Abstract

JONES, MARTHA LOUISE. A Retrospective Method for Inference on Haplotype Main Effects and Haplotype-environment Interactions Using Clustered Haplotypes. (Under the direction of Dr. Jung-Ying Tzeng.)

Many regression-based methods exist for conducting haplotype association analysis in case-control studies. Such methods generally are based on either a prospective framework (modeling the probability of disease conditional on haplotypes and covariates) or a retrospective framework (modeling the probability of haplotypes and covariates conditional on disease). For haplotype analysis, both theoretical and simulation work have demonstrated that a retrospective framework can be more efficient than a prospective framework in this context. Given this result, we aim to improve the performance of the retrospective haplotype framework by more efficiently modeling haplotype information. We do so by clustering evolutionarily close haplotypes and studying the effects of haplotype clusters. Previous work has shown that the strategy of clustering haplotypes under the prospective framework can increase the power of haplotype-based association analysis. This work extends the clustering idea to the retrospective framework and improves the performance of haplotype analysis for case-control studies.

Specifically, we construct a retrospective likelihood that allows for environmental covariates and interactions between haplotypes and covariates. We derive generalized score statistics to test for haplotype main effects and interaction effects at the global and individual levels. We also derive tests for interaction effects that can be applied to case subjects only. Through simulation, we assess the validity of the proposed tests and, where appropriate, compare the power with the power of the retrospective full-dimensional and prospective analyses. We also present a new strategy for

evaluating haplotype specific effects that allows us to identify haplotypes that have similar effects on disease. Finally, we apply our proposed method to real data from a genetic study of hypertriglyceridemia.

A Retrospective Method for Inference on Haplotype Main Effects and Haplotype-environment Interactions Using Clustered Haplotypes

by

Martha L. Jones

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

August 2007

APPROVED BY:

DR. JUNG-YING TZENG (CO-CHAIR)

DR. JEFF THORNE (CO-CHAIR)

DR. MARIE DAVIDIAN

DR. MEG EHM

Dedication

For Mom and Dad.

Biography

Martha (Marti) Jones was born in Lynchburg, VA in 1976, and continued the family legacy by majoring in mathematics at James Madison University in Harrisonburg, VA. (Marti's parents both majored in mathematics at Guilford College in Greensboro, NC and her sister majored in mathematics at the University of North Carolina at Chapel Hill.) On a whim, she minored in statistics and began working as a Mathematical Statistician for the US Census Bureau around the time of the 2000 Decennial Census. She enjoyed her time at the Census Bureau, but did not want to work with survey data for the rest of her career. Following her interest in biostatistics, she came to NC State University to obtain a Master's degree in statistics, and ended up staying for a PhD. She began working in the Statistical Sciences group at GlaxoSmithKline in Research Triangle Park as a Graduate Industrial Trainee in May of 2002. Later her interests expanded to include statistical genetics, and she continued working as an intern in the Pharmacogenetics group at GSK until May 2007.

Acknowledgements

I would first like to thank my adviser, Dr. Jung-Ying Tzeng. She has been the best adviser I can imagine and has made this process as enjoyable as it could be. I am very grateful for her support and guidance over the last 2 years. I would also like to thank Dr. Meg Ehm for her support and the opportunity to gain invaluable experience working as a Graduate Industrial Trainee in the Pharmacogenetics group at GlaxoSmithKline. I also appreciate the feedback and guidance from the other members of my committee: Dr. Marie Davidian, Dr. Jeff Thorne, and Dr. Zhao-Bang Zeng.

I must also acknowledge the support and friendship of Dr. Cheryl Lindeman (aka A.C.). She has been an excellent mentor to me since high school, and was one of the first people to encourage me to consider graduate school. Now she can finally add my name to the plaque at CVGS!

I must thank all of the great friends I have made while at NC State, they are really the only reason this has been worth it. Thanks to Michael, Shufang, Venita, Matt, Karen, Ray, Theresa, Aarthi, Paul, Mike, Kirsten, and Lavanya. And I never would have made it without the support, advice, and friendship of Amanda, Amy, Joe, Lovely, Kristen and Alvin. Thanks also to my non-NC State friends: Heather, Susan, Careyanne, Taylor, and Erin. They have provided an invaluable support network and have always reminded me that I had a life before grad school and will, hopefully, have one afterwards.

Thanks to mom and dad who have told me since I was young that I could be anything I wanted to. They always believed I could do anything I put my mind to, even when I didn't. My sister Amanda has been there for me through all of this. I

don't know what I would have done if she had been any further away than Chapel Hill.

And lastly, thanks to Scott. He is the best witness to the blood, sweat, and (mostly) tears that went into this. He let me cry when I needed to cry, laugh when I needed to laugh, and always reminded me of the light at the end of the tunnel. This was only possible because of his love and support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Problem of Interest

The goal of genetic association studies is to find the disease susceptibility genes that increase the risk for developing complex diseases. Study designs may involve collecting data from groups of related individuals or from a cohort of patients over time, but more often researchers use a case-control study design. A case-control study identifies subjects who have a particular trait (e.g. a disease) and then identifies appropriate control subjects. Researchers are then interested in the relationship between genetic factors and the disease status of the subjects. Many genetic association studies today measure the genetic variation through single nucleotide polymorphisms (SNPs). A SNP is a variation in DNA consisting of a single base change. Through identifying SNPs that are associated with disease, one can approximately locate the region of disease susceptibility genes. This strategy is based on the conjecture that the observed SNPs are causal variants themselves, or more often, are in linkage disequilibrium with a causal variant.

It is generally believed that jointly analyzing multiple SNPs together, such as in the format of haplotypes, may be more effective in identifying gene-trait associations. A haplotype is a set of alleles from several loci that are located on the same chromosome and inherited as a unit. Haplotypes represent a unit of inheritance and preserve the linkage disequilibrium among the loci. Previous studies have shown that haplotype analysis can be more powerful than single SNP analysis, especially when there are multiple disease causing variants (Morris and Kaplan (2002)). But there are practical limitations that limit the use of haplotype analysis in practice. One limitation is that haplotype phase is usually not known. The most commonly used genotyping methods do not provide information on which alleles occur on the same chromosome. Therefore we know which alleles occur at each loci, but not which alleles occur together across multiple loci. Molecular haplotyping methods that can provide this information exist, but they are expensive and not practical for use on large samples of genetic data. Alternatively, statistical methods can handle missing phase information by using the expectation-maximization (EM) algorithm and using genotype data to infer haplotype frequencies. Several early haplotype association studies demonstrated that the EM algorithm is successful at accurately estimating haplotype frequencies (Fallin and Schork (2000), Zhao et al. (2000)).

Haplotype analysis can have a power disadvantage due to an increasing number of parameters as the number of possible haplotypes increases. This problem is even more pronounced when considering interactions between haplotypes and covariates. When relatively rare haplotypes exist, degrees of freedom are expended on haplotypes and interaction terms that we have a limited ability to detect, even if a significant association exists. The common way of dealing with this problem is to exclude the rare haplotypes from the analysis. However, by doing so, one has to discard samples of rare haplotypes. One way to retain all of the sample information while reducing the degrees of freedom is to cluster haplotypes. The advantage of clustering is that it does not throw away the information provided by rare haplotypes. Most haplotype clustering methods fall into two major categories: methods that use haplotype similarity and methods that use evolutionary relationships (also called cladistic methods). Haplotype similarity methods define a pairwise similarity measure and then assess how the similarity differs between cases and controls. The assumption is that if cases have inherited a disease-causing variant from a common founder haplotype, they will also have inherited more of the same alleles at nearby markers, resulting in a greater degree of haplotype similarity than controls (Yu et al. (2004)). Molitor et al. (2003), Yu et al. (2004), and Waldron and Whittaker (2006) have all developed methods for genetic association analysis based on haplotype similarity. Clustering methods that use evolutionary relationships are based on representing how haplotypes evolve from an ancestral haplotype with a cladogram. A cladogram is a tree diagram used to represent a genetic evolutionary history. Templeton et al. (1987), Templeton (1995), Seltman et al. (2003), and Durrant et al. (2004) all use cladistic approaches to cluster haplotypes. Seltman et al. (2003) develop a cladogram-collapsing algorithm that performs 1-df tests to see if two haplotypes (or clusters) occurring at nearby nodes of the cladogram should be collapsed into a new cluster. The challenge of cladistic methods is that they rely on the knowledge of the true, but unknown, evolutionary tree. As a result, most of the cladistic methods have to estimate this tree and then cluster haplotypes based on the estimated tree. The haplotype clustering method of Tzeng (2005) bypasses this issue by probabilistically assigning haplotypes to clusters based on all possible evolutionary relationships. Instead of inferring the most likely relationship and performing the clustering, the method assigns probabilities to all possible relationships based on the relatedness between haplotypes and the ages of haplotypes. Tzeng et al. (2006) incorporate this clustering method into a generalized linear model framework and show that the clustered approach has greater power than the full dimensional approach for detecting global haplotype main effects.

Most complex diseases are thought to be caused by both genetic and environmental factors, as well as interactions between them. Methods to analyze effects of interactions between haplotypes and environmental factors in case-control studies are relatively new, and are generally based on either a prospective framework (modeling the probability of disease status conditional on haplotype and covariates) or a retrospective framework (modeling the probability of haplotype and covariates conditional on disease status). The method of Lake et al. (2003), which is incorporated into the `haplo.stats` package in `R`, estimates and tests for haplotype-covariate interactions using a prospective approach under the generalized linear model framework. The classic result of Prentice and Pyke (1979) says that analyzing case-control data with a logistic regression model and treating the data as if they were sampled prospectively does not affect the estimates of the odds ratios. They show that the retrospective likelihood function can be factored into a piece that looks like the prospective likelihood and a piece that is only a function of the regression variable. But Carroll et al. (1995) show that standard errors calculated assuming a prospective likelihood will only be correct if the distribution of the regression variable is unrestricted. In haplotype analysis, the regression variable of interest is the haplotype, while only genotypes are observable. To reconstruct haplotypes from genotypes, we must make an assumption on the haplotype pair distribution (usually Hardy-Weinberg Equilibrium (HWE)) to ensure that the haplotype frequencies will be identifiable. The HWE assumption results in a restricted distribution for the regression variable. Epstein and Satten (2003) have shown that using a retrospective likelihood is more efficient when making assumptions about the distribution of covariates. As a result, it would be more appropriate and more efficient to take into account the case-control sampling scheme and use a retrospective likelihood to study the effects of haplotypes, environmental covariates, and their interactions. One difficulty with using a retrospective likelihood is specifying the distribution for the environmental covariates. Several approaches, including the work of Chatterjee and Carroll (2005), Spinka et al. (2005), Chen et al. (2007), Lin and Zeng (2006), Chen and Kao (2006), and Kwee et al. (2007), have been developed to study haplotype effects based on a retrospective likelihood while being able to incorporate covariates and interactions. These approaches differ in how the distribution of the covariates is handled and what assumptions are imposed on the haplotype-environment relationship, HWE, and the prevalence of the disease being investigated. Chatterjee and Carroll (2005) and Spinka et al. (2005) assume haplotype-environment independence and HWE in the population, but do not make any assumptions about the disease prevalence. Lin and Zeng (2006) assume that the disease of interest is rare, but allow for Hardy-Weinberg disequilibrium and only assume that haplotype and the environmental covariate are independent conditional on genotype. Chen and Kao (2006) and Kwee et al. (2007) both assume haplotype-environment independence in the population, but Kwee et al. (2007) assume a rare disease and HWE in the population, while Chen and Kao (2006) only assume HWE in the control population.

These methods all recognize gains in efficiency from using a retrospective likelihood when analyzing effects of haplotypes and haplotype-environment interactions, but they are still all limited by large degrees of freedom from including all possible haplotypes in the model. We propose to further improve the performance of the retrospective framework by allowing for haplotype clustering. Kwee et al. (2007) and Spinka et al. (2005) both suggest that using techniques to reduce the haplotype space could improve the power of their methods. But there is no method available yet that incorporates clustering and a retrospective likelihood. We incorporate the clustering method of Tzeng (2005) into a retrospective likelihood that assumes haplotype-environment independence, HWE in the target population, and a rare disease. The clustering algorithm can easily be incorporated into the logistic regression framework, and we derive tests for main haplotype and interaction effects that are practical to implement. Our method handles data with unknown haplotype phase, addresses the power limitation due to an increasing number of parameters, and can be used to study haplotype-environment interactions.

1.2 Literature Review

Schaid et al. (2002) presented a prospective method to test for haplotype association that can be used with binary, quantitative, or ordinal responses. Their method can be used with unphased genetic data and can adjust for the effect of environmental covariates. They use a generalized linear model framework to develop score tests for assessing global haplotype association and haplotype specific effects. The method is implemented in the `haplo.stats` package in `R`. Lake et al. (2003) extended the work of Schaid et al. (2002) to incorporate interactions between haplotype and covariates. Lake et al. (2003) use the EM algorithm to iteratively estimate the haplotype frequencies and regression parameters. Parameter estimates can be obtained from the `haplo.glm` function in `haplo.stats`, and can be used to create multivariate Wald statistics to test for global or specific interaction effects.

Tzeng et al. (2006) incorporated the clustering algorithm of Tzeng (2005) into a generalized linear model framework and derived score tests for global haplotype association. The clustering algorithm constructs a core set of haplotypes and probabilistically incorporates rare haplotypes into the core clusters by evaluating all possible evolutionary relationships. The method uses an entropy-based information criterion to find the balance between information and dimensionality to determine the core set of haplotypes. Through clustering, they assume that all haplotypes in a cluster have the same effect on disease. Tzeng et al. (2006) based their work on the prospective method of Schaid et al. (2002), and allow for adjustment of covariates and modeling a binary or quantitative trait. They find that their clustering method improves power when compared with the full-dimensional method.

Zhao et al. (2003) developed a prospective estimating equation approach to estimate and test for haplotype and interaction effects in case-control studies. They use score equations derived from the prospective likelihood of disease given the environmental covariate, and only require assuming HWE in the control population. Some of the advantages of the method are that it is easy to implement and suitable for estimating haplotype frequencies based on a large number of SNPs. But Satten and Epstein (2004) found the method to be inefficient when compared to more recent retrospective approaches.

Chatterjee and Carroll (2005) developed a retrospective model that assumes independence between haplotypes and environmental covariates in the population, and showed that this model is more efficient when estimating regression parameters than the traditional logistic regression approach. Their method used a semiparametric framework in which a nonparametric distribution is assumed for the covariate. They then used the profile likelihood technique to obtain maximum likelihood estimates for the parameters of interest. Spinka et al. (2005) extend this method to allow for haplotype phase ambiguity and missing genetic data. The method does not explicitly need the rare disease assumption, but does assume HWE in the population. When the rare disease assumption is not made, the true intercept parameter (the log odds of the baseline category, $\log(\frac{P(D=1|baseline)}{P(D=0|baseline)})$) appears in the risk function and must be estimated. Spinka et al. (2005) show that this parameter is identifiable but difficult to estimate without information on the marginal probability of disease ($P(D = 1)$). Because in practice this information on $P(D = 1)$ tends to be unavailable, Spinka et al. (2005) also proposed a modified approach that does not require estimation of the intercept. The modification assumes a rare disease so as to remove the intercept parameter from the likelihood, and is equivalent to some of the methods discussed below.

Lin et al. (2005) and Lin and Zeng (2006) presented a retrospective likelihood-based approach that assumes haplotype-environment independence conditional on the genotype data. Their method allows for Hardy-Weinberg disequilibrium, but does assume that the disease of interest is rare. They also use the profile likelihood technique to calculate the maximum likelihood estimates of the parameters. Assuming that haplotypes and environmental covariates are independent conditional on genotypes allows Lin and Zeng (2006) to avoid having to specify the distribution of the haplotypes for each level of the covariate. However in reality, if haplotypes and covariates are not independent at the population level, it is not likely that they will be independent conditional on the genotypes (Kwee et al. (2007)). Hence it is generally recognized that this conditional independence assumption does not realistically release the haplotype-environment independence assumption, but ends up adding complexity to the method.

Kwee et al. (2007) developed a simpler retrospective likelihood-based approach that assumes both a rare disease and haplotype-environment independence. These assumptions allow them to write the likelihood as a product of the likelihood of haplotypes conditional on disease and environmental covariates and the prospective likelihood of disease conditional on environmental factors. The second piece results from assuming a saturated distribution for the environmental covariate and applying the results of Prentice and Pyke (1979) to the probability of the environmental covariate conditional on disease. Specifying the distribution of the disease conditional on the environmental factors allows them, as the methods above, to avoid specifying the distribution of the covariates. Kwee et al. (2007) also assume HWE in the target population, which along with the rare disease assumption, implies HWE in the controls. It can be shown that the method of Kwee et al. (2007) and the method of Spinka et al. (2005) with the rare disease assumption are equivalent. Kwee et al. (2007) found that their retrospective approach has greater power and results in more efficient estimators than the traditional prospective approach.

Chen and Kao (2006) presented a method that is very similar to the method of Kwee et al. (2007). They introduced a multinomial logistic model where the combinations of disease status and haplotype pairs are considered the outcome variables and the environmental factors are considered as explanatory variables. By using the profile likelihood technique to profile out the distribution of the covariate, they can write their retrospective likelihood as a prospective multinomial likelihood of disease status and haplotype pair given the covariate. Chen and Kao (2006) assumed haplotype-environment independence, but instead of assuming a rare disease as Kwee et al. (2007), they only make an assumption of HWE for the distribution of haplotype pairs in the controls. Their assumption of HWE in the controls is equivalent to the assumptions made by Kwee et al. (2007) of a rare disease and HWE in the target population.

All of these methods recognize that assuming haplotype-environment independence leads to simpler models and easier parameter estimation. But Spinka et al. (2005) and Kwee et al. (2007) both have shown that violation of this assumption can lead to biased inference. There are many examples of environmental covariates of interest that we would expect to be independent of genetic factors, for example, exposure to a certain toxin. But Chatterjee and Carroll (2005) gave an example in which the genetic polymorphisms that may increase a person's risk of developing a disease from smoking can also affect the person's tendency to be addicted to smoking. In this example, the covariate of interest, whether or not a person smokes, may be dependent on the person's genetic makeup. To release the assumption of haplotype-environment independence, Chen et al. (2007) developed a method that directly models the relationship between haplotypes and environmental factors. They used a multinomial logistic regression to describe the relationsip between haplotypes and covariates while still allowing for HWE in the marginal distribution of haplotype pairs. This parametric model, along with a nonparametric model for the environmental covariate, results in a semiparametric model for the distribution of the covariates conditional on the haplotype pairs. They use semiparametric estimating equations to make inferences on the odds ratios of interest. As Spinka et al. (2005), they also proposed the rare disease assumption to avoid having to estimate the true intercept parameter and to simplify the estimation.

1.3 Dissertation Outline

In Chapter 2 we will present a retrospective likelihood that incorporates clustering and allows for environmental covariates in the model. The assumptions we make and the likelihood we use are based on the work of Kwee et al. (2007). We feel that the rare disease assumption is appropriate when considering methods for case-control studies, and the haplotype-environment independence assumption is reasonable since there are many cases where we can expect the assumption to hold. We derive generalized score statistics to test for global and specific haplotype main effects. We compare the tests for haplotype main effects to the retrospective full-dimensional method as well as the clustering and full-dimensional prospective methods. In Chapter 2, we also propose a strategy for evaluating haplotype specific effects that allows us to further group haplotypes that have similar effects on risk of disease.

In Chapter 3 we extend this likelihood to include haplotype-environment interactions, and derive score statistics to test for global and specific interaction effects. Most of the methods described in the literature only present tests for specific interaction effects, assuming the other effects are zero. To implement these interaction tests, we need prior knowledge about which haplotype will interact with the covariate, but in practice this information will not usually be known. Ideally, we wish to perform the same analysis for studying interaction effects as for studying main effects. That is, we would first carry out a global test on all interaction terms; if we detect a significant effect, we would then carry out the specific tests for interactions. However, such a procedure is usually not practical due to a lack of power from a large number of degrees of freedom. With clustered haplotypes, this test strategy is more feasible since we have reduced the parameter space of haplotypes.

In Chapter 4 we present a likelihood for a cases-only analysis and derive score statistics to test for interaction effects. We compare the results from the case-only analysis to results using a case-control sample in Chapter 3. In Chapters 2-4, we assess the validity and power of the proposed tests through simulations.

In Chapter 5 we apply our methods to data from a genetic study of hypertriglyceridemia and confirm previous findings. In Chapter 6 we present a summary of the contributions of our work and discuss future extensions.

Chapter 2

Testing for Haplotype Main Effects

2.1 Introduction

The first step in a haplotype association analysis is to assess the main effects of haplotype on the trait being studied. For the retrospective clustering method, we are interested in evaluating the effects of the core haplotypes. In Chapter 2, we will present the retrospective likelihood and derive the robust score test for testing both for global and specific haplotype effects. Robust score tests are robust to misspecification of the model for the odds of disease and do not require maximization of the observed likelihood under the alternative hypothesis. We will present simulation results evaluating the size and power of the score test, and compare the results to those obtained from the retrospective, full-dimensional analysis and the prospective, clustering and prospective full-dimensional analyses.

The retrospective clustering method incorporates the clustering algorithm of Tzeng et al. (2006) through the allocation matrix, $B(\mathbf{p})$. The algorithm divides the space of haplotypes into a core category, a group that differs from the core category by one mutation, a group that differs by two mutations, and so on until all haplotypes are assigned to a group. The clustering algorithm starts with the last group of haplotypes, and clusters each haplotype in the current group into the group that is one step closer to the core group based on haplotype similarity and age, and this continues until the space has been collapsed to the group of core haplotypes. In each step, Tzeng et al. (2006) used an allocation probability matrix to describe how a haplotype is allocated to each of the haplotypes in the previous group. Specifically, the allocation probability matrices from each step can be combined together by taking their product. The combined matrix, denoted by $B(\mathbf{p})$, describes how each haplotype is grouped into the core haplotypes. By multiplying the design matrix for the full-dimensional space of haplotypes by $B(\mathbf{p})$, we obtain the design matrix for the clustered haplotypes. This design matrix is then incorporated into the logistic regression model for disease.

For detecting the global haplotype-phenotype association, we compare our method to the prospective clustering method of Tzeng et al. (2006) and the retrospective, full-dimensional method of Kwee et al. (2007). For completeness, we also compare the method to the prospective, full-dimensional method of Schaid et al. (2002). Schaid et al. (2002) use a generalized linear model (GLM) framework and derive score tests for testing for haplotype main effects using a prospective likelihood. The method allows for environmental covariates and missing phase information, and is implemented in the `haplo.stats` package in `R`.

We will also present a new framework for evaluating the effects of specific haplotypes. If the global haplotype test is significant, the next step is to examine the effects of specific haplotypes. Investigators commonly employ one of two strategies for investigating haplotype specific tests. They may choose one haplotype as the reference haplotype and then test to see if the other haplotypes have a significant effect relative to the reference. Or they may choose one haplotype of interest and test to see if it is different relative to the pooled group of remaining haplotypes. We present an example where these strategies may lead us to incomplete or misleading conclusions about the effects of specific haplotypes. We propose an alternative procedure which carries out all pairwise comparisons between haplotypes and allows us to partition the clusters into groups that have similar effects on disease.

2.2 Methods

2.2.1 The Retrospective Likelihood for Haplotype Main Effects

Let D represent the disease status for a subject, with $D = 1$ denoting a case and $D = 0$ denoting a control. Let G represent a subject's multilocus genotype for the region of interest and let E represent a subject's value for the environmental covariate. Let H denote the haplotype pair (h, h'), where (h', h) is counted as a separate haplotype pair. We can write the observed retrospective likelihood as

$$L_{obs} = \prod_{i=1}^{n} P(G_i, E_i|D_i) = \prod_{i=1}^{n} \sum_{H \in S(G_i)} P(H, E_i|D_i)$$

where n is the total number of subjects and $H \in S(G_i)$ represents the set of haplotype pairs (h, h') that are consistent with a subject's genotype. We can further factorize the likelihood into

$$L_{obs} = \prod_{i=1}^{n} \sum_{H \in S(G_i)} P(H|E_i, D_i)P(E_i|D_i).$$

Kwee *et al* show that the above can be further written as

$$L_{obs} \propto \prod_{i=1}^{n} \sum_{H \in S(G_i)} P(H|E_i, D_i)P(D_i|E_i). \tag{2.1}$$

This comes from the result of Prentice and Pyke (1979) that says the retrospective likelihood $P(D_i|E_i)$ is proportional to the prospective likelihood $P(E_i|D_i)$ if we assume a saturated distribution for E.

