MARKOV CHAIN MONTE CARLO IN GENETICS: SUBPHENOTYPING, LINKAGE DISEQUILIBRIUM MODELING, AND FINE MAPPING

by

Ziqian Geng

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2014

Doctoral Committee:

Associate Professor Sebastian K. Zöllner, Chair
Professor Michael L. Boehnke
Professor Margit Burmeister
Professor Timothy D. Johnson
Assistant Professor Xiaoquan W. Wen

Dedication

To my parents and Stephanie

Acknowledgments

I would like to thank many people including my advisor, my committee members, friends, colleagues, and my family. Without their guidance and support, this dissertation is never made possible.

First, I would like to express my deepest gratitude to my advisor, Dr. Sebastian Zöllner, who introduced me to the world of statistical genetics, led me to confront interesting problems, and offered me excellent guidance. His rigorous statistical thinking, broad knowledge about human genetics, and his tolerance and patience have helped me a lot to enter this challenging yet exciting field of study.

I also owe my gratitude to my committee members, for their support and guidance during my Ph.D. study: I would like to say thank you to Dr. Boehnke, who is always willing to help and provides me great advice on my research. I really benefit a lot from his profound knowledge in statistical genetics. I would also thank Dr. Burmeister. She gave me so much valuable instructions about human genetics from more biological perspectives, so that I could understand the scientific problems more accurately, and fit into this area of study smoothly. I would also like to express my gratefulness to Dr. Timothy Johnson, for his inspiring guidance about Bayesian modeling and Markov chain Monte Carlo. I would also thank Dr. Xiaoquan Wen. I benefit a lot in every meeting with him, especially on topics about population genetics and hidden Markov models in human genetic studies.

I also owe my gratitude to Dr. Paul Scheet, whose experience in population genetics and clustering methods enlighten me a lot in my research about fine mapping based on the coalescence model. Dr. Scheet kindly provided me with his computational tools, which enabled my coalescence-based method applicable in large-scale studies. During every meeting with him, he is always willing to help. He is so open-minded to discuss about his method, and is so kind that he adjust his own computational tool to implement his clustering method into my study more smoothly.

Next, I would like to thank my parents and my beloved Stephanie. No matter at my good or bad times, they never stopped encouraging and supporting me in pursuing my academic goal. I feel so fortunate to have them in my life.

Then, I would like to say thank you to every present and former member in the Zöllner lab, the Center for Statistical Genetics, and the Department of Biostatistics, especially Shyam Gopalakrishnan, Ming Hu, Youna Hu, Bingshan Li, Yancy Lo, Mark Reppell, Carlo Sidore, Jincheng Shen, Matt Zawistowski, Xiaowei Zhan. All the interaction and discussion that I have got with people here will be the most irreplaceable and most valuable memory in my life.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ARG: Ancestral recombination graph

CNV: Copy number variation

GWAS: Genome-wide association studies

HMM: Hidden Markov model

INDEL: Insertion or deletion

LD: Linkage disequilibrium

MAF: Minor allele frequency

MCMC: Markov chain Monte Carlo

OR: Odds ratio

PCA: Principal component analysis

QTL: Quantitative trait loci

ROI: Region of interest

SNP: Single nucleotide polymorphism

Abstract

MARKOV CHAIN MONTE CARLO IN GENETICS: SUBPHENOTYPING,
LINKAGE DISEQUILIBRIUM MODELING, AND FINE MAPPING

by

Ziqian Geng

Chair: Sebastian Zöllner

The advance of modern genotyping and sequencing technologies makes large scale data available in different genetic studies. Meanwhile, MCMC algorithm provides powerful computational tools in handling these high-dimensional genetic data. In this dissertation, I demonstrate several MCMC applications in emerging genetic studies.

In Chapter 2, I propose a method to identify genetically homogeneous subphenotypes of complex diseases. I assume that different disease subtypes, caused by different risk variants, behave uniquely in clinical characteristics (treated as covariates). I design an algorithm to identify these covariates to define genetically homogeneous subtypes. Conditional on these covariates, this algorithm calculate each affected individual's posterior probability of belonging to each subtype. Using simulated data, I illustrate that my algorithm correctly identifies subtypes, such that affected individuals within each subtype group are likely to carry the same risk variants. I also evaluate whether

stratifying on these estimated subtype memberships improves the power to detect phenotypic association at risk loci attributable to these subtypes.

In Chapter 3, I introduce a novel algorithm to model the linkage disequilibrium (LD) between different genomic positions through shared genealogies. Compared to traditional hidden Markov models (HMM) which might over simplify the evolutionary process of sampled haplotypes, my method allows for more variations in prior probabilities about shared haplotype segments descend from particular ancestors, as well as more variations in population genetic parameters. Through this more careful model, our method improves the accuracy in haplotype reconstruction. Moreover, I propose a fine mapping algorithm based on this model to localize complex trait loci. My algorithm identifies disease causal loci accurately when traditional mapping approaches based on single marker tests have low power.

In Chapter 4, I propose an approach to overcome the computational burden in fine mappings using our coalescent-based modeling. I first estimate a set of clusters of sampled haplotypes such that members within each cluster share one common ancestor. I then make inferences about genealogies of these clusters to localize candidate regions of disease-causing mutations. Using simulated data, I illustrate that this implementation enables my fine mapping approach in large samples with several tens of thousands of individuals.

Chapter 1 Introduction

In recent years, genetics studies involve high dimensional data analysis. In a typical genome-wide association study (GWAS), millions of variant sites are genotyped and analyzed to detect candidate regions associated to certain phenotypes of interest. Along with the development in next-generation sequencing technology, genetic data to be analyzed contains even higher dimensions of variables. Consequently, many statistics of interest in modern genetic studies cannot be calculated analytically.

To approximate these statistics of interest from high-dimensional data, Markov chain Monte Carlo (MCMC), as a powerful computational tool in Bayesian modeling, is widely applied in modern genetic studies. Conditional on the observed phenotypes and genotypes, as well as the family structures of sampled pedigrees, MCMC algorithms pick a large number of random draws from the Markov chain to approximate the posterior distribution of statistics to be analyzed. During the late 1990s and early 2000s, several MCMC approaches have been developed in linkage studies to explore different perspectives of genetic questions, such as what individual in large pedigrees should be genotyped and analyzed (e.g., Fernández et al. 2001; Tier and Henshall 2001), and what loci are attributive to the quantitative trait of interest (e.g., Heath et al. 1997; Uimari and Hoeschele 1997; Daw el al. 1999). More recently, MCMC algorithms have also been applied to association studies using population genotype data of unrelated individuals. Multiple MCMC-based approaches have been developed to model the linkage disequilibrium (LD) in genomic regions and to make inferences about population genetic parameters (Nielsen 2000; Fearnhead and Donnelly 2001; Morris et al. 2003).

Although promising, traditional MCMC approaches require large amount of computations to approximate the posterior probability of interest. To reduce this computational intensity, Li and Stephens have developed a method using hidden Markov models (HMM) to approximate LD and to identify recombination hotspots (Li and Stephens 2003), which has been proved an computationally efficient model. In recent years, similar HMM approaches are widely applied to other LD-based studies, such as haplotype reconstruction (e.g., Stephens and Donnelly 2003; Scheet and Stephens 2006) and genotype imputation (e.g., Marchini et al. 2007; Browning and Yu 2009; Li et al. 2010; Howie et al. 2012).

In this Chapter, we give an overview of MCMC methods in subphenotyping for complex diseases, as well as in modeling linkage disequilibrium and in mapping of complex trait loci.

1.1 Sampling genetically homogeneous subphenotypes

Genetic heterogeneity is one of the major challenges for identifying complex trait loci. In many cases, we observe multiple genetic variants affecting the same phenotype – each with a relatively small effect size. Accordingly, the power to detect every single variant's association to the phenotype is reduced when affected individuals carrying different risk variants are collected into one sample of cases. Indeed, creating genetically more homogeneous subtypes of cases, would increase some risk allele frequencies in these subtype samples, and thus improve the power to detect phenotypic association at relevant risk loci.

Since the underlying risk variant is priori unknown, subtype strategies are valid only if non-genetic indicators for the underlying genetic variants are available. Several studies have hypothesized that different risk variants affect the same disease endpoint through different pathways, such that they create differences in clinical characteristics (Burmeister et al. 2008; Saunders et al. 2008; Saunders et al. 2009). These clinical behaviors, correlated with different risk variants, are then treated as covariates to define different subtypes of the disease in many studies.

While defined subphenotypes can increase the risk variant frequency within each subphenotype sample, they also reduce the sample size of affected individuals within each subtype sample. Hence, the power gain from genetic homogeneity must outweigh the power loss from sample size reduction (Burmeister et al. 2008; Morris et al. 2010).

Various subphenotype approaches have been implemented in recent years. Most commonly, subphenotypes are defined by stratifying affected individuals on single observable traits (e.g., Zandi et al. 2007; Lydall et al. 2009). However, single covariate stratification is less efficient when genetic variants weakly affect many covariates. To define subphenotypes by analyzing multiple covariates, different approaches have been proposed in recent years. Most of these approaches are either based on known comorbidity information (e.g., Schulze et al. 2006; Morris et al. 2010) or pre-selected SNPs (e.g., Wen and Lu 2013); they are thus not generalizable to other complex disease studies.

In Chapter 2, we propose a novel approach to define genetically homogeneous sample of subphenotypes that are generalizable to any complex disease. This approach analyzes the segregation in a large set of covariates, and identifies subsets of co-inherited covariates attributable to the same risk variants to identify sub-samples of cases more likely to carry the same risk variants.

1.2 Modeling of linkage disequilibrium

By definition, LD is the dependence structure of alleles carried at different positions in the genome. Modeling of LD is crucial in many aspects of genetic studies. In practice, LD provides valuable information about which alleles, at different loci, are inherited together. In recent years, many widely used software packages, such as PHASE (Stephens et al. 2001; Stephens and Donnelly 2003) or MaCH (Li et al. 2010), are modeling LD based on an HMM algorithm (Li and Stephens 2003): At each genomic position, the observed state is the genotypes and the latent state is which ancient haplotype in the target population is the observed haplotype descends from.. Conditional on the hidden state, the observed genotypes can be modeled as imperfect copies with mutations at each site. The algorithm then models the whole genomic region site-by-site as a Markov process, where the transition probability between neighboring sites is modeled through recombination rates.

Although successfully applied in many studies, the HMM algorithm oversimplify the latent inheritance pattern. Typically, it assumes uniform prior probabilities that each sampled haplotypes are equally likely to be descendant of any haplotypes within the same target population. In genomic regions where too many mutation and recombination events have resulted in much complicated inheritance patterns, we observe reduced performance when applying these oversimplified HMM algorithms. Likewise, we also observe reduced performance when applying HMM algorithms to individuals sampled from admixed populations.

Alternatively, several studies have focused on making inferences about the evolutionary history of sampled individuals, to model LD through the shared genealogy within short genomic segments that are inherited together (e.g., Zöllner and Pritchard 2005; Burkett et al. 2013b). Compared to the HMM algorithm which assume uniform prior probabilities of inheritance from a "pool of haplotypes", these coalescent-based methods quantify the sampled haplotypes' probability of descending from different ancestors based on the estimated local ancestry of sampled individuals. Moreover, coalescent-based methods analyze different positions in parallel, thus allowing more variability in mutation rates and recombination rates across genomic regions of interests. For each co-inherited short segments in the studied genomic region, these methods provide more careful local LD modeling.

However, coalescent-based methods are computationally extremely intensive when making inferences about genealogies. To reduce computational intensity, existing coalescent-based methods typically assume first or second-degree Markov model to approximate the construction of sampled haplotypes. This simplification ignores the LD between more distant loci, and is thus not applicable in studies using high density SNPs.

In Chapter 3, we propose a more careful model of LD, which combines the local genealogy from coalescent-based model with the more distant modeling of haplotypes using HMM. To demonstrate that our algorithm provides more accurate model of LD than traditional methods, we apply our algorithm to an LD-related study, haplotype reconstruction. According to our simulation studies, haplotypes are more accurately reconstructed based our LD modeling than based on traditional HMM algorithms.

1.3 Mapping of complex trait loci

According to the Catalog of Published Genome-Wide Association Studies (http://www.genome.gov/gwasstudies), GWAS have identified over 1,500 genomic regions associated with hundreds of phenotypes. However, most of variants found by GWAS are located in regions that lack obvious functional consequences. These results are not surprising, as the idea of GWAS is based on the assumption that at least some detected associations are due to high LD between the tested SNPs and the non-genotyped true causal variant site, to identify a set of candidate regions for follow-up studies about functionalities.

This ambiguity in detected associations can be partly solved by whole-exome sequencing studies, where the association at true causal site is directly tested. However, the causal site is not always genotyped even using sequencing data: it can be filtered due to low quality in low coverage sequencing; or the causal mutation may be not a variant but an indel or a copy-number variation (CNV), or it may be a non-coding functional variant. Moreover, many studies are based on imputation-based analysis to detect rare to low-frequent variant associations (Servin and Stephens 2007; Guan and Stephens 2008; Li et al. 2010). Although promising, rare variant association studies still face the challenge that some rare causal variant may be incorrectly imputed when it is not observed in the reference panel for imputation. In all these situations where causal variant is non-genotyped, traditional approaches will have reduced power to detect the phenotypic association at the causal loci. And the follow-up fine-scale mapping can be biased.

Alternatively, several studies have focused on mapping candidate regions through estimating the evolutionary history of causal mutations (Morris et al. 2002; Zöllner and Pritchard 2005; Tachmazidou et al. 2007, Burkett et al. 2013a). Based on the coalescence theory, these methods analyze the genealogy between cases and controls position by position, to locate candidate regions harboring disease causal mutations. These methods jointly analyze genotypes at surrounding sites to make inferences about the genealogy. They are thus more powerful in studies where traditional approaches based on single marker tests have low power, and are more robust to the scenario when genotype is missing at a single causal position.

In Chapter 3, we improve traditional coalescent-based model of LD by an HMM algorithm such that it can be applied to studies using high density SNPs. We then propose a fine-mapping approach based on this HMM-enhanced model to locate complex trait loci. Using simulated data, we demonstrate that our method correctly identifies candidate regions harboring causal variant sites. We also illustrate that our algorithm performs more robustly when genotypes at causal sites are missing, compared to traditional fine mapping algorithms based on initial sets of candidate regions with GWAS-detected associations.

In Chapter 4, we design a novel coalescent-based algorithm for fine mapping, which utilizes a clustering method (Scheet and Stephens 2006) to reduce the sampling space of genealogy inferences. Using simulated data, we illustrate that our implemented algorithm correctly identifies credible regions harboring causal variants in large-scale studies with several thousands of sequenced individuals.

1.4 Outline of dissertation

In this dissertation, we describe several MCMC algorithms that address different statistical genetic challenges. This dissertation is organized as follows: In Chapter 1, we give an overview about our dissertation. In Chapter 2, we describe our algorithm which identifies affected individuals carrying the same risk variants through analyzing the segregation of a large set of observable covariates. In Chapter 3, we focus on our HMM-enhanced coalescent-based modeling of LD, which improves the accuracy of haplotype reconstruction and can be applied to correctly mapping candidate loci of complex traits. In Chapter 4, we implement a clustering method to our coalescent-based model, such that our new method is capable in large-scale studies to identify candidate regions for complex trait loci. In the end, we provide a summary of our studies in Chapter 5.

Chapter 2 A covariate based affected sib-pair endophenotype sampler in association studies

2.1 Introduction

Recent years have seen substantial progress in understanding the genetic basis of common complex diseases. Combining studies to achieve massive sample sizes (Thornton and McPeek 2010; Wen and Stephens 2010) reveals that common risk variants typically have very low effect sizes and each variant explains only a small proportion of the total heritability in many diseases. Identifying further genetic factors for complex traits remains a major challenge in human genetics (Hirschhorn and Daly 2005; McCarthy et al. 2008), which will require even larger sample sizes. Meanwhile, rare variants are now targeted as promising candidates for the missing heritability not explained by common variants. However, rare variant studies also require very large sample sizes to be adequately powered (Zöllner 2012; Nelson et al. 2012).

The large sample size requirement can be at least partly explained by the genetic heterogeneity of the studied phenotype. In many cases, we observe multiple genetic variants affecting the same phenotype – each with a relatively small effect size. Accordingly, the power to detect every single variant's association to the phenotype is small when patients carrying different risk variants are collected into one sample of cases. Indeed, creating genetically more homogeneous sample of cases, would increase some risk variants' allele frequencies in this sample, and thus improve the power of detecting these variants in association studies. As the underlying risk variant is priori unknown, such stratification is only possible if non-genetic indicators for the underlying genotypes are available. Such indicators can be identified if different risk variants affect the same disease endpoint through different pathways, and thus create differences in clinical characteristics (clinical heterogeneity). These clinical characteristics (treated as covariates) are then correlated with different risk variants (Burmeister et al. 2008; Saunders et al. 2008). Properly picked covariates will then contain information about the underlying genetic heterogeneity, and may define genetically more homogeneous subtypes of the disease.

We borrow the endophenotype approaches suggested by many psychiatric studies, to identify genetically more homogeneous subtypes of a disease (Gottesman and Gould 2003). While endophenotype approaches can increase the risk variant frequencies within cases with the same subtype, it also reduces the sample size of cases to the subset of cases showing this subtype. Hence, the power gain from genetic homogeneity must outweigh the power loss from sample size reduction (Burmeister et al. 2008; Morris et al. 2010).

Various endophenotype approaches have been applied to study different psychiatric diseases and other complex diseases (Gottesman and Gould 2003; Flint and Munafo 2007). Most commonly, this idea is applied by stratifying affected individuals on single observable traits (single covariate stratification). For example, stratification by age of onset showed the relation between certain genotypes and early onset Alzheimer's disease is stronger than between the same genotypes and late onset Alzheimer's disease (Dizier et al. 2012; Velez et al. 2013). Likewise, several studies for bipolar (BP) disorders stratified affected individuals on comorbid symptoms (Willour et al. 2007; Zandi et al. 2007;

Lydall et al. 2009; Saunders et al. 2009; Kerner et al. 2011), to identify variants associated with BP that are not reported before. In general, single covariate stratification is most powerful if only one genetic variant has a strong effect on the covariate. It is less efficient if genetic variants weakly affect many covariates, because stratifying on one covariate now insufficiently increases the allele frequency of any risk variant. In this latter situation, stratifying on multiple covariates simultaneously may identify individuals carrying specific risk variants. However, doing so will divide the whole sample of affected individuals into many small subsets (Burmeister et al. 2008; Morris et al. 2010). Such quick sample size reduction will result in an unrecoverable power loss. Therefore, an applicable approach to analyze multiple covariates for endophenotypes, which does not generate a hard limit on the number of individuals within each subtype, becomes a necessity.

In recent years, several advanced approaches based on multiple covariates have been developed to define endophenotypes for complex diseases. For instance, Schulze et al (2003) proposed a mixed-effects regression analysis on BP with 40 phenotypic features. Andreassen et al. (2013) increased the power to detect common variants associated with schizophrenia by leveraging comorbidity risk factors of cardiovascular diseases using a stratified FDR method. And Morris et al. (2010) proposed a multinomial regression framework to allow for genetic heterogeneity between subtypes. However, these methods assume known comorbid and pleiotropic information to pre-select a few covariates. They are not generalizable to studies of other diseases where comorbidity is not well studied. Alternatively, Wen and Lu (2013) recently proposed a method which does not rely on known comorbidity, but focuses on selecting the optimum number of subtypes contained in sampled cases. However, their model selection is based on selected SNPs. It is only applicable to confirm known candidate loci rather than exploiting new gene-phenotype association. Finally, principal component analysis (PCA) is also applied to define endophenotypes in complex diseases (Wang et al. 2009). However, it is difficult to distinguish whether the PC-represented subtypes are truly driven by underlying genetic variants, or are merely artifacts due to some highly correlated, non-heritable factors.

In present study, we propose an algorithm based on the similar intuition as PCA: we assume that each subtype, represented by differences in covariates, is attributable to the constellation of one or more genetic variants. We then study the correlation structure between the set of covariates to identify affected individuals belonging to these subtypes in the study sample. Unlike PCA, our method identifies heritable subtypes by modeling the heritability in related individuals – if truly heritable due to genetic variants, covariates representing these subtypes are more likely to be shared within each family than between unrelated individuals (Freimer and Sabatti 2004). Indeed, we model this sharing pattern of covariates to calculate every affected individual's posterior probability of being affected by each subtype, conditional on the set of covariates. We use this posterior probability to prioritize individuals carrying the same risk variants. We implement our algorithm for sibling-pairs – a Covariate-based Affected Sib-pair Endophenotype Sampler (CASES).

To test CASES, we simulated sib-pair datasets in which at least one sibling is affected by a dichotomous phenotype. For a wide range of disease models, we simulated one or more risk variants which also associated with different sets of covariates. We analyzed these covariates using CASES to identify subtypes among the cases and to evaluate if the subtypes represent the underlying genotypes at the risk loci.

Our results show that our algorithm successfully identifies the informative subset of covariates from other covariates not affected by the underlying risk variants. Based on these covariate subsets, we successfully identify individuals carrying the same endophenotype, who are more likely to be carrying the same risk variants. We then illustrate that using these endophenotypes is helpful in guiding the sampling of affected individuals carrying specific risk variants. The precision of endophenotype identification increases with the increased effect sizes and increased number of co-inherited covariates. Moreover, we are able to use this defined endophenotypes to improve the power in association studies.

2.2 Methods

We propose an algorithm that estimates each affected individual's probability of carrying any subtype of the disease. We model each subtype as attributable to a set of risk variants which also affect a set of distinctive covariates.

