta_{H\,w}\right)}{1+exp\left(\alpha_i + x_E\beta_E + \sum_{w=1}^{W} c\left(h_1, h_2\right)\beta_{H\,w}\right)}. \tag{3.17}$$

To make this happen, we first generated the family-specific parameter α_i to allow for heterogeneity in the risk of disease between families that cannot be accounted for by genetic and environmental factors fitted in the model. For a given family *i*, we assumed that α_i follows a uniform distribution, $U\left(\theta - log(5/4),\ \theta + log(5/4)\right)$, with a particular value of θ chosen to control the baseline penetrance of disease for the family. For example, if we set $\theta = log\left(10^{-3}\right)$, the baseline penetrance of disease will be in some neighborhood of 0.001. *W* is the number of haplotypes. $c\left(h_1, h_2\right)$ is the number of times that a particular haplotype, say, the *w*th haplotype, appears in each drawn haplotype pairs $\left(h_1, h_2\right)$, and it is defined as

$$c\left(h_1, h_2\right) = \begin{cases} 0 & \text{if 0 } w\text{th haplotype} \\ u & \text{if 1 } w\text{th haplotype} \\ 2 & \text{if 2 } w\text{th haplotypes} \end{cases}$$

where *u* can be 0, 1, or 2, which depends on the prespecified genetic model: recessive, additive, or dominant, respectively.

We then assumed an effect size ("true" value), measured as an OR, for each element in the vector of parameters $\boldsymbol{\beta}$ here including β_E and $\boldsymbol{\beta}_H$ (see Subsection 2.4.2 for more details). By varying the effect size in the three scenarios, we assumed that some explanatory variables increased the odds of getting disease and others are not associated with the disease so that we can assess statistical power and type I error rates of the proposed method. To be more specific, we described the three scenarios as follows:

1) In the first scenario, we considered 5 haplotypes within a haplotype block and defined the first and fourth haplotypes as common haplotypes (frequency $\geq$ 0.05), the second haplotype as a rare haplotype (frequency < 0.01), and the third and fifth haplotype as moderately rare haplotypes (0.01 $\leq$ frequency < 0.05). We assumed that the first three haplotypes were associated with the disease with ORs ranging between 2 and 4, and the other two haplotypes were not associated with the disease with ORs being fixed at 1 (Table 3.1).

2) In the second scenario, we considered both the main and interacting effects arising between the two haplotype blocks, and between the haplotypes and the environmental factor. The first and second haplotypes within the 2nd haplotype block were defined as common and moderately rare haplotypes respectively, and the other 5 haplotype within the 1st haplotype block were defined as those in the first scenario. The effect sizes of all the explanatory variables in this scenario were assumed as those in Table 3.1.

3) In the third scenario, we considered twenty haplotype blocks with 10 haplotypes within each block. We fitted the main effects of all the haplotypes as well as the environmental factor and all possible interacting effects between any two haplotype blocks and between the haplotypes and the environmental factor. So there were a total of 19401 explanatory variables in the model, including 201 main-effect variables, 200 haplotype-environment interaction terms, and 19000 haplotype-haplotype interaction terms. However, we only assumed that the environmental factor, the haplotype 10 within the 5th haplotype block (common), the haplotype 10 within the 10th haplotype block (rare), the haplotype 10 within the 15th haplotype block (common), the haplotype 10 within the 20th haplotype block (moderately rare), and five interactions were associated with the

disease (Table 3.1). We created this kind of simulation setting because we wanted to prevent the labels of the explanatory variables on the vertical axis of Figure 3.3 from overlapping one another (seeing Figure 3.3 could help easily understand our consideration). In addition, in this scenario, we formed a hypothesis that some haplotypes affect a disease or a trait mainly through their interactions by assuming that the ninth haplotype within the 15th haplotype block and the ninth haplotype within the 20th haplotype block had no main effects but had an interacting effect.

Table 3.1. Explanatory Variables and Their Effect Sizes in the Model for the Three Scenarios

Scenario 1		Scenario 2		Scenario 3	
Variable	OR	Variable	OR	Variable	OR
haplo1	2	*ef*	2	*ef*	2
haplo2	4	*haplo1.1*	2	*haplo5.10*	2
haplo3	3	*haplo1.2*	4	*haplo10.10*	4
haplo4	1	*haplo1.3*	3	*haplo15.10*	2
haplo5	1	*haplo1.4*	1	*haplo20.10*	3
		haplo1.5	1	*ef*:*haplo10.10*	4
		haplo2.1	2	*ef*:*haplo15.10*	3
		haplo2.2	1	*haplo5.10*:*haplo15.10*	3
		ef:*haplo1.1*	3	*haplo10.10*:*haplo20.10*	5
		haplo1.2:haplo2.1	4	*haplo15.9*:*haplo20.9*	3
		haplo1.3:haplo2.2	1	other 19391 variables	1

haplo1: the haplotype 1, *haplo1.1*: the haplotype 1 in the 1st haplotype block, *ef*: the environmental factor, *haplo2.2*: the haplotype 2 in the 2nd haplotype block, *ef*:*haplo1.1*: the interaction between the environmental factor and the haplotype 1 in the 1st haplotype block, *haplo1.2:haplo2.1*: the interaction between the haplotype 2 in the 1st haplotype block and the haplotype 1 in the 2nd haplotype block, and so on.

Given α_i and $\boldsymbol{\beta}$, we first generated a case (proband) for the *i*th family, and then generated several controls according to varying family sizes as described above for the family. In our simulation study, we considered 250, 500, and 1000 families for the first two scenarios and 2000, 4000, and 8000 families for the third scenario, and a total of 1000 replicates were generated under each of these settings.

All of the generated data were analyzed using the proposed method and the results were compared with those from the traditional conditional logistic regression. For the family data, the traditional conditional logistic regression treats the families as strata and we obtain the estimates of parameters in a model by maximizing the conditional likelihood, which can be done by implementing the `clogit` function in the package `survival` in R (http://cran.r-project.org/web/packages/survival/index.html).

In our evaluation processes, we mainly assessed the statistical accuracy and reliability as follows:

1) We calculated relative bias for each parameter in the model by $\left(b_k - E\left(\hat{b}_k\right)\right)/b_k$, where b_k is the "true" value of the *k*th parameter, $k = 1,\ 2,\ ...,\ K$, $\hat{b}_k$ is an estimate of the *k*th parameter.

2) We calculated empirical power for each parameter in the model by $power = 1\big/R\sum_{r=1}^{R} I_{\left(p_{rk} \le \alpha\right)}$, where R is the number of replicates required, p_{rk} is the *p*-value of the *k*th parameter in the *r*th replicate, α is the statistical significance criterion used in the test. For the first two scenarios, α is selected as being 0.05, 0.01, and 0.001, while for the third scenario, α takes a more stringent genome-wide significance threshold level of 2.6×10-6.

3.4.4 Results

Prior to presenting our simulation results of comparing the statistical properties and performance of the proposed method versus the traditional conditional logistic regression, we reported the relative merits of OMSC, Delta Method, and the Bootstrap technique for estimating standard errors of parameter estimates as described in the Methods section and determined which one is an optimal solution for our new method to correct the standard errors, which is prerequisite for our new method to ensure valid statistical inference.

Standard errors of parameter estimates

We conducted the evaluation process in the first scenario (Table 3.1) and we first assessed the accuracy of the three approaches for computing standard errors of parameter estimates. The common way to assess the accuracy of an estimator for computing standard errors of parameter estimates in the literature is to first set up the "true" values of standard errors of parameter estimates, then simulate data via some simulation schemes and drive the estimated standard errors of parameter estimates based on some statistical models, and finally compare the estimated values with the "true" values. But since the primary goal of our present study was to evaluate the statistical properties and performance, not the accuracy of an estimator for computing standard errors of parameter estimates, of the proposed method by comparing it with the traditional conditional logistic regression, we considered an indirect way to examine the accuracy of the three approaches for computing standard errors of parameter estimates by investigating empirical power or type I error rate, or more precisely, the empirical distribution of *p*-values for each parameter estimate in our simulation study because in the computation of *p*-value for each parameter estimate, in general, the only factor that affects the magnitude of *p*-value is the standard error of the parameter estimate given the hypothesis test, the probability distribution of the test statistic, the parameter estimate, and even the sample size.

For the first scenario, the empirical power was calculated for the first three haplotypes because they were assumed to be associated with the disease and the empirical type I error rates were calculated for the other two haplotypes because they were not assumed to be associated with the disease based on OMSC, Delta Method, and the Bootstrap technique (the right panel of Figure 3.1). For the Bootstrap technique, six numbers of Bootstrap samples (the number of Bootstrap replications) (10, 30, 50, 100, 500, and 1000) were considered for each effect of interest (Figure 3.1). Note that here we show only the results obtained from 500 families, which is a medium-scale sample in our simulation setting and, of course, a realistic sample size.

The results were fairly clear. Out of the three approaches, Delta Method had the highest power and type I error rates under each of three fixed statistical significance criteria used in the test (α = 0.001, 0.01, and 0.05) (the right panel of Figure 3.1). By further checking the variance (the square of standard error) of parameter estimates, we can see that Delta Method had the lowest variance for all of the 5 effects (the left panel of Figure 3.1), which is in agreement with the well-established fact that Delta Method generally tends to underestimate standard errors of parameter estimates (Efron, 1990). On the contrary, OMSC yielded the lowest power under each of three fixed statistical significance criteria used in the test (the right panel of Figure 3.1) and, as imagined, the highest variance for the corresponding effects (the left panel of Figure 3.1). While for the

Bootstrap technique, except the Bootstrap sample being 10, all of the other five Bootstrap samples provided quite good power and type I error rates under each of three fixed statistical significance criteria used in the test (the right panel of Figure 3.1). This result implied that the variances obtained from these Bootstrap samples were much better than those obtained from the other two approaches, which is perfectly in line with the finding that, in general, the Bootstrap technique is superior to Delta Method in the context of estimating standard errors when the sample size is moderate (Efron, 1982; Chernick, 2007). Furthermore, by taking a close look at the graph, it can be seen that these five Bootstrap samples had comparable power at $\alpha = 0.05$ and type I error rates under all of the three fixed statistical significance criteria used in the test; only when the Bootstrap samples went up to 500 and 1000, the powers at $\alpha = 0.01$ and 0.001 were lightly higher than those when the Bootstrap samples were 30, 50, and 100.