We further assume that the disease of interest is rare (defined as having a prevalence <10%) and that haplotypes and the environmental covariate are independent in the target population. We first consider $P(H|E_i, D_i = 0)$, which simplifies to $P(H|D_i = 0)$ under the assumption of haplotype-environment independence. Assuming Hardy-Weinberg Equilibrium (HWE) in the target population, along with the rare disease assumption, allows us to assume that the controls will represent a sample from the population that is also in HWE. Then we can write

$$P(H|D_i = 0) = p_h p_{h'}.$$

Epstein and Satten (2003) show that the distribution of haplotypes conditional on disease and coviarates in cases can be written as

$$\begin{aligned}
P(H|E_i, D_i = 1) &= \frac{P(D_i = 1|H, E_i)P(H)P(E_i)}{\sum_{H'} P(D_i = 1|H', E_i)P(H')P(E_i)} \\
&= \frac{\theta(H, E_i)P(D_i = 0|H, E_i)P(H)}{\sum_{H'} \theta(H', E_i)P(D_i = 0|H', E_i)P(H')} \\
&= \frac{\theta(H, E_i)P(D_i = 0|H)P(H)}{\sum_{H'} \theta(H', E_i)P(D_i = 0|H')P(H')} \\
&= \frac{\theta(H, E_i)P(H|D_i = 0)P(D_i = 0)}{\sum_{H'} \theta(H', E_i)P(H'|D_i = 0)P(D_i = 0)} \\
&= \frac{\theta(H, E_i)P(H|D_i = 0)}{\sum_{H'} \theta(H', E_i)P(H'|D_i = 0)}
\end{aligned}$$

where $\theta(H, E) = \frac{P(D=1|H,E)}{P(D=0|H,E)}$ and is the odds of disease for haplotype pair H and covariate E. Using a logistic regression model for the probability of disease, we write the odds as

$$\theta(H, E) = \exp(\alpha + \mathbf{X_{C_H}}\boldsymbol{\beta} + \mathbf{X_E}\boldsymbol{\gamma})$$

where $\mathbf{X_{C_H}}$ is the row of the clustering design matrix X_C that corresponds to haplotype pair H, $\boldsymbol{\beta}$ is the vector of haplotype cluster effects, $\mathbf{X_E}$ is the design matrix for the environmental covariates, and $\boldsymbol{\gamma}$ is the vector of covariate effects. The clustering design matrix is a product of the design matrix for the full dimensional space of haplotypes and the clustering allocation matrix. We can write $\mathbf{X_C}$ as

$$\mathbf{X_C} = \mathbf{X_F}\mathbf{B}(\mathbf{p}).$$

$\mathbf{X_F}$ has dimension $(L+1)^2$ by $(L+1)$ where $(L+1)$ is the number of observed haplotypes. $\mathbf{X_F}$ has a row for each of the $(L+1)^2$ possible pairs of observed haplotypes and a column for each observed haplotype. Each row of $\mathbf{X_F}$ corresponds to a unique (with respect to order) haplotype pair and each column corresponds to an observed haplotype. The design matrix assumes a multiplicative genetic model, which says that the haplotypes are multiplicative in their effect on the odds of disease. $\mathbf{X_{F_H}}$ is the row corresponding to haplotype pair $H = (h, h')$ and counts the number of each haplotype in the haplotype pair. The matrix $B(\mathbf{p})$ is a function of the haplotype frequencies and has dimension $(L + 1)$ by $(L^* + 1)$ where $(L^* + 1)$ is the number of clusters. Matrix $B(\mathbf{p})$ contains the allocation probabilities that describe how the $(L + 1)$ haplotypes are grouped into the $(L^* + 1)$ clusters. We write $B(\mathbf{p})$ as

$$B(\mathbf{p}) = \begin{pmatrix} B_{11}(\mathbf{p}) & \dots & B_{1(L^*+1)}(\mathbf{p}) \\ B_{21}(\mathbf{p}) & \dots & B_{2(L^*+1)}(\mathbf{p}) \\ \vdots & \vdots & \vdots \\ B_{(L+1)1}(\mathbf{p}) & \dots & B_{(L+1)(L^*+1)}(\mathbf{p}) \end{pmatrix}$$

where each element B_{jk} describes how haplotype j is allocated to cluster k. We can write $\mathbf{X_F B(p)}$ as

$$\mathbf{X_F B(p)} = \begin{pmatrix} \sum_{h=1}^{(L+1)} X_{F_{1h}} B_{h1}(\mathbf{p}) & \dots & \sum_{h=1}^{(L+1)} X_{F_{1h}} B_{h(L^*+1)}(\mathbf{p}) \\ \sum_{h=1}^{(L+1)} X_{F_{2h}} B_{h1}(\mathbf{p}) & \dots & \sum_{h=1}^{(L+1)} X_{F_{2h}} B_{h(L^*+1)}(\mathbf{p}) \\ \vdots & \vdots & \vdots \\ \sum_{h=1}^{(L+1)} X_{F_{(L+1)^2 h}} B_{h1}(\mathbf{p}) & \dots & \sum_{h=1}^{(L+1)} X_{F_{(L+1)^2 h}} B_{h(L^*+1)}(\mathbf{p}) \end{pmatrix}.$$

To determine $P(D_i|E_i)$, Kwee *et al* show we can write the odds of disease given E

as,

$$\theta(E) = \frac{P(D=1|E)}{P(D=0|E)} = \sum_H \theta(H,E)P(H|E,D=0).$$

Therefore

$$\theta(E) = \sum_H \exp(\alpha + \mathbf{X}_{C_H}\boldsymbol{\beta} + \mathbf{X}_E\boldsymbol{\gamma})p_h p_{h'}.$$

Since

$$\begin{aligned} 1+\theta(H,E) &= \frac{P(D=0|E)}{P(D=0|E)} + \frac{P(D=1|E)}{P(D=0|E)} \\ &= \frac{1}{P(D=0|E)}, \end{aligned}$$

we can write $P(D=0|E) = \frac{1}{1+\theta(E)}$ and $P(D=1|E) = \frac{\theta(E)}{1+\theta(E)}$. In logistic regression analysis of case-control data we cannot estimate the true intercept α, so we replace it with a modified intercept that we can estimate. Let $\alpha^* = \alpha - \frac{\text{probability a case is sampled}}{\text{probability a control is sampled}}$ and use it to replace α in $\theta(E)$. We will call this new odds $\theta^*(E)$. We can now write (2.1) as

$$L_{obs} \propto \prod_{i=1}^{n} \left[\frac{\sum_{H\in S(G_i)} \exp(\alpha + \mathbf{X}_{C_H}\boldsymbol{\beta} + \mathbf{X}_{E_i}\boldsymbol{\gamma})p_h p_{h'}}{\sum_H \exp(\alpha + \mathbf{X}_{C_H}\boldsymbol{\beta} + \mathbf{X}_{E_i}\boldsymbol{\gamma})p_h p_{h'}} \frac{\theta^*(E_i)}{1+\theta^*(E_i)} \right]^{d_i} \left[\frac{\sum_{H\in S(G_i)} p_h p_{h'}}{1+\theta^*(E_i)} \right]^{1-d_i} \tag{2.2}$$

The term $\exp(\alpha)$ in the numerator and denominator of the first part of the likelihood for a case will cancel, and we can replace it in the numerator with the α^* from $\theta^*(E_i)$. The remaining terms in $\theta^*(E_i)$ cancel with $\sum_H \exp(X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma})p_h p_{h'}$ in the denominator. This leads us to a simplified version of (2.2)

$$L_{obs} \propto \prod_{i=1}^{n} \left[\frac{\sum_{H\in S(G_i)} \exp(\alpha^* + \mathbf{X}_{C_H}\boldsymbol{\beta} + \mathbf{X}_{E_i}\boldsymbol{\gamma})p_h p_{h'}}{1+\theta^*(E_i)} \right]^{d_i} \left[\frac{\sum_{H\in S(G_i)} p_h p_{h'}}{1+\theta^*(E_i)} \right]^{1-d_i} \tag{2.3}$$

2.2.2 Score Test for Global Haplotype Effects

The global test for main haplotype effects tests the hypothesis that all of the haplotype cluster parameters are 0, $H_0 : \boldsymbol{\beta} = 0$. The generalized score statistic is $S_{\boldsymbol{\beta}} = U_{\boldsymbol{\beta}}^T V_{\boldsymbol{\beta}}^{-1} U_{\boldsymbol{\beta}} \Big|_{\boldsymbol{\beta}=0, \boldsymbol{\xi}=\tilde{\boldsymbol{\xi}}}$ where $U_{\boldsymbol{\beta}}$ is the score function and $V_{\boldsymbol{\beta}}$ is the generalized variance function for $\boldsymbol{\beta}$. $S_{\boldsymbol{\beta}}$ has a χ^2 distribution with L^* degrees of freedom. Let $\boldsymbol{\xi}$ be the vector of nuisance parameters, which consists of α^*, the covariate parameter vector $\boldsymbol{\gamma}$ and the vector of haplotype frequencies $\mathbf{p}$. The score function is defined as

$$U_{\boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}} \log L_{obs}$$

and

$$V_{\boldsymbol{\beta}} = D_{\boldsymbol{\beta\beta}} - I_{\boldsymbol{\beta\xi}} I_{\boldsymbol{\xi\xi}}^{-1} D_{\boldsymbol{\beta\xi}}^T - D_{\boldsymbol{\beta\xi}} I_{\boldsymbol{\xi\xi}}^{-1} I_{\boldsymbol{\beta\xi}}^T + I_{\boldsymbol{\beta\xi}} I_{\boldsymbol{\xi\xi}}^{-1} D_{\boldsymbol{\xi\xi}} I_{\boldsymbol{\xi\xi}}^{-1} I_{\boldsymbol{\beta\xi}}^T,$$

(Boos (1992)). D is the variance-covariance matrix of the score function $U = (U_{\boldsymbol{\beta}}, U_{\boldsymbol{\xi}})^T$ and I is the observed Fisher information matrix which is constructed by taking first derivatives of $U_{\boldsymbol{\beta}}$ and $U_{\boldsymbol{\xi}}$. See Appendix A for the detailed expressions for these quantities.

Estimation

We evaluate the score statistic using estimates of the nuisance parameters under the null hypothesis that $\boldsymbol{\beta} = 0$. Under the null hypothesis, 2.3 becomes

$$L_{obs} = \prod_{i=1}^{n} \left[\frac{\exp(\alpha^* + X_{E_i}\gamma) \sum_{H \in S(G_i)} p_h p_{h'}}{1 + \exp(\alpha^* + X_{E_i}\gamma)} \right]^{d_i} \left[\frac{\sum_{H \in S(G_i)} p_h p_{h'}}{1 + \exp(\alpha^* + X_{E_i}\gamma)} \right]^{1-d_i}.$$

We see that the observed likelihood factors into terms only involving the haplotype frequencies $\mathbf{p}$ and terms only involving the regression parameters α^* and $\boldsymbol{\gamma}$. For estimating $\mathbf{p}$, we consider the terms of the observed likelihood that contain $\mathbf{p}$:

$$L_{obs} \propto \prod_{i=1}^{n} \sum_{H \in S(G_i)} p_h p_{h'}.$$

If we assume haplotype phase is known, we can write the full data likelihood involving $\mathbf{p}$ as

$$L_{full} \propto \prod_{(h,h')} (p_h p_{h'})^{(c_{hh'}+d_{hh'})},$$

where $c_{hh'}$ is the number of controls with haplotype pair (h, h') and $d_{hh'}$ is the number of cases with haplotype pair(h, h'). Therefore, the full data likelihood has a multinomial distribution and we can use the EM algorithm implemented in the `haplo.em` function in R to estimate $\mathbf{p}$.

For estimating α^* and $\boldsymbol{\gamma}$, we write the terms of the observed likelihood that contain the regression parameters:

$$L_{obs} \propto \prod_{i=1}^{n} \frac{\exp(\alpha^* + \mathbf{X}_{E_i}\boldsymbol{\gamma})^{d_i}}{1 + \exp(\alpha^* + \mathbf{X}_{E_i}\boldsymbol{\gamma})}.$$

α^* and $\boldsymbol{\gamma}$ are the parameter estimates from regressing the response d_i on the environmental covariate E. Therefore, we can use the `glm` function in R to obtain estimates for these parameters.

2.2.3 Score Test for Specific Haplotype Main Effect

The test for a specific haplotype tests the hypothesis that the effect of the specific haplotype is 0 (compared to the reference haplotype) while the other haplotype effects are unconstrained. If we are interested in haplotype t, the null hypothesis will be $H_o : \beta_t = 0$. Here the score function is

$$U_{\beta_t} = \frac{\partial}{\partial \beta_t} \log L_{obs}$$

and the generalized variance function is

$$V_{\beta_t} = D_{\beta_t\beta_t} - I_{\beta_t\boldsymbol{\xi}} I^{-1}_{\boldsymbol{\xi}\boldsymbol{\xi}} D^T_{\beta_t\boldsymbol{\xi}} - D_{\beta_t\boldsymbol{\xi}} I^{-1}_{\boldsymbol{\xi}\boldsymbol{\xi}} I^T_{\beta_t\boldsymbol{\xi}} + I_{\beta_t\boldsymbol{\xi}} I^{-1}_{\boldsymbol{\xi}\boldsymbol{\xi}} D_{\boldsymbol{\xi}\boldsymbol{\xi}} I^{-1}_{\boldsymbol{\xi}\boldsymbol{\xi}} I^T_{\beta_t\boldsymbol{\xi}}$$

where $\boldsymbol{\xi}$ now consists of α^*, $\boldsymbol{\gamma}$, $\mathbf{p}$, and the set of haplotype cluster parameters excluding β_t (i.e. $\boldsymbol{\beta}_{(-t)}$). $S_{\beta_t} = U^T_{\beta_t} V^{-1}_{\beta_t} U_{\beta_t}$ and has a χ^2 distribution with 1 degree of freedom.

Estimation

Because we must now estimate all but one of the haplotype cluster parameters under H_0, we use the expectation-conditional-maximization (ECM) algorithm (Meng and Rubin (1993)) to iteratively estimate the regression parameters and haplotype frequencies. Under the null hypothesis that $\beta_t = 0$, the full data likelihood assuming the missing data (haplotype phase) is known is

$$L_{full} = \prod_{k=1}^{e} \prod_{(h,h')} \left[\frac{p_h p_{h'}}{1 + \sum_H \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma}) p_h p_{h'}} \right]^{c_{hh',k}}$$
$$\left[\frac{\exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma}) p_h p_{h'}}{1 + \sum_H \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma}) p_h p_h} \right]^{d_{hh',k}}.$$

The steps for estimating the nuisance parameters are

1. Obtain initial estimates of $\mathbf{p}$, α^*, $\boldsymbol{\gamma}$, and $\boldsymbol{\beta}$.

2. For step s, use E Step to estimate $c_{hh',k}^{(s)}$ and $d_{hh',k}^{(s)}$.

3. Use M step 1 to find $\mathbf{p}^{(s+1)}$ that maximizes L_{full}, using $\boldsymbol{\beta}^{(s)}$, $\boldsymbol{\gamma}^{(s)}$, $\alpha^{*(s)}$, $\mathbf{p}^{(s)}$, $c_{hh',k}^{(s)}$ and $d_{hh',k}^{(s)}$.

4. Use M step 2 to find $\alpha^{*(s+1)}$, $\boldsymbol{\gamma}^{(s+1)}$ and $\boldsymbol{\beta}^{(s+1)}$ that maximize L_{full}, using $\mathbf{p}^{(s+1)}$, $c_{hh',k}^{(s)}$, and $d_{hh',k}^{(s)}$.

5. Check differences between parameters at steps $(s+1)$ and s.

6. If difference for at least one of the parameters is greater than a specified limit, start over with step 2 to estimate $c_{hh',k}^{(s+1)}$, and $d_{hh',k}^{(s+1)}$ using $\mathbf{p}^{(s+1)}$, $\boldsymbol{\beta}^{(s+1)}$, $\boldsymbol{\gamma}^{(s+1)}$, and $\alpha^{*(s+1)}$.

E step

The E step estimates the number of controls with haplotype pair (h, h') and covariate level k as

$$\begin{aligned} E(c_{hh',k}) &= \sum_g c_{g,k} I(H \in S(g)) P(H|g) \\ &= \sum_g c_{g,k} I(H \in S(g)) \frac{p_h p_{h'}}{\sum_{H \in S(g)} p_h p_h} \end{aligned}$$

and the number of cases with haplotype pair (h, h') and covariate level k as

$$\begin{aligned} E(d_{hh',k}) &= \sum_g d_{g,k} I(H \in S(g)) P(H|g) \\ &= \sum_g d_{g,k} I(H \in S(g)) \frac{\exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma}) p_h p_{h'}}{\sum_{H \in S(g)} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma}) p_h p_h}, \end{aligned}$$

where $c_{g,k}$ and $d_{g,k}$ are the number of controls and cases, respectively, with genotype g and covariate level k.

M step 1: Estimating p

To estimate $\mathbf{p}$, we maximize the log of the full data likelihood subject to the constraint $\sum_h p_h = 1$. We introduce a Lagrange multiplier λ and call the new likelihood L_M:

$$\begin{aligned}
L_M &= \log L_{full} + \lambda(\sum_h p_h - 1) \\
&= \sum_{k=1}^{e} \sum_{(h,h')} \Big[c_{hh',k} \log(p_h p_{h'}) + d_{hh',k}[(\alpha^* + X_{C_{hh'}}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma}) + \log(p_h p_{h'})] - \\
&\quad \log(1 + \sum_H \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma}) p_h p_h)(c_{hh',k} + d_{hh',k}) \Big] + \lambda(\sum_h p_h - 1) \\
&= \sum_{k=1}^{e} \sum_{(h,h')} \Big[c_{hh',k} \log(p_h p_{h'}) + d_{hh',k}[(\alpha^* + X_{C_{hh'}}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma}) + \log(p_h p_{h'})] - \\
&\quad \log(1 + \sum_H \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma}) p_h p_h)(c_{hh',k} + d_{hh',k}) \Big] + \lambda(\sum_h p_h - 1) \\
&= \sum_{k=1}^{e} \sum_{(h,h')} \Big[\log(p_h p_{h'})(c_{hh',k} + d_{hh',k}) + d_{hh',k}(\alpha^* + X_{C_{hh'}}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma}) - \\
&\quad n_{hh',k} \log(1 + \sum_H \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma}) p_h p_h) \Big] + \lambda(\sum_h p_h - 1).
\end{aligned}$$

The portion of L_M that depends on $\mathbf{p}$ is

$$\begin{aligned}
L_M &\propto \sum_{k=1}^{e} \sum_{(h,h')} \Big[\log(p_h p_{h'}) n_{hh',k} - \\
&\quad n_{hh',k} \log(1 + \sum_H \exp(\alpha^* + X_H\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma}) p_h p_{h'}) \Big] + \lambda(\sum_h p_h - 1) \\
&\propto \sum_{k=1}^{e} \big(\sum_h m_{h,k} \log(p_h) \big) - \\
&\quad \sum_{k=1}^{e} \sum_{(h,h')} \big(n_{hh',k} \log(1 + \sum_H \exp(\alpha^* + X_H\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma}) p_h p_{h'}) \big) + \lambda(\sum_h p_h - 1),
\end{aligned}$$

where $m_{h,k}$ is the number of haplotypes of type h with covariate level k. The derivative of this expression with respect to a specific p_τ is

$$\begin{aligned}
\frac{\partial}{\partial p_\tau} L_M \propto & \sum_{k=1}^{e} \frac{m_{\tau,k}}{p_\tau} - \\
& \sum_{k=1}^{e} \sum_{(h,h')} \left\{ n_{hh',k} \frac{\sum_H \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma}) I(h=\tau) 2p_{h'}}{1 + \sum_H \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma}) p_h p_{h'}} \right\} + \lambda \\
\propto & \sum_{k=1}^{e} \frac{m_{\tau,k}}{p_\tau} - \\
& \sum_{k=1}^{e} \left\{ n_k \frac{\sum_{h'} \exp(\alpha^* + X_{C_{\tau,h'}}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma}) 2p_{h'}}{1 + \sum_H \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma}) p_h p_{h'}} \right\} + \lambda
\end{aligned}$$

This expression is still difficult to solve for an analytical expression for p_τ, therefore we define the quantity $u(\mathbf{p})_\tau$ as

$$u(\mathbf{p})_\tau = \sum_{k=1}^{e} n_k \frac{\sum_{h'} \exp(\alpha^* + X_{C_{\tau,h'}}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma}) 2p_{h'}}{1 + \sum_H \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma}) p_h p_{h'}}$$

Then we can write the derivative with respect to p_τ as

$$\frac{\partial}{\partial p_\tau} L_m \propto \sum_{k=1}^{e} \frac{m_{\tau,k}}{p_\tau} - u(\mathbf{p})_\tau + \lambda$$

By setting the above equal to 0 and solving for p_τ we obtain the updating equation for p_τ

$$p_\tau = \frac{\sum_{k=1}^{e} m_{\tau,k}}{(u(\mathbf{p})_\tau - \lambda)}. \tag{2.4}$$

We carry out another iteration within the M step for estimating $\mathbf{p}$ by estimating

$u(\mathbf{p})^{(s,k)}$ based on $\mathbf{p}^{(s)}$, then estimating $\mathbf{p}^{(s,k+1)}$ based on $u(\mathbf{p})^{(s,k)}$. This iteration continues until the difference between $\mathbf{p}^{(s,k)}$ and $\mathbf{p}^{(s,k+1)}$ is less than a specified limit. Then $\mathbf{p}^{(s,k+1)}$ becomes $\mathbf{p}^{(s+1)}$.

M step 2: Estimating α^*, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$

The log of the full likelihood that depends on the regression parameters is

$$\begin{aligned}\log L_{full} \propto & \sum_{k=1}^{e} \sum_{(h,h')} -n_{hh',k} \log(1 + \sum_{H} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma}) p_h p_{h'}) \\ & + d_{hh',k}\big[\log(\exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma})) + \log p_h p_{h'}\big] \\ \propto & \sum_{k=1}^{e} \sum_{(h,h')} -n_{hh',k} \log(1 + \sum_{H} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma}) p_h p_{h'}) \\ & + d_{hh',k}(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma})\end{aligned}$$

We obtain maximum likelihood estimates for α^*, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ using the optimization function `nlminb` in R.

2.2.4 Testing Framework for Haplotype Specific Tests

After a global effect is detected, the next step is to make inferences about specific haplotype cluster effects. The ideal outcome from the haplotype cluster specific tests is to be able to partition the clusters into groups that have similar effects on risk of disease. Suppose we have 6 haplotype clusters with clusters 3 and 5 having a significant effect on disease and clusters 1, 2, 4, and 6 having no effect. We want our testing framework to conclude that clusters 1, 2, 4, and 6 belong in a different group than clusters 3 and 5. In addition, we also want to know if clusters 3 and 5 have different or similar effects on disease, and therefore if they should be assigned to the same group or 2 different groups. We propose a new strategy that considers all pairwise comparisons between haplotypes by assigning each of the haplotypes in turn to be the reference, and then testing for relative effects of the others.

To evaluate haplotype specific tests, often researchers choose one haplotype to be the reference and test the effects of the others relative to it. This strategy may be appropriate if we know in advance there is one haplotype of interest. But in practice we usually do not have any prior knowledge of a haplotype of interest, and therefore wish to examine the relationships between all of the haplotypes. Arbitrarily assigning a haplotype as the reference may make our conclusions dependent on which haplotype we designate as the reference. Another common strategy is to compare one haplotype of interest to the group of remaining haplotypes. This strategy may not be adequate to group the haplotypes, especially when haplotypes being grouped together have opposite effects.

As an example, we simulated data where we have 7 haplotype clusters defined by the haplotypes 100010, 011010, 100011, 010000, 010100, 010010, and 100000. Haplotypes 011010 and 010010 have a significant effect on the odds of disease. The odds ratio for haplotype 011010 is 3 and the odds ratio for haplotype 010010 is 2. We first assign the most frequent haplotype, 100010, as the reference and test if the others have different effects relative to it. The results of these tests are in Table 2.1, where the p-value denotes the p-value for the haplotype specific score test described in the previous section. From these results, we can conclude that 011010 and 010010 have different effects from the others, but we do not know anything about the relationship between the effects of these 2 haplotypes. Suppose instead we had assigned haplotype

Table 2.1: Haplotype Specific Testing Example: Reference haplotype=100010

Haplotype	p-values
011010	$\mathbf{3.33 \times 10^{-15}}$
100011	5.58×10^{-01}
010000	7.18×10^{-01}
010100	7.01×10^{-01}
010010	$\mathbf{9.66 \times 10^{-04}}$
100000	3.49×10^{-01}

011010 as the reference haplotype and tested the relative effects of the others. The results from these tests are in Table 2.2. From these results alone, we could only divide the haplotypes into 2 groups, one containing haplotype 011010 and one containing the remaining haplotypes. These results would be misleading because we would not be able to distinguish haplotype 010010 from the group of remaining haplotypes. But putting these results together with those from using haplotype 100010 as a reference, we can conclude that 011010 and 010010 are both different from all of the others and different from each other. If we carry out the remaining pairwise comparisons, shown in Table 2.3, we then have the complete picture of the relationships between all of the haplotype clusters.

Our haplotype specific testing framework can be compared to the cladogram collapsing clustering method of Seltman et al. (2003) discussed in Chapter 1. The method of Seltman et al. (2003) collapses high frequency core haplotypes together if they have similar effects on disease. Their method uses marginal tests by simply comparing haplotype A to haplotype B. Our test is a conditional test, as we compare one haplotype

Table 2.2: Haplotype Specific Testing Example: Reference haplotype=011010

Haplotype	p-values
100010	$\mathbf{1.34 \times 10^{-10}}$
100011	$\mathbf{6.41 \times 10^{-06}}$
010000	$\mathbf{6.20 \times 10^{-06}}$
010100	$\mathbf{5.77 \times 10^{-09}}$
010010	$\mathbf{1.92 \times 10^{-04}}$
100000	$\mathbf{1.60 \times 10^{-06}}$

Table 2.3: Haplotype Specific Testing Example: Remaining Comparisons

Reference	Haplotype	p-values
100011	010000	8.88×10^{-01}
	010100	5.54×10^{-01}
	010010	$\mathbf{3.50 \times 10^{-02}}$
	100000	2.87×10^{-01}
010000	010100	6.77×10^{-01}
	010010	$\mathbf{3.30 \times 10^{-02}}$
	100000	4.14×10^{-01}
010100	010010	$\mathbf{2.37 \times 10^{-02}}$
	100000	5.85×10^{-01}
010010	100000	$\mathbf{1.20 \times 10^{-03}}$

to the reference while conditioning on the effects of the other haplotypes. We choose to include separate terms for all of the core haplotypes in our model, but then collapse these haplotypes together if appropriate according to the results of the haplotype specific tests.

When we conduct all possible pairwise haplotype specific comparisons, we must control the familywise error rate by adjusting for multiple comparisons. We use the Bonferroni correction method, which adjusts the α level by dividing by the number of comparisons being made. The Bonferroni method is a conservative method that is easy to implement. In the example above, if the original α level was 0.05, we would declare a comparison significant if the p-value is less than 0.05/21=0.002.

2.3 Simulations

2.3.1 Scenarios

To assess the size and power of the proposed score tests, we simulated data using two different scenarios. We will describe the coalescent simulation and the FUSION simulation.

Coalescent Simulation

For the coalescent simulation, we used case-control sampling and single SNPs as the disease loci for evaluating the global test. We generated haplotype data using the coalescent model similar to the method used by Roeder et al. (2005) and Tzeng (2005). We generated 100 SNP haplotypes using a variable recombination rate to simulate regions of varying haplotype diversity. The parameters used to simulate the haplotypes were chosen so that the simulated data were similar to the *SELP* gene in the SeattleSNP database in terms of the number of SNPs and the linkage disequilibrium (LD) pattern. We then determined six different liability loci according to allele frequency and haplotype diversity in the region. We chose liability loci with frequencies, denoted by q, equal to 0.1 and 0.3 and in regions with high, moderate, and low haplotype diversity. The high haplotype diversity regions represent recombination hotspots and have an average of 10-16 observed haplotypes. The moderate haplotype diversity regions have 9-12 haplotypes, and the low diversity regions represent haplotype blocks with 5-8 haplotypes. We defined 6-SNP haplotypes using the 3 adjacent SNPs directly to the left of the liability locus and the 3 SNPs directly to the right of the liability locus. To perform the case-control sampling, we sampled 2 haplotypes with replacement from the original sample of 100 haplotypes and sampled the binary environmental covariate from a uniform distribution to form an individual's data. Case-control status is determined using a penetrance function f_{ji}, where f_{ji} is the probability that an individual is a case given they have j copies of the liability allele and level i of a binary environmental covariate. We assumed a logistic model at the liability locus so that $\text{logit}(f_{ji}) = b_0 + b_1 j + b_2 i$. We are assuming a genetic effect that is additive on the log scale by assuming the logit increases by the same amount for each copy of the liability allele. We determined the appropriate penetrance function for an individual, and if a randomly chosen number was less than f_{ji}, the subject was assigned case status, otherwise they were assigned control status. The process was repeated until we obtained 500 cases and 500 controls.