We consider a sample of n families; each family contains at least one affected individual. Moreover, every family member is measured for a set of J covariates in dataset C. We use K as the unknown number of subtypes and G to denote the unknown genotypes at the risk loci affecting different subtypes. We will discuss the selection of K later. The risk allele frequency p, variant effect sizes β, and the between-covariate effects γ are also unknown and to be estimated in our algorithm (Table 2.1)

2.2.1 Data structure

We observe J covariates in dataset C for each individual. These covariates can be either dichotomous or continuous, and are observed in dataset C. We denote $C_i = [C_i^{(1)}, \dots C_i^{(S)}]$ to represent the vector of covariates in family i of size S. Let $C_{i0}^{(s)}$ represent the affection status (phenotype) of family member s in family i, where $C_{i0}^{(1)} = 1$ for all $i = 1,2, \dots n$ since we condition on at least one affected individual per family. We use $C_{ij}^{(s)}$ to denote the observed covariate j ($j = 1,2, \dots J$).

We model K risk variants affecting K subtypes. For family i, let $G_i = [G_i^{(1)}, \dots G_i^{(S)}]$ denote the unknown genotypes at these risk loci in G, where $G_i^{(s)} = [G_{i1}^{(s)}, \dots G_{iK}^{(s)}]^T$ are the underlying genotypes in family member s ($s = 1,2, \dots S$).

2.2.2 Models

Ultimately, our goal is to estimate each affected individual's probability of belonging to each subtype. Under our assumption, this is equivalent as estimate the probability of carrying each risk variant, as in $P(G|C)$. By Bayes rule, it is proportional to

$$P(G|C) \propto P(C|G) \cdot P(G)$$

First, we specify the covariates model $P(C|G)$ and model the co-inheritance pattern in covariates. We assume multiplicative models. For each member s in family i, the phenotype probability follows logistic models:

$$logit\left(P\left(C_{i0}^{(s)} \middle| G_i, C_{i(-0)}^{(s)}\right)\right) = \beta_0^T G_i^{(s)} + \gamma_0^T C_{i(-0)}^{(s)},$$

where $\beta_0 = [\beta_{00}, \beta_{01}, \dots \beta_{0K}]^T$ are risk variant effect sizes on the studied phenotype and β_{00} is the baseline phenotype risk. Similarly, $\gamma_0 = [\gamma_{01}, \dots \gamma_{0J}]^T$ are the other covariates' effects on the phenotype, which models the between-covariate correlation not explained by genetic variants.

We assume similar multiplicative models for all covariates. Let β_j and γ_j denote the risk variants' and other covariates' effect sizes on covariate j with baseline level β_{j0}. We use logistic regression to model any dichotomous covariate:

$$logit\left(P\left(C_{ij}^{(s)}\middle|G_i, C_{i(-j)}^{(s)}\right)\right) = \beta_j^T G_j^{(s)} + \gamma_j^T C_{i(-j)}^{(s)}.$$

For continuous covariates, we substitute linear regression models. Overall, the joint probability of phenotypes and covariates in family i, given their genotypes, can be written as:

$$P(C_i|G_i) = \prod\nolimits_j \prod\nolimits_s [P(C_{ij}^{(s)}|G_i^{(s)}, C_{i(-j)}^{(s)})]$$

To calculate $P(G)$, we define $p = [p_1, p_2, \dots p_K]^T$ as the unknown risk allele frequencies in the target population. We use p to calculate the joint genotype probability $P(G_i|p)$ within each family i, based on the relatedness and genotype transmission probabilities between members in this family.

2.2.3 Algorithm

By Bayes rule, the joint posterior probability of genotypes given covariates in all families is proportional to the total likelihood:

$$P(G|C) \propto \prod_i [P(C_i|G_i, \beta, \gamma) \cdot P(G_i|p) \cdot P(p) \cdot P(\beta) \cdot P(\gamma)].$$

Specifically, we derive $P(C_i|G_i, \beta, \gamma)$ from the previous multiplicative model. Given allele frequencies, the marginal genotype probability in each family, $P(G_i|p)$, can be derived through the kinship coefficient and genotype transmission probabilities between family members.

Furthermore, $P(p)$, $P(\beta)$, and $P(\gamma)$ are prior probabilities of parameters. We choose uniform priors on p. As we expect that many enrolled covariates are not informative about underlying risk variants, many effect sizes in β should be 0. Indeed, we specify the spike and slab priors for any effect size β_{jk} as suggested by Ishwaran and Rao (2005) and other related studies. We also apply the spike-and-slab priors on γ, as the prior probability of between-covariate correlation unexplained by genetic variants.

In the present study, we implement our algorithm for sib-pairs in our own software (CASES). Accordingly, we model $P(G_i|p)$ by integrating over possible identity by descent (IBD) status in each sib-pair i (Appendix A).

Since $P(G|C)$ is of high dimensions ($n \times K$), we use MCMC to approximate this posterior probability. Specifically, we use Metropolis-coupled MCMC to resolve the potentially multi-modal posterior distribution in Markov chains. And we use Gelman and Rubin diagnostics (1992) to assess the MCMC convergence (see Appendix A for spike and slab priors and MCMC details).

2.2.4 Simulated data

For each family, we simulate parental genotypes at K loci based on risk allele frequency p, assuming Hardy-Weinberg equilibrium. We then generate one offspring genotype and assign a phenotype based on this genotype and the risk variant effect sizes in β_0. We discard this family if the offspring is unaffected. For families with affected offspring, we generate a second offspring genotype and phenotype. We repeat this process until we have generated n sib-pairs, each with at least one affected offspring. Given β and γ, we then assign J dichotomous covariates to these n affected individuals and their siblings.

We then mask all genotypes, and use the phenotypes and covariates as the input dataset into our program CASES to generate the estimated posterior probability $P(G|C)$ for these n affected individuals.

2.2.5 Algorithm evaluation

To evaluate if the estimated $P(G|C)$ from CASES is informative in identifying affected individuals carrying the same risk variants, we compare $P(G|C)$ to the true genotypes G. We first rank all affected individuals based on $P(G|C)$, and compare if individuals carrying genetic risk variants are ranked higher than individuals not carrying these variants. We further evaluate the benefit of our algorithm in association studies. Each time, we stratify the top-half ranked cases in one vector in $P(G|C)$, to formulate a sample of cases of one subtype. We compare this sample of cases to a random sample of n controls from the same population in association tests. At the specific locus attributable to each selected subtype, we compute the testing power using our stratified sub-sample of cases under significance level $= 5 \times 10^{-4}$ where the power using all n affected individuals is lower than 0.1.

To evaluate if the subsets of covariates identified by CASES can define genetically homogeneous subtypes in other samples of cases from the same population, we design a two-step simulation study:

First, we simulate a sample of n sib-pairs exactly the same as in the previous subsection. We then apply CASES to this sib-pair data, to draw a large number (M) of effect size estimates $\beta^{(m)}$ and $\gamma^{(m)}$.

Second, we simulate a second sample of N unrelated affected individuals, based on the same effect sizes as in our first sample of sib-pairs. Conditional on their observed covariates, we calculate each individual's probability of belonging to each subtype based on $\beta^{(m)}$ and $\gamma^{(m)}$ in step one by:

$$\hat{P}(G_i|C_i) = \sum\nolimits_m P\big(G_i\big|C_i, \beta^{(m)}, \gamma^{(m)}\big)$$

We then stratify these unrelated affected individuals in this second sample, based on the estimated $\hat{P}(G_i|C_i)$, to formulate sub-samples of cases. Comparing each sub-sample of cases to a sample of N controls from the same population, we perform association tests at the loci attributable to this subtype. Compared to the traditional association tests using all N affected individuals as cases, we evaluate if using sub-samples defined by CASES improves the power in association studies.

2.3 Results

To evaluate our algorithm CASES' ability to identify heritable subtypes, we simulate samples of sib-pairs with at least one affected sibling. For each individual, we simulate a set of dichotomous covariates affected by unobserved underlying risk variants. We apply our program CASES to these covariates to estimate each affected individual's posterior probability of belonging to each subtype. We then stratify cases based on this posterior probability to evaluate if we can identify affected individuals carrying underlying risk variants attributable to this subtype.

We initially assume that all covariates are affected by underlying risk variants, and different variants are affecting distinctive covariate subsets. Also, we assume known number of subtypes K. We will relax these assumptions and consider more complicated scenarios in the second half of section 2.3.2 and section 2.3.3.

2.3.1 Discriminating carriers and non-carriers

We evaluate the resolution of CASES by comparing the estimated probabilities of subtype membership to the true genotyeps at risk loci. Initially, we consider the single locus model with risk allele frequency 5% and disease OR 1.5. We analyze 200 cases with siblings considering 3, 5, 10, or 15 covariates affected by the risk variant, wich effect sizes varying between 1.0 and 2.0.

We rank all 200 cases by their estimated probability of carrying the specific subtype. Based on the true genotypes, we compare the mean ranks between cases that carry either one or two copies of the risk allele (carriers) and cases that are homozygous with the wild type allele (non-carriers). In all simulated scenarios with covariate effect size greater than 1.0, carriers have a higher mean rank than non-carriers (figure 2.1A). This difference in mean rank increases with more covariates affected by the same variant. Our algorithm also better differentiates carriers from non-carriers with an increasing effect size of the risk variant on covariates. Overall, including 3 covariates each with an OR of 1.8 has about the same effect as including 15 covariates each with an OR of 1.2.

We also calculated the observed risk allele frequency in top ranked individuals identified by CASES (figure 2.1B). For all models, the risk allele frequency decreases from higher ranked individuals to lower ranked individuals. Moreover, the ability to identify carriers depends on the number of covariates affected by the same risk variant. Specifically, when we have more than 15 covariates, more than one third of all carriers within the sample are identified within the top 10% ranked affected individuals. If we stratify on the top 10% cases, we increase the risk allele frequency within the sample by more than 3 times. When we stratify the top half of cases, almost two thirds of all carriers are sampled even with only 5 covariates in our model. With increasing stratification size, the observed allele frequency reduces towards the expected allele frequency in cases (7.3%), as more non-carriers are sampled. In general, CASES correctly identifies carriers conditional on co-inherited covariates, which provides an efficient guidance in re-sampling affected individuals carrying rare risk variants.

To illustrate our algorithm performance when more than one risk variants are affecting different covariates, we simulated scenarios with 2 risk loci. We analyzed the resulting dataset using a model of $K = 2$ subtypes. As the assignment of clusters to the underlying risk variant changes between realizations, we evaluated the performance in multiple realizations by comparing the mean pair-wised distances in the space of inferred genotype probabilities. On average, we observe shorter mean pair-wised distance in both carriers of risk loci 1 and 2, rather than the overall mean pair-wised distance among all sampled individuals. This indicates that true carriers at each risk loci can be identified correctly as they have similar inferred genotype probability (see Appendix B for details).

2.3.2 Power improvement in sib-pairs

To evaluate the power gain from our algorithm, we simulate datasets of $n = 500$ cases with siblings each, under different parameter settings with 5, 10, or 15 covariates affected by one risk variant, and with effect sizes on each covariate ranging from 1.0 to 2.0. We analyze the covariate data with CASES and stratify $n/2$ cases with the highest estimated probability of carrying the subtype. Using this stratified sample of cases, we perform association tests on the risk locus, and calculate the power under significance level $\alpha = 5 \times 10^{-4}$. We compare our testing power to association tests using all n cases (figure 2.2).

When no covariate is affected by the risk variant (covariate OR $= 1.0$), the power using stratified subtype of cases is constantly lower than using all cases, reflecting the loss of power due to a reduced sample of half cases. With increasing number of co-inherited covariates and increasing covariate OR, the ability to identify cases carrying the underlying risk variants is strengthened, thus we observe improved power to detect the association when using these stratified cases. Specifically, the power using stratified cases is higher than the power using all cases when analyzing 10 affected covariates with moderate effect sizes (OR < 1.3) or 5 covariates with greater effect sizes (OR < 1.5).

When analyzing 15 covariates strongly affected by the underlying risk variant (OR > 1.8), using our stratified cases enriched the power for even more than ten times, compared to the baseline power using all cases.

So far we assume that all covariates analyzed in our model are affected by the risk variant. In practice, these informative covariates are unknown and most enrolled covariates may be non-informative in identifying subtypes. To evaluate how mixing informative and non-informative covariates would affect the performance of CASES, we consider scenarios where we analyze up to 50 covariates in our model while only 5 covariates are affected by the underlying risk variant. We analyze the dataset blind to the knowledge about which covariates are affected (figure 2.3). Not surprisingly, the difference between using only informative covariates (red dashed line) and using both informative and non-informative covariates is decreasing when the effect size on each informative covariate increases. However, including non-informative covariates only have a small effect on the ability to identify risk variant carriers. Even when only 10 percent (5 out of 50) covariates in our model are truly informative (black dashed line), the power using stratified subtype of cases still outperforms the power using all cases when the risk variant affect each of the 5 informative covariates with OR ≥ 1.6. In general, our algorithm distinguishes non-heritable factors, such that the testing power is mostly driven by covariates truly affected by the heritable risk variants.

We also evaluate the performance of our algorithm for a range of different effect sizes (Appendix C). Most importantly, genetic variants may only affect the observed covariates, but not the phenotype (i.e., not true risk variant). Indeed, we want to verify that our algorithm does not stratify individuals carrying such variants, which can otherwise increase type I error rate in association studies. We apply our algorithm on simulated datasets where the genetic variant has phenotypic OR equals 1.0 to evaluate the type I error rate using our stratified samples. We observe an error rate fluctuating around the nominal significance level.

2.3.3 Models with more than one variants affecting the same covariates

So far we present results based on the assumption that different subtypes are represented by different sets of covariates, where each of these mutually exclusive sets of covariates is affected by one risk variant. To evaluate if more than one risk variants affecting the same set of covariates would affect our algorithm performance, we consider scenarios where 10 covariates are affected by risk variant 1, while 0, 5, or 10 out of these 10 covariates are also affected by another risk variant 2. As in practice, the number of subtypes in unknown, we compare two models with K = 1 and K = 2, and evaluate the power to detect phenotypic association at variant site 1 for each model (Appendix D).

In general, we observe increased power to detect the association at variant site 1 under both $K = 1$ and $K = 2$ models. More importantly, the selection of K results in different algorithm performance: The $K = 1$ model achieves higher power than the $K = 2$ model when 0 or 10 out of the 10 covariates affected by risk variant 1 are also affected by variant 2; while $K = 2$ model outperforms when 5 out of 10 covariates are affected by both variants (Appendix D). In practice, specifying the “best” value of K improves the efficiency of our algorithm. We suggest a sequential procedure for model selection of K (Appendix E).

2.3.4 Power improvement in unrelated individuals

To evaluate if CASES provide valid effect size estimates from sib-pair data, we simulate multiple scenarios with different true effect sizes and different sample sizes of sib-pairs, and compare our estimates from CASES to the true parameters by the mean squared errors (MSE) (figure 2.4). In general, our effect size estimates are more accurate when the underlying risk variant has stronger effect on a covariate, or when we have larger sample sizes of sib-pairs. Specifically, the estimation based on 200 sib-pairs requires the covariate strongly affected by the underlying risk variant with $\mathrm{OR} > 1.8$, to achieve the same accuracy as the estimation based on 500 sib-pairs when a covariate is only moderately affected by the risk variant ($\mathrm{OR} = 1.2$)

Consequently, our algorithm defines heritable disease subtypes based on these identified models of covariates in sib-pairs. To evaluate if these identified indicators (sets of covariates) for each subtype can be generalized and applied to other samples of unrelated affected individuals from the same target population, we simulate two samples, a model simulation sample of sib-pairs and a testing sample of unrelated affected individuals.

In the first step of this analysis, we apply CASES to the sib-pair dataset to estimate the underlying risk variants' effect sizes on covariates.

In the second step, given these effect size estimates from sib-pairs, we calculate any individual's probability of subtype membership, in the sample of unrelated affected individuals, conditional on the same set of covariates. We then stratify these unrelated affected individuals based on their probabilities of subtype membership.

To evaluate the benefit using these stratified subsamples of cases, we consider datasets of 2,000 unrelated affected individuals under the same scenarios as in the sib-pair dataset, and stratify these unrelated affected individuals with the top-half estimated probabilities of belonging to one subtype. We then compute the power in association studies using this stratified sample of cases under genome-wide significance level ($\alpha = 5 \times 10^{-8}$), compared to the power using all 2,000 cases (figure 2.5) for different effect sizes on covariates. Specifically, the power using all cases (dotted line) is constant since it does not depend on information provided by covariates. When no covariate is truly affected by the risk variant (covariates OR $= 1.0$), the power of using the stratified sample is smaller than the power using all cases due to the reduced sample size. Similar to the power in sib-pair datasets, testing power using our stratified sample of cases increases with the number of covariates, and with effect sizes on covariates. If only 5 covariates are affected by the risk variant, using our stratified sample provides higher power than using all cases when the variant has moderate (OR $\approx$ 1.3) effect sizes on these covariates. If we have more than 10 covariates affected by the same risk variant, using stratified cases will lead to a much higher testing power even if the effect on each covariate is as weak as less than OR ≤ 1.2.

2.4 Discussion

We have presented a novel algorithm to identify heritable subtypes of a complex disease. We model clusters of covariates that are co-inherited with the studied phenotype within families to define subtypes attributable to one or more underlying risk variants. We are thus able to use these estimated subtype memberships select subsamples of affected individuals that are genetically more homogeneous. Focusing on the selected sub-sample of cases can increase the power in association tests between the subtype-specific risk loci and the phenotype of interests. Furthermore, our algorithm provides estimates about the risk variant's effect sizes on different covariates. Such estimates can be applied to calculate the posterior probability of subtype membership in other samples from the same target population, given the same set of observed covariates. Indeed, we are able to generalize our estimations to unrelated affected individuals, to define genetically more homogeneous subtypes of the disease.

Compared to traditional endophenotype approaches in complex diseases; our algorithm has at least four advantages. First, we model variation within and between sib-pairs, thus providing evidences that subtypes reflected in covariates are truly heritable. In contrast, methods using traditional mixed models (Schulze et al. 2006) and principal component studies (Wang et al. 2009) do not distinguish if the resulting subtypes are due to genetic or non-genetic factors.

Second, our algorithm identifies informative subsets from a large set of covariates for endophenotypes, before knowing any genotype information. Thus, it is more generalizable to any complex disease of interest, and may discover novel regions associated with the study phenotype. In comparison, existing methods (Andreassen et al. 2013; Wen and Lu 2013) require genotypes in a set of previously reported SNPs to select co-inherited covariates, while other methods require prior pleiotropic knowledge to define subtypes (Morris et al. 2010; Andreassen et al. 2013).

Third, our algorithm obtains a quantitative definition of the modeled subtypes. This subtype definition can then be applied to other samples of unrelated affected individuals from the same target population. We are able to stratify unrelated affected individuals based on their subtype membership, to create genetically more homogeneous sample of cases. Based on our simulation results, using these sub-samples as cases will increase the power to detect associations under a wide range of scenarios.

Fourth, our method is tolerant to missing data as such data can be imputed in MCMC. When an individual has covariates missing at random, the posterior probability of subtypes can be inferred from the same mechanism, such that it is equivalent to the marginal probability conditional on the remaining covariates.

Our algorithm is of interests in different studies. First, it identifies individuals carrying specific risk variants. It thus can be applied to re-sequencing studies to prioritize affected individuals more likely to carry risk variants. Second, it can be applied to association studies to select the sample of cases caused by the same risk variants, to increase the testing power through increased risk allele frequencies in the case group. When diagnosis is based on large sets of clinical measurements and questionnaires, our method fits naturally as such measurements can be used directly as observed covariates to define endophenotypes to identify different genetic basis.

In summary, we have developed a general framework for defining endophenotypes and selecting genetically homogeneous subsamples of affected individuals. Comparing to other approaches, our algorithm is more flexible and can be used to improve the power of single marker tests in GWAS studies. It can also be applied to prioritize affected individuals in re-sequencing studies such that affected individuals from the same genetic origin of the disease are more likely to be sampled.

Table 2.1 Model parameters in CASES.

Notation	**Parameters**
n	Number of sib-pairs
C	Observed covariates
J	Number of covariates
G	Underlying genotypes at risk loci
K	Number of risk loci (number of subtypes)
p	Vector of risk allele frequencies
β	Matrix of risk allele effect sizes on disease and covariates
γ	Matrix of between covariates effect sizes

Figure 2.1 Figure A: Mean rank comparison between carriers and non-carriers when using 3, 5, 10, or 15 covariates. The horizontal axis represents effect sizes on covariates ranging from 1.0 to 2.0. The vertical axis represents mean rank differences. Figure B: Risk allele frequency in stratified cases. The horizontal axis represents stratification sizes, from 20 to 200. The vertical axis is the observed risk allele frequency in sampled cases. Results are shown for a sample of size 200 and risk allele frequency of 5%, the effect size to the disease is 1.5.

