LINKAGE ANALYSIS FOR CATEGORICAL TRAITS AND ANCESTRY ASSIGNMENT IN ADMIXED INDIVIDUALS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Abra Gail Brisbin

May 2010

LINKAGE ANALYSIS FOR CATEGORICAL TRAITS AND ANCESTRY ASSIGNMENT IN ADMIXED INDIVIDUALS

Abra Gail Brisbin, Ph.D.

Cornell University 2010

A major goal in genetics is the identification of loci that contribute to diseases and other traits. With my Ph.D. research, I have developed methods that address two important challenges in this search: First, I addressed the challenge of choosing an appropriate disease model by developing a Gibbs sampler and an elimination algorithm to perform linkage analysis for categorical traits. Second, I addressed the challenge of population stratification due to admixture by developing a Principal Components-based approach to the assignment of ancestry at local regions along the genome of phased haplotypes in admixed individuals.

Choosing an appropriate disease model is critical for maximizing power to detect disease loci. Many complex heritable diseases feature nominal or ordinal phenotypic measurements for which traditional methods of linkage analysis, which model traits as binary or continuous, are not well-suited. To address this challenge, I developed a Gibbs sampling approach (LOCate) and an elimination algorithm approach (LOCate2) to assess linkage for categorical traits. I validated the methods on simulated data and found that my approaches have increased power versus existing methods for ordinal linkage analysis. I also used these methods to analyze several data sets of categorical traits in humans and dogs, and found increased LOD scores at candidate loci when the traits were treated as categorical rather than binary. This will be useful for mapping

genes for many complex traits.

Identifying ancestry along each chromosome in admixed individuals is of interest for admixture mapping, understanding the population genetic history of admixture events, and identifying recent targets of selection. I developed a Principal Components-based forward-backward algorithm for determining local ancestry from a high-density, genomewide set of SNP genotypes of admixed individuals. Simulations show that the method is robust to misspecification of ancestral populations and the number of generations since admixture. I also applied my method to assess 3-way European, Native American, and African admixture among four Latino populations, and identified regions of extreme levels of African and Native American ancestry which may have experienced selection during admixture. This method is fast, accurate, and applicable to phased haplotypes with admixture from two or more populations.

BIOGRAPHICAL SKETCH

Abra Brisbin was born on March 24, 1982 in St. Louis Park, Minnesota. She grew up in Minnesota with her parents, Gary and Gail Brisbin, and younger brother Drew. She attended school in the Spring Lake Park School District, where she was fortunate to have a strong cohort in the Gifted and Talented program and several excellent teachers who encouraged her interest in learning in general, and in mathematics in particular.

Abra attended Carleton College, where after much soul-searching, she decided to major in math. The encouragement of her math professors, combined with her experiences at the Mayo Clinic Summer Undergraduate Research Fellowship and the George Washington University Summer Program for Women in Mathematics, helped her decide to attend graduate school. After a summer spent as a teaching assistant for the Carleton Summer Math Program for Women and a year as a fifth-year intern for Carleton's math department, she moved to Cornell University. She joined the field of Applied Mathematics and, shortly thereafter, the labs of Jason Mezey and Carlos Bustamante. The fruits of her efforts in those labs are the thesis you now hold in your hands.

To my family

ACKNOWLEDGEMENTS

I would like to thank my advisors, Carlos Bustamante and Jason Mezey, for their mentorship over the past five years. I would also like to thank Dexter Kozen for his assistance on my thesis committee. Many thanks go to my labmates in the Mezey and Bustamante labs, especially Kasia Bryc, Jeremiah Degenhardt, Shaila Musharoff, Ben Logsdon, Larsson Omberg, Kirk Lohmueller, Adam Boyko, Ryan Hernandez, and Lan Zhu, for their support and excellent scientific suggestions. I would also like to thank my other collaborators on the papers included in this thesis: Myrna M. Weissman, Abby J. Fyer, Steven P. Hamilton, James A. Knowles, Jenifer Cruickshank, N.Sydney Moïse, Teresa Gunn, and Ausra Milano. They generously allowed me to work with their datasets, putting the “applied” into my Applied Math research.

I thank all of my teachers at the Spring Lake Park School District, who encouraged me in my education and created a learning environment in which being smart was a positive thing. Special thanks go to Jeanne Bennett, Mike Hilst, Tom Young, Mike Crofton, and Jeff Witt, who made science stand out for me and were instrumental in my decision to major in math in college. I also thank my professors at Carleton College, especially Bob Dobrow, Mark Krusemeyer, Deanna Haunsperger, Steve Kennedy, Gail Nelson, and Stephan Zweifel. It is largely due to their inspirational teaching that I aspire to become a professor. I also thank Armando Manduca, my advisor at the Mayo Clinic Summer Undergraduate Research Fellowship, and my mentors at the George Washington Summer Program for Women in Mathematics, especially Murli Gupta and Angela Gallegos. The experience I gained working with them added an extra dimension to my scientific understanding, and helped convince me that graduate school was something I could do.

Finally, I want to thank my family. They set an example of curiosity and critical thinking that will last me a lifetime, and supported me throughout this adventure of graduate school. Special gratitude goes to my grandfather, George Brisbin, the first scientist in my life, and to my parents and brother. Mom, Dad, and Drew: You're still three of the most interesting people to talk with about science that I've ever met.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Recent years have seen an explosion of research into the genetic causes of diseases and other traits, with many successes, including the identification of Complement Factor H as a contributor to age-related macular degeneration [17], BRCA1 and 2 as factors in breast cancer [32, 75], and IGF1 as a factor in small body size in dogs [65]. However, there is still a great deal to be discovered: With the exception of Mendelian traits such as sickle-cell anemia [53, 38] or Huntington's disease [30], most identified loci explain only a small fraction of the heritability their traits exhibit. For example, human height has a heritability of 80%, meaning that the correlation between the average of parents' height and the height of the child (conditional on the child's gender) is $r = .8$, a highly predictive correlation. Yet the 54 loci that have been identified to contribute to height explain only 5% of this heritability [46]. Clearly, there is a need for additional research and improved methods to enhance the search for causative loci.

In this thesis, I present three new methods to aid in the search for loci that contribute to diseases and other traits. The rest of this introduction gives an overview of linkage analysis, categorical traits, and admixture. Chapters 2 and 3 describe approaches for linkage analysis for categorical traits. Chapter 4 describes a method for assigning ancestry along chromosomes in admixed individuals.

1.1 Linkage Analysis

Linkage analysis refers to the identification of co-transmission within a pedigree of a trait and a genetic marker or markers in order to identify genetic linkage between the genetic markers and the quantitative trait locus (QTL), or the genetic location that directly affects the trait. Linkage analysis has been used to identify loci linked to traits as diverse as breast cancer [32], macular degeneration [17], and hip dysplasia [85]. The use of pedigree data affords protection against the reduced power and high type I error rate that can be the result of genetic heterogeneity and population stratification in association mapping studies. For this reason, investigations of loci contributing to a trait often first employ linkage analysis to identify candidate regions, which may be followed by association mapping to obtain finer resolution in candidate regions.

Single Marker Analysis (SMA) is a form of linkage analysis which uses information from one genetic marker at a time. The goal is to infer θ, the probability of recombination between the marker and the QTL. θ can range from 0 to .5. With a map that aligns genetic and physical distances, such as [22], θ can be used as a proxy for the physical distance between the marker and the QTL. To find the maximum likelihood estimate of θ, we compute $P(X \mid \theta)$, the probability of the observed data (marker genotypes and phenotypes) conditional on θ. To do this, we use the equation

$$P(X \mid \theta) = \Sigma_Y P(X, Y \mid \theta) \tag{1.1}$$

where Y is the set of possible configurations of unobserved QTL genotypes on the pedigree. However, $|Y|$ grows exponentially with n, the number of individuals in the family ($|Y|$ is approximately 3^n, ignoring the fact that some of these configurations will involve Mendelian inconsistencies). Therefore, it is important to perform this summation in an efficient way, for example, through the use of Elston-Stewart peeling [18].

Many software programs are available to conduct linkage analysis for binary and quantitative traits. Superlink [21] uses the Elston-Stewart (peeling by nuclear families) and Lander-Green (peeling by locus) algorithms to eliminate variables, in order to perform single marker analysis (SMA) and interval mapping (IM) for binary traits. Merlin [1] performs SMA and IM for binary and quantitative traits, using the Lander-Green algorithm on systematically condensed binary trees describing the gene flow through the pedigree. SOLAR [2] uses a variance-components analysis to perform IM for quantitative traits. Loki [35] is a Markov chain Monte Carlo (MCMC) method for estimating the locations and number of quantitative trait loci (QTL) affecting a continuous trait with normally distributed residuals.

1.2 Categorical Traits

Categorical traits are those in which phenotypes fall into more than two discrete categories. The categories may be ordered, such as “mild,” “moderate,” or “severe,” or unordered, such as color. The former are also known as ordinal traits, and the latter are known as nominal traits. Many traits of interest, from pathogen resistance in plants [77] to panic disorder in humans [23], have a natural means of classification as categorical traits. Methods designed for binary or quantitative traits are not expected to be effective for categorical traits. Assigning the categories of a categorical trait to the dichotomy of a binary trait involves a loss of information, so binary-trait methods are likely to suffer from reduced power when used to search for loci affecting a categorical trait [12, 20]. Methods designed for quantitative traits assume that the phenotype is normally, or otherwise continuously, distributed conditional on the genotype, which is a poor model for the discrete categories of ordinal data. Therefore, continuous-trait methods are also likely to suffer from reduced power for ordinal traits.

Most previous work done on family-based mapping of categorical traits has been restricted to particular types of pedigrees; these include backcross [31, 42, 78] and F2 designs [81, 42, 78, 34], 4-way experimental crosses [60, 77, 80, 79], and sets of independent nuclear families [59, 82, 74]. Many of these can more appropriately be classified as family-based association testing, as they rely on equal levels of relatedness among tested individuals, and their logical extensions involve applying the methods to populations of unrelated individuals, not to extended pedigrees. Recent methods by Zhang *et al.* [83], Dupuis *et al.* [16], and Diao and Lin [15] allow linkage analysis for ordinal traits on arbitrary pedigrees.

QTLlink [16] is designed for continuous quantitative traits, but with a score statistic that is more robust [69] to the departures from normality presented by categorical traits, compared to traditional LOD (log of odds) score calculations. It uses a variance-component model, which models the expectation of individuals' phenotypes as $E(Y|X) = m + aX$, where Y is the vector of phenotypes, m is the mean phenotype, and X is the set of observed covariates . The covariance matrix of $Y \mid X$ depends on $\alpha = \sigma_a^2 + \sigma_d^2$, the variance components due to a particular locus. (α may contain additional terms if testing for interactions between loci.) The goal of the variance-components approach is to test whether $\alpha = 0$, that is, whether a particular locus has an effect on the phenotypic variance. QTLlink does this with a score statistic, taking the derivative of the likelihood with respect to α at $\alpha = 0$, and normalizing by the square root of the variance. The variance is computed conditional on the phenotypes, which Tang and Siegmund [69] found to make the score statistic more robust to departure from normality. Under the null model (true $\alpha = 0$), we expect the likelihood to be maximized at $\alpha = 0$, so we expect that its derivative at $\alpha = 0$ will equal 0. Therefore, a score statistic significantly different from 0 is taken as evidence that the locus is linked to the trait. This approach reduces the number of parameters that must be estimated, compared to using a likelihood ratio test, as the expectation and covariance of the phenotypes must be estimated only at $\alpha = 0$.

Diao and Lin [15] present another variance-component model, which is specifically designed for ordinal traits, by the use of a liability threshold. Unlike QTLlink, Diao and Lin use a likelihood ratio test instead of a score statistic. They use a quasi-Newton method to obtain the maximum likelihood estimate for the parameters. Diao and Lin's model also incorporates between- and within-family association components, which allow for joint association mapping.

LOT [83] does not use a variance-component approach; it models gene transmission through the pedigree via inheritance vectors as in Genehunter [41] rather than via a covariance matrix that depends on kinship coefficients. It uses a proportional-odds logistic model that is similar to the liability threshold model of Diao and Lin, though the latter uses a probit model. Like the method of Diao and Lin, LOT incorporates a family-specific environmental effect, though this effect contributes to the trait mean instead of its variance. (Here, “environmental” refers to any unobserved covariate or genetic background effect on the trait.) LOT uses a likelihood ratio test to assess linkage, like Diao and Lin, and uses an expectation-maximization (EM) algorithm to identify the maximum-likelihood estimates of the parameters. The EM algorithm typically converges more slowly than Newton-type methods such as those used by Diao and Lin. (The accuracy of the EM algorithm is linear in the number of iterations [14], versus quadratic for Newton’s method (p. 207 of [50]).) Unlike Newton’s method, however, iterations of the EM algorithm never decrease the likelihood of the estimates [14], so the EM algorithm is guaranteed to converge to a local maximum or saddle point, while Newton’s method can fail to converge.

The proportional-odds logistic model used by LOT means that if there are no covariates or family-specific environmental effect, the penetrances (probabilities of each phenotype) follow the model

$$\text{logit}(P(Y_j^i \leq k)) = \alpha_k - \gamma U_j^i \tag{1.2}$$

where Y_j^i is the phenotype of individual j in family i, α_k is a trait level-specific intercept, γ is the effect of the locus, and U_j^i is the individual’s genotype at the locus (0, 1, or 2 copies of the disease allele). This means that the penetrance matrix looks like that shown in Table 1.1.

$\alpha_{K=\max k} = \infty$, so the probability of having phenotype less than or equal to the maximum is 1. γ does not depend on the trait level k, which is what makes the model a “proportional-odds” model: The odds (probability/(1-probability)) of having a phenotype larger than a given value k, given U copies of the disease allele, is e^γ times the odds given $U - 1$ copies of the disease allele, for all $k = 1, 2, ...K$ and $U = 1, 2$. This is a sensible model for ordinal traits because the

Table 1.1: Penetrance matrix for a proportional-odds logistic model.

qq, Qq, and QQ represent the genotype at the disease locus.

	qq	Qq	QQ
$P(Y = 1 \mid \text{genotype})$	$\frac{e^{\alpha_1}}{1+e^{\alpha_1}}$	$\frac{e^{\alpha_1-\gamma}}{1+e^{\alpha_1-\gamma}}$	$\frac{e^{\alpha_1-2\gamma}}{1+e^{\alpha_1-2\gamma}}$
$Y = 2$	$\frac{e^{\alpha_2}}{1+e^{\alpha_2}} - \frac{e^{\alpha_1}}{1+e^{\alpha_1}}$	$\frac{e^{\alpha_2-\gamma}}{1+e^{\alpha_2-\gamma}} - \frac{e^{\alpha_1-\gamma}}{1+e^{\alpha_1-\gamma}}$	$\frac{e^{\alpha_2-2\gamma}}{1+e^{\alpha_2-2\gamma}} - \frac{e^{\alpha_1-2\gamma}}{1+e^{\alpha_1-2\gamma}}$
...	...	...	...
$Y = K$	$1 - \Sigma_{k=1}^{K-1} \frac{e^{\alpha_k}}{1+e^{\alpha_k}}$	$1 - \Sigma_{k=1}^{K-1} \frac{e^{\alpha_k-\gamma}}{1+e^{\alpha_k-\gamma}}$	$1 - \Sigma_{k=1}^{K-1} \frac{e^{\alpha_k-2\gamma}}{1+e^{\alpha_k-2\gamma}}$

trait is parameterized according to an underlying "severity" that is a function of $U\gamma$, which must exceed the threshold α_k to produce a phenotype at least as severe as k. The ordering of $\alpha_1 < \alpha_2 < ... < \alpha_K$ reflects the ordered quality of the trait categories. In contrast, a nominal trait has no ordering to its phenotypic categories and thus no restrictions on its penetrance matrix.

The proportional-odds quality of ordinal trait models makes it convenient to characterize these models in terms of odds ratios (ORs). The odds ratio is the ratio of the odds of an event (in this case, the phenotype being larger than k) given a risk factor, to the odds of the event without that risk factor. We are interested in the risk factor of having one disease allele (genotype Qq) compared to having no disease alleles (genotype qq), and in the risk factor of having two disease alleles (genotype QQ) compared to having only one. Proportional odds means that the OR will be the same for both of these comparisons, as well as for each phenotype level k, so the penetrance model can be characterized in terms of a single OR, which depends on γ (but not on α). In contrast, a 3-level nominal trait has 4 (potentially) distinct ORs: $\frac{\text{Odds}(Y>1|Qq)}{\text{Odds}(Y>1|qq)}$, $\frac{\text{Odds}(Y>2|Qq)}{\text{Odds}(Y>2|qq)}$, $\frac{\text{Odds}(Y>1|QQ)}{\text{Odds}(Y>1|Qq)}$, $\frac{\text{Odds}(Y>2|QQ)}{\text{Odds}(Y>2|Qq)}$. Odds ratios are a way of describing the strength of the correlation between the risk factor and the phenotype. With all other factors (such as sample size) being equal, traits with ORs that are very different from 1 will be easier to map to genes.

Because LOT and the method of Diao and Lin use ordinal models, they are apt to suffer reduced power when used to analyze traits which are nominal rather than ordinal. I demonstrate this for LOT in chapters 2 and 3. (I was unable to test Diao and Lin's method, as these authors did not respond to requests for copies of their software.) As I show in chapter 3, LOT also experiences reduced power when analyzing an ordinal trait with incomplete linkage between the marker and trait locus, as LOT does not offer the option of computing linkage scores at ungenotyped loci. QTLlink's robust score statistic is well-powered for analysis of ordinal traits, but experiences reduced power for nominal traits, as I show in chapter 3. In chapters 2 and 3, I present two alternative approaches which allow a unified approach to linkage analysis for ordinal and nominal traits. These approaches have excellent power to analyze nominal traits as well as ordinal traits.

1.2.1 Markov Chain Monte Carlo and Gibbs Sampling

Markov Chain Monte Carlo (MCMC) refers to a broad class of methods for generating simulations from a desired probability distribution. An MCMC algorithm is a stochastic process in which each element of the process (i.e., each assignment of values to the set of random variables) depends stochastically on the previous element–hence, a Markov chain–in such a way that if X_i, the set of random variables on iteration i, was drawn from the desired probability distribution, then the marginal distribution of X_{i+1} will also be the desired probability distribution. When this condition is met, the desired distribution is the stationary distribution for the chain. If the chain is ergodic, then the stationary distribution is unique and is the limiting distribution for the chain. That is, as the number of iterations i goes to infinity, X_i will converge in distribution to the stationary distribution [61].

In practice, MCMC chains are typically run for a burn-in period, which allows the distribution of the first sampled element, X_n, to become close to the stationary distribution P^*, independent of the starting value of the chain:

$$P(X_n \mid X_0) \rightarrow P^*(X_n) \text{ as } n \rightarrow \infty.$$

The length of the burn-in period required depends on how well the chain "mixes". Chains in which the correlation between successive iterations is low are said to mix quickly, and they require shorter burn-in periods than chains with higher levels of autocorrelation. Determining whether a given length of burn-in is sufficient will be discussed below.

After the burn-in period, the chain is sampled. A wide variety of properties of the distribution P^* can be estimated by taking the mean of the property over the sampled values:

$$\hat{f}(Y) = \Sigma_{i=n+1}^{N+n} \frac{f(X_i)}{N}, \tag{1.3}$$

where f is a function and Y is a random variable with distribution P^* (denoted $Y \sim P^*$). By the Ergodic theorem [61], if the X_is are drawn from an ergodic Markov chain with stationary distribution P^* and

$$E_{P^*}(f(Y)) < \infty,$$

where

$$E_{P^*}(f(Y)) = \Sigma_j f(j) P^*(j),$$

then $\hat{f}(Y)$ converges to $E_{P^*}(f(Y))$ with probability 1 as the number of samples $N \rightarrow \infty$. Therefore, samples from the MCMC chain allow estimation of properties of P^*. It is not necessary that the samples be independent for this estimator to be valid, although chains with lower autocorrelations typically require fewer iterations to produce a good estimate. As I will discuss in chapter 2, the likelihood $P(\text{data} \mid \theta)$ cannot be conveniently formulated as the expectation of a function of the unobserved random variables, so it is not amenable to direct estimation by Equation 1.3. However, the conditional probability $P(Y_i \mid X, \theta)$, where Y_i is a configuration of the unobserved variables and X is the observed data, can be formulated as

$$P(Y_i \mid X, \theta) = E_{P^*}(I_{Y=Y_i})$$

where

$$I_{Y=Y_i} = \begin{cases} 1 & \text{if } Y = Y_i \\ \\ 0 & \text{else.} \end{cases}$$

Therefore, I employ Equation 1.3 in the estimation of the likelihood, via the estimation of $P(Y_i \mid X, \theta)$.

To be ergodic, a Markov chain must meet three conditions. First, it must be aperiodic, meaning that the sequence of iterations on which it is possible (probability > 0) to visit any state i must have a greatest common divisor of 1. In practice, this is not a problem if there are many transition probabilities strictly between 0 and 1, which is typically the case in MCMC. Second, the chain must be positive recurrent, meaning that the expected number of iterations required to return to any given state is finite. If the number of possible states is finite, then positive recurrence is guaranteed by irreducibility. This is the case discussed in Chapter 2, where the states consist of configurations of unobserved genotypes, a number that can be large, depending on the size of the pedigree, but is always finite. Finally, the chain must be irreducible, meaning that from any given starting state, any other state has positive probability of being reached at some point in the future. This is the condition of greatest concern in chapter 2, as Markov chains on pedigrees with more than two marker alleles can be reducible if there are individuals with missing marker genotypes, as described in [72, 73]. The problem of reducibility is an extreme case of the problem of slow mixing, in which the Markov chain requires an excessive number of iterations to transition from one part of the sample space to another, and successive samples from the chain are highly correlated. In chapter 2, I address this problem using the technique of simulated tempering.

As mentioned above, MCMC chains are typically run for an initial "burn-in" period, during which no samples are collected, to allow the chain to converge to its stationary distribution. There are several ways to assess whether a burn-in period has been sufficient (reviewed in pp. 370-374 of [50]); the most common of these is the use of Gelman-Rubin statistics, which compare the variance across several chains to the variance within one chain. If the variances are similar, as measured by a ratio close to 1, then the chains are considered to be from the same distribution, and thus the burn-in period was sufficient for the chains to reach their stationary distribution. If the burn-in period was insufficient, then the variance across chains will be larger than the variance within chains, and the ratio will be larger than 1.

To illustrate the meaning of Gelman-Rubin statistics, consider the simple Markov chain shown in table 1.2. There is only one variable, x, which takes values 1 through 6. The probability that $x = j$ on iteration $t + 1$, given that

$x = i$ on iteration t, is given by the entry in the ith row, jth column of the transition matrix. Note that the states fall into two sets: $\{1, 2, 4\}$ and $\{3, 5, 6\}$, and the probability of transitioning between these sets is very low (.01 if $x = 3$ or 4, 0 otherwise). Starting one chain at $x = 1$ and another at $x = 6$ produces very different results for the first 200 iterations (Figure 1.1a), producing a Gelman-Rubin statistic of $\hat{R}^{1/2} = 1.42$. The Gelman-Rubin statistic is much greater than 1, indicating that 200 iterations is not a sufficient burn-in period for the chains to converge to their stationary distribution. In contrast, running the same chains for 1000 iterations produces the results shown in Figure 1.1b. The chains now have similar distributions, independent of the starting value of x. The Gelman-Rubin statistic is $\hat{R}^{1/2} = 1.02$, indicating that a burn-in of 1000 iterations is sufficient. However, the chains still look “blocky”; the values of x on successive iterations are highly correlated. This slow mixing is due to the low rate of transitions between the sets $x \in \{1, 2, 4\}$ and $x \in \{3, 5, 6\}$.