We conduct simulations with $b_1 = 0.0$ for size analysis and $b_1 = 0.5$ for power analysis. $b_1 = 0.5$ implies that the odds ratio for the liability allele is $\exp(0.5) = 1.6$.

This says that the odds of disease for one copy of the liability allele are 1.6 times the odds for zero copies of the liability allele. We set $b_2 = 0.3$ and let b_0 range from -3.0 to -3.3, depending on the frequency of the liability locus. These parameter values were chosen so that the disease prevalence would be approximately 5 % for each scenario.

To evaluate the haplotype specific test, we chose a high haplotype diversity region with 10 haplotypes. We generated data assuming 2 different liability haplotypes with frequencies 0.12 and 0.08. For 2 liability haplotypes, the penetrance function f_{jik} is the probability that an individual is a case given they have j copies of liability haplotype 1, k copies of liability haplotype 2, and level i of a binary environmental covariate. The model at the liability haplotypes becomes $\text{logit}(f_{jik}) = b_0 + b_1 j + b_2 k + b_3 i$. To evaluate power, we generated data assuming the 2 liability haplotypes had the same effect ($b_1 = b_2 = 0.5$) and also assuming they had different effects ($b_1 = 0.7, b_2 = 0.5$ and $b_1 = 0.7, b_2 = 0.3$). We set $b_3 = 0.3$ and varied b_0 to achieve a 5 % disease prevalence.

FUSION Simulation

For the FUSION simulation, we used retrospective sampling and a haplotype as the causal variant. The data simulation was based on the procedure used in Kwee et al. (2007), which uses parameter estimates based on the retrospective analysis of the Finland-United States Investigation of Non-Insulin Dependent Diabetes Mellitus (FUSION) study in Epstein and Satten (2003). Epstein and Satten (2003) analyzed 5 SNPs in a region of interest on chromosome 22 in a sample from the FUSION study and found 2 haplotypes with a significant effect on risk of Type II diabetes. We simulated data based on the observed haplotypes formed from these 5 SNPs. We used the haplotype

01100 as the causal haplotype, as Epstein and Satten (2003) found this haplotype to increase the risk of Type II diabetes using their retrospective approach. We simulated the haplotype data and binary environmental covariate for each individual conditional on disease status. We assumed haplotype-environment independence and HWE for simulating control data. For case subjects, we used the model

$$P(H=(h,h'),E=e|D=1)=\frac{\exp(X_{C_{hh'}}\boldsymbol{\beta}+X_e\gamma)p_hp_{h'}F_0(e)}{\sum_H\sum_{e^*}\exp(X_{C_{hh'}}\boldsymbol{\beta}+X_{e^*}\gamma)p_hp_{h'}F_0(e^*)},\qquad(2.5)$$

where $F_0(e)$ is the probability of having covariate e in controls. Based on the FUSION results, we set $F_0(0)=0.17$, $F_0(1)=0.83$, and $\gamma=1.39$. To evaluate Type I error, we set $\beta_{01100}=0.0$, and evaluated power for $\beta_{01100}=0.2$ and $\beta_{01100}=0.4$. We also used the same sample size as the FUSION study: 727 cases and 415 controls. To evaluate the haplotype specific test, we set $\beta_{01100}=0.5$.

For the global haplotype test, we simulated 2000 datasets to assess Type I error and 1000 datasets to assess power for $\alpha=0.05$ and $\alpha=0.01$ for both simulation scenarios. We also conducted the analyses using the retrospective, full-dimensional approach, and the prospective full-dimensional and clustering approaches on the same datasets for both simulations.

To evaluate the haplotype specific test, we simulated 1000 datasets to assess Type I error and 500 datasets to assess power. For the haplotype specific test, we only evaluated the retrospective clustering method. There is no available haplotype specific test for the prospective clustering method. The retrospective full-dimensional method is not feasible because we must estimate all but one of the full-dimensional haplotype main effect parameters. This estimation becomes prohibitively difficult when there

Table 2.4: Type I error rate for Coalescent Simulation: Global Haplotype Test ($\alpha = 0.05$)

Hap Diversity	RD retro	FD retro	RD prosp	FD prosp
High:				
q=0.1	0.048 (*0.005*)	0.044 (*0.005*)	0.050 (*0.005*)	0.050 (*0.005*)
q=0.3	0.043 (*0.005*)	0.047 (*0.005*)	0.046 (*0.005*)	0.051 (*0.005*)
Moderate:				
q=0.1	0.047 (*0.005*)	0.028 (*0.004*)	0.052 (*0.005*)	0.051 (*0.005*)
q=0.3	0.040 (*0.004*)	0.023 (*0.003*)	0.041 (*0.004*)	0.049 (*0.005*)
Low:				
q=0.1	0.054 (*0.005*)	0.044 (*0.005*)	0.052 (*0.005*)	0.047 (*0.005*)
q=0.3	0.046 (*0.005*)	0.039 (*0.004*)	0.045 (*0.005*)	0.045 (*0.005*)

RD denotes reduced-dimensional (or clustering) analysis. FD denotes full-dimensional analysis. Retro denotes retrospective analysis. Prosp denotes prospective analysis. Numbers in parentheses are Monte Carlo standard deviations.

is a large number of estimated haplotypes. The prospective full-dimensional method can be carried out using `haplo.glm` in `R`, but we did not use this method as it is not comparable to the retrospective clustering method.

2.3.2 Results

Coalescent Simulation

Tables 2.4 and 2.5 presents the Type I error rate analysis for the coalescent simulation. These results are based on 2000 replicates, with each replicate having 500 cases and 500

Table 2.5: Type I error rate for Coalescent Simulation: Global Haplotype Test ($\alpha = 0.01$)

Hap Diversity	RD retro	FD retro	RD prosp	FD prosp
High:				
q=0.1	0.011 (*0.002*)	0.012 (*0.002*)	0.013 (*0.003*)	0.016 (*0.003*)
q=0.3	0.007 (*0.002*)	0.005 (*0.002*)	0.010 (*0.002*)	0.008 (*0.002*)
Moderate:				
q=0.1	0.009 (*0.002*)	0.007 (*0.002*)	0.009 (*0.002*)	0.016 (*0.003*)
q=0.3	0.007 (*0.002*)	0.003 (*0.001*)	0.006 (*0.002*)	0.013 (*0.003*)
Low:				
q=0.1	0.011 (*0.002*)	0.009 (*0.002*)	0.011 (*0.002*)	0.008 (*0.002*)
q=0.3	0.012 (*0.002*)	0.011 (*0.002*)	0.011 (*0.002*)	0.012 (*0.002*)

RD denotes reduced-dimensional (or clustering) analysis. FD denotes full-dimensional analysis. Retro denotes retrospective analysis. Prosp denotes prospective analysis. Numbers in parentheses are Monte Carlo standard deviations.

controls. The Type I error rate for the global tests for main haplotype effects are close to the nominal level for each of the analyses. This is evidence that the χ^2 distribution accurately approximates the asymptotic distribution of the score statistics.

Tables 2.6 and 2.7 shows that the clustering retrospective analysis has the greatest power for all 6 liability loci. For the loci in the low diversity regions we see very little difference between the clustering and full-dimensional analyses. When haplotype diversity is low, the dimension reduction due to clustering is small. For the 2 low diversity regions, the average number of observed haplotypes was 6 and the average dimension reduction was 0.25. Therefore, we do not expect to see a difference in the

Table 2.6: Power for Coalescent Simulation: Global Haplotype Test ($\alpha = 0.05$)

Hap Diversity	RD retro	FD retro	RD prosp	FD prosp
High:				
q=0.1	0.849 (*0.011*)	0.775 (*0.013*)	0.843 (*0.012*)	0.772 (*0.013*)
q=0.3	0.646 (*0.015*)	0.639 (*0.015*)	0.644 (*0.015*)	0.639 (*0.015*)
Moderate:				
q=0.1	0.562 (*0.016*)	0.479 (*0.016*)	0.560 (*0.016*)	0.504 (*0.016*)
q=0.3	0.911 (*0.009*)	0.844 (*0.011*)	0.907 (*0.009*)	0.868 (*0.011*)
Low:				
q=0.1	0.839 (*0.012*)	0.827 (*0.012*)	0.835 (*0.012*)	0.826 (*0.012*)
q=0.3	0.796 (*0.013*)	0.779 (*0.013*)	0.798 (*0.013*)	0.778 (*0.013*)

RD denotes reduced-dimensional (or clustering) analysis. FD denotes full-dimensional analysis. Retro denotes retrospective analysis. Prosp denotes prospective analysis. Numbers in parentheses are Monte Carlo standard deviations.

clustering and full-dimensional analyses. There is also very little difference between the clustering and full-dimensional analyses for the loci with frequency 0.3 in a high haplotype diversity region. This is due to a large number of observed haplotypes and a relatively small dimension reduction due to clustering. For this region, the average number of observed haplotypes was 15 and the average dimension reduction was 2. The increase in power due to clustering for the retrospective method is greatest for the loci in a moderate haplotype diversity region. For the locus with frequency 0.1 in a moderate haplotype diversity region, there was an average of 12 observed haplotypes and an

Table 2.7: Power for Coalescent Simulation: Global Haplotype Test ($\alpha = 0.01$)

Hap Diversity	RD retro	FD retro	RD prosp	FD prosp
High:				
q=0.1	0.642 (*0.015*)	0.541 (*0.016*)	0.640 (*0.015*)	0.549 (*0.016*)
q=0.3	0.401 (*0.015*)	0.377 (*0.015*)	0.403 (*0.016*)	0.384 (*0.015*)
Moderate:				
q=0.1	0.319 (*0.015*)	0.257 (*0.014*)	0.314 (*0.015*)	0.269 (*0.014*)
q=0.3	0.769 (*0.013*)	0.643 (*0.015*)	0.763 (*0.013*)	0.685 (*0.015*)
Low:				
q=0.1	0.660 (*0.015*)	0.645 (*0.015*)	0.663 (*0.015*)	0.644 (*0.015*)
q=0.3	0.598 (*0.016*)	0.573 (*0.016*)	0.597 (*0.016*)	0.565 (*0.016*)

RD denotes reduced-dimensional (or clustering) analysis. FD denotes full-dimensional analysis. Retro denotes retrospective analysis. Prosp denotes prospective analysis. Numbers in parentheses are Monte Carlo standard deviations.

Table 2.8: Type I error Rate and Power for Coalescent Simulation: Haplotype Specific Test ($\alpha = 0.05$)

True Effect	Global	Hap1	Hap2
$\beta_1 = \beta_2$=0.0	0.039 (*0.006*)	0.025 (*0.005*)	0.029 (*0.005*)
$\beta_1 = \beta_2$=0.5	0.958 (*0.009*)	0.892 (*0.014*)	0.794 (*0.018*)
β_1=0.7, β_2=0.3	0.994 (*0.003*)	0.998 (*0.002*)	0.312 (*0.021*)
β_1=0.7, β_2=0.5	0.998 (*0.002*)	1.000 ($<$*0.001*)	0.798 (*0.018*)

Numbers in parentheses are Monte Carlo standard deviations.

Table 2.9: Type I error Rate and Power for Coalescent Simulation: Haplotype Specific Test ($\alpha = 0.01$)

True Effect	Global	Hap1	Hap2
$\beta_1 = \beta_2$=0.0	0.007 (*0.003*)	0.005 (*0.002*)	0.003 (*0.002*)
$\beta_1 = \beta_2$=0.5	0.854 (*0.016*)	0.716 (*0.020*)	0.556 (*0.022*)
β_1=0.7, β_2=0.3	0.976 (*0.007*)	0.988 (*0.005*)	0.120 (*0.015*)
β_1=0.7, β_2=0.5	0.990 (*0.004*)	0.980 (*0.006*)	0.570 (*0.022*)

Numbers in parentheses are Monte Carlo standard deviations.

average dimension reduction of 3. For the locus with frequency 0.3 in a moderate haplotype diversity region, there was an average of 14 observed haplotypes and an average dimension reduction of 4.

The simulation results for the haplotype specific tests are presented in Table 2.8. Data were simulated using 2 causal haplotypes, one with frequency 0.12 (haplotype 1) and the other with frequency 0.08 (haplotype 2). The haplotype specific score statistics test the hypothesis that one of the haplotype effects is 0 while the other effects are unconstrained. The Type I error analysis shows that the haplotype specific test is slightly conservative. The global haplotype test has higher power when there are 2 causal haplotypes with an effect size of 0.5 (0.958 for $\alpha = 0.05$) than when there is one liability SNP with an effect size of 0.5 (0.849 for $\alpha = 0.05$ in Table 2.6). The haplotype specific tests also have acceptable power, with the test for causal haplotype 1 having greater power than the test for causal haplotype 2. We expect to have greater power to detect the effect of causal haplotype 1 because it has a greater frequency. When causal haplotype 1 has an effect size of 0.7 and causal haplotype 2 has an effect size of

Table 2.10: Type I Error Rate and Power for FUSION Simulation: Global Haplotype Test ($\alpha = 0.05$)

True Effect	RD retro	FD retro	RD prosp	FD prosp
β=0.0	0.050 (*0.005*)	0.032 (*0.004*)	0.051 (*0.005*)	0.060 (*0.005*)
β=0.2	0.255 (*0.014*)	0.158 (*0.012*)	0.248 (*0.014*)	0.188 (*0.012*)
β=0.4	0.865 (*0.011*)	0.724 (*0.014*)	0.834 (*0.012*)	0.649 (*0.015*)

RD denotes reduced-dimensional (or clustering) analysis. FD denotes full-dimensional analysis. Retro denotes retrospective analysis. Prosp denotes prospective analysis. Numbers in parentheses are Monte Carlo standard deviations.

0.3, both the global haplotype test and the test for causal haplotype 1 have very high power. Causal haplotype 2 has a smaller frequency and effect size, and therefore there is lower power to detect the effect of this haplotype.

FUSION simulation

For the FUSION simulation, the average number of haplotypes estimated in each sample is 16 and the average number of haplotype clusters is 9. Therefore, the average decrease in the degrees of freedom from the full-dimensional test to the clustering test is 7. The Type I error rate and power results for simulation scenario 2 are presented in Tables 2.10 and 2.11. These results are based on 2000 replicates for Type I error rate analysis ($\beta = 0$) and 1000 replicates for power analysis. The true effect refers to the effect of the causal haplotype (01100) used to generate the data. The global test for haplotype main effects has type I error close to the nominal level for all of the analy-

Table 2.11: Type I Error Rate and Power for FUSION Simulation: Global Haplotype Test ($\alpha = 0.01$)

True Effect	RD retro	FD retro	RD prosp	FD prosp
β=0.0	0.010 (*0.002*)	0.004 (*0.001*)	0.010 (*0.002*)	0.021 (*0.003*)
β=0.2	0.116 (*0.010*)	0.048 (*0.007*)	0.110 (*0.010*)	0.066 (*0.008*)
β=0.4	0.676 (*0.015*)	0.488 (*0.016*)	0.637 (*0.015*)	0.436 (*0.016*)

RD denotes reduced-dimensional (or clustering) analysis. FD denotes full-dimensional analysis. Retro denotes retrospective analysis. Prosp denotes prospective analysis. Numbers in parentheses are Monte Carlo standard deviations.

Table 2.12: Type I Error Rate and Power for FUSION Simulation: Haplotype Specific Test

	$\alpha = 0.05$		$\alpha = 0.01$	
True Effect	Global	Specific	Global	Specific
β=0.0	0.050 (*0.007*)	0.033 (*0.006*)	0.012 (*0.003*)	0.003 (*0.002*)
β=0.5	0.976 (*0.007*)	0.982 (*0.006*)	0.926 (*0.012*)	0.912 (*0.013*)

Numbers in parentheses are Monte Carlo standard deviations.

ses. This indicates that the score statistics under H_0 have the proper χ^2 asymptotic distribution. The full-dimensional retrospective test is slightly conservative, which can be expected because of the large number of degrees of freedom. The full-dimensional prospective test is slightly anti-conservative.

The clustering retrospective analysis has the greatest power for each of the effect sizes, and has slightly greater power than the clustering prospective analysis. For both retrospective and prospective approaches, the clustering analysis has greater power than the full-dimensional analysis. This difference in power is similar for the retrospective and prospective approaches.

The haplotype specific test results are presented in Table 2.12. As with the coalescent simulation, the haplotype specific test is conservative. The global and specific tests have almost identical power, which is expected as there is only one causal haplotype. Therefore the global and specific tests are detecting the same effect.

2.4 Conclusions

We have proposed a method that addresses one of the major limitations to the usefulness of haplotype analysis for detecting genetic associations in complex diseases. Our method reduces the degree of freedom by clustering haplotypes and carrying out inference based on a core set of haplotypes. The method uses unphased genotype data and can incorporate environmental covariates, which is important when studying complex diseases. The method has greater power than the retrospective full-dimensional approach, evidence that reducing the degrees of freedom through clustering improves the performance of haplotype analysis. The greater the dimension reduction due to clustering, the larger the difference in power between the clustering and full-dimensional approaches. This can be seen from the larger power difference in our results from the FUSION simulation, where the dimension reduction is much greater than for any of the coalescent simulation scenarios.

Our simulations show little difference between the prospective and retrospective clustering approaches. This supports the finding of Satten and Epstein (2004) that retrospective and prospective likelihood analyses have similar power when assuming haplotypes have a multiplicative effect on the disease odds. Our clustering retrospective likelihood assumes a multiplicative model of disease odds because the clustering algorithm must make this assumption.

The clustering retrospective method assumes a rare disease, while the clustering algorithm is based on the common disease/common variant hypothesis. The common disease/common variant hypothesis says that common variants are responsible for common diseases. This assumption for the clustering algorithm allows us to reduce the degrees of freedom by concentrating attention on common variants that comprise the set of core haplotypes. We do not expect this contradiction to be a problem as we expect rare causal variants to be oversampled in a case-control sample. In addition, Kwee et al. (2007) found that the retrospective method assuming a multiplicative disease model is robust to the assumption of a rare disease. Case-control studies are usually used when the disease prevalence is less than 10 %, and they found the method has similar size and power for disease prevalences of 5 % and 10 %.

Satten and Epstein (2004) also show that the retrospective approach with a multiplicative model is robust to the assumption of HWE in the target population. They also propose a method to model departure from HWE due to inbreeding and population stratification by incorporating a fixation index. The fixation index is a measurement of how different the subpopulation is from a population in HWE.

The third assumption that our method makes is that haplotypes and the environmental covariate are independent in the population. Spinka et al. (2005) and Kwee et al. (2007) found that the retrospective methods are sensitive to this assumption of haplotype-environment independence. We expect this assumption to be valid in many cases, but if there is evidence that the environmental covariate is also influenced by genetic factors, other methods should be used. Spinka et al. (2005) propose a modified prospective approach that is similar to the estimating equation method of Zhao et al. (2000). This method is more robust to the haplotype-environment independence assumption. Chen et al. (2007) develop a method that allows a direct relationship between haplotypes and environmental covariates. It may be interesting to see if these methods can incorporate haplotype clustering, resulting in a method that can be used if there is a known relationship between haplotypes and covariates.

Chapter 3

Testing for Haplotype-Environment Interactions

3.1 Introduction

It is known that complex diseases are caused by both genetic and environmental factors, and researchers are becoming increasingly interested in studying the interactions between these factors. There are numerous methods available to test for interactions between single SNPs and environmental covariates, but methods for studying interactions with haplotypes are relatively new and still developing. Recent methods have focused on retrospective likelihood based approaches for studying these interactions in case-control studies. It has been shown that a prospective approach for analyzing these data is equivalent to a retrospective approach that respects the case-control sampling design (Prentice and Pyke (1979)), but only if we assume a nonparametric distribution for all of the covariates. But for haplotype-based studies, we must make some assumption about the haplotype distribution to guarantee identifiability when haplotype phase is unknown. In epidemiological studies, it is also common to assume that the genetic factors, or haplotypes, are independent of the environmental covariates. If we wish to makes these assumptions about the covariates, then a prospective approach is no longer the most efficient (Satten and Epstein (2004)). A retrospective likelihood allows us to incorporate these assumptions and improve efficiency.

Some of the retrospective methods presented in Chapter 1 only present tests to test for interaction between a specific haplotype and the covariate of interest. While there are situations where this may be an appropriate test to consider, ideally we would use the same strategy to study interaction effects that we use for studying main effects. We wish to first carry out a global test for interaction effects. Then if a global association is detected, we carry out tests for specific interaction effects. This strategy is usually not practical for methods that consider the full-dimensional haplotype space due to lack of power from large degrees of freedom. But our method uses a reduced set of haplotypes, and therefore this strategy becomes feasible.

In Chapter 2, we derived a retrospective likelihood that incorporates haplotype clustering and is based on assuming a rare disease and HWE and haplotype-environment independence in the target population. In Chapter 3 we extend the methods developed in Chapter 2 to study haplotype-environment interactions. We present score statistics to test for both global and haplotype specific interaction effects. The proposed framework for testing for haplotype specific main effects can also be applied to testing for specific interaction effects.

3.2 Methods

3.2.1 The Retrospective Likelihood for Haplotype-Environment Interaction

Here we extend the likelihood in (2.3) to allow for haplotype-environment interactions. We write the design matrix for haplotype-environment interactions as $\mathbf{X}_C \otimes \mathbf{X}_E$, where $\otimes$ denotes the Kronecker product. If we assume a binary covariate, we write the row of the design matrix corresponding to haplotype pair H as

$$X_{HE} = X_{C_H} \otimes X_E = \begin{pmatrix} X_{C_{H1}} X_E & X_{C_{H2}} X_E & \ldots & X_{C_{HL^*}} X_E. \end{pmatrix}$$

X_{HE} will have $L^*(k-1)$ columns, where k is the number of levels of the environmental covariate and (L^*+1) is the number of haplotype clusters. We write the vector of $L^*(k-1)$ interaction parameters as $\boldsymbol{\nu}$. We can modify the odds of disease as

$$\theta(H, E) = \exp(\alpha + X_C\boldsymbol{\beta} + X_E\boldsymbol{\gamma} + X_{HE}\boldsymbol{\nu})$$

and now the likelihood for testing for interactions becomes

$$L_{obs} \propto \prod_{i=1}^{n} \left[\frac{\sum_{H \in S(G_i)} \exp(\alpha^* + \mathbf{X}_{C_H}\boldsymbol{\beta} + \mathbf{X}_{E_i}\boldsymbol{\gamma} + \mathbf{X}_{HE_i}\boldsymbol{\nu}) p_h p_{h'}}{1 + \theta^*(E_i)} \right]^{d_i} \left[\frac{\sum_{H \in S(G_i)} p_h p_{h'}}{1 + \theta^*(E_i)} \right]^{1-d_i} . \tag{3.1}$$

3.2.2 Score Test for Global Interaction Effect

The null hypothesis to test for global interaction effects for a certain environmental covariate is $H_0 : \boldsymbol{\nu} = 0$. The score statistic is $S_{\boldsymbol{\nu}} = U_{\boldsymbol{\nu}}^T V_{\boldsymbol{\nu}}^{-1} U_{\boldsymbol{\nu}} \Big|_{\boldsymbol{\nu}=0, \boldsymbol{\xi}=\tilde{\boldsymbol{\xi}}}$, and has a χ^2 distribution with $L^*(k-1)$ degrees of freedom.

$$U_{\boldsymbol{\nu}} = \frac{\partial}{\partial \boldsymbol{\nu}} \log L_{obs}$$

and

$$V_{\boldsymbol{\nu}} = D_{\boldsymbol{\nu\nu}} - I_{\boldsymbol{\nu\xi}} I_{\boldsymbol{\xi\xi}}^{-1} D_{\boldsymbol{\nu\xi}}^T - D_{\boldsymbol{\nu\xi}} I_{\boldsymbol{\xi\xi}}^{-1} I_{\boldsymbol{\nu\xi}}^T + I_{\boldsymbol{\nu\xi}} I_{\boldsymbol{\xi\xi}}^{-1} D_{\boldsymbol{\xi\xi}} I_{\boldsymbol{\xi\xi}}^{-1} I_{\boldsymbol{\nu\xi}}^T,$$

where D is the variance-covariance matrix of the score function $U = (U_{\boldsymbol{\nu},\boldsymbol{\xi}})^T$ and I is the observed information matrix (Boos (1992)). The vector of nuisance parameters, $\boldsymbol{\xi}$, now consists of α^*, $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$, and $\mathbf{p}$. See Appendix B for detailed expressions for the score and variance functions.

Estimation

We use the expectation-conditional maximization (ECM) algorithm to estimate all of the nuisance parameters $\boldsymbol{\xi}$ under the null hypothesis. The estimation scheme is the same as that in section 2.2.3 for the score test for a specific haplotype main effect. The difference for the global interaction test is that we must estimate the haplotype main effect parameters $\boldsymbol{\beta}$ for all clusters. The E step estimates the number of cases and controls of each haplotype pair and covariate value. The first M step estimates maximum likelihood estimates of $\mathbf{p}$ using the updating equation (2.4). The second M step uses the optimization function `nlminb` in `R` to obtain MLEs of α^*, $\boldsymbol{\gamma}$, and $\boldsymbol{\beta}$.

3.2.3 Score Test for Specific Haplotype-Environment Interaction Effect

The test for a specific interaction effect tests the hypothesis that the specific effect is 0 while the other interaction effects are unconstrained. If we include one covariate in the model, the null hypothesis for testing for an interaction between haplotype t and the covariate is $H_0 : \nu_t = 0$. The score function is

$$U_{\nu_t} = \frac{\partial}{\partial \nu_t} \log L_{obs}$$

and the generalized variance function is

$$V_{\nu_t} = D_{\nu_t \nu_t} - I_{\nu_t \boldsymbol{\xi}} I^{-1}_{\boldsymbol{\xi}\boldsymbol{\xi}} D^T_{\nu_t \boldsymbol{\xi}} - D_{\nu_t \boldsymbol{\xi}} I^{-1}_{\boldsymbol{\xi}\boldsymbol{\xi}} I^T_{\nu_t \boldsymbol{\xi}} + I_{\nu_t \boldsymbol{\xi}} I^{-1}_{\boldsymbol{\xi}\boldsymbol{\xi}} D_{\boldsymbol{\xi}\boldsymbol{\xi}} I^{-1}_{\boldsymbol{\xi}\boldsymbol{\xi}} I^T_{\nu_t \boldsymbol{\xi}}.$$

The score statistic is $S_{\nu_t} = U^T_{\nu_t} V^{-1}_{\nu_t} U_{\nu_t} \Big|_{\nu_t = 0, \boldsymbol{\xi} = \tilde{\boldsymbol{\xi}}}$, and has a χ^2 distribution with 1 degree of freedom. The vector of nuisance parameters $\boldsymbol{\xi}$ now contains α^*, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\mathbf{p}$, and the interaction parameters $\boldsymbol{\nu}$, excluding ν_t.

