Using Ancestral Information to Search for Quantitative Trait Loci in Genome-wide Association Studies

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Katherine L. Thompson, B.S., B.S., M.S.

Graduate Program in Statistics

The Ohio State University

2013

Dissertation Committee:

Laura Kubatko, Ph.D., Advisor
Radu Herbei, Ph.D.
Shili Lin, Ph.D.
Dennis Pearl, Ph.D.
Abstract

Many quantitative traits, such as blood pressure, are thought to be influenced by specific genes, but are also affected by environmental factors, making the associated genes difficult to identify and locate from genetic data alone. In particular, the application of classical statistical methods to data collected in genome-wide association studies (GWAS) has been especially challenging. Representing information contained in the evolutionary history as a bifurcating tree at each location provides a new way to investigate covariance structures in existing genome-wide association study data. Because the relationships among observations are determined by the tree, many existing statistical methods can be adapted to this new variance-covariance structure.

I propose a novel approach to search for locations along the genome associated with quantitative traits in GWAS data by taking into account the evolutionary history among samples. I evaluate the performance of the new method using simulated data, and find that it outperforms the classical t-test and performs at least as well as other existing methods, with a slight increase in performance in the case of population structure. Application of the methodology to a real data set consisting of high-density lipoprotein cholesterol measurements in mice shows the method performs well for empirical data, as well.
In addition, current statistical methodologies for analyzing GWAS data are especially limited when considering a quantitative trait in the presence of external influences. I expand the basic model to search for associations between a quantitative trait and either a genetic location or an external covariate of interest, and show via a simulation study that the method performs well at determining associations between the quantitative trait and either or both of the genetic and external influences on the quantitative trait.

The advantages of the models presented here are rooted in their flexibility and ready implementation to a wider range of GWAS data sets, including those which contain information about covariates. By pursuing these possible research directions, I can continue to both simulate and analyze data in a wide variety of settings, addressing the current limitations of existing statistical methodology in quantitative trait mapping, especially in the presence of external covariates.
This is dedicated to my family.
Acknowledgements

I would like to thank all of the people who have helped, supported, and guided me throughout my graduate work. In particular, I want extend special thanks to my parents, Rock and Debbie Thompson, to my brothers, Benjamin and Christopher, and to all of my friends for their unending support and patience. Within the Department of Statistics, I would like to thank my candidacy and dissertation committee members, Dr. Radu Herbei, Dr. Shili Lin, and Dr. Dennis Pearl for their time and helpful comments, which have helped guide and improve this work. I would also like to thank the computer support staff for their time and effort in helping me implement these methods, and for allowing me to use their resources extensively. In addition, I want to thank Dr. Elizabeth Stasny for her guidance during my time as a graduate student, especially during my employment search. Lastly, I cannot thank Dr. Laura Kubatko enough for her countless hours, tireless effort, and enthusiastic support while being my advisor. I want to extend my sincere gratitude for her unending patience, advice and guidance both within this research and outside of my graduate work.
Vita

April 1, 1986 .......................... Born - Louisville, Kentucky, USA

2008 ................................. B.S. Mathematics

2008 ................................. B.S. Biology

2010 ................................. M.S. Statistics

2008-present ........................ Graduate Teaching Associate,
                          The Ohio State University.

Fields of Study

Major Field: Statistics
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Vita</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>ix</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xi</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Introduction to Genetics</td>
<td>2</td>
</tr>
<tr>
<td>1.1.1 Modeling the Evolutionary History of Individuals</td>
<td>4</td>
</tr>
<tr>
<td>1.1.2 Ancestral Recombination Graphs</td>
<td>6</td>
</tr>
<tr>
<td>1.2 Quantitative Trait Mapping</td>
<td>10</td>
</tr>
<tr>
<td>1.3 Overview of Dissertation</td>
<td>12</td>
</tr>
<tr>
<td>2. Literature Review</td>
<td>13</td>
</tr>
<tr>
<td>2.1 Quantitative Trait Mapping</td>
<td>13</td>
</tr>
<tr>
<td>2.1.1 Non-tree Based Methods</td>
<td>14</td>
</tr>
<tr>
<td>2.1.2 Tree-based Methods</td>
<td>15</td>
</tr>
<tr>
<td>2.1.3 Ancestral Recombination Graph Estimation</td>
<td>21</td>
</tr>
<tr>
<td>2.1.4 Methods Allowing for External Covariates</td>
<td>26</td>
</tr>
<tr>
<td>2.1.5 Multiple Comparisons Adjustment</td>
<td>27</td>
</tr>
<tr>
<td>2.2 Quantitative Trait Evolution Models</td>
<td>28</td>
</tr>
</tbody>
</table>
### 3. Methods

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Data Analysis</td>
<td>36</td>
</tr>
<tr>
<td>3.1.1 Genetic Model</td>
<td>36</td>
</tr>
<tr>
<td>3.1.2 Modeling External Covariates</td>
<td>49</td>
</tr>
<tr>
<td>3.2 Data Simulation</td>
<td>56</td>
</tr>
<tr>
<td>3.2.1 Simulating Genetic Data</td>
<td>58</td>
</tr>
<tr>
<td>3.2.2 Simulating Quantitative Trait Data</td>
<td>59</td>
</tr>
<tr>
<td>3.2.3 Simulating Quantitative Trait Data with Genetic and Environmental Covariates</td>
<td>61</td>
</tr>
<tr>
<td>3.2.4 Simulation Parameters</td>
<td>62</td>
</tr>
<tr>
<td>3.3 Implementation</td>
<td>64</td>
</tr>
<tr>
<td>3.3.1 Data Analysis</td>
<td>64</td>
</tr>
<tr>
<td>3.3.2 Data Simulation</td>
<td>65</td>
</tr>
</tbody>
</table>

### 4. Results

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Genetic Model</td>
<td>67</td>
</tr>
<tr>
<td>4.1.1 Simulation Studies</td>
<td>67</td>
</tr>
<tr>
<td>4.1.2 Real Data Analysis</td>
<td>76</td>
</tr>
<tr>
<td>4.2 Modeling Environmental Covariates</td>
<td>80</td>
</tr>
<tr>
<td>4.2.1 Simulation Studies</td>
<td>80</td>
</tr>
<tr>
<td>4.3 Summary</td>
<td>96</td>
</tr>
</tbody>
</table>

### 5. Discussions and Conclusions

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1 Genetic Model</td>
<td>99</td>
</tr>
<tr>
<td>5.2 Modeling Environmental Covariates</td>
<td>101</td>
</tr>
<tr>
<td>5.3 Future Directions</td>
<td>102</td>
</tr>
</tbody>
</table>

### Appendices

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Supplementary Results</td>
<td>104</td>
</tr>
</tbody>
</table>

### Bibliography

<table>
<thead>
<tr>
<th></th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>112</td>
</tr>
</tbody>
</table>
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Data example</td>
<td>3</td>
</tr>
<tr>
<td>1.2 Two chromosomes before and after a recombination event</td>
<td>7</td>
</tr>
<tr>
<td>3.1 SNP data matrix example</td>
<td>37</td>
</tr>
<tr>
<td>4.1 Type I error for simulated data sets</td>
<td>68</td>
</tr>
<tr>
<td>4.2 Power and localization distance (bp) for data sets simulated under the additive model</td>
<td>69</td>
</tr>
<tr>
<td>4.3 Power and localization distance (bp) for data sets simulated under the recessive model</td>
<td>70</td>
</tr>
<tr>
<td>4.4 Power and localization distance (bp) for data sets simulated under the dominant model</td>
<td>71</td>
</tr>
<tr>
<td>4.5 Type I error for data sets simulated with population structure</td>
<td>75</td>
</tr>
<tr>
<td>4.6 Power and localization distance for data sets simulated with population structure</td>
<td>76</td>
</tr>
<tr>
<td>4.7 Type I Error of LSS, SNPassoc, and the LRT for simulated data sets with environmental covariates</td>
<td>82</td>
</tr>
<tr>
<td>4.8 Power of LSS, SNPassoc, and the LRT for simulated data sets</td>
<td>83</td>
</tr>
<tr>
<td>4.9 Average localization distance of LSS, SNPassoc, and the LRT for simulated data sets</td>
<td>83</td>
</tr>
</tbody>
</table>
A.1 Standard deviation of localization distances for simulated data sets . 105

A.2 Standard deviation of localization distances for simulated data sets in
the presence of population structure . . . . . . . . . . . . . . . . . . . 106

A.3 Standard deviation of localization distance of LSS, SNPassoc, and the
LRT for simulated data sets . . . . . . . . . . . . . . . . . . . . . . . 106
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Phylogenetic tree at an associated SNP</td>
<td>6</td>
</tr>
<tr>
<td>1.2 Evolutionary history of SNPs along a chromosome</td>
<td>8</td>
</tr>
<tr>
<td>1.3 Example of recombination within a phylogenetic network</td>
<td>9</td>
</tr>
<tr>
<td>2.1 Examples of ARGs</td>
<td>22</td>
</tr>
<tr>
<td>2.2 Phylogeny example</td>
<td>31</td>
</tr>
<tr>
<td>3.1 Example of a local perfect phylogeny</td>
<td>38</td>
</tr>
<tr>
<td>3.2 Example of an MPR of ancestral character states</td>
<td>41</td>
</tr>
<tr>
<td>3.3 Example of a six-taxon tree with branch lengths</td>
<td>44</td>
</tr>
<tr>
<td>3.4 Example of correlation structure defined by a phylogeny</td>
<td>45</td>
</tr>
<tr>
<td>4.1 Power of detection</td>
<td>72</td>
</tr>
<tr>
<td>4.2 Localization distance</td>
<td>73</td>
</tr>
<tr>
<td>4.3 Example of population structure</td>
<td>74</td>
</tr>
<tr>
<td>4.4 Results in the presence of population structure</td>
<td>77</td>
</tr>
<tr>
<td>4.5 Example of detection p-values and localization distances</td>
<td>78</td>
</tr>
<tr>
<td>4.6 Likelihood score statistic plots</td>
<td>79</td>
</tr>
</tbody>
</table>
4.7 Power and localization in covariate analysis

4.8 Example of behavior of LSS across a chromosome

4.9 Example of number of clusters chosen by the LSS across a chromosome

4.10 Example of behavior of estimation of $\beta_1$ across a chromosome

4.11 Example of behavior of baseline estimate across a chromosome

4.12 Example of behavior of within-cluster mean estimates along a chromosome

4.13 Estimates of genetic and environmental variances at the maximally-scored SNP

4.14 Estimated differences in cluster means at the maximally-scored SNP

4.15 Estimates of $\beta_0 + \mu_k$ at the maximally-scored SNP

4.16 Estimates of $\beta_1$ at the maximally-scored SNP
Chapter 1: Introduction

The goal of quantitative trait mapping based on genome-wide association study (GWAS) data is to find locations along the genome that are associated with a set of quantitative traits under study. The biological complexity of the evolutionary history of genes and the environmental factors acting simultaneously on the trait values makes this a challenging task, even with very large data sets. However, distinguishing these two types of associations could aid in researchers’ understanding of the driving influences behind the complex traits under study. This is an important task because finding associations between particular traits and either genetic or external influences could prove useful in the development of preventative tools and treatments for particular complex diseases.

For instance, low-density lipoprotein (LDL) cholesterol levels have been previously investigated in a genome-wide association study [32]. Sandhu et al. (1998) indicate that the causal relationship between LDL cholesterol levels and cardiovascular disease makes identifying ways in which LDL cholesterol levels are regulated useful in the prevention and treatment of cardiovascular disease. This particular study identified a novel location along the genome related to LDL cholesterol levels, which has led other researchers to study this relationship. In fact, it has been determined that the same genomic location identified by [32] is also related to other forms of cardiovascular
disease [27]. In this case, the analysis of GWAS data has allowed researchers to identify genes that play a role in the development of cardiovascular disease, enabling the possibility of improvements in prevention and treatment of cardiovascular disease. In this chapter, a review of genetic concepts and GWAS data is given as background for the description of methods used for quantitative trait mapping in this setting given in Chapter 2.

1.1 Introduction to Genetics

Genetic information is organized into chromosomes. A chromosome is a strand of deoxyribonucleic acid (DNA), which is comprised of a sequence of four nucleotide bases: adenine (A), thymine (T), cytosine (C), and guanine (G). An individual has two copies of each chromosome, one from each parent. The sequence of base pairs present on a particular copy of a chromosome is called a haplotype. Genotypes are the DNA sequences aggregated across chromosomes for each individual. For instance, each row in the second column of Table 1.1 represents a haplotype. Considering two haplotypes from an individual without respect to ordering would produce the individual’s genotype. In Table 1.1 Person 1’s genotype at the DNA sequence site in green could be ‘AC’ or equivalently, ‘CA.’ Thus, the separation of sequences of nucleotidic base pairs into chromosomes is unspecified in genotypic data.

The connection between DNA sequences and characteristics of an organism lies in the function of DNA. Within DNA sequences, particular sets of three bases, called codons, contain the information to code particular amino acids, which are the building blocks for proteins in organisms. Thus, these regions in DNA sequences can affect the characteristics of organisms, so relationships exist between particular base pair
Table 1.1: Data example. In each region of the genome, an individual (first column) has two DNA sequences (one per chromosome, shown in the second column). Many of these locations, shown in black, do not vary across the chromosomes. The locations showing variation, displayed in color, represent SNP sites. The third column shows a binary representation of the SNP data alone. The trait data (fourth column) consist of one trait value per individual, which is assumed to be associated with both chromosomes.

<table>
<thead>
<tr>
<th>Person</th>
<th>DNA Sequences</th>
<th>SNP Data</th>
<th>Trait Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>. . . AACTGGTCCAACGTC . . .</td>
<td>. . . 000 . . .</td>
<td>175.8</td>
</tr>
<tr>
<td>1</td>
<td>. . . AACTGGTCCAACCGTC . . .</td>
<td>. . . 010 . . .</td>
<td>175.8</td>
</tr>
<tr>
<td>2</td>
<td>. . . AACTTGTTCAACATC . . .</td>
<td>. . . 101 . . .</td>
<td>115.6</td>
</tr>
<tr>
<td>2</td>
<td>. . . AACTGGTCCAACCATC . . .</td>
<td>. . . 011 . . .</td>
<td>115.6</td>
</tr>
<tr>
<td>3</td>
<td>. . . AACTTTGTCCAACGTC . . .</td>
<td>. . . 100 . . .</td>
<td>157.3</td>
</tr>
<tr>
<td>3</td>
<td>. . . AACTGGTCCAACATC . . .</td>
<td>. . . 001 . . .</td>
<td>157.3</td>
</tr>
</tbody>
</table>

locations and the traits of an organism. These are the relationships of interest in quantitative trait mapping.

In order to find locations along the genome associated with a quantitative trait, only nucleotides that vary across chromosomes are generally considered. Table 1.1 shows DNA sequences (second column) from six different chromosomes in three individuals (first column). The base pairs in black show no variation across chromosomes, while the colored base pairs do show variation. These sites showing variation are called single nucleotide polymorphisms (SNPs). In this work, each of the two represented states of the SNP is referred to as an allele. By representing the “major” allele (i.e., the allele that occurs with the highest frequency in the population) with a ‘0’ and the “minor” allele (i.e., the allele that occurs with the lowest frequency in the population) with a ‘1’ at each SNP, this genetic data can be represented as a matrix of binary random variables, as in the third column of Table 1.1. In this matrix, each
row corresponds to an observation (a haplotype) and each column corresponds to a SNP. The goal is to find SNPs (columns of this matrix) associated with a quantitative trait, also referred to as a phenotype (example in column four). Because in this case an observation is a haplotype, rather than an individual’s genotype, accurate phasing – separation of genotypes into haplotypes – must be carried out as the first step in an analysis. The phase of sampled chromosomes can be inferred using various statistical techniques, but this task is nontrivial and may lead to inaccuracies in the assigned haplotypes. Alternatively, phasing may be carried out biologically, but this is not typically performed in the laboratory since it is time- and cost-intensive [36]. Typical GWAS data sets include unphased genotypic SNP data and quantitative trait data for many individuals (hundreds to thousands). Some examples of available GWAS data sets are described in Section 1.2.

1.1.1 Modeling the Evolutionary History of Individuals

The two most-commonly used models in the study of the evolutionary history of individuals within a population are the Wright-Fisher model and Moran’s model [37]. An important difference between the models is that the Wright-Fisher model assumes nonoverlapping generations of individuals, while Moran’s model allows generations to overlap. The focus here is on the more commonly used of the two models, the Wright-Fisher model.

Under the Wright-Fisher model, a finite, fixed-size population of $2N$ individuals is considered. At each generation, $2N$ individuals are randomly selected with replacement from the population to serve as the next generation of individuals. Due to the random selection of individuals with replacement, the probability that two of these
individuals have a common ancestor in the preceding generation is $\frac{1}{2N}$. Considering time discretely, the number of generations since two individuals shared a common ancestor follows a Geometric distribution with parameter $\frac{1}{2N}$. By extending this concept and considering time as a continuous random variable, the time (in units of $2N$ generations) that it takes for $i$ individuals to have $(i-1)$ ancestors can be approximated with an exponential random variable with mean $\left(\frac{i}{2}\right)^{-1}$ for a large population size, $N$. This is the underlying result in Kingman’s coalescent [19].

Genetic data can be modeled using this continuous-time coalescent model [15, 37]. The work presented here models the evolutionary history of a set of genes that arises due to neutral mutation - mutations not affecting the reproductive ability nor the survival ability of an individual and occurring independently of both population size and selection. If a gene undergoes a mutation between two time points, then it will produce two descendants. Considering time from present to past, this can be considered to be a coalescent event. A common assumption in modeling DNA sequence data is that each mutation produces a new SNP site with an evolutionary history in the data. Under this assumption, the data are said to follow an infinite-sites model. Under this model, once a DNA base pair has undergone a mutation, it cannot mutate back to its original state.

In the coalescent framework, the evolutionary history at each SNP can be represented as a bifurcating tree, called a phylogenetic tree. Figure 1 shows an example of a phylogenetic tree at a particular SNP. In a phylogenetic tree, each tip represents an observation. Time is shown from past to present moving from left to right across the tree. The root at the far left of the tree shows the most recent common ancestor (MRCA) of all individuals in the tree. The branch lengths represent durations
of time, so that longer branches represent longer amounts of time. Along the tree, wherever two observations share a lineage (collection of branches), they also share a part of their evolutionary history. After a split, or separation from one lineage into two, individuals evolve independently. Each point at which a split is observed in the tree is referred to as an internal node.

Figure 1.1: Phylogenetic tree at an associated SNP. This tree shows the evolutionary history of a SNP for 50 diploid observations of a quantitative trait (colored circles on right). The low values of the trait are blue, and the high values are red. The clustering pattern in the trait values shown by the tree is an indication of the association between the SNP and the trait.