Table 1.2: Transition matrix for a simple Markov chain.

	1	2	3	4	5	6
1	.5	.25	0	.25	0	0
2	.25	.5	0	.25	0	0
3	0	0	.33	.01	.33	.33
4	.33	.33	.01	.33	0	0
5	0	0	.25	0	.5	.25
6	0	0	.25	0	.25	.5

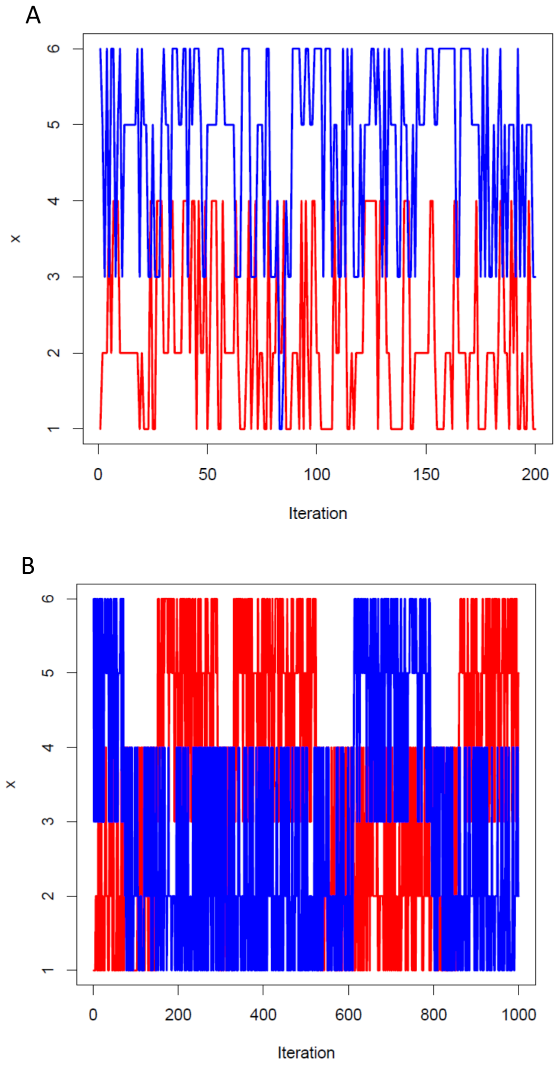

Figure 1.1: Gelman-Rubin statistics for the Markov chain in Table 1.2.

A. After 200 iterations, the pair of chains has a Gelman-Rubin ($\hat{R}^{1/2}$) statistic of 1.42. **B.** After 1000 iterations, the pair of chains has $\hat{R}^{1/2} = 1.02$.

Metropolis-Hastings and Gibbs sampling

The most common form of MCMC is the Metropolis-Hastings algorithm [33]. In the Metropolis-Hastings algorithm, a new state y is proposed according to an arbitrary proposal distribution $Q(y \mid x_i)$, which may depend on the current state x_i. The new state is accepted with probability

$$\alpha = P(x_{i+1} = y \mid x_i) = \min(1, \frac{P^*(y)Q(x_i \mid y)}{P^*(x_i)Q(y \mid x_i)}). \tag{1.4}$$

With probability $1 - \alpha$, the new state is rejected and $x_{i+1} = x_i$. Here $P^*(y)$ is the probability of y under the desired distribution. This probability may be difficult to calculate, which is a common reason to want to sample from the distribution instead. Fortunately, the Metropolis-Hastings algorithm relies on the ratio $\frac{P^*(y)}{P^*(x_i)}$, so there is no need to compute the normalizing constant for $P^*(y)$.

To see that the Metropolis-Hastings algorithm has the desired distribution, P^*, as its stationary distribution, first note that the transition probability, $p(x_{i+1} \mid x_i)$, is reversible with respect to the desired distribution P^*:

$$\begin{aligned}
P^*(x_i)p(x_{i+1} \mid x_i) &= P^*(x_i)Q(x_{i+1} \mid x_i)\alpha(x_{i+1} \mid x_i) \qquad (1.5)\\
&= P^*(x_i)Q(x_{i+1} \mid x_i)\min(1, \frac{P^*(x_{i+1})Q(x_i \mid x_{i+1})}{P^*(x_i)Q(x_{i+1} \mid x_i)})\\
&= \min(P^*(x_i)Q(x_{i+1} \mid x_i), P^*(x_{i+1})Q(x_i \mid x_{i+1}))\\
&= P^*(x_{i+1})Q(x_i \mid x_{i+1})\min(\frac{P^*(x_i)Q(x_{i+1} \mid x_i)}{P^*(x_{i+1})Q(x_i \mid x_{i+1})}, 1)\\
&= P^*(x_{i+1})Q(x_i \mid x_{i+1})\alpha(x_i \mid x_{i+1})\\
&= P^*(x_{i+1})p(x_i \mid x_{i+1})
\end{aligned}$$

Then

$$\begin{aligned}
\Sigma_{x_i} p(x_{i+1} \mid x_i) P^*(x_i) &= \Sigma_{x_i} p(x_i \mid x_{i+1}) P^*(x_{i+1}) \text{ by reversibility} && (1.6) \\
&= P^*(x_{i+1}) \Sigma_{x_i} p(x_i \mid x_{i+1}) \\
&= P^*(x_{i+1}) \cdot 1
\end{aligned}$$

This implies that if $x_i \sim P^*$, then $x_{i+1} \sim P^*$. Therefore, P^* is the stationary distribution.

Gibbs sampling [25] is a subcategory of the Metropolis-Hastings algorithm in which each transition step updates a subset of the variables, and the proposal distribution Q is the desired stationary distribution P^*, conditioned on all other variables: $Q(x_j^{t+1} \mid x^t) = P^*(x_j \mid x_{-j})$, where t is the current iteration, j is the subset of variables being updated, and $-j$ refers to the complement of j. The acceptance rate is then

$$\begin{aligned}
\alpha &= \min(1, \frac{P^*(x^{t+1}) Q(x^t \mid x^{t+1})}{P^*(x^t) Q(x^{t+1} \mid x^t)}) && (1.7) \\
&= \min(1, \frac{P^*(x^{t+1}) P^*(x_j^t \mid x_{-j}^{t+1}) I_{x_{-j}^t = x_{-j}^{t+1}}}{P^*(x^t) P^*(x_j^{t+1} \mid x_{-j}^t) I_{x_{-j}^{t+1} = x_{-j}^t}})
\end{aligned}$$

Using the fact that $x_{-j}^{t+1} = x_{-j}^t$, then

$$\begin{aligned}
\alpha &= \min(1, \frac{P^*(x^{t+1}) P^*(x_j^t \mid x_{-j}^{t+1})}{P^*(x^t) P^*(x_j^{t+1} \mid x_{-j}^t)}) && (1.8) \\
&= \min(1, \frac{P^*(x_j^{t+1}, x_{-j}^{t+1}) P^*(x_j^t \mid x_{-j}^{t+1})}{P^*(x_j^t, x_{-j}^t) P^*(x_j^{t+1} \mid x_{-j}^t)}) \\
&= \min(1, \frac{P^*(x_j^{t+1} \mid x_{-j}^{t+1}) P^*(x_{-j}^{t+1}) P^*(x_j^t \mid x_{-j}^{t+1})}{P^*(x_j^t \mid x_{-j}^t) P^*(x_{-j}^t) P^*(x_j^{t+1} \mid x_{-j}^t)}) \\
&= \min(1, \frac{P^*(x_j^{t+1} \mid x_{-j}^t) P^*(x_{-j}^t) P^*(x_j^t \mid x_{-j}^{t+1})}{P^*(x_j^t \mid x_{-j}^{t+1}) P^*(x_{-j}^t) P^*(x_j^{t+1} \mid x_{-j}^t)}) \\
&= \min(1, 1)
\end{aligned}$$

Therefore, Gibbs sampling is a form of Metropolis-Hastings sampling in which the acceptance rate is 1. In chapter 2, I apply Gibbs sampling to sample disease locus genotypes in families. I also apply the Metropolis-Hastings algorithm to explore different “temperatures” in my implementation of simulated tempering.

1.2.2 The Elimination Algorithm

The elimination algorithm (reviewed in [39]) is an efficient means of summing probabilities over possible states at nuisance variables to compute the exact total probability of observed data. When applied to linkage analysis, as I have done in chapter 3, the nuisance variables are the unknown disease locus genotypes of each member of the family. In this situation, the elimination algorithm can be thought of as a generalized form of the Elston-Stewart algorithm [18], which “peels” information within each nuclear family that is a subset of the larger pedigree onto one member of the nuclear family. The elimination algorithm is more general because it allows this “peeling” even in pedigrees with inbreeding loops. The elimination algorithm can greatly speed up computation in large pedigrees because, by marginalizing over one variable at a time, it reduces the number of terms required for the total summation. An example of this can be found in the Appendix “Supplementary information for chapter 3”.

1.3 Admixture

Admixed individuals are those who have ancestry from two or more distinct populations. For example, African Americans are admixed individuals because they have ancestry from Africa and Europe. This admixture results in distinct blocks of DNA from each population within their genomes. Identifying these distinct blocks, or “ancestry tracts”, is useful for identifying loci associated with traits that occur at different frequencies in the two ancestral populations, via admixture mapping, and for answering population genetic questions about the ancestral populations and the admixture event.

1.3.1 Admixture Mapping

Association mapping refers to identifying loci that are statistically associated with a trait by identifying correlations between individuals’ phenotypes and their genotypes at the loci. Association mapping commonly involves studies of unrelated individuals (though family-based association tests are also in practice). Two of the most common approaches are Fisher’s exact tests for binary traits and linear regression for continuous traits. It is well-known that population structure can introduce false positives in association mapping: If a trait is more frequent in one subpopulation than another, then any locus with different allele frequencies in the two subpopulations will appear to be associated with the trait if subpopulation membership is not taken into account (Figure 1.2). This can be a particular challenge when mapping loci in admixed populations, as subpopulation membership is not composed of individuals, but of portions of individuals’ genomes.

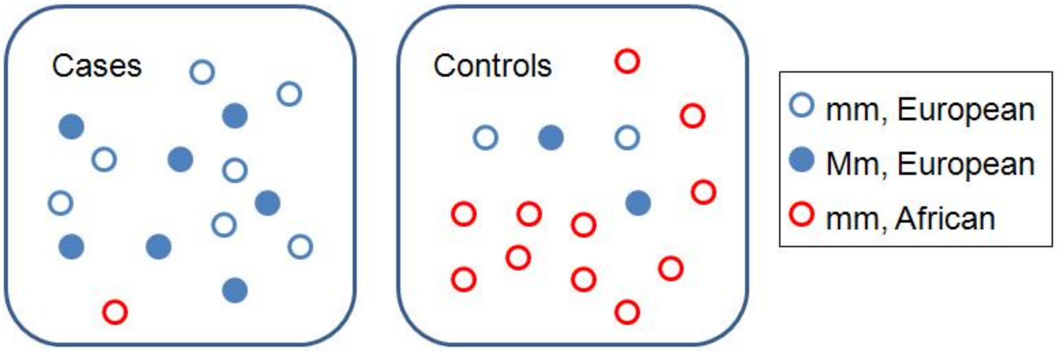

Figure 1.2: Population structure can produce false positives in association mapping.

In this example, the M allele (solid circles) appears to be associated with the disease ("cases"), because both the M allele and the disease are more frequent in European samples.

One way to account for subpopulation membership in this situation is admixture mapping [9]. This refers to identifying correlations between individuals' phenotypes and their ancestry at particular loci (Figure 1.3). In order to do this, it is necessary to identify the ancestry at particular points along the genome.

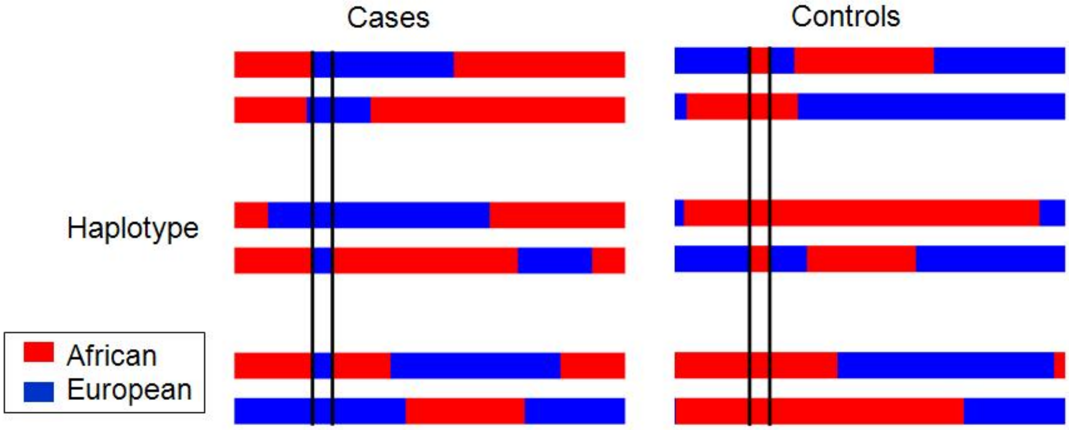

Figure 1.3: Admixture mapping identifies correlations between phenotypes and ancestry.

In this example, European ancestry in the region between the black lines is correlated with the disease phenotype.

1.3.2 Population Genetics of Admixture

Identifying ancestry along the genome is also of interest for population genetic questions. By identifying an individual's segments of ancestry from a particular population, those segments can be used as samples from that population, and the distribution of allele frequencies and stretches of linkage disequilibrium (LD) can contribute to inferences about population size and natural selection. The number of heterozygous sites can be used to infer the time to the most recent common ancestor [70] of the lineages contributing each of the individual's haplotypes. This could be used to infer population divergence times from genomic regions where the individual has mixed ancestry. Finally, the length of the admixture tracts themselves provides information about the number of generations since admixture occurred, as recombination tends to break up long ancestry tracts [55].

1.3.3 Methodology

There are many excellent methods available for assigning ancestry along the genome, including *structure* [57, 19], SABER [67], HAPMIX [56], ADMIXMAP [36], LAMP [63], ANCESTRYMAP [52], HAPAA [64], and the principal components analysis (PCA) based method of Bryc *et al.* [7].

Structure [19] uses MCMC to infer each individual's average ancestry proportions and the "chunk size" of ancestry blocks, which relates to the recombination rate, and then uses a Hidden Markov Model (HMM) to assign ancestry to blocks of the genome. It does not require samples from the ancestral populations.

SABER accounts for linkage disequilibrium between every pair of adjacent loci via a Markov-Hidden Markov model to model each observed allele as dependent upon the current ancestral state and, if the previous allele was generated by the same ancestral state, upon the previous allele.

ADMIXMAP and ANCESTRYMAP are designed specifically for trait mapping in admixed populations. ADMIXMAP uses a generalized linear model to relate a trait value to an individual's level of admixture at the loci, while ANCESTRYMAP uses MCMC to perform inference on the parameters of average ancestry proportion and recombination rate, and an HMM to assign ancestry. Both of them work with a restricted number ($\lesssim 3000$) of ancestry-informative markers.

HAPAA uses an HMM that models linkage within haplotypes. After using a forward-backward algorithm to infer ancestry blocks, it filters the blocks based on length and performs a second HMM to eliminate ancestry blocks that are too short.

Unlike most other methods, LAMP does not use an HMM; instead, it clusters genotypes within windows of Single Nucleotide Polymorphisms (SNPs) and uses a majority vote system to assign the ancestry of individual SNPs based on the windows to which they belong.

HAPMIX models each ancestral genotype as a mosaic of haplotypes from two ancestral populations. It assumes that the ancestral population data are fully phased, but accepts phased or unphased data from the admixed population. If the admixed data are unphased, HAPMIX averages over possible phasings.

Bryc *et al.* developed a method that uses Principle Components Analysis (PCA; see page 24) SNP loadings to weight SNPs based on their informativeness about population distinction, then uses an HMM to assign ancestry to segments of individuals' unphased genomes. In chapter 4, I extend this approach to a haplotype-based method that allows high-accuracy assignment of ancestry to multiple populations from dense genomewide data.

Hidden Markov Models

A Hidden Markov Model (HMM) is a stochastic process in which certain variables are unobserved (or "hidden") and form a Markov chain, and other variables are observed and depend on a single hidden variable apiece. In the context of chapter 4, the hidden variables are the ancestral states (say, African or European) of the windows of SNPs along a chromosome and the observed variables are the window scores, which are determined by the genotypes of that (phased, haploid) chromosome within the window and the PC loadings of the SNPs; that is, how much each SNP contributes to the separation between the populations (Figure 1.4). Conditional on the ancestral state of a window, the score of that window is conditionally independent of all other windows. The ancestral states form a Markov chain because each window's ancestry depends on the ancestry of the previous window, with the strength of the dependence determined by the probability of a recombination between the two windows.

We are interested in the posterior probability of the ancestry at each window, given the window scores at all of the windows. The standard way to compute this in an HMM is with the forward-backward algorithm [4]. For each i from 1 to the number of windows, we compute $f_{ij} = P(x_1, ..., x_i, z_i = j)$, where x_i is the

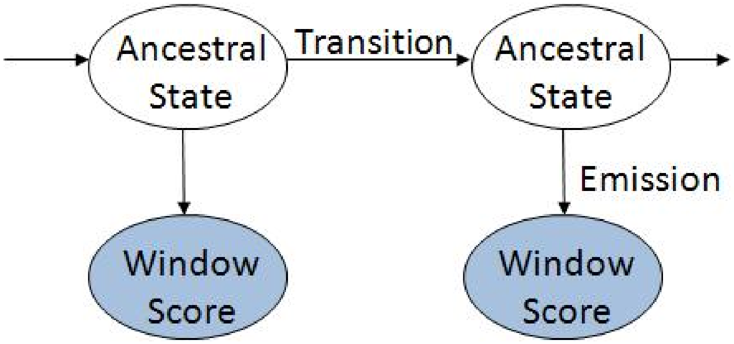

Figure 1.4: The HMM for ancestry assignment.

Observed variables are in blue; hidden variables are in white.

observed window score at window i and z_i is the hidden state at window i. This is the "forward" part of the algorithm, because dynamic programming enables f_{ij} to be computed rapidly by summing over possible states k for the previous window: $f_{ij} = e_{ij}\Sigma_k f_{i-1,k}a_{kj}$, where e_{ij} is the emission probability $P(x_i \mid z_i = j)$, and a_{kj} is the transition probability $P(z_i = j \mid z_{i-1} = k)$. The "backward" part of the algorithm involves computing $b_{ij} = P(x_{i+1}, x_{i+2}...x_L \mid z_i = j)$ (where L is the number of windows) by summing over possible states k for the next window: $b_{ij} = \Sigma_k a_{jk}e_{i+1,k}b_{i+1,k}$. Then

$$f_{ij}b_{ij} = P(x_1, ...x_i, z_i = j)P(x_{i+1}, ...x_L \mid z_i = j) = P(\vec{x}, z_i = j) \tag{1.9}$$

and

$$\Sigma_j f_{ij}b_{ij} = \Sigma_j P(\vec{x}, z_i = j) = P(\vec{x}), \tag{1.10}$$

so

$$\frac{f_{ij}b_{ij}}{\Sigma_j f_{ij}b_{ij}} = \frac{P(\vec{x}, z_i = j)}{P(\vec{x})} = P(z_i = j \mid \vec{x}), \tag{1.11}$$

which is the posterior probability in which we are interested. In chapter 4, I did these computations in logspace to avoid overflow due to very large Gaussian densities.

Principal Components Analysis

Principal Components Analysis (PCA) is a linear algebra technique for identifying the vectors that describe the greatest amount of variation in a set of data. We take the eigenvalue decomposition of the covariance matrix of the SNPs: $XX^T = VS^2V^T$, where X is the matrix of data, with each column corresponding to a data point (in chapter 4, the data points are haploid individuals) and each row corresponding to a dimension of information (in chapter 4, these are SNPs). The rows of V^T (equivalently, the columns of V) are the eigenvectors of X^TX. Because covariance matrices are symmetric, the eigenvectors are orthogonal. Each column of V^T is the coordinates of one data point in the basis of eigenvectors. S^2 is a diagonal matrix where each element is an eigenvalue of the covariance matrix, and is proportional to the amount of variance explained by that principal component.

PCA applied to the covariance matrix of X is equivalent to Singular Value Decomposition (SVD) of X: $X = USV^T$, where X is the data, S is a diagonal matrix of the singular values (the square roots of the elements of S^2), the columns of U are the left singular vectors of X, and the rows of V^T are the right singular vectors. Each column of U gives the SNP loadings for one principal component, which describe how much each dimension (SNP) contributes to that principal component (Figure 1.5).

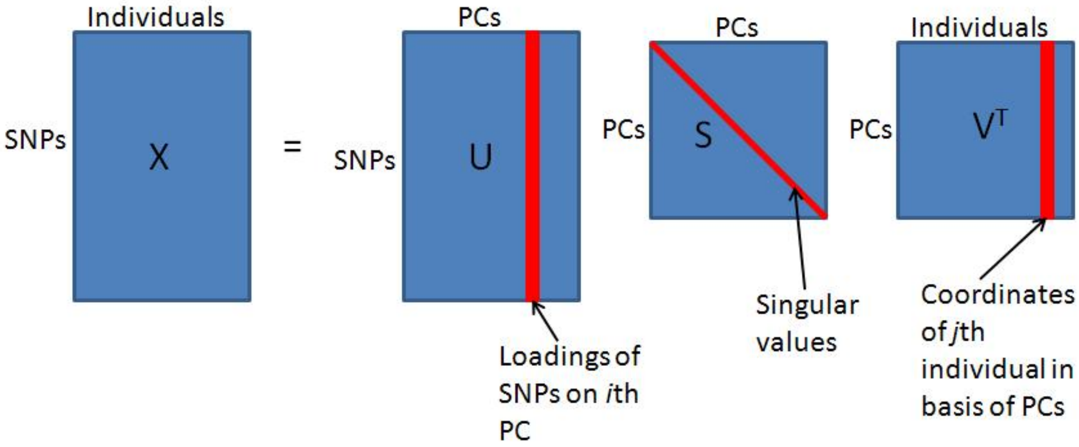
Individuals
SNPs
X
=
PCs
SNPs
U
PCs
PCs
S
Individuals
PCs
V^T
Loadings of SNPs on *i*th PC
Singular values
Coordinates of *j*th individual in basis of PCs

Figure 1.5: SVD of genotype or haplotype data.

When PCA is applied to genetic data from multiple populations, the first few principal components commonly correspond to variation due to genome-wide differences in allele frequency that are due to geographic separation, as for example in [51]. In chapter 4, I utilize this fact to employ PCA in identifying which SNPs contribute most strongly to the ability to distinguish between potential ancestral populations.

CHAPTER 2

BAYESIAN LINKAGE ANALYSIS OF CATEGORICAL TRAITS FOR ARBITRARY PEDIGREE DESIGNS[1]

2.1 Abstract

Pedigree studies of complex heritable diseases often feature nominal or ordinal phenotypic measurements and missing genetic marker or phenotype data. We have developed a Bayesian method for Linkage analysis of Ordinal and Categorical traits (LOCate) that can analyze complex genealogical structure for family groups and incorporate missing data. LOCate uses a Gibbs sampling approach to assess linkage, incorporating a simulated tempering algorithm for fast mixing. While our treatment is Bayesian, we develop a LOD (log of odds) score estimator for assessing linkage from Gibbs sampling that is highly accurate for simulated data. We demonstrate that LOCate exhibits better performance than LOT, an alternative method for ordinal linkage analysis on complex pedigrees, when analyzing simulated data with no family-specific environmental effect. We use our method to analyze a candidate locus for panic disorder in humans, and find evidence that an ordinal model is a better fit to the data than the binary model previously used.