Estimation

The estimation scheme for the test for a specific haplotype-environment effect is similar to that for the test of a specific main haplotype effect presented in 2.2.3. Now we must also estimate the interaction effects $\boldsymbol{\nu}$, excluding the parameter of interest ν_t.

E Step

The E step for estimating the number of controls with haplotype pair (h, h') and covariate level k is the same as that in Section 2.2.3. The E step for estimating the number of cases with haplotype pair (h, h') and covariate level k is now

$$\begin{aligned} E(d_{hh',k}) &= \sum_g d_{g,k} I(H \in S(g)) P(H|g) \\ &= \sum_g d_{g,k} I(H \in S(g)) \frac{\exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma} + X_{HE_k}\boldsymbol{\nu}) p_h p_{h'}}{\sum_{H \in S(g)} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma} + X_{HE_k}\boldsymbol{\nu}) p_h p_{h'}} \end{aligned}$$

where $d_{g,k}$ is the number of cases with genotype g and covariate level k.

M step 1: Estimating p

To estimate $\mathbf{p}$, we use a Lagrange multiplier λ and maximize the log of the full-data likelihood subject to the constraint $\sum_h p_h = 1$.

$$\begin{aligned}
L_M &= \log L_{full} + \lambda(\sum_h p_h - 1) \\
&= \sum_{k=1}^{e} \sum_{(h,h')} \Bigg[c_{hh',k} \log(p_h p_{h'}) + d_{hh',k}[(\alpha^* + X_{C_{hh'}}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma} + X_{(h,h')E_k}\boldsymbol{\nu}) + \\
&\quad \log(p_h p_{h'})] - \log(1 + \sum_H \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma} + X_{HE_k}\boldsymbol{\nu}) p_h p_{h'}) \\
&\quad (c_{hh',k} + d_{hh',k}) \Bigg] + \lambda(\sum_h p_h - 1) \\
&= \sum_{k=1}^{e} \sum_{(h,h')} \Bigg[c_{hh',k} \log(p_h p_{h'}) + d_{hh',k}[(\alpha^* + X_{C_{hh'}}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma} + X_{(h,h')E_k}\boldsymbol{\nu}) + \\
&\quad \log(p_h p_{h'})] - \log(1 + \sum_H \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma} + X_{HE_k}\boldsymbol{\nu}) p_h p_{h'}) \\
&\quad (c_{hh',k} + d_{hh',k}) \Bigg] + \lambda(\sum_h p_h - 1) \\
&= \sum_{k=1}^{e} \sum_{(h,h')} \Bigg[\log(p_h p_{h'})(c_{hh',k} + d_{hh',k}) + d_{hh',k}(\alpha^* + X_{C_{hh'}}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma} + X_{(h,h')E_k}\boldsymbol{\nu}) - \\
&\quad n_{hh',k} \log(1 + \sum_H \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma} + X_{HE_k}\boldsymbol{\nu}) p_h p_{h'}) \Bigg] + \lambda(\sum_h p_h - 1)
\end{aligned}$$

The portion of L_M that depends on $\mathbf{p}$ is

$$
\begin{aligned}
L_M \propto & \sum_{k=1}^{e} \sum_{(h,h')} \Bigg[\log(p_h p_{h'}) n_{hh',k} - \\
& n_{hh',k} \log(1 + \sum_{H} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma} + X_{HE_k}\boldsymbol{\nu}) p_h p_{h'}) \Bigg] + \lambda(\sum_{h} p_h - 1) \\
\propto & \sum_{k=1}^{e} \big(\sum_{h} m_{h,k} \log(p_h) \big) - \\
& \sum_{k=1}^{e} \sum_{(h,h')} \big(n_{hh',k} \log(1 + \sum_{H} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma} + X_{HE_k}\boldsymbol{\nu}) p_h p_{h'}) \big) \\
& + \lambda(\sum_{h} p_h - 1),
\end{aligned}
$$

where $m_{h,k}$ is the number of haplotypes of type h with covariate level k. The derivative with respect to a specific p_τ is

$$
\begin{aligned}
\frac{\partial}{\partial p_\tau} L_M \propto & \sum_{k=1}^{e} \frac{m_{\tau,k}}{p_\tau} - \\
& \sum_{k=1}^{e} \sum_{(h,h')} \left\{ n_{hh',k} \frac{\sum_H \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma} + X_{HE_k}\boldsymbol{\nu}) I(h=\tau) 2 p_{h'}}{1 + \sum_H \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma} + X_{HE_k}\boldsymbol{\nu}) p_h p_{h'}} \right\} + \lambda \\
\propto & \sum_{k=1}^{e} \frac{m_{\tau,k}}{p_\tau} - \\
& \sum_{k=1}^{e} \left\{ n_k \frac{\sum_{h'} \exp(\alpha^* + X_{C_{\tau,h'}}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma} + X_{(\tau,h')E_k}\boldsymbol{\nu}) 2 p_{h'}}{1 + \sum_H \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma} + X_{HE_k}\boldsymbol{\nu}) p_h p_{h'}} \right\} + \lambda.
\end{aligned}
$$

As in Chapter 2, we define the quantity $u(\mathbf{p})_\tau$ as

$$
u(\mathbf{p})_\tau = \sum_{k=1}^{e} n_k \frac{\sum_{h'} \exp(\alpha^* + X_{\tau,h'}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma} + X_{(\tau,h')E_k}\boldsymbol{\nu}) 2 p_{h'}}{1 + \sum_H \exp(\alpha^* + X_H\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma} + X_{HE_k}\boldsymbol{\nu}) p_{h_1} p_{h_2}}
$$

where n_k is the number of subjects with covariate level k. Then the derivative with respect to p_τ becomes

$$\frac{\partial}{\partial p_\tau} L_M \propto \sum_{k=1}^{e} \frac{m_{\tau,k}}{p_\tau} - u(\mathbf{p})_\tau + \lambda$$

and the updating equation for p_τ is

$$p_\tau = \frac{\sum_{k=1}^{e} m_{\tau,k}}{(u(\mathbf{p})_\tau - \lambda)}. \tag{3.2}$$

We carry out another iteration within the M step for estimating $\mathbf{p}$ by estimating $u(\mathbf{p})^{(s,k)}$ based on $\mathbf{p}^{(s)}$, then estimating $\mathbf{p}^{(s,k+1)}$ based on $u(\mathbf{p})^{(s,k)}$. This iteration continues until the difference between $\mathbf{p}^{(s,k)}$ and $\mathbf{p}^{(s,k+1)}$ is less than a specified limit. Then $\mathbf{p}^{(s,k+1)}$ becomes $\mathbf{p}^{(s+1)}$.

M step 2: Estimating α^*, $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$, and $\boldsymbol{\nu}$

The log of the full likelihood that depends on the regression parameters is

$$\begin{aligned}
\log L_{full} \propto & \sum_{k=1}^{e} \sum_{(h,h')} -n_{hh',k} \log(1 + \sum_{H} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma} + X_{HE_k}\boldsymbol{\nu}) p_h p_{h'}) \\
& + d_{hh',k}\big[\log(\exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma} + X_{HE_k}\boldsymbol{\nu})) + \log p_h p_{h'}\big] \\
\propto & \sum_{k=1}^{e} \sum_{(h,h')} -n_{hh',k} \log(1 + \sum_{H} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma} + X_{HE_k}\boldsymbol{\nu}) p_h p_{h'}) \\
& + d_{hh',k}(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_k}\boldsymbol{\gamma} + X_{HE_k}\boldsymbol{\nu}).
\end{aligned}$$

We obtain maximum likelihood estimates for α^*, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\nu}$ using the optimization function `nlminb` in `R`.

3.3 Simulations

3.3.1 Scenarios

To assess the size and power of the proposed score tests for interactions, we use simulation scenarios similar to those described in Chapter 2.

Coalescent Simulation

The coalescent simulation uses haplotype data generated using the coalescent model. We perform case-control sampling and use a single SNP as the disease liability locus. We use the same 6 scenarios for selecting a disease locus as in Chapter 2, which are chosen based on allele frequency (0.1 and 0.3) and haplotype diversity (high, moderate, and low) in the region. The logistic model for the liability variant allows for the effect of an interaction between the liability variant and the environmental covariate, and is defined as $\text{logit}(f_{ji}) = b_0 + b_1 j + b_2 i + b_3 i * j$ where j is the number of copies of the liability SNP and i is the level of the environmental covariate. We set $b_3 = 0.0$ to evaluate Type I error rate and $b_3 = 0.5$ and 0.7 to evaluate power. We set $b_1 = 0.5$, $b_2 = 0.3$, and vary b_0 so that the prevalence for each situation is approximately 5 %. We analyze 6-SNP haplotypes that exclude the liability locus. The haplotypes are formed from the 3 adjacent SNPs directly to the left and right of the liability locus. For each replicate we generate 500 cases and 500 controls.

To evaluate the interaction specific test, we chose the same high haplotype diversity region used to test for specific main haplotype effects in Chapter 2. We generated data assuming 2 different liability haplotypes with frequencies 0.12 and 0.08. The model at the liability haplotypes is $\text{logit}(f_{jik}) = b_0 + b_1 j + b_2 k + b_3 i + b_4 i * j + b_5 i * k$, where

f_{jik} is the probability that an individual is a case given they have j copies of liability haplotype 1, k copies of liability haplotype 2, and level i of the binary environmental covariate. We evaluated power assuming the interaction effects between both of the liability haplotypes and the covariate were the same ($b_4 = b_5 = 0.5$) and different ($b_4 = 0.7, b_5 = 0.5$). We set $b_1 = b_2 = 0.5$, $b_3 = 0.3$, and varied b_0 to achieve a 5 % disease prevalence.

FUSION Simulation

The FUSION simulation uses haplotype data based on 5 SNPs of interest on chromosome 22 in the FUSION study. We allow for interaction between the causal haplotype (01100) and the environmental covariate by modifying equation (2.5) to become

$$P(H = (h, h'), E = e|D = 1) = \frac{\exp(X_{C_{hh'}}\boldsymbol{\beta} + X_e\boldsymbol{\gamma} + X_{hh',e}\boldsymbol{\nu})p_h p_{h'} F_0(e)}{\sum_H \sum_{e^*} \exp(X_{C_{hh'}}\boldsymbol{\beta} + X_{e^*}\boldsymbol{\gamma} + X_{hh',e^*}\boldsymbol{\nu})p_h p_{h'} F_0(e^*)}. \tag{3.3}$$

We set the probability for each level of the covariate in the controls to be 0.5. This eliminates estimation problems that could result from having low counts for certain haplotype-environmental covariate combinations. We used the same sample size as in scenario 1: 500 cases and 500 controls. We set $\beta_{01100} = 0.65$, $\gamma = 1.39$, and $\nu_{01100,1} = 0.0$ to evaluate Type I error rate and $\nu_{01100,1} = 0.3$, 0.5, and 0.7 to evaluate power. We evaluated the interaction specific test using the simulated data with an effect size of 0.7.

For all simulations, we simulated 1000 datasets to assess Type I error rate and 500 datasets to assess power for $\alpha = 0.05$ and $\alpha = 0.01$. For the low diversity regions in the coalescent simulation we also conducted the analyses using the retrospective full-dimensional method of Kwee et al. (2007). For this analysis, we must estimate the full-dimensional main effect parameters. In the case of the interaction specific test, we must estimate all of the full-dimensional interaction parameters except the one that is being tested. Therefore the estimation for these parameters of main and interaction effects is not feasible for the FUSION simulation and the high and moderate haplotype diversity regions of the coalescent simulation. Furthermore, there is no available interaction test for the prospective clustering method. Consequently, in this chapter we will evaluate our proposed test under the retrospective framework, and report the results of the clustering analysis only for most scenarios.

3.3.2 Results

Coalescent Simulation

Table 3.1 presents the Type I error rate analysis for the global interaction test from the coalescent simulation. The Type I error rate is at the nominal level for all six of the liability loci for the global interaction test, indicating that the χ^2 distribution accurately estimates the asymptotic distribution of the score statistic.

Power results for an interaction effect of 0.5 and 0.7 are presented in Tables 3.2 and 3.3. The interaction effects represent the effect of the interaction between the liability locus and the environmental covariate. We set $\nu = 0.5$, meaning that the odds ratio of the covariate for individuals with the liability locus is $\exp(0.5) = 1.65$. We see an increase in power for $\nu = 0.7$ compared to $\nu = 0.5$. For the liability loci where we are able to conduct the retrospective full-dimensional analysis, we do see an increase in power for the retrospective clustering analysis as compared to the retrospective

Table 3.1: Type I Error Rate for Coalescent Simulation: Global Interaction Test

	$\alpha = 0.05$		$\alpha = 0.01$	
Hap Diversity	RD retro	FD retro	RD retro	FD retro
High:				
q=0.1	0.054 (*0.007*)	NA	0.015 (*0.004*)	NA
q=0.3	0.048 (*0.007*)	NA	0.010 (*0.003*)	NA
Moderate:				
q=0.1	0.056 (*0.007*)	NA	0.010 (*0.003*)	NA
q=0.3	0.067 (*0.008*)	NA	0.014 (*0.004*)	NA
Low:				
q=0.1	0.057 (*0.007*)	0.042 (*0.006*)	0.009 (*0.003*)	0.008 (*0.003*)
q=0.3	0.047 (*0.007*)	0.038 (*0.006*)	0.014 (*0.004*)	0.011 (*0.003*)

RD denotes reduced-dimensional (or clustering) analysis. FD denotes full-dimensional analysis. Retro denotes retrospective analysis. NA indicates that the analysis was not conducted. Numbers in parentheses are Monte Carlo standard deviations.

full-dimensional analysis. The retrospective full-dimensional analysis is only feasible in regions with low haplotype diversity. As with the test for global main haplotype effects, we do not expect a large difference in power between the clustering and full-dimensional analyses for low haplotype diversity regions since the reduction in degrees of freedom is small. We are still able to observe slight power improvement. This result suggests potential power gain that can be brought by the clustering strategy when the haplotype diversity is moderate and high.

The interaction specific test results are presented in Tables 3.4 and 3.5. Data

Table 3.2: Power for Coalescent Simulation: Global Interaction Test, $\nu = 0.5$

	$\alpha = 0.05$		$\alpha = 0.01$	
Hap diversity	RD retro	FD retro	RD retro	FD retro
High:				
q=0.1	0.422 (*0.022*)	NA	0.222 (*0.019*)	NA
q=0.3	0.188 (*0.017*)	NA	0.060 (*0.011*)	NA
Moderate:				
q=0.1	0.240 (*0.019*)	NA	0.088 (*0.013*)	NA
q=0.3	0.262 (*0.020*)	NA	0.110 (*0.014*)	NA
Low:				
q=0.1	0.496 (*0.022*)	0.420 (*0.022*)	0.244 (*0.019*)	0.208 (*0.018*)
q=0.3	0.280 (*0.020*)	0.264 (*0.020*)	0.104 (*0.014*)	0.084 (*0.012*)

RD denotes reduced-dimensional (or clustering) analysis. FD denotes full-dimensional analysis. Retro denotes retrospective analysis. NA indicates that the analysis was not conducted. Numbers in parentheses are Monte Carlo standard deviations.

were simulated using 2 causal haplotypes with frequencies 0.12 and 0.08. The Type I error rate analysis shows that the specific interaction test is conservative, similar to the specific haplotype main effect test. This is likely due to having to estimate all of the main effect haplotype parameters as well as all but one of the interaction parameters for the specific interaction test. We can compare the power of the global test in Table 3.4 to the power in Table 3.2 for the liability locus with frequency 0.1 and in a high haplotype diversity region. The power of the global test is greater with 2 causal haplotypes than with 1 causal SNP. When the interaction effect between the

Table 3.3: Power for Coalescent Simulation: Global Interaction Test, $\nu = 0.7$

	$\alpha = 0.05$		$\alpha = 0.01$	
Hap diversity	RD retro	FD retro	RD retro	FD retro
High:				
q=0.1	0.717 (*0.020*)	NA	0.469 (*0.022*)	NA
q=0.3	0.310 (*0.021*)	NA	0.112 (*0.014*)	NA
Moderate:				
q=0.1	0.452 (*0.022*)	NA	0.244 (*0.019*)	NA
q=0.3	0.436 (*0.022*)	NA	0.204 (*0.018*)	NA
Low:				
q=0.1	0.790 (*0.018*)	0.752 (*0.019*)	0.586 (*0.022*)	0.560 (*0.022*)
q=0.3	0.422 (*0.022*)	0.388 (*0.022*)	0.196 (*0.018*)	0.178 (*0.017*)

RD denotes reduced-dimensional (or clustering) analysis. FD denotes full-dimensional analysis. Retro denotes retrospective analysis. NA indicates that the analysis was not conducted. Numbers in parentheses are Monte Carlo standard deviations.

environmental covariate and each of the causal haplotypes is the same, we see the power for detecting the interaction effect for causal haplotype 1 ($q = 0.12$) is greater than that for the interaction effect for causal haplotype 2 ($q = 0.08$). When the effect of haplotype 1 increases from 0.5 to 0.7, the power of the global test and the test for haplotype 1 increases.

Table 3.4: Type I Error Rate and Power for Coalescent Simulation: Interaction Specific Test ($\alpha = 0.05$)

True Effect	Global	Hap 1	Hap 2
$\nu_1 = \nu_2$=0.0	0.051 (*0.007*)	0.022 (*0.005*)	0.027 (*0.005*)
$\nu_1 = \nu_2$=0.5	0.550 (*0.022*)	0.456 (*0.022*)	0.366 (*0.022*)
ν_1=0.7, ν_2=0.5	0.782 (*0.018*)	0.738 (*0.020*)	0.340 (*0.021*)

Numbers in parentheses are Monte Carlo standard deviations.

Table 3.5: Type I Error Rate and Power for Coalescent Simulation: Interaction Specific Test ($\alpha = 0.01$)

True Effect	Global	Hap 1	Hap 2
$\nu_1 = \nu_2$=0.0	0.007 (*0.003*)	0.004 (*0.002*)	0.004 (*0.002*)
$\nu_1 = \nu_2$=0.5	0.324 (*0.021*)	0.206 (*0.018*)	0.162 (*0.016*)
ν_1=0.7, ν_2=0.5	0.536 (*0.022*)	0.470 (*0.022*)	0.130 (*0.015*)

Numbers in parentheses are Monte Carlo standard deviations.

FUSION Simulation

The Type I error rate and power results for the global interaction test in the FUSION simulation are presented in Table 3.6. The global interaction test is anti-conservative for the FUSION simulation. We found this is caused by low haplotype frequencies which lead to near-singularities in the observed information matrix. The observed information matrix for the haplotype frequencies $\mathbf{p}$ has eigenvalues close to 0, which results in an inflated estimate for the variance of the score function.

For the FUSION simulation, we also compare the results from the global interaction

Table 3.6: Type I Error Rate and Power for FUSION Simulation: Global Interaction Test

	RD retro	
True Effect	$\alpha = 0.05$	$\alpha = 0.01$
ν=0.0	0.075 (*0.008*)	0.020 (*0.004*)
ν=0.3	0.197 (*0.018*)	0.077 (*0.007*)
ν=0.5	0.418 (*0.022*)	0.214 (*0.018*)

Numbers in parentheses are Monte Carlo standard deviations.

test to results from the test assuming only one interaction term is included in the model. For this test, we assume all of the interaction effects except the interaction with the causal haplotype are 0, and therefore the global test becomes a 1-df test. From Table 3.7, we see this test has appropriate Type I error and greater power than the full global test. This is evidence that if we have prior knowledge of a specific interaction term of interest, we will have greater power to detect a significant interaction. However, in practice we rarely have this prior knowledge; in this case it would be more practical to start with the global interaction test for investigating the interaction effects.

The interaction specific test results are presented in Table 3.8. As with the coalescent simulation, we see that the interaction specific test is conservative. The power of the global test for an interaction effect size of 0.7 can be compared to the results for effect sizes 0.3 and 0.5 in Table 3.6. For $\alpha = 0.05$, the power improves to 0.653 for an effect size of 0.7 as compared to 0.418 for an effect size 0.5. Finally, the power of the interaction specific test is slightly lower than the power for the global test. This implies

Table 3.7: Type I Error Rate and Power for FUSION Simulation: One df Global Test

	RD retro	
True Effect	$\alpha = 0.05$	$\alpha = 0.01$
ν=0.0	0.049 (*0.007*)	0.010 (*0.003*)
ν=0.3	0.415 (*0.022*)	0.197 (*0.018*)
ν=0.5	0.766 (*0.019*)	0.520 (*0.022*)

Numbers in parentheses are Monte Carlo standard deviations.

Table 3.8: Type I Error Rate and Power for FUSION Simulation: Interaction Specific Test

	$\alpha = 0.05$		$\alpha = 0.01$	
True Effect	Global	Specific	Global	Specific
ν=0.0	0.087 (*0.013*)	0.017 (*0.006*)	0.025 (*0.007*)	0.000 (<*0.001*)
ν=0.7	0.653 (*0.021*)	0.529 (*0.022*)	0.382 (*0.022*)	0.268 (*0.020*)

Numbers in parentheses are Monte Carlo standard deviations.

that a global interaction effect is easier to detect than a specific haplotype-environment effect.

3.4 Conclusions

We have extended our retrospective framework of haplotype clustering to incorporate interactions between haplotypes and environmental covariates. When testing for interaction effects, the nuisance parameter space includes the full-dimensional set of main haplotype effect parameters and the haplotype frequencies. The large dimension of the nuisance parameters can cause problems when estimating them under the null hypothesis. The estimation is especially unstable for low frequency haplotypes, and as a result, full-dimensional methods can only include specific interaction terms in the model. This makes the full-dimensional analysis less realistic, as usually we do not have prior information about what specific haplotypes may interact with the environmental covariate. The advantage of clustering is that we reduce the dimension of the parameter space and do not have to estimate main effect parameters for low frequency haplotypes, as these haplotypes will have been incorporated into clusters represented by more frequent haplotypes.

A significant contribution of our work is that by reducing the degrees of freedom through clustering, we can derive a global test for interaction that is more feasible in reality. The global test we refer to is a test involving the interaction terms between all of the haplotype clusters and a designated environmental covariate of interest. We do not refer to a test that involves the interaction terms across all haplotype clusters and all environmental covariates. We wish to use the same testing strategy we use for testing main haplotype effects to also test for interaction effects. That is, we first wish to conduct a global test to detect if there is any significant interaction between the covariate of interest and the haplotype clusters. If a global association is detected, we would then evaluate specific interaction tests.

The proposed strategy for evaluating specific haplotype effects can also be applied when evaluating specific interation effects. Our method is especially useful when we do not have prior knowledge of a haplotype or interaction term of interest. In this situation, the strategy proposed in Chapter 2 will be useful to evaluate the relationship between all of the interaction effects and group together interaction terms with a similar effect on risk of disease.

Another strategy used to evaluate interaction effects is to first test for main haplotype effects, and then test for interaction terms involving haplotypes with a significant main effect. The method proposed here can also be used for this testing strategy. We can use the haplotype specific tests from Chapter 2 to determine which haplotypes have a significant main effect and then include interaction terms for these haplotypes only.

The tests for haplotype-environment interactions presented in this chapter are subject to the same assumptions described in Chapter 2. We assume haplotype-environment independence in the target population, Hardy-Weinberg equilibrium (HWE) in the controls, and a rare disease. Satten and Epstein (2004) show that the retrospective likelihood is robust to the assumption of HWE when assuming a genetic effect that is multiplicative with respect to the odds of disease. It has also been shown by Kwee et al. (2007) that the retrospective method is robust to the rare disease assumption when the genetic effect is assumed to be multiplicative. If we expect the haplotype-environment independence assumption is violated, we suggest using another method (such as that of Spinka et al. (2005) or Chen et al. (2007)) that can incorporate a direct relationship between the covariate and haplotypes.

Chapter 4

Case-only Analysis for Testing for Haplotype-Environment Interactions

4.1 Introduction

Interactions between haplotypes and environmental covariates can also be assessed in samples of case individuals only. By assuming a saturated distribution for the environmental covariate and a multiplicative genetic effect with respect to the odds of disease, it can be shown that the retrospective likelihood factors into a piece that contains information about the haplotype and interaction parameters and a piece that does not (Kwee et al. (2007)). It can further be shown that the piece containing information about these parameters only involves case data. Thus, through a reparameterization involving the haplotype main effect parameters and haplotype frequencies, we can develop a case-only likelihood to study the effects of haplotype-environment interactions.

A method for evaluating interaction effects using only cases has several practical applications, more commonly seen in cancer therapy efficacy or drug adverse event studies. In these cases, researchers may be interested in studying how genetic factors modify the effect of treatment, but may only have data from those who responded to the treatment. For example, in pharmaceutical research one main focus in studying drug adverse events is to investigate whether the adverse reaction is different for individuals with different genetic variants (i.e. gene-drug interactions). These studies are often conducted post-marketing, and therefore we may have data for individuals who experienced an adverse reaction but not for individuals who did not experience one. In this situation, the environmental covariate is now the drug treatment and cases are defined as individuals who experience a specific adverse reaction.

To take advantage of such data, in Chapter 4 we construct a retrospective approach based on case subjects only to study haplotype-environment interaction effects. We derive score statistics to test for both global and interaction specific effects and compare the performance of these tests to those based on a sample of cases and controls presented in Chapter 3.