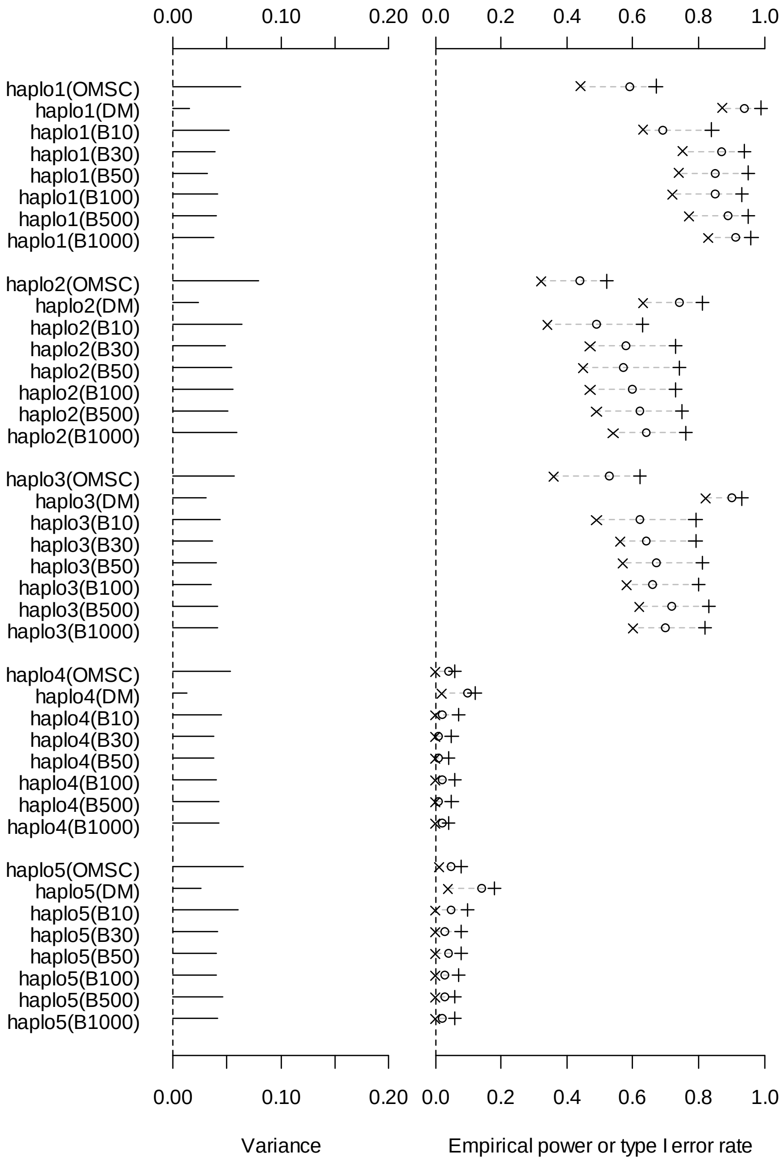


Figure 3.1. Accuracy of the Three Approaches for Computing Standard Errors. Variances and empirical powers or empirical Type I error rates (× indicates the empirical powers or Type I error rates for $\alpha = 0.001$, ○ for $\alpha = 0.01$, and + for $\alpha = 0.05$) for each of 5 main-effect predictors based on the three approaches under the sample sizes of 500. *haplo1*(DM) stands for the effect of the haplotype 1 with the variance estimated by Delta Method, *haplo2*(B10) stands for the effect of the haplotype 2 with the variance estimated by the Bootstrap technique with the Bootstrap sample being 10, and so on.

This demonstrates that Bootstrap technique provides a fairly good measurement of standard errors of parameter estimates. However, to choose among these six Bootstrap samples we had to use further information about their performance. In this regard, one key aspect is the efficiency of an estimator in terms of execution time, which answers: How fast is the estimator speed of computing estimates? To do so, we ran our model 1000 times with each of the three approach for computing standard errors of parameter estimates on a desktop computer with a single 3.6GHz Intel Pentium 4 CPU and 2GB RAM, which is not a mainstream configuration of the desktop computer at present. We measured the execution time, in seconds, and summarized it in Table 3.2.

Table 3.2. Average Time (Second) of Computing Standard Error for Different Approaches

OMSC	Delta Method	Bootstrap					
		Six numbers of Bootstrap samples					
		10	30	50	100	500	1000
1.02	46.27	19.07	52.15	96.38	191.02	952.34	1908.64

As can be seen in Table 3.2, the execution time taken by the Bootstrap sample being 30 was almost the same as that taken by Delta Method and much less than those taken by the Bootstrap sample being 50, 100, 500, and 1000. Hence, it is evident from this result together with the accuracy results above that the Bootstrap sample being 30, with realistic sample sizes, can provide adequate accuracy for computing standard errors of parameter estimates and is sufficiently efficient in terms of execution time that are satisfactory for most of applied researches. Therefore, we used only the Bootstrap technique with the Bootstrap sample being 30 to compute the standard errors of parameter estimates in the following analysis.

Small-scale model

Turning to the process of evaluating the statistical properties and performance of the proposed method compared with the traditional conditional logistic regression, we first investigated the potential bias in the parameter estimates in the second scenario (Table 3.1) that is typically encountered in the candidate gene analysis in which usually several, not many, genes, sometimes in combination with environmental factors, and/or their interactions are studied for complex diseases or traits. We did not directly compare the bias in the parameter estimates from the two methods because the parameter estimates from different methods are not comparable due to difference in the scale of measurement. Thus, we assessed the two methods based on their relative bias (the three left panels of Figure 3.2), which provided a measure of the magnitude of the bias on the same scale for the two methods.

The top left panel of Figure 3.2 shows the corresponding results under the sample size of 250, from which we can see that all of the parameter estimates obtained from the two methods were biased, though the relative bias was quite small, ranging between 0.03 and 0.15. We also observed that, for both the methods, the estimates for the effects of the (moderately) rare haplotypes (*haplo1.2*, *haplo1.3*, *haplo1.5*, and *haplo2.2*) and the interacting effects were more biased than the others, which is obviously attributed to the fact that rare predictors and interaction terms in a model usually have larger estimated standard errors than the common ones in model fit. Furthermore, we noticed that the proposed method had less bias than the traditional conditional logistic regression for all of the parameter estimates, especially for those of the (moderately) rare haplotypes and the interactions.

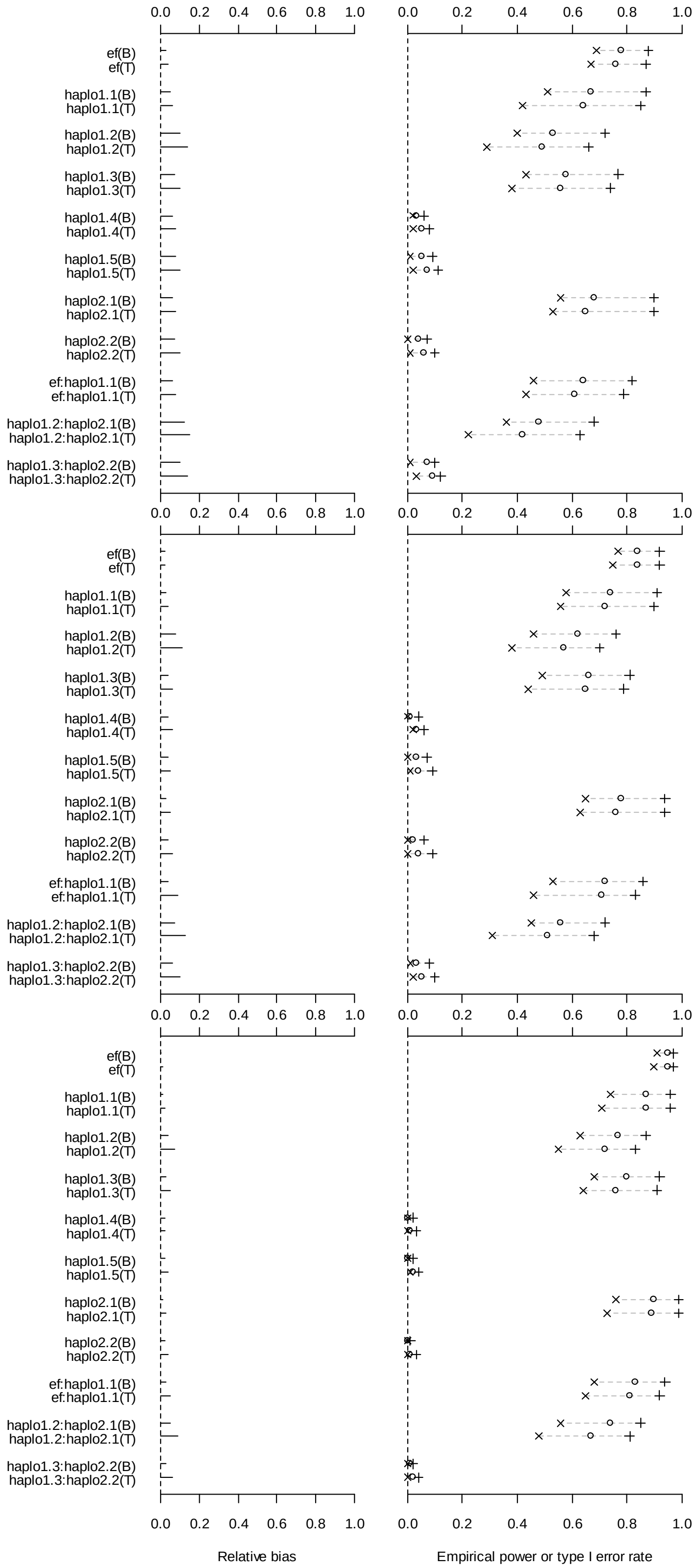


Figure 3.2. Small-Scale Model. Relative biases and empirical powers or empirical Type I error rates (× indicates the empirical powers or Type I error rates for $\alpha = 0.001$, ○

for $\alpha = 0.01$, and + for $\alpha = 0.05$) for each of 11 explanatory variables based on the two methods under the sample sizes of 250 (top), 500 (middle), and 1000 (bottom). *ef* (B) stands for the effect of the environmental factor to be estimated using the proposed method, *haplo1.1*(T) stands for the effect of the haplotype 1 within the 1st haplotype block to be estimated using the traditional conditional logistic regression, *ef:haplo1.1*(B) stands for the interacting effect between the environmental factor and the haplotype 1 within the 1st haplotype block to be estimated using the proposed method, *haplo1.2*(B)*:* *haplo2.1*(T) stands for the interacting effect between the haplotype 2 within the 1st haplotype block and the haplotype 1 within the 2nd haplotype block to be estimated using the traditional conditional logistic regression, and so on.

To see the influence of the sample size, we also displayed the corresponding results under the sample size of 500 and 1000, from which it can be seen that with the number of families selected increased, the relative bias of both the methods declined but that of the proposed method went down faster and, for some effects, even shrank towards zero for the sample size up to 1000 (the middle and bottom left panels of Figure 3.2).

We next assessed the two methods in terms of empirical power for the main effects of *ef*, *haplo1.1*, *haplo1.2*, *haplo1.3*, and *haplo2.1* and the interacting effects of *ef*:*haplo1.1*, *haplo1.2*:*haplo2.1*, and *haplo1.3*:*haplo2.2* because we assumed that they were associated with the disease (the three right panels of Figure 3.2). Here we tried to evaluate the ability of the methods to declare any disease-predisposing factors when some effects really existed. For the effect of *ef,* the empirical powers under each of three fixed statistical significance criteria used in the test were comparable for the two methods no matter what sample sizes were considered (the top two lines in each of the three right panels of Figure 3.2). This phenomenon would seem to be reasonable because, for a common environmental factor with a decent frequency, any valid statistical test can obtain a similar power for detecting it and the possible variation of powers from different tests can be explained by the random variability. For the other effects, the proposed method provided higher probabilities for correctly identifying them under each of three fixed statistical significance criteria used in the test compared with the traditional conditional logistic regression irrespective of what sample sizes were used (the three right panels of Figure 3.2). Although the superiority of the proposed method in the statistical validity was diminishing for all of the explanatory variables with the increase in sample sizes, it still persisted, especially for the (moderately) rare haplotypes and the interactions and for the empirical powers at $\alpha = 0.01$ and 0.001 (the three right panels of Figure 3.2). For the proposed method at $\alpha = 0.05$, a sample size of 500 was sufficient to detect a common haplotype with a statistical power of 90% approximately, and a sample size of 1000 was sufficient to identify a rare haplotype or an interaction with a statistical power of 80% approximately.