[1] Brisbin, A., M.M. Weissman, A.J. Fyer, S.P. Hamilton, J.A. Knowles, C.D. Bustamante, J.G. Mezey. Submitted to PLoS One. PLoS One is an Open Access journal; if accepted, this paper will be freely reproducible with attribution.

2.2 Introduction

Many heritable traits, from pathogen resistance in plants [77] to panic disorder in humans [23], are described using discrete categories such as color or are quantified using discrete, ordered scales such as "mildly," "moderately," or "severely" affected. When performing linkage analysis of categorical traits, it is well appreciated that recoding measurements as binary can lead to decreased power [12, 20]. Recoding measurements as continuous can lead to the same problem. Use of the most widely applied software for linkage analysis such as Superlink [21], Merlin [1], Genehunter [41], and LOKI [35] that do not employ categorical trait models is therefore not the most appropriate strategy for analyzing categorical diseases.

Most previous work done on family-based mapping of categorical traits has been restricted to particular types of pedigrees. These include backcross [31, 42, 78] and F2 designs [81, 42, 78, 34], 4-way experimental crosses [60, 77, 80, 79], and sets of independent nuclear families [59, 82, 74]. Recent methods by Zhang *et al.* [83], Dupuis *et al.* [16], and Diao and Lin [15] allow linkage analysis for ordinal traits on arbitrary pedigrees. To date, there is no Bayesian framework for ordinal and nominal linkage analysis on pedigrees with inbreeding loops and missing data.

In this paper, we develop a Bayesian statistical framework for linkage analysis of a categorical trait with a user-specified penetrance function of arbitrary form. We implement this framework in the software LOCate (Linkage for Ordinal and Categorical traits). Our method can analyze an ordinal or nominal trait with any number of categories, can handle missing genotype and phenotype data, and can analyze pedigrees with inbreeding loops. In our analysis, we compare the performance of our method to LOT [83], the method of Zhang *et al.*, on simulated pedigrees. We also demonstrate the use of our method to reanalyze a study of panic disorder in humans previously analyzed as a binary trait [23].

2.3 Methods

In our linkage analysis framework, we seek the probability of a pedigree conditional on θ, the recombination rate between a single marker locus and the unknown disease locus:

$$P(X \mid \theta) = \Sigma_Y P(X, Y \mid \theta),$$

where the observed data X consists of individuals' phenotypes and unphased marker genotypes, and the unobserved data Y consists of all individuals' disease locus and phased marker genotypes, as well as any unobserved phenotypes and unphased marker genotypes. As the number of individuals in the family increases, the sum over all possible genotype assignments Y can grow unwieldy. Instead of considering all possible values of Y, Gibbs sampling is used to randomly explore the space of genotype configurations, emphasizing those configurations Y which have the highest values of $P(X, Y \mid \theta)$, and therefore contribute the most to the summation. Below, we describe the model, demonstrate the use of simulated tempering to improve the mixing of the Gibbs sampler, and introduce a novel estimator for the likelihood of the data from Gibbs sampling.

2.3.1 The Model

Figure 2.1 shows the graphical model for our Gibbs sampler. Following this model, the joint probability of the observed data (X) and unobserved data (Y), conditional on the recombination rate θ, is as follows:

$$\begin{aligned} P(X, Y \mid \theta) \quad \propto \quad & [\Pi_{i \in founders} P(Q_{fi}, Q_{mi} \mid HWE) \cdot P(M_{fi}, M_{mi} \mid HWE)] \\ \cdot \quad & [\Pi_{i \in nonfounders} P(Q_{fi}, Q_{mi} \mid parents, selectors) \\ \cdot \quad & P(M_{fi}, M_{mi} \mid parents, selectors) \\ \cdot \quad & P(sel_Q \mid sel_M, \theta) \cdot P(sel_M)] \\ \cdot \quad & [\Pi_{i \in all} P(M_{obs} \mid M_{fi}, M_{mi}) \cdot P(d_i \mid \overrightarrow{Q_i}, penetrance)] \\ \cdot \quad & [\Pi_{missing} P(M_{fi}) \cdot P(M_{mi})] \\ \cdot \quad & P(penetrance) \end{aligned} \tag{2.1}$$

where Q_{fi}, Q_{mi} are the disease alleles individual i received from its father and mother; M_{fi}, M_{mi} are the marker alleles i received from its father and mother; sel_Q and sel_M are "selector" variables that tell whether i received the grandpaternal or grandmaternal allele from each parent at the disease locus and the marker, respectively; $M_{i,obs}$ is i's observed, unphased marker genotype; d_i is i's phenotype; and $penetrance$ refers to the matrix of Pr(phenotype | genotype) used to model the disease. HWE refers to the genotype frequencies assuming the founders are drawn from a population under Hardy-Weinberg Equilibrium.

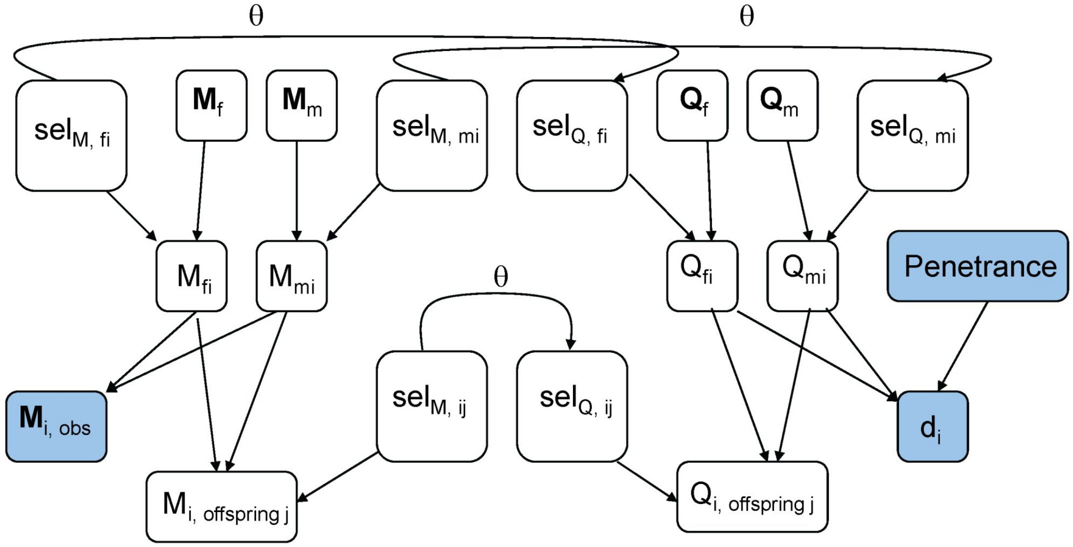

Figure 2.1: The graphical model for the Gibbs sampler.

All variables shown here are involved in updating the information for individual i. Filled-in variables are typically observed, and held constant throughout the run of the sampler. M_{fi}, M_{mi} = marker alleles that i received from its father and mother. Q_{fi}, Q_{mi} = disease locus alleles that i received from its father and mother. $M_{i,offspring=j}$, $Q_{i,offspring=j}$ = marker and disease locus alleles that individual i passed to its jth offspring. (Only one offspring is shown for illustration.) d_i = individual i's phenotype. $sel_{M,fi}$ = Selector variable: tells whether i's paternal marker allele comes from its paternal grandfather or grandmother. $M_{i,observed}$ = i's unphased marker genotype. $\mathbf{M}_f$, $\mathbf{M}_m$ = marker genotype vectors of i's mother and father. If i is a founder, replace by a constant node describing the population allele frequencies. Penetrances = matrix of the probabilities of each phenotype, conditional on disease genotype. The penetrances are held constant.

We derived a Gibbs sampler to sample genotype configurations Y in proportion to the probability in equation 2.1. In our Bayesian implementation, we used a uniform prior on the marker genotypes of individuals with missing data. We also used $P(sel_M) = .5$, which assumes unbiased inheritance; e.g., no meiotic drive. With the availability of additional information, it would be straightforward to change these priors. The penetrance parameters, which describe the probability of each phenotype category conditional on each disease locus genotype, are assumed to have a point prior, that is, to be fixed. We used a grid of values for θ in the current implementation.

The Gibbs sampler updates each set of variables conditional on its Markov blanket [39]. For example, individual i's marker alleles and selectors M_{fi}, M_{mi}, $sel_{marker,fi}$, $sel_{marker,mi}$ are updated by a draw from the distribution

$$\begin{aligned} P(M_{fi}, M_{mi}, \quad sel_{marker,fi}, \quad & sel_{marker,mi} \mid \text{Markov Blanket}) \propto \\ & P(M_{fi} \mid \mathbf{M}_f, sel_{marker,fi}) \cdot P(M_{mi} \mid \mathbf{M}_m, sel_{marker,mi}) \\ & \cdot P(M_{i,obs} \mid M_{fi}, M_{mi}) \\ & \cdot P(sel_{Q,fi} \mid sel_{marker,fi}) \cdot P(sel_{Q,mi} \mid sel_{marker,mi}) \\ & \cdot \Pi_{offspring=j} P(M_{ij} \mid M_{fi}, M_{mi}, sel_{marker,ij}) \end{aligned} \tag{2.2}$$

where $\mathbf{M}_f$ indicates the vector of marker alleles held by i's father in the current iteration.

Here,

$$P(M_{i,obs} \mid M_{fi}, M_{mi}) = \begin{cases} 0 & \text{if } M_{i,obs} \text{ is not a permutation of } M_{fi}, M_{mi} \\ 1 & \text{if } M_{fi} = M_{mi} \text{ (i is a homozygote)} \\ 1/2 & \text{if } M_{fi} \neq M_{mi} \text{ (i is a heterozygote)} \\ 1 & \text{if } M_{i,obs} \text{ is unobserved.} \end{cases} \tag{2.3}$$

In setting $P(M_{i,obs} \mid M_{fi}, M_{mi}) = 1$ if $M_{i,obs}$ is unobserved, we assume that this individual's genotype had probability 1 of being unobserved, independent of the individual's true phased genotype. If another model for gene dropouts were available, it could be employed here.

The calculation of $P(M_{ij} \mid M_{fi}, M_{mi}, sel_{M,ij})$ for each of i's offspring is analogous to this.

Also,

$$P(M_{fi} \mid \mathbf{M}_f, sel_{marker,fi}) = \begin{cases} 1-\mu & \text{if } M_{fi} \text{ matches } M_{f,sel} \\ \mu & \text{if they do not match} \end{cases}$$

where the mutation rate μ depends on the current "temperature" of simulated tempering (see below). If individual i's parents are not included in the pedigree, then i is a founder, and $P(M_{fi} \mid \mathbf{M}_f, sel_{marker,fi})$ is replaced by $P(M_{fi}) = 1/m$, where m is the number of distinct marker alleles.

2.3.2 Improving the Speed of the Method

Slow mixing is a chronic problem in Gibbs samplers for linkage analysis [71, 73]. This can result in inadequate exploration of the sample space and excessively long times to reach the stationary distribution. Even more of a concern is the fact that in cases with missing marker data and more than two possible marker alleles, the Markov chain may be reducible, rendering portions of the sample space inaccessible from a given starting point [72, 73].

To ameliorate this problem, we implemented simulated tempering [27, 28] in our Gibbs sampling algorithm. In simulated tempering, the Markov chain is run at several different "temperatures" λ, ranging from $\lambda = 0$, at which the chain's stationary distribution is the desired probability distribution, to $\lambda = 1$, at which the chain's distribution is very "relaxed," or smoothed, to increase the chance of the chain traversing regions of low probability density to reach different modes of the distribution. The most common way of relaxing the probability distribution is to raise the distribution to a power; however, this method is ineffective when some states to be traversed have zero probability. Geyer and Thompson [1995] performed simulated tempering by varying the disease penetrances at different values of λ. We extended their approach to a more general parameter relaxation, in which each value of λ features its own penetrances, recombination rate, mutation rate, and disease-allele frequency (see supplement). This greatly improved the mixing of our Gibbs sampler (Figure 2.2). Without simulated tempering (black line), distantly separated iterations of the Gibbs sampler remained highly correlated. With simulated tempering, the autocorrelation reached near-independence ($< .05$, below blue line) for $k > 15$, demonstrating improved mixing of the Gibbs sampler. Simulated tempering also reduced the time to stationarity of our Gibbs sampler (Figure 2.3). Without simulated tempering (blue bars), the Gelman-Rubin statistics at a burn-in of 64000 iterations were significantly greater than 1, indicating that the chains had not reached stationarity. With simulated tempering (red bars), a burn-in of 1000 iterations was sufficient to achieve Gelman-Rubin statistics very close to 1.

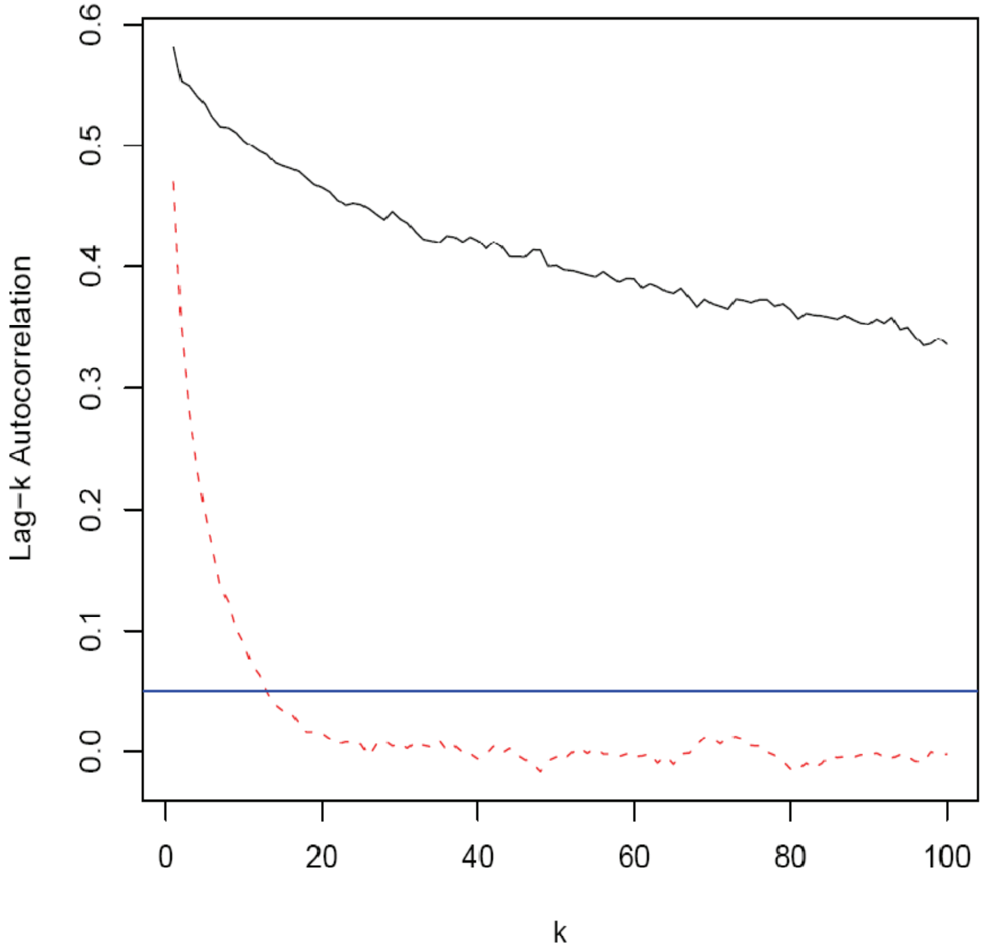

Figure 2.2: Lag-k autocorrelation with and without simulated tempering.

We show the correlation between $Pr(X, Y_i)$ (the joint probability of the observed and unobserved data at iteration i) and $Pr(X, Y_{i+k})$ (the probability k iterations later), for the simulated pedigree in Figure 2.5a. Black line = autocorrelation without simulated tempering; red line = autocorrelation with simulated tempering; blue line = .05, “near-independence” level.

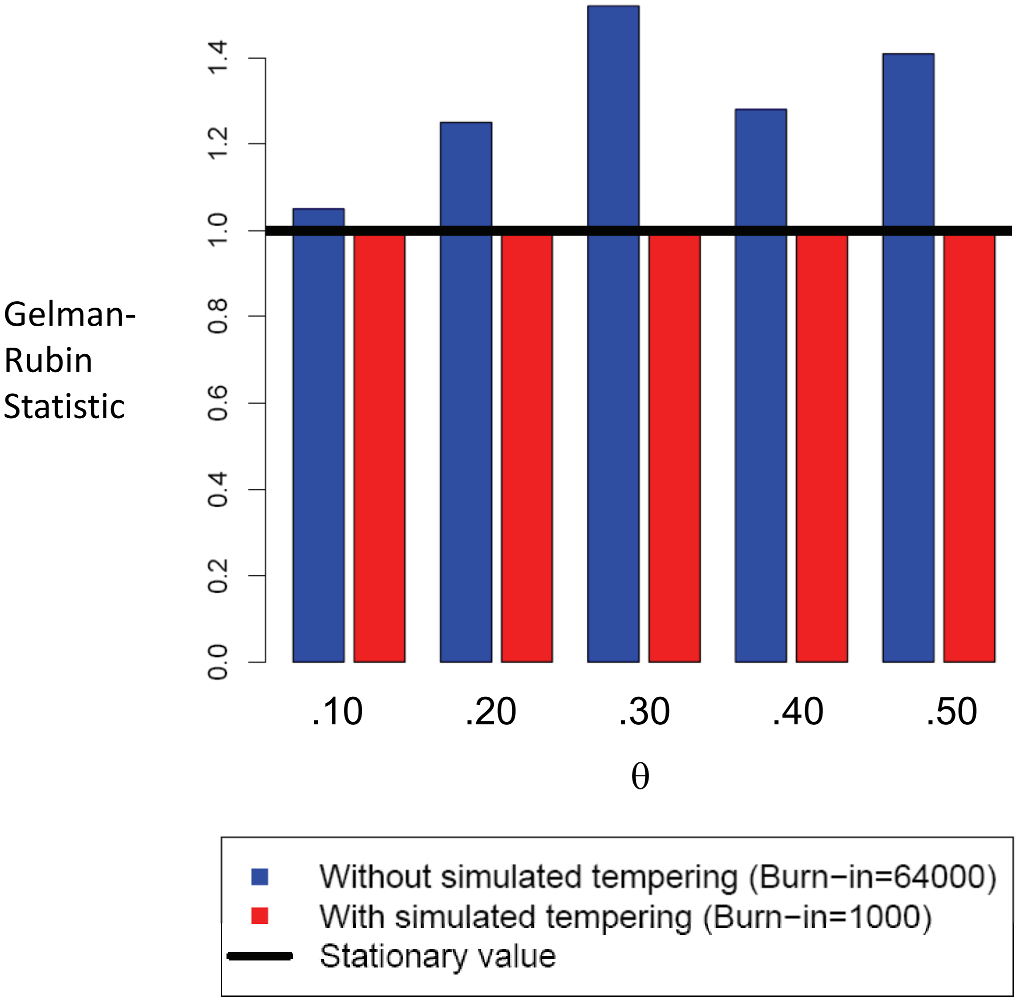

Figure 2.3: Gelman-Rubin statistics for the likelihood of a simulated pedigree.

Shown are the Gelman-Rubin statistics for the likelihood of the pedigree in Figure 2.5d.

2.3.3 Estimating the LOD Curve

While results of an analysis using our framework may be interpreted entirely from a Bayesian perspective by assuming a prior over the grid values of θ, we wished to provide a log of odds (LOD) score for convenient linkage assessment. Likelihood-based parameter inference from Markov chain Monte Carlo is prone to sampling bias [72, 45]. To avoid this bias, we developed a linear regression-based estimator (LinReg) which takes advantage of the relation

$$P(X \mid \theta) = \frac{P(X, Y \mid \theta)}{P(Y \mid X, \theta)}.$$

The numerator can be computed exactly (equation 2.1). We estimate the denominator $P(Y \mid X, \theta)$ by the proportion of iterations which visit each configuration Y. The LinReg estimator of $P(X \mid \theta) = L(\theta \mid X)$ is the slope of the best fit line (with intercept 0) through a plot of $P(X, Y \mid \theta)$ vs $\hat{P}(Y \mid X, \theta)$, as shown in Figure 2.4.

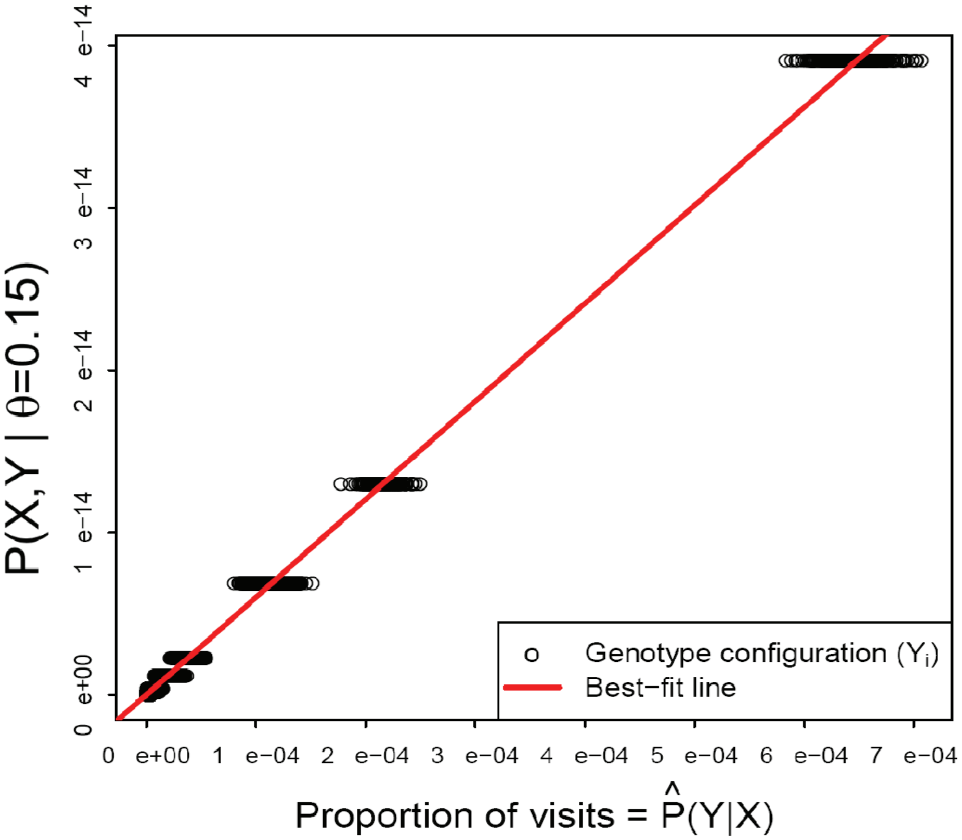

Figure 2.4: The Linear Regression estimator of $P(X \mid \theta)$.

X=observed data, Y=unobserved data. Shown is the estimator of $P(X \mid \theta)$ for the pedigree structure in Figure 2.5c, but with a binary trait simulated according to Table 2.1. $P(X, Y)$ is calculated using equation 2.1; $\hat{P}(Y \mid X)$ is estimated by the proportion of iterations which visit configuration Y, given the observed genotypes X. The slope of the regression line (red) is an estimate of $P(X \mid \theta)$.