4.2 Methods

4.2.1 The Case-Only Retrospective Likelihood for Haplotype-Environment Interaction

We derive the case-only likelihood for testing for interaction effects by first examining which pieces of the likelihood contain information about the interaction parameters $\boldsymbol{\nu}$. Recall equation (2.1) from Chapter 2 that writes the observed likelihood of the data as

$$L_{obs} \propto \prod_{i=1}^{n} \sum_{H \in S(G_i)} P(H|E_i, D_i)P(D_i|E_i).$$

$P(D|E)$ is a function of the odds of disease given E,

$$\theta(E) = \sum_{H} \exp(\alpha + X_{C_H}\boldsymbol{\beta} + X_E\boldsymbol{\gamma} + X_{HE}\boldsymbol{\nu})p_h p_{h'}.$$

Kwee et al. (2007) show that we can rewrite $\theta(E)$ as $\exp(\alpha + X_E\boldsymbol{\gamma} + \phi_E)$ where $\phi_E = \sum_H \exp(X_{C_H}\boldsymbol{\beta} + X_{HE}\boldsymbol{\nu})p_h p_{h'}$. Because we assume a saturated distribution for E, we must estimate $k-1$ main environmental effects, where k is the number of levels of E. Therefore, $\boldsymbol{\beta}$ and $\boldsymbol{\nu}$ are incorporated into a reparameterized set of parameters, $\tilde{\boldsymbol{\gamma}}$, where $\tilde{\gamma_E} = \gamma_E + \phi_E$. These $\tilde{\boldsymbol{\gamma}}$ parameters are a function of $\boldsymbol{\beta}$ and $\boldsymbol{\nu}$, but do not provide us information about them. Thus we see that $P(D|E)$ does not contain information about $\boldsymbol{\beta}$ and $\boldsymbol{\nu}$, and we can concentrate on $P(H|E,D)$ when interested in interaction effects. Because of the assumption of haplotype-environment independence in the target population, we can further say that all of the information about $\boldsymbol{\beta}$ and $\boldsymbol{\nu}$ is contained in $P(H|E,D=1)$. Therefore we can conduct inference on the interaction parameters $\boldsymbol{\nu}$ using the cases only. We can write $P(H|E,D=1)$ as

$$\begin{aligned}\sum_{H\in S(G)} P(H|E,D=1) &= \frac{\sum_{H\in S(G)} \theta(H,E)P(H|D=0)}{\sum_{H'} \theta(H',E)P(H'|D=0)} \\ &= \frac{\sum_{H\in S(G)} \exp(X_{C_H}\boldsymbol{\beta} + X_{HE}\boldsymbol{\nu})p_h p_{h'}}{\sum_{H'} \exp(X_{C_{H'}}\boldsymbol{\beta} + X_{H'E}\boldsymbol{\nu})p_h p_{h'}}. \qquad (4.1)\end{aligned}$$

We can write $\exp(X_{C_H}\boldsymbol{\beta})$ as $\exp((B(p)[h,] + B(p)[h',])\boldsymbol{\beta})$, where $B(p)[h,]$ is the *hth* row of the clustering allocation matrix. $B(p)[h,]$ describes how haplotype h is allocated to each of the $(L^* + 1)$ core clusters. We can further expand this as

$$\exp(X_{C_H}\boldsymbol{\beta}) = \exp(\sum_{c=1}^{L^*} B(p)[h,c]\beta_c)\exp(\sum_{c=1}^{L^*} B(p)[h',c]\beta_c).$$

Incorporating this into (4.1) , we can rewrite $\sum_{H\in S(G)} P(H|E, D = 1)$ as

$$\frac{\sum_{H\in S(G)} \exp(\sum_{c=1}^{L^*} B(p)[h,c]\beta_c)\exp(\sum_{c=1}^{L^*} B(p)[h',c]\beta_c)\exp(X_{HE}\nu)p_h p_{h'}}{\sum_{H'} \exp(\sum_{c=1}^{L^*} B(p)[h,c]\beta_c)\exp(\sum_{c=1}^{L^*} B(p)[h',c]\beta_c)\exp(X_{H'E}\nu)p_h p_{h'}}. \tag{4.2}$$

Define the quantity $\tilde{p}_h$ as

$$\tilde{p}_h = \frac{\exp(\sum_{c=1}^{L^*} B(p)[h,c]\beta_c)p_h}{\sum_{h^*}\exp(\sum_{c=1}^{L^*} B(p)[h^*,c]\beta_{c^*})p_{h^*}}.$$

Now we write the case-only likelihood as

$$L_{obs} \propto \prod_{i=1}^{d} \frac{\sum_{H\in S(G_i)} \exp(X_{HE_i}\boldsymbol{\nu})\tilde{p}_h\tilde{p}_{h'}}{\sum_{H'}\exp(X_{H'E_i}\boldsymbol{\nu})\tilde{p}_h\tilde{p}_{h'}} \tag{4.3}$$

where d is the number of cases. Thus we can conduct inference on the interaction parameters $\boldsymbol{\nu}$ using (4.3) and case data only.

4.2.2 Score Test for Global Interaction Effect

The null hypothesis to test for global interaction effects is $H_0 : \boldsymbol{\nu} = 0$. The score statistic is $S_{\boldsymbol{\nu}} = U_{\boldsymbol{\nu}}^T V_{\boldsymbol{\nu}}^{-1} U_{\boldsymbol{\nu}}\Big|_{\boldsymbol{\nu}=0,\boldsymbol{\xi}=\tilde{\boldsymbol{\xi}}}$, and has a χ^2 distribution with $L^*(k-1)$ degrees of freedom. The score function is

$$U_{\boldsymbol{\nu}} = \frac{\partial}{\partial\boldsymbol{\nu}}\log L_{obs}$$

and the generalized variance function is

$$V_{\boldsymbol{\nu}} = D_{\boldsymbol{\nu\nu}} - I_{\boldsymbol{\nu\xi}} I_{\boldsymbol{\xi\xi}}^{-1} D_{\boldsymbol{\nu\xi}}^T - D_{\boldsymbol{\nu\xi}} I_{\boldsymbol{\xi\xi}}^{-1} I_{\boldsymbol{\nu\xi}}^T + I_{\boldsymbol{\nu\xi}} I_{\boldsymbol{\xi\xi}}^{-1} D_{\boldsymbol{\xi\xi}} I_{\boldsymbol{\xi\xi}}^{-1} I_{\boldsymbol{\nu\xi}}^T,$$

where D is the variance-covariance matrix of the score function $U = (U_{\boldsymbol{\nu}}, U_{\boldsymbol{\xi}})^T$ and I is the observed information matrix (Boos (1992)). For the case-only global analysis, the vector of nuisance parameters, $\boldsymbol{\xi}$, only contains the new parameters $\tilde{\mathbf{p}}$. The parameters $\tilde{\mathbf{p}}$ do not have a straightforward interpretation since they are a function of both the main effect parameters $\boldsymbol{\beta}$ and the haplotype frequencies $\mathbf{p}$. See Appendix C for detailed expressions for the score and variance functions for the case-only analysis.

Estimation

We use the EM algorithm to maximize the case-only likelihood with respect to $\tilde{\mathbf{p}}$, while using a Lagrange multiplier to constrain $\sum_h \tilde{p}_h = 1$. We first assume that haplotype phase is known and construct the full data likelihood:

$$L_{full} = \prod_{k=1}^{e} \prod_{(h,h')} \left[\frac{\exp(X_{HE_k}\boldsymbol{\nu})\tilde{p}_h\tilde{p}_{h'}}{\sum_{H'} \exp(X_{H'E_k}\boldsymbol{\nu})\tilde{p}_h\tilde{p}_{h'}} \right]^{d_{hh',k}},$$

where $d_{hh',k}$ is the number of cases with haplotype pair (h, h') and covariate level k. Under the null hypothesis that $\boldsymbol{\nu} = 0$, the full data likelihood becomes

$$L_{full} = \prod_{k=1}^{e} \prod_{(h,h')} \left[\frac{\tilde{p}_h\tilde{p}_{h'}}{\sum_{H'} \tilde{p}_h\tilde{p}_{h'}} \right]^{d_{hh',k}}.$$

Taking the log of the full likelihood and incorporating the Lagrange multiplier, we maximize the new likelihood, L_M:

$$\begin{aligned}
L_M &= \log L_{full} + \lambda(\sum_h \tilde{p}_h - 1)\\
&= \sum_{k=1}^{e} \sum_{(h,h')} \left[d_{hh',k} \log(\tilde{p}_h \tilde{p}_{h'}) - d_{hh',k} \log(\sum_{H'} \tilde{p}_h \tilde{p}_{h'}) \right] + \lambda(\sum_h \tilde{p}_h - 1)\\
&= \sum_{k=1}^{e} \sum_{(h,h')} d_{hh',k} \log(\tilde{p}_h \tilde{p}_{h'}) + \lambda(\sum_h \tilde{p}_h - 1)\\
&= \sum_{k=1}^{e} (\sum_h d_{h,k} \log(\tilde{p}_h)) + \lambda(\sum_h \tilde{p}_h - 1),
\end{aligned}$$

where $d_{h,k}$ is the number of cases with haplotype h and covariate level k. We estimate $d_{hh',k}$ in the E step:

$$\begin{aligned}
E(d_{hh',k}) &= \sum_g d_{g,k} I(H \in S(g)) P(H|g)\\
&= \sum_g d_{g,k} I(H \in S(g)) \frac{\tilde{p}_h \tilde{p}_{h'}}{\sum_{H \in S(g)} \tilde{p}_h \tilde{p}_{h'}}
\end{aligned}$$

where $d_{g,k}$ is the number of cases with genotype g and covariate level k. The derivative of L_M with respect to a specific $\tilde{p}_\tau$ is

$$\frac{\partial}{\partial \tilde{p}_\tau} L_M \propto \sum_{k=1}^{e} \frac{d_{\tau,k}}{\tilde{p}_\tau} + \lambda.$$

By setting the above equal to 0 and solving for $\tilde{p}_\tau$, we obtain the updating equation for $\tilde{p}_\tau$

$$\tilde{p}_\tau = \frac{\sum_{k=1}^{e} d_{\tau,k}}{\lambda}.$$

4.2.3 Score Test for Specific Haplotype-Environment Interaction Effect

The null hypothesis for testing for an interaction between haplotype t and the environmental covariate is $H_0 : \nu_t = 0$. The score function U_{ν_t} is

$$U_{\nu_t} = \frac{\partial}{\partial \nu_t} \log L_{obs}$$

and the generalized variance function is

$$V_{\nu_t} = D_{\nu_t \nu_t} - I_{\nu_t \boldsymbol{\xi}} I^{-1}_{\boldsymbol{\xi}\boldsymbol{\xi}} D^T_{\nu_t \boldsymbol{\xi}} - D_{\nu_t \boldsymbol{\xi}} I^{-1}_{\boldsymbol{\xi}\boldsymbol{\xi}} I^T_{\nu_t \boldsymbol{\xi}} + I_{\nu_t \boldsymbol{\xi}} I^{-1}_{\boldsymbol{\xi}\boldsymbol{\xi}} D_{\boldsymbol{\xi}\boldsymbol{\xi}} I^{-1}_{\boldsymbol{\xi}\boldsymbol{\xi}} I^T_{\nu_t \boldsymbol{\xi}}.$$

The score statistic is $S_{\nu_t} = U^T_{\nu_t} V^{-1}_{\nu_t} U_{\nu_t} \Big|_{\nu_t = 0, \boldsymbol{\xi} = \tilde{\boldsymbol{\xi}}}$, and has a χ^2 distribution with 1 degree of freedom. The vector of nuisance parameters $\boldsymbol{\xi}$ now contains $\tilde{\mathbf{p}}$ and the vector of interaction parameters $\boldsymbol{\nu}$, excluding the parameter of interest ν_t.

Estimation

We use the ECM algorithm to estimate the nuisance parameters $\boldsymbol{\xi}$ under the null hypothesis. Assuming that haplotype phase is known, the full data likelihood is

$$L_{full} = \prod_{k=1}^{e} \prod_{(h,h')} \left[\frac{\exp(X_{HE_k} \boldsymbol{\nu}) \tilde{p}_h \tilde{p}_{h'}}{\sum_{H'} \exp(X_{H'E_k} \boldsymbol{\nu}) \tilde{p}_h \tilde{p}_{h'}} \right]^{d_{hh',k}}.$$

The steps for estimating the nuisance parameters are

1. Obtain initial estimates of $\tilde{\mathbf{p}}$ and $\boldsymbol{\nu}$.

2. For step s, use E Step to estimate $d^{(s)}_{hh',k}$.

3. Use M step 1 to find $\tilde{\mathbf{p}}^{(s+1)}$ that maximizes L_{full}, using $\boldsymbol{\nu}^{(s)}$ and $d^{(s)}_{hh',k}$.

4. Use M step 2 to find $\boldsymbol{\nu}^{(s)}$ that maximizes L_{full}, using $\tilde{\mathbf{p}}^{(s+1)}$ and $d^{(s)}_{hh',k}$.

5. Check differences between parameters at steps $(s+1)$ and s.

6. If difference for at least one of the parameters is greater than a specified limit, start over with step 2 to estimate $d^{(s+1)}_{hh',k}$ using $\tilde{\mathbf{p}}^{(s+1)}$ and $\boldsymbol{\nu}^{(s+1)}$.

E step

The E step estimates the number of cases with haplotype pair (h, h') and covariate level k as

$$
\begin{aligned}
E(d_{hh',k}) &= \sum_g d_{g,k} I(H \in S(g)) P(H|g) \\
&= \sum_g d_{g,k} I(H \in S(g)) \frac{\exp(X_{HE_k}\boldsymbol{\nu})\tilde{p}_h\tilde{p}_{h'}}{\sum_{H \in S(g)} \exp(X_{HE_k}\boldsymbol{\nu})\tilde{p}_h\tilde{p}_{h'}}
\end{aligned}
$$

where $d_{g,k}$ is the number of cases with genotype g and covariate level k.

M step 1

Similar to the estimation of $\tilde{\mathbf{p}}$ for the global interaction test, we maximize the full data likelihood subject to the constraint that $\sum_h \tilde{p}_h = 1$. Incorporating the Lagrange multiplier, the new likelihood L_M is

$$\begin{aligned}L_M &= \log L_{full} + \lambda(\sum_h \tilde{p}_h - 1)\\ &= \sum_{k=1}^{e}\sum_{(h,h')}\Bigg[d_{hh',k}[(X_{HE_k}\boldsymbol{\nu}) + \log(\tilde{p}_h\tilde{p}_{h'})] -\\ &\quad d_{hh',k}\log(\sum_H \exp(X_{HE_k}\boldsymbol{\nu})\tilde{p}_h\tilde{p}_{h'})\Bigg] + \lambda(\sum_h \tilde{p}_h - 1)\end{aligned}$$

The portion of L_M that depends on $\tilde{\mathbf{p}}$ is

$$\begin{aligned}L_M &\propto \sum_{k=1}^{e}\sum_{(h,h')}\Bigg[d_{hh',k}\log(\tilde{p}_h\tilde{p}_{h'}) -\\ &\quad d_{hh',k}\log(\sum_H \exp(X_{HE_k}\boldsymbol{\nu})\tilde{p}_h\tilde{p}_{h'})\Bigg] + \lambda(\sum_h \tilde{p}_h - 1)\\ &\propto \sum_{k=1}^{e}(\sum_h d_{h,k}\log(\tilde{p})) + \sum_{k=1}^{e}\sum_{(h,h')} d_{hh',k}\log(\sum_H \exp(X_{HE_k}\boldsymbol{\nu})\tilde{p}_h\tilde{p}_{h'}) + \lambda(\sum_h \tilde{p}_h - 1)\end{aligned}$$

The derivative of this expression with respect to a specific p_τ is

$$\begin{aligned}\frac{\partial}{\partial p_\tau}L_M &\propto \sum_{k=1}^{e}\frac{d_{\tau,k}}{\tilde{p}_\tau} - \sum_{k=1}^{e}\sum_{(h,h')} d_{hh',k}\frac{\sum_H \exp(X_{HE_k}\boldsymbol{\nu})I(h=\tau)2\tilde{p}_{h'}}{\sum_H \exp(X_{HE_k}\boldsymbol{\nu})\tilde{p}_h\tilde{p}_{h'}} + \lambda\\ &\propto \sum_{k=1}^{e}\frac{d_{\tau,k}}{\tilde{p}_\tau} - \sum_{k=1}^{e} d_k\frac{\sum_{h'} \exp(X_{HE_k}\boldsymbol{\nu})2\tilde{p}_{h'}}{\sum_H \exp(X_{HE_k}\boldsymbol{\nu})\tilde{p}_h\tilde{p}_{h'}} + \lambda\end{aligned}$$

As in Chapters 2 and 3, we define a quantity $u(\tilde{\mathbf{p}})_\tau$ as

$$u(\tilde{\mathbf{p}})_\tau = \sum_{k=1}^{e} d_k\frac{\sum_{h'} \exp(X_{HE_k}\boldsymbol{\nu})2\tilde{p}_{h'}}{\sum_H \exp(X_{HE_k}\boldsymbol{\nu})\tilde{p}_h\tilde{p}_{h'}}.$$

Then we write the derivative with respect to p_τ as

$$\frac{\partial}{\partial p_\tau} L_M \propto \sum_{k=1}^{e} \frac{d_{\tau,k}}{\tilde{p}_\tau} - u(\tilde{\mathbf{p}})_\tau + \lambda,$$

and by setting the above equal to 0 and solving for p_τ we obtain the updating equation for p_τ:

$$p_\tau = \frac{\sum_{k=1}^{e} d_{\tau,k}}{u(\tilde{\mathbf{p}})_\tau - \lambda}.$$

We carry out another iteration within the M step for estimating $\tilde{\mathbf{p}}$ by estimating $u(\tilde{\mathbf{p}})^{(s,k)}$ based on $\tilde{\mathbf{p}}^{(s)}$, then estimating $\tilde{\mathbf{p}}^{(s,k+1)}$ based on $u(\tilde{\mathbf{p}})^{(s,k)}$. This iteration continues until the difference between $\tilde{\mathbf{p}}^{(s,k)}$ and $\tilde{\mathbf{p}}^{(s,k+1)}$ is less than a specified limit. Then $\tilde{\mathbf{p}}^{(s,k+1)}$ becomes $\tilde{\mathbf{p}}^{(s+1)}$.

M step 2

The log of the full data likelihood that depends on the interaction parameters $\boldsymbol{\nu}$ is

$$\log(L_{full}) \propto \sum_{k=1}^{e} \sum_{(h,h')} \Big[d_{hh',k} \log(\exp(X_{HE_k}\boldsymbol{\nu})\tilde{p}_h\tilde{p}_{h'}) - d_{hh',k} \log(\sum_{H} \exp(X_{HE_k}\boldsymbol{\nu})\tilde{p}_h\tilde{p}_{h'}).$$

We use the optimization function `nlminb` in `R` to obtain maximum likelihood estimates of $\boldsymbol{\nu}$.

4.3 Simulations

4.3.1 Scenarios

The simulations for assessing the size and power of the proposed score tests based on a case-only likelihood use the same simulation scenarios described in Chapter 3.

Coalescent Simulation

The coalescent simulation simulates case-control data using 6 single-SNP liability loci chosen based on allele frequency and haplotype diversity in the surrounding region. The logistic model at the liability locus is defined as $\text{logit}(f_{ji}) = b_0 + b_1 j + b_2 i + b_3 i * j$ where j is the number of copies of the liability allele and i is the level of the environmental covariate. To evaluate the Type I error rate, we set $b_3 = 0$. To evaluate power, we run the case-only analysis on the same data used for evaluating the case-control method using an effect size of $b_3 = 0.7$. We set $b_1 = 0.5$, $b_2 = 0.3$, and vary b_0 so that the prevalance for each situation is approximately 5 %. As in Chapter 3, we analyze 6-SNP haplotypes that exclude the liability locus. For each replicate, we generate 500 cases and 500 controls.

We evaluated the case-only haplotype specific test using the same data used to evaluate the case-control haplotype specific test in Chapter 3. These data were simulated using two causal haplotypes with frequencies 0.12 and 0.08. The model at the liability haplotypes is $\text{logit}(f_{jik}) = b_0 + b_1 j + b_2 k + b_3 i + b_4 i * j + b_5 i * k$, where f_{jik} is the probability that an individual is a case given they have j copies of liability haplotype 1, k copies of liability haplotype 2, and level i of the binary environmental covariate. We evaluated power assuming the interaction effects between both of the liability haplotypes and the covariate were the same ($b_4 = b_5 = 0.5$) and different ($b_4 = 0.7, b_5 = 0.5$). We set $b_1 = b_2 = 0.5$, $b_3 = 0.3$, and varied b_0 to achieve a 5 % disease prevalence.

FUSION Simulation

The FUSION simulation uses haplotype data based on 5 SNPs of interest on chromosome 22 in the FUSION study. We set the interaction effect between the causal haplotype (01100) and the environmental covariate equal to 0 to evaluate the Type I error rate. We evaluated the power of the global and specific interaction tests using the same data used to evaluate the case-control tests using an effect size of 0.7.

We simulated 1000 datasets to assess Type I error rate and 500 datasets to assess power for $\alpha = 0.05$ and $\alpha = 0.01$. For the case-only analysis of the global interaction test, we also conducted the analyses using the retrospective full-dimensional method of Kwee et al. (2007). This full-dimensional analysis is implementable for the case-only analysis because the estimation is the same for both the clustering and full-dimensional analyses. For the global interaction test, the nuisance parameters only consist of $\tilde{\mathbf{p}}$.

4.3.2 Results

Coalescent Simulation

Table 4.1 contains the Type I error rate analysis results for the case-only global test of interaction from the coalescent simulation. The Type I error rate is at the nominal level for the retrospective clustering analysis, which is evidence that the χ^2 distribution accurately estimates the asymptotic distribution of the case-only score statistic. The Type I error rate is conservative for the retrospective full-dimensional analysis.

Table 4.1: Type I Error Rate for Coalescent Simulation: Case-only Global Test of Interaction

	$\alpha = 0.05$		$\alpha = 0.01$	
Hap Diversity	RD retro	FD retro	RD retro	FD retro
High:				
q=0.1	0.054 (*0.007*)	0.043 (*0.006*)	0.009 (*0.003*)	0.006 (*0.002*)
q=0.3	0.047 (*0.007*)	0.044 (*0.006*)	0.010 (*0.003*)	0.005 (*0.002*)
Moderate:				
q=0.1	0.054 (*0.007*)	0.039 (*0.006*)	0.019 (*0.004*)	0.008 (*0.003*)
q=0.3	0.046 (*0.007*)	0.014 (*0.004*)	0.010 (*0.003*)	0.002 (*0.001*)
Low:				
q=0.1	0.053 (*0.007*)	0.045 (*0.007*)	0.013 (*0.004*)	0.009 (*0.003*)
q=0.3	0.050 (*0.007*)	0.040 (*0.006*)	0.010 (*0.003*)	0.008 (*0.003*)

RD denotes reduced-dimensional (or clustering) analysis. FD denotes full-dimensional analysis. Retro denotes retrospective analysis. Numbers in parentheses are Monte Carlo standard deviations.

The power results for the case-only global test of interaction are presented in Table 4.2. These results are based on the same simulated data used to evaluate the case-control global test of interaction for an effect size $\nu = 0.7$. The power results for the case-control global test of interaction are in Table 3.3. The case-only test is as powerful as the case-control test for all of the liability loci, and in some cases the case-only test has slightly greater power. We also see an improvement in power for the case-only retrospective clustering method as compared to the case-only retrospective full-dimensional method. We see the largest reduction in degrees of freedom as a result of clustering for the locus with frequency 0.3 in a region of moderate haplotype diversity.

Table 4.2: Power for Coalescent Simulation: Case-only Global Test of Interaction

	$\alpha = 0.05$		$\alpha = 0.01$	
Hap Diversity	RD retro	FD retro	RD retro	FD retro
High:				
q=0.1	0.721 (*0.020*)	0.638 (*0.021*)	0.481 (*0.022*)	0.396 (*0.022*)
q=0.3	0.328 (*0.021*)	0.278 (*0.020*)	0.116 (*0.014*)	0.088 (*0.013*)
Moderate:				
q=0.1	0.452 (*0.022*)	0.366 (*0.022*)	0.244 (*0.019*)	0.170 (*0.017*)
q=0.3	0.494 (*0.022*)	0.270 (*0.020*)	0.266 (*0.020*)	0.088 (*0.013*)
Low:				
q=0.1	0.802 (*0.018*)	0.768 (*0.019*)	0.574 (*0.022*)	0.556 (*0.022*)
q=0.3	0.430 (*0.022*)	0.380 (*0.022*)	0.204 (*0.018*)	0.178 (*0.017*)

RD denotes reduced-dimensional (or clustering) analysis. FD denotes full-dimensional analysis. Retro denotes retrospective analysis. Numbers in parentheses are Monte Carlo standard deviations.

For this locus, there was an average of 14 observed haplotypes and an average of 10 haplotype clusters. As a result, we see the largest power improvement for this locus. For $\alpha = 0.05$, the power for the full-dimensional analysis is 0.27 and for the clustering analysis is 0.49.

Results for the case-only interaction specific test are presented in Tables 4.3 and 4.4. As with the case-control interaction specific test, we see that the case-only interaction specific test is conservative. When the effect sizes of the 2 causal haplotypes are the same, the power of the global test is greater than for each of the interaction specific tests. The test for causal haplotype 1 is more powerful than the test for causal

Table 4.3: Type I Error Rate and Power for Coalescent Simulation: Case-only Haplotype Specific Test ($\alpha = 0.05$)

True Effect	Global	Hap 1	Hap 2
$\nu_1 = \nu_2$=0.0	0.054 (*0.007*)	0.024 (*0.005*)	0.024 (*0.005*)
$\nu_1 = \nu_2$=0.5	0.566 (*0.022*)	0.460 (*0.022*)	0.370 (*0.022*)
ν_1=0.7, ν_2=0.5	0.786 (*0.018*)	0.746 (*0.019*)	0.358 (*0.021*)

Numbers in parentheses are Monte Carlo standard deviations.

Table 4.4: Type I Error Rate and Power for Coalescent Simulation: Case-only Haplotype Specific Test ($\alpha = 0.01$)

True Effect	Global	Hap 1	Hap 2
$\nu_1 = \nu_2$=0.0	0.008 (*0.003*)	0.004 (*0.002*)	0.004 (*0.002*)
$\nu_1 = \nu_2$=0.5	0.328 (*0.021*)	0.210 (*0.018*)	0.162 (*0.016*)
ν_1=0.7, ν_2=0.5	0.544 (*0.022*)	0.482 (*0.022*)	0.144 (*0.016*)

Numbers in parentheses are Monte Carlo standard deviations.

haplotype 2, which is expected as causal haplotype 1 is more frequent. When the effect size of haplotype 1 increases from 0.5 to 0.7, the power of the global test and the test for haplotype 1 both increase.

FUSION Simulation

The Type I error rate and power results for the clustering and full-dimensional case-only global interaction tests from the FUSION simulation are presented in Table 4.5. As found for the case-control global interaction test, the clustering case-only interaction

Table 4.5: Type I Error Rate and Power for FUSION Simulation: Case-only Global Interaction Test

	$\alpha = 0.05$		$\alpha = 0.01$	
True Effect	RD retro	FD retro	RD retro	FD retro
ν=0.0	0.066 (*0.008*)	0.088 (*0.009*)	0.017 (*0.004*)	0.024 (*0.005*)
ν=0.7	0.644 (*0.021*)	0.540 (*0.022*)	0.376 (*0.022*)	0.212 (*0.018*)

Numbers in parentheses are Monte Carlo standard deviations.