Meanwhile, we assessed the two methods regarding empirical type I error rates for the main effects of *haplo1.4*, *haplo1.5*, and *haplo2.2* and the interacting effect of *haplo1.3*:*haplo2.2* because we assumed that they were not associated with the disease (the right panel of Figure 3.2). Here we tried to evaluate the probability of observing a disease-associated factor when in truth there was none. Under the sample sizes of 250 and 500, the traditional conditional logistic regression yielded a little higher empirical Type I error rates that the proposed method but they are acceptable based on the practical consideration (the top and middle right panels of Figure 3.2). As the sample size went up to 1000, all empirical Type I error rates shrank towards almost zero (the bottom right panel of Figure 3.2).

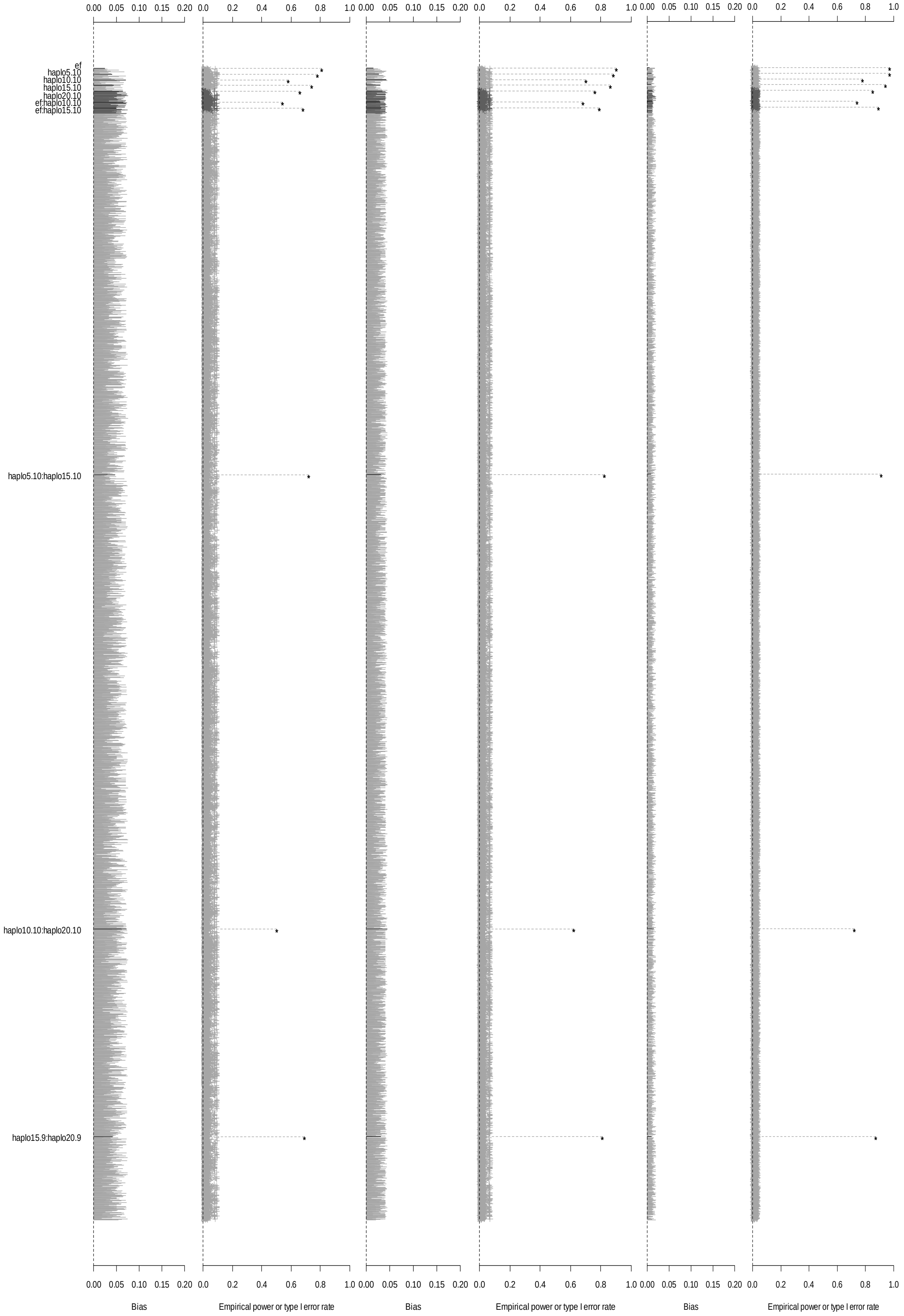
High-dimensional model

In order to obtain a more comprehensive picture about the performance of the proposed method in the case where a very large number of haplotypes and/or environmental factors are investigated, e.g., in genome-wide association studies that involve testing numerous genes across the complete sets of DNA of many people to find genetic variations associated with a particular disease, we conducted a sophisticated simulation study in the third scenario (Table 3.1), in which there were a total of 19401 explanatory variables jointly considered, including 201 main-effect variables, 200 gene-environment interaction terms, and 19000 gene-gene interaction terms. For this huge data, as one would expect, the implementation of analysis must require a large amount of memory and high-performance computing resources. Therefore, we ran our simulation on a computer cluster, named Cheaha, at the University of Alabama at Birmingham (UAB) (http://docs.uabgrid.uab.edu/wiki/Cheaha), which includes 192 3.0GHz Intel-based compute cores with 386GB of RAM interconnected via a DDR Infiniband network. A high-performance, 60TB Lustre parallel file system built on a Direct Data Network (DDN) hardware platform is also connected to these cores via the Infiniband fabric. An additional 40TB of traditional shared storage and an auxiliary 120 1.6GHz AMD-based compute cores are available via a 1GigE network fabric.

In addition, since `clogit`, which is an R function created to carry out the traditional conditional logistic regression as mentioned earlier, did not work for the data, specifically, the program appeared to freeze, we show only the results from the proposed method (Figure 3.3).

As in the preceding small-scale model subsection, the relative biases in the parameter estimates of all the explanatory variables in the model were first computed for each of three sample sizes respectively (the first, third, and fifth panels of Figure 3.3). From the graphs we can see that, under the sample size of 2000, the maximum relative bias was not more than 0.07. Along with the increase in sample sizes, there existed an overall tendency towards the decrease in the relative bias. This finding was consistent with that observed in the preceding subsection.

The empirical power was next computed for a total of ten disease-associated explanatory variables in the model at a genome-wide significance threshold level of 2.6×10^{-6} (the second, fourth, and sixth panels of Figure 3.3). As can be seen from the graphs, with the sample size increased, there could be some significant gains in the empirical power, with the large gain happening for the environmental factor and the common haplotypes (*haplo5.10* and *haplo15.10*) and the small gain happening for the (moderately) rare haplotypes (*haplo10.10* and *haplo20.10*) and interactions. This verified the statement we made in the foregoing power analysis in the small-scale model. Furthermore, this result also proved our hypothesis that some haplotypes may affect a disease or a trait mainly through their interactions, and the interacting effects play a more significant role than does the main effects in regulating the genetic variation of the disease or the trait. However, we noticed that the empirical power started at a relatively low level and maintained a quite small growth rate over the three sample sizes compared with those in the foregone small-scale model subsection.



0.00 0.05 0.10 0.15 0.20
0.0 0.2 0.4 0.6 0.8 1.0
ef
haplo5.10
haplo10.10
haplo15.10
haplo20.10
ef:haplo10.10
ef:haplo15.10
haplo5.10:haplo15.10
haplo10.10:haplo20.10
haplo15.9:haplo20.9
*
Bias
Empirical power or type I error rate
Bias
Empirical power or type I error rate
Bias
Empirical power or type I error rate

Figure 3.3. High-Dimensional Model. Relative biases and empirical powers at $\alpha = 2.6 \times 10^{-6}$ (*) or empirical Type I error rates at $\alpha = 0.001$ (×), $\alpha = 0.01$ (○), and $\alpha = 0.05$ (+) for each of 19401 explanatory variables based on the proposed method under the sample sizes of 2000 (the first two panels), 4000 (the third and fourth panels), and 8000 (the last two panels). The three categories of the explanatory variables (main-effect, gene-environment, and gene-gene) were distinguished from one another through different colors (gray, dark gray, and gray, respectively). Only the disease-associated explanatory variables were labeled on the vertical axis. *ef* stands for the environmental factor, *haplo5.10* stands for the haplotype 10 within the 5th haplotype block, *ef:haplo10.10* stands for the interaction between the environmental factor and the haplotype 10 within the 10th haplotype block, *haplo5.10: haplo15.10* stands for the interaction between the haplotype 10 within the 5th haplotype block and the haplotype 10 within the 15th haplotype block, and so on.

The empirical Type I error rates were also calculated for a total of 19391 non-disease-associated explanatory variables in the model (the second, fourth, and sixth panels of Figure 3.3). The result shows that there was a substantial decline in the empirical Type I error rates over the three sample sizes. Under the sample size of 2000, the magnitudes of the empirical Type I error rates were not more than 8%. As the sample size went up to 8000, all of the Type I error rates shrank to almost zero.

3.4.5 Discussion

We have developed a Bayesian framework for detecting gene-gene and gene-environment interactions, particularly involving rare variants, using family-based case-control data. Since susceptibility to the majority of human diseases is complex and multifactorial, involving both genetic and environmental factors, jointly considering all these factors and their possible interactions in analysis just like we do could enhance the statistical power for identifying genetic variants that are involved in the etiology of disease mainly through an interacting effect, and ascertaining rare variants that act primarily in genetically susceptible individuals. However, analyzing both gene-related interactions and rare variants remains a big challenge because we must simultaneously handle high-dimensional data arising from numerous marginal and interacting effects fitted in a model, sparse data arising from both a large number of interactions and rare variants, and the computational burden of analysis. This motivates sophisticated approaches, nevertheless until recently few studies have had success. We have been making a great effort trying to fill the gap in available methods in the context of both population-based (Li *et al.*, 2011) and family-based association studies.

Our method is created on the basis of the generalized linear model and thus can take advantage of the well-established theory, algorithm, and software developed for the generalized linear model, and include various models as special cases. To fit a large number of terms, including common and rare variants, environmental factors, and their possible interactions, in a model, we assume weakly informative priors on the parameter estimates because the priors can induce strong shrinkage for near-zero effects but weak shrinkage for large effects (Gelman *et al*., 2003; Gelman *et al*., 2008; Yi and Banerjee, 2009). To enhance the efficiency of computing maximum likelihood estimates, we employ the Multinomial-Poisson transformation technique by substituting a Poisson likelihood with an additional parameter (e.g., Baker, 1994; Gelman *et al*., 2003). We consider three ways to correct the bias in the estimation of standard errors of parameter estimates happened after introducing the Multinomial-Poisson transformation technique and treating the additional parameter as offset in model fit. The result yielded evidence that the Bootstrap technique can provide a better estimate of standard errors than Delta Method and OMSC (Figure 3.1), which is in agreement with the previous observations (Efron, 1982; Efron, 1990; Chernick, 2007). Furthermore, we used the Bootstrap sample of 30 to calculate the estimate of standard errors in the present study because the results suggest that the Bootstrap sample of 30 is adequate and efficient for estimating standard errors of parameter estimates in terms of accuracy and execution time based on the practical consideration (Figure 3.1 and Table 3.2).