1.1.2 Ancestral Recombination Graphs

At a single locus, the coalescent process describes the genealogical history among sampled individuals in the form of a phylogenetic tree. In the case of GWAS data, however, the goal is to model the complete evolutionary history of a set of DNA
Table 1.2: Two chromosomes before and after a recombination event. The original two separate chromosomes are shown (first column), along with a portion of their DNA sequences (second column), before a recombination event occurs. After recombination, the newly-formed chromosomes are shown in the third column, along with the revised DNA sequences (fourth column). The recombination point is denoted by the red dash (fourth column).

<table>
<thead>
<tr>
<th>Before Recombination Event</th>
<th>After Recombination Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromosome</td>
<td>DNA Sequence</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------</td>
</tr>
<tr>
<td>□□</td>
<td>...CTGGTCCACCGT</td>
</tr>
<tr>
<td>□□</td>
<td>...CTGGTCCAACGT</td>
</tr>
<tr>
<td>□□</td>
<td>...CTGGGTCC-ACCGT</td>
</tr>
<tr>
<td>□□</td>
<td>...CTGGGTCC-AACGT</td>
</tr>
</tbody>
</table>

sequences along an entire chromosome instead of at a particular SNP site. In this case, a single phylogenetic tree cannot be used to model the genealogical history among all individuals for the entire chromosome because two competing processes, coalescence and recombination, are occurring simultaneously along a chromosome.

Biologically, a *recombination event* refers to the event when a pair of chromosomes break and exchange material during cell reproduction. For example, the first column of Table 1.2 shows a pair of chromosomes before a recombination event has occurred. The second column has an excerpt of a DNA sequence along each chromosome. The third column of the table shows the chromosomes after the recombination event, and the corresponding DNA sequences (with the recombination point in red) are in the fourth column. Both the third and fourth columns show the exchange of genetic material that occurred during the recombination event.

When recombination occurs between two genetic sequences and the sequences exchange genetic material at a recombination point, a situation occurs where the portion of the genetic sequence on one side of the recombination point is the same
Figure 1.2: Evolutionary history of SNPs along a chromosome. This is a representation of evolutionary history of each SNP along a chromosome (indicated by blue arrows) as phylogenetic trees. These marginal trees are shown at the six SNPs along the sequence of base pairs from the chromosome. Time moves from present to past from bottom to top along the phylogenetic trees.

as the sequence present before the recombination, while the portion of the sequence on the remaining side is new [38]. When recombination occurs within a phylogenetic network, a single lineage has two ancestors, so that a split in a lineage occurs looking from present to past. This is what removes the ability to represent the genetic history of multiple SNPs with a single phylogenetic tree. For instance, Figure 1.3 shows two examples of such lineages which have two ancestors due to recombination events. In each case, at least one internal node in the tree has two ancestors at the preceding time point, so that the evolutionary history cannot be modeled as a bifurcating tree.

In this case, the genealogical history of a chromosome can be represented by an Ancestral Recombination Graph (ARG), a phylogenetic network representing both recombination and coalescent events [40]. ARGs provide a model that accommodates
Figure 1.3: Example of recombination within a phylogenetic network. This is an example of a phylogenetic network from multiple SNP sites. Within the evolutionary history, two recombination events have occurred.

the fact that while a (true) local tree exists at each site along the chromosome, neighboring trees may be not be the same due to recombination events. These discrepancies among sites in the evolutionary history of a sequence of DNA bases make the local trees incompatible, meaning that the evolutionary history of that particular sequence cannot be represented in one single phylogenetic tree. For instance, in Figure 1.2, a phylogenetic tree exists at each SNP site along the chromosome, but neighboring trees may be incompatible due to the occurrence of recombination events. Here, the underlying ARG is a graphical representation of these six phylogenetic trees. In general, ARGs can be used to represent clusters of incompatible local trees, and they contain the information needed to determine the marginal tree at each SNP along the chromosome [40].

In some cases, a single phylogenetic tree, or local perfect phylogeny, can be built for multiple SNP sites that are compatible. Compatibility means that the evolutionary
history of the set of sites can be represented by a single phylogenetic tree, and it can be determined by the four-gamete test \[14\]. The four-gamete test considers pairs of SNP sites, and examines whether all four possible combinations of SNP states (00, 01, 10, and 11) are observed at the two sites. If so, the sites are incompatible, meaning that either a back mutation (mutation from the current to the ancestral allele) or a recombination event occurred during the evolutionary history of the two sites. If all four possible combinations of the SNP states are observed, a single phylogenetic tree cannot represent the evolutionary history of the two SNPs under the assumption of a single mutation at each SNP \[37\]. SNP sites that are compatible under the four-gamete test can be used together to build a local perfect phylogeny.

Local perfect phylogenies can be used as an estimate of the local marginal phylogeny at a particular SNP site in the presence of mutation and recombination events along the genealogy. ARGs contain the information in neighboring local trees, which is typically more information than in the largest local perfect phylogeny around that site. However, the number of possible tree topologies at each SNP, much less the size of the space of possible ARGs, makes an ARG difficult to estimate from GWAS data. Due to their complexity, ARGs are difficult to model and estimate from genetic data, and methodology has been proposed to estimate ARGs from genetic data using approximation techniques. Descriptions of these techniques are given in Chapter \[2\].

### 1.2 Quantitative Trait Mapping

Quantitative trait mapping refers to the search through a collection of genetic information on a group of individuals with the goal of finding regions of the genome associated with quantitative traits. This search has two distinct goals: detection
and localization. Detection is achieved if any SNP in a certain region is found to be statistically significantly associated with the trait, while localization addresses how close the detected SNP(s) are to the true causative SNP(s). Localization performance is usually measured by the distance between the associated SNP and the true SNP. Since most data sets will include a large number of SNPs, it is very unlikely that any statistical method will pick up a truly causal SNP, but clearly methods that can provide relatively precise localization will be the most useful.

Due to the size of GWAS data, computational cost is a concern in this problem. Current methods of quantitative trait mapping either have straightforward implementations and fail to use the information from the evolutionary relationships among SNPs, or incorporate this information at a computational expense. Ideally, a quantitative trait mapping method will use the information contained in the relationships among SNPs and remain computationally feasible in the case of large data sets.

For quantitative trait and genome-wide SNP data, a data set usually contains genotypes at the SNP locations in the genome, and a phenotypic trait value for each individual. Genotypic data are typically not phased into haplotypes before use. Examples of available mammalian GWAS data include mouse data [43, 29, 42] and human data [2]. In mice, examples of previously-analyzed phenotypes include average daily running distance, white head spot, and cardiovascular traits such as high-density lipoprotein cholesterol level, systolic blood pressure, diastolic blood pressure, and mean arterial pressure. These studies included anywhere from 55 to 288 mice, and analyzed between 44,428 and approximately 7.87 million SNPs. Several of these studies analyze publicly-available mouse data sets [20, 5]. In human GWAS data,
Besenbacher et al. (2009) analyzed height data for 5,000 individuals with approximately 317,000 polymorphic loci [2]. The size of GWAS data sets such as these can often limit the feasibility of implementation of statistical methodology.

1.3 Overview of Dissertation

The goal of this dissertation is to develop a novel method to search for SNPs associated with a quantitative trait which is flexible enough to detect and localize associated SNPs in the presence of external influences on the quantitative trait. I will begin by introducing previous methods used in quantitative trait mapping and provide a literature review of these methods in Chapter 2. Chapter 3 includes a description of the proposed methods of data analysis and data simulation for this setting, and Chapter 4 gives the results of simulation studies and real data analysis. The discussions and conclusions in Chapter 5 provide a synthesis of the results and implications of the proposed techniques in quantitative trait mapping and outline some future research directions.
Chapter 2: Literature Review

In this chapter, a review of the current statistical methodology used in the analysis of quantitative traits is given. These methods include non-tree based and tree-based methods used in quantitative trait mapping and models for quantitative trait evolution.

2.1 Quantitative Trait Mapping

Methods of quantitative trait mapping can be broadly classified into two groups: those that model the shared evolutionary history among sampled individuals, usually in the form of a phylogenetic tree, and those that do not. Further, tree-based methods may or may not use an ARG to model relationships along a chromosome for genetic sequence data. However, since non-tree based methods fail to consider the evolutionary relationships among SNPs, they may have difficulty detecting some associations between SNPs and quantitative traits. This leaves room for improvement in the power of detection of associated SNPs. By using information contained in the relationships among SNPs, tree-based methods gain power in detection and localization. However, this gain in power comes at the expense of an increased computational cost. Several methods in each category are reviewed below.
2.1.1 Non-tree Based Methods

Non-tree based methods used in quantitative trait mapping include methods that analyze each marker independently (e.g. the $t$-test), and those that analyze groups of markers together (e.g., Haplotype Association Mapping [25] and Single Marker Analysis [43]). Recall the data in Table 1.1 in which each SNP in the third column is represented as a binary random variable. The $t$-test simply groups samples according to allele type at each SNP, and uses a two-sided alternative to look for a significant difference in mean trait value between groups. Both Haplotype Association Mapping (HAM) and Single Marker Association (SMA) perform an ANOVA on particular groupings of genotypic samples to assess the significance of the groupings [43]. SMA partitions samples into groups using the genotypic data at each single locus, while HAM groups samples based on sequences in sliding windows of three SNPs [43]. Using these groupings, each method performs an ANOVA on the phenotypic data to test for a difference in the average phenotype among the grouped samples. Since HAM uses a fixed-size sliding window, it cannot avoid the use of regions containing incompatible SNPs during inference [43]. In the simulation studies presented in [29], HAM tended to outperform SMA in most cases with respect to detection of an associated SNP and location of the causative locus, with particularly strong performance in the presence of multiple causative SNPs within a small region.

Since non-tree based methods fail to consider the evolutionary relationships among SNPs, they have difficulty detecting associations between SNPs and quantitative traits, and have a lower power than some tree-based methods. For instance, in Figure
the quantitative trait observations represented by colored dots along the phylo-
geny exhibit a clustering pattern along the tree, providing evidence of the underly-
ning relationship. Performing a classical $t$-test with a two-sided alternative hypothesis

gives an insignificant p-value for this data set. This example demonstrates that the
$t$-test may be deficient because it fails to account for the evolutionary history within
this particular SNP.

2.1.2 Tree-based Methods

By using information contained in the relationships among SNPs, tree-based meth-
ods have the potential to gain power in detection and localization. However, this gain
in power comes at the expense of an increased computational cost. In spite of the
computational issues, many tree-based methods have recently been proposed for this
problem. In the case of discrete trait data, these methods include LATAG [44],
MARGARITA [26], and Blossoc [24]. For quantitative traits, tree-based methods in
common use include TreeQA [29], QBlossoc [2], and HTreeQA [43].

One of the earlier tree-based methods used to study this type of association is
LATAG (implemented in the software TreeLD) [44]. This method approximates the
underlying ARG along a chromosome using a Bayesian approach to sample genealogies
consistent with the genotypic data at each locus in a set of locations along the chromo-
some. Specifically, LATAG uses a Markov chain Monte Carlo (MCMC) algorithm to
sample from the posterior distribution of genealogies at each locus conditional on the
genotypic data. This MCMC technique is performed at particular locations, called
focal points, along the chromosome. At each focal point, a local approximation to
the ARG is built using the markers in the surrounding region, under the assumption
that these genetic markers follow a finite-sites mutation model along the branches of the phylogeny. At each locus, the trees sampled from the posterior distribution are used to compute the posterior probability of the SNP being causative conditional on the phenotypic and genotypic data. Then, LATAG uses the sampled genealogies and their estimated posterior distribution for inference about each focal point. The mode of the estimated posterior distribution can be used to estimate the location of the truly-causative SNP. To assess significance in detection, the following likelihood ratio statistic is calculated:

$$LR = \frac{L_A(\Phi; x, \hat{P}_{alt}, G)}{L_O(\Phi; \hat{P}_0, G)}$$

(2.1)

where $\Phi$ is the trait data, $x$ is a SNP location, and $G$ is the genotypic data, and $\hat{P}_{alt}$ and $\hat{P}_0$ are the vectors of penetrance parameters. To approximate the alternative likelihood, $L_A(\Phi; x, \hat{P}_{alt}, G)$, the likelihood of the phenotypic data conditional on $\hat{P}_{alt}, x$, and each sampled genealogy is calculated, and the sample average of these likelihoods is used. The penetrance parameters describe the probability that a chromosome of a particular genotype leads to an individual with a particular phenotype in the case that $x$ is either associated with the phenotype ($\hat{P}_{alt}$) or not associated with the phenotype ($\hat{P}_0$). When the locus, $x$, is not associated with the phenotype, the penetrances of each genotype are equal. To use $LR$ to assess significance, the numerator and denominator in Equation 2.1 are each maximized with respect to the penetrance parameters in the case of an association or no association between the focal point and the phenotype, respectively. If the location, $x$, is not truly associated with the phenotype, then $2 \ln LR$ is said to asymptotically follow a $\chi^2$ distribution. This distribution is used to calculate a p-value at each location along the chromosome.
to address detection. The flexibility of LATAG to model either a binary or a continuous trait provides a distinct advantage over other proposed methods. However, like other previous methods, LATAG requires phased data for analysis in its current implementation. Additionally, even using the local approximation to estimate the underlying ARG, the Bayesian framework of LATAG is computationally intensive. The largest data set analyzed contained data from 85 SNPs within 220 diploid individuals, and is the upper limit of what the algorithm can feasibly handle, taking approximately 36 hours for analysis. In order to address these computational deficiencies, several non-Bayesian tree-based techniques have been proposed, including QBlossoc [2], TreeQA [29], and HTreeQA [43].

The tree-based methods vary in the way that the phylogenetic information is used in the subsequent analysis. One method of particular interest to the present study is QBlossoc [2]. After local phylogenies are estimated for each SNP using Blossoc’s approach [24] (details are given in Section 3.2.1), QBlossoc uses this information to partition the sampled individuals into some number of clusters, $k$, and calculates a score for each possible set of $k$ clusters defined by the phylogenetic tree. QBlossoc tests two types of scoring to assess each site. The first is an ‘allelic’ score, where each haplotype is analyzed independently. In the second score, chromosomes are paired during analysis so that clusters are further partitioned so that both tips for an individual are observed in the same group. Thus, a group can include a subset of observations from multiple clusters. In either procedure, the score calculated is a penalized likelihood (the penalty is determined by the number of clusters), where the likelihood is a multivariate normal with a different mean in each cluster and an overall shared variance, with zero covariance among observations. The maximum
score over all possible sets of \( k \) clusters defined by the phylogeny is used to assess the significance of each SNP.

Simulation studies using data from either an additive model (where the distribution of the trait is a function of both alleles within an individual at the disease location) or a recessive model (where the trait comes from a different distribution only if both alleles within an individual are mutant alleles) indicate that QBlossoc has a comparable power of detection in comparison to the classical \( t \)-test when a small amount of variation in the quantitative trait is attributed to the truly-causal SNP. As the amount of variation attributed to the truly-causal SNP increases, QBlossoc's power of detection tends to increase. In addition, QBlossoc shows an ability to find high-scoring markers closer to the causal SNP than the single-marker approach, especially when the quantitative trait nucleotide frequency is low. Overall, QBlossoc shows higher accuracy than the single-marker \( t \)-test in the localization of true positives as well as the ranking of regions with disease SNPs, and is efficient in terms of both computing time and memory. Although the clustering technique of QBlossoc accounts for the shared evolutionary history among SNPs, QBlossoc has two weaknesses rooted in its assumptions during the score calculation: namely, QBlossoc assumes both independence and a common variance among the quantitative trait values.

Similarly to the technique used by QBlossoc, TreeQA \cite{29} first estimates the local phylogeny at each SNP using a local perfect phylogeny based on the maximal compatible region. In a simulation study, a maximum of five compatible sites was used to infer the local perfect phylogeny at each SNP \cite{29}. To search for associations between markers and the quantitative trait under study, TreeQA performs a one-way
Analysis of Variance (ANOVA) between the quantitative trait and all possible groupings of samples determined by the edges in the local perfect phylogeny, using the corresponding $F$ test statistic. Significance at each site is assessed using permutation testing. For each grouping, phenotype values are permuted across the tips of the phylogeny, and the corresponding ANOVA statistic is calculated. For each grouping, a significance score is defined:

$$P(G(S')) = \log_{10} \left( \frac{nPerm}{p} \right)$$

(2.2)

where $G(S')$ is the grouping considered, $nPerm$ is the number of permuted data sets, and $p$ is the number of scores at least as large as the $F$-statistic from the observed data set. The P-score for each locus is the maximum value of the significance score in (2.2) over all possible groupings compatible with the local perfect phylogeny. Higher P-scores are evidence of stronger associations between the SNP and quantitative trait.

TreeQA accounts for the information in the phylogenetic history, while maintaining relative computational efficiency. One chromosome containing approximately 40,000 SNPs from 55 diploid individuals is analyzed in approximately 10 minutes \[44\]. Due to its use of haplotypic data in building local perfect phylogenies, TreeQA requires that genotypic data be phased before analysis in order to build local perfect phylogenies at each site.

Current implementations of the tree-based techniques described above require the SNP data be phased into haplotypes prior to analysis. Phasing genotypic data into haplotypes is a nontrivial process, often leading to errors. Even if haplotypes are reconstructed accurately, analyzing haplotypes independently can allow one diploid individual to appear in two different clades in an estimated phylogeny. HTreeQA aims to address this limitation of other tree-based techniques by analyzing unphased
To estimate local phylogenies directly from unphased genotypic data, HTreeQA builds a tristate semi-perfect phylogeny tree at each site. The tristate semi-perfect phylogeny tree is an application of the multistate perfect phylogeny in which the three states are the homozygous wild, homozygous mutant, and heterozygous alleles.

To begin, a binary matrix is constructed, with three columns corresponding to each SNP in the genotypic data. At a particular SNP, each of the three columns serves as an indicator variable for each of the three possible genotypes. This matrix is used to identify compatible regions enabling the construction of a tristate semi-perfect phylogeny. At a location, compatible SNPs are added successively as in the case of local perfect phylogenies. Here, two SNPs are compatible if any two of the three corresponding columns for each SNP in the binary matrix pass the four-gamete test described in Section 1.1.2. A matrix containing the binary columns used to construct the compatibility matrix is then used to build a local perfect phylogeny using the same technique as for binary (or phased SNP) data. The resulting tristate semi-perfect phylogeny is a tree with the same number of tips as diploid individuals (rather than the number of haplotypes) in the study. Then, the same scoring technique used by TreeQA (described above) is applied to the tristate semi-perfect phylogeny estimated at each location along the chromosome, where an observation corresponds to a diploid individual.

In addition to the elimination of a phasing step during data analysis, HTreeQA is able to detect a wider range of genetic effects than other methods since it considers two haplotypes from an individual simultaneously. In the case of an additive model, HTreeQA’s performance in detection of associated SNPs is comparable to
other methods, but it shows an increased performance in the case of traits from a recessive model and an overdominant model (a model in which heterozygous individuals show extreme trait values) while maintaining lower error rates than previous methods in all models considered. In fact, it has recently been noted that the application of a phylogenetic framework to analysis of GWAS data may be beneficial. In terms of computational time, HTreeQA runs faster than previous tree-based techniques.