Table 4.6: Type I Error Rate and Power for FUSION Simulation: Case-only Interaction Specific Test

	RD retro	
True Effect	$\alpha = 0.05$	$\alpha = 0.01$
ν=0.0	0.019 (*0.004*)	0.000 (<*0.001*)
ν=0.7	0.505 (*0.022*)	0.227 (*0.019*)

Numbers in parentheses are Monte Carlo standard deviations.

test is anti-conservative for the FUSION simulation for both the clustering and full-dimensional methods. For the FUSION simulation, clustering reduces the degrees of freedom by 7 on average, and we see a corresponding increase in power compared to the full-dimensional analysis. In Table 4.6 we see that the interaction specific test is conservative. The power of the global and interaction specific tests is similar to the power for the case-control tests evaluated on the same data (Table 3.8).

4.4 Conclusions

We have derived tests for haplotype-environment interactions at the global and specific levels that can be applied to a sample of case individuals only. The tests based on the case-only likelihood have similar power to tests based on a sample of cases and controls. These results are consistent with those of Kwee et al. (2007), who found the case-only analysis to have similar power to the case-control analysis for the retrospective full-dimensional analysis. A test for haplotype-environment interactions in a case-only sample will be relevant in situations where it is not practical or possible to collect a control sample. The test may be especially useful in cancer research where collecting controls is difficult and interactions with environmental covariates are of particular interest.

We also conclude that the case-only clustering analysis is more powerful than the case-only full-dimensional analysis. This is an important and encouraging result, as we were unable to directly verify the power improvement from clustering for the interaction tests in Chapter 3. In most situations, it is not feasible to carry out the full dimensional interaction analysis for a case-control sample. For the case-only global interaction test, we only need to estimate the new parameters $\tilde{\mathbf{p}}$ for both the clustering and full-dimensional analyses. Therefore we are able to compare the power of the two methods. The power difference we see for the case-only analysis is further evidence of the advantage of reducing degrees of freedom through clustering.

The case-only analysis method is limited by the assumptions that haplotypes have a multiplicative effect on the odds of disease and that the distribution for covariate E is saturated. Assuming that the model for E is saturated ensures that $P(E|D)$

will not contain any information about the haplotype and interaction parameters, and allows us to conduct inferences on interaction parameters using cases only through $P(G|E, D = 1)$. If a dominant or recessive model is required, then we are not able to reparameterize our model using the parameters $\tilde{\mathbf{p}}$, and we must make inferences based on the full case-control likelihood. Because the clustering algorithm also assumes a multiplicative effect on the odds of disease, our method will not be further limited by this assumption for the case-only analysis.

Chapter 5

Application to Hypertriglyceridemia Study

5.1 Background

We apply the proposed score tests for testing for global and specific haplotype and interaction effects to data from a study of hypertriglyceridemia conducted by the National Taiwan University Hospital. This dataset was also analyzed by Tzeng et al. (2006) and Chen and Kao (2006). Our goal in analyzing this dataset is to compare our results with previous findings.

Hypertriglyceridemia is a metabolic disorder characterized by high levels of triglycerides in the blood, and is known to be a risk factor for coronary heart disease. The study consisted of 290 cases, defined as individuals having serum triglycerides >400mg/dl, and 303 controls who were recruited through health examinations at National Taiwan University Hospital. Controls were excluded if they had secondary hyperlipoproteinemia, hypertension, diabetes mellitus, lipid-lowering medications, and endocrine and metabolic disorders. All participants provided consent for DNA samples.

From this study, Kao et al. (2003) identified a novel variant in the coding region of the APOA5 gene on chromosome 11 that increases the risk of developing hypertriglyceridemia. We will analyze haplotypes comprised of 5 SNPs in this region of interest, including the SNP identified by Kao et al. (2003).

5.2 Analysis

Our analysis was motivated by the results of Chen and Kao (2006), who found a significant interaction between haplotype and age. We dichotomized age and included it as an environmental covariate in our model. We used the mean age in controls of 49 as the cutoff for the dichotomization. Therefore we set X_{age} for individual i as $I(Age_i > 49)$. We excluded participants with missing covariate information or missing genetic information at all 5 SNPs analyzed. For our analysis, there were 210 cases and 287 controls. We first analyzed global and specific haplotype main effects. We then included the interactions between all of the haplotype clusters and age in the model and tested for interaction effects.

5.3 Results

The region of interest contains 12 estimated haplotypes with frequency $>1\times10^{-5}$. The clustering algorithm creates 4 haplotype clusters represented by the core haplotypes GGGCT, GGTCT, AGGCC, and GAGTT.

The global test for haplotype main effects is highly significant. The global score statistic is 133.71 and the p-value is $<1 \times 10^{-6}$. Table 5.1 contains the haplotype specific results from all pairwise comparisons.

All but one of these comparisons is highly significant using the Bonferroni corrected α level of $0.05/6 = 8.3\times10^{-3}$. Therefore we group the 4 haplotype clusters into 3 groups

Table 5.1: Haplotype Specific Results for Age as Covariate

Reference	Haplotype	Score Statistic	P-value
GGTCT	GAGTT	20.50	5.97×10^{-6}
	GGGCT	59.79	$< 1 \times 10^{-6}$
	AGGCC	8.94	2.79×10^{-3}
GGGCT	GAGTT	12.65	3.75×10^{-4}
	AGGCC	58.88	$< 1 \times 10^{-6}$
AGGCC	GAGTT	5.92	1.50×10^{-2}

based on similar effects on disease risk. Group 1 contains haplotype GGTCT, group 2 contains haplotype GGGCT, and group 3 contains haplotypes AGGCC and GAGTT.

The global test for interactions is close to significant at the nominal level. The score statistic is 6.62 and the p-value is 0.08. The interaction specific results are presented in Table 5.2. Before correcting for multiple testing, we see some evidence that the interaction effect between age and the haplotype GGTCT is different from the interaction effects between age and haplotypes GGGCT and AGGCC. The interaction specific results are inconclusive about the interaction effect with haplotype GAGTT. Our results show that the interaction between age and GAGTT is not different from any of the other interaction effects. Based on the uncorrected results, we can group the interaction term involving age and GGTCT into one group and the interactions with GGGCT and AGGCC into another. None of the interaction specific results would be significant after using the Bonferroni correction for multiple testing.

Table 5.2: Interaction Specific Results for Age as Covariate

Reference	Haplotype	Score Statistic	P-value
GGTCT	GAGTT	1.59	2.08×10^{-1}
	GGGCT	3.42	6.46×10^{-2}
	AGGCC	3.55	5.97×10^{-2}
GGGCT	GAGTT	0.21	6.43×10^{-1}
	AGGCC	0.03	8.67×10^{-1}
AGGCC	GAGTT	0.11	7.39×10^{-1}

5.4 Conclusions

Our analysis of the hypertriglyceridemia study confirms the findings of Tzeng et al. (2006) that there is a highly significant global haplotype main effect in the region of interest on chromosome 11 in the APOA5 gene. Our method is able to further group the 4 haplotype clusters into 3 groups based on similar effects on risk of hypertriglyceridemia.

Our method is able to confirm the finding of Chen and Kao (2006) of a significant interaction between age and haplotype GGTCT. But whereas Chen and Kao (2006) only include the interaction between age and this specific haplotype in their model, we include all of the haplotype clusters and first test for a global effect. We then evaluate the interaction specific tests, and are also able to conclude that the effect of the interaction between age and GGTCT is different from the effects of interactions involving haplotypes GGGCT and GAGTT.

Chapter 6

Summary and Future Work

6.1 Contributions

Haplotypes have the potential to play an important role in the search for genetic factors that cause complex human diseases. However, due to power limitations from the large number of degrees of freedom required to test for haplotype effects, haplotype analyses may not be performed in practice. Also, because of restrictions placed on the haplotype distribution to ensure identifiability when haplotype phase is unknown, the traditional method of using a prospective likelihood to analyze haplotype data from a case-control sample may not be the most efficient. This work increases the potential for haplotype analysis to find regions of disease susceptibility genes by reducing the degrees of freedom through haplotype clustering and improving efficiency through use of a retrospective likelihood.

We have derived a retrospective likelihood that incorporates the haplotype clustering algorithm of Tzeng et al. (2006) and can accomodate environmental covariates and their interactions with haplotypes. Through clustering, we are able to retain information from all observed haplotypes, without using degrees of freedom for rare haplotypes we will have little power to detect. We assume that haplotype and the environmental covariate are independent in the population, the target population is in Hardy-Weinberg Equilibrium, and that the disease of interest is rare. These assumptions lead to a method that is practical to implement and appropriate for many interesting situations.

We have presented score statistics to test for haplotype main effects at both the global and specific levels. The global test for main effects based on our retrospective clustering method has greater power than the test based on the retrospective full-dimensional method and is at least as powerful as the prospective clustering method. We have also suggested a new strategy for testing for haplotype specific effects. Commonly used strategies may not give us a complete picture of which haplotype clusters have similar effects on the risk of disease, and the conclusions may be dependent on which haplotype is chosen as the reference. The strategy we propose looks at all pairwise comparisons by assigning all haplotypes in turn to be the reference haplotype and allows us to group together haplotypes that have similar effects on disease. This strategy can also be applied to tests for specific haplotype-environment interaction terms.

One of the most significant contributions of this work is the ability to test for global interaction effects. As the study of complex diseases becomes an important focus of scientific research, it is critical that methods for genetic analysis be able to incorporate environmental covariates and interactions between genetic factors and these covariates. Including interaction terms involving the full-dimensional space of haplotypes limits power and causes estimation difficulties for low frequency haplotypes. As a result, many full-dimensional methods are restricted to only including interaction terms involving certain pre-determined haplotypes of interest. Our method works with a reduced parameter space, thereby making a global test more practical. Thus we can employ the testing strategy of first testing for a global interaction effect and then testing for specific haplotype-environment effects. We have presented score statistics to test for both global and specific haplotype-environment interaction effects.

We have also derived haplotype-environment interaction tests that can be applied to case-only samples. We have shown that the case-only interaction test is as powerful as the case-control test. This is an important result in the context of cancer drug efficacy and adverse event studies, where the drug treatment is the environmental covariate. In these studies, the main focus is often on gene-drug interactions, and data may only be available from individuals who respond to a treatment or experience an adverse event. Our results also show that the retrospective case-only test using clustered haplotypes is more powerful than the case-only full-dimensional test.

We also plan to make our `R` code available to users interested in implementing our methods. We will provide code for testing for haplotype main effects and interactions, as well as code for carrying out our strategy for haplotype specific tests. Our code for haplotype specific tests can also be used to obtain estimates of haplotype or interaction specific effects. Instead of constraining the haplotype of interest to have no effect on disease and calculating the corresponding score statistic, we can leave all of the effects unconstrained and estimate effects relative to the reference haplotype.

6.2 Future Work

We plan for future work to improve the clustering algorithm, specifically in the context of testing for interactions. When assessing interaction effects, we can further reduce the degrees of freedom by limiting the number of core haplotypes. The clustering algorithm

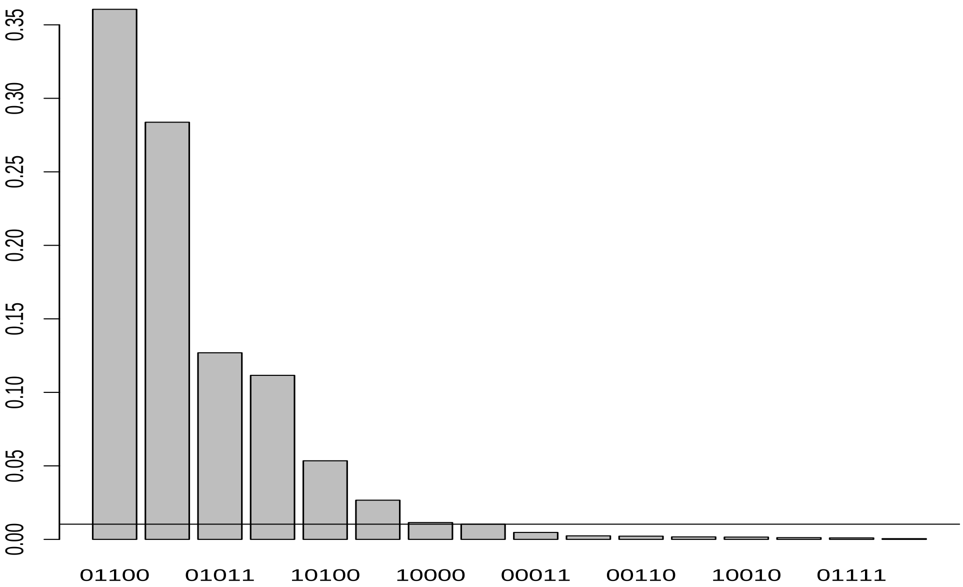

Figure 6.1: Haplotype frequency distribution for FUSION simulation. Haplotypes with frequencies greater than the horizontal cut-off line will be designated as core haplotypes. The clustering algorithm uses the original penalty function to determine the core haplotypes.

was designed for assessing main haplotype effects, where we are interested in including all of the core haplotypes in our model. One factor in determining the number of core clusters is the penalty function used to determine the information criterion. The penalty function is a function of the sample size. The rationale is that the larger the sample size, the more information we have about the haplotypes and the more likely we are to allow for lower frequency core haplotypes. But for interaction tests, we are willing to sacrifice core haplotypes in return for greater power to detect interactions between the most common haplotypes and the covariate. One ad-hoc way to reduce the number of core haplotypes is to use a larger penalty term in the clustering algorithm. For example, Figure 6.1 presents the haplotype distribution from a sample dataset from the FUSION simulation. The horizontal line represents the cut-point for determining

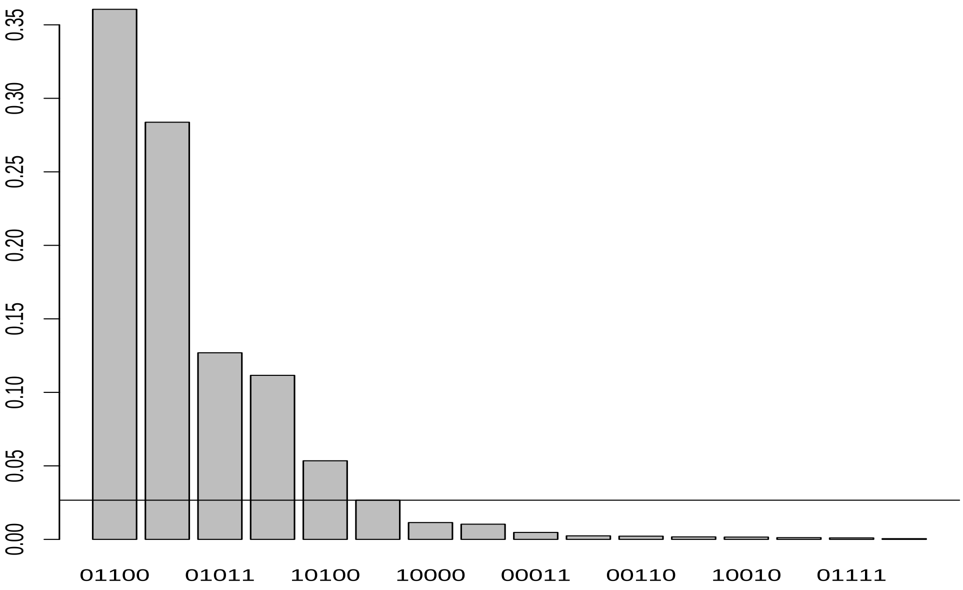

Figure 6.2: Haplotype frequency distribution for FUSION simulation. Haplotypes with frequencies greater than the horizontal cut-off line will be designated as core haplotypes. The clustering algorithm uses a doubled penalty term to determine the core haplotypes.

the core haplotypes, which is based on the original penalty function. The 8 haplotypes with frequency above this cut-point will be included as the core haplotypes. Figure 6.2 presents the distribution from the same dataset, using a doubled penalty term for determining the core haplotypes. The clustering algorithm assumes a larger penalty function, and as a result only designates 6 core haplotypes. Another ad-hoc method is to determine the desired number of core haplotypes by inspecting the haplotype distribution before performing the clustering. Using this method, the user can directly choose which haplotypes to include as core haplotypes. Our future work will look at developing a more rigorous method for determining the core haplotypes in the context of interaction tests.

We also plan to improve the clustering algorithm by incorporating it into the haplotype frequency estimation. Currently the allocation matrix $B(\mathbf{p})$ is determined based on original estimates of the haplotype frequencies in the combined sample of cases and controls. Thus we are assuming one evolutionary tree for the combined sample. For testing for global haplotype main effects, this is acceptable since cases and controls are the same under the null hypothesis. But for testing for interaction effects, we may want to estimate two separate evolutionary trees, and therefore two separate allocation matrices, for cases and controls. We will explore the feasibility of using two evolutionary trees and what effect it might have on the performance of the method. We also plan to investigate how the determination of the core haplotypes and creation of the allocation matrix can be combined with the estimation of the haplotype frequencies into one process.

Bibliography

Boos, D. (1992), 'On generalized score tests', *The American Statistician* **46**, 327–333.

Carroll, R., Wang, S. and Wang, C. (1995), 'Prospective analysis of logistic case-control studies', *Journal of the American Statistical Association* **90**, 157–169.

Chatterjee, N. and Carroll, R. (2005), 'Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies', *Biometrika* **92**, 399–418.

Chen, Y.-H., Chatterjee, N. and Carroll, R. (2007), 'Retrospective analysis of haplotype-based case-control studies under a flexible model for gene-environment association', *Biostatistics* [**Epub ahead of print**].

Chen, Y.-H. and Kao, J.-T. (2006), 'Multinomial logistic regression approach to haplotype association analysis in population-based case-control studies', *BMC Genetics* **7:43**.

Durrant, C., Zondervan, K., Cardon, L., Hunt, S., Deloukas, P. and Morris, A. (2004), 'Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes', *American Journal of Human Genetics* **75**, 35–43.

Epstein, M. and Satten, G. (2003), 'Inference on haplotype effects in case-control studies using unphased genotype data', *American Journal of Human Genetics* **73**, 1316–1329.

Fallin, D. and Schork, N. (2000), 'Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data', *American Journal of Human Genetics* **67**, 947–959.

Kao, J.-T., Wen, H.-C., Chien, K.-L., Hsu, H.-C. and Lin, S.-W. (2003), 'A novel genetic variant in the apolipoprotein a5 gene is associated with hypertriglyceridemia', *Human Molecular Genetics* **12**, 2533–2539.

Kwee, L., Epstein, M., Manatunga, A., Duncan, R., Allen, A. and Satten, G. (2007), 'Simple methods for assessing haplotype-environment interactions in case-only and case-control studies', *Genetic Epidemiology* **31**, 75–90.

Lake, S., Lyon, H., Tantisira, K., Silverman, E., Weiss, S., Laird, N. and Schaid, D. (2003), 'Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous', *Human Heredity* **55**, 56–65.

Lin, D. Y., Zeng, D. and Millikan, R. (2005), 'Maximum likelihood estimation of haplotype effects and haplotype-environment interactions in association studies', *Genetic Epidemiology* **29**, 299–315.

Lin, D. and Zeng, D. (2006), 'Likelihood-based inference on haplotype effects in genetic association studies', *Journal of the American Statistical Association* **101**, 89–104.

Meng, X. and Rubin, D. (1993), 'Maximum likelihood estimation via the ecm algorithm: a general framework', *Biometrika* **80**, 267–278.

Molitor, J., Marjoram, P. and Thomas, D. (2003), 'Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques', *American Journal of Human Genetics* **73**, 1368–1384.

Morris, R. and Kaplan, N. (2002), 'On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles', *Genetic Epidemiology* **23**, 221–233.

Prentice, R. and Pyke, R. (1979), 'Logistic disease incidence models and case-control studies', *Biometrika* **66**, 403–411.

Roeder, K., Bacanu, S., Sonpar, V., Zhang, X. and Devlin, B. (2005), 'Analysis of single-locus tests to detect gene/disease associations', *Genetic Epidemiology* **28**, 207–219.

Satten, G. A. and Epstein, M. P. (2004), 'Comparison of prospective and retrospective methods for haplotype inference in case-control studies', *Genetic Epidemiology* **27**, 192–201.

Schaid, D., Rowland, C., Tines, D., Jacobson, R. and Poland, G. (2002), 'Score tests for association between traits and haplotypes when linkage phase is ambiguous', *American Journal of Human Genetics* **70**, 425–434.

Seltman, H., Roeder, K. and Devlin, B. (2003), 'Evolutionary-based association analysis using haplotype data', *Genetic Epidemiology* **25**, 48–58.

Spinka, C., Carroll, R. and Chatterjee, N. (2005), 'Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity', *Genetic Epidemiology* **29**, 108–127.

Templeton, A. (1995), 'A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping or dna sequencing. v. analysis of case/control sampling designs: Alzheimer's disease and the apoprotein e locus', *Genetics* **140**, 403–409.

Templeton, A., Boerwinkle, E. and Sing, C. (1987), 'A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. i. basic theory and an analysis of alcohol dehydrogenase activity in drosophila', *Genetics* **117**, 343–351.

Tzeng, J. (2005), 'Evolutionary-based grouping of haplotypes in association analysis', *Genetic Epidemiology* **28**, 220–231.

Tzeng, J.-Y., Wang, C.-H., Kao, J.-T. and Hsiao, C. (2006), 'Regression-based assocation analysis with clustered haplotypes through use of genotypes', *American Journal of Human Genetics* **78**, 231–242.

Waldron, E. and Whittaker, J.and Balding, D. (2006), 'Fine mapping of disease genes via haplotype clustering', *Genetic Epidemiology* **30**, 170–179.

Yu, K., Gu, C., Province, M., Xiong, C. and Rao, D. (2004), 'Genetic association mapping under founder heterogeneity via weighted haplotype similarity analysis in candidate genes', *Genetic Epidemiology* **27**, 182–191.

Zhao, J., Curtis, D. and Sham, P. (2000), 'Model-free analysis and permutation tests for allelic associations', *Human Heredity* **50**, 133–139.

Zhao, L., Li, S. and Khalid, N. (2003), 'A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies', *American Journal of Human Genetics* **72**, 1231–1250.

Appendix

Appendix A

Score Statistic Details for Testing for Haplotype Main Effects

The score statistic for main haplotype effects is

$$S_{\boldsymbol{\beta}} = U_{\boldsymbol{\beta}}^T V_{\boldsymbol{\beta}}^{-1} U_{\boldsymbol{\beta}} \Big|_{\substack{\boldsymbol{\beta}=0 \\ \boldsymbol{\xi}=\tilde{\boldsymbol{\xi}}}}$$

where

$$U_{\boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}} \log L_{obs}$$

and

$$V_{\boldsymbol{\beta}} = D_{\boldsymbol{\beta}\boldsymbol{\beta}} - I_{\boldsymbol{\beta}\boldsymbol{\xi}} I_{\boldsymbol{\xi}\boldsymbol{\xi}}^{-1} D_{\boldsymbol{\beta}\boldsymbol{\xi}}^T - D_{\boldsymbol{\beta}\boldsymbol{\xi}} I_{\boldsymbol{\xi}\boldsymbol{\xi}}^{-1} I_{\boldsymbol{\beta}\boldsymbol{\xi}}^T + I_{\boldsymbol{\beta}\boldsymbol{\xi}} I_{\boldsymbol{\xi}\boldsymbol{\xi}}^{-1} D_{\boldsymbol{\xi}\boldsymbol{\xi}} I_{\boldsymbol{\xi}\boldsymbol{\xi}}^{-1} I_{\boldsymbol{\beta}\boldsymbol{\xi}}^T,$$

We write the variance-covariance matrix D as

$$D = \begin{pmatrix} D_{\boldsymbol{\beta}\boldsymbol{\beta}} & D_{\boldsymbol{\beta}\boldsymbol{\xi}} \\ D_{\boldsymbol{\beta}\boldsymbol{\xi}}^T & D_{\boldsymbol{\xi}\boldsymbol{\xi}} \end{pmatrix}$$

where

$$D_{\boldsymbol{\beta}\boldsymbol{\xi}} = \begin{pmatrix} D_{\boldsymbol{\beta}_{\alpha^*}} & D_{\boldsymbol{\beta}\boldsymbol{\gamma}} & D_{\boldsymbol{\beta}_{p_1}} & \dots & D_{\boldsymbol{\beta}_{p_{L+1}}} \end{pmatrix}$$

and

$$D_{\boldsymbol{\xi}\boldsymbol{\xi}} = \begin{pmatrix} D_{\alpha^*\alpha^*} & D_{\alpha^*\boldsymbol{\gamma}} & D_{\alpha^* p_1} & \dots & D_{\alpha^* p_{L+1}} \\ D'_{\alpha^*\boldsymbol{\gamma}} & D_{\boldsymbol{\gamma}\boldsymbol{\gamma}} & D_{\boldsymbol{\gamma} p_1} & \dots & D_{\boldsymbol{\gamma} p_{L+1}} \\ D'_{\alpha^* p_1} & D'_{\boldsymbol{\gamma} p_1} & D_{p_1 p_1} & \dots & D_{p_1 p_{L+1}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \dots & \dots & \vdots & D_{p_i p_j} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ D'_{\alpha^* p_{L+1}} & D'_{\boldsymbol{\gamma} p_{L+1}} & D'_{p_1 p_{L+1}} & \dots & D_{p_{L+1} p_{L+1}} \end{pmatrix}$$

Let

$s_i(y_i, g_i, \alpha^*) = \frac{\partial}{\partial \alpha^*} \log L_i$

$s_i(y_i, g_i, \boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log L_i$

$s_i(y_i, g_i, \boldsymbol{\gamma}) = \frac{\partial}{\partial \boldsymbol{\gamma}} \log L_i$

$s_i(y_i, g_i, p_\tau) = \frac{\partial}{\partial p_\tau} \log L_i.$

Then we define the elements of D as

$D_{\alpha^*\alpha^*} = \sum_{i=1}^n s_i(y_i, g_i, \alpha^*) s_i(y_i, g_i, \alpha^*)$

$D_{\alpha^*\boldsymbol{\beta}} = \sum_{i=1}^n s_i(y_i, g_i, \alpha^*) s_i(y_i, g_i, \boldsymbol{\beta})$

$D_{\alpha^*\boldsymbol{\gamma}} = \sum_{i=1}^n s_i(y_i, g_i, \alpha^*) s_i(y_i, g_i, \boldsymbol{\gamma})$

$D_{\alpha^* p_\tau} = \sum_{i=1}^n s_i(y_i, g_i, \alpha^*) s_i(y_i, g_i, p_\tau)$