As noted, there has been a long-standing debate within the scientific community about the Bootstrap sample size. How many Bootstrap samples we need to take in a study? Unfortunately, there are no any general guidelines that have been proposed for how large the Bootstrapped sample should be relative to the total number of observations in the dataset from which it is drawn? Efron (1987) pointed out that there is little improvement for the accuracy of estimates when the Bootstrap sample size is more than 100. In fact, The Bootstrap sample size as small as 25 gives reasonable results. Certainly, the Bootstrap sample size can be large. But, in general, determining the Bootstrap sample size depends on the extent to which we care to establish the accuracy of estimates, available computing resources, and other practical consideration. If the results really matter, as many samples as is reasonable given available computing resources and time should be used. However, it is noteworthy that increasing the Bootstrap sample size cannot increase the amount of information in the original data. It can only reduce the effects of random sampling errors arising possibly from the Bootstrap procedure itself.

In our simulation study, we compared the results from our method with those from the traditional conditional logistic regression. The traditional conditional logistic regression is a classical method for analyzing matched case-control data in epidemiology (Breslow and Day, 1980; Breslow, 1982; Rothman *et al*., 2008a,b), and it is also appropriate for testing genetic association using family data (e.g., Goldstein *et al*., 1989; Andrieu and Goldstein, 1996; Witte *et al*., 1999; Kraft and Thomas, 2000; Siegmund *et al*., 2000). However, we have not compared our method to the other existing methods because, primarily, the other existing methods are built upon different philosophies and thus it would be difficult to compare directly; in addition, for some existing methods, their implementations have not yet been publicly available or not easily implemented though.

To ensure the statistical reliability of parameter estimates, we evaluated the proposed method and the traditional conditional logistic regression based on their relative bias in the parameter estimates of all the explanatory variables in the model. The simulation results clearly demonstrate that the reliability advantage of the proposed method is consistently over the traditional conditional logistic egression in both the small-scale and high-dimensional models, especially for the interactions and (moderately) rare haplotypes, no matter what sample sizes were considered (Figure 3.2 and Figure 3.3). This indicates that the proposed method can produce statistically reliable and robust models and can be used in both the candidate gene and genome-wide association studies.

Further, we empirically assessed the two competing methods regarding their statistical power and Type I error rates. The simulation study shows that the proposed method is more powered than the traditional conditional logistic regression, especially for the interactions and (moderately) rare haplotypes (Figure 3.2). When much more explanatory variables are fitted in the model, however, the proposed method suffers from loss of power (Figure 3.3). But the power is still acceptable in practice. This should not be surprising because accommodating the high-dimensionality in a model comes at the price of reduced statistical power. We also see that the proposed method produced little higher Type I error rates than the traditional conditional logistic regression under the sample sizes of 250 and 500 (Figure 3.2). In addition, we observed that the proposed method incorporating analysis of interactions can identify causal haplotypes, which might have a weak marginal effect but a strong interacting effect with other haplotypes (Figure 3.3). Therefore, we can say that the proposed method has reasonable power to detect true effects, while controlling the rate of false positives.

CHAPTER 4

SOFTWARE

4.1 Overview of R/BhGLM

The proposed methods have been and are being incorporated into R/BhGLM by creating some new functions to conduct haplotype-based association analysis. R/BhGLM is a publicly available and distributable package (http://www.ssg.uab.edu/bhglm), and is implemented as an add-on package for the software R, which is a free development environment for statistical computing and graphics (Ihaka and Gentleman, 1996). R/BhGLM provides an extensible, interactive programming environment for haplotype-based association analysis except several previously built-in functions for some other statistical genetic analyses, e.g., SNP-based association analysis and quantitative trait loci (QTL) mapping, in the Bayesian framework.

There are many packages available for haplotype-based association analysis, e.g., hapassoc, haplo.stats, and gap. Among them, haplo.stats is a popular tool for haplotype-based association analysis because it is available free and easy to implement (http://mayoresearch.mayo.edu/mayo/research/schaid_lab/software.cfm). haplo.stats performs likelihood inference of trait associations with haplotypes in the generalized linear model framework (Lake *et al.*, 2003), but, as almost all of the existing software implementing haplotype-based association analysis, it can only fit the main effects of haplotypes and haplotype-environment interactions with a relatively small number of effects needed fitting in the model. In addition, if there are numerous effects needed fitting in a model and/or some rare haplotypes present, `haplo.stats` as well as the other software might encounter some serious problems such as nonidentifiability of parameters.

Our package performing haplotype-based association analysis is based on Bayesian hierarchical generalized linear models with a continuous prior distribution on the coefficients that favors sparsity in the fitted model and facilitates intensive computation (Gelman *et al*., 2008; Yi and Banerjee, 2009). A fast EM algorithm is built into the iteratively weighted least squares for classical generalized linear models to fit our models by estimating posterior modes of coefficients (Gelman *et al*., 2008; Yi and Banerjee, 2009), which allows us to simultaneously fit environmental effects, main effects of numerous common and rare haplotypes, and haplotype-haplotype and haplotype-environment interactions.

Currently, `R/BhGLM` can perform population-bases and family-based association analyses with haplotypes, and provide a unified approach to explore haplotype associations with continuous, binary, or ordinal traits. `R/BhGLM` incorporates several functions for data simulation, data manipulation, and result summaries including graphics. `R/BhGLM` is programmed to accept original data in a variety of input formats and is accessible for most platforms including Windows, MacOS, and UNIX/Linux. The computationally intensive algorithms were written in C, while data manipulation and graphics were written in R language.

`R/BhGLM` is under continual development.

4.2 R Functions for Haplotype-Based Association Analysis

Here we briefly describe the use of some functions incorporated in `R/BhGLM` for haplotype-based association analysis. A more extensive tutorial and a help file on their use are distributed with the software and are also available at the website cited in the preceding section. To this end, we consider an example dataset named `haplo`, which includes a total of 100 subjects from a cases-control study, with four htSNPs, some covariate variables including age, race, weight, and so on, and the disease status of breast cancer. We hope that the demonstrations presented here will be helpful to understand how to use the functions with little prior knowledge of R, especially because we neglect to explain the syntax and some basic functions of R. There are lots of free resources available on the R project website (http://www.r-project.org/) or some other relevant websites that can assist the user in learning and using R.

4.2.1 Getting Started

The procedure for using the functions for haplotype-based association analysis is the same as any other one in R. So, in order to use the `R/BhGLM` package, one must download it from the website cited in the preceding section and install it properly, which can be easily done just following the instruction of package installation on the same website where you download the `R/BhGLM` package. After installing the `R/BhGLM` package, the routines are available by starting an R session and loading the package as done below. Here we assume that the user is running either Windows or Mac OS X.

`> library(BhGLM)` (type `library(BhGLM)`within R following the prompt “>”)

Then we can use the function `data()` to load the data. `data()` is a basic function for inputting data in `R`.

```
> data(haplo)
```

Now the dataset or sub-dataset can be accessed by using some functions within R.

4.2.2 Creating a Genotype Matrix

The datasets of genetic markers, e.g., SNPs, are often arranged in a one column format that looks like:

```
      rs9939609 rs1477196 rs7206790 rs8047395
 [1,]         2         1         0         1
 [2,]         1         1         1         1
 [3,]         2         1        NA         2
 [4,]         1         2         1         0
 [5,]        NA        NA         0         2
 [6,]         1         2         2         0
 [7,]        NA         1         1         1
 [8,]         1        NA         2         0
 [9,]         1         2         1         1
[10,]         1         1         1        NA
```

These are the first ten records of SNP alleles of our example dataset displayed by executing the function `geno()`. The numbers in each cell of the table above are the count of the minor allele of a SNP, and the symbol `NA` (not available) represents missing values. Rows represent the measurement of genotypes at four loci for each subject. However, lots of software for genetic analysis requires a special matrix of genotypes, which is arranged such that each locus has a pair of adjacent columns of alleles, and the order of columns corresponds to the order of loci on a chromosome. If there are *m* loci, the total number of columns of genotypes is 2*m*. To convert the format of one column to the format of two columns, a function, `geno.2cols()`, has been added to the package.

```
> geno.2cols(genodata, v1 = 0, label = NULL)
```

The first argument of the function is the name of the original dataset that is needed converting, the second specifies the smallest value that genetic markers take in the original dataset with a default value of zero (it can be a character such as “M” or anything else defined by the investigator), and the third is loci label which is optional. See its help file for more details.

After converting, the layout of the dataset looks like this (only the first ten records were shown corresponding the exhibition in the preceding subsection):

```
    m1.1 m1.2 m2.1 m2.2 m3.1 m3.2 m4.1 m4.2
1      2    2    1    2    1    1    1    2
2      1    2    1    2    1    2    1    2
3      2    2    1    2   NA   NA    2    2
4      1    2    2    2    1    2    1    1
5     NA   NA   NA   NA    1    1    2    2
6      1    2    2    2    2    2    1    1
7     NA   NA    1    2    1    2    1    2
8      1    2   NA   NA    2    2    1    1
9      1    2    2    2    1    2    1    2
10     1    2    1    2    1    2   NA   NA
```

4.2.3 Estimating Haplotype Frequencies

As discussed before, if the genetic data are collected from unrelated individuals, the phase information of haplotypes can not be easily obtained, and hence we usually need to estimate haplotype frequencies by using some statistical methods. There are many statistical algorithms have been developed and implemented in a number of software packages to estimate haplotype frequencies (Schaid *et al*., 2002), among which `haplo.em` is a well-known R function and it is nested in `haplo.stats` described in the preceding section. `haplo.em` has some strengths such as computational efficiency (http://mayoresearch.mayo.edu/mayo/research/schaid_lab/software.cfm). However, it also has some weakness. For instance, it does not provide haplotype frequencies for each level of a grouping (categorical) variable, e.g., sex, disease status, or medical treatment groups, and its output is not ready to use as an intermediate result by other functions within the same session. So, we only consider `haplo.em` as a basic module in our function to estimate haplotype frequencies.

For the converted data (two column format), we use our function `haplo.freq` to construct haplotype pattern and estimate haplotype frequencies by typing the following script:

```
> haplo.freq(geno2col, group = FALSE, group.var = list(y))
```

Here the first argument of the function is the name of the converted dataset. The second one determines whether haplotype frequencies are estimated separately or not by a grouping variable. To do this, the whole dataset must be sorted by the grouping variable first. The third argument is a list of grouping variable(s) by which the haplotype frequencies are estimated separately.