2.1.3 Ancestral Recombination Graph Estimation

In addition to modeling ARGs using forms of local phylogenies built on compatible regions of SNPs, methods exist to estimate ARGs in a larger capacity. Recall that, in the absence of recombination events in a phylogenetic network, the network can be represented as a bifurcating tree. However, in the presence of a recombination event, at least one internal node in the tree will have two ancestors in addition to its two descendants, so that a perfect phylogeny will not accommodate this case, and a more general procedure is needed to estimate the underlying ARG. Due to the complications that naturally arise in estimating ARGs, previous methods allowing for recombination within phylogenetic networks have been aimed at two special cases of this problem.

The first case is the estimation of ARGs that can be represented as galled trees. In this case, recombination events can be represented within the phylogenetic network at a particular SNP if no internal nodes are shared among the recombination events. For instance, the ARG in Figure 2.1(b) is not a galled tree due to the shared internal node (marked red) between the two recombination events present in the tree. However, Figure 2.1(a) shows a case where two recombination events occur but the
Figure 2.1: Examples of ARGs. (a) shows a genealogy along which two recombination events have occurred. Each of the recombination nodes (shown in blue) has two ancestors and one descendant. Since no nodes are shared within the two recombination events, this is a galled tree. Another genealogy containing two recombination events is shown in (b). This ARG violates the assumption that no recombination nodes are shared among recombination events because the recombination node in red is shared by both recombination events. Thus, the ARG in (b) is not a galled tree and cannot be estimated using the techniques in [38, 11].
two events do not share any nodes. This is an example of a galled tree. Note that in the absence of any recombination events, a galled tree will reduce to a perfect phylogeny, so if a perfect phylogeny exists, a galled tree exists as well.

In [38], an algorithm is given to check a sufficient condition for existence of a galled tree. If a galled tree exists, an algorithm exists to build a phylogenetic network allowing for recombination. During initial implementation of this algorithm, the special case appeared to be too restrictive for estimating ARGs in a disease mapping setting. Gusfield et al. [11] extend the results in [38] and study a more general case of a galled tree called a conflict graph. This method applies when the number of recombination events is sufficiently small (at most half of sites considered) or if most recombinations occurred recently, both of which are believed to hold for human data [11]. See [11] for details.

Instead of focusing on the case of galled trees, other methods exist to estimate the ARG for a data set requiring the fewest number of recombinations, the minimal ARG [35]. This is the goal of the tree scan method [35], which attempts to estimate ARGs by constructing near minimum ARGs using subtree-prune-regraft (SPR) methods, which are often used to build phylogenetic trees. In an ARG, recombination events (cycles) can be considered as a movement of a subtree connected to one node to be connected to another, in other words, an SPR operation performed on a phylogeny. In order to reconstruct the minimal ARG at each site, all binary trees consistent with the observed SNP site are enumerated. The estimated ARG is taken to be the collection of trees across the set of sites that requires the fewest number of SPR operations along the path of trees. This is a parsimony-based method due to its minimization of number of SPR operations across a sequences of sites. Additionally, due to
enumeration of all consistent binary trees, the method is only feasible for a small number of sequences. In fact, the authors considered at most nine sequences in their implementation [35].

In order to address this deficiency and incorporate a probabilistic component into the tree scan method, a Hidden Markov Model approach can be incorporated into modeling the underlying ARG topology [40]. To introduce a probabilistic component, the method assumes that recombination events occur according to a Poisson process within small regions with some constant recombination rate. Then, the estimated ARG is the set of local trees which maximizes the probability of observing a particular set of local trees along the region of sites conditional on the recombination rate. A Hidden Markov Model (HMM) can be implemented in this setting with a Markovian assumption that a local tree at a particular SNP site, conditional on the local tree at its left-neighboring site, is independent of the local trees any farther to its left. However, implementing this model still requires the same enumeration of consistent trees as in the tree scan method, rendering HMM unusable in the case of large GWAS data sets.

REfining Neighboring Trees (RENT) is an ARG estimation technique which addresses this shortcoming of previous methods through its heuristic technique that allows it to scale to large data sets [40]. RENT uses a series of heuristic rules, applied to sites from left to right across a region in order to infer local trees at each site. These rules are ordered from most to least reliable and are used to update the local trees at each site, where reliability was determined empirically. The effect of this on
the resulting inferred local phylogenies is unclear and has not been explored. However, RENT shows a more reasonable runtime than previous techniques, and tends to perform best when markers are dense along the region of interest [40].

MARGARITA also searches for a most plausible ARG for a particular data set using a set of rules [26]. The algorithm begins at the present time and considers time increments in the past to further partition the set of sequences. These partitions are created by grouping or splitting the set of sequences according to one of three evolutionary events: coalescence, mutation, or recombination. If more than one of these evolutionary events is possible at a particular time step, one of the possible events is randomly selected using a set of heuristic rules. This leads to some randomness in the ARG inferred each time the software is implemented, so during inference, MARGARITA averages over multiple inferred ARGs to account for these differences. In addition, since there is no Bayesian component to the approach, the likelihood of the inferred ARG is not measured, so a ‘plausible’ ARG is not explicitly defined. However, it is computationally feasible to implement this heuristic algorithm on large data sets. This algorithm is also easily extended to handle unphased and missing data, providing an advantage over other ARG estimation methods.

In terms of ARG estimation, MARGARITA had a larger percentage of mis-inferred splits in local trees than RENT [40]. In addition, RENT tended to outperform both HMM and the tree scan method in terms of similarity of inferred and true local trees for small data sets, while still performing well when larger numbers of haplotypes (15–30) were simulated [40]. The largest data set for which RENT was tested included 50 sequences, which is much smaller than the number of sequences included in many GWAS data sets.
In addition to its use in ARG estimation, MARGARITA also includes methodology used to make inference about associations between genome-wide SNP data and binary trait data in an ARG framework [26]. In terms of marker-wise association detection, MARGARITA was only outperformed by the $\chi^2$ test when failing to correct for multiplicity, but upon further investigation, MARGARITA had lower error rates than both the $\chi^2$ test and CLADH, another non-ARG based inference tool [6]. MARGARITA consistently outperformed both CLADH and the $\chi^2$ test in terms of localization of the true-causative SNP.

In general, due to the biological complexity of the genetic evolution of SNPs and the compounding environmental factors often simultaneously acting on traits, it is hard to detect the SNPs associated with quantitative traits. Thus, using all information contained in SNP data, including the relationships among SNPs, can aid in finding associations between SNPs and quantitative traits. However, there is a trade-off between the gain from the evolutionary information and the computational tractability of certain tree-based techniques.

2.1.4 Methods Allowing for External Covariates

Previously-proposed methods for the analysis of GWAS SNP data including external covariates during analysis are more limited, and many suggest the use of generalized linear models [33, 41, 9, 34]. In these methods, each SNP is inserted as a covariate into a generalized linear model (GLM). In the case of the models in [9], the genotypic SNP data can be collapsed in any of five models for trait inheritance (codominant, dominant, recessive, overdominant, or log-additive). A p-value is reported from the likelihood ratio test in comparison with the null (intercept-only)
model. In the GLM, González et al. (2007) use a binomial link function for case-control data, and an identity link function for a continuous phenotype. One of the advantages of the model proposed in Chapter 3 is its flexibility to incorporate the covariance structure present from the evolutionary history within a SNP in order to address the limited development of statistical methodology to address quantitative trait mapping in the presence of external covariates.

2.1.5 Multiple Comparisons Adjustment

In quantitative trait mapping, data sets include many SNPs, and thus detection analyses include a large number of association tests. This makes multiple testing errors and controlling the family-wise error rate of each study a concern. The simplest approach taken to adjust for the multiple tests performed at each SNP along a chromosome is to apply a Bonferroni correction, in which the significance level is divided by the number of tests performed, and decisions are made by comparing the p-value of each test to the adjusted significance level. However, due to the correlation present among test statistics at nearby sites, this approach is overly conservative to use in GWAS analysis [44]. QBlossoc simulates a null distribution empirically and assesses significance based on a cutoff determined by the empirical distribution [2]. However, this requires the knowledge of the particular distribution of data under the null hypothesis.

Most methods use permutation testing to control the overall error rate in GWAS analysis [43, 29, 44]. In this case, phenotype values are permuted across individuals, and an appropriate test statistic is calculated for each permuted data set. The p-value
for the entire chromosome is taken to be the proportion of test statistics that are as
or more extreme than the test statistic computed based on the observed data.

2.2 Quantitative Trait Evolution Models

During biological studies of a quantitative trait, stochastic processes are often used
to model the evolution of a quantitative trait along a phylogeny [4, 8, 7]. This presents
an avenue to connect biological quantitative trait evolution models with quantitative
trait models in GWAS data using a stochastic process along a marginal phylogenetic
tree at a particular SNP.

Many of these models for quantitative trait evolution along a phylogeny are based
on Brownian Motion processes. Brownian Motion (BM) was first proposed as a
stochastic process by Robert Brown in the early 1800’s [8]. For an individual with a
trait following a Brownian Motion model, the underlying stochastic process for the
change in the trait over some time interval has a mean of 0 and a constant variance
per unit time. Also, the change in the trait is independent over any small intervals of
time. Since BM is a Gaussian process, the change in trait over a small interval in time,
\textit{dt}, follows a Normal distribution with mean zero and a variance proportional to \textit{dt}.
This stochastic process is nowhere differentiable, meaning many inference techniques
used in classical statistics no longer apply.

It was not until 1964 that Edwards and Cavalli-Sforza applied Brownian Motion
as a model for quantitative traits along a phylogeny [7, 8]. For a particular individual,
the model treats the changes in trait value over small time increments of length \textit{dt} as
independent draws from a normal distribution with mean 0 and variance \( \sigma^2 \textit{dt} \). For
each individual, the process can be represented by the stochastic differential equation,

\[ dY_i(t) = \sigma dB_i(t) \]  \hspace{1cm} (2.3)

This model specifies that the standard deviation of the process per unit time, \( \sigma \), be positive. In the models discussed in this section, values of \( dB_i(t) \) represent “white noise,” or independent, identically distributed random variables following a mean-zero normal distribution with variance \( \sigma^2 dt \).  

Applying the model along a phylogenetic tree is straightforward. Whenever two individuals share a history, they share the path of the stochastic process as well. After two lineages split, their evolutionary trajectories are independent. Thus, for a sample of \( n \) individuals with a trait evolving along a phylogeny with initial value \( y_0 \), at time \( t \), \( Y(t) = (Y_1(t), Y_2(t), \ldots, Y_n(t)) \sim N(y_0 \mathbf{1}, \sigma^2 \mathbf{T}) \), where \( \mathbf{T} \) is the matrix of times of shared evolutionary history for each pair of individuals. Thus, the covariance in trait values between two individuals at time \( t \) is proportional to the sum of the lengths of branches shared by the two individuals. The maximum likelihood estimate (MLE) of the parameter \( \sigma \) maximizes the multivariate normal likelihood when the branch lengths along the phylogeny are known. By applying the BM model to a biological trait, the analyst implicitly assumes that the trait evolves according to a neutral model, where no selection operates on the trait. For many biological traits, this is not a realistic assumption \[4\].

Instead of a BM process, an Ornstein-Uhlenbeck (OU) process can be used to model the selection of a trait toward a fixed optimum value, \( \theta \), along a phylogeny \[12\]. For the \( i^{th} \) individual, the following stochastic differential equation represents
the process by which the trait evolves,

\[ dY_i(t) = \alpha (\theta - Y_i(t)) \, dt + \sigma dB_i(t) \tag{2.4} \]

The deterministic term in this equation, \( \alpha (\theta - Y_i(t)) \, dt \), models the selection of a trait toward the optimum value. Notice that \( \alpha \) is the strength of selection, and the larger \( \alpha \) is, the faster selection toward the optimum will occur. The rate at which the trait moves toward the optimum is also proportional to the distance between the trait and the optimum. By setting \( \alpha = 0 \), the OU model reduces to the BM model in Equation 2.3. The random component of the OU process, \( \sigma dB_i(t) \), is the same as in the BM process, with \( \sigma \), the diffusion coefficient, representing the standard deviation of the “random noise” of the process.

Since the OU process is a Gaussian process, maximum likelihood estimates for parameters in the OU model can be calculated by maximizing the corresponding multivariate normal likelihood at time \( t \). Since the likelihood is a nonlinear function with respect to \( \alpha \), a numerical optimization technique is required to calculate the MLE of \( \alpha \), \( \hat{\alpha} \). However, once \( \hat{\alpha} \) has been determined, the corresponding maximum likelihood estimates of \( \theta \) and \( \sigma \) are directly calculable.

Hansen [12] extends the OU model to allow the optimal trait value, \( \theta_i(t) \), to vary across individuals, \( i \), and time, \( t \), as in Equation 2.5, adding some flexibility to the model,

\[ dY_i(t) = \alpha (\theta_i(t) - Y_i(t)) \, dt + \sigma dB_i(t) \tag{2.5} \]

However, since the optimum is now a function of both the lineage and time, its estimation is not as straightforward, especially for the most general case when no restrictions are placed on \( \theta_i(t) \). This model can be restricted to particular cases
needed for biological application. By restricting the optimum value to be constant across each epoch, or branch directly connecting two internal nodes in a phylogeny (as in Figure 2.2), parameter estimation is more straightforward than in the general case [12].

To reduce the number of parameters in the model, Butler and King (2004) restrict this model even further, to the case when only $r$ different optima, or selective regimes, exist along the phylogeny [4]. These optima are denoted $\theta = (\theta_0, \theta_1, \ldots, \theta_r)'$. It is noted that, wherever two individuals share a lineage along the phylogeny, they will necessarily share the same selective regime as well. Using the multivariate normal likelihood of the parameters $\theta, \alpha,$ and $\sigma,$ given the selective regimes and phylogeny, the MLEs of these parameters can be determined. The MLE for $\alpha$ is determined numerically, while, as a function of $\alpha$, the MLEs for $\theta$ and $\sigma$ can be calculated exactly.

The OU model along the phylogeny can be extended once more to allow the strength of selection, $\alpha$, and the standard deviation of the random component, $\sigma$, to
vary over time for each observation \[1\]. This creates more flexible models which are
applicable to more complicated biological scenarios. These models allow the strength
of selection and standard deviation of the stochastic process per unit time to vary
over the epochs of the tree. Under the Generalized Hansen Model, the change in
a quantitative trait over a lineage can be represented by the stochastic differential
equation,

\[
dY_i(t) = \alpha_i(t) (\theta_i(t) - Y_i(t)) \, dt + \sigma_i(t) dB_i(t).
\] (2.6)

In order to estimate the parameters of the Generalized Hansen Model, the likelihood
can be maximized using numerical optimization.

A more flexible model than the OU model allows the optimum to evolve according
to a stochastic process, as in the two layer OU-BM model \[13\]. In this case, the
quantitative trait value along each lineage can be represented as in (2.5), but the
optimum trait value (also modeled along each lineage), is a function of a collection
of predictor variables, \(x\). Since the predictor variables also each follow a stochastic
process, this yields a two-layer model referred to as an evolutionary regression model,

\[
dy_t = \alpha (\theta_t(x) - y_t) \, dt + \sigma_y dB^y_t
\] (2.7)

\[
dx_t = \sigma_x dB^x_t.
\] (2.8)

where \(t\) represents time and \(x\) represents a predictor variable that evolves via a BM
process as in Equation (2.7) along the phylogenetic tree. This model is referred to as
an evolutionary regression model since the optimum trait value is specified to be a
linear function of a predictor variable:

\[
\theta(x) = b_0 + b_1 x.
\] (2.9)
The subscript $i$, corresponding to the individual, is suppressed from Equations 2.7–2.9 for simplicity. Parameter estimation in this setting is more complicated than in the settings of previous models due to the presence of a second stochastic process. Since the emphasis of this model is on regression, the authors focus on estimating functions of the process parameters and focus their inference on the parameters of interest.

Similar to the setup above, Jhuweng and Maroulas [17] use a two layer OU-OU model in a phylogenetic regression setting. The setup is similar to the OU-BM model in (2.7), but the SDE for the predictor variable, $x$, follows an OU process. Here, $\theta$ is taken to be a linear function of $x$, and can be represented by the SDE,

$$
\begin{align*}
    dy_t &= -\alpha_1 (y_t - \theta_t(x)) \, dt + \sigma_y dB^y_t \\
    dx_t &= -\alpha_2 (x_t - \gamma) \, dt + \sigma_x dB^x_t \\
    \theta_t &= b_0 + b_1 x_t; \gamma = 0 \Rightarrow \theta_t &= -\alpha_3 \theta_t dt + \sigma_{\theta} dB^\theta_t
\end{align*}
$$

Although in the general case, the optimal value for $x$ is $\gamma$, the authors assume $\gamma = 0$ in their derivations. Since the goal here is still phylogenetic regression, the solution to a system of two coupled SDEs is of interest. Parameter estimation can still be completed via an iterative procedure. In order to find the first and second moments of the trait value and trait optimum at time $t$, Itô’s formula can be used to find the system of SDEs for the moments of the random variables. Then, expectations of the SDEs are taken, leaving a system of ordinary differential equations which can be solved explicitly for the moments of interest.

Although the developments in methodology to model the evolution of a quantitative trait along a phylogeny are quite extensive, the models used here will be based on the OU process with a time-varying optimum presented in Equation 2.5.
The methodology developed beyond the considered process gives many avenues to study a quantitative trait according to a variety of models of evolution specific to the quantitative trait.
Chapter 3: Methods

Since the goal of quantitative trait mapping is to search for SNPs associated with a quantitative trait, both detection and localization are considered. In particular, I begin by developing a method for searching for causal SNPs based on the estimation of a local phylogeny at each SNP and a subsequent calculation of a score associated with this phylogeny. Then this method is extended to include external covariates during analysis. This extension of the model allows the simultaneous search for SNPs and/or external covariates associated with a quantitative trait. The proposed analysis techniques include significance assessment based on permutation tests.

In order to examine the performance of the newly-proposed and previously-proposed methods, I use a novel data simulation technique where the location of the SNP truly associated with the quantitative trait is known (if one exists). These data analysis and simulation methods combine ideas from stochastic processes and phylogenetics, providing an innovative avenue to address this problem. The details of the method of data analysis are given in Section 3.1, and the simulation technique is given in Section 3.2. Section 3.3 describes the implementation of the methods in Sections 3.1 and 3.2.
3.1 Data Analysis

3.1.1 Genetic Model

Let $\Theta$ represent a phylogenetic tree, including both the topology and branch lengths. The proposed method of analysis involves the use of the estimated phylogeny at each SNP, so the first step in the proposed method is to estimate a local phylogeny at each SNP. This is done using Blossoc’s approach [24], which assumes the SNP data can be modeled using the infinite-sites model. First, a region of compatible SNPs around the site of interest is determined. Then the SNPs in this compatible region are used to reconstruct a local perfect phylogeny for the considered SNP site.