2.3.4 Simulations

We assessed the performance of our method using two sets of simulated data. First, we tested the accuracy of LOD score estimation for single, small simulated pedigrees. Since any errors that occur in the analysis of one pedigree will be multiplied when multiple pedigrees are aggregated in a typical linkage analysis study, it is important that our method perform accurately when only a small amount of data is available. The simulated pedigrees included from 4 to 18 individuals; some examples are shown in Figure 2.5. These included pedigrees with missing genotype data and with inbreeding loops. Each pedigree has a simulated binary or trichotomous trait. We computed the LOD scores for these pedigrees using the disease penetrances in Table 2.1. For the simulated binary traits, we compared the LOD scores estimated by our method to the LOD scores calculated by Superlink [21]. For trichotomous traits, we compared our estimated LOD scores to the theoretical LOD scores under a model of complete penetrance. We also compared our estimated LOD scores to those obtained by treating the trichotomous trait as binary (in Superlink) or continuous (in Merlin and SOLAR [2]).

Table 2.1: Penetrance models used in our small-family simulations.
qq, Qq, and QQ represent the genotype at the disease locus.

Model	Phenotype	qq	Qq	QQ
Binary	$d = 1$	.9991	.9989	.0008
	$d = 2$	.0009	.0011	.9992
Trichotomous	$d = 1$	.9764	.0228	.0020
	$d = 2$	.0226	.9545	.0225
	$d = 3$	.0010	.0227	.9755

For our second set of simulations, we assessed the ability of our method to

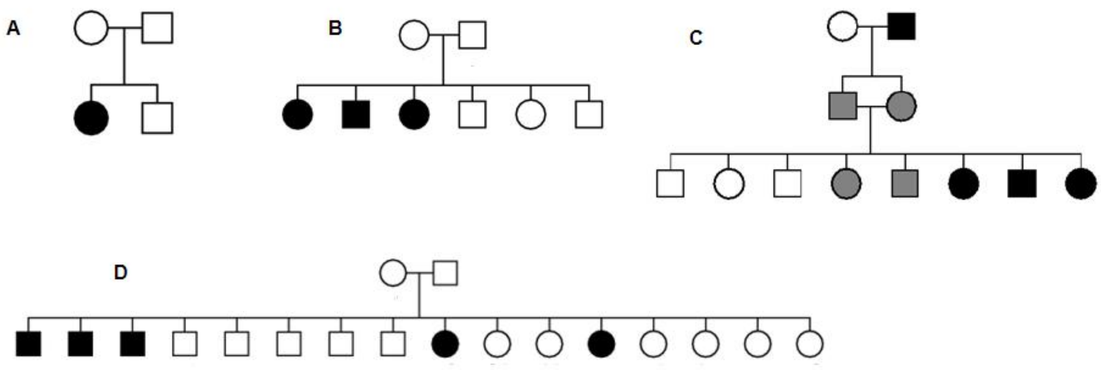

Figure 2.5: Examples of simulated pedigrees.

Black=affected, white=unaffected, gray=moderately affected. Each individual's unphased marker genotype is listed below the individual. A, B, and D are examples of simulated pedigrees with binary traits; C shows a simulated pedigree with a trichotomous trait and an inbreeding loop.

detect linkage in cases where the pedigree(s) may be reasonably broken into a large number of small family groups or where the study includes a large number of small families. For these simulations, we considered linkage studies of 100 families, each family consisting of 2 parents and 2 offspring. We simulated a trichotomous trait with penetrances as given in Table 2.2 (Model A). The trait locus was either tightly linked ($\theta = .01$) or unlinked ($\theta = .50$) to the observed marker locus. We required that each simulated family be informative for linkage (at least one parent heterozygous) and exhibit at least 2 levels of the phenotype among its 4 members. We simulated 100 such studies, and examined the power vs. type I error of our method and that of LOT [83]. Because LOCate requires an estimate of the penetrances as input, we tested our method with a range of penetrances (Table 2.2, Models A, B, C).

Table 2.2: Penetrance models used to analyze simulated linkage studies. Model A was used to generate the simulations.

Model	Phenotype	qq	Qq	QQ
A	$d = 1$	.99	0	0
	$d = 2$	.01	.99	.01
	$d = 3$	0	.01	.99
B	$d = 1$	.8	.1	.1
	$d = 2$	.1	.8	.1
	$d = 3$	.1	.1	.8
C	$d = 1$	.7	.3	0
	$d = 2$	.3	.4	.3
	$d = 3$	0	.3	.7

2.3.5 Application to Data

Panic disorder is a common illness in humans, characterized by periods of intense anxiety. Because individuals exhibit varying degrees of symptoms of panic disorder, this psychiatric illness is a natural choice for analysis as an ordinal trait. We used LOCate to perform ordinal linkage analysis on the Panic disorder data set of Fyer *et al.* [23]. This dataset consists of 1591 individuals in 120 pedigrees, classified into six categories: definitely affected by panic disorder, probably affected, possibly affected, any symptoms of panic, unaffected, or unknown. The dataset has missing data among both phenotypes and microsatellite marker genotypes. We used LOCate to analyze marker D2S1788, which Fyer *et al.* found to have a two-point HLOD(.2)=3.20, allowing for heterogeneity, when treating the trait as binary (treating categories “definite”, “probable”, and “possible” as affected).

We used LOCate to replicate the binary analysis of D2S1788 of Fyer *et al.*, and analyzed the trait under 4 trichotomous models (see page 114). Due to the exponential increase in the sample space with increasing pedigree sizes, we analyzed a reduced set of 96 families. Table 2.3 shows the penetrances used in the binary analysis of Fyer *et al.* [23], and the penetrances of our best-fitting trichotomous model. This work involved a re-analysis of anonymous data on human subjects, for which Institutional Review Board approval was not required.

Table 2.3: Penetrance models used in our analysis of Panic Disorder data.

Model	Phenotype	qq	Qq	QQ
Binary	Unaffected	.99	.5	.5
	Definite, Probable, Possible	.01	.5	.5
Trichotomous	Unaffected	.99	.5	.5
	Possible, Any symptoms	.005	.125	.125
	Definite, Probable	.005	.375	.375

2.4 Results

2.4.1 Estimating the LOD Curve

We compared our LinReg estimator to the Reverse Logistic Regression (RLR) estimator of Geyer [1991]. The LinReg estimator is faster to compute than the RLR estimator, because LinReg involves a simple linear regression, while RLR requires a complex optimization over many values of θ. We used both estimators to estimate the LOD curve for several simulated pedigrees, for 5 different runs of our Gibbs sampler. We found that the two estimators have comparable mean squared error (Figure 2.6), and the error for both methods is very low. Given the speed and accuracy of LinReg, we used this estimator for the rest of the analyses described below.

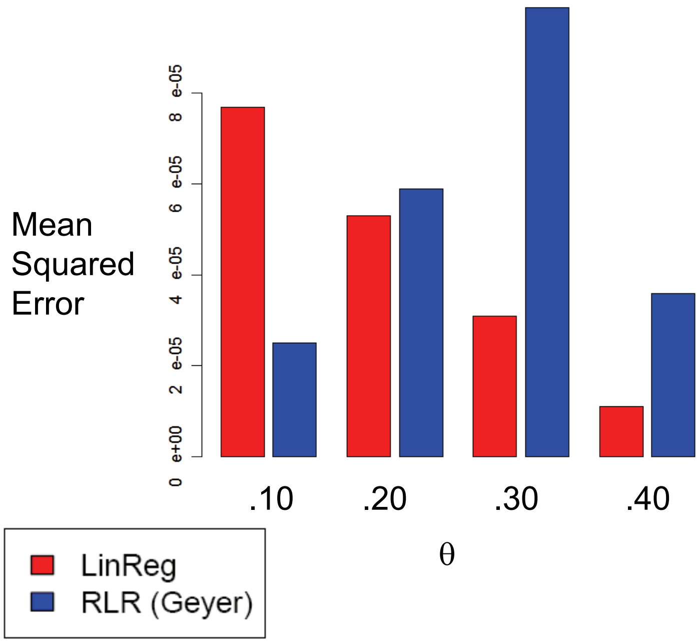

Figure 2.6: LinReg and RLR estimators of LOD(θ).

Shown are the empirical mean squared errors of the LinReg and RLR estimators of LOD(θ) for the simulated pedigree in Figure 3.2b. We used Superlink to compute the target value for each LOD(θ).

2.4.2 Simulations

LOCate accurately estimated LOD curves for individual simulated pedigrees with binary traits (Figure 2.7) and trichotomous traits (Figure 2.8). Previous studies have shown that treating a categorical trait as binary leads to a loss of power [12, 20]. Our results concur with this (Figure 2.9). We also examined the effect of treating categorical traits as continuous by analyzing simulated pedigrees with Merlin [1] and SOLAR [2]. These methods' continuous-trait models were unable to estimate the LOD curves accurately, while LOCate succeeded (Figure 2.8). Transforming the phenotypes using Merlin's *inverseNormal* option was also not effective in improving the fit of the continuous model.

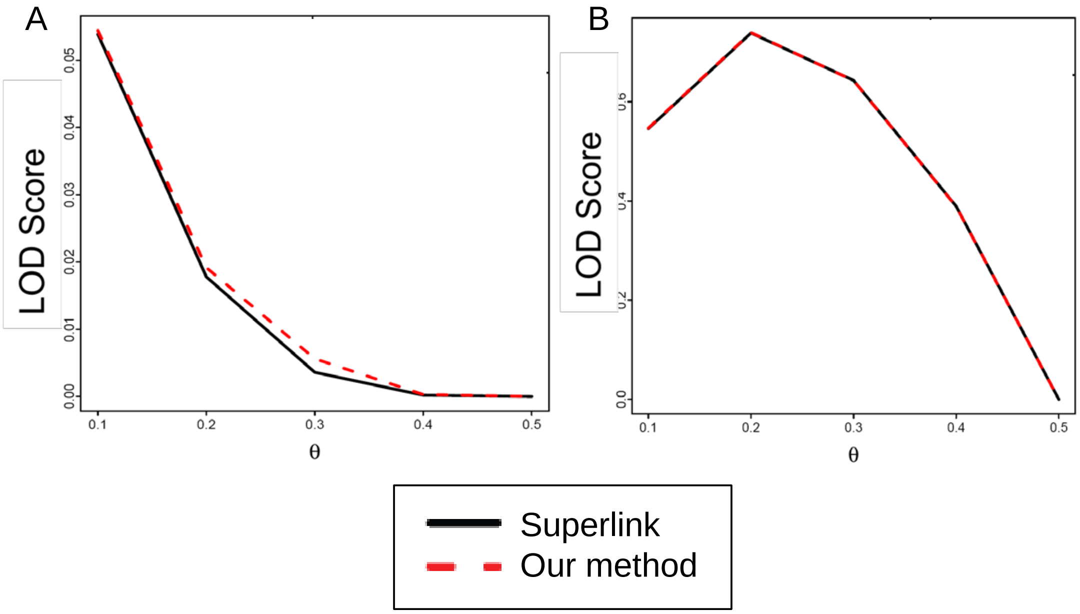

Figure 2.7: Estimated LOD curves for simulated pedigrees with binary traits.

Shown are the LOD curves computed by our method (red) and by Superlink (black) for (**A.**) the simulated pedigree in Figure 3.2a and (**B.**) a simulated pedigree with the structure shown in Figure 3.2c and a binary trait simulated according to the penetrances in Table 2.1.

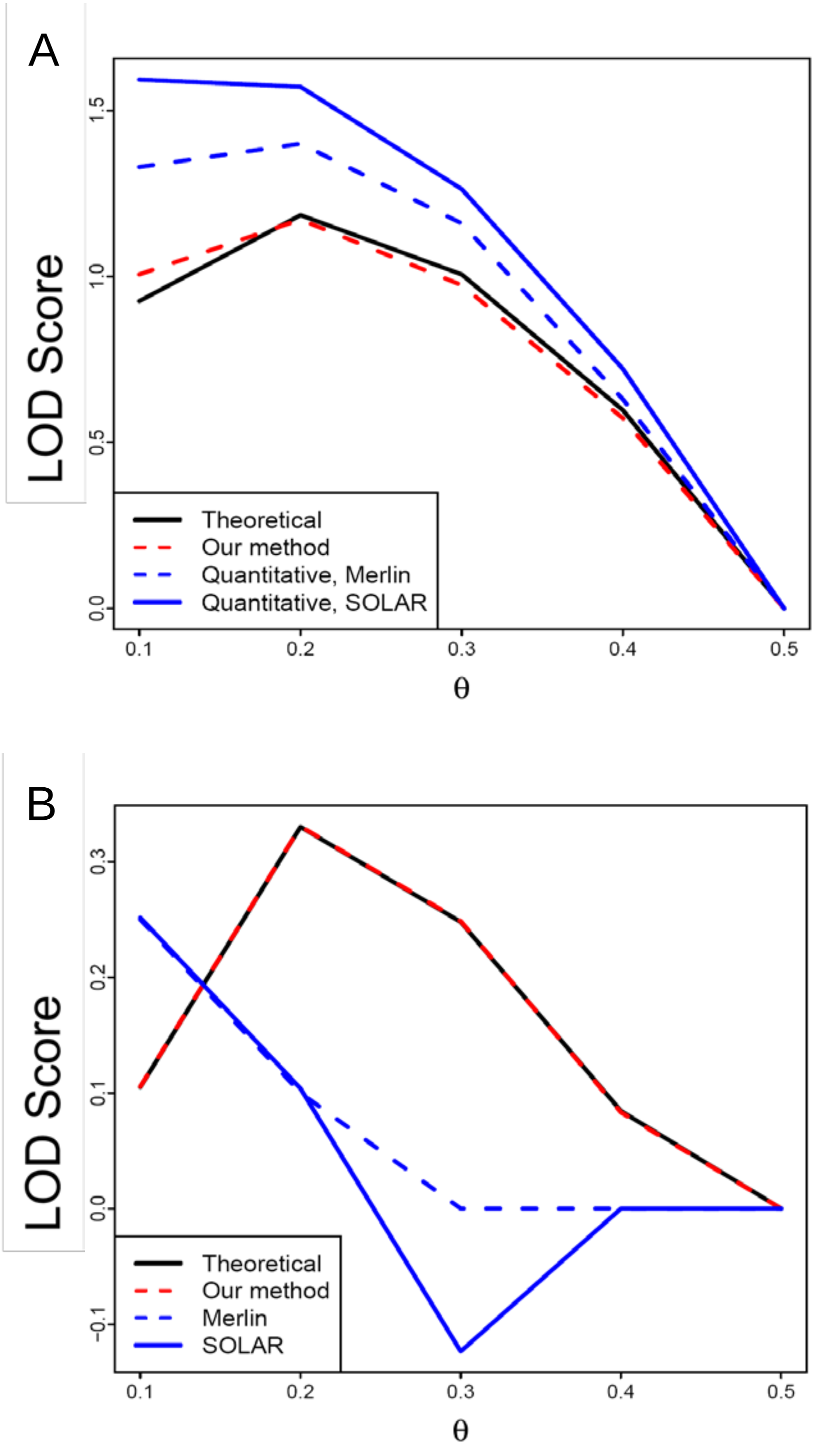

Figure 2.8: Accuracy of LOCate.

Shown are the results of linkage analysis on single, simulated pedigrees with trichotomous traits: **A.** Simulated pedigree shown in Figure 3.2c; **B.** A pedigree with 2 parents and 8 offspring (not shown), with trichotomous trait simulated according to Table 2.1. Merlin (dashed blue) and SOLAR (solid blue) were used for analysis as if the trait were continuous.

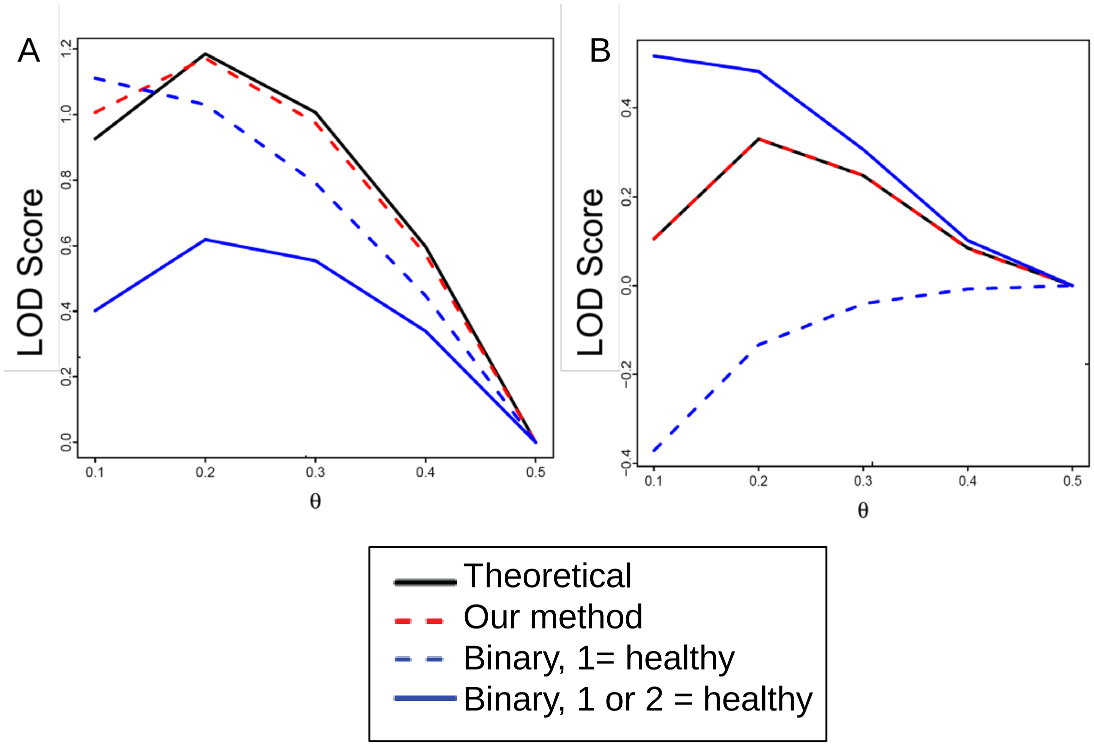

Figure 2.9: Treating trichotomous traits as binary.

Shown are the results of linkage analysis on single, simulated pedigrees with trichotomous traits: **A.** Simulated pedigree shown in Figure 3.2c; **B.** A pedigree with 2 parents and 8 offspring (not shown), with trichotomous trait simulated according to Table 2.1. Superlink was used for analysis as if the trait were binary (solid and dashed blue).

We present the results of our analysis of simulated 100-family linkage studies in Figure 2.10, which compares the receiver operator characteristic (ROC) curves for our method and for LOT. Our method has substantially higher power than LOT for the three penetrance models. Therefore, we find our method retains excellent discriminating power even when the penetrance model used is not the true model. A highly inaccurate penetrance model does reduce the magnitude of the estimated LOD scores, giving low power at a LOD threshold of 3 (Figure 2.11). This reinforces the value of considering alternative penetrance models in situations when LOD scores are close to zero genomewide, especially when analyzing categorical traits.

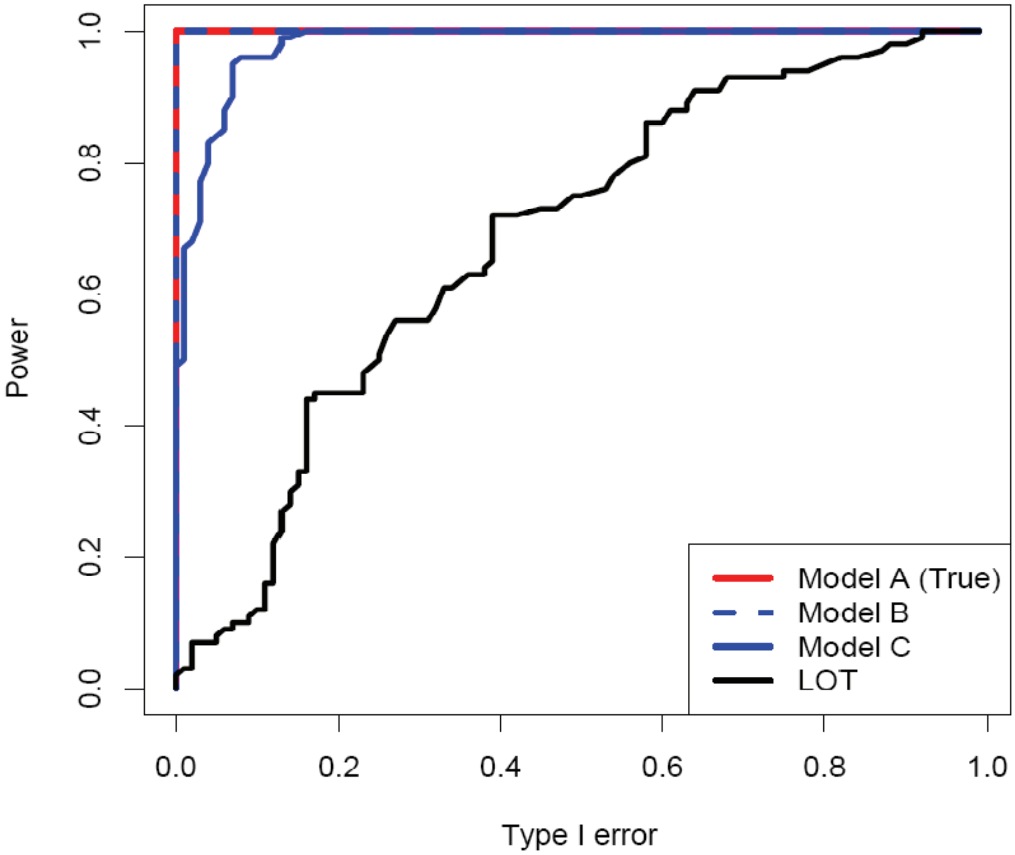

Figure 2.10: ROC plot from simulated linkage studies.

Model A, B, and C refer to analyses done with our method using the penetrance models in Table 2.2.

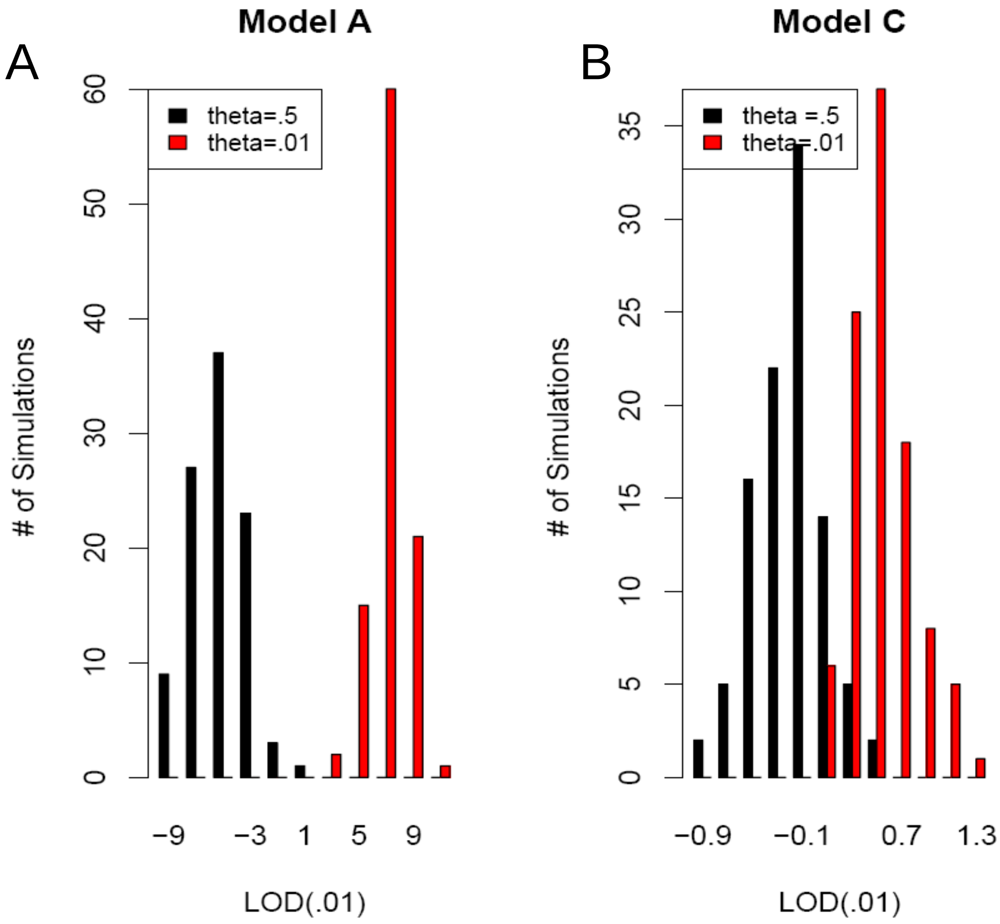

Figure 2.11: LOD scores from simulated linkage studies.