After executing the function `haplo.freq`, we have

```
      haplo.code  loc-1  loc-2  loc-3  loc-4    haplo.prob
1              1      1      1      1      1       0.00283
2              2      1      1      1      2       0.00089
3              3      1      1      2      2       0.00000
4              4      1      2      1      1       0.03084
5              5      1      2      1      2       0.00074
6              6      1      2      2      1       0.32629
7              7      1      2      2      2       0.01179
8              8      2      1      1      1       0.00234
9              9      2      1      1      2       0.32626
10            10      2      1      2      1       0.00909
11            11      2      1      2      2       0.01533
12            12      2      2      1      1       0.07893
13            13      2      2      1      2       0.11370
14            14      2      2      2      1       0.07053
15            15      2      2      2      2       0.01043
```

The first column of the table above is the index of the inferred haplotypes. There are a total of 15 haplotypes inferred from the observed genotype data. In general, for 4 SNPs, a total of 16 possible haplotypes might be obtained. But since some subjects had even half of genotypes missed, and so there is no enough information available for them to estimate their haplotype frequencies. The second through the fifth columns represent the inferred haplotype patterns based on the observed genotype data. For example, for the first haplotype, the haplotype pattern is "`1111`", which means there are 4 minor alleles on the chromosome, and by analogy to others. The last column is the estimated haplotype frequencies.

4.2.4 Computing Haplotype Posterior Probabilities

Since the ambiguity of haplotype phase exists, there may be more than one pair of haplotypes that are consistent with the observed genotype data. To account for this ambiguity, we compute posterior probabilities of haplotype pairs for each subject using the function `haplo.post`.

```
> haplo.post(geno2col, group = FALSE, group.var = list(y))
```

For our example data, by executing the script above, we have more than 300 haplotype pairs for a total of 100 subjects (here only ten rows are shown as below). For each subject, there is at least one pair of haplotypes, e.g., for the first subject in the following table, there are two pairs of haplotypes, i.e., haplotype 8/haplotype 13, and haplotype 9/haplotype 12 (see the preceding subsection for more details of haplotype patterns). The last column of the following table is the posterior probabilities for each haplotype pairs.

```
     subj.id hap1code hap2code posterior
1          1        8       13   0.01024
2          1        9       12   0.98976
3          2        1       15   0.00028
4          2       11        4   0.00441
5          2        6        9   0.99440
```

```
6          2       14        2   0.00059

7          2        5       10   0.00006

8          2        8        7   0.00026

9          3        9       13   0.87486

10         3        9       15   0.08027
```

4.2.5 Constructing a Design Matrix of Haplotypes

To apply some regression models including ours in haplotype-based association analysis, we need to create a proper design matrix based on estimated haplotype dosage for each subject to facilitate our model fit (see Subsections 2.4.2 and 3.4.2 for more details). We created a function, `haplo.matrix`, to do this work, in which another R function named `reshape` is required. `reshape` can be downloaded from the R project website via the link: http://cran.r-project.org/web/packages/reshape/index.html.

Using the function `haplo.matrix` as in the following script:

`> haplo.matrix(poterioal)`,

we have (only the first two subjects are shown)

```
id         haplo1        haplo2        haplo3        haplo4
 1   0.0000000000  0.0000000000  0.000000e+00  0.0000000000
 2   0.0002762031  0.0005881906  0.000000e+00  0.0044149300


           haplo5        haplo6        haplo7        haplo8
 1   0.000000e+00  0.0000000000  0.0000000000  0.0102387934
 2   6.265698e-05  0.9943999000  0.0002580755  0.0002580755
```

```
        haplo9       haplo10       haplo11       haplo12
1    0.9897612  0.000000e+00  0.0000000000  0.9897612070
2    0.9943999  6.265698e-05  0.0044149300  0.0000000000

       haplo13       haplo14       haplo15
1  1.023879e-02  0.000000e+00  0.0000000000
2  0.000000e+00  5.881906e-04  0.0002762031
```

4.2.6 Randomly Sampling Genotypes

In haplotype-based association studies, we often use the inferred haplotype frequencies published in the scientific literature to do simulation studies. Since usually we can only get the inferred haplotype frequencies free, not the original genotype data, we need to simulate the genotype data based on the haplotype frequencies. Using our function `geno.samp`, we randomly draw 2 haplotypes (phased haplotype pairs) for each subject with replacement from inferred haplotypes based on estimated haplotype frequencies by using a multinomial distribution. Then the phased information of sampled haplotype pairs is eliminated to obtain SNP genotype data (see Subsections 2.4.3 and 3.4.3 for more details).

Executing the following script:

```
> geno.samp(hapfreq = hf, group = FALSE, group.var =
list(y), n.geno = 100),
```

we have

```
    m1.1 m1.2 m2.1 m2.2 m3.1 m3.2 m4.1 m4.2
1      2    2    2    2    1    2    2    2
2      1    1    1    2    1    2    1    1
3      1    2    2    2    2    1    1    2
4      1    1    1    2    1    2    2    1
5      1    1    1    1    1    1    1    1
6      1    2    1    2    1    1    1    2
7      1    2    2    2    2    1    1    2
8      1    2    1    2    1    1    1    2
9      1    2    2    2    1    1    1    2
10     1    1    1    2    1    1    1    1
```

Here only the first ten records are shown. These are similar as those in Subsection 4.2.2. Note that the first argument of the function above is the name of the dataset consisting of inferred haplotype frequencies. The last one, `n.geno`, specifies the sample size.

4.2.7 Testing Hardy-Weinberg Equilibrium

We wrote a function, `hwe.test,` to perform HWE test for individual genetic locus. The function calls `chisq.test` to compute a *p*-value for HWE test and the null hypothesis is that HWE holds. To run our function, we need to specify genotype data and chromosomes as below:

```
> hwe.test(cross.control, 1).
```

4.2.8 Simulating genetic data

We created a function, `geno.sim`, to generate haplotype data for both family-based and population-based association analyses using the estimated haplotype frequencies from real data or fake data. The function can also simulate qualitative and quantitative traits based on some genetic and statistical models with various genetic and environmental factors designated by users.

4.2.9 Computing standard errors of parameter estimates

To correct the standard errors of parameter estimates in the family-based haplotype-association analysis, three ways (OMSC, Delta Method, and the Bootstrap technique) are considered in the present study and implemented by running R functions: `omsc.se`, `delta.se`, and `boot.se`.

LIST OF REFERENCES

Akey, J., Jin, L., Xiong, M. (2001). Haplotypes vs. single marker linkage disequilibrium tests, what do we gain? Eur. J. Hum. Genet. 9, 291-300.

Albert, A., and Anderson, J.A. (1984). On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71, 1-10.

Allen, A.S., and Satten, G.A. (2007). Inference on haplotype/disease association using parent-affected child data: the projection conditional on parental haplotypes method. Genet. Epidemiol. 31, 211-223.

Altmuller, J., Palmer, L.J., Fischer, G., Scherb, H., Wjst, M. (2001). Genomewide scans of complex human diseases: true linkage is hard to find. Am. J. Hum. Genet. 69, 936-50.

Altshuler, D., Daly, M.J., Lander, E.S. (2008). Genetic mapping in human disease. Science 322, 881-888.

Andrieu, N., and Goldstein, A.M. (1996). Use of relatives of cases as controls to identify risk factors when an interaction between environmental and genetic factors exists. Int. J. Epidemiol. 25, 649-657.

Asimit, J. and Zeggini, E. (2010). Rare variant association analysis methods for complex traits. Ann. Rev. Genet. 44, 293-308.

Azzopardi, D., Dallosso, A.R., Eliason, K., Hendrickson, B.C., Jones, N., Rawstorne, E., Colley, J., Moskvina, V., Frye, C., Sampson, J.R., Wenstrup, R., Scholl, T., Cheadle, J.P.

(2008). Multiple rare nonsynonymous variants in the adenomatous polyposis coli gene predispose to colorectal adenomas. Cancer Res. 68, 358-363.

Baker, S.G. (1994). The multinomial-Poisson transformation. The Statistician 43, 495-504.

Bansal, V., Libiger, O., Torkamani, A., Schork, N.J. (2010). Statistical analysis strategies for association studies involving rare variants. Nat. Rev. Genet. 11, 773-785.

Basu, S., and Pan, W. (2011). Comparison of statistical tests for disease association with rare variants. Genet. Epidemiol. doi: 10.1002/gepi.20609.

Becker, T., Schumacher, J., Cichon, S., Baur, M.P., Knapp, M. (2005). Haplotype interaction analysis of unlinked regions. Genet. Epidemiol. 29, 313-322.

Bernardinelli, L., Murgia, S.B., Bitti, P.P., Foco, L., Ferrai, R., Musu, L., Prokopenko, I., Pastorino, R., Saddi, V., Ticca, A., Piras, M.L., Cox, D.R., Berzuini, C. (2007). Association between the ACCN1 gene and multiple sclerosis in Central East Sardinia. PLoS ONE 2(5):e480. doi:10.1371/journal.pone.0000480.

Bochukova, E.G., Huang, N., Keogh, J., Henning, E., Purmann, C., Blaszczyk, K., Saeed, S., Hamilton-Shield, J., Clayton-Smith, J., O'Rahilly, S., Hurles, M.E., Farooqi, I.S. (2010). Large, rare chromosomal deletions associated with severe early-onset obesity. Nature 463, 666-670.

Bodmer, W., and Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. Nat. Genet. 40, 695-701.

Boks, M.P.M., Schipper, M., Schubart, C.D., Sommer, I.E., Kahn, R.S., Ophoff, R.A. (2007). Investigating gene environment interaction in complex diseases: increasing power by selective sampling for environmental exposure. Int. J. Epidemiol. 36, 1363-1369.

Botstein, D., and Risch, N. (2003). Discovering genotypes underlying human phenotypes: Past successes for mendelian disease, future approaches for complex disease. Nat. Genet. 33(Suppl.), 228-237.

Boyle, P., and Levin, B. (2008). World cancer report 2008. In: Boyle, P., and Levin, B., editors. World Health Organization, International Agency for Research on Cancer, Lyon, France.

Breslow, N.E. (1982). Covariance adjustment of relative-risk estimates in matched studies. Biometrics 38, 661-672.

Breslow, N.E., and Day, N.E. (1980). Statistical methods in cancer research, Vol. 1: The analysis of case-control studies. International Agency for Research on Cancer, Lyon.

Breslow, N.E., Day, N.E., Halvorsen, K.T., Prentice, R.L., Sabal, C. (1978). Estimation of multiple relative risk functions in matched case-control studies. Am. J. Epidemiol. 108, 299-307.

Broman, K.W., Wu, H., Sen, Ś., Churchill, G.A. (2003). R/qtl: QTL mapping in experimental crosses. Bioinformatics 19, 889-890.

Cardon, L.R., and Palmer, L.J. (2003). Population stratification and spurious allelic association. Lancet 361, 598-604.

Carlborg, Ö., and Haley, C.S. (2004). Epistasis: too often neglected in complex trait studies? Nat. Rev. Genet. 5, 618-625.

Carlson, C.S., Newman, T.L., Nickerson, D.A. (2001). SNPing in the human genome. Curr. Opin. Chem. Biol. 5, 78-85.

Chapman, J., and Clayton, D. (2007). Detecting association using epistatic information. Genet. Epidemiol. 31, 894-909.

Chapman, N.H., and Wijsman, E.M. (1998). Genome screens using linkage disequilibrium tests: optimal marker characteristics and feasibility. Am. J. Hum. Genet. 63, 1872-1885.

Chatterjee, N., Kalaylioglu, Z., Carroll, R.J. (2005). Exploiting gene-environment independence in family-based case-control studies: increased power for detecting associations, interactions and joint effects. Genet. Epidemiol. 28, 138-156.

Chen, Y.H., Chatterjee, N., Carroll, R.J. (2008). Retrospective analysis of haplotype-based case control studies under a flexible model for gene environment association. Biostatistics 9, 81-99.