The compatible region for a considered SNP is a region surrounding this marker that contains compatible SNPs. Two SNPs are compatible if they pass the four-gamete test described in Section 1.1.2 and a region of SNPs is compatible if all combinations of SNPs pass the four-gamete test. To build a compatible region at each SNP, the algorithm considers markers successively, in order of proximity to the considered SNP. Once a SNP either to the left or right is incompatible, the algorithm continues considering SNPs only on the other side (right or left, respectively) of the considered marker until no more SNPs are compatible with the region. After the compatible region has been determined for a particular SNP, each SNP in the region is used to define a single split of lineages in the tree, where the split separates individuals with a ‘0’ allele and a ‘1’ allele at the compatible SNP considered. By defining tree edges in this way, a local perfect phylogeny is built for each SNP site. In this phylogeny, the usual assumptions described in Section 1.1.1 still hold.

In large SNP data sets, many recombinations may have occurred, resulting in compatible regions that contain very few SNPs. In this case, using only SNPs in
Table 3.1: SNP data matrix example. These example SNP data contain 6 observations for 3 individuals, with SNP states observed at five SNP sites (A–E). By the four-gamete test, the compatible region for SNP site B is 3 SNPs in width (namely, across sites A–C).

The compatible region surrounding the site of interest may not fully resolve the tree topology, leaving the local perfect phylogeny with many multifurcations. Blossoc’s tree estimation technique requires the use of a minimum number of SNPs in each region determining a local tree topology. This allows the algorithm to more fully resolve estimated local tree topologies in large data sets. However, incompatible SNPs cannot define single new splits in the tree lineages, so as many splits as necessary are added to the tree to separate individuals with the wild type and mutant type alleles at the incompatible SNP. This is referred to as a local near-perfect phylogeny.

An example SNP data set for six observations (three individuals) is shown in Table 3.1. Here, the largest compatible region determined for SNP site B by the four-gamete test is 3 SNPs in width (namely, across SNPs A–C). This means that the local perfect phylogeny, shown in Figure 3.1(a), is left with multifurcations. The letters along the branches of the local topology show which SNP site was used to make that particular split in the tree.
Figure 3.1: Example of a local perfect phylogeny. For SNP site B in Table 3.1, (a) shows the local perfect phylogeny built from the compatible region of SNPs, SNPs A–C. Notice that by including only three compatible SNPs, the tree is not fully resolved. (b) shows the near-perfect local phylogeny for SNP site B built by including an incompatible SNP site, D, in addition to SNP sites in the compatible region. Site D defines two splits in the estimated tree topology. Including an incompatible site allows the estimation of a fully-resolved topology for the local phylogeny. In both trees, the letters placed at the nodes of the tree represent the sites used for particular splits in the tree.

By forcing four SNPs to be used to build the tree, a near-perfect local phylogeny is produced, as in Figure 3.1(b). Notice that by using SNP site D to form two edges (even though it is incompatible with site B by the four-gamete test), the estimated topology becomes fully resolved. The approach Blossoc uses to build local tree topologies on data sets with many incompatible SNP sites allows the estimation of fully-resolved local topologies in large data sets. This is the approach I use to estimate a local tree topology at each SNP site.

Once the phylogeny is constructed, the branch lengths of each local phylogeny are estimated using a modification of the Rogers-Swofford method in [30], which yields an approximation to the maximum likelihood estimates of the branch lengths.
Specifically, an estimate of the branch length is obtained from a genetic distance estimate under an appropriate model, such as the Jukes-Cantor model [18].

I modify this method to handle SNP data as follows. The distance equation is derived under the M2 model, a two-state Markov model for discrete character data, which is a specific case of the more general Mk model described in [21]. The Mk model is a generalization of the Jukes-Cantor model to \( k \) states. Under the M2 model with 0-1 states, the probability of beginning in state \( i \) and being in state \( j \) after a particular time increment, \( t \), is:

\[
P_{ij}(t) = \frac{1}{2} - \frac{1}{2}e^{-2\alpha t}
\]  

(3.1)

By estimating the transition probability with the observed proportion of differences in SNP sites, \( \hat{p} \), across a branch and defining the branch length estimate \( \hat{d} = \alpha t \), it follows that

\[
\hat{p} = \frac{1}{2} - \frac{1}{2}e^{-2\hat{d}}
\]

(3.2)

\[
\Leftrightarrow \hat{d} = -\frac{1}{2} \ln(1 - 2\hat{p})
\]

(3.3)

Notice that the branch length estimate, \( \hat{d} \), increases as the proportion of differing SNPs between two nodes increases, as expected. This yields an estimate of each branch length based on the estimated distance between a pair of internal nodes in the tree. The distance between a pair of nodes in the tree, \( \hat{p} \), is calculated to be the proportion of the sites in the sequence that differ between the ancestral character states at each internal node across SNPs. If \( \hat{p} = 0 \) for a branch, then I set \( \hat{p} = \frac{1}{\text{number of SNPs}} \). If \( \hat{p} > 1/2 \) for a branch, then I set \( \hat{p} \) to the largest proportion of SNPs still less than 1/2. In this analysis method, the branch length estimates are used to give a broad-scale view of the variance-covariance structure of the data, so having
exact estimates is not of particular interest to this study. This means that even though the Mk model used assumes allele frequencies are equal at equilibrium, the method yields branch length estimates which behave intuitively when considered as functions of the proportion of differing SNPs between the ends of the branch.

Because SNP states are generally unknown at the internal nodes of the tree, a most parsimonious reconstruction (MPR) of ancestral character states is used to calculate \( \hat{p} \) for each set of internal nodes in the tree. An MPR is an assignment of SNPs to the internal nodes of the tree so that the total number of mutations along the phylogeny is minimized [28]. The algorithm in [28] is used to find an MPR of the SNP state at the internal nodes of the phylogeny. This algorithm uses a “two-pass” approach to build MPRs, first moving up the tree from tips to root, then down the tree from root to tips.

During the first pass over the tree from tips to root, a characteristic interval is determined at each node and tip of the tree. At the tips of the tree, this interval is merely an interval with both endpoints equal to the observed SNP state. At each internal node, the set of numbers comprised of the endpoints of the characteristic intervals of both children (direct ancestors) of the node is considered. The characteristic interval at each node is the median interval of this set. The median interval is a closed interval whose endpoints are the middle two numbers of the endpoints of a set of closed intervals. Finding the characteristic intervals at each node leaves a characteristic interval map on the tree. For example, in the tree in Figure 3.2(a), a phylogeny is given with the observed SNP states at the tips of the tree. Next to each node and tip in the tree, the corresponding characteristic intervals are given. For the node shown in red, the characteristic intervals of the child nodes (shown in blue) are
[0,1] and [1,1], respectively. This means the characteristic interval at the red node is the median interval of \{[0, 1], [1, 1]\}, which is [1,1].

Figure 3.2: Example of MPR of ancestral character states. (a) shows the characteristic intervals at all nodes and tips in the tree, according to the algorithm in [28]. At the tips of the tree, the characteristic interval is the closed interval containing the observed SNP states. At the internal nodes of the tree, a tips-to-root pass is taken, and the characteristic intervals are the median intervals of the children of each node. (b) gives the character states which could yield an MPR of ancestral states along the tree. (c) gives one possible MPR of the internal SNP states for this phylogeny. See text for details regarding the determination of these character states.

After the first pass has been completed, a second pass from the root to the tips of the tree is taken to determine the possible character states at the internal nodes which form an MPR along the topology. To determine the minimum value a character state at a particular node can take, the median interval of the set containing three closed intervals is determined. The three intervals are: the left endpoint of the characteristic interval of the parent node and both characteristic intervals of the child nodes. The minimum of the middle two numbers of this is taken to be the smallest possible
character state. The analogous procedure is performed to find the upper endpoint of the possible ancestral character state at a particular node. If these two bounds are the same at all internal nodes, a unique MPR exists for the phylogeny.

For example, consider the node shown in red in Figure 3.2(b). Here, the parent node (shown in green) has the interval $[0,0]$ from the second pass of the algorithm. The closed interval containing only the smaller bound of this interval is $[0,0]$. The characteristic intervals at the two blue “child” nodes (shown in Figure 3.2(a)) are $[0,1]$ and $[1,1]$. This means that the set of intervals considered is $\{[0,0],[0,1],[1,1]\}$. The median interval of this set is $[0,1]$, and thus the minimum ancestral SNP state that could produce an MPR is 0. Using the analogous procedure for the upper endpoint, the set of intervals considered is $\{[0,0],[0,1],[1,1]\}$, which has median interval $[0,1]$, and the maximum of this interval is 1. This leaves us with the interval $[0,1]$ as possible ancestral SNP states at this node in the MPR of the tree. Since not all bounds of the character states in Figure 3.2(b) are unique, multiple MPRs of the ancestral character states might exist. The SNP states at the internal nodes in Figure 3.2(c) represent one possible MPR of ancestral character states for this particular phylogeny.

In the case that multiple MPRs of ancestral SNP states may exist, only certain combinations of these non-unique states will yield a truly MPR on the tree. Whenever both allele types are included in the MPR set at a particular node, the reconstruction containing the wild type allele is used. At times, this choice may not give a truly most parsimonious reconstruction, but biologically, mutations at SNP sites are expected to occur less often than the wild type allele. This choice will reduce the number of back mutations estimated in the ancestral history of each SNP. This type of reconstruction
will be referred to as a nearly most parsimonious reconstruction, and an example is given in Figure 3.2(c).

Once the nearly MPR of ancestral character states has been obtained, the proportion of differing sites between each set of internal nodes in the considered phylogeny can be obtained. Then the branch length estimates can be calculated using Equation 3.3. With each marginal tree topology, all SNPs are used to estimate each branch length. This ensures as much information as possible about each branch length is included, even though SNPs outside the compatible region may have evolved along a different, but related, phylogeny. Ideally, the estimates, $\hat{p}$, would be constrained to be the proportion of differing SNPs within regions not showing recombination. However, the branch length estimates require enough sequence data to find a reasonable estimate, $\hat{p}$, and so the entire sequences are used in this estimation. For each marginal tree, a most parsimonious set of ancestral character states is calculated for each SNP based on the given tree topology.

For each estimated local phylogeny, the branch lengths are used to estimate the variance-covariance matrix of the tree, $V(\Theta)$, as shown in the example in Figure 3.3. The covariance between two quantitative trait values is defined to be proportional to the time of shared evolutionary history between those two observations. Given the local estimated phylogeny at the SNP of interest, the quantitative trait data are assumed to follow a multivariate normal distribution, with covariance structure determined by the local phylogeny.

For phylogenies with a large number of tips, as in the case of GWAS data, this variance-covariance matrix will be very complex with many entries close to zero and
Figure 3.3: Example of a six-taxon tree with branch lengths. The overall tree, \( \Theta \), along with its variance-covariance matrix, \( V(\Theta) \), is shown in (a). The corresponding clustered tree along with the variance-covariance matrices for \( k = 2 \) and \( k = 3 \) clusters are in (b) and (c), respectively.

Many entries taking similar values due to the underlying coalescent model. For example, Figure 3.4(a) shows a surface plot of the correlation structure defined by the phylogeny in Figure 1.1. Many of the correlations among individuals are very near zero. Thus, to account for the most important portion of this evolutionary history, the proposed method uses a clustering technique, considering the estimated phylogeny as a group of clusters with a broad-scale variance-covariance structure. Figures 3.4(b)-(d) show the correlation structures as determined by the clustered tree for 2, 3, and 4 clusters, respectively. Notice that for as few as three or four clusters, most of the correlation structure in the data is preserved.

To accomplish this clustering, at each SNP site, the tree can be partitioned into \( k \) clusters using only the earliest \((k - 1)\) edges in the tree. This provides a computational
Figure 3.4: Example of correlation structure defined by a phylogeny. (a) shows the correlation structure imposed by the phylogeny in Figure 1.1. (b)-(d) show the correlation structures of the corresponding clustered trees for 2, 3, and 4 clusters, respectively. The largest correlations among individuals appear to be preserved in the three- and four-cluster trees. Details of clustering are given in the text.
advantage over QBlossoc \cite{2}, which considers all possible clusters of size $k$ consistent with the estimated phylogeny. I choose to use the earliest $(k - 1)$ edges of the tree in order to capture the broad-scale evolutionary relationships among observations. In the case that one of the $k$ clusters contains few observations, then considering more than $k$ clusters improves the refinement of the variance-covariance structure. However, rather than using the earliest $(k - 1)$ edges to partition the tree, other approaches could be considered in this case. An example of our clustering via only the earliest $(k - 1)$ edges in the phylogeny is shown in Figure 3.3. For a fixed partition of the tree into $k$ clusters, a matrix, $D$, is defined with elements:

$$D_{ij} = \begin{cases} 
1, & \text{if observation } i \text{ falls in cluster } j \\
0, & \text{otherwise}
\end{cases}$$

(3.4)

for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, k$, where $n$ is the number of observations (for diploid individuals, this is twice the number of individuals in the study). Then the trait data, $Y = (Y_1, Y_2, \ldots, Y_n)$ are such that:

$$Y \sim N(D\mu, \sigma^2V)$$

(3.5)

Here, as in QBlossoc, each cluster has its own mean, denoted $\mu = (\mu_1, \mu_2, \ldots, \mu_k)$. However, instead of assuming independence, the variance-covariance matrix of the tree, $\sigma^2V = \sigma^2V(\Theta)$, allows for covariance structure to be present among the quantitative trait observations.

Using the distribution in Equation 3.5, the maximum likelihood estimates of the parameters are straightforward to calculate by maximizing this likelihood (or the
equivalent log-likelihood) function:

\[ L(\mu, \sigma|V, D, y) = \frac{1}{(2\pi)^{n/2}\sigma^2V^{1/2}} \exp \left\{ -\frac{(y - D\mu)^T V^{-1}(y - D\mu)}{2\sigma^2} \right\} \] (3.6)

\[ \ln L(\mu, \sigma|V, D, y) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln(|\sigma^2V|)
- \frac{(y - D\mu)^T V^{-1}(y - D\mu)}{2\sigma^2} \] (3.7)

\[ = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln(|V|) - \frac{n}{2} \ln(\sigma^2)
- \frac{(y - D\mu)^T V^{-1}(y - D\mu)}{2\sigma^2} \] (3.8)

By taking the first partial derivative with respect to each parameter, this system of equations is obtained,

\[ \frac{\partial \ln L(\mu, \sigma|V, D, y)}{\partial \mu} = 2D^T V^{-1} - 2D^T V^{-1} D\hat{\mu} \] (3.9)

\[ \frac{\partial \ln L(\mu, \sigma|V, D, y)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{(y - D\hat{\mu})^T V^{-1}(y - D\hat{\mu})}{2\hat{\sigma}^4} \] (3.10)

Setting these partial derivatives equal to zero, this system of equations is obtained,

\[ 0 = 2D^T V^{-1} y - 2D^T V^{-1} D\hat{\mu} \] (3.11)

\[ 0 = -\frac{n}{2\hat{\sigma}^2} + \frac{(y - D\hat{\mu})^T V^{-1}(y - D\hat{\mu})}{2\hat{\sigma}^4} \] (3.12)

The solutions obtained, \( \hat{\mu} \) and \( \hat{\sigma}^2 \), are extrema of the log-likelihood in Equation 3.8

\[ \hat{\mu} = (D^T V^{-1} D)^{-1} D^T V^{-1} y \] (3.13)

\[ \hat{\sigma}^2 = \frac{(y - D\hat{\mu})^T V^{-1}(y - D\hat{\mu})}{n} \] (3.14)

By considering the determinant of the Hessian and the second partial derivative of the likelihood with respect to \( \mu \) at \( \mu = \hat{\mu} \) and with respect to \( \sigma \) at \( \sigma^2 = \hat{\sigma}^2 \), it can be seen that these solutions are the maximum likelihood estimates of \( \mu \) and \( \sigma^2 \), respectively, for the likelihood conditional on the entire data set, shown in Equation 3.8.
Hypothesis testing is carried out using a likelihood framework. In particular, a penalized likelihood similar to that proposed by \cite{2} is used. The Likelihood Score Statistic (LSS) is defined to be

$$LSS = \max_{k \in \{2, \ldots, 2n-2\}} \{2 \ln L(\hat{\mu}, \hat{\sigma}^2 | Y, \Theta, V) - k \ln(n)\}. \quad (3.15)$$

To calculate the LSS, the maximum likelihood is penalized by subtracting a penalty as in the Bayesian Information Criterion (BIC). Calculation of the likelihood involves estimation of \((k + 1)\) parameters, including the mean trait value in each cluster and the variance, \(\sigma^2\). The BIC criterion penalizes for \(k\) of these parameters. In practice, a local tree is scored according to (3.15) at each SNP, for varying numbers of clusters, \(k = 2, \ldots, k_{\text{max}}\), and the resulting tree score is the maximum score over the number of clusters. I choose \(k_{\text{max}} < 2n - 2\) because accounting for the broad-scale evolutionary relationships requires fewer than \(2n - 2\) clusters, and this will decrease the number of computations required by the method. Results are checked to ensure that \(k_{\text{max}}\) is chosen to be larger than the \(k\) which maximizes the penalized likelihood score at each SNP.

After the score in Equation (3.15) is calculated for the phylogenetic tree at each site along a chromosome, permutation testing can be used to evaluate significance. The null hypothesis is that there is no association between the SNP and quantitative trait under study. At each SNP, any evidence that the SNP is associated with the quantitative trait under study is of interest, so this is the alternative hypothesis. To perform hypothesis testing at each SNP, the observed trait values among the tips of the estimated phylogenetic tree are shuffled to obtain permuted data sets, and the LSS is calculated for each permuted data set. The p-value for detection at each locus is the proportion of data sets scoring higher than the observed data set at each.
particular locus. To address localization, the location of the maximally-scored locus is used as an estimate of the location of the truly-associated SNP.

The goal of detection is to detect the presence of at least one significant SNP in a region that truly contains a SNP linked to the quantitative trait. One method of assessing this significance is to consider the significance of the maximum test statistic along each chromosome, and requires a careful consideration of multiple comparisons. The same permutation testing technique as in [29, 43] is used to control the overall error rate in detection analysis. In this case, for each permuted data set, LSS is calculated at each SNP, and the maximum value of LSS across the considered chromosome is recorded. Then, the p-value is taken to be the proportion of permuted data sets with a maximum LSS value larger than the maximum value of LSS for the observed data. By using permutation testing, significance of the entire chromosome can be assessed without applying the overly-conservative Bonferroni correction.