Shown are the frequencies of values of LOD(.01) for simulated sets of 100 4-person families. Red bars show the frequency of LOD scores for simulations with a linked QTL; black bars show the frequency for simulations with an unlinked QTL. Model A (Table 2.1) was used to generate the simulations. In **(A)**, model A was used to analyze the simulations; in **(B)**, model C was used to analyze the simulations.

2.4.3 Application to Data

For the subset of pedigrees we used, LOCate's estimated binary heterogeneous LOD score (HLOD) at $\theta = .2$ was 0.24 (LOD(.2)= -0.26); this value is lower than the HLOD in Fyer *et al.* [2006] because we used a reduced subset of pedigrees for computational speed. Our best-fitting trichotomous analysis, using the penetrances shown in Table 2.3, yielded HLOD(.2)=0.55 (LOD(.2)=0.19). This HLOD is higher than that found for the binary analysis on the same subset of pedigrees, suggesting that this trichotomous penetrance model is a better fit to the data than the binary model.

2.5 Discussion

Bayesian methods for linkage analysis are useful because they allow for incorporation of prior information about allele frequencies, meiotic drive, and other factors important to linkage calculations. This, along with LOCate's versatility for ordinal and nominal traits, makes our method a valuable complementary tool to existing frequentist methods.

Even in a Bayesian framework, it is desirable to have a means of computing LOD scores, as they are commonly used to assess linkage. We developed a new, linear-regression based estimator for L(θ), which has similar mean squared error to the RLR estimator, and is faster to compute. Our LinReg estimator will be useful for parameter inference in any situation in which MCMC is used and it is possible to calculate $P(X, Y \mid \theta)$, the joint probability of the observed and unobserved data, conditional on the parameter. For example, it could be used in the problem of population structure [57] to infer K, the number of populations represented by an observed sample of genotypes.

In our simulations, LOCate exhibits better power than LOT, and this is the case even when only a rough estimate of the penetrances is used as input to our method. The difference in power is likely to be partly due to the fact that LOT estimates a within-family environmental effect, which we did not include in our simulations. Our results demonstrate that when researchers do not expect a strong within-family environmental effect in their data, our method affords better power. It is worth noting that LOT, which uses a proportional-odds model, explicitly models traits as ordinal, and thus will gain power when the data contain a strong signal of ordering among the phenotypes. In contrast, the effectiveness of our approach is dependent upon the user-specified penetrance matrix, and many such matrices inherently model the trait as nominal rather than ordinal. In these cases, our method would not gain power in the presence of a strong signal of phenotypic ordering, but neither will it lose power when analyzing a truly nominal trait.

We used LOCate to perform ordinal linkage analysis on a dataset of humans affected by panic disorder, which had previously been analyzed as a binary trait. We found that a model which treats the trait as trichotomous is a better fit to the data than the binary model. This is consistent with the ordinal quality of the data and the well-demonstrated loss of power in treating a categorical trait as binary [12, 20]. Further investigation of panic disorder as an ordinal trait is warranted, including increasing the computational speed of LOCate to enable analysis of the full set of pedigrees.

We have implemented our method in the software LOCate, available at https://sourceforge.net/projects/categorical. LOCate is an effective and versatile approach for single marker analysis of nominal, ordinal, and binary traits on arbitrary family-sized pedigrees, including those with inbreeding loops and missing phenotypes and/or genotypes. While our method currently has scaling limitations for larger pedigrees, we are developing extensions for LOCate that make use of Elston-Stewart peeling to make the method available for the analysis of arbitrarily sized linkage studies. Other potential extensions include the random exploration of penetrance parameters and θ values within the Gibbs sampler.

2.6 Acknowledgements

This study was supported in part by NIH grant R01 GM083606-01 and NIMH Grants MH28274 (to MMW), MH37592 (to AJF), MH30906 (to DFK, AJF), MH48858 (to SEH). Genotyping services were provided to JAK by the Center for Inherited Disease Research (CIDR). CIDR is fully funded through a federal contract from the National Institutes of Health to Johns Hopkins University, contract number N01-HG-65403. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We would like to thank Adam Siepel for discussions of how to estimate $P(X \mid \theta)$. We thank the families who participated in this study.

CHAPTER 3

AN ELIMINATION ALGORITHM FOR CATEGORICAL LINKAGE ANALYSIS[2]

3.1 Abstract

Pedigree studies of complex heritable diseases often feature nominal or ordinal phenotypic measurements and missing genetic marker or phenotype data. We have developed a fast method for linkage analysis of categorical traits (LOCate2) that can analyze complex genealogical structure for family groups and incorporate missing data. LOCate2 uses an elimination algorithm to compute exact likelihoods efficiently, even in the presence of inbreeding loops. We demonstrate that LOCate2 is able to analyze simulated 100-individual pedigrees without pedigree cutting, which increases its power versus LOT, an alternative method for ordinal linkage analysis, when used to analyze ordinal and nominal traits on such large pedigrees. LOCate2 also exhibits better performance than LOT and QTLlink, another method for ordinal linkage analysis analyzing simulated nominal traits on large or small pedigrees. We use our method to conduct a segregation analysis for a cataract trait in Labrador Retriever dogs, and are able to reject the hypothesis of complete recessive inheritance efficiently in a large pedigree. We also analyze candidate loci for cardiac arrhythmia in a complex pedigree of German Shepherd Dogs, and find an increased LOD score at FH2525 on chromosome 6.

[2] Brisbin, A., J. Cruickshank, N.S. Moïse, T. Gunn, A. Milano, C. D. Bustamante, J.G. Mezey. In preparation.

3.2 Introduction

Many heritable traits, from pathogen resistance in plants [77] to panic disorder in humans [23], are described using discrete categories such as color or are quantified using discrete, ordered scales such as “mildly,” “moderately,” or “severely” affected. When performing linkage analysis of categorical traits, recoding measurements as binary or continuous can lead to decreased power [12, 20, 6]. Therefore, use of the most widely applied software for linkage analysis such as Superlink [21], Merlin [1], Genehunter [41], and LOKI [35] that do not employ categorical trait models is not the most appropriate strategy for analyzing categorical diseases.

Most previous work done on family-based mapping of categorical traits has been restricted to particular types of pedigrees. These include backcross [31, 42, 78] and F2 designs [81, 42, 78, 34], 4-way experimental crosses [60, 77, 80, 79], and sets of independent nuclear families [59, 82, 74]. Recent methods by Zhang *et al.* [83], Dupuis *et al.* [16], and Diao and Lin [15] allow linkage analysis for ordinal traits on arbitrary pedigrees. In a previous paper, we presented LOCate (chapter 2), a unified method for linkage analysis of ordinal and nominal traits on arbitrary pedigrees. LOCate, however, is a Gibbs sampling-based method that suffered from long computation times for pedigrees with more than 10-20 individuals.

In this paper, we present LOCate2, which employs an elimination algorithm to perform exact LOD score inference for single marker analysis of ordinal and nominal traits. The elimination algorithm allows efficient summing of joint probabilities, enabling us to analyze larger pedigrees and to examine multiple markers and penetrance schemes. We test our method on simulations and on two real data sets consisting of large pedigrees of dogs.

3.3 Methods

3.3.1 Computational Model

In linkage analysis, we are interested in computing $P(X \mid \theta)$, the probability of observed pedigree data X (phenotypes and marker genotypes) conditional on the recombination rate θ between a marker and the disease locus. If Y is a configuration of unobserved data (disease locus genotypes, as well as any unobserved phenotypes or marker genotypes), then $P(X \mid \theta) = \Sigma_Y P(X, Y \mid \theta)$. However, the space of possible configurations $\{Y\}$ can become quite large, even for moderate-sized pedigrees. Therefore, to quickly compute $P(X \mid \theta)$ exactly, an efficient means of summation is necessary.

To sum efficiently over all configurations Y, we used an elimination algorithm [39], a generalized form of the Elston-Stewart algorithm [18] in which inbreeding loops can be handled without cutting. By summing over the possible disease locus genotypes of one individual at a time, we can reduce the total number of terms required. For an example of this, see the Appendix (page 115).

We implemented this algorithm in R. Each possible combination of marker genotypes within a trio (for example, “offspring and one parent homozygous; offspring inherited heterozygous parent’s first marker allele”) is represented by a 3-dimensional table (Table 3.1), in which $P(m_i, q_i | m_{parents}, q_{parents})$, the probability of individual i's disease and marker locus genotypes, conditional on i's parents, is represented as a function of θ, the recombination rate between the observed marker and the disease locus. In the case of missing marker genotypes, we use tables with 4 or more dimensions, allowing one dimension to represent the possible values for an individual's missing marker genotype. Probabilities of founders' genotypes (assuming Hardy-Weinberg equilibrium) and the probability of each individual's phenotype, conditional on his or her disease locus genotype, are represented as 1-dimensional tables. We assume that there is no mutation at the marker or disease locus.

Table 3.1: Elimination table for homozygous offspring and one homozygous parent.

We suppose that the child's marker genotype is (1,1) and the parents' marker genotypes are (1,1) and (1,2). (Note that (1,2) is considered different from (2,1).) The probability of the child's marker and disease locus genotype, conditional on the parents' genotypes, is a function of the child's disease genotype (boxes), the homozygous parent's disease genotype (rows), and the heterozygous parent's disease genotype (columns).

Offspring genotype	Homozygous parent genotype	Heterozygous parent genotype			
		qq	qQ	Qq	QQ
qq	qq	.5	$\frac{1-\theta}{2}$	$\frac{\theta}{2}$	0
	qQ/Qq	.25	$\frac{1-\theta}{4}$	$\frac{\theta}{4}$	0
	QQ	0	0	0	0
qQ/Qq	qq	0	$\frac{\theta}{2}$	$\frac{1-\theta}{2}$	.5
	qQ/Qq	.25	.25	.25	.25
	QQ	.5	$\frac{1-\theta}{2}$	$\frac{\theta}{2}$	0
QQ	qq	0	0	0	0
	qQ/Qq	0	$\frac{\theta}{4}$	$\frac{1-\theta}{4}$	.25
	QQ	0	$\frac{\theta}{2}$	$\frac{1-\theta}{2}$	.5

Our algorithm first initializes the tables for all individuals; the penetrance and HWE tables are initialized based on the penetrances and disease allele frequency chosen by the user. On each iteration of the algorithm, the program eliminates one individual; the order is chosen in advance by the user (see below). During each iteration, the program first identifies all tables that depend in some way upon the individual i to be eliminated. It calculates a "pre-sum" table which is the product of all these tables; for example, if tables $f_1(g_i, g_j, g_k)$ and $f_2(g_i, g_j)$ are the only tables that depend on individual i, then the algorithm finds the pre-sum table

$$T : T_{a,b,c} = f_1(g_i = a, g_j = b, g_k = c)f_2(g_i = a, g_j = b).$$

The algorithm then sums over the possible values for g_i, producing a post-sum table with one fewer dimension than the pre-sum table:

$$f_3 : f_3(b, c) = \Sigma_a T_{a,b,c}.$$

If the post-sum table has only 1 element, then no remaining individuals' genotypes are dependent upon i. The table now contains the marginal probability of the phenotypes and marker genotypes of all individuals that were eliminated in the process of producing this table. The value in the table is multiplied by the current total probability for the pedigree, and the algorithm proceeds to the next iteration. Alternatively, if the post-sum table has more than one element, this table is saved (it is now a function of the individuals upon whom i depended) and the algorithm proceeds.

Note that if individual i's marker genotype is unobserved, it is represented by an additional dimension in every table involving i. In this case, the variable representing i's marker genotype is eliminated immediately after eliminating i's disease locus genotype.

3.3.2 Elimination Order

The choice of elimination order can have a profound effect on the speed of the algorithm. The number of terms involved in the summation $P(X \mid \theta) = \Sigma_Y P(X, Y \mid \theta)$, and therefore the speed of the algorithm, depends on the the size of the tables created during the elimination or, equivalently, the number of individuals involved in each function $f_j(g_A, g_B, ...)$. Another way to interpret this is to model the pedigree as a graph. Eliminating a variable i is equivalent to deleting the vertex v_i that represented i and adding edges between every pair of vertices that were connected to v_i. This creates a clique (complete subgraph) representing the dependency among variables that were conditionally independent given i. For example, in a nuclear family, siblings' genotypes are conditionally independent given their parents' genotypes, so eliminating a sibling creates a function $f(g_{mother}, g_{father})$ that depends only on the parents.

However, eliminating a parent first would involve marginalizing over the parent's possible genotypes, removing the conditional independence. All the siblings' genotypes would then be interdependent, resulting in a table $f(g_{\text{other parent}}, g_1, ...g_n)$ of dimension $1 + n$, where n is the number of siblings. If this table is large, subsequent calculations involving any of the siblings will be slow. For example, the pedigree in Figure 3.1 can be analyzed in 30 seconds for 5 values of θ on a 2 GHz desktop computer if a good elimination order is chosen: By first eliminating the siblings in the last generation, the largest clique size is 3. In contrast, first eliminating one of the parents in the second generation results in cliques as large as 11, which would take a predicted 23.5 hours to analyze.

In pedigrees without inbreeding loops, a good strategy is to eliminate first those individuals without offspring and any founders with small numbers of

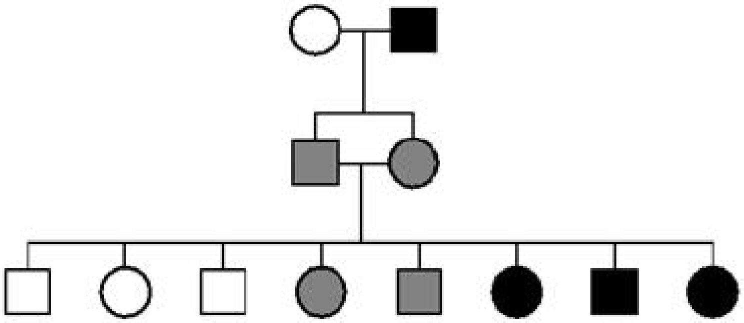

Figure 3.1: A sample pedigree with a full-sib mating.

offspring, and work in to the center of the pedigree. The method can deal exactly with inbreeding loops, provided the size of the cliques created by the elimination order are not too large. In situations with multiple, large inbreeding loops, it may be necessary to cut the loops to allow analysis in a reasonable length of time. In our experience, for pedigrees with approximately 200 individuals, elimination orders with several cliques of size 5 will run in minutes to hours on a desktop computer, depending on the amount of missing data and the number of marker alleles. In contrast, elimination orders with even one clique of size 7 will require hours to days. Elimination orders with a clique of size 12 will require more memory than typical installations of R are equipped to handle.

3.3.3 Simulations

To test this method, we simulated large pedigrees with inbreeding loops and ordinal and nominal traits. To mimic the structures of the dog pedigrees we analyzed in our data analysis, we simulated 10 pedigree structures with 100 members each, of which 20% were founders and 70% were leaf individuals (had no offspring). We did this by simulating pedigrees with 45 individuals in PyPedal [10] and adding 55 leaf individuals, distributed uniformly at random among mating pairs (Figure 3.2). For each pedigree structure, we simulated 50 sets of disease genotypes, for a total of 500 simulated pedigrees. We then simulated phenotypes based on the ordinal and nominal ”true penetrance” functions in Tables 3.2 and 3.3 and simulated genotypes at a marker that was either linked ($\theta = .10$) or unlinked ($\theta = .50$) to the disease genotype. We randomly combined sets of 5 pedigrees to produce simulated 500-individual linkage studies, which we analyzed using LOCate2, LOT [83], and QTLlink [16]. We allowed LOT and QTLlink to estimate the penetrances from the data; since LOCate2 requires that the penetrances be estimated in advance, we tested 3 models, as shown in Tables 3.2 and 3.3: the true penetrance, a somewhat misspecified penetrance model, and a very misspecified penetrance model.

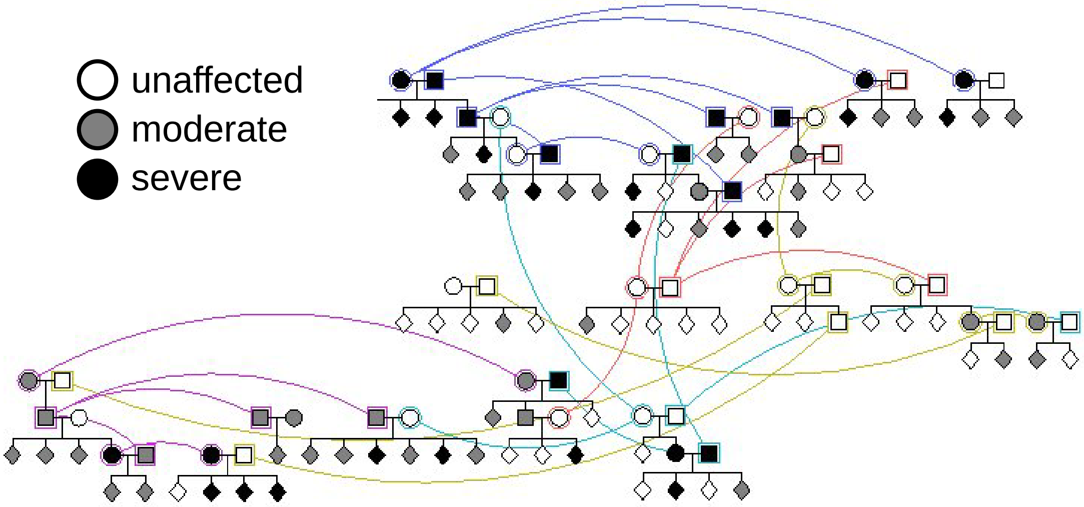

Figure 3.2: Example of a large simulated pedigree with inbreeding. Colored lines connect multiple representations of the same individual.

Table 3.2: Penetrance models used to simulate an ordinal trait with OR=4.95.

Shown are the penetrance models used to generate and analyze an ordinal trait on our large and small simulated pedigrees. The model used to simulate the ordinal trait is a proportional-odds logistic model with $\alpha = (.8, 2.8)$ and $\gamma = 1.6$, producing an odds ratio of 4.95. The misspecified analysis model uses $\gamma = .7$, giving OR=2.01. The "very misspecified" analysis model is a nearly-complete codominant nominal model.

Analysis Model	Phenotype	P(pheno|qq)	P(pheno|Qq)	P(pheno|QQ)
true penetrance	unaffected	.6900	.3100	.0832
	moderate	.2527	.4585	.3181
	severe	.0573	.2315	.5987
misspecified	unaffected	.6900	.5250	.3543
	moderate	.2527	.3660	.4478
	severe	.0573	.1091	.1978
very misspecified	unaffected	.9000	.0500	.0500
	moderate	.0500	.9000	.0500
	severe	.0500	.0500	.9000

Table 3.3: Penetrance models used to simulate a nominal trait.

Shown are the penetrance models used to generate and analyze a nominal trait on our large and small simulated pedigrees. The true penetrance model has a codominant penetrance structure. The misspecified analysis model for this simulation has a codominant structure with weaker penetrance, and the "very misspecified" model is an ordinal model with OR=4.95.

Analysis Model	Phenotype	P(pheno|qq)	P(pheno|Qq)	P(pheno|QQ)
true penetrance	unaffected	.8000	.1000	.1000
	moderate	.1000	.8000	.1000
	severe	.1000	.1000	.8000
misspecified	unaffected	.6000	.2000	.2000
	moderate	.2000	.6000	.2000
	severe	.2000	.2000	.6000
very misspecified	unaffected	.6900	.3100	.0832
	moderate	.2527	.4585	.3181
	severe	.0573	.2315	.5987

An odds ratio of 4.95, such as that used for our simulated ordinal trait, is large, but not unheard-of: Various studies have found odds ratios larger than this for a variety of traits, including coronary artery disease [3], susceptibility to bacterial disease [44], diabetic Charcot neuroarthropathy [54], developmental delay [29], and bipolar disorder [84]. We chose this large OR to best demonstrate the use of our method on large, complex pedigrees for a sample size of 500. Our method could also be used to analyze traits with smaller ORs if the sample size were greatly increased. Because our method computes exact LOD scores, and no loop-cutting was required to analyze these simulated pedigrees, the large OR we chose represents the real need for strong evidence in order to identify loci using linkage analysis, not a limitation of our method.

LOT was unable to analyze the full pedigrees, returning the error message "Error: num_aff>16". This error also occurred for some of the subpedigrees when we used Pedcut [43] to split the pedigrees into subpedigrees with a maximum of 25 bits and 20 bits, where the bitsize of a pedigree is measured by $2*$the number of nonfounders minus the number of founders. When we used a maximum of 15 bits and 10 bits, LOT froze. Instead, we split the pedigrees into nuclear families for analysis in LOT. QTLlink was able to analyze the large pedigrees when IBD calculations were performed with Loki [35].

We also tested the three methods on simulated small-family linkage studies, such as might be used in humans. For each of 100 simulations, we simulated 100 families consisting of 5 individuals each: 2 parents and 3 offspring. The sample size (500) is the same as in our large-family simulations. We simulated phenotypes for these pedigrees according to the same models used for the large-family simulations, plus an additional ordinal model with a larger odds ratio

(Table 3.4).

Table 3.4: Penetrance model used to simulate an ordinal trait with OR=12.06.

Due to low power to detect the ordinal trait in table 3.2 using small families, we also simulated data under $\gamma = 2.49$, corresponding to an OR of 12.06. The misspecified penetrances used to analyze this model followed an ordinal model with OR=4.95, and the "very misspecified" model was a codominant nominal model with weak penetrance.

Analysis Model	Phenotype	P(pheno|qq)	P(pheno|Qq)	P(pheno|QQ)
true penetrance	unaffected	.6900	.1558	.01507
	moderate	.2527	.4211	.0865
	severe	.0573	.4231	.8984
misspecified	unaffected	.6900	.3100	.0832
	moderate	.2527	.4585	.3181
	severe	.0573	.2315	.5987
very misspecified	unaffected	.6000	.2000	.2000
	moderate	.2000	.6000	.2000
	severe	.2000	.2000	.6000

3.3.4 Data Analysis

To illustrate the use of our method on a large, inbred pedigree, we used LOCate2 to perform a segregation analysis on a pedigree of 177 Labrador Retriever dogs (Figure 3.3) affected by juvenile hereditary cataracts, a binary phenotype [49]. By visualizing the pedigree in GraphViz [24], we were able to choose a tractable elimination order without cutting any loops. We performed segregation analysis by setting θ to 0 and all individuals' marker genotypes to (1,1), so that the "marker genotypes" were uninformative about disease locus inheritance. We analyzed the data under a completely penetrant recessive model (P(affected | qq) = P(affected | Qq) = 0, P(affected | QQ) = 1), an incompletely penetrant recessive model (P(affected | qq) = P(affected | Qq)), and a free model (no restrictions on penetrances). For the incomplete recessive and free models, we used a grid of penetrance values from (0,0,0) to (1,1,1) in intervals of .1. For all three models, we assumed that the frequency of the disease allele, Q, was .25. This is a reasonable value for the pedigree, which was taken from a colony of dogs in which the cataract phenotype segregates at relatively high frequency.