Chen, J., and Rodriguez, C. (2007). Conditional likelihood methods for haplotype-based association analysis using matched case-control data. Biometrics 63, 1099-1107.

Chen, H.-S., Zhu, X., Zhao, H., Zhang, S. (2003). Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. Ann. Hum. Genet. 67, 250-264.

Chernick, M.R. (2007). Bootstrap methods: A guide for practitioners and researchers, 2nd Edition. Wiley, New York.

Cheverund, J.M., and Routman, E.J. (1995). Epistasis and its contribution to genetic variance components. Genetics 139, 1455-1461.

Clark, A.G. (1990). Inference of haplotypes from PCR-amplified samples of diploid populations. Mol. Biol. Evol. 7, 111-122.

Clark, A.G. (2004). The role of haplotypes in candidate gene studies. Genet. Epidemiol. 27, 321-333.

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

Clayton, D. (2007). Population association. In: Balding, D.J., Bishop, M., Cannings, C., editors, Handbook of statistical genetics, third edition. Wiley.

Cohen, J.C., Kiss, R.S., Pertsemlidis, A., Marcel, Y.L., McPherson, R., Hobbs, H.H. (2004). Multiple rare alleles contribute to low plasma levels of hdl cholesterol. Science 305, 869-872.

Collins, F.S., Green, E.D., Guttmacher, A.E., Guyer, M.S. (2003). A vision for the future of genomics research. Nature 422, 835-47.

Cordell, H.J. (2009). Genome-wide association studies: Detecting gene-gene interactions that underlie human diseases. Nat. Rev. Genet. 10, 392-404.

Cordell, H.J., Barratt, B.J., Clayton, D.G. (2004). Case/pseudo control analysis in genetic association studies: A unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. Genet. Epidemiol. 26, 167-185.

Cordell, H.J. and Clayton, D.G. (2002). A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. Am. J. Hum. Genet. 70, 124-141.

Cordell, H.J., and Clayton, D.G. (2005). Genetic association studies. Lancet 366, 1121-1131.

Cordell, H.J., Barratt, B.J., Clayton, D.G. (2004). Case/pseudocontrol analysis in genetic association studies: A unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. Genet. Epidemiol. 26, 167-185.

Costanza, M.C. (1995). Matching. Preventive Medicine 24, 425-433.

Cox, D.R. (1970). The analysis of binary data. Methuen, London.

Curtis, D. (1997). Use of siblings as controls in case-control association studies. Ann. Hum. Genet. 61, 319-333.

Davidson, S. (2000). Research suggests importance of haplotypes over SNPs. Nat. Biotechnol. 18, 1134-1135.

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. J. R. Statist. Soc. B 39, 1-38.

Devlin, B., and Roeder, K. (1999). Genomic control for association studies. Biometrics 55, 997-1004.

Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H., Goldstein, D.B. (2010). Rare variants create synthetic genome-wide associations. PLoS Biol. 8:e1000294. doi:10.1371/journal.pbio.1000294.

Dudbridge, F. (2003). Pedigree disequilibrium tests for multilocus haplotypes. Genet. Epidemiol. 25, 115-121.

Dudbridge, F. (2007). Family-based association. In: Balding, D.J., Bishop, M., Cannings, C., editors, Handbook of statistical genetics, 3rd edition. Wiley.

Durrant, C., Zondervan, K.T., Cardon, L.R., Hunt, S., Deloukas, P., Morris, A.P. (2004). Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. Am. J. Hum. Genet. 75, 35-43.

Efron, B. (1982). The Jackknife, the Bootstrap, and other resampling plans. SIAM, Philadelphia.

Efron, B. (1987). Better bootstrap confidence intervals. JASA 82, 171-200.

Efron, B. (1990). Six questions raised by the bootstrap. Technical Report No. 350, Department of Statistics, Stanford University.

Efron, B., and Tibshirani, R.J. (1993). An Introduction to the Bootstrap. Chapman and Hall, New York.

Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., Nadeau, J.H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. 11, 446-450.

Eitan, Y., and Kashi, Y. (2002). Direct micro-haplotyping by multiple double PCR amplifications of specific alleles (MD-PASA). Nucleic Acids Res. 30:e62.

Epstein, M.G., Allen, A., Satten, G.A. (2007). A simple and improved correction for population stratification in case-control studies. Am. J. Hum. Genet. 80, 921-930.

Epstein, M.G., and Satten, G.A. (2003). Inference on haplotype effects in case-control studies using unphased genotype data. Am. J. Hum. Genet. 73, 1316-1329.

Ewens, W.J., and Spielman, R.S. (1995). The transmission/disequilibrium test: history, subdivision and admixture. Am. J. Hum. Genet. 57, 455-464.

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

Fallin, D., Cohen, A., Essioux, L., Chumakov, I., Blumenfield, M., Cohen, D., Schork, N. (2001). Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. Genome Res. 11, 143-151.

Fallin, D., and Schork, N. (2000). Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid data. Am. J. Hum. Genet. 67, 947-959.

Feng, T., Elston, R.C., Zhu, X. (2011). Detecting rare and common variants for complex traits: sibpair and odds ratio weighted sum statistics (SPWSS, ORWSS). Genet. Epidemiol. 35, 398-409.

Finkelstein, E.A., Trogdon, J.G., Brown, D.S., Allaire, B.T., Dellea, P.S., Kamal-Bahl, S.J. (2008). The lifetime medical cost burden of overweight and obesity: implications for obesity prevention. Obesity 16, 1843-1848.

Fitze, G., Cramer, J., Ziegler, A., Schierz, M., Schreiber, M., Kuhlisch, E., Roesner, D., Schackert, H.K. (2002). Association between c135G/A genotype and RET proto-oncogene germline mutations and phenotype of Hirschsprung's disease. Lancet 359, 1169-1170.

Fitzmaurice, G.M., Laird, N.M., Ware, J.H. (2004). Applied longitudinal analysis. John Wiley and Sons, New York.

Gauderman, W.J. (2002). Sample size requirements for matched case-control studies of gene-environment interaction. Stat. Med. 21, 35-50.

Gauderman, W.J., Witte, J.S., Thomas, D.C. (1999). Family-based association studies. J. Natl. Cancer Inst. Monogr. 26, 31-37.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B. (2003). Bayesian data analysis, 2nd edition. Chapman and Hall, London.

Gelman, A., Jakulin, A., Pittau, M.G., Su, Y.S. (2008). A weakly informative default prior distribution for logistic and other regression models. Ann. Appl. Stat. 2, 1360-1383.

Gelman, A., and Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, New York.

Goldstein, A.M., Hodge, S.E., Haile, R.W. (1989). Selection bias in case-control studies using relatives as the controls. Int. J. Epidemiol. 18, 985-989.

Gorlov, I.P., Gorlova, O.Y., Sunyaev, S.R., Spitz, M.R., Amos, C.I. (2008). Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am. J. Hum. Genet. 82, 100-112.

Greenland, S., and Neutra, R. (1981). An analysis of detection bias and proposed corrections in the study of estrogens and endometrial cancer. J. Chronic. Dis. 34, 433-438.

Guan, W., Liang, L., Boehnke, M., Abecasis, G.R. (2009). Genotype-based matching to correct for population stratification in large-scale case-control genetic association studies. Genet. Epidemiol. 33, 508-517.

Guo, W., and Lin, S. (2009). Generalized linear modeling with regularization for detecting common disease rare haplotype association. Genet. Epidemiol. 33, 308-316.

Hahn, R.A., Truman, B.I., Barker, N.D. (1996). Identifying ancestry: the reliability of ancestral identification in the United States by self, proxy, interviewer, and funeral director. Epidemiology 7, 75-80.

Hartl, D.L., and Clark, A.G. (2006). Principles of population genetics, 4th edition. Sinauer Associates, Inc, Sunderland, MA.

Hein, R., Beckmann, L., Chang-Claude, J. (2009). Comparison of different haplotype-based association methods for gene-environment (G×E) interactions in case-control studies when haplotype-phase is ambiguous. Hum. Hered. 68, 252-267.

Hindorff, L.A., Junkins, H.A., Hall, P.N., Mehta, J.P., Manolio, T.A. (2010). A catalog of published genome-wide association studies. Available at: www.genome.gov/gwastudies. Accessed December 9, 2010.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA 106, 9362-9367.

Hoffmann, T.J., Marini, N.J., Witte, J.S. (2010). Comprehensive approach to analyzing rare genetic variants. PLoS ONE 5:e13584. doi:10.1371/journal.pone.0013584.

Hoggart, C.J., Parra, E.J., Shriver, M.D., Bonilla, C., Kittles, R.A., Clayton, D.G., McKeigue, P.M. (2003). Control of confounding of genetic associations in stratified populations. Am. J. Hum. Genet. 72, 1492-1504.

Horvath, S., Xu, X., Lake, S.L., Silverman, E.K., Weiss, S.T., Laird, N.M. (2004) Family-based tests for associating haplotypes with general phenotype data: application to asthma genetics. Genet. Epidemiol. 26, 61-69.

Hsu, L., Zhao, L.P., Aragaki, C. (2000). A note on a conditional-likelihood approach for family-based association studies of candidate genes. Hum. Hered. 50, 194-200.

Huang, W., Wang, P., Liu, Z., Zhang, L. (2009). Identifying disease associations via genome-wide association studies. BMC Bioinformatics 10(Suppl 1), S68. doi:10.1186/1471-2105-10-S1-S68.

Ihaka, R., and Gentleman, R. (1996). R: a language for data analysis and graphics. J. Comp. Graph. Stat. 5, 299-314.

Jablon, S., Neel, J.V., Gershowitz, H., Atkinson, G.F. (1967). The NAS-NRC twin panel: Methods of construction of the panel, zygosity diagnosis, and proposed use. Am. J. Hum. Genet. 19, 133-161.

Jewell, N.P. (2003). Statistics for epidemiology. Chapman & Hall, New York.

Ji, W., Foo, J.N., O'Roak, B.J., Zhao, H., Larson, M.G., Simon, D.B., Newton-Cheh, C., State, M.W., Levy, D. Lifton, R.P. (2008). Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat. Genet. 40, 592-99.

Joosten, P.H., Toepoel, M., Mariman, E.C., Van Zoelen, E.J. (2001). Promoter haplotype combinations of the platelet-derived growth factor alpha-receptor gene predispose to human neural tube defects. Nat. Genet. 27, 215-217.

Kaplan, N.L., and Morris, R.W. (2001). Issues concerning association studies for fine mapping a susceptibility gene for a complex disease. Genet. Epidemiol. 20, 432-457.

Khoury, M.J., and Yang, Q. (1998). The future of genetic studies of complex human diseases: an epidemiologic perspective. Epidemiology 9, 350-354.

King, H., Aubert, R.E., Herman, W.H. (1998). Global burden of diabetes, 1995-2025: prevalence, numerical estimates, and projections. Diab. Care 21, 1414-31.

Kitsios, G.D., and Zintzaras, E. (2010). An NOS3 haplotype is protective against hypertension in a Caucasian population. Int. J. Hypertens. 25;2010:865031 doi:10.4061/2010/865031.

Kooperberg, C., LeBlanc, M.L., Dai, J.Y., Rajapakse I. (2009). Structures and assumptions: strategies to harness gene × gene and gene × environment interactions in GWAS. Statistical Science 24, 472-488.