$D_{\boldsymbol{\beta}\boldsymbol{\beta}} = \sum_{i=1}^n s_i(y_i, g_i, \boldsymbol{\beta}) s_i(y_i, g_i, \boldsymbol{\beta})^T$

$D_{\boldsymbol{\beta}\boldsymbol{\gamma}} = \sum_{i=1}^n s_i(y_i, g_i, \boldsymbol{\beta}) s_i(y_i, g_i, \boldsymbol{\gamma})^T$

$D_{\boldsymbol{\beta} p_\tau} = \sum_{i=1}^n s_i(y_i, g_i, \boldsymbol{\beta}) s_i(y_i, g_i, p_\tau)$

$D_{\boldsymbol{\gamma}\boldsymbol{\gamma}} = \sum_{i=1}^n s_i(y_i, g_i, \boldsymbol{\gamma}) s_i(y_i, g_i, \boldsymbol{\gamma})^T$

$D_{\boldsymbol{\gamma} p_\tau} = \sum_{i=1}^n s_i(y_i, g_i, \boldsymbol{\gamma}) s_i(y_i, g_i, p_\tau)$

$D_{p_\tau p_\tau} = \sum_{i=1}^n s_i(y_i, g_i, p_\tau) s_i(y_i, g_i, p_\tau)$

$D_{p_\tau p_\theta} = \sum_{i=1}^n s_i(y_i, g_i, p_\tau) s_i(y_i, g_i, p_\theta).$

We can write the observed information matrix I as

$$I = \begin{pmatrix} I_{\boldsymbol{\beta}\boldsymbol{\beta}} & I_{\boldsymbol{\beta}\boldsymbol{\xi}} \\ I^T_{\boldsymbol{\beta}\boldsymbol{\xi}} & I_{\boldsymbol{\xi}\boldsymbol{\xi}} \end{pmatrix}$$

where

$$I_{\boldsymbol{\beta}\boldsymbol{\xi}} = \begin{pmatrix} I_{\boldsymbol{\beta}\alpha^*} & I_{\boldsymbol{\beta}\boldsymbol{\gamma}} & I_{\boldsymbol{\beta}p_1} & \dots & I_{\boldsymbol{\beta}p_{L+1}} \end{pmatrix}$$

and

$$I_{\boldsymbol{\xi}\boldsymbol{\xi}} = \begin{pmatrix} I_{\alpha^*\alpha^*} & I_{\alpha^*\boldsymbol{\gamma}} & I_{\alpha^* p_1} & \dots & I_{\alpha^* p_{L+1}} \\ I'_{\alpha^*\boldsymbol{\gamma}} & I_{\boldsymbol{\gamma}\boldsymbol{\gamma}} & I_{\boldsymbol{\gamma}p_1} & \dots & I_{\boldsymbol{\gamma}p_{L+1}} \\ I'_{\alpha^* p_1} & I'_{\boldsymbol{\gamma}p_1} & I_{p_1p_1} & \dots & I_{p_1p_{L+1}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \dots & \dots & \vdots & I_{p_ip_j} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ I'_{\alpha^* p_{L+1}} & I'_{\boldsymbol{\gamma}p_{L+1}} & I'_{p_1p_{L+1}} & \dots & I_{p_{L+1}p_{L+1}} \end{pmatrix}$$

and the elements of I as

$I_{\alpha^*\alpha^*} = -\frac{\partial}{\partial\alpha^*}\frac{\partial}{\partial\alpha^*}\log L_i$

$I_{\boldsymbol{\beta}\alpha^*} = -\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial}{\partial\alpha^*}\log L_i$

$I_{\boldsymbol{\gamma}\alpha^*} = -\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\alpha^*}\log L_i$

$I_{p_\tau \alpha^*} = -\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial \alpha^*}\log L_i$

$I_{\boldsymbol{\beta}\boldsymbol{\beta}} = -\frac{\partial}{\partial \boldsymbol{\beta}}\frac{\partial}{\partial \boldsymbol{\beta}}\log L_i$

$I_{\boldsymbol{\gamma}\boldsymbol{\beta}} = -\frac{\partial}{\partial \boldsymbol{\gamma}}\frac{\partial}{\partial \boldsymbol{\beta}}\log L_i$

$I_{p_\tau \boldsymbol{\beta}} = -\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial \boldsymbol{\beta}}\log L_i$

$I_{\boldsymbol{\gamma}\boldsymbol{\gamma}} = -\frac{\partial}{\partial \boldsymbol{\gamma}}\frac{\partial}{\partial \boldsymbol{\gamma}}\log L_i$

$I_{p_\tau \boldsymbol{\gamma}} = -\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial \boldsymbol{\gamma}}\log L_i$

$I_{p_\tau p_\tau} = -\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial p_\tau}\log L_i$

$I_{p_\tau p_\theta} = -\frac{\partial}{\partial p_\theta}\frac{\partial}{\partial p_\tau}\log L_i$

To simplify notation, define the quantities v, u, and w for each individual as

$v = \sum_{H \in S(G_i)} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma}) p_h p_{h'}$

$u = \sum_{H} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma}) p_h p_{h'}$

$w = \sum_{H \in S(G_i)} p_h p_{h'}$

The derivatives of v with respect to the parameters α^*, $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$, and a specific p_τ are

$\frac{\partial}{\partial \alpha^*} v = \sum_{H \in S(G_i)} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma}) p_h p_{h'}$

$\frac{\partial}{\partial \boldsymbol{\beta}} v = \sum_{H \in S(G_i)} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma}) X_{C_H}^T p_h p_{h'}$

$\frac{\partial}{\partial \boldsymbol{\gamma}} v = \sum_{H \in S(G_i)} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma}) X_{E_i}^T p_h p_{h'}$

$\frac{\partial}{\partial p_\tau} v = \sum_{H \in S(G_i)} 2I(h = \tau) \left\{ p_{h'} exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma}) \right\}$

$\frac{\partial}{\partial \alpha^*}\frac{\partial}{\partial \alpha^*} v = v$

$\frac{\partial}{\partial \boldsymbol{\beta}}\frac{\partial}{\partial \alpha^*} v = \frac{\partial}{\partial \boldsymbol{\beta}} v$

$\frac{\partial}{\partial \boldsymbol{\gamma}}\frac{\partial}{\partial \alpha^*} v = \frac{\partial}{\partial \boldsymbol{\gamma}} v$

$\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial \alpha^*} v = \frac{\partial}{\partial p_\tau} v$

$\frac{\partial}{\partial \boldsymbol{\beta}}\frac{\partial}{\partial \boldsymbol{\beta}} v = \sum_{H \in S(G_i)} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma}) X_{C_H}^T X_{C_H} p_h p_{h'}$

$$\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\boldsymbol{\beta}}v = \textstyle\sum_{H\in S(G_i)} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma})X_{C_H}^T X_{E_i} p_h p_{h'}$$
$$\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\boldsymbol{\beta}}v = \textstyle\sum_{H\in S(G_i)} 2I(h=\tau)\left\{ \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma})X_{C_H}^T p_{h'}\right\}$$
$$\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\boldsymbol{\gamma}}v = \textstyle\sum_{H\in S(G_i)} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma})X_{E_i}^T X_{E_i} p_h p_{h'}$$
$$\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\boldsymbol{\gamma}}v = \textstyle\sum_{H\in S(G_i)} 2I(h=\tau)\left\{ \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma})X_{E_i}^T p_{h'}\right\}$$
$$\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial p_\tau}v = I(h=h'=\tau)\left\{2\exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma})\right\}$$
$$\frac{\partial}{\partial p_\theta}\frac{\partial}{\partial p_\tau}v = \textstyle\sum_{H\in S(G_i)} 2I(h_j=\tau, h'_j=\theta)\left\{ \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma})\right\}$$

The derivatives of u are the same as those for v, except they are summed over all haplotype pairs H instead of $H \in S(G_i)$. The derivates of w are

$$\frac{\partial}{\partial p_\tau}w = \textstyle\sum_{H\in S(G_i)} 2I(h=\tau)p_{h'}$$
$$\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial p_\tau}w = 2I(h=h'=\tau)$$
$$\frac{\partial}{\partial p_\theta}\frac{\partial}{\partial p_\tau}w = \textstyle\sum_{H\in S(G_i)} 2I(h_j=\tau, h'_j=\theta)$$

Now we can write the components of D and I in terms of derivatives of v, u, and w.

$$\frac{\partial}{\partial\alpha^*}\log L_i = y_i\frac{\frac{\partial}{\partial\alpha^*}v}{v} - \frac{\frac{\partial}{\partial\alpha^*}u}{u}$$
$$\frac{\partial}{\partial\boldsymbol{\beta}}\log L_i = y_i\frac{\frac{\partial}{\partial\boldsymbol{\beta}}v}{v} - \frac{\frac{\partial}{\partial\boldsymbol{\beta}}u}{u}$$
$$\frac{\partial}{\partial\boldsymbol{\gamma}}\log L_i = y_i\frac{\frac{\partial}{\partial\boldsymbol{\gamma}}v}{v} - \frac{\frac{\partial}{\partial\boldsymbol{\gamma}}u}{u}$$
$$\frac{\partial}{\partial p_\tau}\log L_i = (1-y_i)\frac{\frac{\partial}{\partial p_\tau}w}{w} + y_i\frac{\frac{\partial}{\partial p_\tau}v}{v} - \frac{\frac{\partial}{\partial p_\tau}u}{u}$$
$$\frac{\partial}{\partial\alpha^*}\frac{\partial}{\partial\alpha^*}\log L_i = y_i\frac{v(\frac{\partial}{\partial\alpha^*}\frac{\partial}{\partial\alpha^*}v)-(\frac{\partial}{\partial\alpha^*}v)(\frac{\partial}{\partial\alpha^*}v)}{v^2} - \frac{u(\frac{\partial}{\partial\alpha^*}\frac{\partial}{\partial\alpha^*}u)-(\frac{\partial}{\partial\alpha^*}u)(\frac{\partial}{\partial\alpha^*}u)}{u^2}$$
$$\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial}{\partial\alpha^*}\log L_i = y_i\frac{v(\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial}{\partial\alpha^*}v)-(\frac{\partial}{\partial\alpha^*}v)(\frac{\partial}{\partial\boldsymbol{\beta}}v)}{v^2} - \frac{u(\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial}{\partial\alpha^*}u)-(\frac{\partial}{\partial\alpha^*}u)(\frac{\partial}{\partial\boldsymbol{\beta}}u)}{u^2}$$
$$\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\alpha^*}\log L_i = y_i\frac{v(\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\alpha^*}v)-(\frac{\partial}{\partial\alpha^*}v)(\frac{\partial}{\partial\boldsymbol{\gamma}}v)}{v^2} - \frac{u(\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\alpha^*}u)-(\frac{\partial}{\partial\alpha^*}u)(\frac{\partial}{\partial\boldsymbol{\gamma}}u)}{u^2}$$
$$\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\alpha^*}\log L_i = y_i\frac{v(\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\alpha^*}v)-(\frac{\partial}{\partial\alpha^*}v)(\frac{\partial}{\partial p_\tau}v)}{v^2} - \frac{u(\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\alpha^*}u)-(\frac{\partial}{\partial p_\tau}u)(\frac{\partial}{\partial\alpha^*}u)}{u^2}$$

$$\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial}{\partial\boldsymbol{\beta}}\log L_i = y_i\frac{v(\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial}{\partial\boldsymbol{\beta}}v)-(\frac{\partial}{\partial\boldsymbol{\beta}}v)(\frac{\partial}{\partial\boldsymbol{\beta}}v)}{v^2}-\frac{u(\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial}{\partial\boldsymbol{\beta}}u)-(\frac{\partial}{\partial\boldsymbol{\beta}}u)(\frac{\partial}{\partial\boldsymbol{\beta}}u)}{u^2}$$

$$\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\boldsymbol{\beta}}\log L_i = y_i\frac{v(\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\boldsymbol{\beta}}v)-(\frac{\partial}{\partial\boldsymbol{\beta}}v)(\frac{\partial}{\partial\boldsymbol{\gamma}}v)}{v^2}-\frac{u(\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\boldsymbol{\beta}}u)-(\frac{\partial}{\partial\boldsymbol{\beta}}u)(\frac{\partial}{\partial\boldsymbol{\gamma}}u)}{u^2}$$

$$\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\boldsymbol{\beta}}\log L_i = y_i\frac{v(\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\boldsymbol{\beta}}v)-(\frac{\partial}{\partial\boldsymbol{\beta}}v)(\frac{\partial}{\partial p_\tau}v)}{v^2}-\frac{u(\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\boldsymbol{\beta}}u)-(\frac{\partial}{\partial\boldsymbol{\beta}}u)(\frac{\partial}{\partial p_\tau}u)}{u^2}$$

$$\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\boldsymbol{\gamma}}\log L_i = y_i\frac{v(\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\boldsymbol{\gamma}}v)-(\frac{\partial}{\partial\boldsymbol{\gamma}}v)(\frac{\partial}{\partial\boldsymbol{\gamma}}v)}{v^2}-\frac{u(\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\boldsymbol{\gamma}}u)-(\frac{\partial}{\partial\boldsymbol{\gamma}}u)(\frac{\partial}{\partial\boldsymbol{\gamma}}u)}{u^2}$$

$$\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\boldsymbol{\gamma}}\log L_i = y_i\frac{v(\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\boldsymbol{\gamma}}v)-(\frac{\partial}{\partial\boldsymbol{\gamma}}v)(\frac{\partial}{\partial p_\tau}v)}{v^2}-\frac{u(\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\boldsymbol{\gamma}}u)-(\frac{\partial}{\partial\boldsymbol{\gamma}}u)(\frac{\partial}{\partial p_\tau}u)}{u^2}$$

$$\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial p_\tau}\log L_i = (1-y_i)\frac{w(\frac{\partial^2}{\partial p_\tau^2}w)-(\frac{\partial}{\partial p_\tau}w)(\frac{\partial}{\partial p_\tau}w)}{w^2}+y_i\frac{v(\frac{\partial^2}{\partial p_\tau^2}v)-(\frac{\partial}{\partial p_\tau}v)(\frac{\partial}{\partial p_\tau}v)}{v^2}-\frac{u(\frac{\partial^2}{\partial p_\tau^2}u)-(\frac{\partial}{\partial p_\tau}u)(\frac{\partial}{\partial p_\tau}u)}{u^2}$$

$$\frac{\partial}{\partial p_\theta}\frac{\partial}{\partial p_\tau}\log L_i = (1-y_i)\frac{w(\frac{\partial}{\partial p_\theta}\frac{\partial}{\partial p_\tau}w)-(\frac{\partial}{\partial p_\tau}w)(\frac{\partial}{\partial p_\theta}w)}{w^2}+y_i\frac{v(\frac{\partial}{\partial p_\theta}\frac{\partial}{\partial p_\tau}v)-(\frac{\partial}{\partial p_\tau}v)(\frac{\partial}{\partial p_\theta}v)}{v^2}-$$
$$\frac{u(\frac{\partial}{\partial p_\theta}\frac{\partial}{\partial p_\tau}u)-(\frac{\partial}{\partial p_\tau}u)(\frac{\partial}{\partial p_\theta}u)}{u^2}$$

Appendix B

Score Statistic Details for Testing for Interaction Effects

The score statistic for haplotype-environment interaction effects is

$$S_{\boldsymbol{\nu}} = U_{\boldsymbol{\nu}}^T V_{\boldsymbol{\nu}}^{-1} U_{\boldsymbol{\nu}} \Big|_{\substack{\boldsymbol{\nu}=0 \\ \boldsymbol{\xi}=\tilde{\boldsymbol{\xi}}}}$$

where

$$U_{\boldsymbol{\nu}} = \frac{\partial}{\partial \boldsymbol{\nu}} \log L_{obs}$$

and

$$V_{\boldsymbol{\nu}} = D_{\boldsymbol{\nu}\boldsymbol{\nu}} - I_{\boldsymbol{\nu}\boldsymbol{\xi}} I_{\boldsymbol{\xi}\boldsymbol{\xi}}^{-1} D_{\boldsymbol{\nu}\boldsymbol{\xi}}^T - D_{\boldsymbol{\nu}\boldsymbol{\xi}} I_{\boldsymbol{\xi}\boldsymbol{\xi}}^{-1} I_{\boldsymbol{\nu}\boldsymbol{\xi}}^T + I_{\boldsymbol{\nu}\boldsymbol{\xi}} I_{\boldsymbol{\xi}\boldsymbol{\xi}}^{-1} D_{\boldsymbol{\xi}\boldsymbol{\xi}} I_{\boldsymbol{\xi}\boldsymbol{\xi}}^{-1} I_{\boldsymbol{\nu}\boldsymbol{\xi}}^T,$$

We write the variance-covariance matrix D as

$$D = \begin{pmatrix} D_{\boldsymbol{\nu}\boldsymbol{\nu}} & D_{\boldsymbol{\nu}\boldsymbol{\xi}} \\ D_{\boldsymbol{\nu}\boldsymbol{\xi}}^T & D_{\boldsymbol{\xi}\boldsymbol{\xi}} \end{pmatrix}$$

where

$$D_{\boldsymbol{\nu}\boldsymbol{\xi}} = \begin{pmatrix} D_{\boldsymbol{\nu}\alpha^*} & D_{\boldsymbol{\nu}\boldsymbol{\beta}} & D_{\boldsymbol{\nu}\boldsymbol{\gamma}} & D_{\boldsymbol{\nu}p_1} & \ldots & D_{\boldsymbol{\nu}p_{L+1}} \end{pmatrix}$$

and

$$D_{\boldsymbol{\xi}\boldsymbol{\xi}} = \begin{pmatrix} D_{\alpha^*\alpha^*} & D_{\alpha^*\boldsymbol{\beta}} & D_{\alpha^*\boldsymbol{\gamma}} & D_{\alpha^* p_1} & \dots & D_{\alpha^* p_{L+1}} \\ D'_{\alpha^*\boldsymbol{\beta}} & D_{\boldsymbol{\beta}\boldsymbol{\beta}} & D_{\boldsymbol{\beta}\boldsymbol{\gamma}} & D_{\boldsymbol{\beta} p_1} & \dots & D_{\boldsymbol{\beta} p_{L+1}} \\ D'_{\alpha^*\boldsymbol{\gamma}} & D'_{\boldsymbol{\beta}\boldsymbol{\gamma}} & D_{\boldsymbol{\gamma}\boldsymbol{\gamma}} & D_{\boldsymbol{\gamma} p_1} & \dots & D_{\boldsymbol{\gamma} p_{L+1}} \\ D'_{\alpha^* p_1} & D'_{\boldsymbol{\beta} p_1} & D'_{\boldsymbol{\gamma} p_1} & D_{p_1 p_1} & \dots & D_{p_1 p_{L+1}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \dots & \dots & \vdots & \vdots & D_{p_i p_j} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ D'_{\alpha^* p_{L+1}} & D'_{\boldsymbol{\beta} p_{L+1}} & D'_{\boldsymbol{\gamma} p_{L+1}} & D'_{p_1 p_{L+1}} & \dots & D_{p_{L+1} p_{L+1}} \end{pmatrix}$$

Let

$s_i(y_i, g_i, \alpha^*) = \frac{\partial}{\partial \alpha^*} \log L_i$

$s_i(y_i, g_i, \boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log L_i$

$s_i(y_i, g_i, \boldsymbol{\gamma}) = \frac{\partial}{\partial \boldsymbol{\gamma}} \log L_i$

$s_i(y_i, g_i, p_\tau) = \frac{\partial}{\partial p_\tau} \log L_i$

$s_i(y_i, g_i, \boldsymbol{\nu}) = \frac{\partial}{\partial \boldsymbol{\nu}} \log L_i.$

Then we define the elements of D as

$D_{\alpha^*\alpha^*} = \sum_{i=1}^n s_i(y_i, g_i, \alpha^*) s_i(y_i, g_i, \alpha^*)$

$D_{\alpha^*\boldsymbol{\beta}} = \sum_{i=1}^n s_i(y_i, g_i, \alpha^*) s_i(y_i, g_i, \boldsymbol{\beta})$

$D_{\alpha^*\boldsymbol{\gamma}} = \sum_{i=1}^n s_i(y_i, g_i, \alpha^*) s_i(y_i, g_i, \boldsymbol{\gamma})$

$D_{\alpha^* p_\tau} = \sum_{i=1}^n s_i(y_i, g_i, \alpha^*) s_i(y_i, g_i, p_\tau)$

$D_{\alpha^*\boldsymbol{\nu}} = \sum_{i=1}^n s_i(y_i, g_i, \alpha^*) s_i(y_i, g_i, \boldsymbol{\nu})$

$D_{\boldsymbol{\beta}\boldsymbol{\beta}} = \sum_{i=1}^n s_i(y_i, g_i, \boldsymbol{\beta}) s_i(y_i, g_i, \boldsymbol{\beta})^T$

$D_{\boldsymbol{\beta}\boldsymbol{\gamma}} = \sum_{i=1}^{n} s_i(y_i, g_i, \boldsymbol{\beta}) s_i(y_i, g_i, \boldsymbol{\gamma})^T$

$D_{\boldsymbol{\beta} p_\tau} = \sum_{i=1}^{n} s_i(y_i, g_i, \boldsymbol{\beta}) s_i(y_i, g_i, p_\tau)$

$D_{\boldsymbol{\beta}\boldsymbol{\nu}} = \sum_{i=1}^{n} s_i(y_i, g_i, \boldsymbol{\beta}) s_i(y_i, g_i, \boldsymbol{\nu})^T$

$D_{\boldsymbol{\gamma}\boldsymbol{\gamma}} = \sum_{i=1}^{n} s_i(y_i, g_i, \boldsymbol{\gamma}) s_i(y_i, g_i, \boldsymbol{\gamma})^T$

$D_{\boldsymbol{\gamma} p_\tau} = \sum_{i=1}^{n} s_i(y_i, g_i, \boldsymbol{\gamma}) s_i(y_i, g_i, p_\tau)$

$D_{\boldsymbol{\gamma}\boldsymbol{\nu}} = \sum_{i=1}^{n} s_i(y_i, g_i, \boldsymbol{\gamma}) s_i(y_i, g_i, \boldsymbol{\nu})^T$

$D_{p_\tau p_\tau} = \sum_{i=1}^{n} s_i(y_i, g_i, p_\tau) s_i(y_i, g_i, p_\tau)$

$D_{p_\tau p_\theta} = \sum_{i=1}^{n} s_i(y_i, g_i, p_\tau) s_i(y_i, g_i, p_\theta)$

$D_{p_\tau \boldsymbol{\nu}} = \sum_{i=1}^{n} s_i(y_i, g_i, p_\tau) s_i(y_i, g_i, \boldsymbol{\nu})$

$D_{\boldsymbol{\nu}\boldsymbol{\nu}} = \sum_{i=1}^{n} s_i(y_i, g_i, \boldsymbol{\nu}) s_i(y_i, g_i, \boldsymbol{\nu})^T.$

We can write the observed information matrix I as

$$I = \begin{pmatrix} I_{\boldsymbol{\nu}\boldsymbol{\nu}} & I_{\boldsymbol{\nu}\boldsymbol{\xi}} \\ I_{\boldsymbol{\nu}\boldsymbol{\xi}}^T & I_{\boldsymbol{\xi}\boldsymbol{\xi}} \end{pmatrix}$$

where

$$I_{\boldsymbol{\nu}\boldsymbol{\xi}} = \begin{pmatrix} I_{\boldsymbol{\nu}\alpha^*} & I_{\boldsymbol{\nu}\boldsymbol{\beta}} & I_{\boldsymbol{\nu}\boldsymbol{\gamma}} & I_{\boldsymbol{\nu}p_1} & \dots & I_{\boldsymbol{\nu}p_{L+1}} \end{pmatrix}$$

and

$$I_{\boldsymbol{\xi}\boldsymbol{\xi}} = \begin{pmatrix} I_{\alpha^*\alpha^*} & I_{\alpha^*\boldsymbol{\beta}} & I_{\alpha^*\boldsymbol{\gamma}} & I_{\alpha^* p_1} & \dots & I_{\alpha^* p_{L+1}} \\ I'_{\alpha^*\boldsymbol{\beta}} & I_{\boldsymbol{\beta}\boldsymbol{\beta}} & I_{\boldsymbol{\beta}\boldsymbol{\gamma}} & I_{\boldsymbol{\beta} p_1} & \dots & I_{\boldsymbol{\beta} p_{L+1}} \\ I'_{\alpha^*\boldsymbol{\gamma}} & I'_{\boldsymbol{\beta}\boldsymbol{\gamma}} & I_{\boldsymbol{\gamma}\boldsymbol{\gamma}} & I_{\boldsymbol{\gamma} p_1} & \dots & I_{\boldsymbol{\gamma} p_{L+1}} \\ I'_{\alpha^* p_1} & I'_{\boldsymbol{\beta} p_1} & I'_{\boldsymbol{\gamma} p_1} & I_{p_1 p_1} & \dots & I_{p_1 p_{L+1}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \dots & \dots & \vdots & \vdots & I_{p_i p_j} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ I'_{\alpha^* p_{L+1}} & I'_{\boldsymbol{\beta} p_{L+1}} & I'_{\boldsymbol{\gamma} p_{L+1}} & I'_{p_1 p_{L+1}} & \dots & I_{p_{L+1} p_{L+1}} \end{pmatrix}$$

and the elements of I as

$I_{\alpha^*\alpha^*} = -\frac{\partial}{\partial\alpha^*}\frac{\partial}{\partial\alpha^*}\log L_i$

$I_{\boldsymbol{\beta}\alpha^*} = -\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial}{\partial\alpha^*}\log L_i$

$I_{\boldsymbol{\gamma}\alpha^*} = -\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\alpha^*}\log L_i$

$I_{p_\tau\alpha^*} = -\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\alpha^*}\log L_i$

$I_{\boldsymbol{\nu}\alpha^*} = -\frac{\partial}{\partial\boldsymbol{\nu}}\frac{\partial}{\partial\alpha^*}\log L_i$

$I_{\boldsymbol{\beta}\boldsymbol{\beta}} = -\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial}{\partial\boldsymbol{\beta}}\log L_i$

$I_{\boldsymbol{\gamma}\boldsymbol{\beta}} = -\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\boldsymbol{\beta}}\log L_i$

$I_{p_\tau\boldsymbol{\beta}} = -\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\boldsymbol{\beta}}\log L_i$

$I_{\boldsymbol{\nu}\boldsymbol{\beta}} = -\frac{\partial}{\partial\boldsymbol{\nu}}\frac{\partial}{\partial\boldsymbol{\beta}}\log L_i$

$I_{\boldsymbol{\gamma}\boldsymbol{\gamma}} = -\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\boldsymbol{\gamma}}\log L_i$

$I_{p_\tau\boldsymbol{\gamma}} = -\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\boldsymbol{\gamma}}\log L_i$