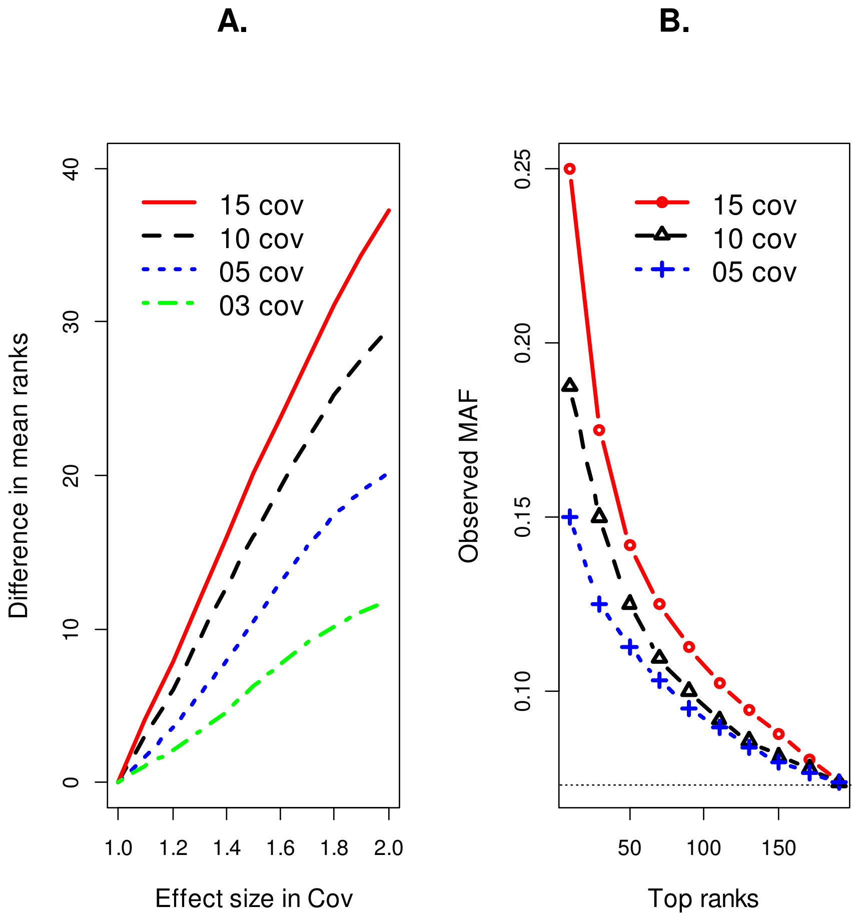


Figure 2.2 Power comparison in association tests between using full samples (dotted horizontal line) and using stratified samples. The horizontal axis represents effect sizes on covariates ranging from 1.0 to 2.0. The vertical axis shows testing power under significance level $\alpha = 5 \times 10^{-4}$. The red, green, and black solid curves indicate the power using stratified subtypes when 15, 10, and 5 covariates are in our model.

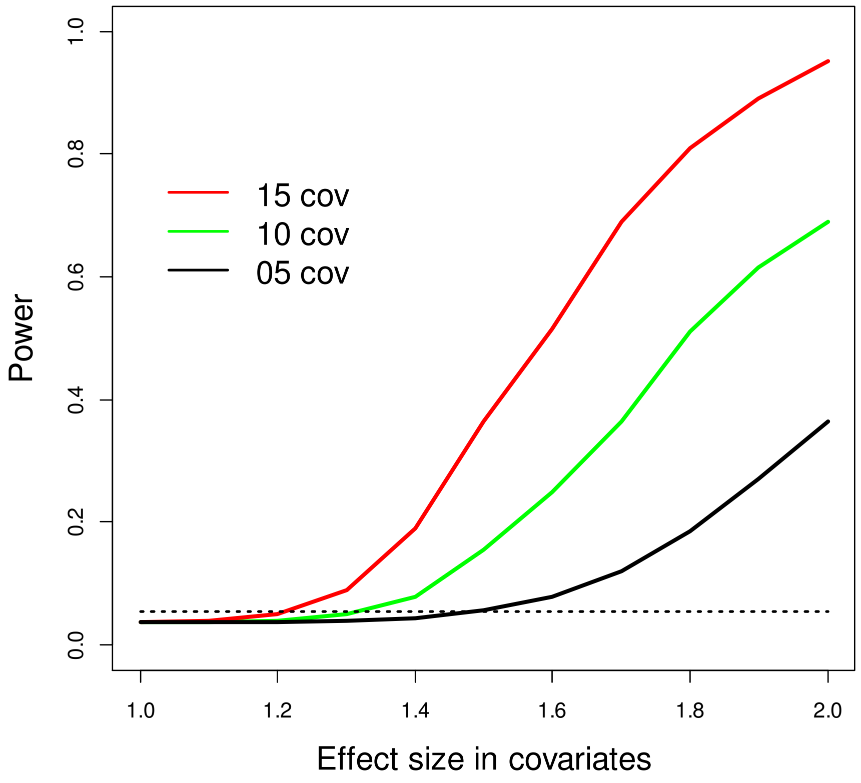


Figure 2.3 Power comparison in association tests between using full samples (dotted horizontal line) and using stratified samples based on 5 informative covariates. The horizontal axis represents effect sizes on covariates ranging from 1.0 to 2.0. The vertical axis shows testing power under significance level $\alpha = 5 \times 10^{-4}$. The red, blue, brown, green, and black dashed curves indicate the power using stratified subtypes when 0, 5, 10, 20, and 45 non-informative covariates are analyzed jointly with 5 informative covariates.

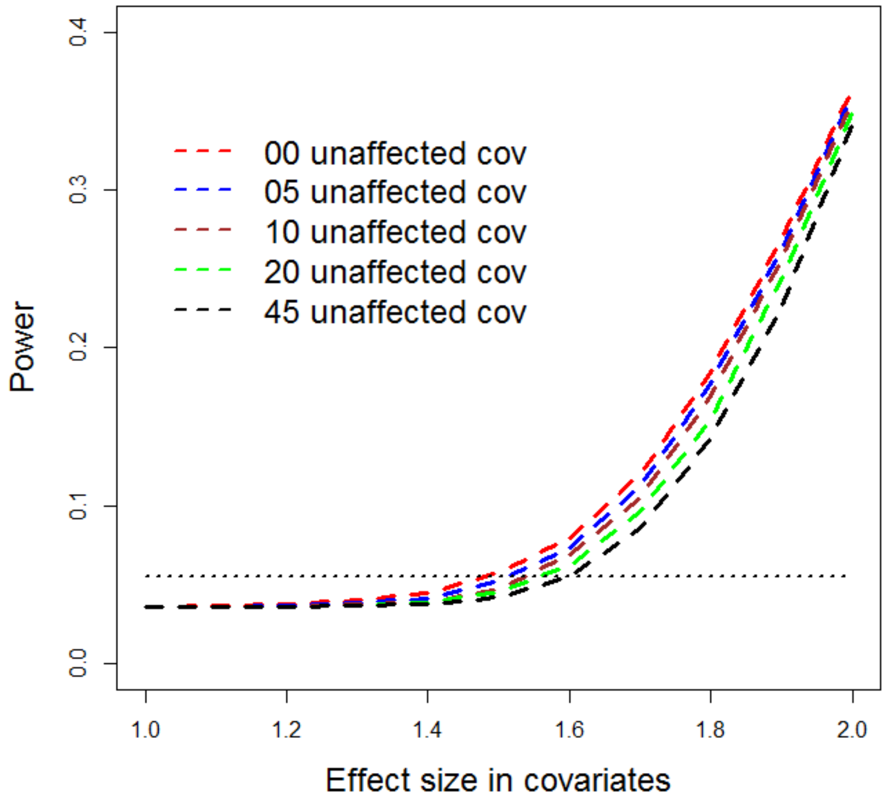


Figure 2.4 Mean squared error (MSE) comparing the risk variant's estimated effect sizes and the true value. The vertical axis is the mean squared error. The horizontal axis represents the increased sample size. The three red curves are the MSE of the effect size estimates on covariates, where the true values are 1.2, 1.5, and 1.8. The black curve is the effect sizes on phenotype whose true value is also 1.5. The estimated effect on the phenotype shows the same accuracy as the estimated effect on covariates when true effects are of same level.

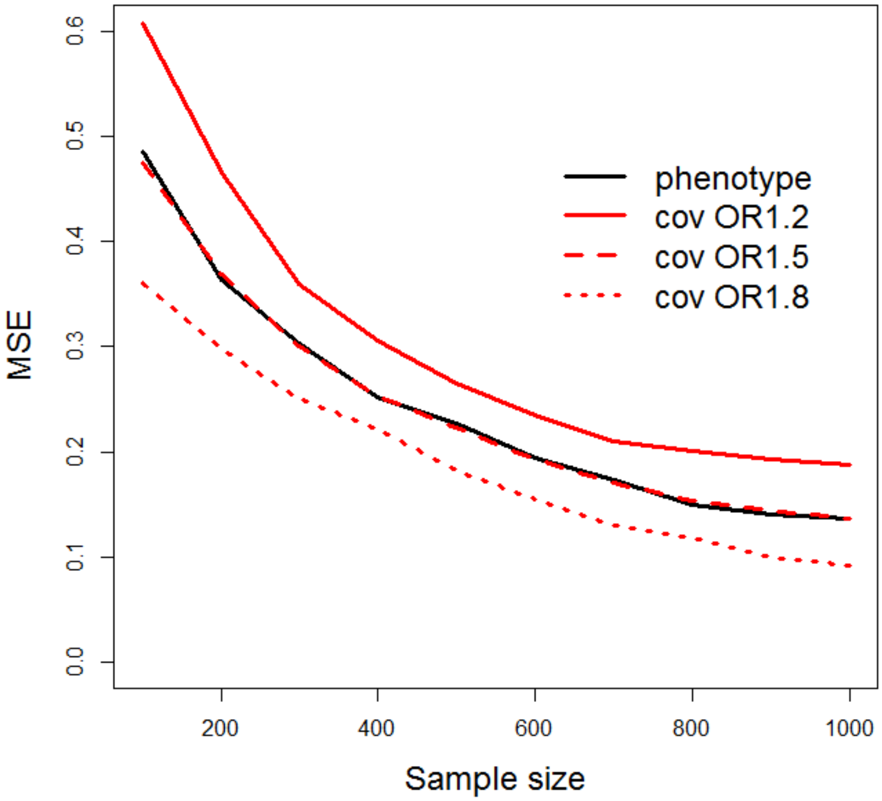
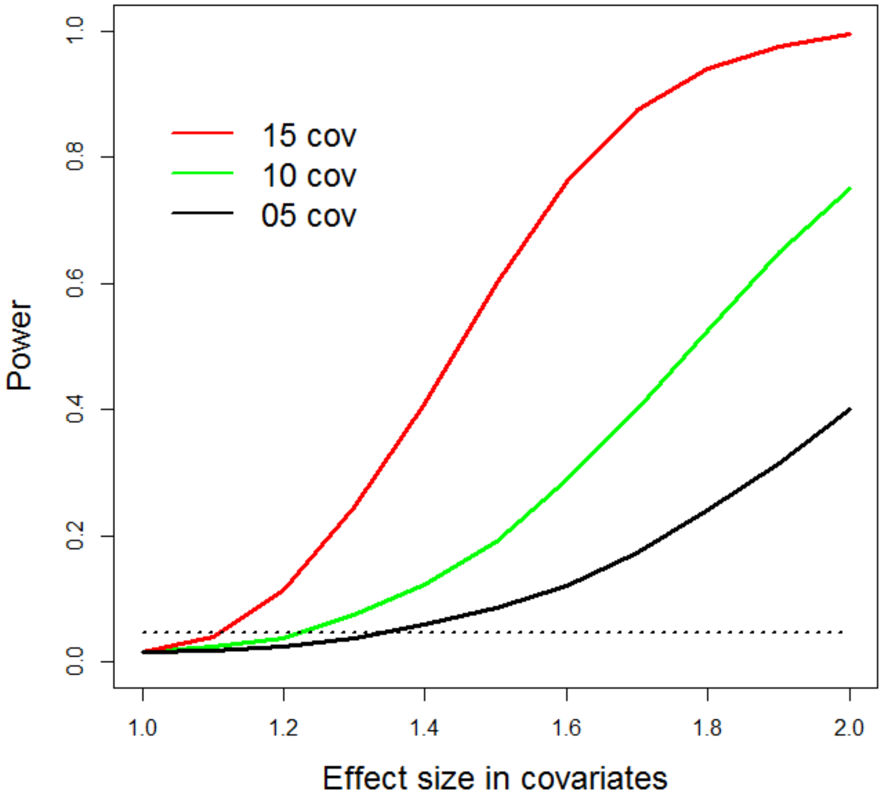


Figure 2.5 Power curves in external datasets of 2,000 unrelated cases and 2,000 controls. The horizontal axis represents effect sizes on covariates ranging from 1.0 to 2.0; and the vertical axis shows testing power under genome-wide significance level ($\alpha = 5 \times 10^{-8}$). The power using all 2,000 cases is fixed at 0.045 and is represented by the black dashed line. The red, green, and black solid lines indicate power curves using stratified samles when 15, 10, and 5 covariates are enrolled and affected by the risk variant.



Chapter 3 Modeling linkage disequilibrium for population genotype data using a coalescent-based method

3.1 Introduction

In population genetics, linkage disequilibrium (LD) represents the non-random association structure of two or more alleles, at different loci, which descend together from the same ancestral chromosomes. Modeling of LD, which approximates the evolutionary history of sampled haplotypes, is crucial in many aspects of genetic studies such as haplotype reconstruction (Stephens et al. 2001; Stephens and Donnelly 2003), genotype imputation (e.g., Howie et al. 2009; Li et al. 2010; Howie et al. 2012), variant calling (e.g., Li et al. 2011; Bizon et al. 2014), and mapping of complex trait loci (Wellcome Trust Case Control Consortium 2012; Edwards et al. 2013).

Traditionally, LD is modeled sequentially across the genomic region of interest (ROI). For instance, several studies have modeled LD in neighboring SNPs as a Markov process when estimating recombination rates and other population genetic parameters (Nielsen 2000; Fearnhead and Donnelly 2001). More recently, Li and Stephens (2003) have proposed a hidden Markov model (HMM) to approximate LD across different loci. Such HMM algorithm has been proved an efficient approach, and is widely implemented in many existing software packages for genetic studies to model LD of sampled haplotypes, such as PHASE (Stephens et al. 2001; Stephens and Donnelly 2003), BEAGLE (Browning and Yu 2009), MaCH (Li et al. 2010) or THUNDER (Li et al. 2011).

However, the HMM algorithm oversimplifies the variation of mutation and recombination rates across ROI – especially if the ROI is highly polymorphic and/or has high recombination rates, or if sampled individuals are from an admixed population. Potentially, this over-simplification can be reduced by implementing larger reference panels from the same or similar target populations as a larger "pool of haplotypes," to achieve higher accuracy in reconstructing sampled haplotypes (Huang et al. 2009; Huang et al. 2011). Most studies performed to date have selected reference panels from external databases such as the International HapMap Project (The International HapMap Consortium 2005; Frazer et al. 2007) or the 1000 Genomes Project (The 1000 Genomes Project Consortium 2010). Following this path, several studies have also proposed other strategies to select more efficient reference panels (e.g., Jewett et al. 2009; Pasaniuc et al. 2010; Liu et al. 2013) through local ancestry inferences (e.g., RFMix; Maples et al. 2013) to improve the HMM-based model of LD. Unfortunately, HMM algorithms still face challenges in modeling LD in certain complex genomic regions lacking efficient reference panels, such as HLA genes in the human MHC region (Cao et al. 2013).

As an alternative approach, when no efficient reference panel is available, several studies have proposed coalescent-based methods to improve the modeling of LD through the inferred ancestral recombination graph (ARG; Nordborg 2000; Nordborg and Tavare 2002). At each position across the ROI, these methods estimate the genealogy of sampled haplotypes by analyzing the genotype similarity at surrounding variant sites. LD is then modeled through shared and non-shared genealogies between different loci. Because genealogies are analyzed in parallel position-by-position, these coalescent-based methods allow more variation in LD across the ROI, compared to HMM algorithms.

Although promising, the coalescent-based model of LD is only fulfilled in a few studies under specific assumptions, due to its computational complexity. For instance, Burkett et al. (2013b) proposed a computational tool based on Zöllner and Pritchard (2005) to make inferences about genealogies using unphased genotype data, and compared the accuracy between their inferences to the haplotypes estimated by HMM algorithms to evaluate their model. However, their method assumes the same first-order Markov process as in Zöllner and Pritchard (2005), which ignores the LD between more distant loci to simplify its computation. Therefore, these methods are not applicable in studies using dense SNP data or using sequencing data.

In the present study, we implement an HMM-enhanced algorithm to coalescent-based modeling of LD, which combines the local LD information from shared genealogies with the more distant LD approximated by the HMM algorithm. Based on this more careful modeling, we are able to apply this coalescent-based algorithm to estimate LD using densely genotyped or sequenced data.

To evaluate the accuracy of our model, we focus on two applications highly related to the LD modeling: reconstructing haplotypes and mapping of complex trait loci. In simulation studies, we compare our results in these two applications to results using traditional HMM-based algorithms, to determine whether our new model improves the understanding of sampled haplotypes.

3.1.1 Haplotype reconstruction

Traditionally, HMM-based algorithms have reconstructed haplotypes through an MCMC, by approximating each haplotype as a mosaic of short segments imperfectly copied from one or more ancestors in the target population. Without any local ancestry information, traditional HMM algorithms (e.g., PHASE; Stephens et al. 2001; Stephens and Donnelly 2003) assume uniform prior probability of being copied from any specific ancestor haplotype.

As an alternative in the present study, we apply more informative prior probability of ancestries for each haplotype segment conditional on the inferred genealogy through our modeling of LD. Thus, we are able to update the inferred haplotypes more efficiently.

We set up simulation studies across a wide range of population genetic models to evaluate our method. Specifically, we reconstruct haplotypes from these simulated diploid samples through our inferred genealogies, and compare them to the true haplotypes to calculate the miscalling rate of the inferred haplotypes provided by our algorithm. We then compare our results to the miscalling rate of inferred haplotypes using HMM algorithms (MaCH; Li et al. 2010) to evaluate if our model of LD extracts more information about the construction of sampled haplotypes.

According to our simulation results, our algorithm provides comparable haplotyping accuracy with MaCH in sequenced regions under neutral recombination rates. In sequenced regions where the recombination rate is more than 10 times higher than the neutral rates, the inferred haplotypes based on our model of LD are more accurate than the haplotypes inferred by MaCH.

3.1.2 Mapping of complex trait loci

Traditionally, fine mapping is applied subsequent to genome-wide association studies (GWAS), to identify candidate regions harboring disease-causing mutations. Focusing on a set of GWAS-selected loci with detected phenotypic associations, fine mapping algorithms calculate the Bayes factor in favor of associations at each variant site, to calculate the posterior probability of causality for fine mapping (Servin and Stephens 2007; Stephens and Balding 2009). Typically, these methods require that functional loci to be sought must be included in the GWAS-selected set.

Alternatively in the present study, we make inferences about genealogies based on our model of LD. Conditional on these estimated genealogies, we then identify candidate regions harboring causal mutations by evaluating the distribution of cases and controls on the ARG across ROI (figure 3.1). Instead of focusing on certain GWAS-selected candidate regions, we analyze the entire ROI directly for causal loci.

To evaluate our algorithm, we simulate samples of several hundred cases and controls from the target population. Under different simulated scenarios, we compare our mapping results to the results provided by traditional fine-mapping approaches based on single marker association tests (BIMBAM; Servin and Stephens 2007).

According to our simulation results, our algorithm correctly localizes sub-regions harboring causal mutations under scenarios where single marker tests have low power to detect any association signal. Moreover, since we make inferences about genealogies through jointly analyzing genotypes at multiple sites, our algorithm performs robustly when the genotype at the causal site is missing.

3.2 Methods

3.2.1 Estimating genealogies

Suppose we have a sample of n individuals, each sequenced in a region of L Mbs containing J variant sites. We let G denote the genotypes of the $2n$ diploids in J variant sites, where $M = [m_1, \dots m_J]$ is the physical positions (in base pairs) from each variant site to the left end of the sequenced ROI. We denote T_x as the sample genealogy at a position x base pairs to the left end of the target ROI, which contains the ancestral relationship topology, time to each coalescent event, as well as the length of each sampled haplotype sharing the same local genealogy with position x.

At each position, we make inferences about the posterior genealogy probability, $P(T_x|G)$, conditional on the observed genotypes at all J variant sites. We approximate this posterior distribution $P(T_x|G)$ by an MCMC sampling method similar to Zöllner and Pritchard (2005).

To begin, we assume known haplotypes $H = [h^1, \dots h^{2n}]$ from genotypes G in n diploids and let h_j^i denotes haplotype i's allele type at variant site j. To make inferences about $P(T_x|H)$, we treat these haplotypes as the external nodes of a coalescent tree. We then let $H' = [h^{2n+1}, \dots h^{4n-1}]$ denote the unknown haplotypes of the $2n - 1$ common ancestors (treated as internal nodes) of the genealogy, where h^{4n-1} is the haplotype of all sampled individuals' most recent common ancestor (MRCA).

Without loss of generality, we assume the position x locates at the left end of the ROI ($x = 0$), such that each variant site j is located m_j based pairs to the right of position x.

In practice, for variant sites locate to the left of x, a similar algorithm can be applied easily as a mirror image. As in Zöllner and Pritchard (2005), we denote the recombination rate and mutation rate as $\rho/2$ and $\theta/2$. We let $D = [d_1, \dots d_J]$ denote the physical distances between two consecutive variant sites, such that $d_1 = m_1$ and $d_j = m_j - m_{j-1}$ for site $j = 2, \dots J$. Moreover, we denote Ω as the vector of coalescent times in all $2n - 1$ coalescent event on T_x (Table 3.1).

Moreover, we denote r_i as the variant site closest to x at haplotype i that is not inherited from parental node on T_x (due to recombination). We then use $S = [s_1, \dots s_{4n-1}]$ to denote the lengths of haplotypes descending together with position x in genealogy T_x, such that all $s_i = J$ for all observed haplotypes (external nodes) $i = 1, \dots 2n$. For any internal node P with left and right descendant nodes L and R, $s_P = max(r_L, r_R)$.