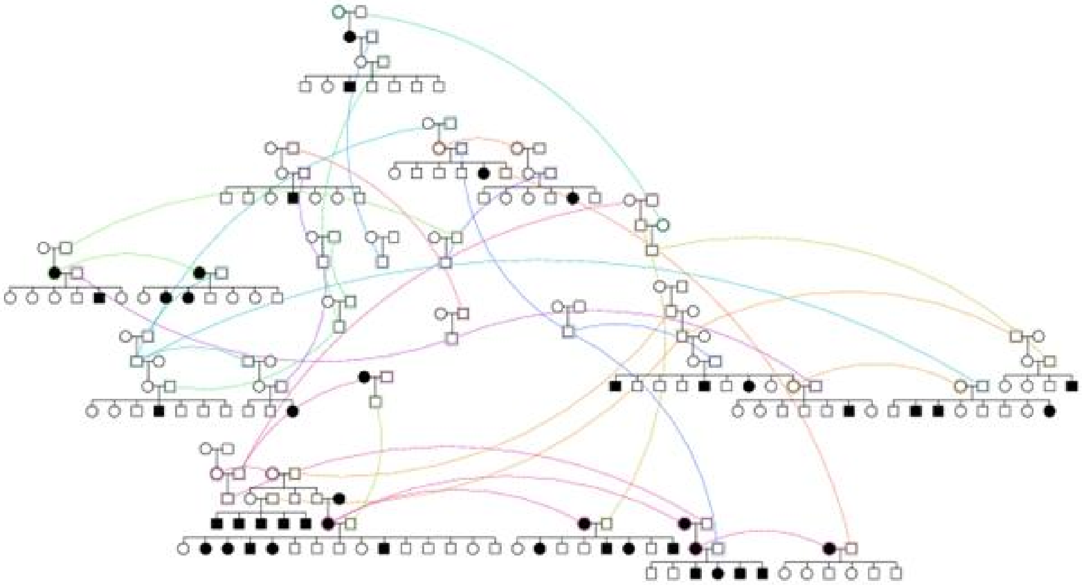

Figure 3.3: Pedigree of Labrador Retrievers used for segregation analysis of juvenile hereditary cataracts.

Black=affected, white=unaffected. Colored lines connect different representations of the same individual.

To demonstrate the effectiveness of our method for linkage analysis of a categorical trait, we analyzed a pedigree of 155 German Shepherd dogs (Figure 3.4a) affected by ventricular cardiac arrhythmias [13]. This is a complex phenotype which was measured as the number of single, double, triple, and "runs" of premature ventricular complexes a Holtered dog experienced in a 24-hour period. In order to make the phenotype more tractable for analysis, we (Teresa Gunn, Jenifer Cruickshank, and Sydney Moïse) combined these four values into an ordinal assessment of no arrhythmia, or "mild," "moderate," or "severe" arrhythmia. Because this pedigree was highly complex, with large inbreeding loops, our initial attempt at analysis resulted in clique sizes of up to 12, too large for R to handle. We cut all inbreeding loops by duplicating 8 individuals (Figure 3.4b). We assigned elimination priorities by hand to produce a pedigree with maximum clique size 5.

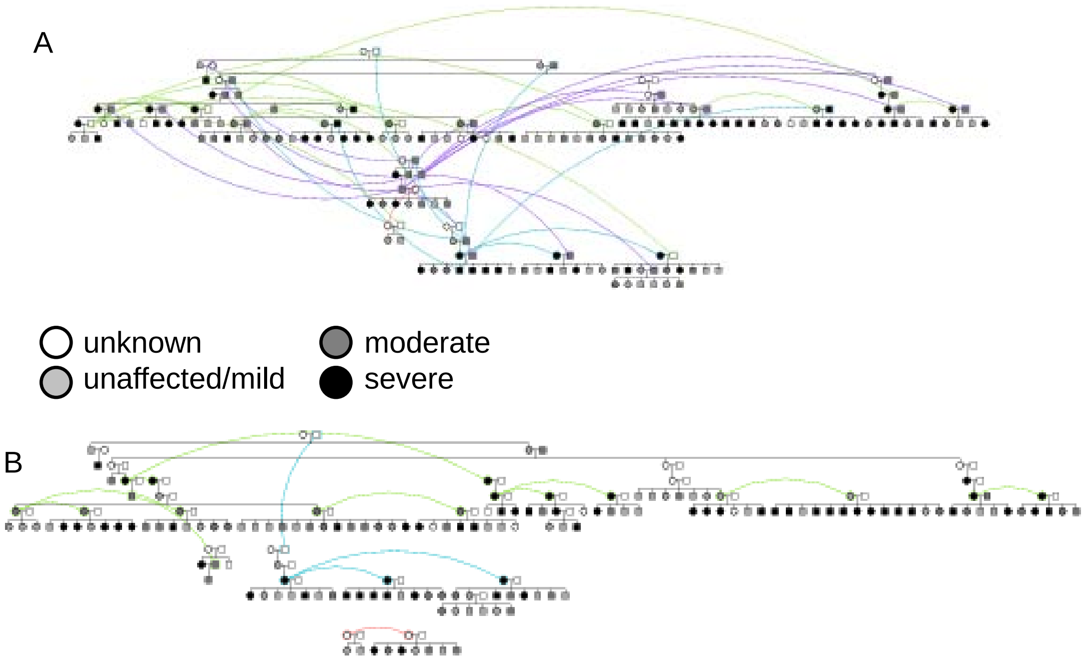

Figure 3.4: Pedigrees of German Shepherd dogs affected by cardiac arrhythmia.

The original pedigree (a) and the same pedigree after duplicating individuals to cut inbreeding loops (b). Colored lines connect different representations of the same individual.

We first used the elimination algorithm to conduct a segregation analysis for the loop-cut version of the pedigree, to determine the maximum-likelihood binary penetrance function (out of the 29 recessive and dominant functions allowed by Superlink Online [21]). This was a dominant model with Pr(affected | Qq) = Pr(affected | QQ) = .9 and Pr(affected | qq) = 0. For the binary analyses, we treated phenotypes “moderate” and “severe” as being affected, and “mild” or “no arrhythmia” as being unaffected. We chose to combine “mild” with the unaffected class in both the binary and the trichotomous analyses because of the small number (11, or 7.1%) of dogs that were truly unaffected. We used freq(Q) = .25, as in the cataract analysis.

We used this penetrance model to perform a preliminary binary analysis on all 302 microsatellites in the data using Superlink Online, for both the loop-cut and the original versions of the pedigree. We selected the 11 markers with a LOD score $>$ .7 in either analysis to analyze using LOCate2 (Table 3.5). We ran LOCate2 on the selected markers using 3 penetrance models, shown in Table 3.6, for 4 values of θ: 0, 0.1, 0.2, and 0.3, as well as 0.5, as this value is necessary to convert Pr(data | θ) into LOD(θ). Of the 11 selected markers, 7 had 5 or fewer alleles and 32 or fewer missing genotypes; these markers could be fully analyzed for each penetrance model in a few hours. For the remaining markers, we collapsed alleles with frequency $< 10\%$ into 1 category (markers 2, 9, 10) or iteratively removed founder pairs with missing genotypes (markers 2, 3, 9) in order to expedite analysis.

Table 3.5: Markers tested for linkage with cardiac arrhythmia.

Shown are the markers that had LOD(θ) ≥ 0.7 for either binary analysis (i.e., with or without cutting inbreeding loops), with the maximum LOD score achieved and the value of θ where it was achieved.

Marker #	Chromosome	Name	θ	LOD(θ)	Pedigree version
1	6	FH2525	.3	.8918	original
1	6	FH2525	.2	1.8691	cut
2	11	FH2319	.3	.8490	cut
3	12	REN213F01	.2	.8939	cut
4	21	FH2441	0	1.0617	cut
5	21	REN37A15	.2	.7298	cut
6	21	FH2312	.1	.7167	cut
6	21	FH2312	.2	1.5506	original
7	36	REN179H15	0	.7271	cut
8	1	FH2793	.3	.7046	original
9	1	FH2294	.3	.9711	original
10	7	FH3972	.3	.7886	original
11	20	REN93E07	.3	.7450	original

Table 3.6: Penetrance models used to evaluate German Shepherd dog pedigree.

Shown are the 3 trichotomous penetrance models we used to analyze the German Shepherd dog pedigree. Model A is closely based on the dominant(.9,.9) model we used for our binary analyses; model B is similar to model A, but provides a more codominant approach to distinguishing the moderate and severe phenotypes. Model C follows a proportional-odds ordinal model.

Model	Phenotype	P(pheno|qq)	P(pheno|Qq)	P(pheno|QQ)
A	unaffected/mild	1	.1	.1
	moderate	0	.35	.35
	severe	0	.55	.55
B	unaffected/mild	1	.1	.1
	moderate	0	.8	.1
	severe	0	.1	.8
C	unaffected/mild	.8176	.3775	.0759
	moderate	.1350	.3535	.1931
	severe	.0474	.2690	.7310

3.4 Results

3.4.1 Simulations

Figure 3.5 shows the LOD curve for the inbred pedigree in Figure 3.1, demonstrating that our method computes exact LOD scores, even for pedigrees with inbreeding loops. The method is also much faster than our previous, Gibbs sampling-based approach, which required over 125 hours to approximate this LOD curve on a 2 GHz desktop computer, as compared to 30 seconds by the elimination algorithm. Figure 3.6 shows the power vs. type I error of LOCate2, LOT, and QTLlink on the large simulated pedigrees (3.6a=ordinal trait, 3.6b=nominal trait). Our method, which is able to analyze the full pedigrees without cutting, has excellent power, even when the penetrance model used for analysis is only a rough approximation to the true penetrances. QTLlink also has excellent power on the ordinal trait, but is outperformed by LOCate2 on the nominal trait. In contrast, LOT suffers from reduced power due to necessary pedigree cutting, as well as the poor fit of LOT's ordinal trait model to the nominal penetrances used to generate the simulations for Figure 3.6b. It is also likely that LOT is hindered by the incomplete linkage ($\theta = .1$) between the marker and the simulated QTL, as, unlike LOCate2 and QTLlink, LOT does not offer the option of calculating linkage statistics for ungenotyped locations.

Figures 3.7 and 3.8 show the power vs. type I error of the three methods on the simulated small-family linkage studies. We found that there was lower power to detect loci using many small families compared to few large families, with a constant sample size (Figure 3.9), which agrees with previous results [76, 40]. When the trait is ordinal (Figures 3.7 and 3.9), LOCate2 has similar

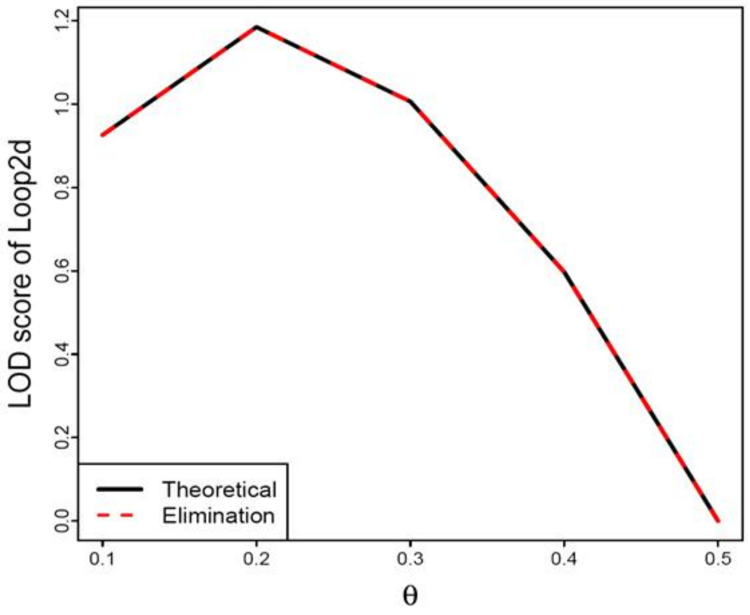

Figure 3.5: LOD curve for the inbred pedigree in Figure 3.1.

The black line is the theoretical LOD curve, and the red line is the LOD curve calculated by LOCate2.

power to QTLlink. LOCate2 and QTLlink outperform LOT on the ordinal trait with OR=12.06. When the trait is nominal (Figure 3.8), LOCate2 outperforms QTLlink and LOT when the penetrances are correctly specified or somewhat misspecified, and has similar power when the penetrances are very misspecified.

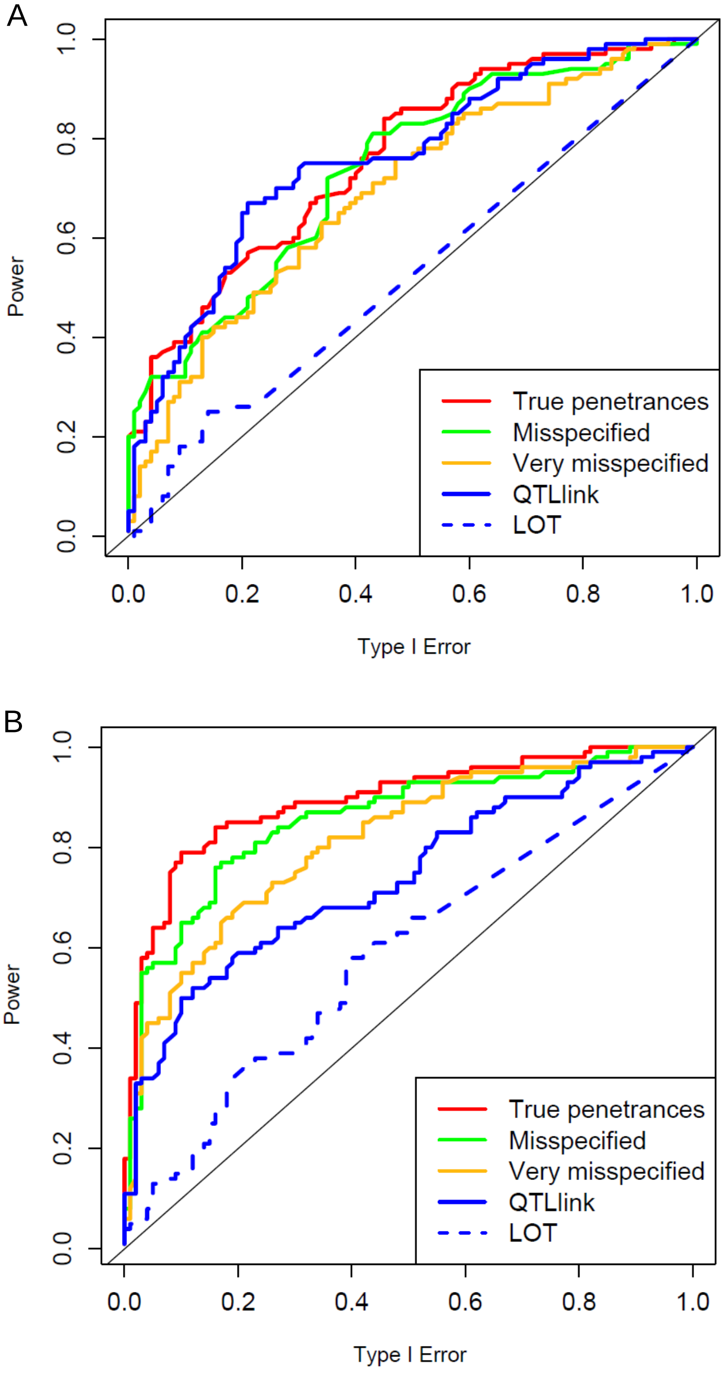

Figure 3.6: Power vs. Type I error of our method on large simulated pedigrees.

A. A simulated ordinal trait (Table 3.2). **B.** A simulated nominal trait (Table 3.3). Each simulation consists of 5 pedigrees, each containing 100 individuals, as in Figure 3.2.

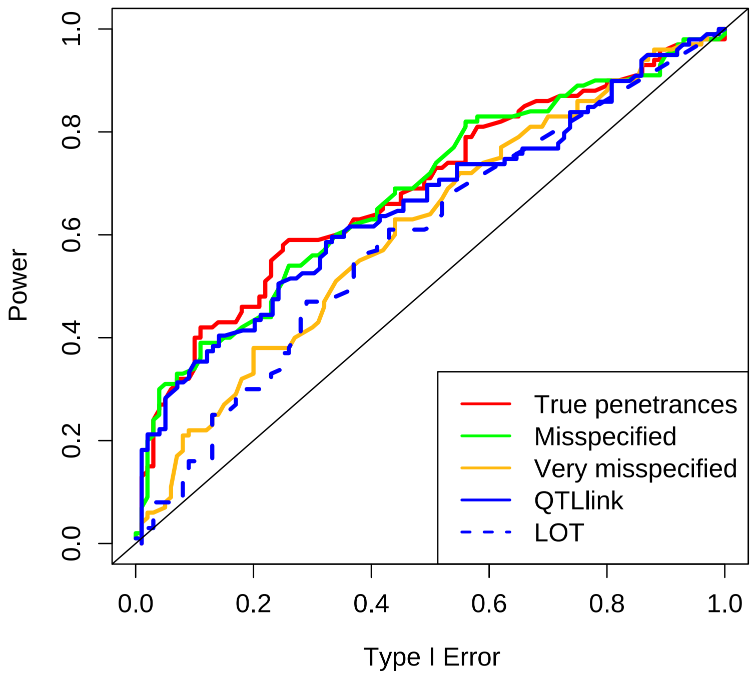

Figure 3.7: Power vs. Type I error on simulated small pedigrees with an ordinal trait.

Shown are the ROC curves for LOCate2, QTLlink, and LOT on an ordinal trait with OR=12.06 (Table 3.4). Each simulation consists of 100 pedigrees, each containing 5 individuals (two parents and three offspring).

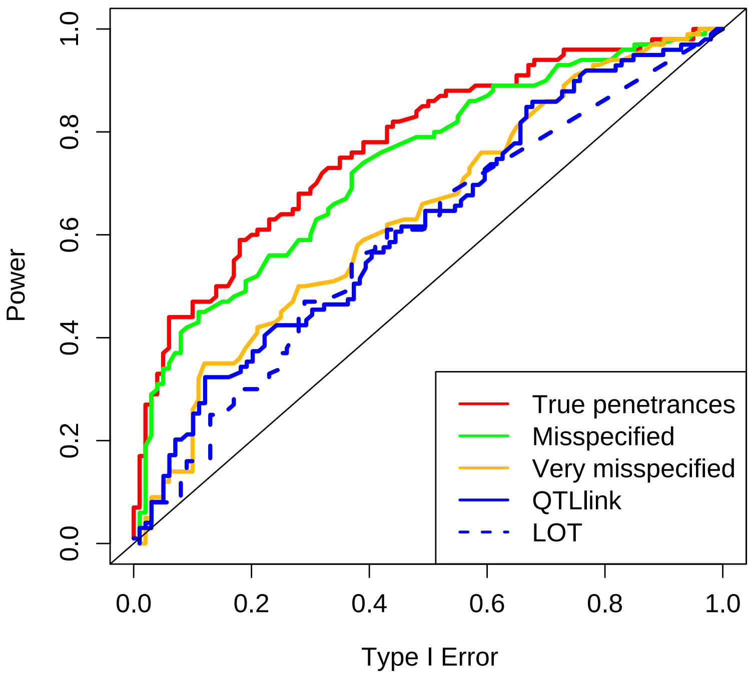

Figure 3.8: Power vs. Type I error on simulated small pedigrees with a nominal trait.

Shown are the ROC curves for LOCate2, QTLlink, and LOT on a nominal trait with OR=12.06 (Table 3.3). Each simulation consists of 100 pedigrees, each containing 5 individuals (two parents and three offspring).

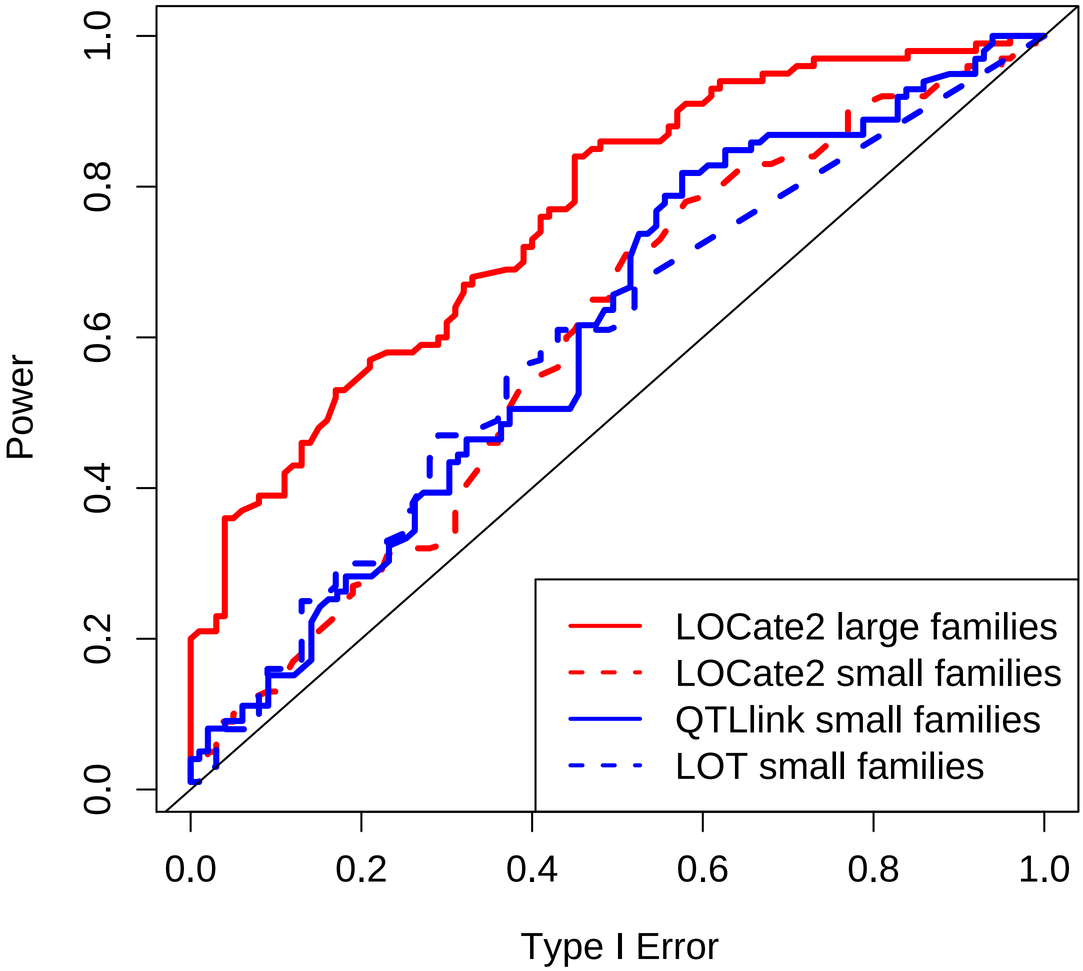

Figure 3.9: Lower power to detect QTLs in small families than large families, for a constant sample size.