$I_{\boldsymbol{\nu}\boldsymbol{\gamma}} = -\frac{\partial}{\partial\boldsymbol{\nu}}\frac{\partial}{\partial\boldsymbol{\gamma}}\log L_i$

$$I_{p_\tau p_\tau} = -\tfrac{\partial}{\partial p_\tau}\tfrac{\partial}{\partial p_\tau}\log L_i$$
$$I_{p_\tau p_\theta} = -\tfrac{\partial}{\partial p_\theta}\tfrac{\partial}{\partial p_\tau}\log L_i$$
$$I_{\boldsymbol{\nu} p_\theta} = -\tfrac{\partial}{\partial \boldsymbol{\nu}}\tfrac{\partial}{\partial p_\tau}\log L_i.$$

To simplify notation, define the quantities v, u, and w for each individual as

$$v = \textstyle\sum_{H\in S(G_i)} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma} + X_{HE_i}\boldsymbol{\nu})p_h p_{h'}$$
$$u = \textstyle\sum_{H} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma} + X_{HE_i}\boldsymbol{\nu})p_h p_{h'}$$
$$w = \textstyle\sum_{H\in S(G_i)} p_h p_{h'}$$

The derivatives of v with respect to the parameters α^*, $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$, $\boldsymbol{\nu}$, and a specific p_τ are

$$\tfrac{\partial}{\partial \alpha^*}v = \textstyle\sum_{H\in S(G_i)} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma} + X_{HE_i}\boldsymbol{\nu})p_h p_{h'}$$
$$\tfrac{\partial}{\partial \boldsymbol{\beta}}v = \textstyle\sum_{H\in S(G_i)} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma} + X_{HE_i}\boldsymbol{\nu})X_{C_H}^T p_h p_{h'}$$
$$\tfrac{\partial}{\partial \boldsymbol{\gamma}}v = \textstyle\sum_{H\in S(G_i)} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma} + X_{HE_i}\boldsymbol{\nu})X_{E_i}^T p_h p_{h'}$$
$$\tfrac{\partial}{\partial p_\tau}v = \textstyle\sum_{H\in S(G_i)} 2I(h=\tau)\left\{p_{h'}exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma} + X_{HE_i}\boldsymbol{\nu})\right\}$$
$$\tfrac{\partial}{\partial \nu}v = \textstyle\sum_{H\in S(G_i)} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma} + X_{HE_i}\boldsymbol{\nu})X_{HE_i}^T p_h p_{h'}$$

$$\tfrac{\partial}{\partial \alpha^*}\tfrac{\partial}{\partial \alpha^*}v = v$$
$$\tfrac{\partial}{\partial \boldsymbol{\beta}}\tfrac{\partial}{\partial \alpha^*}v = \tfrac{\partial}{\partial \boldsymbol{\beta}}v$$
$$\tfrac{\partial}{\partial \boldsymbol{\gamma}}\tfrac{\partial}{\partial \alpha^*}v = \tfrac{\partial}{\partial \boldsymbol{\gamma}}v$$
$$\tfrac{\partial}{\partial p_\tau}\tfrac{\partial}{\partial \alpha^*}v = \tfrac{\partial}{\partial p_\tau}v$$
$$\tfrac{\partial}{\partial \alpha^*}\tfrac{\partial}{\partial \nu}v = \textstyle\sum_{H\in S(G_i)} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma} + X_{HE_i}\boldsymbol{\nu})X_{HE_i}^T p_h p_{h'}$$
$$\tfrac{\partial}{\partial \boldsymbol{\beta}}\tfrac{\partial}{\partial \boldsymbol{\beta}}v = \textstyle\sum_{H\in S(G_i)} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma} + X_{HE_i}\boldsymbol{\nu})X_{C_H}^T X_{C_H} p_h p_{h'}$$
$$\tfrac{\partial}{\partial \boldsymbol{\gamma}}\tfrac{\partial}{\partial \boldsymbol{\beta}}v = \textstyle\sum_{H\in S(G_i)} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma} + X_{HE_i}\boldsymbol{\nu})X_{C_H}^T X_{E_i} p_h p_{h'}$$
$$\tfrac{\partial}{\partial p_\tau}\tfrac{\partial}{\partial \boldsymbol{\beta}}v = \textstyle\sum_{H\in S(G_i)} 2I(h=\tau)\left\{\exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma} + X_{HE_i}\boldsymbol{\nu})X_{C_H}^T p_{h'}\right\}$$

$$\frac{\partial}{\partial \boldsymbol{\nu}}\frac{\partial}{\partial \boldsymbol{\beta}} v = \textstyle\sum_{H\in S(G_i)} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma} + X_{HE_i}\boldsymbol{\nu}) X_{C_H}^T X_{HE_i} p_h p_{h'}$$

$$\frac{\partial}{\partial \boldsymbol{\gamma}}\frac{\partial}{\partial \boldsymbol{\gamma}} v = \textstyle\sum_{H\in S(G_i)} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma} + X_{HE_i}\boldsymbol{\nu}) X_{E_i}^T X_{E_i} p_h p_{h'}$$

$$\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial \boldsymbol{\gamma}} v = \textstyle\sum_{H\in S(G_i)} 2I(h=\tau)\Bigg\{ \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma} + X_{HE_i}\boldsymbol{\nu}) X_{E_i}^T p_{h'} \Bigg\}$$

$$\frac{\partial}{\partial \boldsymbol{\nu}}\frac{\partial}{\partial \boldsymbol{\gamma}} v = \textstyle\sum_{H\in S(G_i)} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma} + X_{HE_i}\boldsymbol{\nu}) X_{E_i}^T X_{HE_i} p_h p_{h'}$$

$$\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial p_\tau} v = 2I(h=h'=\tau)\Bigg\{ \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma} + X_{HE_i}\boldsymbol{\nu}) \Bigg\}$$

$$\frac{\partial}{\partial p_\theta}\frac{\partial}{\partial p_\tau} v = \textstyle\sum_{H\in S(G_i)} 2I(h_j=\tau, h'_j=\theta)\Bigg\{ \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma} + X_{HE_i}\boldsymbol{\nu}) \Bigg\}$$

$$\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial \nu} v = \textstyle\sum_{H\in S(G_i)} 2I(h=\tau\neq h')\Bigg\{ \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma} + X_{HE_i}\boldsymbol{\nu}) p_{h'} X_{HE_i} \Bigg\} +$$

$$2I(h=h'=\tau)\Bigg\{ \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma} + X_{HE_i}\boldsymbol{\nu}) p_\tau X_{HE_i} \Bigg\}$$

$$\frac{\partial}{\partial \nu}\frac{\partial}{\partial \nu} v = \textstyle\sum_{H\in S(G_i)} \exp(\alpha^* + X_{C_H}\boldsymbol{\beta} + X_{E_i}\boldsymbol{\gamma} + X_{HE_i}\boldsymbol{\nu}) X_{HE_i}^T X_{HE_i} p_h p_{h'}$$

The derivatives of u are the same as those for v, except they are summed over all haplotype pairs H instead of $H \in S(G_i)$. The derivates of w are

$$\frac{\partial}{\partial p_\tau} w = \textstyle\sum_{H\in S(G_i)} 2I(h=\tau) p_{h'}$$

$$\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial p_\tau} w = 2I(h=h'=\tau)$$

$$\frac{\partial}{\partial p_\theta}\frac{\partial}{\partial p_\tau} w = \textstyle\sum_{H\in S(G_i)} 2I(h_j=\tau, h'_j=\theta)$$

Now we can write the components of D and I in terms of derivatives of v, u, and w.

$$\frac{\partial}{\partial \alpha^*} \log L_i = y_i \frac{\frac{\partial}{\partial \alpha^*} v}{v} - \frac{\frac{\partial}{\partial \alpha^*} u}{u}$$

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log L_i = y_i \frac{\frac{\partial}{\partial \boldsymbol{\beta}} v}{v} - \frac{\frac{\partial}{\partial \boldsymbol{\beta}} u}{u}$$

$$\frac{\partial}{\partial \boldsymbol{\gamma}} \log L_i = y_i \frac{\frac{\partial}{\partial \boldsymbol{\gamma}} v}{v} - \frac{\frac{\partial}{\partial \boldsymbol{\gamma}} u}{u}$$

$$\frac{\partial}{\partial p_\tau} \log L_i = (1-y_i) \frac{\frac{\partial}{\partial p_\tau} w}{w} + y_i \frac{\frac{\partial}{\partial p_\tau} v}{v} - \frac{\frac{\partial}{\partial p_\tau} u}{u}$$

$$\frac{\partial}{\partial \boldsymbol{\nu}} \log L_i = y_i \frac{\frac{\partial}{\partial \boldsymbol{\nu}} v}{v} - \frac{\frac{\partial}{\partial \boldsymbol{\nu}} u}{u}$$

$$\frac{\partial}{\partial\alpha^*}\frac{\partial}{\partial\alpha^*}\log L_i = y_i\frac{v(\frac{\partial}{\partial\alpha^*}\frac{\partial}{\partial\alpha^*}v)-(\frac{\partial}{\partial\alpha^*}v)(\frac{\partial}{\partial\alpha^*}v)}{v^2} - \frac{u(\frac{\partial}{\partial\alpha^*}\frac{\partial}{\partial\alpha^*}u)-(\frac{\partial}{\partial\alpha^*}u)(\frac{\partial}{\partial\alpha^*}u)}{u^2}$$

$$\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial}{\partial\alpha^*}\log L_i = y_i\frac{v(\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial}{\partial\alpha^*}v)-(\frac{\partial}{\partial\alpha^*}v)(\frac{\partial}{\partial\boldsymbol{\beta}}v)}{v^2} - \frac{u(\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial}{\partial\alpha^*}u)-(\frac{\partial}{\partial\alpha^*}u)(\frac{\partial}{\partial\boldsymbol{\beta}}u)}{u^2}$$

$$\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\alpha^*}\log L_i = y_i\frac{v(\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\alpha^*}v)-(\frac{\partial}{\partial\alpha^*}v)(\frac{\partial}{\partial\boldsymbol{\gamma}}v)}{v^2} - \frac{u(\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\alpha^*}u)-(\frac{\partial}{\partial\alpha^*}u)(\frac{\partial}{\partial\boldsymbol{\gamma}}u)}{u^2}$$

$$\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\alpha^*}\log L_i = y_i\frac{v(\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\alpha^*}v)-(\frac{\partial}{\partial\alpha^*}v)(\frac{\partial}{\partial p_\tau}v)}{v^2} - \frac{u(\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\alpha^*}u)-(\frac{\partial}{\partial p_\tau}u)(\frac{\partial}{\partial\alpha^*}u)}{u^2}$$

$$\frac{\partial}{\partial\alpha^*}\frac{\partial}{\partial\boldsymbol{\nu}}\log L_i = y_i\frac{v(\frac{\partial}{\partial\alpha^*}\frac{\partial}{\partial\boldsymbol{\nu}}v)-(\frac{\partial}{\partial\boldsymbol{\nu}}v)(\frac{\partial}{\partial\alpha^*}v)}{v^2} - \frac{u(\frac{\partial}{\partial\alpha^*}\frac{\partial}{\partial\boldsymbol{\nu}}u)-(\frac{\partial}{\partial\boldsymbol{\nu}}u)(\frac{\partial}{\partial\alpha^*}u)}{u^2}$$

$$\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial}{\partial\boldsymbol{\beta}}\log L_i = y_i\frac{v(\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial}{\partial\boldsymbol{\beta}}v)-(\frac{\partial}{\partial\boldsymbol{\beta}}v)(\frac{\partial}{\partial\boldsymbol{\beta}}v)}{v^2} - \frac{u(\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial}{\partial\boldsymbol{\beta}}u)-(\frac{\partial}{\partial\boldsymbol{\beta}}u)(\frac{\partial}{\partial\boldsymbol{\beta}}u)}{u^2}$$

$$\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\boldsymbol{\beta}}\log L_i = y_i\frac{v(\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\boldsymbol{\beta}}v)-(\frac{\partial}{\partial\boldsymbol{\beta}}v)(\frac{\partial}{\partial\boldsymbol{\gamma}}v)}{v^2} - \frac{u(\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\boldsymbol{\beta}}u)-(\frac{\partial}{\partial\boldsymbol{\beta}}u)(\frac{\partial}{\partial\boldsymbol{\gamma}}u)}{u^2}$$

$$\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\boldsymbol{\beta}}\log L_i = y_i\frac{v(\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\boldsymbol{\beta}}v)-(\frac{\partial}{\partial\boldsymbol{\beta}}v)(\frac{\partial}{\partial p_\tau}v)}{v^2} - \frac{u(\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\boldsymbol{\beta}}u)-(\frac{\partial}{\partial\boldsymbol{\beta}}u)(\frac{\partial}{\partial p_\tau}u)}{u^2}$$

$$\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial}{\partial\boldsymbol{\nu}}\log L_i = y_i\frac{v(\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial}{\partial\boldsymbol{\nu}}v)-(\frac{\partial}{\partial\boldsymbol{\nu}}v)(\frac{\partial}{\partial\boldsymbol{\beta}}v)}{v^2} - \frac{u(\frac{\partial}{\partial\boldsymbol{\beta}}\frac{\partial}{\partial\boldsymbol{\nu}}u)-(\frac{\partial}{\partial\boldsymbol{\beta}}u)(\frac{\partial}{\partial\boldsymbol{\beta}}u)}{u^2}$$

$$\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\boldsymbol{\gamma}}\log L_i = y_i\frac{v(\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\boldsymbol{\gamma}}v)-(\frac{\partial}{\partial\boldsymbol{\gamma}}v)(\frac{\partial}{\partial\boldsymbol{\gamma}}v)}{v^2} - \frac{u(\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\boldsymbol{\gamma}}u)-(\frac{\partial}{\partial\boldsymbol{\gamma}}u)(\frac{\partial}{\partial\boldsymbol{\gamma}}u)}{u^2}$$

$$\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\boldsymbol{\gamma}}\log L_i = y_i\frac{v(\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\boldsymbol{\gamma}}v)-(\frac{\partial}{\partial\boldsymbol{\gamma}}v)(\frac{\partial}{\partial p_\tau}v)}{v^2} - \frac{u(\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\boldsymbol{\gamma}}u)-(\frac{\partial}{\partial\boldsymbol{\gamma}}u)(\frac{\partial}{\partial p_\tau}u)}{u^2}$$

$$\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\boldsymbol{\nu}}\log L_i = y_i\frac{v(\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\boldsymbol{\nu}}v)-(\frac{\partial}{\partial\boldsymbol{\nu}}v)(\frac{\partial}{\partial\boldsymbol{\gamma}}v)}{v^2} - \frac{u(\frac{\partial}{\partial\boldsymbol{\gamma}}\frac{\partial}{\partial\boldsymbol{\nu}}u)-(\frac{\partial}{\partial\boldsymbol{\nu}}u)(\frac{\partial}{\partial\boldsymbol{\gamma}}u)}{u^2}$$

$$\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial p_\tau}\log L_i = (1-y_i)\frac{w(\frac{\partial^2}{\partial p_\tau^2}w)-(\frac{\partial}{\partial p_\tau}w)(\frac{\partial}{\partial p_\tau}w)}{w^2} + y_i\frac{v(\frac{\partial^2}{\partial p_\tau^2}v)-(\frac{\partial}{\partial p_\tau}v)(\frac{\partial}{\partial p_\tau}v)}{v^2} - \frac{u(\frac{\partial^2}{\partial p_\tau^2}u)-(\frac{\partial}{\partial p_\tau}u)(\frac{\partial}{\partial p_\tau}u)}{u^2}$$

$$\frac{\partial}{\partial p_\theta}\frac{\partial}{\partial p_\tau}\log L_i = (1-y_i)\frac{w(\frac{\partial}{\partial p_\theta}\frac{\partial}{\partial p_\tau}w)-(\frac{\partial}{\partial p_\tau}w)(\frac{\partial}{\partial p_\theta}w)}{w^2} + y_i\frac{v(\frac{\partial}{\partial p_\theta}\frac{\partial}{\partial p_\tau}v)-(\frac{\partial}{\partial p_\tau}v)(\frac{\partial}{\partial p_\theta}v)}{v^2} - \frac{u(\frac{\partial}{\partial p_\theta}\frac{\partial}{\partial p_\tau}u)-(\frac{\partial}{\partial p_\tau}u)(\frac{\partial}{\partial p_\theta}u)}{u^2}$$

$$\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\boldsymbol{\nu}}\log L_i = y_i\frac{v(\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\boldsymbol{\nu}}v)-(\frac{\partial}{\partial\boldsymbol{\nu}}v)(\frac{\partial}{\partial p_\tau}v)}{v^2} - \frac{u(\frac{\partial}{\partial p_\tau}\frac{\partial}{\partial\boldsymbol{\nu}}u)-(\frac{\partial}{\partial\boldsymbol{\nu}}u)(\frac{\partial}{\partial p_\tau}u)}{u^2}$$

Appendix C

Score Statistic Details for Case-only Test for Interaction Effects

The score statistic for the case-only analysis of haplotype-environment interaction effects is

$$S_{\boldsymbol{\nu}} = U_{\boldsymbol{\nu}}^T V_{\boldsymbol{\nu}}^{-1} U_{\boldsymbol{\nu}} \Big|_{\substack{\boldsymbol{\nu}=0 \\ \boldsymbol{\xi}=\tilde{\boldsymbol{\xi}}}}$$

where

$$U_{\boldsymbol{\nu}} = \frac{\partial}{\partial \boldsymbol{\nu}} \log L_{obs}$$

and

$$V_{\boldsymbol{\nu}} = D_{\boldsymbol{\nu\nu}} - I_{\boldsymbol{\nu\xi}} I_{\boldsymbol{\xi\xi}}^{-1} D_{\boldsymbol{\nu\xi}}^T - D_{\boldsymbol{\nu\xi}} I_{\boldsymbol{\xi\xi}}^{-1} I_{\boldsymbol{\nu\xi}}^T + I_{\boldsymbol{\nu\xi}} I_{\boldsymbol{\xi\xi}}^{-1} D_{\boldsymbol{\xi\xi}} I_{\boldsymbol{\xi\xi}}^{-1} I_{\boldsymbol{\nu\xi}}^T,$$

We write the variance-covariance matrix D as

$$D = \begin{pmatrix} D_{\boldsymbol{\nu\nu}} & D_{\boldsymbol{\nu\xi}} \\ D_{\boldsymbol{\nu\xi}}^T & D_{\boldsymbol{\xi\xi}} \end{pmatrix}$$

where

$$D_{\boldsymbol{\nu}\boldsymbol{\xi}} = \begin{pmatrix} D_{\boldsymbol{\nu}\tilde{p}_1} & \dots & D_{\boldsymbol{\nu}\tilde{p}_{L+1}} \end{pmatrix}$$

and

$$D_{\boldsymbol{\xi}\boldsymbol{\xi}} = \begin{pmatrix} D_{\tilde{p}_1\tilde{p}_1} & \dots & D_{\tilde{p}_1\tilde{p}_{L+1}} \\ \vdots & \vdots & \vdots \\ \dots & D_{\tilde{p}_i\tilde{p}_j} & \dots \\ \vdots & \vdots & \vdots \\ D'_{\tilde{p}_1\tilde{p}_{L+1}} & \dots & D_{\tilde{p}_{L+1}\tilde{p}_{L+1}} \end{pmatrix}$$

Let

$s_i(y_i, g_i, \tilde{p}_\tau) = \frac{\partial}{\partial \tilde{p}_\tau} \log L_i$

$s_i(y_i, g_i, \boldsymbol{\nu}) = \frac{\partial}{\partial \boldsymbol{\nu}} \log L_i$.

Then we define the elements of D as

$D_{\tilde{p}_\tau\tilde{p}_\tau} = \sum_{i=1}^n s_i(y_i, g_i, \tilde{p}_\tau) s_i(y_i, g_i, \tilde{p}_\tau)$

$D_{\tilde{p}_\tau\tilde{p}_\theta} = \sum_{i=1}^n s_i(y_i, g_i, \tilde{p}_\tau) s_i(y_i, g_i, \tilde{p}_\theta)$

$D_{\boldsymbol{\nu}\tilde{p}_\tau} = \sum_{i=1}^n s_i(y_i, g_i, \boldsymbol{\nu}) s_i(y_i, g_i, \tilde{p}_\tau)$

$D_{\boldsymbol{\nu}\boldsymbol{\nu}} = \sum_{i=1}^n s_i(y_i, g_i, \boldsymbol{\nu}) s_i(y_i, g_i, \boldsymbol{\nu})^T$.

We can write the observed information matrix I as

$$I = \begin{pmatrix} I_{\boldsymbol{\nu}\boldsymbol{\nu}} & I_{\boldsymbol{\nu}\boldsymbol{\xi}} \\ I^T_{\boldsymbol{\nu}\boldsymbol{\xi}} & I_{\boldsymbol{\xi}\boldsymbol{\xi}} \end{pmatrix}$$

where

$$I_{\boldsymbol{\nu}\boldsymbol{\xi}} = \begin{pmatrix} I_{\boldsymbol{\nu}\tilde{p}_1} & \dots & I_{\boldsymbol{\nu}\tilde{p}_{L+1}} \end{pmatrix}$$

and

$$I_{\boldsymbol{\xi}\boldsymbol{\xi}} = \begin{pmatrix} I_{\tilde{p}_1\tilde{p}_1} & \dots & I_{\tilde{p}_1\tilde{p}_{L+1}} \\ \vdots & \vdots & \vdots \\ \dots & I_{\tilde{p}_i\tilde{p}_j} & \dots \\ \vdots & \vdots & \vdots \\ I'_{\tilde{p}_1\tilde{p}_{L+1}} & \dots & I_{\tilde{p}_{L+1}\tilde{p}_{L+1}} \end{pmatrix}$$

and the elements of I as

$I_{\tilde{p}_\tau\tilde{p}_\tau} = -\frac{\partial}{\partial\tilde{p}_\tau}\frac{\partial}{\partial\tilde{p}_\tau}\log L_i$

$I_{\tilde{p}_\tau\tilde{p}_\theta} = -\frac{\partial}{\partial\tilde{p}_\theta}\frac{\partial}{\partial\tilde{p}_\tau}\log L_i$

$I_{\boldsymbol{\nu}\tilde{p}_\tau} = -\frac{\partial}{\partial\boldsymbol{\nu}}\frac{\partial}{\partial\tilde{p}_\tau}\log L_i.$

To simplify notation, define the quantities v, u, and w for each individual as

$v = \sum_{H\in S(G_i)}\exp(X_{HE}\boldsymbol{\nu})\tilde{p}_h\tilde{p_{h'}}$

$u = \sum_{H}\exp(X_{HE}\boldsymbol{\nu})\tilde{p}_h\tilde{p_{h'}}$

The derivatives of v with respect to the parameters $\boldsymbol{\nu}$ and a specific $\tilde{p}_\tau$ are

$\frac{\partial}{\partial\boldsymbol{\nu}}v = \sum_{H\in S(G_i)}\exp(X_{HE}\boldsymbol{\nu})\tilde{p}_h\tilde{p_{h'}}X_{HE}^T$

$\frac{\partial}{\partial\tilde{p_\tau}}v = \sum_{H\in S(G_i)}\exp(X_{HE}\boldsymbol{\nu})2I(h=\tau)\tilde{p}_{h'}$

$\frac{\partial}{\partial\tilde{p_\theta}}\frac{\partial}{\partial\tilde{p_\tau}}v = \sum_{H\in S(G_i)}2I(h=\tau,h'=\theta)\exp(X_{HE}\boldsymbol{\nu})$

$\frac{\partial}{\partial\tilde{p_\tau}}\frac{\partial}{\partial\boldsymbol{\nu}}v = \sum_{H\in S(G_i)}\exp(X_{HE}\boldsymbol{\nu})2I(h=\tau)\tilde{p_{h'}}X_{HE}^T$

The derivatives of u are the same as those for v, except they are summed over all haplotype pairs H instead of $H \in S(G_i)$.

Now we can write the components of D and I in terms of derivatives of v, u, and w.

$$\frac{\partial}{\partial \tilde{p}_\tau} \log L_i = d_i \frac{\frac{\partial}{\partial \tilde{p}_\tau} v}{v} - d_i \frac{\frac{\partial}{\partial \tilde{p}_\tau} u}{u}$$

$$\frac{\partial}{\partial \boldsymbol{\nu}} \log L_i = d_i \frac{\frac{\partial}{\partial \boldsymbol{\nu}} v}{v} - d_i \frac{\frac{\partial}{\partial \boldsymbol{\nu}} u}{u}$$

$$\frac{\partial}{\partial \tilde{p}_\tau} \frac{\partial}{\partial \tilde{p}_\tau} \log L_i = d_i \frac{v(\frac{\partial^2}{\partial \tilde{p}_\tau^2} v) - (\frac{\partial}{\partial \tilde{p}_\tau} v)(\frac{\partial}{\partial \tilde{p}_\tau} v)}{v^2} - d_i \frac{u(\frac{\partial^2}{\partial \tilde{p}_\tau^2} u) - (\frac{\partial}{\partial \tilde{p}_\tau} u)(\frac{\partial}{\partial \tilde{p}_\tau} u)}{u^2}$$

$$\frac{\partial}{\partial \tilde{p}_\theta} \frac{\partial}{\partial \tilde{p}_\tau} \log L_i = d_i \frac{v(\frac{\partial}{\partial \tilde{p}_\theta} \frac{\partial}{\partial \tilde{p}_\tau} v) - (\frac{\partial}{\partial \tilde{p}_\tau} v)(\frac{\partial}{\partial \tilde{p}_\theta} v)}{v^2} - d_i \frac{u(\frac{\partial}{\partial \tilde{p}_\theta} \frac{\partial}{\partial \tilde{p}_\tau} u) - (\frac{\partial}{\partial \tilde{p}_\tau} u)(\frac{\partial}{\partial \tilde{p}_\theta} u)}{u^2}$$

$$\frac{\partial}{\partial \tilde{p}_\tau} \frac{\partial}{\partial \boldsymbol{\nu}} \log L_i = d_i \frac{v(\frac{\partial}{\partial \tilde{p}_\tau} \frac{\partial}{\partial \boldsymbol{\nu}} v) - (\frac{\partial}{\partial \boldsymbol{\nu}} v)(\frac{\partial}{\partial \tilde{p}_\tau} v)}{v^2} - d_i \frac{u(\frac{\partial}{\partial \tilde{p}_\tau} \frac{\partial}{\partial \boldsymbol{\nu}} u) - (\frac{\partial}{\partial \boldsymbol{\nu}} u)(\frac{\partial}{\partial \tilde{p}_\tau} u)}{u^2}$$