Then the posterior probability of genealogy, conditional on the sampled haplotypes, is proportional to the total likelihood:

$$P(T_x|H) = P(\tau, \Omega, H', R, \rho, \mu|H) \propto P(H, H'|R, \rho, \mu, \tau)P(R|\rho, \mu, \tau)P(\rho)P(\mu)P(\tau)$$

Similar to Zöllner and Pritchard (2005), we apply uniform prior probabilities on ρ, μ, and τ. Given these population genetic parameters, we propose possible recombination events in the MCMC using a Metropolis-Hasting algorithm to approximate $P(R|\rho, \mu, \tau)$. For the joint haplotype probability $P(H, H'|R, \rho, \mu, \tau)$, we re-write it as a series of conditional probabilities of haplotypes:

$$P(H, H'|R, \rho, \mu, \tau) = P(h^{4n-1}|R, \rho, \mu, \tau) \cdot \dots \cdot P(h^1|h^2, \dots, h^{4n-1}, R, \rho, \mu, \tau)$$

For each conditional probability term, we split this haplotype probability into two parts as follows: For any node q with parental node p, we define h_{inh}^{q} as the segment of all h_i^q with $i < r_q$, which is inherited from h_i^p. We also define h_{rec}^{q} as the segment of all h_i^q with $r_q \leq i \leq s_q$, which is copied from other haplotypes due to recombination (figure 3.2). Consequently, we model the conditional probability of h^q as follows:

$$P(h^q|h^{q+1}, \dots, h^{4n-1}, R, \rho, \mu, \tau) = P\left(h_{inh}^{q}\middle|h^p, \mu\right) \cdot P\left(h_{rec}^{q}\middle|h^{q+1}, \dots, h^{4n-1}, R, \rho, \mu, \tau\right)$$

The probability for inherited segment h_{inh}^{q} can then be modeled through mutation models. We then consider the recombined segment, h_{rec}^{q}, as a mosaic of imperfect copies from a “pool of haplotypes” in the target population, and model this segment through an HMM model (Li and Stephens 2003).

We use an MCMC to draw a large number (M) of possible coalescent trees $T_x^{(m)}$ ($m = 1, \dots M$), to approximate the posterior probability of genealogy $P(T_x|H)$, conditional on the sampled haplotypes.

3.2.2 Haplotype reconstruction

In practice, genotyping and sequencing technologies provide diploid genotypes at each variant site, but do not provide haplotype information. Given genotypes G, we propose a method to make inference about haplotypes H through our estimation of genealogies.

Let $H^{(0)}$ be the initial state of estimated haplotypes conditional on G. During our MCMC, we begin with $H^{(0)}$ to make inferences about genealogies at each position x. Suppose $T_x^{(0)}$ is the initial draw of possible genealogy conditional on $H^{(0)}$, we then propose possible updates of haplotypes $H^{(1)}$ in accordance with G, to the MCMC: We let $h^{q1(0)}$ and $h^{q2(0)}$ denote the two haplotypes of individual q in $H^{(0)}$, where $h_{inh}^{q1(0)}$ and $h_{inh}^{q2(0)}$ are the inherited haplotype segments from parental node $p1$ and $p2$ in estimated genealogy $T_x^{(0)}$. We then randomly choose a variant site $i < \min(r_{q1(0)}, r_{q2(0)})$ where the two haplotypes are ambiguous in individual q, and propose a switch of alleles at h_i^{q1} and h_i^{q2} based on a Metropolis-Hasting algorithm to update the haplotypes. We update these haplotypes within the Markov chain that makes inferences about genealogies, such that the marginal probability of haplotypes in short segments surrounding x, $P(H_x|G)$, can also be approximated by the random draws from the MCMC. We sum over these random draws from $P(H_x|G)$ at each analyzed position x, and map the draws at different positions as a mosaic to reconstruct the entire haplotype within ROI.

3.2.3 Mapping of complex trait loci

Traditionally, fine mapping is applied to an initial set of candidate loci reported by association studies, which is based on the assumption that at least one causal variant must be carried by this initial set of loci with reported phenotypic association. As an alternative, we propose an algorithm mapping causal variants through our estimated genealogies across the entire ROI, which is not based on this assumption.

We consider a study with n cases and n controls, each sequenced in an ROI of L Mbs. We let Φ denote the observed phenotypes, where $\varphi_i = 1$ if individual i is an affected individual (case), and $\varphi_i = 0$ if otherwise. We assume there is one causal variant within this ROI, and let X denote the indicator of the causal variant, such that $X = x$ represents that a causal mutation is carried at the position x base pairs to the left end of the ROI. We then make inferences about the genealogies of cases and controls at each position x, to map candidate regions harboring disease causal variants.

Theoretically, we could make inferences about the sample genealogy T_x at all L positions within the ROI. However, this is unnecessary since positions extremely close to each other are highly likely to share the same genealogy. Moreover, we are not able to distinguish the genealogies between two closely located positions, where no variant site is observed and genotyped in between. Instead, we borrow the idea of "focal points" suggested by Zöllner and Pritchard (2005) as the set of positions where we make genealogy inferences. Specifically, we estimate the genealogy at one position in every short genomic segment of 10Kbs.

At each focal point x, our goal is to calculate the posterior probability of causality $P(X = x|\Phi, G)$. By Bayes rule, this is proportional to:

$$P(X = x|\Phi, G) \propto P(\Phi, G|X = x)P(X = x).$$

The analyzed positions, the "focal points", are uniformly selected in every 10Kbs. Therefore, we assume they are equally likely to be harboring the causal mutation, and apply uniform prior probability of causality $P(X = x)$ to each focal point x. We then calculate the joint probability of phenotypes and genotypes, $P(\Phi, G|X = x)$, similar to Zöllner and Pritchard (2005):

$$P(\Phi, G|X = x) = \int P(\Phi|T_x, X = x)P(T_x|G)dT_x$$

We approximate $P(T_x|G)$, the posterior probability of genealogies given genotypes, by sampling possible genealogies from our MCMC described above. We then approximate the probability above by:

$$P(\Phi, G|X = x) = \frac{1}{M}\sum\nolimits_{m=1}^{M} P(\Phi|T_x^{(m)}, X = x)$$

We assume a multiplicative model, and denote p_0 and p_1 as the phenotype penetrance level on any individual carrying homozygous wild type alleles and variant type alleles. Therefore, the joint probability of phenotypes and genotypes is equivalent to:

$$P\left(\Phi\middle|T_x^{(m)}, X = x\right) = \iint_{p_0 \neq p_1} P\left(\Phi\middle|T_x^{(m)}, p_0, p_1\right) dp_0 dp_1$$

We then evaluate $P(\Phi|T_x^{(m)}, p_0, p_1)$ by integrating over possible mutations on any branch of the tree $T_x^{(m)}$ as Zöllner and Pritchard (2005).

In the end, we apply Monte Carlo integration over all focal points x, to estimate the posterior probability of causality $P(X = x|\Phi, G)$ at each position, and build a credible region to fine map candidate loci harboring causal variants.

3.2.4 Algorithm evaluation

To evaluate our algorithm performance in haplotype reconstruction, we simulate datasets containing 400 haplotypes using a coalescent haplotype simulator (Hudson 2002). Each haplotype is sequenced in a 1Mbs with 500Kbs flanking spaces on both ends. All singletons are removed from the data since these single observations provide no information of LD. We apply neutral mutation rate, and consider different recombination rates ranging from 1×10^{-8} to 2×10^{-7}. We then randomly assign these haplotypes into 200 diploid individuals. We mask the haplotype information, and apply the diploid genotypes into our algorithm to make inferences about the sample genealogies.

We then apply both our method and a HMM algorithm (MaCH; Li et al. 2010) to these genotype data, and compare the consequent estimated most likely haplotypes to the true haplotypes using both algorithms. For each simulated scenario, we categorize variant sites into 4 bins minor allele frequency levels: less than 1%, from 1% to 5 $\%$, from 5% to 20%, and greater than 20%. We compare our reconstructed haplotypes to the true haplotypes in each simulated dataset, and calculate the ratio of the number of alleles mislabeled into the opposite haplotype over the total number of heterozygous sites as the miscalling rate. We calculate this miscalling rate of our haplotype estimation within each category, and compare it to the same rate calculated from haplotype reconstruction using MaCH, to evaluate if our algorithm reconstructs haplotypes more accurately.

To evaluate our algorithm performance in mapping of complex trait loci, we simulate a dataset of $2N$ of haplotypes, and randomly assign these haplotypes into N diploid individuals. Each individual is sequenced in a 1Mbs ROI with 500Kbs flanking region on both ends. We then randomly pick one variant site within ROI as the causal variant under different minor allele frequency (MAF) levels. We consider 3 MAF levels with different risk allele frequency levels, 1%, 5%, or 20%, to evaluate our algorithm performance. We then assign a phenotype to each individual conditional on his genotype at this causal site under a multiplicative model. The risk allele effect size is OR $=$ 2.0 for the rare risk allele with MAF $=$ 1%. We set OR $=$ 1.5 for risk allele with MAF $=$ 5%, and OR =1.2 for the common risk allele with MAF $= 20\%$. Under each scenario, we randomly sample

400 cases and 400 controls, and apply these sampled individuals into our algorithm to make inferences about genealogies. Conditional on the estimated genealogy, we calculate the posterior probability of causality across the ROI, and construct 90% Bayesian credible regions based on the top ranked focal points to evaluate the mapping accuracy using our algorithm. Under each scenario, we simulate 5 datasets, and calculate the median width of our credible regions, as well as the distance between causal site and the position with greatest posterior probability of causality, to evaluate our mapping accuracy. We also calculate the coverage probability to determine if the credible region provided by our algorithm correctly locates causal sites.

To evaluate if our algorithm provide comparable results regardless of the causal site is genotyped or not, we mask the genotype at causal site in each simulated dataset, and re-apply our method to calculate the same statistics for fine mappings. We compare our results with and without genotypes at causal sites, to evaluate our algorithm robustness.

3.3 Results

3.3.1 Reconstructing haplotypes

To evaluate if our algorithm improves the accuracy in haplotype reconstruction, we simulate multiple datasets under different recombination rates, ranging from a neutral rate of 1×10^{-8} to as high as twenty times of the neutral level (2×10^{-7}). Under each recombination level, we simulate datasets each containing 200 diploid individuals sequenced in an ROI of 1Mbs. We also include 500Kbs flanking regions on both sides of the ROI. We then categorize the variant sites within ROI into 4 MAF levels: smaller than 1%, from 1% to 5%, from 5% to 20%, and greater than 20% (figure 3.3).

We then compare the most likely haplotype estimates using our method to the true simulated haplotypes. Under each MAF level, we calculate the rate of miscalled variant sites over the total number of heterozygous sites within our sample provided by our algorithm, and compare it to the same miscalling rate provided by MaCH (Li et al. 2010).

Using both our method and MaCH, we observe that haplotype reconstruction is more accurate with lower recombination rates, using both our method and using MaCH. This is reasonable since higher rates result in more recombination events along the genomic region, cutting the ROI into much smaller segments where each segment contains less LD information to the neighboring loci. When recombination rate is less than 2×10^{-8} (black and green lines), both our method and MaCH provide comparable accuracy where less than 2 percent heterozygous alleles are mislabeled in the inferred haplotypes. Under low recombination rates, we also observe lowest haplotyping accuracy in rare allele sites with MAF less than 1%. Specifically, using our method slightly improves the accuracy in these sites where variant alleles are only observed a few times in our sample.

When the recombination rate is 5 to 20 times higher, both methods experience reduced accuracy. Using MaCH, almost 18 percent of all heterozygous alleles are mislabeled into the opposite inferred haplotype when the recombination rate is 2×10^{-7} (red solid line). In contrast, using our coalescent-based algorithm correctly identifies around 1/5 of these mislabeled sites, and reduces the overall miscalling rate to 14 percent across all variant sites (red dashed line). Moreover, such improvement is strongest in rare allele sites (MAF

< 1%), where miscalling rate is reduced by 1/3, and is below 10 percent in these variant sites where variant alleles are only observed 2 to 4 times in our sample.

3.3.2 Mapping complex trait loci

To evaluate if our algorithm can be applied to fine mapping for candidate region harboring disease causal variants, we simulate datasets with 400 cases and 400 controls under different risk allele models. We consider rare to common risk alleles with different effect sizes as described in the methods section.

For a small sample of a few hundred individuals, traditional single marker tests are not sufficiently powered to detect the phenotypic association from either rare variants with MAF less than 1% or common variants with moderate to weak effect sizes (OR less than 1.2). To begin with, we illustrate a single realizations from our simulated datasets, for such low-powered results in single marker tests when the risk variant is rare with 1% MAF and disease effect size 2.0 (figure 3.5). We apply both Fisher's exact test (PLINK; Purcell et al. 2007) and Bayesian approaches (BIMBAM; Servin and Stephens 2007) to detect single marker associations, and compare the consequent p-values (figure 3.5A, black dots) and posterior probability of associations (PPA; figure 3.5B, grey dots). Both methods provide similar results that we observe no signal associated to the disease.

When association studies have low power, loci harboring causal variants are hardly reported as candidate regions with phenotypic associations. If so, traditional fine mapping approaches are rarely performed since no signal has been found in the target region, and are not applicable to localize sub-regions for causal variants; In contrast, our algorithm of fine mapping does not depend on the power of single marker association tests.

Specifically, we apply our algorithm to make inferences about genealogies at different positions across ROI. Based on the estimated genealogies at a set of focal points across ROI, we then calculate each focal point's posterior probability of carrying causal variants (figure 3.5B, red dotted line), and construct 90% credible regions to localize causal loci (figure 3.5B, red underlines). According to our result, the posterior probability of causality increases when moving towards the causal position, and peaks at focal points less than 20Kbs from the exact site carrying the disease risk variant. Moreover, based on our credible interval, we correctly localize the causal variant site in two sub-regions covering a region less than 350Kbs in the ROI. Compared to traditional mapping approaches, our algorithm successfully identifies cumulative weak signals around the causal region, and down-weights other noise signals from more distant loci, to correctly localize sub-regions harboring disease causal variant.

To evaluate our algorithm performance in mapping candidate regions harboring causal variants, we simulate 5 datasets for each considered risk variant type. Under each scenario, we calculate the median distance between the causal site and the focal point with the strongest estimated posterior probability of causality (round-up in 5Kbs, as the mid-points in two consecutive focal points in our analysis). Under this small sample size, the accuracy of our fine mapping algorithm is affected by the frequency of risk alleles observed in the study sample. On the one hand, when the causal variants are common, our estimated most likely position harboring these causal variants is on average less than 15 to 20Kbs from the exact position of the causal variant. On the other hand, when causal variants are rare (MAF < 1%), our estimated most likely positions is on average 30Kbs from the causal site. Across these scenarios, the estimated most likely position is consistently within 2 to 3 focal points from the exact position – given that we only make inferences at focal points in every 10Kbs in present study.

Furthermore, we construct the 90% credible region for each realization, and summarize the median credible region width per scenario (table 3.2) to evaluate our algorithm performance. We observe similar trend as in point estimates of most likely positions that the accuracy of fine mapping is determined by the frequency of risk alleles observed in study sample. When we are mapping rare risk variants with MAF less than 1%, our method suggests a region covering over 350Kbs across the ROI. Our credible region width decreases by half when the risk allele frequency is around 5%, and further reduces to less than 140Kbs when the risk allele frequency is 20% (even when a diseased effect size to the disease).

To evaluate if our algorithm provides robust mapping accuracy when the causal variant is not genotyped in sampled data, we mask the genotype at the true position in our sample of sequenced individuals. We then re-apply our algorithm to make inferences about the genealogy, and to mapping candidate regions covering causal variant site. Compared to previous results when the causal variant is genotyped, we obtain comparable mapping accuracy with credible interval widths (table 3.2) slightly increased by only a few focal points. In general, our results suggest that our fine mapping accuracy is robust to the missing genotypes at the exact causal site.

3.4 Discussion

We have presented a novel algorithm to model LD through estimating sample genealogies across ROI. At each position, we make inferences about the genealogy at one position at a time, based on the genotype similarities at surrounding variant sites co-inherited with this position. We model the LD between more distant sites using an HMM algorithm through estimated mutation and recombination models. By evaluating the genealogy at every short segment, we are then able to improve the haplotype reconstruction by treating the entire haplotype within an ROI as a mosaic of these co-inherited short segments. Moreover, by evaluating the distribution of cases and controls at genealogies in different positions across an ROI, we are also able to identify candidate regions harboring disease risk variants in case-control studies.

Compared to existing methods, our algorithm has at least three advantages. First, we make genealogy inferences at every position independently, such that we allow more variations in population genetic parameters such as mutation rates and recombination rates across different positions. In contrast, traditional HMM-based modeling of LD (Li and Stephens 2003) may over simplify the distribution of these parameters across the ROI by assuming certain prior probabilities, thus may experience challenges in studying highly polymorphic genomic regions with high recombination rates.

Second, our algorithm models both local LD and LD between more distant loci carefully: We combine the local LD through estimated genealogies with the distant LD modeling through a HMM algorithm. Thus our algorithm is applicable to studies using genotype data with dense SNPs and studies using sequencing data. In comparison, many existing coalescent-based approaches (Zöllner and Pritchard 2005; Burkett et al. 2013b) simplify the distant LD by assuming first-degree Markov models, which is only appropriate when SNPs are sparsely selected across the ROI.

Third, our algorithm makes inferences about genealogies based on unphased genotype data. Our genealogy estimation is then not dependent to pre-phasing quality. This is especially beneficial for fine mapping of complex traits loci through estimated genealogies, where some traditional mapping approaches are highly sensitive to pre-phasing accuracy (Morris et al. 2004; Burkett et al. 2013a). Moreover, allowing genealogy inferences based on unphased genotype data also improves the haplotype reconstruction in regions where traditional phasing algorithms provide less accurate reconstructed haplotypes – since each short haplotype segment co-inherited along the genealogy provides more local LD information about which alleles are descend from the same ancestor.

Our LD modeling is of interest in different types of genetic studies. First, it contributes to fine mapping studies to identify disease causal loci. When practitioners have small sample sizes, our algorithm can be especially beneficial in localizing regions harboring rare causal variant or common variants with weak effect sizes – where traditional algorithms based on single marker tests are not sufficiently powered. According to our study, fine mapping based on our genealogy inferences correctly identifies credible regions covering causal sites under multiple disease genetic architectures, and outperforms traditional algorithm when the risk variant to be sought is rare in study sample. Second, it improves the accuracy of haplotype reconstruction through more accurately modeled genealogies. It thus can be applied to studies analyzing highly polymorphic regions with high gene-densities where traditional approaches experience challenges in correctly identifying haplotype information.

Future implementations of the present study include that one may improve the accuracy in identifying recombination hotspots, as well as improve the accuracy in identifying de novo rare variants in sequencing studies, through our more careful model of LD based on estimated genealogies. Moreover, in the present study, we only select one focal point per 10Kbs to make genealogy inferences, and apply uniform prior probability of causality at every focal point across the ROI. Potentially, higher resolution of focal points can be applied to improve the power in estimating genealogies if computational time allows. Likewise, more informative priors based on gene densities and variant and more adaptive selection of focal points may also improve our mapping accuracy through up-weighting positions near causal sites. Furthermore, we evaluate our algorithm performance under scenarios where traditional single marker tests have limited power. For instance, in testing for rare variants, one may consider using collapsing methods (Burden tests, Li and Leal 2008) to jointly analyze multiple rare variants instead of applying single marker tests. In future studies, we will make further comparisons to these strategies, to evaluate under what conditions our coalescent-based algorithm is preferable.

In summary, we have developed a general framework to model linkage disequilibrium through estimated genealogy. Compared to existing approaches, our algorithm combines the advantages in HMM algorithms and coalescent-based algorithms, and allows more variation in population genetic parameters. It then improves the accuracy in haplotype reconstruction, and enables using unphased genotype data for fine mapping of complex trait loci. We also project future applications such as identifying recombination hotspots and calling de novo rare variants in sequencing studies, through this genealogy-based modeling of LD.

Table 3.1 Model parameters in estimating genealogies at position x.

Notation	Parameters
h_i^j	Haplotype i's allele type at variant site j
H	Haplotypes of all $2n$ observed haplotypes
H'	Haplotypes of $2n$-1 ancient haplotypes
s_i	Length of haplotype i restored in T_x
S	Vector of haplotype lengths
r_i	SNP closest to x not inherited to haplotype i
R	Vector of not inherited SNP boundaries
m_j	Distance of variant site j to the left end of ROI
M	Vector of physical distances to the left end of ROI
d_j	Distance of between variant site j and $j-1$
D	Vector of physical distances between SNPs
$\rho/2$	Recombination rate
$\mu/2$	Mutation rate
Ω	Vector of coalescent times
τ	Tree topology of T_x

Table 3.2 Mapping of complex trait loci using our algorithm through estimated genealogies using 400 cases and 400 controls.

MAF	**OR**	**Distance**	**CI width**	
			w. causal	**o. causal**
1%	2.0	30	350	370
5%	1.5	20	180	190
20%	1.2	15	140	150

Distance: between positions with strongest signals to causal variant sites in Kbs.

CI width: median 90% credible region width in 5 realizations in Kbs.

w. causal: when causal variant is directly genotyped in study sample

o. causal: when causal variant site is missing in genotyped data.

Figure 3.1 Intuition of mapping of complex traits through genealogy. Figure A: position extremely close to causal site with no recombination in between, sharing identical genealogy. Figure B: position slightly distant from causal site where one recombination occurred, slightly different genealogy. Figure C: position far from causal site where multiple recombination occurred, genealogy of phenotypes.