Shown are the power vs. type I error of the three methods on simulated small pedigrees (100 pedigrees x 5 individuals) with an ordinal trait with OR=4.95 (Table 3.2). The solid red line shows the power vs. type I error of LOCate2 on simulated large pedigrees (5 pedigrees x 100 individuals) with the same sample size and OR.

3.4.2 Data Analysis

In our segregation analysis of the cataracts pedigree, the elimination algorithm ran in 45 seconds for 1 penetrance function on a 2 GHz desktop computer. Calculating the probability of the data under a free penetrance model (ranging from Pr(affected | qq,Qq,QQ)=(0,0,0) to (1,1,1) in intervals of .1 for each term) therefore required 12.5 hours. We found that the probability of the data under the completely penetrant recessive model was $6.38 * 10^{-46}$, under the incomplete recessive model was $1.995 * 10^{-41}$, and under the free model was $3.096 * 10^{-41}$. Using a likelihood ratio test, we can reject the complete recessive model in favor of the free model ($p < .0001$ by likelihood ratio test with 3 degrees of freedom), but we cannot reject the incomplete recessive model ($p = .3488$, 1 degree of freedom). This illustrates that our method can be used for efficient segregation analysis on large, complex pedigrees.

In our trichotomous analysis of the cardiac arrhythmia data set, we obtained the LOD scores shown in Figure 3.10. As expected, Model A (solid red line) gave LOD scores very similar the binary loopcut analysis (dotted black line) for most markers, as shown for marker 1 (FH2525) in Figure 3.10a. Markers 4 (FH2441) and 9 (FH2294) were exceptions to this trend. The discrepancy between model A and binary-loopcut for marker 9 may be due to the modifications we made to expedite the analysis; however, the discrepancy at marker 4 is harder to explain. As the LOD curves for marker 4 under models A and C are closer to that for the binary-unloopcut analysis, it appears that the LOD(0)=1.06 we observed for marker 4 in the loopcut analysis was an artifact of the loop-cutting, and this marker is not really linked to a QTL contributing to the ordinal nature of this trait.

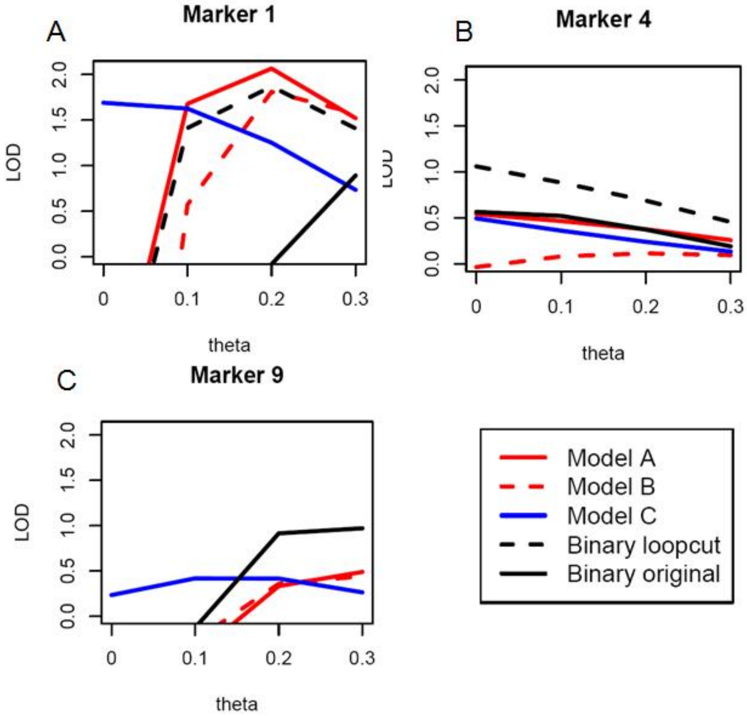

Figure 3.10: LOD scores plotted by marker.

Models A, B, and C are trichotomous models described in Table 3.6. Details of the markers are shown in Table 3.5.

As shown in Figure 3.11, marker 1 shows the highest LOD score in all 3 penetrance models, with LOD(.2)=2.066 under model A. This is perhaps not surprising, as marker 1 had the highest LOD score in the binary loopcut analysis (LOD(.2)=1.87). Since this marker's LOD increased under the nominal penetrance model, we consider it an excellent candidate for future investigation.

The LOD curves under model C tend to be flatter than the LOD curves under models A and B. This is not surprising: Model C has a lower odds ratio than models A and B, so more of the phenotypic variation is attributed to random noise rather than being used as evidence in favor of a particular value of θ. (Model C, an ordinal model based on a proportional-odds logistic model, has odds ratio = 7.39 for both phenotypes "severely affected" and "moderately or

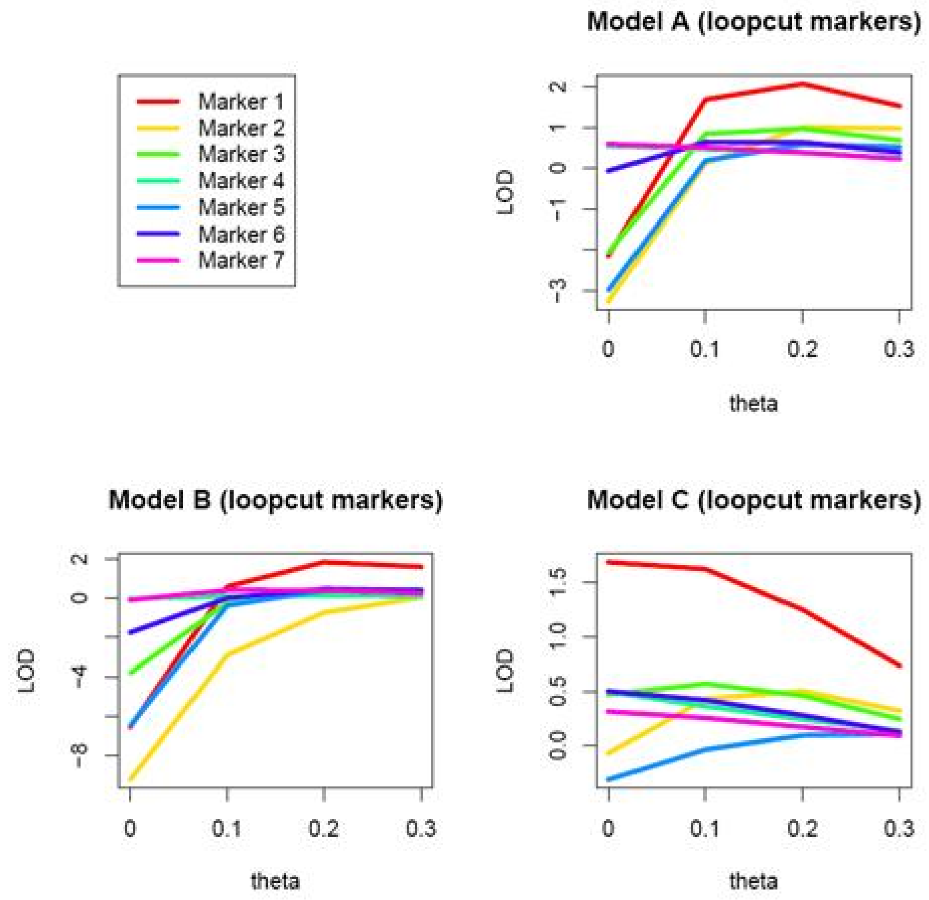

Figure 3.11: LOD scores plotted by penetrance model.

Shown are the LOD scores for the markers that had max LOD > .7 in the binary analysis of the cardiac arrhythmia data after inbreeding loops in the pedigree had been cut. Models A, B, and C are trichotomous models described in Table 3.6. Details of the markers are shown in Table 3.5.

severely affected", and the OR is independent of the genetic background onto which the additional disease allele is placed; that is, the ratio is the same for QQ vs Qq as for Qq vs qq. In contrast, models A and B have OR=∞ for either phenotype set when comparing Qq vs qq.) Because we have reason to believe this trait is ordinal rather than nominal, it would be valuable to explore other ordinal penetrance models, as well as to compare nominal models with similar odds ratios.

3.5 Discussion

In this paper, we present a fast, exact method for linkage analysis of ordinal and nominal traits. Our method is robust to missing data and computes the exact likelihood of the recombination rate θ even in pedigrees with some inbreeding loops. When used to analyze simulated large and small families, our method performs as well as QTLlink and better than LOT on ordinal traits, and better than both methods on nominal traits when the penetrances are correctly specified or somewhat misspecified.

When used to analyze real datasets, our method allowed efficient segregation analysis of a large, inbred pedigree of Labrador Retrievers. We also used our method to perform linkage analysis on a large pedigree of German Shepherd Dogs, but found it necessary to cut inbreeding loops to achieve a computationally feasible elimination order for this complex pedigree. Based on this analysis, we found additional evidence for linkage at microsatellite FH2525, and rejected the suggestion of linkage at FH2441 which was suggested by the binary analysis of the loop-cut pedigree. In the future, it would be beneficial to perform additional tests to refine the penetrance model, as well as to perform a trichotomous analysis on the other markers in this dataset, in case an important disease locus was left out of our candidate set due to the reduction in power that is expected when treating a categorical trait as binary.

In the future, this method could be enhanced by automating the choice of elimination order. *In general, choosing an optimal elimination order is an NP-*

complete problem, but a feasible elimination order can be identified efficiently by taking advantage of the small treewidth [5] of graphs corresponding to pedigrees with few inbreeding loops (or whose inbreeding loops have been cut). The method could also be enhanced by embedding the elimination algorithm inside a Markov chain Monte Carlo algorithm for Bayesian inference of θ.

We have implemented our method in the software LOCate2, available upon request. LOCate2 is a fast, accurate, and versatile approach for single marker analysis of nominal, ordinal, and binary traits on arbitrary pedigrees, including those with inbreeding loops and missing phenotypes and/or genotypes.

3.6 Acknowledgements

Financial support for this study was provided by NIH grant R01 GM083606-01 and the Consolidated Research Grant Program and the Collaborative Research in Preclinical and Clinical Sciences Program of the Cornell University College of Veterinary Medicine.

CHAPTER 4

PRINCIPAL COMPONENTS-BASED ASSIGNMENT OF ANCESTRY ALONG EACH CHROMOSOME IN INDIVIDUALS WITH ADMIXED ANCESTRY FROM 2 OR MORE POPULATIONS[3]

4.1 Abstract

Identifying ancestry along each chromosome in admixed individuals is of great interest for admixture mapping, understanding the population genetic history of admixture events, and identifying recent targets of selection. We present a Principal Components-based forward-backward algorithm for determining ancestry along each chromosome from a high-density, genomewide set of SNP genotypes of admixed individuals. We test our method on simulations which show that the method is robust to misspecification of ancestral populations and the number of generations since admixture. We apply our method to a dataset of Hispanic/Latino populations and identify regions of shared ancestry that may be recent targets of selection and could serve as candidate regions for admixture-based association mapping.

4.2 Introduction

Identifying ancestry along each chromosome in admixed individuals is of great interest for admixture mapping, understanding the population genetic history of admixture events, and identifying recent targets of selection. Several methods

[3] Brisbin, A., K. Bryc, L. Omberg, J. Degenhardt, A. Reynolds, J.G. Mezey, C.D. Bustamante. In preparation.

for identifying ancestry along each chromosome have been developed, including *structure* [57, 19], SABER [67], HAPMIX [56], and the principal components analysis (PCA) based method of Bryc *et al.* [7].

The problem of identifying ancestry along each chromosome involves a trade-off between speed and the number of parameters estimated. Methods such as SABER [67], which accounts for linkage disequilibrium between every pair of adjacent loci, and *structure* [57, 19], which estimates each individual's average ancestry proportions by Markov chain Monte Carlo, are too slow to be run on dense genome-wide data. HAPMIX [56] is a more recent method which is much faster; however, HAPMIX is not designed to assign ancestry to more than 2 ancestral populations.

In this paper, we expand upon the PCA-based method of Bryc *et al.* to produce PCAdmix, a method which uses phased genotype data to determine exact posterior probabilities of ancestry along each chromosome. The method is applicable for populations with admixture from 2 or more populations. We test our method on simulations which show that the method is robust to misspecification of ancestral populations and the number of generations since admixture. We also apply our method to assess 3-way European, Native American, and African admixture among Puerto Ricans, Ecuadorians, Dominicans, and Colombians in the NYULatino dataset [8] and identify 12 regions of extreme ancestry levels shared among multiple Latino populations, which may have experienced selection during admixture.

4.3 Methods

4.3.1 Haplotype Forward-Backward Algorithm

Our approach uses Principal Components Analysis (PCA) to identify how much information each SNP contributes to distinguishing the ancestry of a region. These PC loadings are used as weights in a weighted average of the allele values in a window of 10-20 SNPs, and the resulting window scores are used as the observed values in a Hidden Markov Model (HMM) to assign posterior probabilities to the ancestry in each window.

We first filtered out SNPs with high missingness, low minor allele frequency, and high linkage disequilibrium (LD $> .80$) with other SNPs in the dataset. The LD filtering step is not required, but was helpful in reducing the number of short regions of spurious ancestry assignment when applied to simulated chromosomes.

We used Singular Value Decomposition (the svd function) in R [58] to perform PCA on the phased genotypes of the ancestral representatives. The HapMap data had too many SNPs for R to admit a matrix of all 22 chromosomes, even after LD filtering. Therefore, we performed PCA separately for each chromosome. This approach has the advantage of not artificially combining haplotypes across chromosomes, as, in the absence of family data, such combined haplotypes have no meaning. We projected the admixed individuals onto the basis of principal components, and compute the observed ancestry "score" for haplotype i in window j as the weighted average $L_j g_{ij}$, where g_{ij} is a column vector of the haplotype's alleles (coded as $0/1$) in window j, standardized by the mean and standard deviation of that SNP in the ancestral populations, and L_j is a matrix in which the entry in the kth row, lth column is the loading of SNP l in window j on PC k.

For the forward-backward algorithm on our HMM, we used a haploid version of the transition probabilities in the Viterbi algorithm of Bryc *et al.* [7]. It is convenient to think of the probability of transitioning to a different ancestry as the probability of a recombination, times the probability that the recombination occurs with a chromosome from the target population, out of a "pool" of possible chromosomes in which the target population is represented with relative frequency q_j:

$$P(anc_i = j \mid anc_{i-1} = k) = \begin{cases} q_j\pi & \text{if } k \neq j \\ q_j\pi + (1-\pi) & \text{if } k = j \end{cases} \tag{4.1}$$

where anc_i is the ancestry at window i, q_j is the chromosome-wide proportion of population j ancestry in this haploid chromosome, and $\pi = 1 - e^{-d\hat{G}}$ is the probability of a single recombination having occurred in the distance d (in Morgans) between the midpoints of windows $i-1$ and i, during the estimated $\hat{G}$ generations since admixture. We assume that the windows are sufficiently dense that the probability of two or more recombinations is negligible, and this assumption is borne out by our method's robustness to mis-estimation of G, as demonstrated below.

For a given haplotype, q_j is estimated by $\frac{d_j}{\Sigma_i d_i}$, where d_j is the distance from the haplotype to the hyperplane containing the mean window scores of all ancestral populations other than j, as shown in Figure 4.1. In this way, a haplotype which falls far from the mean of population j will have a small value for d_j, and a small estimated proportion of ancestry from population j. To ensure that all transitions are possible in the HMM, we restricted $.01 \leq q_j \leq .99$ for all j.

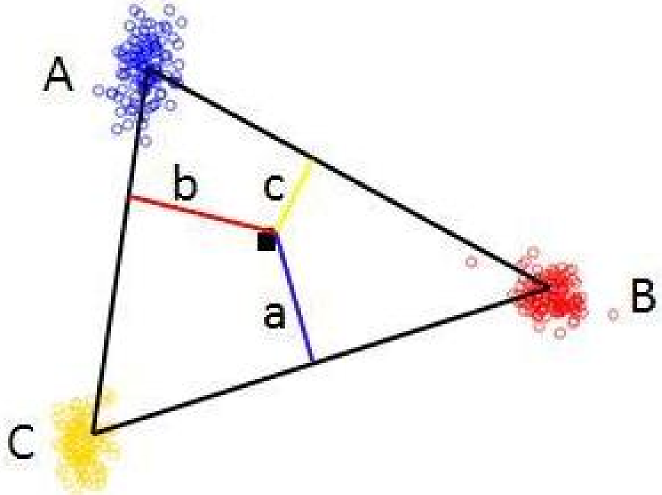

Figure 4.1: Estimation of average ancestry proportion for a haplotype.

For $k = 3$ ancestral populations, the population A average ancestry proportion of a haplotype (black square) is estimated by that haplotype's distance from the line connecting the PC1 and 2 means of the other two populations, as a proportion of the haplotype's total distance from all edges: $P(A) = \frac{a}{a+b+c}$.

The haplotype emission probabilities are similar to those used for genotypes in Bryc *et al.*: $w \mid anc_i \sim N(\mu_i, \Sigma_i)$, where w is the vector of window scores for an admixed haplotype; μ_i is a vector of length $k-1$ (where k is the number of ancestral populations), containing ancestral population i's mean scores for this window on the first $k-1$ PCs; and Σ_i is the covariance matrix of the scores for this window among population i. (In practice, each entry of Σ_i is the maximum of the relevant empirical covariance and .0001. This prevents Pr(w) from going to zero if, for example, all the sampled African haplotypes are identical within a particular window.)

Using the transition and emission probabilities described above, we use a forward-backward algorithm to find the posterior probability that the ancestry of a given window in a particular haplotype is population 1, 2, ... k. We can work with these probabilities directly or make hard assignments of ancestry if the posterior probability exceeds a calling threshold of $.50$, $.90$, or $.99$.

4.3.2 Simulations

We tested our method on simulated admixed individuals that were based on data from the HapMap3 project [11]. We used $G = 8$, freq(African) $\sim$ Beta(12, 3) to generate ancestry breakpoints and then copied haplotype segments from two ancestral individuals, one European (CEU) and one Yoruba (YRI). We analyzed these simulated data using a set of HapMap CEU and YRI founders as ancestral representatives, from which we had removed all individuals used to generate the simulations. We used $\hat{G}$=8, 20 SNPs per window, and LD $< .80$. We also tested our method's robustness to these parameters by allowing $\hat{G}$ to vary from 1 to 128 and the number of SNPs per window to vary from 1 to 160. We also tested our method by using various combinations of ancestral population representatives, including the true populations (YRI and CEU), sets of 3 populations (YRI, CEU, and Han Chinese and Japanese (CHB-JPT) or Italians (TSI)), and alternative populations with varying levels of relatedness, using Luhya (LWK) or Maasai (MKK) to represent YRI ancestry. The F_{ST} between Luhya and Yoruba is .0080 and between Maasai and Yoruba is .0270 [11]. For each of these analyses, we assessed the accuracy of our method, defined as the proportion of SNPs assigned to the correct (simulated) ancestry. We ignored SNPs that fell before the first window on the chromosome or after the last window, and SNPs that fell between windows assigned to different ancestries.

4.3.3 Application

We used our method to assign ancestry to chromosomal segments in individual NA19836, an African American individual in the HapMap3 [11] dataset. We used the transmitted haplotypes from the individual's parents from the HapMap trio-phased data. We used haplotypes from CEU and YRI trio founders as representatives of the ancestral groups. We ran the forward-backward algorithm with $\hat{G}$=4 and window size=20 SNPs.

We also used our method to assign ancestry in HapMap3 Mexican individual NA19730. HapMap phasing was not available for this individual, so we phased this individual's parents using IMPUTE v.2.1.0 [37], using a set of unrelated Mexican individuals from the HapMap phasing results as a reference panel. We used the same IMPUTE parameters as in the HapMap project: 110 iterations, 10 iterations of burn-in, and 120 conditioning states. We then used trio information to do deterministic phasing where possible, and used these SNPs to categorize the parents' haplotypes from IMPUTE as transmitted or untransmitted. The transmitted strands were used as the phased haplotypes for NA19730.

We compared three sets of ancestral representatives for NA19730. In each case, we assessed European vs. Native American ancestry, using HapMap3 CEU individuals as the European ancestral representatives. For the first set of Native American ancestral representatives, we used Maya, Pima, Karitiana, Surui, and Colombian individuals from the Human Genome Diversity Project (HGDP) [62] which a FRAPPE [68] analysis found to have $< 5\%$ European ancestry. For the second set, we used Maya, Nahua, and Aymara individuals from the Mao *et al.* data set [47] which a FRAPPE analysis found to have $< 1\%$ European ancestry. Because the Nahua individuals originated from the same general part of Mexico as the HapMap3 Mexican samples, the Mao *et al.* data set is expected to be a more accurate set of ancestral representatives. Finally, we used just the Nahua as the ancestral representative set. We phased the Native Americans using IMPUTE, using the same parameters as for phasing NA19730, but without trio information for validation. We used $\hat{G} = 8$ for NA19730 as well as for the other Latino individuals we studied.

We examined 3-way European, Native American, and African admixture in Hispanic individuals from the NYULatino project. We examined individuals from Ecuador, Colombia, Puerto Rico, and the Dominican Republic. We used CEU and YRI individuals from HapMap3 as European and African ancestral representatives, and Native Americans from HGDP as Native American ancestral representatives (matching the HGDP Native Americans used to analyze NA19730). Based on FRAPPE analyses and historical information, we expected Dominicans to have the greatest proportion of African ancestry, and Ecuadorians to have the greatest proportion of Native American ancestry [8]. We phased the Latino and Native American individuals in IMPUTE, and used HapMap phasings for the CEU and YRI individuals. We then computed the genomewide (autosomal) mean and standard deviation of the proportion of ancestry each Latino sample had from the African, European, and Native American ancestral groups, and normalized each window's ancestry proportion by the genomewide mean and standard deviation. Regions with ancestry proportions falling more than 3 standard deviations from the mean were considered to have “extreme” ancestry.

4.4 Results

4.4.1 Simulations

Our method is highly accurate in assigning ancestry along simulated chromosomes (Table 4.1), with increasing accuracy at more stringent calling thresholds. The method is robust to the choice of number of SNPs per window (Table 4.2), with any window size between 15 and 80 SNPs having accuracy $> 98\%$. This demonstrates that we pick up on consistent signals in the data, not artifacts of window subdivisions. Using fewer than 10 SNPs per window increases the number of spurious short ancestry regions identified (Figure 4.2).

Table 4.1: Comparison of our method and HAPMIX on simulated data.

Accuracy of our method and HAPMIX on simulated data, at several different probability thresholds.

Calling Threshold	Our method	HAPMIX
.5	98.1	99.2
.8	98.6	99.3
.9	98.8	99.3
.95	99.0	99.3
.99	99.2	99.5
.999	99.5	99.8

Table 4.2: Accuracy of our method under different window sizes.