### 3.1.2 Modeling External Covariates

In many cases, quantitative traits are affected by both a genetic component and external conditions. The model given in Section 3.1.1 can be extended to the case where the effects of both a genetic component and external covariates are of interest. Here, an external (or environmental) covariate refers to any covariate related to an individual’s environment which is unrelated to the individual’s genetics.

If the trait follows an additive model, the quantitative trait values can be considered as the sum of a genetic component \( Y_g \) and an environmental component \( Y_e \). In this case, the observed trait value is the sum of these two components, \( Y \equiv Y_g + Y_e \). This additive trait structure is the same as that assumed in the general mixed model
in classical genetics (see [23]), where the trait is the sum of a genetic and environmental component. Further, each component is assumed to follow a multivariate normal distribution, so that the mean of the trait is the sum of the means of each component, and the analogous is true for the variance-covariance matrix of the trait. In one case of this model, the environmental component has a variance-covariance matrix proportional to the identity matrix, and the genetic component of the trait has a variance-covariance structure determined by genetic relationships (for example, genetic relationships from pedigree data) [23]. The same idea will be used in this method, but the genetic relationship matrix will be determined by the evolutionary history represented by the phylogenetic tree at a particular SNP.

The genetic component of the trait is modeled as in Section 3.2.1. This means that at a SNP site, for an evolutionary history represented by a phylogenetic tree, Θ, the genetic components of the trait data, \( Y_g \), are assumed to follow a multivariate normal distribution along tree Θ:

\[
Y_g \sim N(D\mu, \sigma^2 V). \tag{3.16}
\]

Under the assumptions that the environmental component of the trait follows a multivariate normal distribution, where the mean for the \( i \)th observation can be modeled as a linear combination of \( r + 1 \) fixed covariate values, \( X_{i0}, X_{i1}, X_{i2}, \ldots, X_{ir} \), where \( X_{i0} = 1 \) for all \( i = 1, \ldots, n \). Then, the environmental component of the trait data, \( Y_e \), follows a multivariate normal distribution:

\[
Y_e \sim N(X\beta, \nu^2 I) \tag{3.17}
\]

where \( X_{ij} \) is the value of the \( j \)th covariate for observation \( i \) for \( i = 1, 2, \ldots, n \) and \( j = 1, 2, \ldots, r \), \( X_{i0} = 1 \) for \( i = 1, 2, \ldots, n \), and the parameters \( \beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_r) \)
give the linear combination of covariates which determine the mean of the multivariate normal distribution. Notice that, here, an assumption of independence of the environmental component values across alleles is made. Although this may not be valid in real data analysis, this model was implemented initially to test performance of LSS using a simple computational technique. The performance of LSS appears to be robust to this assumption, but certainly the next model extension should apply to the more general case when the environmental component of a trait may not be independent across alleles, but is independent across diploid individuals.

Using the currently-developed model, under the assumptions of independence and normality of the genetic and environmental components in the additive model, the trait data, \( Y = (Y_1, Y_2, \ldots, Y_n) \), are also assumed to follow a multivariate normal distribution:

\[
Y \sim N(D\mu + X\beta, \sigma^2V + \nu^2I) \quad (3.18)
\]

By defining \( W = [D X], \gamma^T = (\mu^T, \beta^T) \), and \( R \equiv R(\sigma, \nu, V(\Theta)) = \sigma^2V + \nu^2I \), then:

\[
Y \sim N(W\gamma, R(\sigma, \nu, V)) \quad (3.19)
\]

This yields the following likelihood and log-likelihood for the trait data:

\[
L(\gamma, \sigma, \nu|V, W, y) = \frac{1}{(2\pi)^{n/2}|R|^{1/2}} \exp \left\{ -\frac{(y - W\gamma)^TR^{-1}(y - W\gamma)}{2} \right\} \quad (3.20)
\]

\[
\ln L(\gamma, \sigma, \nu|V, W, y) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln(|R|)
- \frac{(y - W\gamma)^TR^{-1}(y - W\gamma)}{2} \quad (3.21)
\]

Detection and localization analyses are carried out in the same likelihood framework as under the genetic-only model, so it is necessary to maximize the penalized likelihood with respect to the parameters \( \gamma, \sigma, \) and \( \nu \). By taking the first partial
derivative of the log-likelihood in Equation 3.21 with respect to each parameter, this
system of equations is obtained,

\[
\frac{\partial \ln L(\gamma, \sigma, \nu | V, X, y)}{\partial \gamma} = 2W^T R^{-1} y - \frac{2W^T R^{-1} y}{2}
\]

(3.22)

\[
\frac{\partial \ln L(\gamma, \sigma, \nu | V, X, y)}{\partial \sigma^2} = -\frac{\partial |R|/\partial \sigma^2}{2|\hat{R}|} - \frac{(y - W\gamma)^T \frac{\partial R^{-1}}{\partial \sigma^2} (y - W\gamma)}{2}
\]

(3.23)

\[
\frac{\partial \ln L(\gamma, \sigma, \nu | V, X, y)}{\partial \nu^2} = -\frac{\partial |R|/\partial \nu^2}{2|\hat{R}|} - \frac{(y - W\gamma)^T \frac{\partial R^{-1}}{\partial \nu^2} (y - W\gamma)}{2}
\]

(3.24)

Setting these partial derivatives equal to zero, this system of equations is obtained,

\[
0 = W^T \hat{R}^{-1} y - W^T \hat{R}^{-1} W\gamma
\]

(3.25)

\[
0 = \frac{1}{|\hat{R}|} \begin{bmatrix} \frac{\partial |R|}{\partial \sigma^2} \end{bmatrix}_{\nu = \hat{\nu}, \sigma = \hat{\sigma}} + \begin{bmatrix} (y - W\hat{\gamma})^T \frac{\partial R^{-1}}{\partial \sigma^2} (y - W\hat{\gamma}) \end{bmatrix}_{\nu = \hat{\nu}, \sigma = \hat{\sigma}}
\]

(3.26)

\[
0 = \frac{1}{|\hat{R}|} \begin{bmatrix} \frac{\partial |R|}{\partial \nu^2} \end{bmatrix}_{\nu = \hat{\nu}, \sigma = \hat{\sigma}} + \begin{bmatrix} (y - W\hat{\gamma})^T \frac{\partial R^{-1}}{\partial \nu^2} (y - W\hat{\gamma}) \end{bmatrix}_{\nu = \hat{\nu}, \sigma = \hat{\sigma}}
\]

(3.27)

where \( \hat{R} \equiv R(\hat{\sigma}, \hat{\nu}, V) \).

Once the MLE’s of \( \sigma^2 \) and \( \nu^2 \), and thus \( \hat{R} \) have been determined, Equation 3.25 can be solved to show that the MLE of \( \gamma \) is

\[
\hat{\gamma} = (W^T \hat{R}^{-1} W)^{-1} W^T \hat{R}^{-1} y.
\]

(3.28)

Since Equations 3.26 and 3.27 cannot be solved analytically, the likelihood must be maximized numerically using software.

When maximizing the log-likelihood in Equation 3.21 notice that \( (W^T R^{-1} W) \) could be a singular matrix, since summing the columns of \( D \) produces an \( n \times 1 \) vector of ones, and the first column of \( X \), by nature of the linear regression model, is also a vector of ones. To avoid computational singularity, the mean of the distribution can be reparameterized during implementation of the numerical maximization of the
log-likelihood,

\[ W \gamma = D \mu + X \beta = [D \ X] \begin{bmatrix} \mu \\ \beta \end{bmatrix} \]  
(3.29)

\[ = D \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix} + X \beta \]  
(3.30)

\[ = D \begin{bmatrix} \mu_1 - \mu_k \\ \mu_2 - \mu_k \\ \vdots \\ \mu_k - \mu_k \end{bmatrix} + X \beta + D \begin{bmatrix} \mu_k \\ \mu_k \\ \vdots \\ \mu_k \end{bmatrix} \]  
(3.31)

\[ = D \begin{bmatrix} \mu_1 - \mu_k \\ \mu_2 - \mu_k \\ \vdots \\ \mu_k - \mu_k \end{bmatrix} + \begin{bmatrix} 1 & x_1 & \cdots & x_r \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{bmatrix} + D \mathbf{1} \mu_k \]  
(3.32)

where \( x_p \) is the \((p + 1)\)th column of \( X \). Since the sum of the columns of \( D \) is a vector of ones, \( D \mathbf{1} = 1 \). Then,

\[ W \gamma = D \begin{bmatrix} \mu_1 - \mu_k \\ \mu_2 - \mu_k \\ \vdots \\ \mu_k - \mu_k \end{bmatrix} + \begin{bmatrix} 1 & x_1 & \cdots & x_r \end{bmatrix} \begin{bmatrix} \beta_0 + \mu_k \\ \beta_1 \\ \vdots \\ \beta_r \end{bmatrix} \]  
(3.33)

\[ = D \begin{bmatrix} \mu_1 - \mu_k \\ \mu_2 - \mu_k \\ \vdots \\ \mu_k - \mu_k \end{bmatrix} + X \begin{bmatrix} \beta_0 + \mu_k \\ \beta_1 \\ \vdots \\ \beta_r \end{bmatrix} \]  
(3.34)

\[ \equiv X \beta^* + D^* \mu^*, \]  
(3.35)

where \( \beta^* = \begin{bmatrix} \beta_0 + \mu_k \\ \beta_1 \\ \vdots \\ \beta_r \end{bmatrix} \) and \( \mu^* = \begin{bmatrix} \mu_1 - \mu_k \\ \mu_2 - \mu_k \\ \vdots \\ \mu_k - 1 - \mu_k \end{bmatrix} \),

and \( D^* = \) the first \( k - 1 \) columns of \( D \)

\[ \Rightarrow W \gamma \equiv W^* \gamma^*, \text{ where } W^* = [X \ D^*] \text{ and } \gamma^* = \begin{bmatrix} \beta^* \\ \mu^* \end{bmatrix}. \]  
(3.36)
Thus, maximizing the log-likelihood with mean \( W^*\gamma^* \) is equivalent to maximizing the log-likelihood with mean \( W\gamma \) (in Equation 3.19), and will avoid issues with the singularity of the matrix \( (W^TR^{-1}W) \) during estimation.

Once the likelihood has been maximized for a particular number of clusters, \( k \), the proposed LSS statistic is defined exactly as in Section 3.2.1, with the log-likelihood defined as in Equation 3.21:

\[
LSS = \max_{k \in \{2, \ldots, 2n-2\}} \{2 \ln L(\hat{\gamma}, \hat{\sigma}^2, \hat{\nu}^2 | Y, \Theta, V) - (k + r) \ln(n) \}. \tag{3.37}
\]

To calculate the LSS, the maximum likelihood is penalized by subtracting a penalty as in the Bayesian Information Criterion (BIC). Calculation of the likelihood involves estimation of \((k+r+1)\) parameters, including the difference in mean trait value in each cluster relative to the \(k^{th}\) cluster, the variance of the genetic component, the baseline mean, the regression parameters, and the variance of the environmental component. The LSS penalizes for \((k+r)\) of these parameters. In practice, a local tree is scored according to (3.37) at each SNP, for varying numbers of clusters, \( k = 2, \ldots, k_{\text{max}} \), and the resulting tree score is the maximum score over the number of clusters. Again, \( k_{\text{max}} \) is chosen so that \( k_{\text{max}} < 2n - 2 \) for the same reasons mentioned in Section 3.2.1.

After the score in Equation 3.37 is calculated for the phylogenetic tree at each locus along a chromosome, permutation testing can be used to evaluate significance according to the technique in Section 3.2.1. The null hypothesis is that there is no association between the SNP and quantitative trait nor between the environmental covariate and quantitative trait. At each SNP, any evidence that the SNP or the covariate is associated with the quantitative trait under study is of interest, so this is the alternative hypothesis. To perform hypothesis testing at each SNP, the observed trait values among the tips of the estimated phylogenetic tree are shuffled to obtain
permuted data sets, and the LSS is calculated for each permuted data set. The covariate values are fixed, and thus not permuted when determining the permuted data sets. The p-value for detection at each locus is the proportion of data sets scoring higher than the observed data set at each particular locus. To address localization, the distance (in DNA base pairs) between the maximally-scored locus and the disease locus is calculated.

In this case, the statistic aims to detect the presence of at least one significant SNP in a region that truly contains a SNP linked to the quantitative trait or a relationship between the external covariate and the quantitative trait. It should be noted that, in the case of a quantitative trait related to only the SNP and not the environmental covariate, letting \( \nu = 0 \) and \( \beta = 0 \), the distribution given in Equation 3.18 reduces to the correct distribution given in Equation 3.5 from the genetic-only model in Section 3.2.1. In the analogous case if the quantitative trait is only related to the environmental covariate, by letting \( \sigma = 0 \) and \( \mu = 0 \), the distribution given in Equation 3.18 reduces to that of the true multiple linear regression model.

Since the relationship between the trait and the covariate and/or a SNP is tested at each location along the chromosome, and significance at any SNP defines a relationship, this is equivalent to considering the significance of the maximum test statistic along each chromosome, and requires a careful consideration of multiple comparisons. The same permutation testing technique as in [29, 43] is used to control the overall error rate in detection analysis. In this case, for each permuted data set, LSS is calculated at each SNP, and the maximum value of LSS across the chromosome is recorded. Then, the p-value is taken to be the proportion of permuted data sets with an LSS value larger than the maximum value of LSS for the observed data. By using
permutation testing, significance of the entire chromosome can be assessed without applying the overly-conservative Bonferroni correction.

### 3.2 Data Simulation

To assess the performance of the proposed methods and make comparisons with the existing methods, data sets are needed with known genetic and/or environmental signals. To ensure the permutation testing controls the Type I Error of all three approaches, data sets where there is no known signal are required. Thus, a simulation technique to generate data in the above cases is proposed. This technique is more flexible than the additive model presented in [22], and also allows the simulation of data related to both a truly-causative SNP site and an environmental covariate.

To simulate quantitative trait data with only a genetic signal under the assumption of random mating within the population, the steps followed are:

1. Simulate SNP data in an ARG framework using the software ms [16].

2. Choose one SNP location as the “truly-causative” SNP.

3. Simulate quantitative trait data for each observation along the local phylogenetic tree at the truly-causative SNP using a two-target OU process. This requires knowledge of the ancestral SNP states at each node of the local phylogeny.

4. Pair tips of the tree randomly to create diploid individuals, and determine the trait value for each diploid individual under the assumption of a codominant, recessive, or dominant trait.
To simulate data from the null distribution, with no association between the SNP and the quantitative trait, Steps 1 through 4 are implemented, with this modification. In Steps 2 and 3, a SNP unrelated to the chromosome, along with its evolutionary history, is used to simulate the null quantitative trait.

To simulate quantitative trait data with both a genetic signal and relation to an environmental covariate, the steps followed are:

1. Simulate the genetic component of the trait using Steps 1–4 above under the codominant trait model.

2. Simulate the environmental component of the trait by:
   
   (a) Choosing covariate values randomly over some interval.

   (b) Use a regression equation to determine the environmental component value for each diploid individual.

3. Pair the environmental and genetic components randomly for the diploid individuals.

4. Use a weighted average of the genetic and environmental components of the trait to determine each diploid individual’s observed quantitative trait value.

Under the null hypothesis, the quantitative trait is related to neither a SNP nor the environmental covariate. For this case, a quantitative trait should be simulated so that it is not related to a SNP nor to the environmental covariate under study. Thus, the simulation of genetic data during Step 1 is modified as in the genetic-only model. In addition, in Step 2(a), the covariate used in the simulation of the environmental
component is unrelated to the covariate under study. More detail about each step is given below.

### 3.2.1 Simulating Genetic Data

To assess the performance of the proposed likelihood technique, simulated data sets are used. This provides a setting where the presence and location of the SNP truly associated with the quantitative trait is known. In this simulation study, SNP data are simulated for 100 replicate data sets from a diploid population using the program \texttt{ms} (without selection) \cite{16}.

The simulation algorithm used in \texttt{ms} assumes the infinite-sites model of mutation within each SNP, and simulates the evolutionary history within each SNP under the coalescent approximation to the Wright-Fisher model. First, a genealogy is generated randomly for a chromosome. Then, mutations are placed randomly along the chromosome, providing SNP sites. Within a genealogy in which a mutation occurred, the number of mutations along a genealogy follows a Poisson distribution with a mean equal to the product of the mutation rate and the total length of the genealogy. Recombination is simulated under a uniform finite-sites model, where the number of sites on the chromosome and recombination rate are specified by the user.

Each data set consists of the SNP data corresponding to one chromosome for each individual. For each simulated replicate, a single DNA base pair location is randomly chosen to be associated with the trait. This choice of “disease” locus is restricted so that the minor allele frequency is between 10\% and 30\%. If a simulated data set does not contain a SNP with this minor allele frequency, a new set of SNP data are simulated.
3.2.2 Simulating Quantitative Trait Data

For each haplotype, quantitative data are simulated along the phylogenetic tree at the disease locus according to a generalized version of the Ornstein-Uhlenbeck (OU) process described by [12],

\[ dY_i(t) = \alpha (\delta - Y_i(t)) \, dt + \sigma_Y dB_i(t) \]  

(3.38)

where \( Y_i(t) \) is the quantitative trait value for the \( i^{th} \) lineage at time \( t \), \( \delta \) is the target trait value, \( \alpha \) is the strength of selection toward the target value, \( \sigma_Y \) is the standard deviation of the process per unit time, and \( dB_i(t) \) represents a Brownian Motion process for lineage \( i \), so that values of \( dB_i(t) \) for small time increments, \( dt \), are independent, identically distributed random variables from a normal distribution with mean zero and variance \( dt \). Thus, the OU process is a mean-reverting process with a deterministic component, \( \alpha (\delta - Y_i(t)) \, dt \), modeling the selection of a trait toward the optimum target value, and a stochastic component, \( \sigma_Y dB_i(t) \), providing the “random noise” for the process. Notice that the deterministic portion of this process implies that the movement of the trait toward the target is proportional to the distance between the trait and the target value, \( \delta \).

When this process is applied in the setting of phylogenetics, the stochastic process gives the same value during the time when the evolutionary history is shared for any two observations. However, after two lineages split, their trait values evolve independently from one another. This implies that before a split, two observations are perfectly correlated, while after the split, they evolve in an uncorrelated manner.

For this study, a more flexible form of the Ornstein-Uhlenbeck process is used, the Generalized Hansen model [12]. This allows the trait to evolve toward a non-constant
optimum (target) as follows,

\[ dY_i(t) = \alpha (\delta_i(t) - Y_i(t)) \, dt + \sigma_Y dB_i(t) \]  
(3.39)

\[ \delta_i(t) = \begin{cases} 
\delta_1, & \text{if } S_i(t) = 0 \\
\delta_2, & \text{if } S_i(t) = 1 
\end{cases} \]  
(3.40)

where \( S_i(t) \) is the SNP state for the \( i^{th} \) observation at time \( t \). Thus, the trait is simulated according to this stochastic process with the target trait value determined solely by the SNP state at any time in the evolutionary history at that SNP. For the SNP associated with the quantitative trait, the target value for a particular epoch along the phylogeny is solely determined by the SNP state at the start of the epoch.