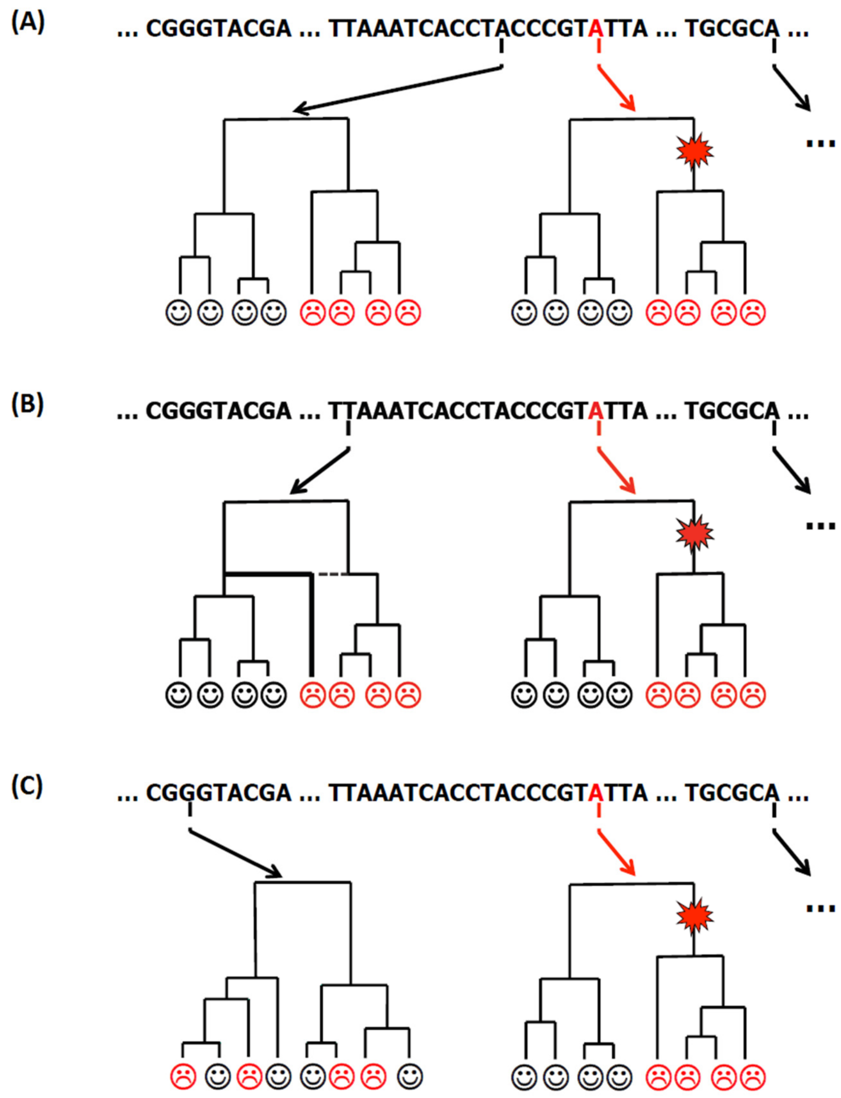

(A)
... CGGGTACGA ... TTAAATCACCTACCCGTATTA ... TGCGCA ...
...
(B)
... CGGGTACGA ... TTAAATCACCTACCCGTATTA ... TGCGCA ...
...
(C)
... CGGGTACGA ... TTAAATCACCTACCCGTATTA ... TGCGCA ...
...

Figure 3.2 Illustration of estimated genealogy at the left end of a sample of 6 haplotypes. Each sampled haplotype (external nodes) contains 9 SNPs, while the inferred ancestor haplotype (internal nodes) lengths differ due to recombination events. For each external and internal node, the segment colored in green indicates it is inherited from the parental node, while the segment in red means it is copied from other ancestors due to recombination.

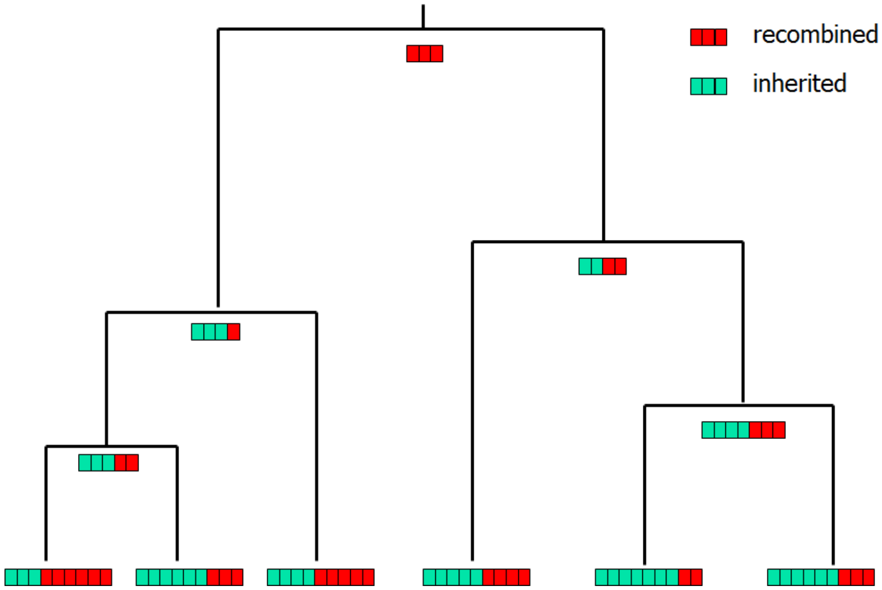


Figure 3.3 Distribution of variant sites in different MAF levels (binned by 4 categories: <1%, <5%, <20%, and <50%) across 5 simulated recombination rate levels.

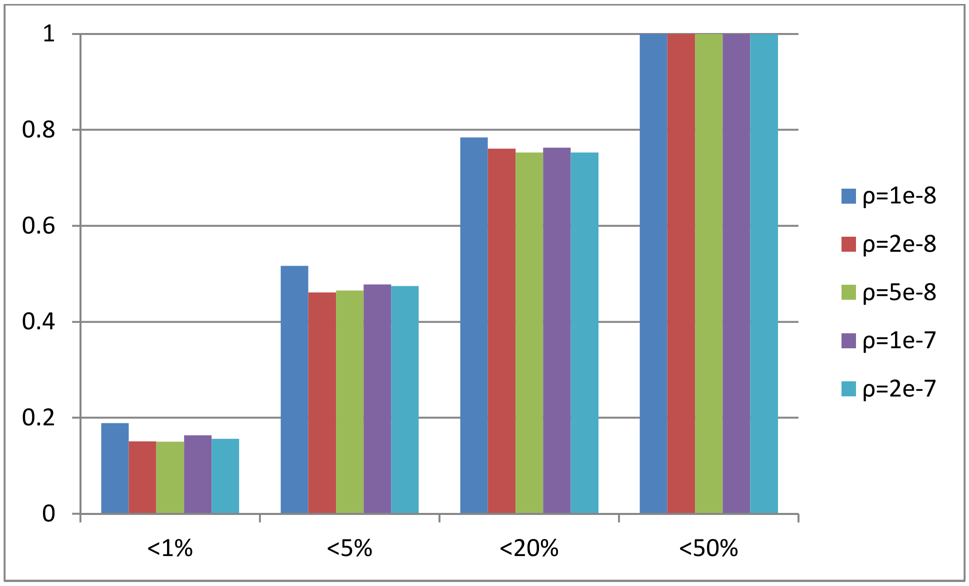


Figure 3.4 Miscalling rate in haplotype reconstruction under different simulated recombination rates, the solid lines represent miscalling rate using HMM based approaches (MaCH), and the dashed lines are miscalling rate through our inferred genealogies (Tree). The horizontal axis represents the 4 levels of MAF categories (<1%, <5%, <20%, and <50%).

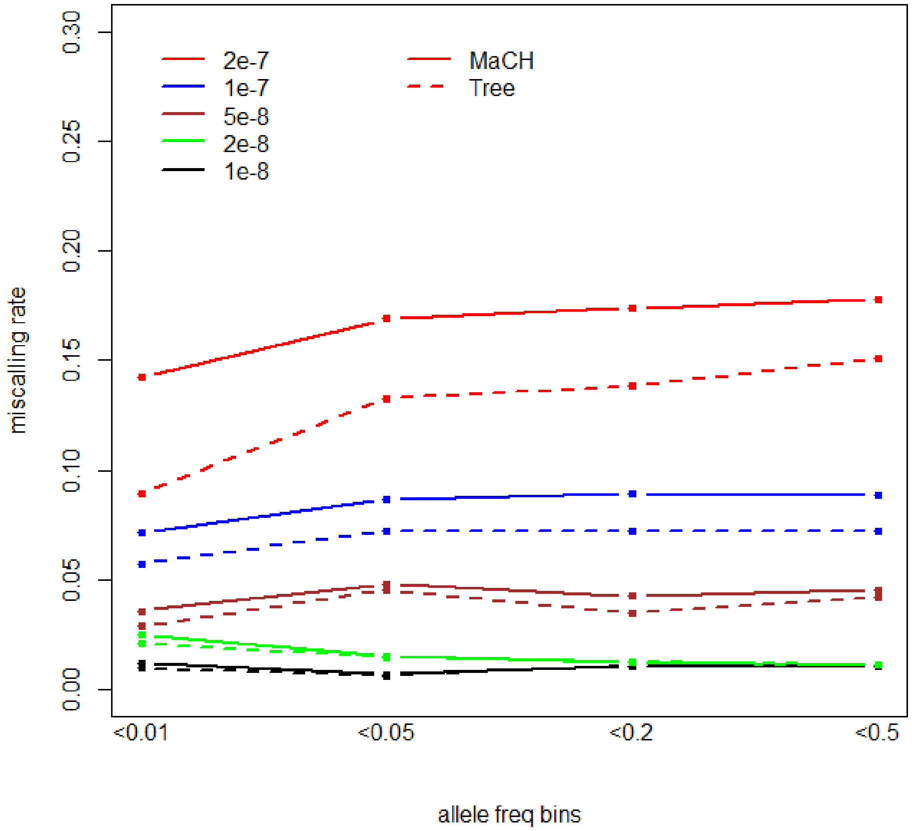
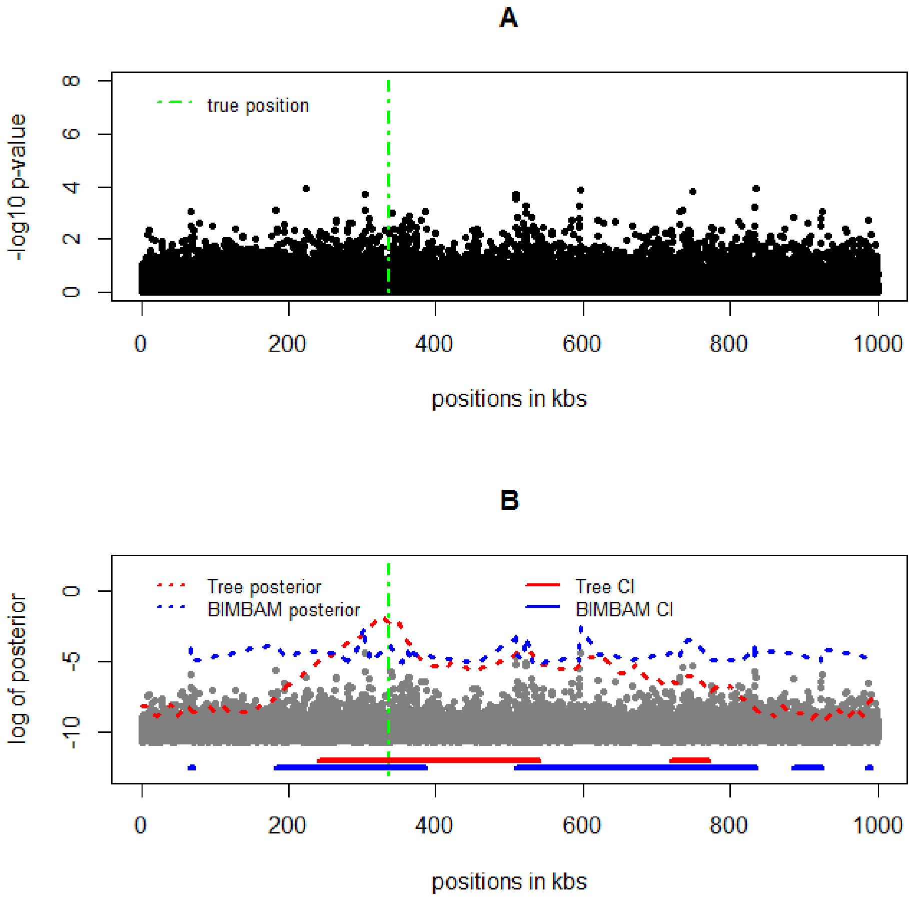


Figure 3.5 Results in identifying risk variants. The green line represents the true causal position. Black dots in figure A represent p-values using Fisher's exact tests. In figure B D, grey dots represent posterior probability of association. The red dashed lines are posterior probabilities provided by our coalescent-based algorithm, while the red and blue underlines are the 90% credible regions provided by our algorithm and traditional single marker test based fine mappings.



Chapter 4 A coalescent-based model using sequencing data in large-scale association mapping and fine mapping

4.1 Introduction

In recent years, GWAS have identified over 1,500 genomic regions associated with hundreds of phenotypes (the Catalog of Published Genome-Wide Association Studies, http://www.genome.gov/gwasstudies). Thousands of SNPs have been found associated with various complex diseases. However, approximately 88% of the variants found in GWAS are located in regions that lack obvious functional consequences. Such ambiguity leads to unclear conclusions about how these variants affect gene expression and/or the regulation of gene expression (Edwards et al. 2013). This uncertainty is in fact not surprising, since most GWAS tag-SNPs are not selected for having likely functional consequences. Rather, the interpretation of a typical GWAS result is based on the assumption that the phenotypic associations detected at tag-SNPs are due to high LD (in short physical distance) between these SNPs and the non-genotyped causal variant site. Indeed, a successful GWAS study only suggests an initial set of candidate loci marked by tag-SNPs with detected associations. A consequent fine mapping study can then be performed to make causal inferences about this initial set, to identify one or more loci harboring functional variants.

Traditionally, existing fine mapping algorithms are based on results from GWAS studies, which only analyze a set of GWAS-selected candidate loci with detected phenotypic associations (Stephens and Balding 2009; Edwards et al. 2013). To ensure valid causal inferences, these GWAS-based fine mapping approaches require that the causal variant site must be included in the initial set of candidate regions reported by GWAS.

With the advances in next generation sequencing technologies, the risk that no causal site is reported in the initial candidate set can be reduced by substituting traditional GWAS by whole-genome or whole-exome sequencing studies – where each variant site, including causal sites, will be directly tested for association. However, causal sites are not always genotyped even using sequencing data; they can be filtered due to low quality in low coverage sequencing, or the causal variant can be not a SNP but an indel or a CNV. Moreover, due to the high cost of sequencing technology, imputation-based approaches have been proved efficient to detect rare to low-frequent variant associations (Servin and Stephens 2007; Guan and Stephens 2008; Li et al. 2010), yet some rare causal variants may still be incorrectly imputed if not observed in the reference panel. After all, traditional fine mapping based on single marker association tests still faces the challenge that the causal site may remain unidentified.

Alternatively, several studies have focused on mapping causal variants through the estimated genealogies (e.g., Morris et al. 2002; Zöllner and Pritchard 2005; Tachmazidou et al. 2007; Burkett et al. 2013). For a genomic ROI undergoing recombination, different positions have different ancestries. Therefore, by evaluating the distribution of cases and controls on the genealogies across different positions, we are able to identify positions where cases are separating from controls on the ancestry, to locate candidate regions likely to be harboring causal variants. Compared to traditional fine mapping algorithms, these coalescent-based approaches directly analyze sampled genotypes across the ROI, and thus do not rely on the power in single marker association tests. Moreover, the genealogy at each position is inferred through jointly analyzing genotypes at surrounding regions. Indeed, these approaches are also more robust to scenarios when causal loci are not genotyped – since the missing information in a few non-genotyped variant sites can be minimized by the shared genealogies at neighboring positions.

Unfortunately, existing coalescent-based approaches are computationally intensive. Typically, they require MCMC to approximate the posterior distribution of possible genealogies conditional on sampled genotypes. With the increased sample size, the sample space of possible genealogies grows even faster than exponentially and thus creates unfeasible computational intensity in large-scale studies. Therefore, existing coalescent-based approaches tend to oversimplify the modeling of LD in haplotypes and are restricted to samples of no more than a few hundred individuals.

In the present study, we propose a framework which utilizes a clustering algorithm to reduce the sample space in genealogy inferences, which then enables fine-scale mapping of complex trait loci through estimated genealogies applicable to large scaled studies.

Intuitively, this method is motivated by the observation that sampled haplotypes tend to coalesce into a number of common ancestors in a very short period of time, compared to the entire coalescent process (Scheet and Stephens 2006). Therefore, if the causal mutation occurred at a time far prior to the formation of these common ancestors, we can then treat the sub-genealogy under each common ancestor as a cluster, and ignore these lower-end coalescent patterns within each cluster. Consequently, we only need to make inferences about the upper-end genealogy of clusters, such that the sample space of genealogies is reduced from a function of $2n$ haplotypes to a function of K clusters (figure 4.1). We make inferences about the K clusters similar to the clustering algorithm suggested by Scheet and Stephens (fastPHASE; 2006). Given these clusters, we then use the same HMM-enhanced model, introduced in Chapter 3, to estimate the genealogies of clusters at each position across the ROI. We implement the estimation of genealogies of clusters into our own computational package, fastTREE.

Conditional on the estimated genealogy at any position x within the ROI, we propose a Bayesian approach to calculate the posterior probability of association at x, to detect the association between genotypes at this position and the study phenotypes. We also design a fine mapping approach similar to our method in Chapter 3, which makes causal inferences for fine mapping through estimated genealogies across different positions in the ROI, to identify candidate regions harboring disease-causal variants.

To evaluate the performance of our method, we simulate case-control samples using a coalescent haplotype simulator (Hudson 2002). We consider sample sizes up to tens of thousands of individuals, each sequenced for a ROI in a few Mbs. We also consider multiple disease models with different risk allele frequencies and different effect sizes. We then compare the mapping accuracy of our method to the accuracy of traditional single marker association tests and subsequent fine-scale mapping (Purcell et al. 2007; Servin and Stephens 2007). To evaluate our method's robustness to the non-genotyped causal variants, we mask the true variant site and re-apply our method. We then evaluate if our method provides mapping accuracy comparable to scenarios when the causal site is genotyped and analyzed.

When genotypes at the causal variant site are available in the study sample, both our method and traditional approaches based on single marker tests correctly identify candidate regions covering these causal sites. Specifically, credible regions provided by our algorithm are slightly wider than the intervals suggested by traditional methods when single marker tests have extremely high power. However, when the site carrying the causal variant is not genotyped, single marker tests becomes less accurate, while our method still correctly identifies credible regions covering disease-causing sites. According to our simulation results, we obtain comparable mapping accuracy and coverage probability, regardless of whether the causal variant is genotyped or not. These results suggest that our algorithm performs robustly when the genotype at causal sites is missing.

4.2 Methods

We consider a case-control study where we have several thousand individuals, to locate candidate regions harboring causal variants in a sequenced region of interest (ROI). Suppose we have n cases and n controls, each sequenced in a ROI of L Mbs containing J variant sites; we denote G as the genotypes of these $2n$ individuals in J variant sites. We denote Φ as the observed phenotypes, where $\varphi_i = 1$ if individual i is an affected individual (case), and $\varphi_i = 0$ if otherwise.

At any position x within the ROI, we define p_0 and p_1, the phenotype probability for any individual carrying homozygous wild type alleles and for any individual carrying homozygous variant type alleles at this position. At a position carrying disease-causal variants, p_0 and p_1 at this position are equivalent to the penetrance levels conditional on the genotype. At a position independent to the phenotype, true values of p_0 and p_1 are both equal to the phenotype prevalence in the target population.

Under different models of p_0 and p_1, we would like to calculate the joint probability of observed phenotypes and genotypes at position x, $P(\Phi, G|p_0, p_1)$. In Chapter 3, we propose an algorithm similar to Zöllner and Pritchard (2005) or Burkett et al. (2013), which calculates $P(\Phi, G|p_0, p_1)$ through integrating over possible genealogies at this position x, which is denoted as T_x:

$$P(\Phi, G|p_0, p_1) = \int P(T_x|G)P(\Phi|T_x, p_0, p_1)dT_x$$

However, the sample space of possible genealogies grows too fast with increased sample size n, such that this previous MCMC algorithm does not converge in large samples.

In this Chapter, we separate the entire genealogy T_x of any position x into two levels. In the lower-end the genealogy, sampled haplotypes have coalesced into K common ancestor haplotypes where the value of K is pre-determined in our model. We treat these ancient haplotypes as cluster, and approximate the lower-end coalescent process by estimating each sampled haplotype's probability of descending from each cluster. We then denote T'_x as the upper-end genealogy of the K clusters as. Since the number of clusters K is pre-selected such that the sample space of genealogies of T'_x does not increase with the increased sample size n.

Specifically, we denote $\Theta = [\theta^1, \dots \theta^K]$ as the vector of ancient cluster haplotypes, where $\theta^k = [\theta_1^k, \dots \theta_J^k]$ represents the allele types of the J variant sites carried in ancient cluster k $(k = 1, 2, \dots, K)$. We then denote α, the clusters of origin of sampled individuals, where $\alpha_j^{i(1)}(k)$ denotes the probability of individual i's $(i = 1, \dots 2n)$ first haplotype is a descendant from cluster k at site j $(j = 1, 2, \dots, J)$.

Consequently, we calculate $P(\Phi, G|p_0, p_1)$ through T'_x as follows:

$$P(\Phi, G|p_0, p_1) = \iint P(\Theta, \alpha|G) \cdot \int P(T'_x|\Theta)P(\Phi|T'_x, p_0, p_1, \alpha)dT'_x \cdot d\Theta d\alpha$$

Specifically, we make inference about $P(\Theta, \alpha|G)$ by sampling possible clusters using fastPHASE (Scheet and Stephens 2006), such that we approximate $P(\Phi, G|p_0, p_1)$ by:

$$P(\Phi, G|p_0, p_1) = \frac{1}{S}\sum\nolimits_{s=1}^{S} \int P(T'_x|\Theta^{(s)})P(\Phi|T'_x, p_0, p_1, \alpha^{(s)})dT'_x$$

Given estimated K cluster haplotypes $\Theta^{(s)}$, the posterior probability of upper-end genealogies $P(T'_x|\Theta^{(s)})$ can be approximated using the same MCMC algorithm as we used to approximate the genealogy of sampled haplotypes in Chapter 3. We then evaluate the probability of observed phenotypes conditional on the estimated genealogy, $P(\Phi|T'_x, p_0, p_1, \alpha^{(s)})$ by integrating over possible mutations at every branch on T'_x (Appendix F).