Kraft, P., Cox, D.G., Paynter, R.A., Hunter, D., De Vivo, I. (2005). Accounting for haplotype uncertainty in matched association studies: a comparison of simple and flexible techniques. Genet. Epidemiol. 28, 261-272.

Kraft, P., and Stram, D.O. (2007). Re: The use of inferred haplotypes in downstream analysis. Am. J. Hum. Genet. 81, 863-865.

Kraft, P., and Thomas, D.C. (2000). Bias and efficiency in family-matched gene-characterization studies: Conditional, prospective, retrospective, and joint likelihoods. Am. J. Hum. Genet. 66, 1119-1131.

Kupper, L.L., Karon, J.M., Kleinbaum, D.G., Morgenstern, H., Lewis, D.K. (1981). Matching in epidemiologic studies: validity and efficiency considerations. Biometrics 37, 271-292.

Kwee, L.C., Epstein, M.P., Manatunga, A.K., Duncan, R., Allen, A.S., Satten, G.A. (2007). Simple methods for assessing haplotype-environment interactions in case-only and case-control studies. Genet. Epidemiol. 31, 75-90.

Laird, N.M., and Lange, C. (2006). Family-based designs in the age of large-scale gene-association studies. Nat. Rev. Genet. 7, 385-394.

Laird, N.M., and Lange, C. (2009). The role of family-based designs in genome wide association studies. Statistical Science 24, 388-397.

Lake, S.L., Lyon, H., Tantisira, K., Silverman, E.K., Weiss, S.T., Laird, N.M,, Schaid, D.J. (2003). Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. Hum. Hered. 55, 56-65.

Lange, E.M., Sun, J., Lange, L.A., Zheng, S.L., Duggan, D., Carpten, J.D., Gronberg, H., Isaacs, W.B., Xu, J., Chang, B.L. (2008). Family-based samples can play an important role in genetic association studies. Cancer Epidemiol. Biomarkers Prev. 17, 2208-14.

Lee, W.C. (2004). Case-control association studies with matching and genomic controlling. Genet. Epidemiol. 27, 1-13.

Lei, Z., Liu, R.Y., Zhao, J., Liu, Z., Jiang, X., You, W., Chen, X., Liu, X., Zhang, K., Pasche, B., Zhang, H. (2009). TGFBR1 haplotypes and risk of non-small-cell lung cancer. Cancer Res. 69, 7046-52.

Lesaffre, E., and Albert, A. (1989). Partial separation in logistic discrimination. J. R. Statist. Soc. B 51, 109-116.

Levenstien, M.A., Ott, J., Gordon, D. (2006). Are molecular haplotypes worth the time and expense? A cost-effective method for applying molecular haplotypes. PLoS Genet. 2(8):e127.

Li, C.C. (1969). Population subdivision with respect to multiple alleles. Ann. Hum. Genet. 33, 23-29.

Li, C., and Boehnke, M. (2006). Haplotype association analysis for late onset diseases using nuclear family data. Genet. Epidemiol. 30, 220-230.

Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. 83, 311-321.

Li, J., Zhang, K., Yi, N. (2011). A Bayesian hierarchical model for detecting haplotype-haplotype and haplotype-environment interactions in genetic association studies. Hum. Hered. 71, 148-160.

Li, X., and Li, J. (2007). Comparison of haplotyping methods using families and unrelated individuals on simulated rheumatoid arthritis data. BMC Proc. 1(Suppl 1), S55.

Lin, D. Y., and Huang, B. E. (2007). The use of inferred haplotypes in downstream analyses. Am. J. Hum. Genet. 80, 577-579.

Lin, D.Y., and Zeng, D. (2006). Likelihood-based inference on haplotype effects in genetic association studies. J. Am. Stat. Assoc. 101, 89-118.

Lin, D.Y., Zeng, D., Millikan, R.. (2005). Maximum likelihood estimation of haplotype effects and haplotype-environment interactions in association studies. Genet. Epidemiol. 29, 299-312.

Liu, N., Zhang, K., Zhao, H. (2008). Haplotype-association analysis. Adv Genet. 60, 335-405.

Liu, J., Papasian, C., Deng, H.W.. (2007). Incorporating single-locus tests into haplotype cladistic analysis in case-control studies. PLoS Genet. 3:e46.

Liu, P.Y., Zhang, Y.Y., Lu, Y., Long, J.R., Shen, H., Zhao, L.J., Xu, F.H., Xiao, P., Xiong, D.H., Liu, Y.J., Recker, R.R., Deng, H.W. (2005). A survey of haplotype variants at several disease candidate genes: the importance of rare variants for complex diseases. J. Med. Genet. 42, 221-227.

Lopez, A.D., Mathers, C.D., Ezzati, M., Jamison, D.T., Murray, C.J.L. (2006). Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. Lancet 367, 1747-1757.

Luan, J.A., Wong, M.Y., Day, N.E., Wareham, N.J. (2001). Sample size determination for studies of gene-environment interaction. Int. J. Epidemiol. 30, 1035-40.

MacLean, C.J., Sham, P.C., Kendler, K.S. (1993). Joint linkage of multiple loci for a complex disorder. Am. J. Hum. Genet. 53, 353-366.

Maher, B. (2008). Personal genomes: the case of the missing heritability. Nature 456, 18-21.

Malosetti, M., van der Linden, C.G., Vosman, B., van Eeuwijk, F.A. (2007). A mixed-model approach to association mapping using pedigree information with an illustration of resistance to phytophthora infestans in potato. Genetics 175, 879-889.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J.H., Guttmacher, A.E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, E.N., Slatkin, M., Valle, D., Whittemore, A.S., Boehnke, M., Clark, A.G., Eichler, E.E., Gibson, G., Haines, J.L., Mackay, T.F.C., McCarroll, S.A., Visscher, P.M. (2009). Finding the missing heritability of complex diseases. Nature 461, 747-753.

Martin, E.R. (2006). Linkage disequilibrium and association analysis. In: Haines, J.L., Pericak-Vance, M., editors, Genetic analysis of complex diseases. 2nd edition. John Wiley and Sons, New York.

McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P.A., Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty, and challenges. Nat. Rev. Genet. 9, 356-369.

McCullagh, P., and Nelder, J.A. (1989). Generalized linear models, 2nd edition. Chapman and Hall, London.

McGinnis, R., Shifman, S., Darvasi, A. (2002). Power and efficiency of the TDT and case-control design for association scans. Behav. Genet. 32, 135-144.

McKeigue, P.M. (2007). Population admixture and stratification in genetic epidemiology. In: Balding, D.J., Bishop, M., Cannings, C., editors, Handbook of statistical genetics, third edition. Wiley.

Michalatos-Beloin, S., Tishkoff, S., Bentley, K., Kidd, K., Ruano, G. (1996). Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR. Nucleic Acids Res. 24, 4841-4843.

Miettinen, O.S. (1970). Estimation of relative risk from individually matched series. Biometrics, 26, 75-86.

Moore, J.H. (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum. Hered. 56, 73-82.

Moore, J.H. (2005). A global view of epistasis. Nat. Genet. 37, 13-14.

Morris, R.W., and Kaplan, N.L. (2002). On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genet. Epidemiol. 23, 221-233.

Morris, A.P., and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet. Epidemiol. 34, 188-193.

Mukherjee, B., Ahn, J., Gruber, S.B., Rennert, G., Moreno, V., Chatterjee, N. (2008). Tests for gene-environment interaction from case-control data: a novel study of type I error, power and designs. Genet. Epidemiol. 32, 615-26.

Nejentsev, S.,Walker, N., Riches, D., Egholm, M., Todd, J.A. (2009). Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324, 387-89.

Niu, T., Qin, Z.S., Xu, X., Liu, J.S. (2002). Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am. J. Hum. Genet. 70, 157-169.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association. Nat. Genet. 38, 904-909.

Price, A.L., Zaitlen, N.A., Reich, D., Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. Nat. Rev. Genet. 11, 459-463.

Pritchard, J.K. (2001). Are rare variants responsible for susceptibility to complex diseases? Am. J. Hum. Genet. 69, 124-137.

Pritchard, J.K., Stephens, M., Donnelly, P. (2000b). Inference of population structure using multilocus genotype data. Genetics 155, 945-959.

Pritchard, J.K., Stephens, M., Rosenberg, N.A., Donnelly, P. (2000a). Association mapping in structured populations. Am. J. Hum. Genet. 67, 170-181.

Purcell, S., and Sham, P. (2004). Properties of structured association approaches to detecting population stratification. Hum. Hered. 58, 93-107.

Purcell, S., Sham, P., Daly, M.J. (2005). Parental phenotypes in family-based association analysis. Am. J. Hum. Genet. 76, 249-259.

Qin, Z., Niu, T., Liu, J. (2002). Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. Am. J. Hum. Genet. 71, 1242-1247.

Rabinowitz, D., and Laird, N. (2000). A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Hum. Hered. 504, 227-233.

Ramahi, T.M. (2010). Cardiovascular disease in the Asia middle east region: global trends and local implications. Asia-Pacific Journal Public Health 22(3 suppl), 83S-89S.

Reich, D. E., and Goldstein, D. B. (2001). Detecting association in a case-control study while correcting for population stratification. Genet. Epidemiol. 20, 4-16.

Risch, N.J. (2000). Searching for genetic determinants in the new millennium. Nature 405, 847-856.

Risch, N.J., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. Science 273, 1516-1517.

Robinson, R. (2010). Common disease, multiple rare (and distant) variants. PLoS Biol. 8:e1000293. doi:10.1371/journal.pbio.1000293.

Rothman, K.J., Greenland, S., Lash, T.L. (2008a). Design strategies to improve study accuracy. In: Rothman KJ, Greenland S, Lash TL, editors, Modern epidemiology, 3rd edition. Lippincott Williams & Wilkins, Philadelphia.

Rothman, K.J., Greenland, S., Lash, T.L. (2008b). Applications of stratified analysis methods. In: Rothman, K.J., Greenland, S., Lash, T.L., editors, Modern epidemiology, 3rd edition. Lippincott Williams & Wilkins, Philadelphia.

Rothman, K.J., Greenland, S., Lash, T.L. (2008c). Case-control studies. In: Rothman, K.J., Greenland, S., Lash, T.L., editors, Modern epidemiology, 3rd edition. Lippincott Williams & Wilkins, Philadelphia.

Satten, G.A., Flanders, W.D., Yang, Q. (2001). Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. Am. J. Hum. Genet. 68, 466-477.

Schaid, D.J. (1996). General score tests for associations of genetic markers with disease using cases and their parents. Genet. Epidemiol. 13, 423-449.

Schaid, D. J. (2004). Evaluating associations of haplotypes with traits. Genet. Epidemiol. 27, 348-364.

Schaid, D.J., Rowland, C.M., Tines, D.E., Jacobson, R.M., Poland, G.A. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am. J. Hum. Genet. 70, 425-434.

Schlesselman, J.J. (1982). Case-control studies: design, conduct, analysis. Oxford University Press, Oxford.

Scott, W.K., Pericak-Vance, M.A., Haines, J.L. (1997). Genetic analysis of complex diseases. Science 275, 1327-1330.

Self, S.G., Longton, G., Kopecky, K.J., Liang, K.Y. (1991). On estimating HLA/disease association with application to a study of aplastic-anemia. Biometrics 47, 53-61.