As in previous models [4], the assumption of a constant target value along each epoch of the tree limits the number of parameters in the model and is an appropriate way to model quantitative trait data evolving toward target values based on the SNP state throughout the evolutionary history of the SNP.

Using the Generalized Hansen Model allows the simulation of a quantitative trait value for each haplotype that has both a (deterministic) genetic component, determined by the SNP, and a stochastic component. This process imposes an evolutionary history of the quantitative trait which can be portrayed by the phylogenetic tree at the disease locus, and allows the two haplotypes of a diploid individual to evolve independently along the phylogeny at the disease locus. Since chromosomes are paired to make individuals independent of the trait, this is intuitive since the pairing of chromosomes to make individuals should be unrelated to the quantitative trait.

In order to simulate data for each individual, or diplotype, based on the haplotypic data, three models are used. In the codominant model, the trait value for each diplotype is the average simulated trait value across the two copies of the trait for each individual at the disease location. For the recessive and dominant models, any
diploid individual showing the same state in both alleles of the SNP takes a trait value equal to the average trait value across the two copies of the trait for each individual at the disease location. In the dominant model, individuals showing two different alleles at the disease SNP take the trait value associated with the minor allele, while in the recessive model, individuals showing two different alleles at the disease SNP take the trait value associated with the major allele.

3.2.3 Simulating Quantitative Trait Data with Genetic and Environmental Covariates

This simulation technique allows the simulation of trait data affected by both an environmental covariate and a genetic influence; affected by an environmental covariate but no genetic influence; affected by a genetic influence but no environmental covariate; or affected by neither an environmental covariate nor a genetic influence.

In order to simulate data under this additive model, a quantitative trait is simulated with both a genetic component, denoted $Y_g$, and an environmental component, denoted $Y_e$, where the trait is defined to be $Y = Y_g + Y_e$. Here, a genetic effect is defined to be $Y_g = \rho Z_g$, for some parameter $\rho$, $0 \leq \rho \leq 1$, where $Z_g$ is simulated using the technique described in Section 3.2.2. In order to simulate the environmental component, $Y_e$, first observations, $Z_e$, are simulated using the formula $Z_e = X\eta + \epsilon$, where $\epsilon$ is a collection of random variables from a $N(0, \tau^2)$ distribution. Here, the first column of $X$ is a vector of ones, and the remaining covariate values for the diploid individuals are drawn uniformly over a bounded space (bounds are given in Section 3.2.4). Parameters associated with the simulated environmental component of the trait include $\tau^2$, the variance, and $\eta$, the vector of regression coefficients. Lastly, $Y_e$ is defined to be $(1 - \rho)Z_e$. 

61
Thus, the trait has been simulated using the following method:

1. Simulate $Z_g$ according to the OU process in Section 3.2.2.

2. Simulate $Z_e$ using the formula $Z_e = X\eta + \epsilon$, where $\epsilon$ is a collection of random variables from a $N(0, \tau^2)$ distribution.

3. Calculate the trait values per observation, $Y$, using $Y = \rho Z_g + (1 - \rho) Z_e$, $0 \leq \rho \leq 1$.

Thus, by defining $Y_g = \rho Z_g$ and $Y_e = (1 - \rho) Z_e$, a trait is simulated under the assumptions in Section 3.2.2, where:

$$\beta = (1 - \rho)\eta$$  \hspace{1cm} (3.41)

$$\nu^2 = (1 - \rho)^2\tau^2$$  \hspace{1cm} (3.42)

The flexibility of this model is advantageous to the study. By writing the trait value as a weighted average of the genetic and environmental covariate components initially simulated, the proportion of signal in the trait due to the genetic and the environmental components can be controlled. This can allow a thorough study of how the LSS performs in these situations. By setting $\rho = 0$, a trait with only an environmental component is simulated, and by setting $\rho = 1$, a trait with only a genetic component is simulated.

### 3.2.4 Simulation Parameters

During the simulation studies, the SNP data are simulated using these parameters in ms: the diploid population size is $N_0 = 20,000$, the neutral mutation rate for each DNA base pair is $\mu = 2.0 \times 10^{-10}$, the rate of recombination per generation per
DNA base pair is $\nu = 10^{-8}$, and each simulated chromosome is 1,000,000 base pairs long. During simulation of the quantitative trait with a solely genetic influence, the parameters which vary are the strength of selection, $\alpha$, and the standard deviation of the quantitative trait per unit time, $\sigma_Y$. The two target trait values considered are $\delta_1 = 80$ and $\delta_2 = 100$. The choice to investigate these two target values is based on exploratory analysis of cardiovascular data in mice [42], which seem to have similar trait values.

In the simulation studies which include environmental covariates, the SNP data are simulated using the same parameters in $\textit{ms}$. During simulation of the genetic component of the trait, parameters set include the strength of selection, $\alpha = 10$, the standard deviation of the quantitative trait per unit time, $\sigma_Y = 20$, and the two target trait values, $\delta_1 = 80$ and $\delta_2 = 100$. In the environmental component, the trait has a baseline mean, and one covariate value is simulated per individual (so that $r = 1$), from a Uniform(25, 35) distribution. Define the covariate matrix:

$$X = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$$

(3.43)

where $n$ is the number of observations (twice the number of individuals in the study). There are $n/2$ unique values in the set of covariate observations, $X_1, X_2, \ldots, X_n$, since each individual has two observations at each SNP location. Each individual’s covariate value is replicated in the covariate matrix since the number of observations is twice the number of individuals in the study. Then, the parameters used in simulating the environmental component of the trait are $\eta = (\eta_0, \eta_1)^T = (10, 2.5)^T$. In this study, $\tau$, the standard deviation of the environmental component of the trait, is varied, and $Z_e = X\eta + \epsilon$ where $\epsilon$ is a set of independent observations from a normal distribution
with mean zero and variance $\tau^2$. Lastly, $\rho$, the parameter describing the weight of the genetic and environmental components of the trait, is varied, and the quantitative trait value is defined to be $Y = \rho Z_g + (1 - \rho) Z_e$.

### 3.3 Implementation

#### 3.3.1 Data Analysis

To analyze the data, local tree topologies were estimated using the SNP positions and data using the `blossoc.trees` command in the software Blossoc. Setting the `--unphased` option to 0 in `blossoc.trees` specified that the software should analyze the data as phased data. Blossoc was installed and run on a Virtual Machine using the Ubuntu Hardy Heron platform.

Once the tree topologies were estimated, R was used to estimate branch lengths on the local tree topologies. The MPR of ancestral character states required was computed using the `MPR()` function in the package ape. The proportions of SNPs differing at the ends of each branch in the topology were calculated using the `dist.nodes()` function in the package geiger. Since the branch lengths produced can yield a tree with root-to-tips lengths which differ among observations, the external branches of the tree were cut to ensure the root-to-tip length for each tree was the same for all tips. The “cut” distance was taken to be the smallest root-to-tip length that allowed all internal nodes to remain in the tree. No matter how the tree is cut, some of the branch lengths will be altered in the process of ensuring that the root-to-tip length of the tree is the same for all tips of the tree. However, since the altered branch lengths will always be those at the end of the tree, this will not alter the among-cluster covariances among observations once the tree is clustered. In addition, according to
the coalescent model, these branch lengths are expected to be shorter than the edges occurring earlier in the tree. This left us with an estimated local phylogeny (both a topology and branch lengths) at each SNP site.

At each SNP site, the local phylogeny was simplified to contain some number of clusters of observations, $k$. To convert the estimated tree to a tree with $k$ clusters, the earliest $(k - 1)$ nodes in the tree were kept, and the remaining nodes discarded, leaving a multifurcating tree with $k$ clusters. Lastly, the penalized likelihood score was calculated for the site with $k$ clusters. After repeating this process for values of $k \epsilon \{2, \ldots, k_{\text{max}}\}$, LSS was taken to be the maximum penalized likelihood score over the values of $k$.

### 3.3.2 Data Simulation

#### Simulating Genetic Data and a Truly-Causative SNP

To simulate SNP data along a chromosome, \texttt{ms} was used with parameters given in the above sections. Using \texttt{R}, a SNP site from those produced by \texttt{ms} showing a mutant allele frequency between 10% and 30% was chosen randomly to be the “disease” location. The phylogenetic tree closest to the “disease” location was taken to be the disease tree.

#### Simulating Trait Data Associated with the Truly-Causative SNP

Simulating the two-target OU process in \texttt{R} along the disease tree requires the SNP state at each internal node of the tree. The \texttt{MPR()} function also in the \texttt{ape} package was used to determine these states. The \texttt{MPR()} function determines the possible states forming an MPR of the ancestral character states along the phylogeny using
the algorithm in [28]. When either state (‘0’ or ‘1’) was possible at an internal node, the wild type allele was chosen.

The \texttt{rTraitCont()} function from the \texttt{ape} package was used to simulate phenotypic data using a two-target OU process. This function requires the phylogenetic tree be ordered in a “pruningwise” fashion, which is easily accomplished using the \texttt{reorder()} function along the disease tree.

\textbf{Simulating Trait Data Associated with an Environmental Covariate}

To simulate trait data associated with the environmental component of the data, covariate values were drawn from a Uniform distribution. For the covariate data, $X$, the environmental component of the trait data was determined by the sum of the deterministic regression component of the data ($X\eta$) and a collection of draws from a mean-zero normal distribution with variance $\tau^2$. 
Chapter 4: Results

In this chapter, results are presented from simulation studies and real data analysis using the analysis method in Section 3.1.1 and from simulation studies using the analysis method in Section 3.1.2.

4.1 Genetic Model

4.1.1 Simulation Studies

In order to assess the performance of the proposed technique in terms of power and Type I Error, I begin by analyzing data sets which either include a truly-associated SNP, or do not include an associated SNP. These data sets are simulated under the additive, dominant, and recessive models using the techniques in Sections 3.2.1 and 3.2.2. After simulating the data sets for specified parameter values, the local phylogenetic tree at each SNP is estimated using Blossoc. Next, the score in Equation 3.15 is calculated, using \( k_{\text{max}} = 15 \). The same technique is applied to 200 permutation data sets created by using a permutation of trait values across individuals. The type I error, power, and localization for the two-sample \( t \)-test, QBlossoc (QB), and LSS are presented in Tables 4.1–4.4.

Permutation testing results showed that all three methods control the type I error around 0.05 (see Table 4.1). In terms of power of detection, LSS is competitive with
Table 4.1: Type I error for simulated data sets. For the values of $\alpha$ (the strength of selection) and $\sigma_Y$ (the standard deviation of the quantitative trait per unit time) considered here, the type I error rates are all near 0.05 for these simulation results under the three trait simulation models: additive, dominant, and recessive.

QBlossoc under all three models, as shown in the fifth and seventh columns of Tables 4.2–4.4. These power results are also shown in Figures 4.1(a)-(c), where the red, blue, and black lines show the power of QBlossoc, LSS, and the $t$-test, respectively. The $t$-test is clearly outperformed in all cases by both QBlossoc and LSS (column 3 in Tables 4.2–4.4). In the case of the recessive model, both QBlossoc and LSS show a decrease in power of detection (Figure 4.1(b)), which is expected since a recessive trait has a weaker signal compared to a trait simulated under an additive or dominant model. In the dominant model, there is a notable increase in power of the $t$-test (Figure 4.1(c)), which is evidence of the ability of the $t$-test to detect strong signals. In the dominant model, the strength of selection, $\alpha$, appears to have a larger impact on the power of detection for all three methods, while in the recessive and additive models, different

| Parameters | Additive Model | | | | Recessive Model | | | | | Dominant Model | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | $\sigma_Y$ | $t$-test | QB | LSS | $t$-test | QB | LSS | $t$-test | QB | LSS | $t$-test | QB | LSS |
| 5 | 10 | 0.04 | 0.03 | 0.03 | 0.04 | 0.08 | 0.04 | 0.05 | 0.07 | 0.09 | | | |
| 5 | 20 | 0.03 | 0.06 | 0.07 | 0.07 | 0.05 | 0.04 | 0.06 | 0.06 | 0.04 | | | |
| 5 | 30 | 0.05 | 0.04 | 0.10 | 0.07 | 0.04 | 0.02 | 0.04 | 0.08 | 0.01 | | | |
| 5 | 40 | 0.07 | 0.03 | 0.03 | 0.07 | 0.08 | 0.09 | 0.05 | 0.06 | 0.06 | | | |
| 7.5 | 10 | 0.06 | 0.06 | 0.04 | 0.01 | 0.10 | 0.10 | 0.02 | 0.04 | 0.08 | | | |
| 7.5 | 20 | 0.09 | 0.06 | 0.06 | 0.06 | 0.02 | 0.06 | 0.06 | 0.05 | 0.02 | | | |
| 7.5 | 30 | 0.01 | 0.05 | 0.05 | 0.06 | 0.11 | 0.09 | 0.05 | 0.05 | 0.05 | | | |
| 7.5 | 40 | 0.09 | 0.07 | 0.06 | 0.05 | 0.06 | 0.04 | 0.07 | 0.05 | 0.03 | | | |
| 10 | 10 | 0.03 | 0.08 | 0.03 | 0.06 | 0.08 | 0.06 | 0.08 | 0.04 | 0.05 | | | |
| 10 | 20 | 0.09 | 0.04 | 0.07 | 0.05 | 0.06 | 0.08 | 0.07 | 0.10 | 0.05 | | | |
| 10 | 30 | 0.06 | 0.03 | 0.03 | 0.08 | 0.02 | 0.05 | 0.11 | 0.06 | 0.04 | | | |
| 10 | 40 | 0.05 | 0.02 | 0.03 | 0.03 | 0.06 | 0.09 | 0.03 | 0.09 | 0.04 | | | |
Table 4.2: Power and localization distance (bp) for data sets simulated under the additive model. For the values of $\alpha$ (the strength of selection) and $\sigma_Y$ (the standard deviation of the quantitative trait per unit time) considered here, the power and localization distance (in DNA base pairs) are comparable for QBlossoc and LSS.

Values of $\alpha$ seem to have little effect on the power of detection for all considered methods. Under all three models, as $\sigma_Y$ increases, the power of detection decreases, which is expected since larger amounts of variation in the trait will make any signal harder to detect.

The average localization distance (LocDist) is the shortest distance between the most highly-scored SNP and the associated SNP in DNA base pairs. The localization distances are given in the sixth and eighth columns of Table 4.2. The localization distances are also shown in Figures 4.2(a)-(c), where the red, blue, and black lines show the power of QBlossoc, LSS, and the $t$-test, respectively. Smaller distances indicate a better statistic, and the two methods show approximately the same performance in...
### Table 4.3: Power and localization distance (bp) for data sets simulated under the recessive model.

For the values of $\alpha$ (the strength of selection) and $\sigma_Y$ (the standard deviation of the quantitative trait per unit time) considered here, the power and localization distance (in DNA base pairs) are comparable for QBlossoc and LSS.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>$t$-test</th>
<th>QBlossoc</th>
<th>LSS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Power</td>
<td>Power</td>
<td>Power</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>LocDist</td>
<td>LocDist</td>
<td>LocDist</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>0.06</td>
<td>0.54</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>0.06</td>
<td>0.43</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td>0.08</td>
<td>0.39</td>
</tr>
<tr>
<td>5</td>
<td>40</td>
<td>0.09</td>
<td>0.35</td>
</tr>
<tr>
<td>7.5</td>
<td>10</td>
<td>0.12</td>
<td>0.40</td>
</tr>
<tr>
<td>7.5</td>
<td>20</td>
<td>0.08</td>
<td>0.38</td>
</tr>
<tr>
<td>7.5</td>
<td>30</td>
<td>0.14</td>
<td>0.38</td>
</tr>
<tr>
<td>7.5</td>
<td>40</td>
<td>0.10</td>
<td>0.35</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>0.06</td>
<td>0.35</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>0.09</td>
<td>0.35</td>
</tr>
<tr>
<td>10</td>
<td>30</td>
<td>0.04</td>
<td>0.37</td>
</tr>
<tr>
<td>10</td>
<td>40</td>
<td>0.05</td>
<td>0.39</td>
</tr>
</tbody>
</table>

Terms of localization distance, except in the case of traits simulated under a recessive model, where LSS is slightly outperformed by QBlossoc.

Under all three trait simulation models, the $t$-test is clearly outperformed by QBlossoc and LSS, while LSS is competitive with QBlossoc in this case. All three methods perform more poorly in the case of the recessive model, showing the same trend as in the performance with respect to power. All methods also show an increase in average localization distance as $\sigma_Y$ increases, which is the same trend as in the power of detection. The standard deviation of the localization distances under each trait model are similar for the three considered methods (Table A.1 in Appendix A).

The special case of population stratification was also investigated in the case of the additive model. By using $ms$, population structures involving six subpopulations
Table 4.4: Power and localization distance (bp) for data sets simulated under the dominant model. For the values of $\alpha$ (the strength of selection) and $\sigma_Y$ (the standard deviation of the quantitative trait per unit time) considered here, the power and localization distance (in DNA base pairs) are comparable for QBlossoc and LSS.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>$t$-test</th>
<th>QBlossoc</th>
<th>LSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$\sigma_Y$</td>
<td>Power</td>
<td>LocDist</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>0.29</td>
<td>98280</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>0.19</td>
<td>203367</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td>0.09</td>
<td>177192</td>
</tr>
<tr>
<td>5</td>
<td>40</td>
<td>0.08</td>
<td>232659</td>
</tr>
<tr>
<td>7.5</td>
<td>10</td>
<td>0.47</td>
<td>57881</td>
</tr>
<tr>
<td>7.5</td>
<td>20</td>
<td>0.28</td>
<td>72527</td>
</tr>
<tr>
<td>7.5</td>
<td>30</td>
<td>0.19</td>
<td>146050</td>
</tr>
<tr>
<td>7.5</td>
<td>40</td>
<td>0.16</td>
<td>163952</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>0.62</td>
<td>45764</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>0.28</td>
<td>139669</td>
</tr>
<tr>
<td>10</td>
<td>30</td>
<td>0.17</td>
<td>117105</td>
</tr>
<tr>
<td>10</td>
<td>40</td>
<td>0.19</td>
<td>123131</td>
</tr>
</tbody>
</table>

were specified. Here, the trees are constrained so that after a population splits into two subpopulations, no gene flow exists between the subsequent clades, and the subpopulations evolve independently of one another. This constraint is achieved through specification of times when the splits in populations occur, which are given below. Specifying the number of subpopulations and divergence times allows the investigation of the effect of the complexity of the population structure on the results. An example of a true phylogenetic tree for a particular replication with six specified subpopulations is shown in Figure 4.3.