In the present study, we implement our own program, fastTREE, to calculate the joint probability of phenotypes and genotypes under different penetrance models, $P(\Phi, G|p_0, p_1)$, through estimated genealogy of clusters. We then apply our program to Bayesian tests of associations and fine-scale mappings, as described in the following subsections.

4.2.1 Bayesian test of association

At any position x across the ROI, we propose a method which calculates the posterior probability of association (PPA), as the Bayesian test statistic of phenotypic association at this position (Stephens and Balding 2009):

$$PPA = \frac{PO}{1 + PO},$$

where PO represents the posterior odds of association:

$$PO = BF \times \frac{\pi}{1 - \pi}.$$

Specifically, π is the prior probability of association. We apply uniform priors across the ROI in the present study, such that any position x within the ROI is equally likely to be associated with the study phenotype.

Moreover, BF is the Bayes factor in favor of association, which evaluate the probability of observed phenotype and genotype data under penetrance model with association ($p_0 \neq p_1$) over the same probability under model with no association ($p_0 = p_1$) as follow:

$$BF = \frac{P(\Phi, G|p_0 \neq p_1)}{P(\Phi, G|p_0 = p_1)}$$

Under either model, we apply our own algorithm, fastTREE, to calculate $P(\Phi, G|p_0, p_1)$ through the estimated genealogy of clusters T'_x at any tested position x. Therefore, BF is calculated by:

$$BF = \frac{P(\Phi, G|p_0 \neq p_1)}{P(\Phi, G|p_0 = p_1)}$$

$$= \iint_{p_0 \neq p_1} P(\Phi, G|p_0, p_1) dp_0 dp_1 \,/ \iint_{p_0 = p_1} P(\Phi, G|p_0, p_1) dp_0 dp_1$$

By definition, higher BF indicates that the observed phenotype and genotype data is more likely following the model with association ($p_0 \neq p_1$), in contrast to the model without association ($p_0 = p_1$). Consequently, any position x with extremely high values in BF will result in a posterior probability of association close to 1, which can be interpreted as strong evidence of association at this position.

4.2.2 Fine mapping of causal loci

Similar as testing of associations, we can also apply our algorithm, which calculates the joint phenotype and genotype probability through genealogy of clusters, to fine-scale mappings of causal variants.

To begin with, we denote X as the indicator of causal position, where $X = x$ represents that a causal mutation is carried at position x in the ROI. In the present study, we assume uniform prior probabilities $P(X = x)$ such that any position within the ROI is equally likely to be carrying disease-causal mutations.

For each position across the ROI, our goal in fine mapping is to calculate each position's posterior probability of carrying disease-causal mutations, conditional on the observed genotypes and phenotypes, to identify candidate regions harboring causal variants. By Bayes rule, this posterior probability of causality is proportional to:

$$P(X = x|\Phi, G) \propto P(\Phi, G|X = x)P(X = x).$$

We calculate $P(\Phi, G|X = x)$ by integrating over penetrance levels p_0 and p_1 as described in previous subsections:

$$P(\Phi, G|X = x) = \int_0^1 \int_0^1 P(\Phi, G|p_0, p_1) dp_0 \, dp_1$$

Given each penetrance level, $P(\Phi, G|p_0, p_1)$ is calculated through estimated genealogy of clusters, T'_x, using our algorithm fastTREE.

Similar as in Chapter 3, we select a set of focal points uniformly distributed across the entire ROI. At each focal x, we first calculate $P(\Phi, G|X = x)$ through our estimated genealogy of clusters. We then apply a Monte Carlo integration to calculate the posterior probability of causality $P(X = x|\Phi, G)$ at each focal point x:

$$P(X = x|\Phi, G) \approx P(\Phi, G|X = x) / \sum_{y \in ROI} P(\Phi, G|X = y).$$

4.2.3 Algorithm evaluation

To evaluate our algorithm performance in mapping of complex trait loci, we first simulate a set of $2N$ haplotypes sequenced in a 1Mbs ROI with 500Kbs flanking region on both ends. We then randomly assign these haplotypes into N diploid individuals.

Within the ROI, we randomly pick one variant site as the causal variant based on its minor allele frequency. We consider 3 scenarios with different risk allele frequencies, 1%, 5%, or 20%, to evaluate our algorithm performance. The risk allele effect size is OR = 2.0 for the rare risk allele with MAF = 1%. We let OR = 1.5 for the low frequent risk allele with MAF = 5%, and OR =1.2 for the common risk allele with MAF = 20%. Consequently, we assign a phenotype to each individual conditional on his genotype at this causal site under a multiplicative model. We consider different sample sizes in large-scaled case-control studies, and randomly sample n cases and n controls where n equals 2,000, 5,000, or 10,000. Under each scenario, we simulate 5 realizations to evaluate our algorithm performance.

To evaluate our algorithm performance in detecting phenotypic associations, we apply both our algorithm and the traditional single marker Bayesian approach (BIMBAM; Servin and Stephens 2007), to each simulated sample of cases and controls, to identify association signals across the ROI. Using BIMBAM, we test for associations at every marker individually. While using our algorithm, we make genealogy inferences by fastTREE in a set of focal points in every 10kbs, and calculate each focal point's posterior probability of association (PPA). Under each simulated scenario, we compare the median distance between the causal sites and the estimated most likely positions (focal points with highest PPA) in 5 realizations. We then compare our results to the distance using BIMBAM, to evaluate our algorithm performance in detecting phenotypic associations.

To evaluate our algorithm performance in fine mapping of complex trait loci, we also apply our fine mapping algorithm to each simulated data. Under each simulated scenario, we calculate the median width of 90% credible regions provided by our algorithm, as well as the coverage probability, to evaluate our algorithm performance. We also apply traditional fine mapping approaches suggested by Servin and Stephens (2007) and Stephens and Balding (2009) to each simulated dataset, and calculate the median widths of consequent 90% credible regions. We compare the credible region width using our algorithm to the width using traditional methods, to evaluate our algorithm performance in fine mapping.

Moreover, to evaluate our algorithm performance when the causal site is not genotyped, we mask the genotypes at the exact causal sites, and re-apply our algorithm to each simulated dataset. We calculate the same distance in association tests and credible region width in fine mappings. We then compare these new results to the previous results to illustrate that our method performs robustly when genotypes at causal variant sites are not available in study sample.

4.3 Results

To evaluate our algorithm performances, we simulate large-scale samples with n cases and n controls (n = 2,000, 5,000, or 10,000), each sequenced in a region of 1Mbs with 500Kbs flanking regions on both sides. Under each sample size, we consider different risk allele frequencies with different disease effect sizes: rare risk variants with MAF $= 1\%$ and effect size OR $= 2.0$, low frequent risk variants with MAF $= 5\%$ and OR $= 1.5$, or common risk variants with MAF $= 20\%$ and OR 1.2. We apply both our algorithm and the traditional single marker test based approaches to these simulated datasets, to evaluate our algorithm performance in testing for associations and mapping for disease causal variants.

4.3.1 Testing for associations

To evaluate our algorithm performance in testing of associations, we make inferences about genealogies by fastTREE in a set of focal points per 10Kbs in each simulated dataset. We then calculate the posterior probability of associations at each focal point. We then compare our results to the results provided by single marker tests using either Fisher's exact test (PLINK; Purcell et al. 2007) or Bayesian approaches (BIMBAM; Servin and Stephens 2007).

To begin with, we illustrate a single realization using 5,000 cases and 5,000 controls, while the risk allele to be sought is rare with MAF less than 1% and OR equals 2.0 (figure 4.2). Using single marker tests, variant sites with smallest p-values in Fisher's exact tests (figure 4.2A) are identical to the sites with greatest posterior probability of association (PPA) in Bayesian approaches (figure 4.2B), while both tests provide strongest signal at the exact position of the causal variant. Using our algorithm, the posterior probability of association at the set of focal points also increases towards the causal site, and peaks in the few focal points closest to the causal site. While single marker tests provide strong evidence of association at the exact variant site with PPA close 1 (grey dots, figure 4.2B) in this realization, the result using our method provides slightly lower PPA – since we do not test on the causal site directly (figure 4.2B, red line).

When the causal site is not genotyped, no signal is detected when using single marker tests. To illustrate that our algorithm performance with the absence of causal genotypes, we mask the causal site, and re-calculate the PPA at each focal point to detect any evidence of association (figure 4.2B, blue line). Compared to our previous results, we obtain almost the same low PPA in regions far away from the causal site since the missing genotype at the single causal site have no influence on the genealogy estimation in these regions. Moreover, we observe the PPA increases in focal points closer to the causal site, and peaks at the position closest to the exact site carrying disease causal mutations. Compared to our previous results with genotyped causal site, our algorithm identifies the same position with the strongest evidence of association, although the PPA is reduced since the information carried by the causal site is missing in our data.

In each simulated dataset, we apply both our algorithm and the Bayesian single marker tests using BIMBAM (Servin and Stephens 2007) to 5 simulated datasets per scenario, and calculate the median distance between the exact causal site and the position showing strongest evidence of associations. We then compare these calculated distances using our algorithm to the distances using BIMBAM (table 4.1). We also calculate the highest posterior probability of association reported in each dataset using our algorithm and using BIMBAM, and compare the median value of 5 realizations under each scenario, to evaluate under what scenario our method is more powerful than using traditional single marker tests.

Using single marker tests, the results are driven by single marker testing power. When the sample size increases, we have higher power to detect the association at the exact causal site with higher posterior probability of associations. Under each sample size level, we have higher power to detect low frequent risk variants with MAF = 5% and OR = 1.5, rather than rare risk variants with MAF = 1% and OR = 2.0 and common risk variants with MAF = 20% and OR = 1.2. Specifically, in smaller sample of 2,000 cases and 2,000 controls, single marker test is most powerful in detecting low frequent variants with moderate effect sizes (MAF = 5%, OR = 1.5), with the highest detected PPA equals 0.832 in a site 24Kbs from the exact causal site. The strongest signal is detected with higher PPA, and is detected more closely to the exact positions when we have larger sample sizes. With 10,000 cases and 10,000 controls, the median value, we detect a posterior probability of association extremely close to 1 at the exact position of the causal variant (distance equals 0).

Using our method, the power to detect phenotypic associations is also increasing with increased sample sizes. Moreover, our method is more powerful in detecting rare variant associations when risk variant allele frequency is less than 1% with the effect size OR = 2.0 in these simulated samples of thousands of cases and controls. When we only have 2,000 cases and 2,000 controls, our algorithm outperforms traditional single marker tests with the strongest signal detected closer to the exact position of causal variants – in all 3 scenarios of allele frequencies. When we have more than 5,000 cases and 5,000 controls, although our algorithm correctly detect the strongest association signal within 2 to 3 focal points closest to the causal site; it is less powerful than the traditional single marker test which directly tests on the exact causal variant site.

To evaluate if our algorithm performs robustly when the causal site is not genotyped, we mask the causal site genotypes, and re-apply our algorithm and the single marker tests to each simulated dataset. We illustrate such results in realizations with 5,000 cases and

5,000 controls (figure 4.3). Using traditional single marker tests, the distance between the causal site and the position with the strongest detected association signal (black dots) merely depends on whether neighboring sites under high LD with the causal variant outweigh noise markers at random positions across the ROI. Therefore, the variation in single marker test performances is very high in the 5 simulated datasets per scenario. In contrast, using our algorithm, the strongest signal is always detected within 50Kbs (red triangles) from the exact position of the non-genotyped causal variant.

Moreover, we calculate the median distance between the non-genotyped causal site and the focal point with the highest PPA using our method (table 4.2, in parenthesis) under each scenario. Compared to previous results, we obtain similar results where the focal point with the highest PPA only moved by one or two units, such that the median distance to the causal site barely changes.

4.3.2 Fine mapping of causal variants

Traditionally, association-based fine mapping approaches assume that at least one causal variant is covered by the initial set of candidate regions with reported phenotypic associations. In contrast, our algorithm does not make such assumption, and can be applied directly through our estimated genealogies across the entire ROI, to construct credible regions harboring disease causal variants.

We apply our fine mapping algorithm and the traditional fine mapping approach to simulated datasets. Under each scenario, we first calculate the median width of the 90% credible regions (table 4.2) in 5 realizations. We then compare our results to the same widths from using traditional fine mapping approaches based on candidate regions with detected associations from single marker tests.

Using traditional approaches, fine mapping accuracy is driven by the power in association tests. Therefore, the trend in mapping accuracy across different sample sizes and different allele frequencies is identical to the trend in association tests: We obtain shorter credible regions with increased sample sizes. And we obtain shortest credible regions under each sample size when the causal variant is of low frequent with MAF = 5% and OR = 1.5.

Using our fine mapping algorithm, the resulting credible region is also narrowing with increased sample size. Under each sample size level, we obtain the narrowest credible region when the causal variant is rare (MAF = 1%) with strong effects (OR = 2.0). These trends are also identical as the trends in our coalescent-based association tests – since both applications are based on the same estimated genealogies.

Under all scenarios, our algorithm identifies credible regions covering a candidate region around 100Kbs. Moreover, our method provides credible regions correctly covering the causal variant site in all realizations, indicating that our credible interval may be too conservative compared to its nominal credible level. Specifically, when mapping causal variants using 2,000 cases and 2,000 controls, our algorithm outperforms traditional methods when the causal variant is rare (MAF = 1%) or is of weak effects (OR = 1.2). When the sample size is greater than 5,000 cases and 5,000 controls, traditional approaches provide more accurate mapping of candidate regions.

Moreover, according to our results in association tests, single marker tests do not consistently detect association in regions surrounding the causal site. If an association study fails to correctly report an initial set of candidate regions covering the causal site, the consequent fine mapping will be biased.

To evaluate if our algorithm provides valid fine mapping results when the causal site is not genotyped, we mask the genotypes at causal positions in each simulated dataset, and re-apply our algorithm for fine mapping. Consequently, we calculate the new medians of credible interval widths under each scenario (table 4.2, in parenthesis). Under all scenarios, we observe slightly wider credible regions in scenarios with smaller sample sizes, compared to our previous results. Likewise, we observe the same coverage probability that our algorithm still correctly identifies candidate regions harboring disease causal variants.

4.3.3 Mapping accuracy when more than one causal variants exist in ROI

So far we assume that only one causal site is located in the study ROI. In practice, multiple risk variants can occur in neighboring regions to affect the study disease. To evaluate if our algorithm is capable in identifying more than one risk variants, we simulate datasets with 2,000 cases and 2,000 controls where two risk variants exist in the 1Mbs region of interests. We consider two conditions: either two closely located risk variants with mutual distance less than 50Kbs, or two distantly located risk variants further than 200Kbs from each other. Likewise, we also consider two scenarios of risk allele frequencies: both risk variants are rare with MAF 1% and OR 2.0; or one such rare allele with strong effect sizes and one common allele with MAF 20% and OR 1.2.

We then apply our algorithm to construct 90% credible regions to evaluate if we have correctly covered both risk variants. When two rare risk alleles are closely located, we observe one continuous credible region around 100Kbs, which correctly covers both risk variant sites (figure 4.4A). In contrast, when two rare alleles are distantly located, our credible region covers two discrete sub-regions; each covers one risk variant in a 100Kbs area (figure 4.4B). Under the scenario when one common variant and one rare variant are closely located from each other (figure 4.4C), the posterior probability curve is skewed to the rare variant site as it has stronger effect size to the disease. Our algorithm still provides a credible region with similar widths, to correctly identify both risk variants. When the common risk variant is distantly located from the rare risk variant (figure 4.4D), we observes signals on both risk positions, while the estimated posterior probability at the rare variant site is about 10 times higher than at the common variant site. Although close to the boundary, the common variant is still correctly covered by our credible region (figure 4.3D).

4.4 Discussion

We have presented and implemented an algorithm to identify complex trait loci through estimated genealogies, to enable coalescent-based mapping on studies using large-scaled case-control samples. Conditional on observed genotypes, we first make inferences about a set of K ancient haplotypes, as common ancestors of sampled individuals, based on a clustering algorithm (Scheet and Stephens 2006). We then treat these K common ancestors as clusters, and make inferences about this position-specific genealogy of clusters. Consequently, we reduce the sampling space of genealogies from a function of sample size ($2n$, in haplotypes) to a function of pre-determined number of clusters, K. We implement this algorithm for genealogy inferences into our own computational tool fastTREE.

Focusing on detecting phenotypic association at any position across the genotyped or sequenced ROI, we make inferences about the position-specific genealogy of clusters. Conditional on this inferred genealogy, we calculate this position's Bayes factor (BF) in favor of association, and then derive the posterior probability of association based on BF, at this position.

Furthermore, we are also able to make causal inferences for fine mapping of compelx trait loci, based on the assumption that our ROI contains at least one disease causal variant. In a set of focal points across the ROI, we evaluate the inferred genealogies at every focal point. We then make Bayesian inferences for causality at these positions, and establish credible regions to locate causal variants.

Compared to traditional coalescent-based approaches for fine mappings, which restrict the sample size to a few hundred individuals, our method enables fine mapping in large scale case-control studies with up to several ten thousand individuals. Moreover, our algorithm can be running in parallel to make genealogy inferences at different positions across ROI. Therefore, we are able to apply our method for fine mapping in a sequenced ROI of a few Mbs with up to twenty thousand sampled individuals in less than ten days if running on 40 to 50 modern computer nodes.

Compared to traditional fine mapping approaches based on detected associations, our method performs more robustly when genotypes at the causal site are missing. Because our method make inferences about genealogies at each position by analyzing genotypes at surrounding variants sites jointly, information carried by the missing genotype at one site complemented by genotypes within co-inherited haplotype segments. In comparison, traditional fine mapping algorithms are based on an initial set of candidate regions with detected associations, and calculates individual Bayes factors site-by-site to make causal inferences. Based on our simulation, we suggest using our algorithm for fine-scale mappings in genomic region of interests where single marker tests are not sufficiently powerful in suggesting candidate regions for traditional fine mapping.

In the present study, we fix the number of clusters K at 400. In practice, K can be determined by two factors. First, we want K as high as possible such that our inferred genealogy carries more information about the sampled haplotypes' evolutionary history. The upper bound of possible K is then determined by computational burden. Second, we want K greater than the number of branches left on the coalescent tree when causal mutation occurred in history. Thus the lower bound of K is determined by the disease genetic architecture – suppose our goal is to seek some rare causal variant at certain allele frequency level (e.g., less than 1%), the lower bound of K is then determined as it must be greater than number of branches when all risk allele carriers have met the same common ancestor. In coalescent theories, the expected number of branches left in the coalescent tree, when all risk allele carriers have coalesced to a common ancestor, can be derived sequentially. Overall, the number of K can be pre-selected between the upper and lower bound.

According to our simulation, our algorithm identifies credible regions covering a candidate region around 100Kbs in most scenarios. In practice, this is potentially helpful to locate 2 to 3 candidate genes in human genome for follow up studies to understand their biological functions. Our algorithm is of interests in studies trying to identify complex trait loci using different types of genetic data, such as low-coverage sequencing studies or imputation-based studies – in which sites carrying disease causal variants are of higher probabilities of not genotyped.

In present study, we have only considered uniform prior probability of causality across the ROI. In practice, more informative priors can be applied based on gene densities or annotation of variants, to enhance the efficiency of our algorithm. Other future extensions for our algorithm include using mapping of QTLs through estimated genealogies, as well as under other generalized phenotypes models. Along with the advances in sequencing technologies and the decreasing sequencing costs, we also project more large-scale sequencing study samples will be formed in near future. Yet our fine mapping program provides a robust and computationally feasible package to identify candidate regions for these large-scale studies to improve our understanding about the genetic basis of complex diseases.

Table 4.1 Estimated posterior probability and the distance between causal sites and positions with strongest signal of associations, using single marker tests (BIMBAM) and using our algorithm through estimated genealogies (Tree).

Sample size	**MAF**	**BIMBAM**		**Tree**	
		PPA	**Dist**	**PPA**	**Dist(*)**
2,000	1%	0.756	45	0.853	15(25)
2,000	5%	0.832	24	0.788	20(25)
2,000	20%	0.625	50	0.747	25(30)
5,000	1%	0.998	2	0.906	15(20)
5,000	5%	1.000	0	0.887	20(25)
5,000	20%	1.000	3	0.872	20(25)
10,000	1%	1.000	0	0.953	15(20)
10,000	5%	1.000	0	0.946	15(15)
10,000	20%	1.000	0	0.881	20(25)

Dist: between positions with strongest signals and causal sites in Kbs.

PPA: Posterior probability of association.

*In parenthesis: Distance when causal variant site not genotyped.

Table 4.2 Median 90% credible region width in 5 realizations per scenario, using fine mapping algorithms based on single marker tests (BIMBAM) and using our algorithm through estimated genealogies (Tree).

Sample size	**MAF**	**CI**	
		BIMBAM	**Tree(*)**
2,000	1%	122	100(120)
2,000	5%	76	120(130)
2,000	20%	140	130(130)
5,000	1%	35	80(90)
5,000	5%	24	90(90)
5,000	20%	42	100(100)
10,000	1%	8	80(80)
10,000	5%	5	80(90)
10,000	20%	22	90(90)

CI: 90% credible region widths in Kbs

*In parenthesis: CI width constructed when causal variant site not genotyped.

Figure 4.1 Illustration of clustering method when disease mutation occurred prior to the formation of clusters. Figure A: genealogy of 16 sampled haplotypes, where disease mutation occurred in time when only 3 branches exist. Figure B: genealogy of 5 clusters, which reduces the sampling space by ignoring the bottom coalescent details within each cluster.