Seltman, H., Roeder, K., Devlin, B.. (2001). Transmission/disequilibrium test meets measured haplotype analysis: Family-based association analysis guided by evolution of haplotypes. Am. J. Hum. Genet. 68, 1250-1263.

Seltman, H., Roeder, K., Devlin, B. (2003). Evolutionary-based association analysis using haplotype data. Genet. Epidemiol. 25, 48-58.

Semsei, A.F., Erdélyi, D.J., Ungvári, I., Kámory, E., Csókay B., Andrikovics, H., Tordai, A., Ságoly, E., Falus, A., Kovács, G.T., Szalai, C. (2008). Association of some rare haplotypes and genotype combinations in the MDR1 gene with childhood acute lymphoblastic leukaemia. Leuk. Res. 32, 1214-20.

Sha, Q., Dong, J., Jiang, R., Zhang, S. (2005). Tests of association between quantitative traits and haplotypes in a reduced-dimensional space. Ann. Hum. Genet. 69, 715-732.

Sham, P. (1998). Statistics in human genetics. Arnold, London.

Shastry, B.S. (2006). Pharmacogenetics and the concept of individualized medicine. Pharmacogenomics Journal 6, 16-21.

Siegmund, K.D., Langholz, B., Kraft, P., Thomas, D.C. (2000). Testing linkage disequilibrium in sibships. Am. J. Hum. Genet. 67, 244-248.

Sinha, S., Gruber, S.B., Mukherjee, B., Rennert, G. (2008). Inference of the haplotype effect in a matched case-control study using unphased genotype data. The International Journal of Biostatistics 4(1), Article 6. doi: 10.2202/1557-4679.1079.

Song, P.X.-K. (2007). Correlated data analysis: modeling, analytics, and applications. Springer, New York.

Souverein, O.W., Zwinderman, A.H., Jukema, J.W., Tanck, M.W. (2008). Estimating effects of rare haplotypes on failure time using a penalized Cox proportional hazards regression model. BMC Genet. 9, 9. doi:10.1186/1471-2156-9-9.

Spielman, R.S., and Ewens, J.E. (1998). A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. Am. J. Hum. Genet. 62, 450-458.

Spielman, R.S., McGinnis, R.E., Ewens, W.J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am. J. Hum. Genet. 52, 506-513.

Spinka, C., Carroll, R.J., Chatterjee, N. (2005). Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. Genet. Epidemiol. 29, 108-127.

Stephens, J.C., Schneider, J.A., Tanguay, D.A., Choi, J., Acharya, T., Stanley, S.E., Jiang, R., Messer, C.J., Chew, A., Han, J.H., Duan, J., Carr, J.L., Lee, M.S., Koshy, B., Kumar, A.M., Zhang, G., Newell, W.R., Windemuth, A., Xu, C., Kalbfleisch, T.S., Shaner, S.L., Arnold, K., Schulz, V., Drysdale, C.M., Nandabalan, K., Judson, R.S., Ruano, G., Vovis, G.F. (2001). Haplotype variation and linkage disequilibrium in 313 human genes. Science 293, 489-493.

Stephens, M., Smith, N.J., Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. Am. J. Hum. Genet. 68, 978-989.

Stram, D.O., Pearce, C.L., Bretsky, P., Freedman, M., Hirschhorn, J.N., Altshuler, D., Kolonel, L.N., Henderson, B.E., Thomas, D.C. (2003). Modeling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. Hum. Hered. 55, 179-190.

Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. Scandinavian Journal of Statistics 1, 49-58.

Tavtigian, S.V., Simard, J., Teng, D.H., Abtin, V., Baumgard, M., Beck, A., Camp, N.J., Carillo, A.R., Chen, Y., Dayananth, P., Desrochers, M., Dumont, M., Farnham, J.M., Frank, D., Frye, C., Ghaffari, S., Gupte, J.S., Hu, R., Iliev. D., Janecki, T., Kort, E.N., Laity, K.E., Leavitt, A., Leblanc, G., McArthur-Morrison, J., Pederson, A., Penn, B., Peterson, K.T., Reid, J.E., Richards, S., Schroeder, M., Smith, R., Snyder, S.C., Swedlund, B., Swensen, J., Thomas, A., Tranchant, M., Woodland, A.M., Labrie, F., Skolnick, M.H., Neuhausen, S., Rommens, J., Cannon-Albright, L.A. (2001). A candidate prostate cancer susceptibility gene at chromosome 17p. Nat. Genet. 27, 172-180.

The International HapMap Consortium (2005). A haplotype map of the human genome. Nature 437, 1299-1320.

Thomas, D.C. (2004). Statistical methods in genetic epidemiology. Oxford University Press, Oxford.

Thomas, D.C. (2010a). Gene-environment-wide association studies: emerging approaches. Nat. Rev. Genet. 11, 259-272.

Thomas, D.C. (2010b). Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. Annu. Rev. Public. Health. 31, 21-36.

Thomas, D.C., and Greenland, S. (1983). The relative efficiencies of matched and independent sample designs for case-control studies. J. Chronic. Dis. 36, 685-697.

Tzeng, J.Y. (2005). Evolutionary-based grouping of haplotypes in association analysis. Genet. Epidemiol. 28, 220-231.

Umbach, D.M., and Weinberg, C.R. (2000). The use of case-parent triads to study joint effects of genotype and exposure. Am. J. Hum. Genet. 66, 251-261.

Valle, T., Tuomilehto, J., Bergman, R.N., Ghosh, S., Hauser, E.R., Eriksson, J., Nylund, S.J., Kohtamäki K., Tuomilehto-Wolf, E., Toivanen, L., Vidgren, G., Ehnholm, C., Blaschak, J., Langefeld, C.D., Watanabe, R.M., Magnuson, V., Ally, D.S., Hagopian, W.A., Ross, E., Buchanan, T.A., Collins, F., Boehnke, M. (1998). Mapping genes for non-insulin dependent diabetes mellitus: design of the Finland-United States investigation of NIDDM genetics (FUSION) study. Diab. Care 21, 949-958.

Vansteelandt, S., DeMeo, D.L., Lasky-Su, J., Smoller, J.W., Murphy, A.J., McQueen, M., Schneiter, K., Celedon, J.C., Weiss, S.T., Silverman, E.K., Lange, C. (2008). Testing and estimating gene-environment interactions in family-based association studies. Biometrics, 64, 458-467.

van't Veer, L.J., and Bernards, R. (2008). Enabling personalized cancer medicine through analysis of gene-expression patterns. Nature 452, 564-70.

Waldman, I.D., Robinson, B.F. Rowe, D.C. (1999). A logistic regression based extension of the TDT for continuous and categorical traits. Ann. Hum. Genet. 63, 329-340.

Witte, J.S., Gauderman, W.J., Thomas, D.C. (1999). Asymptotic bias and efficiency in case-control studies of candidate genes and gene environment interactions: basic family designs. Am. J. Epidemiol. 149, 693-705.

Wolf, J.B., Brodie III, E.D., Wade, M.J. (2000). Epistasis and the Evolutionary Process. Oxford University Press, New York.

Wray, N.R., and Goddard, M.E. (2010). Multi-locus models of genetic risk of disease. Genome Medicine 2:10 doi:10.1186/gm131.

Wu, C.F.J. (1983). On the convergence properties of the EM algorithm. Annals of Statistics 11, 95-103.

Yende, S., Angus, D.C., Ding, J., Newman, A.B., Kellum, J.A., Li, R., Ferrell, R.E., Zmuda, J., Kritchevsky, S.B., Harris, T.B., Garcia, M., Yaffe, K., Wunderink, R.G., for the Health ABC Study (2007). 4G/5G plasminogen activator inhibitor-1 polymorphisms and haplotypes are associated with pneumonia. Am. J. Respir. Crit. Care Med. 176, 1129-37.

Yi, N. (2010). Statistical analysis of genetic interactions. Genet. Res. (Camb). 92: 443-459.

Yi, N., and Banerjee, S. (2009). Hierarchical generalized linear models for multiple quantitative trait locus mapping. Genetics 181, 1101-1113.

Yi, N., Kaklamani, V.G., Pasche, B. (2010). Bayesian analysis of genetic interactions in case-control studies, with application to adiponectin genes and colorectal cancer risk. Ann. Hum. Genet. 75, 90-104.

Yi, N., and Xu, S. (2008). Bayesian LASSO for quantitative trait loci mapping. Genetics 179, 1045-1055.

Yu, J., Pressoir, G., Briggs, W.H., Vroh, B.I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S., Buckler, E.S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38, 203-208.

Zaykin, D.V., Westfall, P.H., Young, S.S., Karnoub, M.A., Wagner, M.J., Ehm. M.G. (2002). Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum. Hered. 53, 79-91.

Zhang, F., Wang, Y., Deng, H-W. (2008). Comparison of population-based association study methods correcting for population stratification. PLoS ONE 3:e3392. oi:10.1371/journal.pone.0003392.

Zhang, K., and Zhao, H. (2006). A comparison of several methods for haplotype frequency estimation and haplotype reconstruction for tightly linked markers for general pedigrees. Genet. Epidemiol. 30, 423-437.

Zhang, K., and Zhao, H. (2010). Family-based association studies. In: Lin, S., and Zhao, H., editors, Handbook on analyzing human genetic data: computational approaches and software. Springer, New York.

Zhang, H., Zhang, H., Li, Z. Zheng, G. (2007). Statistical methods for haplotype-based matched case-control association studies. Genet. Epidemiol. 31, 316-326.

Zhang, H., Zheng, G., Li, Z. (2006). Statistical analysis for haplotype-based matched case-control studies. Biometrics 62, 1124-1131.

Zhao, K., Aranzana, M.J., Kim, S., Lister, C., Shindo, C., Tang, C., Toomajian, C., Zheng, H., Dean, C., Marjoram, P., Nordborg, M. (2007). An Arabidopsis example of association mapping in structured samples. PLoS Genet. 3:e4.

Zhao, J.H., Curtis, D., Sham, P.C. (2000). Model-free analysis and permutation tests for allelic associations. Hum. Hered. 50, 133-139.

Zhao, L.P., Li, S.S., Khalid, N. (2003). A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. Am. J. Hum. Genet. 72, 1231-1250.

Zhao, H., Zhang, S., Merikangas, K.R., Trixler, M., Wildenauer, D.B., Sun, F., Kidd, K.K. (2000). Transmission/disequilibrium tests using multiple tightly linked markers. Am. J. Hum. Genet. 67, 936-946.

Zhao, J., Jin, L., Xiong, M. (2006). Test for interaction between two unlinked loci. Am. J. Hum. Genet. 79, 831-845.

Zhu, X., Fejerman, L., Luke, A., Adeyemo, A., Cooper, R.S. (2005). Haplotypes produced from rare variants in the promoter and coding regions of angiotensinogen contribute to variation in angiotensinogen levels. Human Molecular Genetics 14, 639-643.

Zhu, X., Feng, T., Li, Y., Lu, Q., Elston, R.C. (2010). Detecting rare variants for complex traits using family and unrelated data. Genet. Epidemiol. 34, 171-187.

Ziegler, A., and Koenig, I. (2007). A statistical approach to genetic epidemiology. Wiley-VCH.

Zondervan, K,T., and Cardon, L.R. (2004). The complex interplay among factors that influence allelic association. Nat. Rev. Genet. 5, 89-100.