Results for selected studies are shown in Tables 4.5 - 4.6 and Figure 4.4. The times shown are the divergence times at which population splits occurred (looking from present to past). Times are given in evolutionary time units of $4N_0\mu = 1.6 \times 10^{-5}$.
Figure 4.1: Power of detection. (a), (b), and (c) show the power of detection achieved under each of the three models: additive, recessive, and dominant, respectively. Black lines show the power of the \( t \)-test, red lines show the power of QBlossoc, and blue lines show the power of LSS.

In each case, six subpopulations were specified. Divergence times for the case with six subpopulations are 0.1, 0.1, 0.1, 0.2, and 1.0, and are omitted from Tables 4.5 and 4.6 for simplicity. The results in Table 4.5 show that the type I error is controlled by all three considered methods in this special case. Also, the power of detection and localization distance (in DNA bp) are very similar for QBlossoc and
Figure 4.2: Localization distance. (a), (b), and (c) show the average localization distance in bp of each method under the three models: additive, recessive, and dominant, respectively. Black lines show the average localization distance of the $t$-test, red lines show the average localization distance of QBoost, and blue lines show the average localization distance of LSS.

LSS, while the $t$-test is outperformed by these two methods (see Table 4.6). The standard deviation of the localization distances under each trait model are similar for the three considered methods (Table A.2 in Appendix A). In Figure 4.4, we see the
Figure 4.3: Example of population structure. This phylogenetic tree shows a form of population structure which impacts the correlation structure among the trait values. Six particular subpopulations are shown through the six groups of individuals present in the tree. The splits showing the six groups are denoted by hash marks. Notice that the underlying population structure is not necessarily shown directly in the evolutionary history of a SNP, but has a strong influence on the phylogenetic tree.

Upon further investigation, it appears that even though QBlossoc and LSS have comparable powers of detection and average localization distances, QBlossoc and LSS sometimes detect different data sets. Figure 4.5(a) shows the p-values for each simulated data set from the study with $\alpha = 5$ and $\sigma_Y = 30$ shown in Table 4.6. The horizontal and vertical lines in (a) and (b) represent the cutoff values for significance. The 65 observations in the lower left corner were detected by both QBlossoc and LSS, while the 26 observations in the upper right corner were detected by neither method. However, the seven observations in the upper left corner were detected by the proposed LSS but not QBlossoc, while the two observations in the lower right corner were detected by QBlossoc but not LSS. This is an indication that the proposed technique sometimes identifies different types of associations than QBlossoc.
Table 4.5: Type I error for data sets simulated with population structure. For the values of $\alpha$ (the strength of selection) and $\sigma_Y$ (the standard deviation of the quantitative trait per unit time) considered here, the type I error rates are all near 0.05 for these simulation results in the case of six subpopulations. Population divergence times for the six subpopulations are omitted for simplicity. See text for details.

Further, Figure 4.5(b) shows the localization distances for each data set. Observations below the diagonal line indicate data sets in which QBlossoc was able to better localize the associated SNP, while observations above the diagonal line indicate data sets in which LSS was able to better localize the associated SNP. Twenty-three observations were better localized by LSS, while forty-three observations were better localized by QBlossoc. These simulation study results indicate that the proposed method is comparable with QBlossoc in the general case, and detecting different types of relationships between SNPs and quantitative traits in the case of population structure. Additionally, both QBlossoc and LSS appear to control the Type I Error in
<table>
<thead>
<tr>
<th>Parameters</th>
<th>$t$-test</th>
<th>QBlossoc</th>
<th>LSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$\sigma_Y$</td>
<td>Power</td>
<td>LocDist</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>0.12</td>
<td>92849</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>0.11</td>
<td>131605</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td>0.13</td>
<td>200771</td>
</tr>
<tr>
<td>5</td>
<td>40</td>
<td>0.12</td>
<td>193688</td>
</tr>
<tr>
<td>7.5</td>
<td>10</td>
<td>0.17</td>
<td>87300</td>
</tr>
<tr>
<td>7.5</td>
<td>20</td>
<td>0.21</td>
<td>96049</td>
</tr>
<tr>
<td>7.5</td>
<td>30</td>
<td>0.14</td>
<td>99418</td>
</tr>
<tr>
<td>7.5</td>
<td>40</td>
<td>0.08</td>
<td>162016</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>0.33</td>
<td>83512</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>0.14</td>
<td>77187</td>
</tr>
<tr>
<td>10</td>
<td>30</td>
<td>0.13</td>
<td>134557</td>
</tr>
<tr>
<td>10</td>
<td>40</td>
<td>0.10</td>
<td>151681</td>
</tr>
</tbody>
</table>

Table 4.6: Power and localization distance for data sets simulated with population structure. For the values of $\alpha$ (the strength of selection) and $\sigma_Y$ (the standard deviation of the quantitative trait per unit time) considered here, the power and localization distance (in DNA base pairs) are comparable for QBlossoc and LSS in the case of six subpopulations. Population divergence times for the six subpopulations are omitted for simplicity. See text for details.

these simulation studies as shown in the second and third columns of Table 4.5.

### 4.1.2 Real Data Analysis

Having seen that the proposed method performs well for simulated data, next the method is applied to a GWAS data set. The data set from [42] includes both SNP data and phenotypic data for 288 outbred mice. Phenotypic data for each mouse include observations about eight quantitative cardiovascular traits. Here, the trait under study is the high-density lipoprotein cholesterol level (HDL), and I will set $k_{max} = 15$ in LSS. The SNP sites on two chromosomes with previously-detected
Figure 4.4: Results in the presence of population structure. (a) and (b) show the power of detection and average localization distance in bp, respectively, of each method under the additive trait model in the presence of population structure. Black lines show the average localization distance of the t-test, red lines show the average localization distance of QBlossoc, and blue lines show the average localization distance of LSS.

Strong signals and one chromosome without a previously-detected strong signal are analyzed. In order to phase the data from genotypes into haplotypes, Beagle [3] was used, as in the original data analysis [42].

Chromosome 1 included data for 4,165 SNPs, and was analyzed using the proposed method. The method detected the chromosome with a p-value less than 0.005. The score results, presented in Figure 4.6(a), show that of the two sites detected as highly significant by [42], LSS shows a peak very near to one of these sites. In addition, three other sites not previously detected show very large peaks in LSS. These results support the simulation study results, in which LSS tends to detect different signals than previous methods.
Figure 4.5: Example of detection p-values and localization distances. These plots show the detection p-values and localization distance (in DNA bp) in (a) and (b), respectively, for a particular replicate. (a) shows that QBlossoc and LSS are picking up different associations during simulation. In (a), values in the upper left and lower right corners of the plot show associations detected only by LSS and QBlossoc, respectively. Seven observations fall in the upper left corner, while two observations fall in the lower right corner of the plot. The sixty-five observations in the lower left corner were detected by both methods, while the remaining twenty-six observations were detected by neither method. (b) shows the associations better localized by LSS and QBlossoc in the upper left and lower right regions, respectively. Twenty-three observations were better localized by LSS and forty-three observations were better localized by QBlossoc.

Zhang et al. [42] also found a strong genetic signal on Chromosome 5. Chromosome 5 included data for 3,185 SNPs, and was analyzed using the proposed method. The method detected the chromosome with a p-value less than 0.005. The results presented in Figure 4.6(b), show a peak in LSS near the SNP site previously detected as highly significant [42]. In addition, two other regions on the chromosome not previously detected show very large peaks in LSS.
Figure 4.6: Likelihood score statistic plots. For the analysis of the HDL level in mice, these plots show LSS plotted against the location of the SNP in DNA base pairs along Chromosomes 1, 5, and 8 in (a), (b), and (c), respectively. The vertical lines in (a) and (b) represent locations detected as highly significant by Zhang et al. [42]
Chromosome 8 was also analyzed, and results are presented in Figure 4.6(c). Chromosome 8 included data for 1,159 SNP sites. Zhang et al. [42] did not detect any highly significant SNP sites on Chromosome 8. The likelihood analysis resulted in a detection p-value of 0.055 for this chromosome, which is not significant.

4.2 Modeling Environmental Covariates

4.2.1 Simulation Studies

In order to assess the performance of the proposed technique when there is information about fixed external covariates in addition to genetic SNP data, data sets that either include a truly-associated SNP or do not include an associated SNP and either include a related external covariate or do not include such a covariate are analyzed. Here, detecting any relationship between the quantitative trait and either the truly-associated SNP or the truly-associated external covariate is of interest. In the case that \( \rho = 0 \), the only relationship present in the data is between the trait and covariate. Since there is no truly-associated SNP, localization distance need not be studied in this case. In the other extreme case, when \( \rho = 1 \), the external covariate is unrelated to the trait, and the alternative model reduces to the genetic-only model presented in Section 3.1.1.

For comparison, the simulated data were also analyzed using the SNPassoc package [9] implemented in R, which looks for associations between any SNP along the chromosome and the quantitative trait while adjusting for the environmental covariate. In the case of a continuous response variable, SNPassoc fits a generalized linear model (GLM) using an identity link function to the response variable (the quantitative trait) using the state of the considered SNP as the explanatory variable while
adjusting for the environmental covariate. In the case of the codominant model, at each SNP, \textit{SNPassoc} uses the genotypes of individuals as the explanatory variable in the GLM. The null model is a GLM fit to the phenotype using the identity link function while adjusting for the environmental covariate. The test statistic used is twice the log of the ratio of the likelihood under the alternative model to the likelihood under the null model. At each SNP, the p-value reported is the probability of observing a test statistic at least as large as that observed. This probability is calculated using a \(\chi^2\) distribution, with its degrees of freedom calculated from the number of genotypes observed at the considered SNP. Permutation testing is performed using these p-values to control the overall error rate.

Since the null hypothesis of LSS is that there is neither an association between the quantitative trait and the environmental covariate nor an association between the quantitative trait and the considered SNP, a likelihood ratio test (LRT) is also performed at each SNP in order to compare to a method with the same hypotheses. Under the alternative hypothesis, the likelihood is calculated using the GLM fit to the quantitative trait data with the genotypic SNP data and environmental covariate as explanatory variables. Then, the null likelihood is the likelihood of the GLM with the quantitative trait as the response variable and only the intercept as an explanatory variable. The likelihood ratio test uses the ratio of the likelihoods of the alternative to the null model as the test statistic. To control the overall error rate, permutation testing is performed using this test statistic in the same way as it is performed using LSS.

The Type I Error, power, and localization distance of LSS, \textit{SNPassoc}, and the LRT are assessed using data sets simulated with the technique in Section 3.2.3. After
Table 4.7: Type I Error of LSS, SNPassoc, and the LRT for simulated data sets with environmental covariates. For the values of $\rho$ (the weight of the contribution of the genetic component to the quantitative trait values) and $\tau$ (standard deviation of the environmental component of the trait) considered, the Type I Error rates are all near 0.05 for these simulation studies. Note that when $\rho = 1$, there is no environmental influence acting on the trait, so the Type I Errors presented in the last row of this table correspond to two simulation studies under the same settings.

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>SNPassoc $\tau = 5$</th>
<th>SNPassoc $\tau = 15$</th>
<th>LRT $\tau = 5$</th>
<th>LRT $\tau = 15$</th>
<th>LSS $\tau = 5$</th>
<th>LSS $\tau = 15$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.00</td>
<td>0.04</td>
<td>0.04</td>
<td>0.02</td>
<td>0.08</td>
<td>0.08</td>
</tr>
<tr>
<td>0.25</td>
<td>0.06</td>
<td>0.06</td>
<td>0.02</td>
<td>0.10</td>
<td>0.00</td>
<td>0.10</td>
</tr>
<tr>
<td>0.50</td>
<td>0.02</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.08</td>
<td>0.04</td>
</tr>
<tr>
<td>0.75</td>
<td>0.06</td>
<td>0.06</td>
<td>0.04</td>
<td>0.08</td>
<td>0.04</td>
<td>0.06</td>
</tr>
<tr>
<td>1.00</td>
<td>0.06</td>
<td>0.02</td>
<td>0.04</td>
<td>0.06</td>
<td>0.08</td>
<td>0.04</td>
</tr>
</tbody>
</table>

simulating the data sets for specified parameter values, the local phylogenetic tree at each SNP is estimated using Blossoc. Next, the score in Equation 3.37 is calculated via numerical optimization, with $k_{max} = 10$. The same technique is applied to 50 permutation data sets created by using a permutation of trait values across individuals.

Tables 4.7–4.9 and Figure 4.7 show the Type I Error, power, and localization performance of LSS, SNPassoc, and the LRT for this simulation study. All considered methods have Type I Error rates around 0.05 in the studied parameter settings (see Table 4.7). In addition, when $\tau = 5$, the LSS detects the relationship between the trait and the environmental covariates and/or true-causative SNP extremely well, as shown in the sixth columns of Table 4.8. LSS outperforms both SNPassoc (column 2 of Table 4.8) and the LRT (column 4 of Table 4.8) in all but two cases.
Table 4.8: Power of LSS, SNPassoc, and the LRT for simulated data sets. For the values of $\rho$ (the weight of the contribution of the genetic component to the quantitative trait values) considered here, the power of detection of each method is given for the case when $\tau = 5$ (columns 2, 4, and 6), and the case when $\tau = 15$ (columns 3, 5, and 7). In all but two cases, LSS outperforms both the LRT and SNPassoc. Also, in the case that $\rho = 1$, there is no environmental influence acting on the trait, so the results presented in the last row of each table are two simulation studies under the same settings.

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$\tau = 5$</th>
<th>$\tau = 15$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SNPassoc</td>
<td>LRT</td>
</tr>
<tr>
<td>0.00</td>
<td>0.08</td>
<td>1.00</td>
</tr>
<tr>
<td>0.25</td>
<td>0.10</td>
<td>1.00</td>
</tr>
<tr>
<td>0.50</td>
<td>0.60</td>
<td>1.00</td>
</tr>
<tr>
<td>0.75</td>
<td>0.74</td>
<td>0.98</td>
</tr>
<tr>
<td>1.00</td>
<td>0.92</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Table 4.9: Average localization distance of LSS, SNPassoc, and the LRT for simulated data sets. For the values of $\rho$ (the weight of the contribution of the genetic component to the quantitative trait values) considered here, the average localization distance (average number of DNA bp separating the most highly-scored and true-causative SNPs) of each method is given for the case when $\tau = 5$ and (columns 2, 4, and 6) and for the case when $\tau = 15$ (columns 3, 5, and 7). Note that the localization distance does not exist when $\rho = 0$ since there is no truly associated SNP in this case. Also, in the case that $\rho = 1$, there is no environmental influence acting on the trait, so the results presented in the last row of the table for $\tau = 5$ and $\tau = 15$ are two simulation studies under the same settings.

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$\tau = 5$</th>
<th>$\tau = 15$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SNPassoc</td>
<td>LRT</td>
</tr>
<tr>
<td>0.25</td>
<td>277926</td>
<td>302534</td>
</tr>
<tr>
<td>0.50</td>
<td>106454</td>
<td>108296</td>
</tr>
<tr>
<td>0.75</td>
<td>35738</td>
<td>44970</td>
</tr>
<tr>
<td>1.00</td>
<td>734</td>
<td>734</td>
</tr>
</tbody>
</table>
Figure 4.7: Power and localization in covariate analysis. (a) and (b) show the power of detection and average localization distance in bp, respectively, of each method when analyzing data with covariates. Black lines show results from \textit{SNPassoc}, red lines show results from the LRT, and blue lines show results from LSS. LSS outperforms the other two techniques in most cases with respect to both power of detection and average localization distance.

When this variation is much larger ($\tau = 15$), LSS still performs well in terms of detection of any association in the data (column four of Table 4.8 and Figure 4.7(a)). When the influence from the external covariate on the trait is more variable, the association will be harder to detect, so lower power is expected in this case relative to the case presented in Table 4.8. Even when the variation associated with the environmental component of the trait is large ($\tau = 15$), LSS still performs well in detecting an influence between the quantitative trait and either the covariate or the genetic location.

In the case that $\rho = 1$, the performance of LSS is similar to the reduced model including only a genetic component of the quantitative trait presented in Section 4.1.1. The similarity of results across the two simulation studies with $\rho = 1$ also
points toward the reliability of the method. This shows that the proposed model including environmental covariates does, indeed, detect a true-causative SNP when the environmental covariate under study is unrelated to the quantitative trait.

The average localization distance (Table 4.9 and Figure 4.7(b)) is the average number of DNA bp between the true-causative SNP and the most SNP most highly-scored by each method. Here, an increase in the performance of LSS occurs as a larger amount of the trait is attributed to the genetic component (columns 6 and 7). However, SNPassoc (columns 2 and 3) and the LRT (columns 4 and 5) show a decrease in performance when there is a mixture of environmental and genetic influences on the quantitative trait (when $\rho = 0.5$). In addition, the standard deviations in localization distances for each method show this same pattern (Table A.3 in Appendix A). Note that the localization distance when $\rho = 0$ does not exist since there is no true-causative SNP in this case, and is omitted from Tables 4.9 and A.3. These initial results show the strength of LSS in detecting environmental covariates and detecting and localizing SNP sites significantly associated with a quantitative trait under study and its advantages over the previously-proposed techniques used by SNPassoc and the LRT.

Example Results from a Single Replicate

In order to see how well the separate components are estimating using LSS, one replicate (from the $\rho = 0.5$ and $\tau = 5$ case) was chosen to study. In the case where there is a genetic and environmental signal in the data (Figure 4.8), the LSS (black line) shows a strong peak near the truly-causative SNP, and remains low, comparatively, at other locations across the chromosome. In the case where there
is no signal, the LSS (blue line) remains fairly constant and very low across the chromosome. This is the expected behavior of the statistic across a chromosome.

Figure 4.8: Example of behavior of LSS across a chromosome. The plot shows the value of the LSS across a chromosome for a particular replication with $\rho = 0.5$. When there is both an environmental and genetic effect, the LSS (black line) shows a high peak near the truly-causative SNP, and remaining low otherwise. The red line shows the location of the truly-associated SNP. When the trait is not related to the environmental covariate nor a SNP on this chromosome, LSS remains low across the chromosome (blue line).

LSS shows a similar pattern of behavior across the chromosome in the number of clusters it chooses to maximize the penalized likelihood across the chromosome. In the presence of a genetic and environmental signal (Figure 4.9), the number of clusters chosen to maximize the penalized likelihood is high near the truly-causative SNP, and relatively low otherwise. In the case shown by the blue line in Figure 4.9, there is no signal in the trait. The number of clusters chosen to maximize the penalized likelihood remains low across the entire chromosome.
Figure 4.9: Example of number of clusters chosen by the LSS across a chromosome. This plot shows the value of the number of clusters chosen by the LSS across a chromosome for a particular replication with $\rho = 0.5$. In the case of a relationship, the number of clusters chosen is large near the truly-causative SNP (black line), while when there was no relationship (blue line), the number of clusters remained low across most of the chromosome. The red line shows the location of the truly-assocated SNP corresponding to the replication with the black line.