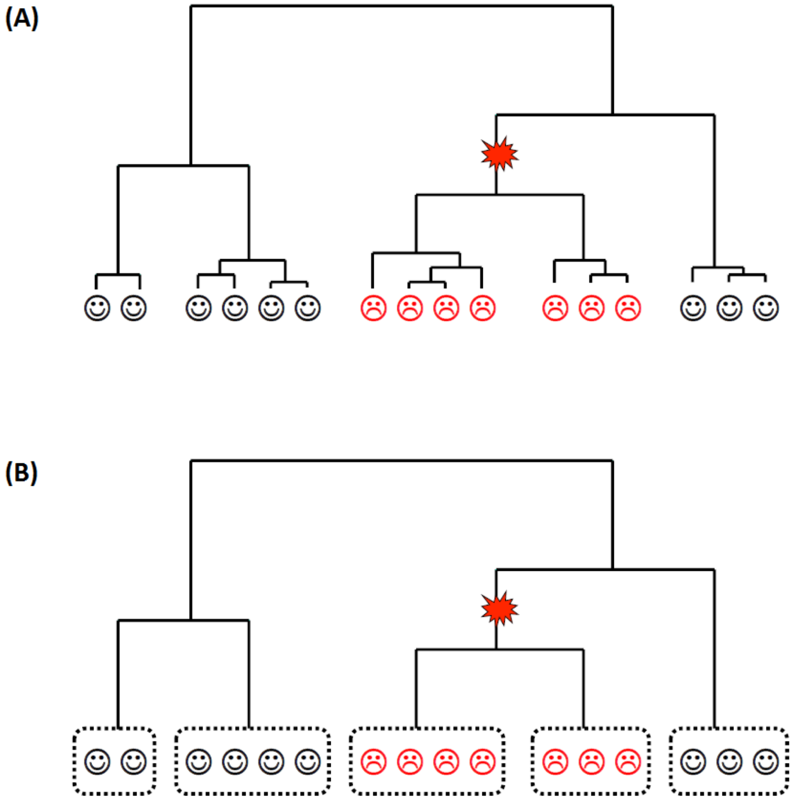


Figure 4.2 Test of associations. In figure A, black dots represent the p-values using Fisher's exact tests. In figure B, grey dots are posterior probabilities of associations provided by Bayesian approaches for association tests (BIMBAM); while the red line indicates the posterior probability of association provided by our algorithm. The true causal position is labeled by the green line.

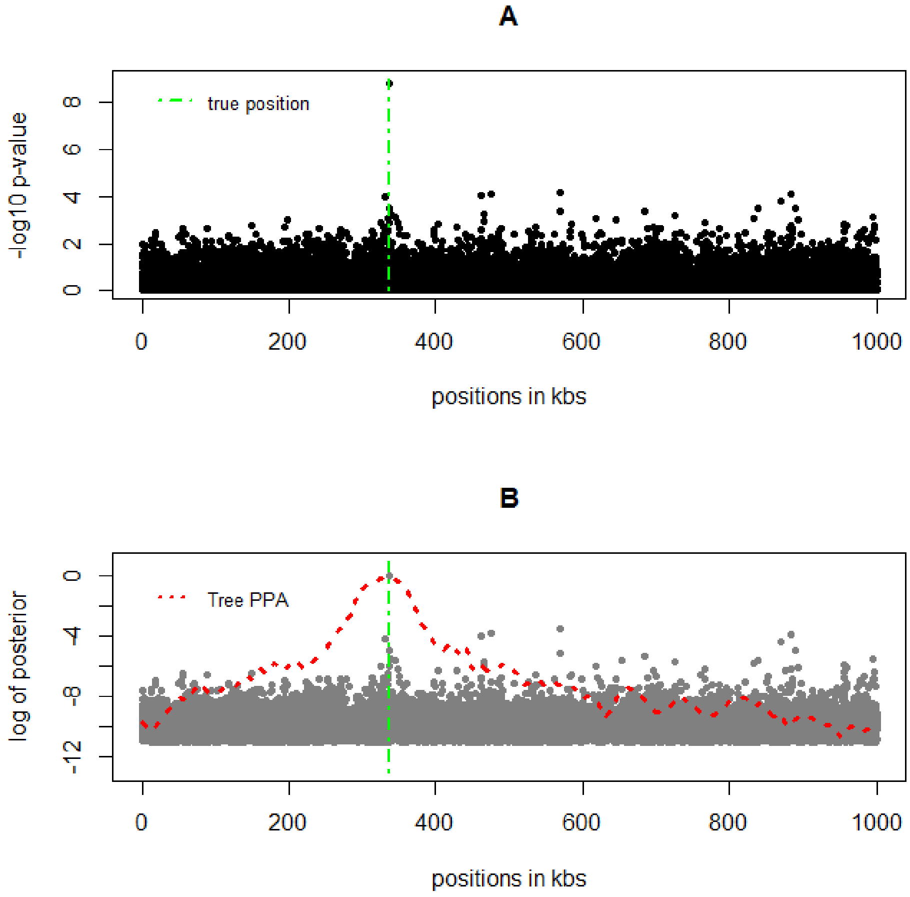


Figure 4.3 Distance between the causal site and the position with strongest detected associations, when causal site is not genotyped in study samples of 5,000 cases and 5,000 controls. The black dots represent the distances obtained using single marker tests, while the red triangles represent the distances using our algorithm through estimated genealogies. The horizontal axis is the physical distances in base-pairs, while the vertical axis indicates the three considered allele frequency level: 1%, 5%, and 20% risk alleles.

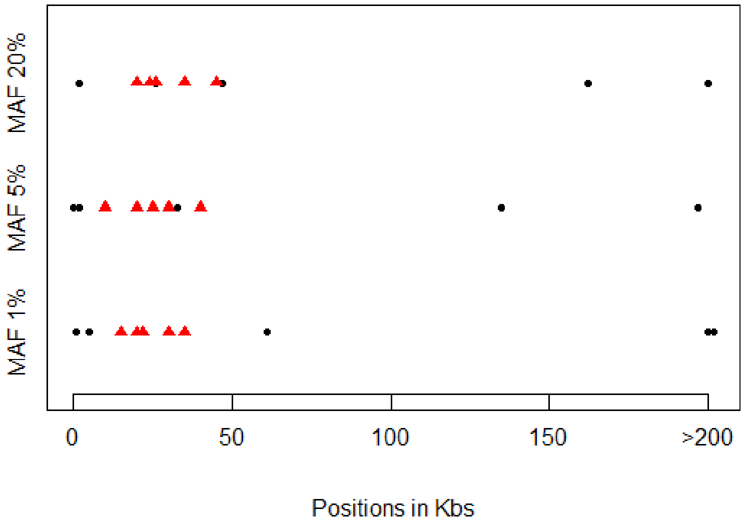
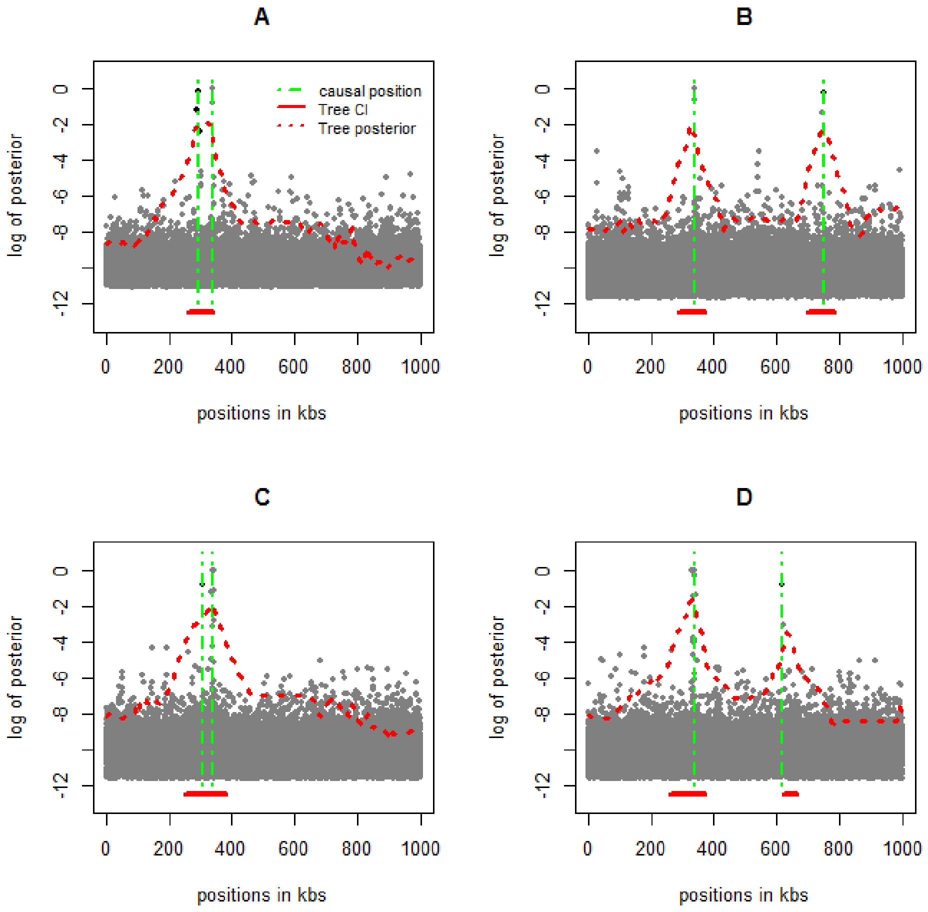


Figure 4.4 Fine mapping results when more than one causal variants exist in the region of interest. In figure A, two rare risk variants are closely located with less than 50Kbs. In figure B, two rare variants are distantly located further than 200Kbs. In figure C we simulate one rare variant and one common risk variant, closely located. In figure D, one rare and one common risk variants are distantly located. Posterior curves are represented by red solid lines, while the red underlines represent the estimated 90% credible regions.



Chapter 5 Conclusion

5.1 Summary

Markov chain Monte Carlo (MCMC) algorithms are widely used to analyze high-dimensional data, and to sample posterior probabilities of interests in modern genetic studies. In this dissertation, we have applied an MCMC algorithm on a framework about sampling genetically homogeneous sub-samples of affected individuals. We have also proposed an MCMC method to model the linkage disequilibrium in genotype data, which can be applied to improve the accuracy in haplotype reconstruction, as well as to identify candidate regions harboring causal variants of a complex disease.

In Chapter 2, we propose a novel approach to identify affected individuals carrying the same underlying genetic risk variants, which can be applied on studies of any complex disease. Using family data with at least one affected individual per family, we first analyze the segregation in a large set of observed covariates, to identify subsets of covariates correlated to the same risk variants to define subphenotypes. Conditional on these covariates subsets, we then calculate each affected individual's posterior probability of belonging to each subtype such that they are more likely to carry the same risk variants attributable to this subtype. We further stratify affected individuals based on their subtype memberships, and illustrate that using these stratified subsamples of cases provide improved power to detect the phenotypic association in relevant loci, compared to the power using all sampled cases. We implement this algorithm into our own program based on sibling pair data, CASES. We recommend using CASES to identify genetically homogeneous subtypes of affected individuals to improve our understanding of any complex disease in follow-up studies.

In Chapter 3, we propose HMM-enhanced method which improves the modeling of LD in coalescent-based estimation of genealogies. Given genotype data, we first make inferences about genealogies at different focal points across ROI, and model the LD between more distant loci not sharing the same genealogy due to recombination events through a HMM algorithm. We then apply this modeling of LD to improve the accuracy in haplotype reconstructions. Compared to existing methods, we recommend our model in applications for haplotype reconstructions where traditional HMM-based methods experience low accuracy due to more complicated population genetic parameters. We also apply this coalescent-based LD modeling into case-control studies, to identify candidate regions harboring disease causal variants. We compare our results to traditional fine mapping algorithms based on an initial set of candidate loci with detected associations, and illustrate that our method outperforms traditional approaches when causal site is not directly genotyped and tested.

In Chapter 4, we implement a clustering method to this coalescent-based algorithm to mapping complex trait loci in large-scale studies. We first make inferences about a set of common ancestors to the study sample, as a set of clusters. Conditional on the inferred ancient haplotypes, we then make inferences about the genealogies of clusters at different positions across the ROI to identify candidate regions with causal variants by our own program fastTREE. We define a Bayesian approach to test for associations at any position through the inferred genealogies. Based on the assumption that at least one causal variant exists within the ROI, we then propose a fine mapping framework through the inferred genealogy, to make causal inferences for candidate regions harboring causal variants. We evaluate our algorithm under different disease genetic architectures: under rare to common causal variants with different effect sizes. We compare our mapping results to traditional Bayesian approaches for association tests and fine mappings, and illustrate that our algorithm behaves more robustly, and outperforms traditional methods, when sample genotypes at causal sites are missing.

5.2 Relevance and future work

This dissertation covers MCMC applications on two distinct areas of genetic studies: subphenotyping complex diseases to identify genetically homogeneous subsamples of cases, and modeling LD through estimated genealogies for reconstructing sample haplotypes and mapping complex trait loci.

We implement our subphenotyping algorithm based on sib-pairs into our own software CASES. This sampler will be helpful for practitioners in sequencing studies to prioritize individuals more likely to carry specific risk variants, to reduce the sequencing cost. It is also helpful in stratifying affected individuals to formulate genetically homogeneous subsamples of cases, to improve the power in association studies. Future extension includes implementing CASES more generalizable to other family structures and then modeling genotypes by kinship coefficients, such that the power to identify genetically homogeneous subphenotypes is potentially improved by larger family sizes with parentoffspring trios. Meanwhile, CASES currently requires dichotomous diseases, and we could further explore the algorithm's potential in QTL studies by allowing continuous phenotype measurements in future works.

We implement the HMM algorithm onto our own approach similar to TreeLD (Zöllner and Pritchard 2005). Our approach enables making inferences about genealogies based on unphased genotype data using high density SNPs or sequencing data. In practice, our estimation of genealogies is helpful to improve the accuracy of reconstructing sample haplotypes in genomic regions where traditional HMM-based algorithms experience limited performances. In practice, practitioners can also apply our approach to make inferences about genealogies across a genomic region of interests to identify candidate loci harboring disease causal variants. Future extension for our algorithm includes using the local information from genealogies to improve genotyping accuracies to identify de novo rare variants, as well as identifying recombination hotspots in genomic regions through the varying length of shared and non-shared genealogies. Furthermore, we have only considered uniform prior probability of causality across the ROI in current study. In practice, more informative priors can be applied based on gene densities or annotation of variants, to enhance the efficiency of our algorithm.

When applying our estimated genealogy to identify complex trait loci, we compare our algorithm performance to traditional algorithms based on single marker tests. According to our results, we suggest using our method to identify candidate regions when single marker tests have low power. We would also like to compare our results to other advanced strategies, such as collapsing method for rare variant association tests (Li and Leal 2008), to further evaluate our algorithm's potential in different types of studies.

We also apply a clustering method to reduce the computational intensity in estimating genealogies. We implement this updated method into our own program fastTREE, which enables mapping of complex trait loci in large-scale studies with sample sizes up to ten thousands. Along with the current advance of sequencing technologies and the trend of reduced per individual cost in sequencing studies, we project more large-scale sequencing study samples will be formed in near future. Yet our fine mapping program provides a robust and computationally feasible package to identify candidate regions for these large-scale studies to improve our understanding about the genetic basis of complex diseases.

Appendix

A: Metropolis-coupled MCMC setup.

Genotype probability in sib-pairs: For each sib-pair i, we model the joint genotype probability as:

$$P(G_i|p) = \prod\nolimits_k P(G_{ik}|p_k)$$

For each variant locus k, $P(G_{ik}|p_k)$ is calculated by integrating over possible identity by descent (IBD) status, where IBD ∈ {0,1,2}:

$$P(G_{ik}|p_k) = \prod\nolimits_{x=0}^{2} P(G_{ik}|p_k, IBD = x)P(IBD = x)$$

The prior probabilities of IBD equals 0, 1, or 2 are 0.25, 0.5, and 0.25, respectively.

The spike and slab prior: For the covariate j, the probability of affected by risk variant in subtype k is denoted as $P_{jk} = P(\beta_{jk} \neq 0)$, is arbitrarily chosen between 0 and 1. Conditional on it is effective, the effect size is uniformly distributed in (–log5, log5). Therefore, the entire prior probability is:

$$I(\beta_{jk} \neq 0) \cdot (1 - P_{jk}) + I(\beta_{jk} = 0) \cdot P_{jk}/2log5$$

We use the same prior on between-covariate-effects.

Parallel Markov chains: We run Y Markov chains with different initial status in parallel. For each chain y, the total likelihood is powered by a “heat”: $H_y = [1 + h \cdot (y - 1)]^{-1} < 1$, where the heat-factor h is arbitrarily chosen from 0 to 1.

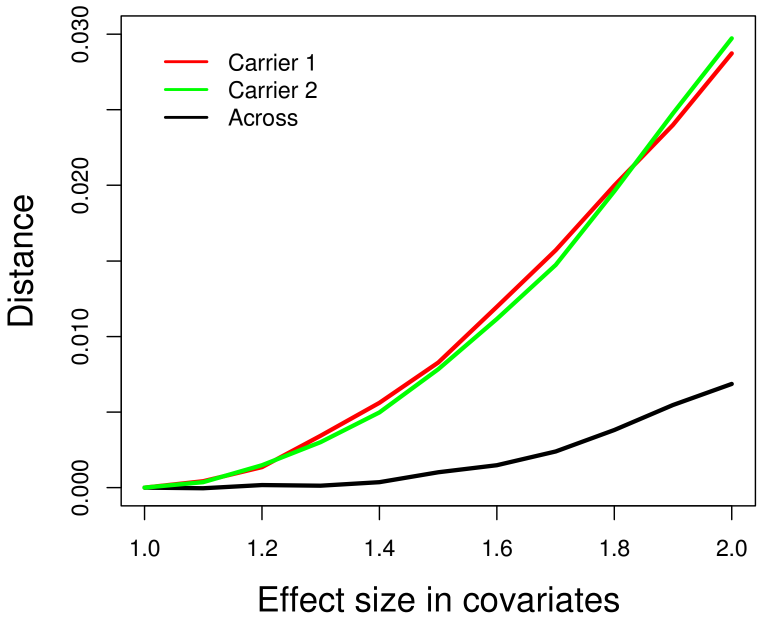
Chain switching: Let L_y be the likelihood in chain y. We randomly choose 2 consecutive chains per one thousand rounds in MCMC, and copy the information from chain a to chain b with probability $min(1, r)$ where $r = (L_a/L_b)^{H_b - H_a}$.

MCMC updates: In each Markov chain, we update all sib-pairs' genotype based on Metropolis-Hasting (M-H) algorithm in each round. We randomly choose one β_{jk} from all effect sizes in every 20 rounds of the MCMC, and update this effect based on the M-H algorithm. We update the between covariate effects using the same mechanism. For allele frequency, randomly pick locus k per 50 rounds, and update allele frequency p_k based on the conjugate posterior $B(T + 1, 2n - T - 1)$ distribution, where T is the total risk allelic type counts in all sibling 1s (or 2s), and $2n - T$ is the total counts for non-risk alleles.

B: Discriminate risk variant carriers from non-carriers

When there are two subtypes ($K = 2$) affected by two risk variants, CASES gives 2 vectors of inferred posterior probability about carrying variant 1 and 2.

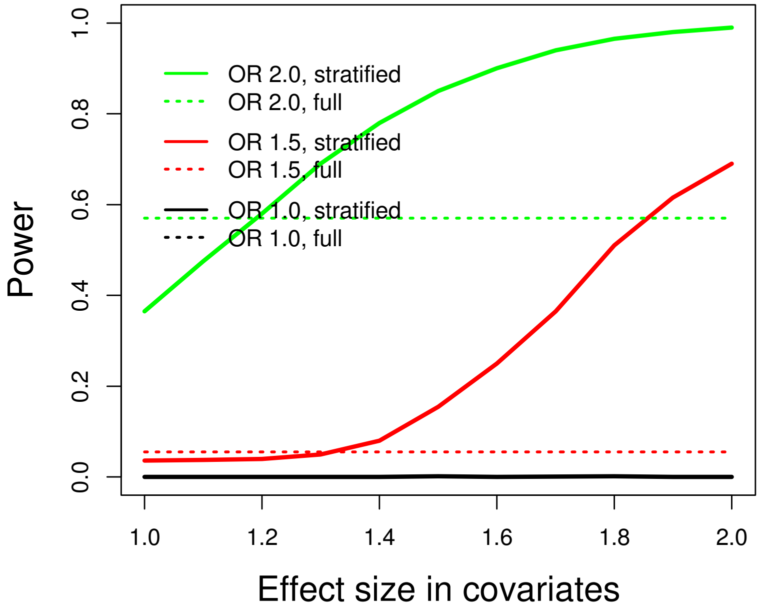
To determine if carriers from different risk variants can be discriminated from each other, we compare the difference between pair-wised distance within carriers and among any pair of individuals in the figure below. The horizontal axis represents effect sizes on covariates ranging from 1.0 to 2.0, and the vertical axis is the difference in pair-wised distances. The red (or green) line is the difference between within carriers of variant 1 (or 2) and within all samples, and the black line is such difference between within all carriers and within all samples.



In general, the mean pair-wised distances within either variant 1 or 2 carriers are shorter than the overall pair-wised distance. The distance is further shortened along with the increased covariate effect sizes. In other words, individuals carrying the same variant are more likely to formulate a cluster and stay close with each other with respect to the inferred posterior probability of carrying each subtype. In addition, the mean distance for pairs of individuals with one carrying variant 1 and the other carrying variant 2 is also shorter than the overall mean pair-wised distance, because carriers are either clustered at the high end of inferred subtype 1 or 2.

C: Power comparison under different effect sizes.

We compare the under different risk variant effects to the phenotype in the figure below. The underlying risk variant affects 10 covariates, whose allele frequency is 5%. The effect size to the phenotype is 1.0, 1.5 or 2.0. The horizontal axis represents effect sizes on covariates ranging from 1.0 to 2.0, and the vertical axis is power under significance level $\alpha = 5 \times 10^{-4}$.

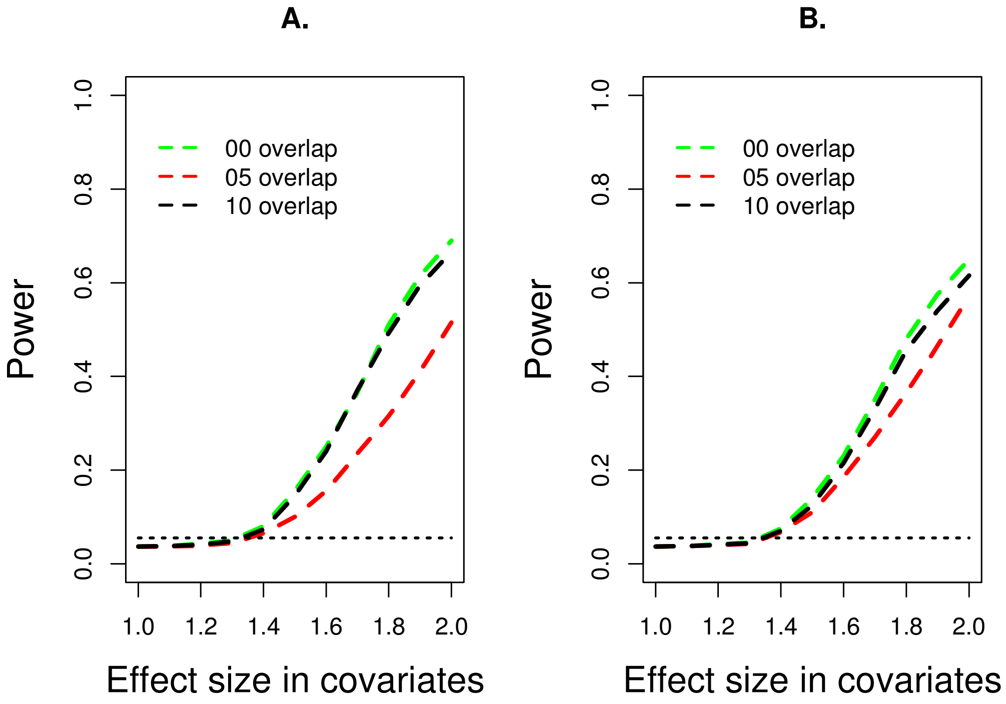


In general, power is increasing along with the covariate effect sizes. When the risk variant is truly affecting the phenotype, the power using our stratified sample outperforms the power using full samples. More importantly, when the effect size on phenotype is OR 1.0 (no association), the power curve using stratified subtype of cases fluctuates around the nominal significance level – indicating that our algorithm does not inflate the false discovery rate.

D: Testing power, multiple risk variants affecting the same covariates.

To evaluate if more than one risk variants affecting the same set of covariates would affect our algorithm performance, we consider scenarios where 10 covariates are affected by risk variant 1, while 0, 5, or 10 out of these 10 covariates are also affected by another risk variant 2. We summarize the power to detect phenotypic association at variant site 1 below. Since the true number of subtypes (K) is unknown, we compared the power under $K = 1$ and $K = 2$ model in figure A and B, respectively. The horizontal axis represents effect sizes on covariates ranging from 1.0 to 2.0, and the vertical axis is power under significance level $\alpha = 5 \times 10^{-4}$.

When risk variant 2 affect no covariate, the power to detect association at variant site 1 is identical to the previous single locus model with 10 covariates under $K = 1$. Alternatively, when we specify $K = 2$, CASES correctly identifies the second underlying variant affects no covariate. Thus, the power is only slightly reduced due to the extra parameters to be estimated, compared to the K=1 model.



When risk variant 2 affect 5 out of the 10 covariates affected by risk variant 1, we observe reduced power in both $K = 1$ and $K = 2$ models, due to the more complicated correlation structures between covariates. Specifically, the $K = 2$ model correctly specifies the number of underlying risk variants, and thus outperforms the $K = 1$ model which treats total variation unexplained by the single risk variant as between-covariate correlation.

When risk variant 2 affects all 10 covariates as variant 1 does, we cannot distinguish these two variants. Instead, these two variants are identified as one variant with elevated allele frequency under the $K = 1$ model. Since the resulting stratified cases are more likely to be carriers of either risk variant 1 or 2, we still observe increased risk allele frequency and improved testing power using this stratified subsample. Under $K = 2$ model, our algorithm struggles to distinguish the following 2 possibilities: 1. Two subtypes affected by two variants are represented by exactly the same covariates. 2. One subtype, affected by a more common risk variant, is represented by all 10 covariates; while the other subtype affected by another variant is not represented in any covariate. Due to the choice of the spike-and-slab prior, we put more weights on effect sizes equal 0. Thus, our algorithm converges to the second possibility such that similar testing power as in the $K = 1$ model.

Overall, we see different specification of K results in different algorithm performance. In reality, we do not know the true number of subtypes carried in our sample. Thus we always want to specify the "best" value of K in CASES, to make our algorithm most efficient. We suggest a sequential procedure for model selection below.

E: Selection of K, the number of subtypes.

In practice, we do not know the true number of subtypes due to underlying risk variants carried in our sample. We would like to choose the best value of K to identify individuals carrying risk variants most efficiently.

Essentially, how to determine the value of K is a model selection problem. We begin from the extreme case where underlying genetic variants affect only covariates (no subtype). Then the posterior probability of the phenotypic effect should be identical to its prior spike and slab distribution, which peaks at 0. If so, we have evidence that no underlying genetic variant is affecting the phenotype.

Similarly, when we set K greater than the true number of underlying risk variants in our model, some effect size to the phenotypes will always show the same spike at 0 since these pseudo, excessive factors are in fact unrelated to the disease. Indeed, we suggest that when using CASES, one could determine K under a sequential manner:

We start running CASES from $K' = 1$, and check the histogram on the phenotypic effect size to see if it peaks at 0. If so, stop CASES since no heritable subtype can be identified. If not, we increase K' by 1 and re-run CASES and evaluate all modeled variants' phenotypic effects. We increase K' iteratively until we see at least one variant's effect size estimate shows the spike at 0. Accordingly, a possibly reasonable choice about the number of risk sites is then $K = K' - 1$, which can be used in the follow-up inference and stratification.

F: Posterior probability of phenotypes given genealogy.

Given any penetrance level p_0 and p_1 at each focal point x, we propose possible occurrence of causal mutation on every possible branch on the genealogy of the K clusters T'_x. There are $2K - 2$ possible positions to allow for mutations, we define an indicator $B = b$ to represent that the disease mutation is occurred on branch b ($b = 1, \dots 2K - 2$), whose prior probability $P(B = b)$ is proportional to the branch length.

$$P(\Phi | T'_x, p_0, p_1, \alpha) = \sum P(\Phi | T'_x, p_0, p_1, \alpha, B = b) P(B = b)$$

Conditional on each proposed mutation on T'_x, the genotype at focal point x of each ancient haplotype is fully determined. We use $h_x^{i(1)}$ and $h_x^{i(2)}$ to denote the two allele types of individual i at this position. We then define the probability that individual i is a case as follows:

$$p^{(i)}(k_1, k_2) = P\left(\varphi_i = 1 \middle| h_x^{i(1)}, h_x^{i(2)}\right) = p_1{}^{(h_x^{i(1)}+h_x^{i(2)})/2} \cdot p_0{}^{(2-h_x^{i(1)}-h_x^{i(2)})/2}$$

where the two alleles are inherited from ancient cluster k_1 and k_2. Similarly, the probability that individual i is a control is $1 - p^{(i)}(k_1, k_2)$.

Given the cluster of origin probabilities in α, we then sum over all possible cluster of origins, to calculate the marginal phenotype probability if individual i is a case:

$$\sum_{k_1=1}^{K} \sum_{k_2=1}^{K} p^{(i)}(k_1, k_2) \cdot \alpha_x^{i(2)}(k_2) \cdot \alpha_x^{i(1)}(k_1)$$

We calculate the marginal phenotype probability for each control similarly, such that the joint phenotype probability of all individuals is:

$$P(\Phi | T'_x, p_0, p_1, \alpha, B = b)$$

$$= \prod_{i \in case} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} p^{(i)}(k_1, k_2) \cdot \alpha_x^{i(2)}(k_2) \cdot \alpha_x^{i(1)}(k_1)$$

$$\cdot \prod_{i \in ctrl} \sum_{k_1=1}^{K} \sum_{k_2=1}^{K} [1 - p^{(i)}(k_1, k_2)] \cdot \alpha_x^{i(2)}(k_2) \cdot \alpha_x^{i(1)}(k_1)$$

Consequently, we integrate over mutations occurring on all possible branches to calculate the joint phenotype probability conditional on genealogies:

$$P(\Phi | T'_x, p_0, p_1, \alpha) = \sum P(\Phi | T'_x, p_0, p_1, \alpha, B = b) P(B = b)$$

Bibliography

Andreassen OA, Djurovic S, Thompson WK, Schork AJ, Kendler KS, O'Donovan MC, Rujescu D, Werge T, Van de Bunt M, Morris AP, McCarthy MI; International consortium for blood pressure GWAS; Diabetes genetics replication and meta-analysis consortium; Psychiatric genomics consortium schizophrenia working group, Roddney JC, McEvoy LK, Desikan RS, Dale AM. 2013. Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. *Am J Hum Genet* **92**(2): 197-209.

Browning BL, Yu Z. 2009. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am J Hum Genet* **85**(6): 847-861.

Burkett KM, Greenwood CM, McNeney B, Graham J. 2013. Gene genealogies for genetic association mapping, with application to Crohn's disease. *Front Genet* **4**: 260, eCollection 2013.

Burkett KM, McNeney B. Graham J. 2013. Markov chain Monte Carlo sampling of gene genealogies conditional on unphased SNP genotype data. *Stat Appl Genet Mol Biol* **12**(5): 559-581.

Burmeister M, McInnis MG, Zöllner S. 2008. Psychiatric genetics: progress amid controversy. *Nat Rev Genet* **9**(7): 527-540.

Cao H, Wu J, Wang Y, Jiang H, Zhang T, Liu X, Xu Y, Liang D, Gao P, Sun Y, Gifford B, D'Ascenzo M, Liu X, Tellier LC, Yang F, Tong X, Chen D, Zheng J, Li W, Richmond T, Xu X, Wang J, Li Y. 2013. An integrated tool to study MHC region: accurate SNV detection and HLA genes typing in human MHC region using targeted high-throughput sequencing. *PLoS One* **8**(7): e69388.

Daw EW, Kumm J, Snow GL, Thompson EA, Wijsman EM. 1999. Monte Carlo Markov chain methods for genome screening. *Genet Epidemiol* **17**(suppl): S133-138.

Dizier MH, Etain B, Lajnef M, Lathrop M, Grozeva D, Craddock N, Henry C, Gard S, Jamain S, Leboyer M, Bellivier F, Mathieu F. 2012. Genetic heterogeneity according to age at onset in bipolar disorder: a combined positional cloning and candidate gene approach. *Am J Med Genet B Neuropsychiatr Genet* **159B**(6): 653-659.

Edwards SL, Beesley J, French JD, Dunning AM. 2013. Beyond GWASs: illuminating the dark road from association to function. *Am J Hum Genet* **93**(5): 779-797.

Fearnhead P, Donnelly P. 2001. Estimating recombination rates from population genetic data. *Genetics* **159**(3): 1299-1318.

Fernández SA, Fernando RL, Guldbrandtsen B, Totir LR, Carriquiry AL. 2001. Sampling genotypes in large pedigrees with loops. *Genet Sel Evol* **33**(4): 337-367.

Flint J, Munafo MR. 2007. The endophenotype concept in psychiatric genetics. *Psycol Med* **37**(2): 163-180.

Frazer KA Ballinger DG Cox DR Hinds DA Stuve LL Gibbs RA Belmont JW Boudreau A Hardenbol P Leal SM et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**(7164): 851-861.

Freimer N, Sabatti C. 2004. The use of pedigree, sib-pair and association studies of common disease for genetic mapping and epidemiology. *Nat Genet* **36**(10): 1045-1051.

Gelman A, Rubin DB. 1992. Inference from iterative simulation using multiple sequences. *Stat Sci* **7**(4): 457-472.

Gottesman II, Gould TD. 2003. The endophenotype concept in psychiatry: etymology and strategic intentions. *Am J Psychiatry* **160**(4): 636-645.

Guan Y, Stephens M. 2008. Practical issues in imputation-based association mapping. *PLoS Genet* **4**(12): e1000279.

Heath SC, Snow GL, Thompson EA, Tseng C, Wijsman EM. 1997. MCMC segregation and linkage analysis. *Genet Epidemiol* **14**(1): 1011-1016.

Hirschhorn JN, Daly MJ. 2005. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**(2): 95-108.

Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* **44**(8): 955-959.

Huang L, Jakobsson M, Pemberton TJ, Ibrahim M, Nyambo T, Omar S, Pritchard JK, Tishkoff SA, Rosenberg NA. 2011. Haplotype variation and genotype imputation in African populations. *Genet Epidemiol* **35**(8): 766-780.

Huang L, Li Y, Singleton AB, Hardy JA, Abecasis G, Rosenberg NA, Scheet P. 2009. Genotype-imputation accuracy across worldwide human populations. *Am J Hum Genet* **84**(2): 235-250.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**(2): 337-338.

Ishwaran H, Rao JS. 2005. Spike and slab variable selection: frequentist and Bayesian strategies. *Ann Stat* **33**(2): 730-773.

Jewett EM, Zawistowski M, Rosenberg NA, Zöllner S. 2012. A coalescent model for genotype imputation. *Genetics* **191**(4): 1239-1255.

Kerner B, Lambert CG, Muthen BO. 2011. Genome-wide association study in bipolar patients stratified by co-morbidity. *PLoS One* **6**(12): e28477.

Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* **83**(4): 311-321.

Li N, Stephens M. 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**(4): 2213-2233.

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. 2010. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* **34**(8): 816-834.

Li Y, Sidore C, Kang HM, Boehnke M, Abecases GR. 2011. Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res* **21**(6): 940-951.

Liu EY, Li M, Wang W, Li Y. 2013. MaCH-admix: genotype imputation for admixed populations. *Genet Epidemiol* **37**(1): 25-37.

Lydall GJ, Bass NJ, McQuillin A, Lawrence J, Anjorin A, Kandaswamy R, Pereira A, Guerrini I, Curtis D, Vine AE, Sklar P, Purcell SM, Gurling HM. 2009. Confirmation of prior evidence of genetics susceptibility to alcoholism in a genome-wide association study of comorbid alcoholism and bipolar disorder. *Psychiatr Genet* **21** (6): 294-306.

Maples BK, Gravel S, Kenny EE, Bustamante CD. 2013. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet* **93**(2): 278-288.

Marchini J, Howie B, Myers S, McVean G, Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**(7): 906-913.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* **9**(5): 356-369.

Morris AP, Lindgren CM, Zeggini E, Timpson NJ, Frayling TM, Hattersley AT, McCarthy MI. 2010. A powerful approach to sub-phenotype analysis in population-based genetic association studies. *Genet Epidemiol* **34**(4): 335-343.

Morris A, Pedder A, Ayres K. 2003. Linkage disequilibrium assessment via log-linear modeling of SNP haplotype frequencies. *Genet Epidemiol* **25**(2): 106-114.

Morris AP, Whittaker JC, Balding DJ. 2002. Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *Am J Hum Genet* **70**(3): 686-707.

Morris AP, Whittaker JC, Balding DJ. 2004. Little loss of information due to unknown phase for fine-scale linkage-disequilibrium mapping with single-nucleotide-polymorphism genotype data. *Am J Hum Genet* **74**(5): 945-953.

Nelsen MR, Wegmann D, Ehm MG, Kessner D, St Jean P, Verzilli C, Shen J, Tang Z, Bacanu SA, Fraser D, Warren L, Aponte J, Zawistowski M, Liu X, Zhang H, Zhang Y, Li J, Li Y, Li L, Woollard P, Topp S, Hall MD, Nangle K, Wang J, Abecasis G, Cardon LR, Zöllner S, Whittaker JC, Chissoe SL, Novembre J, Mooser V. 2012. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* **337**: 100-104.

Nielsen R. 2000. Estimation of population parameters and recombination rates for single nucleotide polymorphisms. *Genetics* **154**(2): 931-942.

Nordborg M. 2000. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics* **154**(2): 923-929.

Nordborg M, Tavare S. 2002. Linkage disequilibrium: what history has to tell us. *Trends Genet* **18**(2): 83-90.

Pasaniuc B, Avinery R, Gur T, Skibola CF, Bracci PM, Halperin E. 2010. A generic coalescent-based framework for the selection of a reference panel for imputation. *Genet Epidemiol* **34**(8): 773-782.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. 2007. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet* **81**(3): 559-575.

Saunders EH, Scott LJ, McInnis MG, Burmeister M. 2008. Familiality and diagnostic patterns of subphenotypes in the National Institutes of Mental Health bipolar sample. *Am J Med Genet B Neuropsychiatr Genet* **147B**(1): 18-26.

Saunders EF, Zhang P, Copeland JN, McLnnis MG, Zöllner S. 2009. Suggestive linkage at 9p22 in bipolar disorder weighted by alcohol abuse. *Am J Med Genet B, Neuropsychiatr Genet* **150B**(8): 1133-1138.

Scheet P, Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78**(4): 629-644.

Schulze TG, Hedeker D, Zandi P, Rietschel M, McMahon FJ. 2003. What is familial about familial bipolar disorder? Resemblance among relatives across a broad spectrum of phenotypic characteristics. *Arch Gen Psychiatry* **63**(12): 1368-1376.

Servin B, Stephens M. 2007. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet* **3**(7): e114.

Stephens M, Balding DJ. 2009. Bayesian statistical methods for genetic association studies. *Nat Rev Genet* **10**(10): 681-690.

Stephens M, Donnelly P. 2003. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* **73**(5): 1162-1169.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* **68**(4): 978-989.

Tachmazidou I, Verzilli CJ, De lorio M. 2007. Genetic associatoin mapping via evolution-based clustering of haplotypes. *PloS Genet* **3**(7): e111.

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**(7319): 1061-1073.

The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437**(7063): 1299-1320.

Thornton T, McPeek MS. 2010. ROADTRIPS: Case-control association testing with partially or completely unknown population and pedigree structure. *Am J Hum Genet* **86**(2): 172-184.

Tier B, Henshall J. 2001. A sampling algorithm for segregation analysis. *Genet Sel Evol* **33**(6): 587-603.

Uimari P, Hoeschele I. 1997. Mapping-linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms. *Genetics* **146**(2): 735-743.

Velez JI, Chandrasekharappa SC, Henao E, Martinez AF, Harper U, Jones M, Solumon BD, Lopez L, Garcia G, Aguirre-Acevedo DC, Acosta-Baena N, Correa JC, Lopera-Gomez CM, Jaramillo-Elorza MC, Rivera D, Kosik KS, Schork NJ, Swanson JM, Lopera F, Arcos-Burgos M. 2013. Pooling/bootstrap-based GWAS (pbGWAS) identifies new loci modifying the age of onset in PSEN1 p.Glu280Ala Alzheimer's disease. *Mol Psychiatry* **18**(5): 568-575.

Wang X, Kammerer CM, Anderson S, Lu J, Feingold E. 2009. A comparison of principal component analysis and factor analysis strategies for uncovering pleiotropic factors. *Genet Epidemiol* **33**(4): 325-331.

Wellcome Trust Case Control Consortium, Maller JB, McVean G, Byrnes J, Vukcevic D, Palin K, Su Z, Howson JM, Auton A, Myers S, Morris A, Pirinen M, Brown MA, Burton PR, Caulfield MJ, Compston A, Farrall M, Hall AS, Hattersley AT, Hill AV, Mathew CG, Pembrey M, Satsangi J, Stratton MR, Worthington J, Craddock N, Hurley M, Ouwehand W, Parkes M, Rahman N, Duncanson A, Todd JA, Kwiatkowski DP, Samani NJ, Gough SC, McCarthy MI, Deloukas P, Donnelly P. 2012. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet* **44**(12): 1294-1301.

Wen X, Stephens M. 2010. Using linear predictors to impute allele frequencies from summary or pooled genotype data. *Ann Appl Stat* **4**(3): 1158-1182.

Wen Y, Lu Q. 2013. A multiclass likelihood ratio approach for genetic risk prediction allowing for phenotypic heterogeneity. *Genet Epidemiol* **37**(7): 715-725.

Willour VL, Zandi PP, Badner JA, Steele J, Miao K, Lopez V, MacKinnon DF, Mondimore FM, Schweizer B, McInnis MG, Miller EB, DePaulo JF Jr, Gershon ES, McMahon FJ, Potash JB. 2007. Attempted suicide in bipolar disorder pedigrees: evidence for linkage to 2p12. *Biol Psychiatry* **61**(5): 725-727.

Zandi PP, Badner JA, Steele J, Willour VL, Miao K, MacKinnon DF, Mondimore FM, Schweizer B, McInnis MG, DePaulo JR Jr, Gershon E, McMahon FJ, Potach JB. 2007. Genome-wider linkage scan of 98 bipolar pedigrees and analysis of clinical covariates. *Mol Psychiatry* **12** (7): 630-639.

Zöllner S. 2012. Sampling strategies for rare variant tests in case-control studies. *Eur J Hum Genet* **20**(10): 1085-1091.

Zöllner S, Pritchard JK. 2005. Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics* **169**(2): 1071-1092.