Figures 4.10–4.12 show the estimated parameter values, plotted with respect to the SNP position across a chromosome for the same replication used above. Figure 4.10 shows the values of $\hat{\beta}_1$ remain fairly stable across the chromosome in both the setting where the environmental covariate is related to the quantitative trait (black line), and the case where it is not related (blue line). In the case of a relationship, the estimated slope is estimated to be larger than zero at all loci. (The true slope for this parameter setting is 1.25.) When there is no signal in the trait, the slope estimates remain close to zero.

A similar constancy in estimates is present in the baseline mean estimates across this particular chromosome, shown in Figure 4.11. The baseline estimates (sum of
Figure 4.10: Example of behavior of estimation of $\beta_1$ across a chromosome. These plots show the estimated slope parameter at each SNP along a chromosome for a particular replication with $\rho = 0.5$ when the data has both a genetic signal from this chromosome and a relationship with the environmental covariate (black line) and when there is no signal (blue line).

the intercept of the environmental component of the trait and an arbitrary cluster mean from the genetic component) remain constant, but take very different values in the case of the presence and absence of a genetic and environmental signal. In the absence of a signal (blue line), the estimates remain close to the sample mean trait value in this replication (83.71), which could be an artifact of the estimated differences in cluster means and estimated slopes remaining close to zero. In the case of a relationship between the trait and the SNP and environmental covariate, the estimates of these parameters are far from zero, which could be the reason that the baseline mean (black line) is so far below the sample mean of the quantitative trait data in this replication (84.75).
Figure 4.11: Example of behavior of baseline estimate across a chromosome. These plots show the estimated baseline mean ($\hat{\beta}_0 + \hat{\mu}_k$) across a chromosome for a particular replication with $\rho = 0.5$ when the data has both a genetic signal from this chromosome and a relationship with the environmental covariate (black line) and when there is no signal (blue line). The red line shows the location of the truly-associated SNP.

Figure 4.12 shows the estimated means within each cluster at each location along the chromosome. At each location along the chromosome, the black circle represents the estimated difference in mean trait value of individuals between the first and $k^{th}$ clusters. In the case that at least 3 clusters were chosen, the red circle represents the estimated difference in mean trait value of individuals between the second and $k^{th}$ clusters. This pattern continues for the different colored circles at each location, up until the estimated difference between the $(k - 1)^{th}$ and $k^{th}$ cluster means. In the presence of a signal (Figure 4.12(a)), the estimated within-cluster means appear to differ at most locations along the chromosome. In Figure 4.12(b), the estimated within-cluster means in the absence of a genetic signal are shown. Notice that there
Figure 4.12: Example of behavior of within-cluster mean estimates along a chromosome. These plots show the estimated within-cluster means at each SNP along a chromosome for a particular replication with $\rho = 0.5$ when the data has both a genetic signal from this chromosome and a relationship with the environmental covariate (as in (a)) and when there is no signal (as in (b)). The red line in (a) shows the location of the truly-associated SNP. At each location, the black dot represents the estimated difference in the first and $k^{th}$ cluster means. If at least three clusters were chosen, the red dot represents the estimated difference between the second and $k^{th}$ cluster means, and so on for different colors, up to the estimated difference between the $(k - 1)^{th}$ and $k^{th}$ cluster means.

is less variability in the cluster means across the chromosome as compared to the case of presence of a genetic signal.
The findings of this particular replication indicate that LSS is behaving as expected in the presence of both a genetic and environmental signal, as well as in the absence of a signal. Now, the analysis is continued by considering the parameter estimates more generally for each parameter setting.

**Parameter Estimates at the Maximally-Scored SNP**

For each replicate shown in Table 4.8, the parameter estimates at the maximally-scoring SNP across the chromosome are studied. These SNPs were the SNPs used in the localization analysis. The estimates of the genetic and environmental variances, regression parameters, and within-cluster means are studied. This will allow the exploration of the behavior of LSS in the case that different amounts of contribution from the genetic signal and environmental covariate can be attributed to the quantitative trait.

The estimates of the log of the genetic and environmental trait variances are shown for the $\tau = 5$ case (Figure 4.13(a)) and the $\tau = 15$ case (Figure 4.13(b)). For both settings of $\tau$, as the contribution from the environmental component of the trait increases (equivalent to a decrease in $\rho$), the amount of variability attributed to the environmental component of the trait decreases as expected. This trend is more clear in the $\tau = 15$ case compared to the $\tau = 5$ case. As the contribution of the genetic component of the trait increases, the estimates of the log of the variances associated with the genetic component of the trait stay fairly constant until $\rho = 1$, when there is an increase in the estimated parameter values. When $\rho = 0$ and the quantitative trait is related solely to the environmental covariate, the estimate of the log of the variance of the genetic component of the trait is larger in the case that $\tau = 15$ compared to the case when $\tau = 5$, which shows the response of the model when there is a weaker
Figure 4.13: Estimates of genetic and environmental variances at the maximally-scored SNP. This plot shows boxplots of the log of the estimated genetic variances and the log of the estimated environmental variances, $\log(\hat{\nu})$, for each of the 50 replications performed for data simulated using different values of $\rho$ and $\tau$. (a) shows the case when $\tau = 5$ and (b) shows the case when $\tau = 15$.

relationship between the trait and the environmental covariate. The bias seen in these estimates could be due to the downward bias often seen in the maximum likelihood estimates of variances associated with correlation structures in mixed models [39], but
these biases do not appear to affect the method’s performance in terms of detection or localization.

Figure 4.14 shows boxplots of the estimates of the differences in within-cluster means for each of the 50 replications performed when $\tau = 5$ (Figure 4.14(a)) and when $\tau = 15$ (Figure 4.14(b)). The differences shown are differences from the mean of the $k^{th}$ cluster. As the contribution from the genetic component of the trait increases (equivalent to an increase in $\rho$), the estimated within-cluster means tend to show larger differences. There are also larger differences among cluster means when $\tau = 15$, which could be attributed to the larger overall variability in quantitative trait values in this setting. Another notable feature of these plots is the increase in number of clusters chosen by LSS for the genetic component of the trait when the portion of variability attributable to the genetic component is large.

In this model, the baseline mean refers to the sum of the intercept term in the regression model and the mean of the $k^{th}$ cluster since the model is fit without the term for the last cluster to avoid singularity in the computation of parameter estimates (see Section 3.1.2). The boxplots of the baseline mean trait value show an increase in the estimated parameter values as more variation in the quantitative trait is attributed to a genetic signal and environmental covariate in the case that $\tau = 5$ and the case that $\tau = 15$ (Figures 4.15 (a) and (b), respectively). Since $\beta_0 = (1 - \rho)\eta_0 = (1 - \rho)10$, and each cluster mean is expected to remain somewhere between 80 and 100 (the two target values from the OU process used in simulation of the genetic component), a range for the baseline mean for each value of $\rho$ can be calculated, namely, $[80\rho + 10(1 - \rho), 100\rho + 10(1 - \rho)]$. These are the values shown by green
Figure 4.14: Estimated differences in cluster means at the maximally-scored SNP. This plot shows boxplots of the estimated differences of within-cluster means for each of the 50 replications performed for data simulated using different values of $\rho$ when $\tau = 5$ (a) and when $\tau = 15$ (b). As the contribution from the genetic component of the trait increases (equivalent to an increase in $\rho$), the estimated within-cluster means tend to show larger differences. There are also larger differences among cluster means when $\tau = 15$. Note that the estimated means shown are differences from the mean of the considered cluster from the mean of the $k^{th}$ cluster.

dashes in each setting. For both settings of $\tau$, the estimates of the baseline mean appear to fall near the expected means shown by the green dashes.
Figure 4.15: Estimates of $\beta_0 + \mu_k$ at the maximally-scored SNP. This figure shows boxplots of the estimated baseline mean for each of the 50 replications performed for data simulated using different values of $\rho$. (a) shows the case when $\tau = 5$ and (b) shows the case when $\tau = 15$. The mean is the sum of the intercept term in the regression model and the mean of the $k^{th}$ cluster. The green dashes represent the ranges of baseline means expected in each parameter setting. See text for details on their calculation.

Figure 4.16 shows boxplots of the estimated slope parameter (contributed by the environmental component of the trait) for each of the simulated data sets. As $\rho$ increases, the contribution of the environmental covariate tends to decrease in both
the case that \( \tau = 5 \) (Figure 4.16(a)) and the case that \( \tau = 15 \) (Figure 4.16(b)). When \( \rho = 1 \), the boxplot of slope estimates is centered about 0, which is expected since the environmental covariate is not related to the trait in this case. In each case, the slope estimates seem to be centered about their true values in both cases (shown as blue dots). The increased variability in estimates in Figure 4.16(b) compared to those in Figure 4.16(a) is expected since (b) shows the case with larger variance in the environmental component of the trait. This is evidence that the model is behaving appropriately when different amounts of variation in the trait are attributable to external influences.

### 4.3 Summary

In the case of only a genetic component to the trait (Section 4.1.1), LSS outperforms the \( t \)-test and is competitive with QBlossoc in terms of power of detection and localization. The similarity in performance of LSS and QBlossoc is also displayed in the presence of a complicated population substructure. Upon closer inspection, it appears that LSS and QBlossoc sometimes detect different types of signals during data analysis.

In the mouse data analysis (Section 4.1.2), the same quality of LSS and previous methods is portrayed, in that LSS shows high peaks near one location previously-detected as linked to HDL in mice, and several other previously undetected locations.

One advantage of LSS lies in its direct extension to covariates. Results from Section 4.2.1 show that LSS performs well in terms of power of detection when the contribution of genetic and external influences on a trait is mixed and has both a genetic and an environmental component. LSS shows an increase in performance as
Figure 4.16: Estimates of $\beta_1$ at the maximally-scored SNP. These boxplots show the estimated slope parameter (contributed by the environmental component of the trait) for each of the 50 replications performed for data simulated using different values of $\rho$. (a) shows the case when $\tau = 5$ and (b) shows the case when $\tau = 15$. As $\rho$ increases, the contribution of the environmental covariate tends to decrease. The blue dots represent the true values of $\beta_1$ for each parameter setting.

The amount of variation attributable to the genetic influence on the trait increases, which is expected and intuitive in this setting.

The ability of LSS to analyze GWAS data which includes external covariates addresses a current limitation of statistical methodology in quantitative trait mapping.
In addition, the modeling framework that LSS provides can be easily extended and implemented to analyze a variety of types of GWAS data including covariates.
Chapter 5: Discussions and Conclusions

In this dissertation, a method is presented to search for SNPs associated with quantitative traits in GWAS data in a framework with a quantitative trait related only to a truly-causative SNP, and in a framework with a quantitative trait related to both a truly-causative SNP and an environmental covariate. This chapter includes a discussion of results under the genetic-only model in Section 5.1, a discussion of results in the case of modeling environmental covariates in Section 5.2, and some future directions of this work in Section 5.3.

5.1 Genetic Model

The proposed methods to analyze quantitative trait data in the genetic model stem from the framework of QBlossoc, but relax the assumptions of independence and common variance between observations. The proposed method looks at this problem using a realistic framework which accounts for the evolutionary relationships among SNPs. However, as opposed to previous techniques using these evolutionary relationships, the method here remains computationally feasible by using only the broad-scale relationships present in the evolutionary history among SNPs. These evolutionary relationships impact results especially in the presence of strong population structure.
Using a biologically-sensible technique, simulated data sets were obtained in both the general case and in the presence of population structure. Simulation results showed that LSS is competitive with QBlossoc in terms of localization and power of detection, and that different chromosomes are sometimes detected by LSS and by QBlossoc. In the presence of population stratification, the proposed score shows a slightly higher power of detection than QBlossoc. For the real data example studying 288 outbred mice, analysis using the proposed tree estimation and likelihood score showed that LSS detects two SNPs previously linked to HDL in mice. In addition, LSS also detected several SNPs not previously mentioned in the literature.

One of the advantages of this proposed method is its use of ancestral information to approach this problem. This framework is more realistic than other previous approximations. Also, the use of the broad-scale evolutionary relationships among SNPs makes the technique computationally feasible. Computation times for the branch length estimation and LSS analysis, including permutation testing, ranged from approximately 3.5 to 5.5 seconds per SNP for the simulated data sets with 100 observations, which typically included between 65 and 105 SNPs. Although these times tend to take longer than QBlossoc and the $t$-test, LSS outperforms the $t$-test in terms of power. In addition, LSS performs fewer contributions in the clustering portion of the approach than QBlossoc does by using the $(k - 1)$ earliest edges in the tree. For the real data analysis, with 576 observations, these computation times ranged from approximately 6 to 35 minutes per SNP, depending on the number of SNPs along the chromosome (ranging from 1,159 for Chromosome 8 to 4,165 for Chromosome 1). It should be noted that individual SNP computations are easily parallelized in this setting.
5.2 Modeling Environmental Covariates

By extending the model presented in Section 3.2 to model environmental covariates as well as genetic data, the model is much more flexible, and showed a strong power of detection. In the case of a quantitative trait related to both an environmental covariate and a SNP, LSS performed well with respect to power of detection of an association with either component, showing its ability to detect associations which are attributed to a mixture of genetic and environmental influences. In addition, in the case that the quantitative trait was related to only a genetic or an environmental influence, LSS maintains its good performance, which is expected as stated in Section 3.2.2.

In addition to being able to detect associations between SNPs and the quantitative trait, the proposed model showed an ability to localize the SNPs, especially when a large component of the trait is due to a truly-causative SNP. At the most highly-scored locus, the genetic and environmental variance estimates of the model showed the same limitations previously mentioned in the literature in models with additive variance components [39]. In terms of the remaining parameters, estimates showed little to no bias, indicating that the model can accurately separate the means due to the genetic and environmental components, within the contraints of estimable parameters.

The flexibility that adding covariates into the model enables addresses a serious limitation in quantitative trait mapping. For instance, many environmental covariates affect cardiovascular traits independently of an individual’s genotype. Thus, by adding these covariates into the model, different types of signals can be detected when a mixture of genetic and environmental factors affect the trait.
5.3 Future Directions

Although this technique begins to address the limitations of current statistical methodology in the problem of quantitative trait mapping, the technique has several avenues that could be pursued in order to extend the method to even more general cases.

First, the current likelihood score requires that genotypic data be phased into haplotypes prior to analysis. Phasing is a nontrivial process which is subject to error. By extending the tree estimation method and likelihood score to be computed on genotypic data, these methods will be more easily applied to real data sets. Another nontrivial step of the data analysis process is the tree topology and branch length estimation procedures. Any improvement to the accuracy of the estimated topology or branch lengths will improve the estimation of the covariance structure, and thus could improve the power of detection of LSS, specifically in the case where only a genetic component is influencing the trait value.

In terms of types of environmental covariates, only continuous covariates have been tested so far under the assumptions of normality and independence. However, other data types could be built into the model. For instance, in the setting of survival analysis, an additive model exists to model a trait in terms of covariates, and it appears that method of analysis in Section 3.2.2 could be extended to this case.

The idea of using the branch lengths from the phylogeny representing the evolutionary history at each SNP provides a new way to investigate covariance structures in existing data. Because this variance-covariance structure is determined by the
phylogeny, many existing statistical methods can be implemented with this variance-
covariance structure. This is what allows the model to be flexible and readily im-
plemented in a wider range of GWAS data sets, including those which contain in-
formation about covariates. By pursuing these possible research directions, I can
continue to both simulate and analyze data in a variety of settings, contributing to
the statistical methodology in quantitative trait mapping.
Appendix A: Supplementary Results
Table A.1: Standard deviation of localization distances for simulated data sets. For the values of $\alpha$ (the strength of selection) and $\sigma_Y$ (the standard deviation of the quantitative trait per unit time) considered, the standard deviations of localization distances are similar among the methods considered for these simulation results under the three trait simulation models: additive, dominant, and recessive.
Table A.2: Standard deviation of localization distances for simulated data sets in the presence of population structure. For the values of $\alpha$ (the strength of selection) and $\sigma_Y$ (the standard deviation of the quantitative trait per unit time) considered in the presence of population structure, the standard deviations of localization distances are similar among the three considered methods.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$\sigma_Y$</th>
<th>$t$-test</th>
<th>QB</th>
<th>LSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>10</td>
<td>20097.6</td>
<td>73833.4</td>
<td>37352.9</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>218435.8</td>
<td>147787.0</td>
<td>168181.5</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td>272409.4</td>
<td>218625.2</td>
<td>219278.1</td>
</tr>
<tr>
<td>5</td>
<td>40</td>
<td>220375.9</td>
<td>208881.2</td>
<td>228044.1</td>
</tr>
<tr>
<td>7.5</td>
<td>10</td>
<td>179903.6</td>
<td>46119.2</td>
<td>22144.7</td>
</tr>
<tr>
<td>7.5</td>
<td>20</td>
<td>194579.3</td>
<td>80875.8</td>
<td>74975.1</td>
</tr>
<tr>
<td>7.5</td>
<td>30</td>
<td>218435.9</td>
<td>52083.5</td>
<td>111441.6</td>
</tr>
<tr>
<td>7.5</td>
<td>40</td>
<td>230468.8</td>
<td>176974.1</td>
<td>204945.1</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>187340.4</td>
<td>54290.0</td>
<td>59528.6</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>160971.4</td>
<td>26873.9</td>
<td>57027.2</td>
</tr>
<tr>
<td>10</td>
<td>30</td>
<td>246745.7</td>
<td>103736.6</td>
<td>53913.7</td>
</tr>
<tr>
<td>10</td>
<td>40</td>
<td>210393.3</td>
<td>89756.7</td>
<td>116860.2</td>
</tr>
</tbody>
</table>

Table A.3: Standard deviation of localization distance of LSS, $\text{SNPassoc}$, and the LRT for simulated data sets. For the values of $\rho$ (the weight of the contribution of the genetic component to the quantitative trait values) considered here, the standard deviation of the localization distance of each method is given for the case when $\tau = 5$ and (columns 2, 4, and 6) and for the case when $\tau = 15$ (columns 3, 5, and 7). Note that the localization distance does not exist when $\rho = 0$ since there is no truly associated SNP in this case. Also, in the case that $\rho = 1$, there is no environmental influence acting on the trait, so the results presented in the last row of the table for $\tau = 5$ and $\tau = 15$ are two simulation studies under the same settings.

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$\tau = 5$</th>
<th>$\tau = 15$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\text{SNPassoc}$</td>
<td>$\text{LRT}$</td>
</tr>
<tr>
<td>0.25</td>
<td>265193.6</td>
<td>247633.0</td>
</tr>
<tr>
<td>0.50</td>
<td>213031.1</td>
<td>216268.8</td>
</tr>
<tr>
<td>0.75</td>
<td>110317.9</td>
<td>113554.0</td>
</tr>
<tr>
<td>1.00</td>
<td>3092.4</td>
<td>3092.4</td>
</tr>
</tbody>
</table>
Bibliography