SNPs per window	Accuracy (threshold=.5)
1	92.2
2	93.5
5	96.2
10	97.3
15	98.4
20	98.1
40	98.6
80	98.7
160	97.6

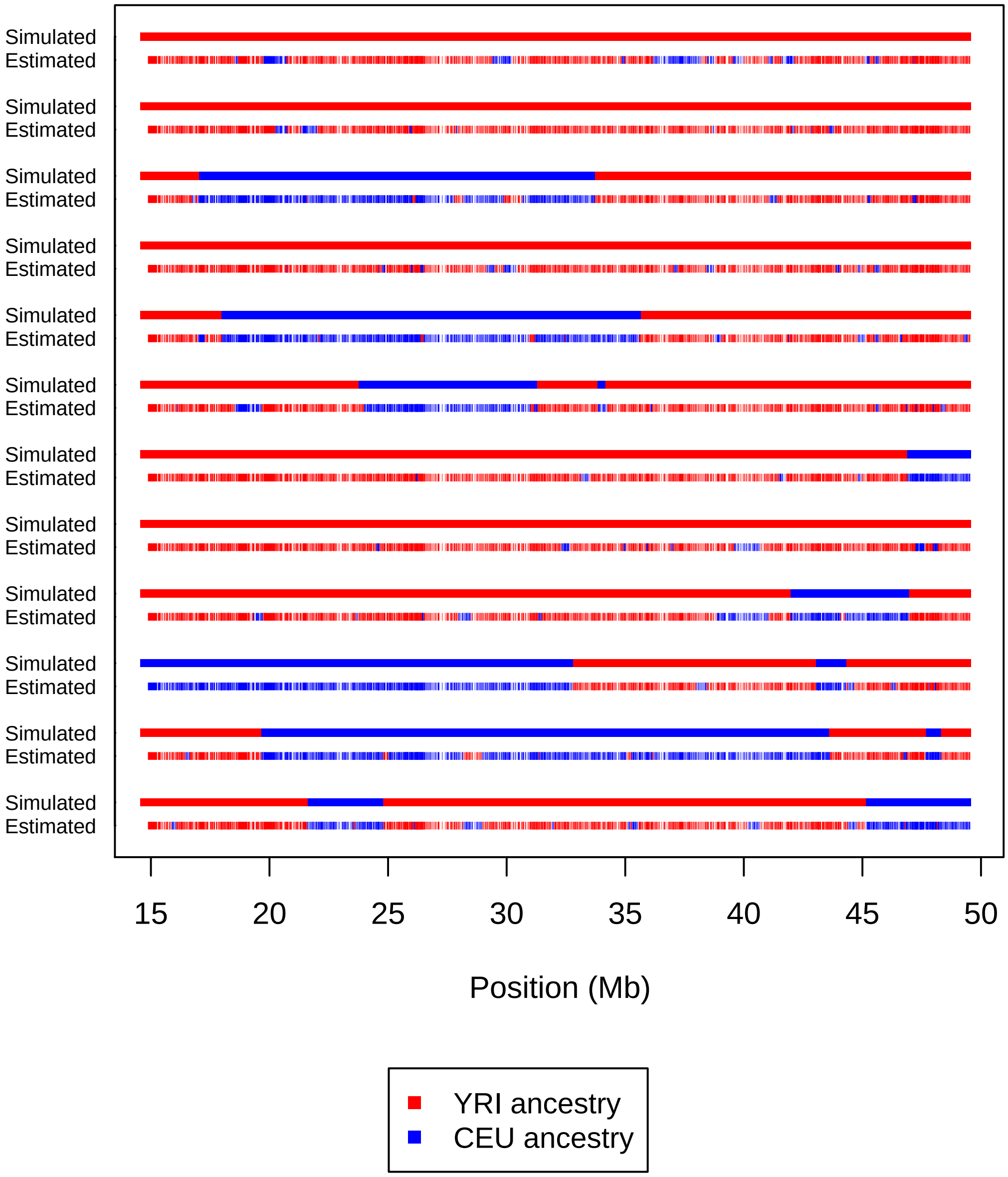

Figure 4.2: Ancestry segments assigned to simulated chromosomes using 2 SNPs per window.

Simulated chromosomes were formed from segments of CEU and YRI chromosomes (see text).

Our method, like HAPMIX [56], is robust to estimation of G, the number of generations since admixture (Table 4.3). This robustness is an advantage to researchers interested in mapping ancestry tracts, but may prove a challenge for fine-scale estimation of the timing of admixture events. We note that in our simulations, accuracy was slightly higher when G was somewhat underestimated than when $\hat{G} = G$. This is likely due to the improved smoothing over noisy window scores. Figure 4.3 demonstrates the difference in posterior probabilities calculated using $\hat{G} = G = 8$, containing a "spike" of intermediate posterior probability which would result in the incorrect inference of a short region of European ancestry, compared to $\hat{G} = 1$, where the lower transition probability has smoothed the spike.

Table 4.3: Accuracy under different values of $\hat{G}$.

The true value of G, the number of generations since admixture, for the simulations was 8.

$\hat{G}$	Accuracy (threshold=.5)	Accuracy (threshold=.9)
1	98.6	99.1
2	98.5	98.9
4	98.2	98.9
8	**98.1**	**98.8**
16	97.7	98.7
32	97.4	98.6
64	96.7	98.3
128	95.7	98.1

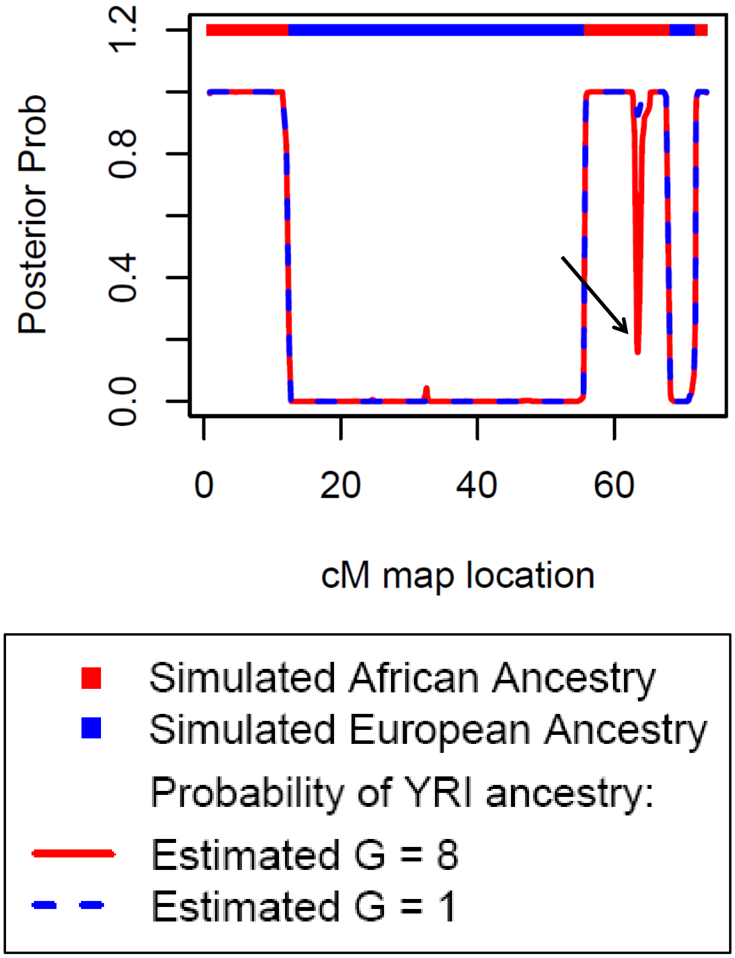

Figure 4.3: Posterior probabilities for a simulated chromosome under different values of $\hat{G}$.

The bar at the top indicates the true simulated ancestry of each chromosomal segment (red=YRI, blue=CEU). Red and dashed blue lines indicate the posterior probability of YRI ancestry at that window, using $\hat{G} = 8$ (red line) and $\hat{G} = 1$ (dashed blue) as the estimated number of generations since admixture. The true value of G used for the simulation was 8. The black arrow indicates a short region that has been incorrectly assigned to European ancestry when $\hat{G} = 8$.

HAPMIX performed slightly better than our method (Table 4.1) due to a lower number of incorrectly inferred short ancestry regions (Figure 4.4). However, it was also less sensitive to short regions of true ancestry (Figure 4.5, black oval). It is interesting to note that HAPMIX agreed with our method in the two longest tracts of incorrect ancestry assignment made by either method (one of which is depicted in Figure 4.5), suggesting that the Yoruba individuals used to simulate these segments may in fact have some European ancestry.

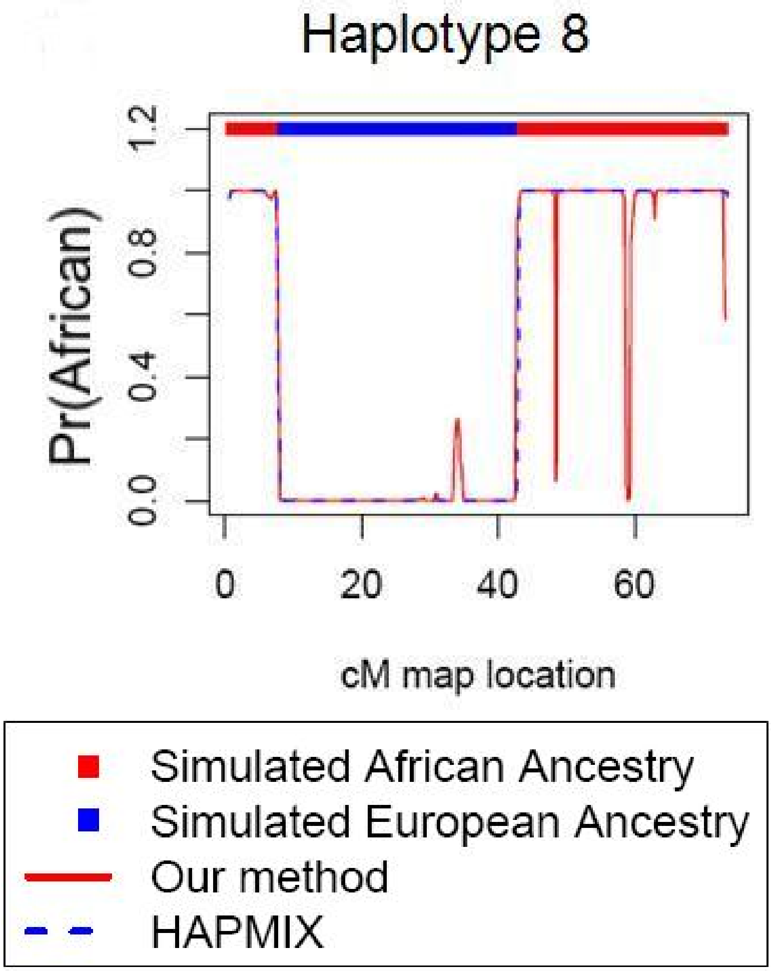

Figure 4.4: Posterior probabilities computed by HAPMIX and our method on simulated haplotype 8.

The bar at the top indicates the true simulated ancestry of each chromosomal segment (red=YRI, blue=CEU). Red and dashed blue lines indicate the posterior probability of YRI ancestry at that window, using our method (red line) and HAPMIX (dashed blue).

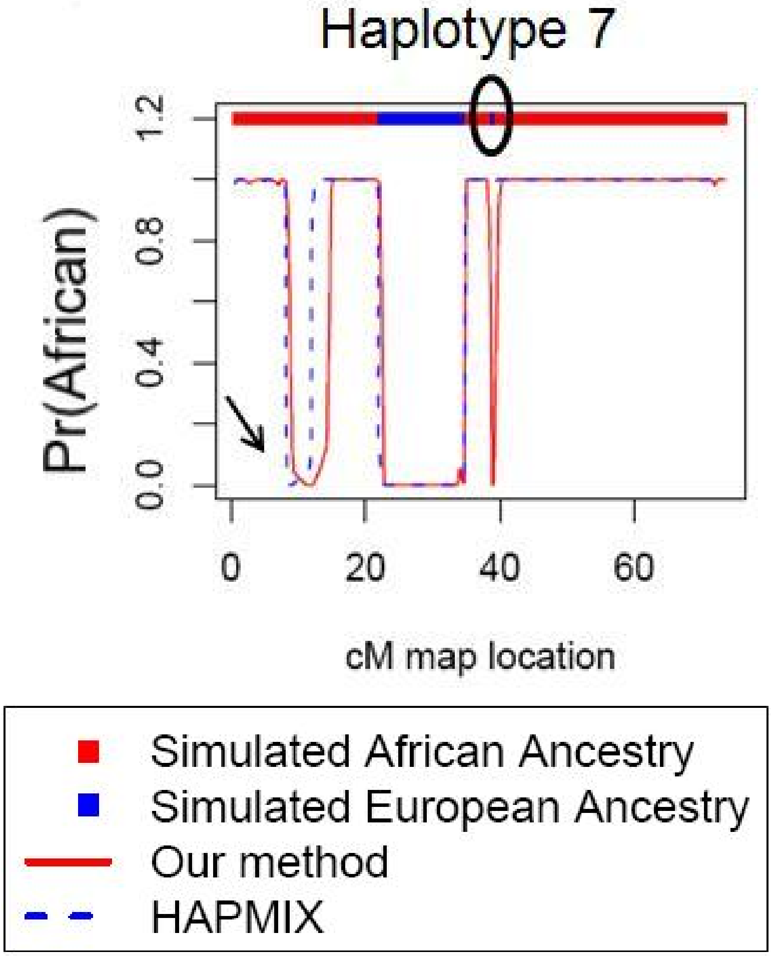

Figure 4.5: Posterior probabilities computed by HAPMIX and our method on simulated haplotype 7.

The bar at the top indicates the true simulated ancestry of each chromosomal segment (red=YRI, blue=CEU). Red and dashed blue lines indicate the posterior probability of YRI ancestry at that window, using our method (red line) and HAPMIX (dashed blue). The black oval indicates a short region of European ancestry. The black arrow indicates a region where both methods inferred European ancestry, although the segment was simulated from a YRI haplotype.

When the Luhya (LWK) or Maasai (MKK) were used as ancestral representatives for the Yoruba, our method's accuracy was essentially unchanged (Table 4.4), despite F_{ST} values of .0080 and .0270 between the true ancestral population and the population used to represent it. A simple Wright-Fisher simulation shows that 97.9% of the time, the F_{ST} between a population with effective population size = 5000 and the same population after 100 generations of drift is less than .027. This suggests that modern-day sampled individuals can be used as representatives for ancestral populations from previous generations without loss of accuracy due to genetic drift.

Table 4.4: Accuracy under different assumptions about the ancestral populations.

Accuracy listed is for a calling threshold of .5 (for 2 ancestral populations) or 1/3 (for 3 ancestral populations). The true ancestry of the simulations was YRI-CEU.

Tested ancestry	Accuracy when True ancestry = YRI	Accuracy when True ancestry = CEU	Overall Accuracy
YRI-CEU	97.7	99.3	98.1
MKK-CEU	98.3	97.9	98.2
LWK-CEU	97.5	99.1	97.9
YRI-CEU-(CHB-JPT)	96.8	98.7	97.2
YRI-CEU-TSI	97.5	51.6	86.1

Simulations show that our method can accurately assign ancestry to continent-level population groupings even when one of the ancestral representative groups is spurious, that is, when the admixed population contains no admixture from that group. Our method retained excellent accuracy when HapMap3 Han Chinese and Japanese (CHB-JPT) individuals were used as a third, spurious ancestral population, with only two African American haplotypes showing small regions assigned to CHB-JPT ancestry (Figure 4.6). In contrast, when the spurious ancestral population is closely related to one of the true ancestral populations, as in the YRI-CEU-TSI analysis (F_{ST}(CEU-TSI)=.004 [11]), our method experiences reduced accuracy due to the expected "splitting" of CEU ancestry into CEU and TSI assignments (Figure 4.7). The accuracy for SNPs whose true background is YRI remains high (97.5% at a calling threshold of 1/3, that is, assigning all SNPs; Table 4.4), but the accuracy for SNPs whose true background is CEU is no better than random guessing (51.6% for calling threshold=1/3), and is not improved by using a more stringent calling threshold.

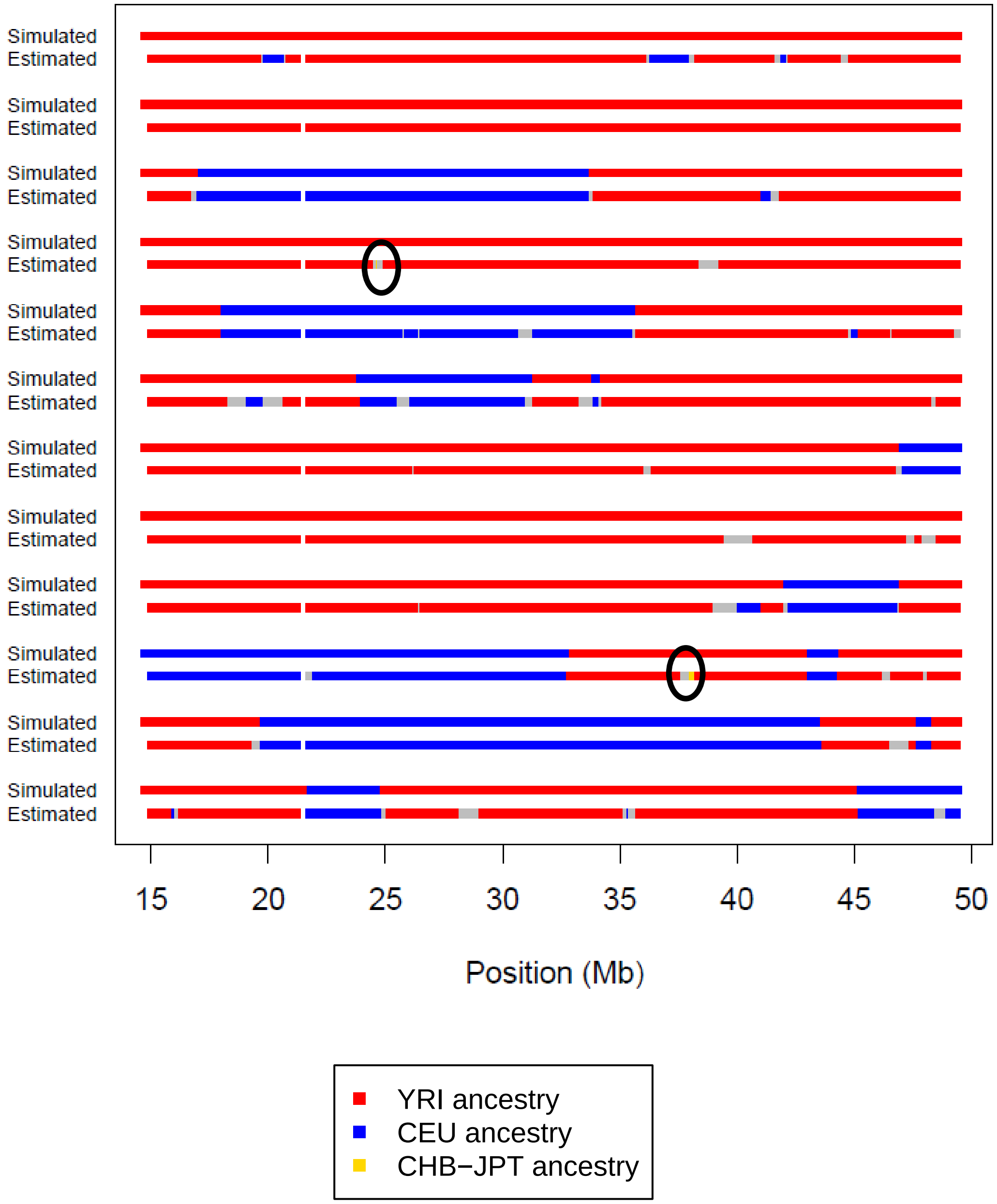

Figure 4.6: Ancestry assignments of YRI-CEU simulated haplotypes when analyzed using YRI, CEU, and CHB-JPT ancestral representatives.

The top line in each pair of chromosomes gives the simulated ancestry, and the bottom line shows the ancestry estimated by PCAdmix, using a calling threshold of .9. The black ovals indicate regions where our method incorrectly inferred CHB-JPT ancestry.

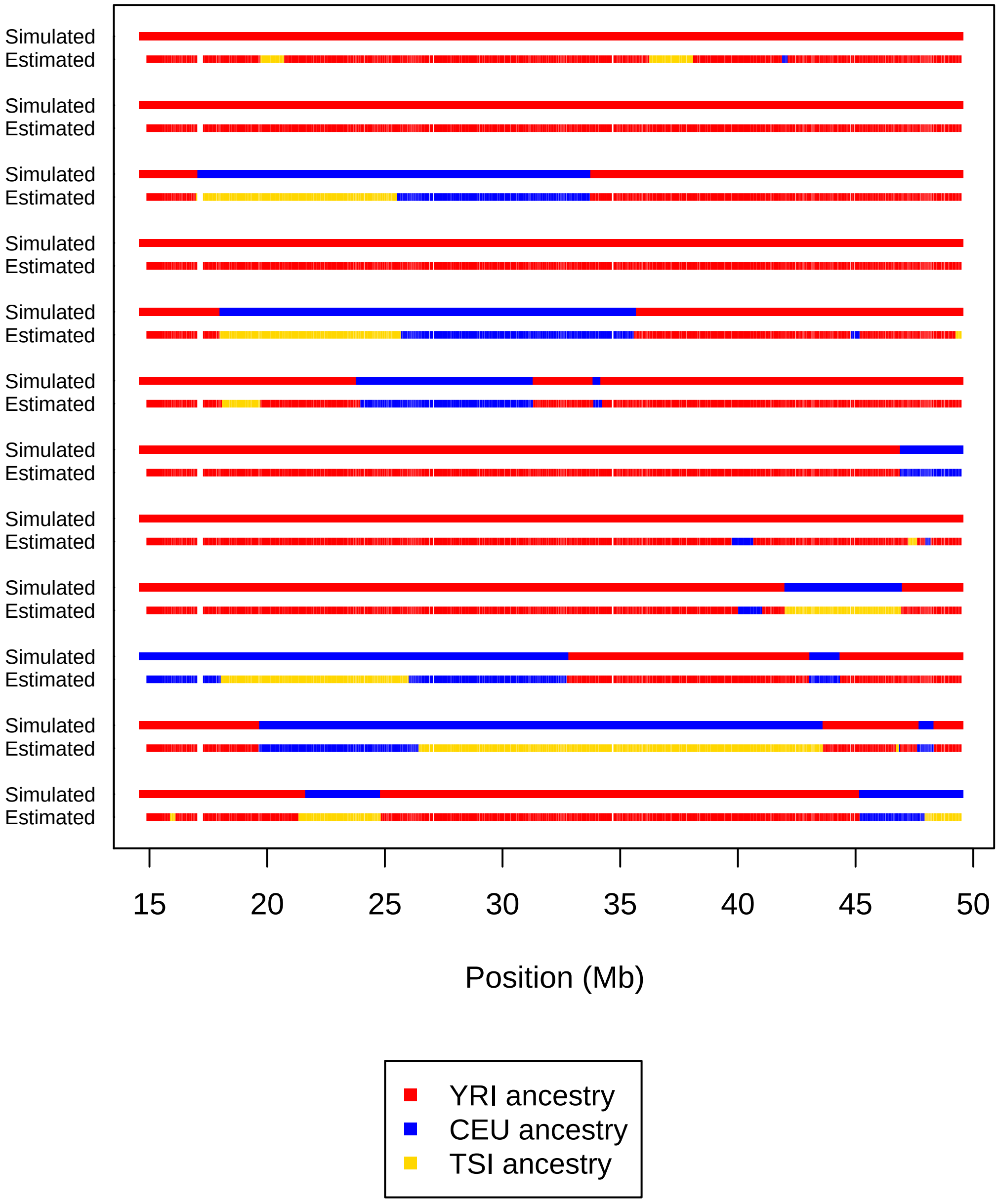

Figure 4.7: Ancestry assignments of YRI-CEU simulated haplotypes when analyzed using YRI, CEU, and TSI ancestral representatives.

The top line in each pair of chromosomes gives the simulated ancestry, and the bottom line shows the ancestry estimated by PCAdmix, using a calling threshold of .5.

4.4.2 Application

We were able to assign the ancestry of nearly all windows in the haplotypes of the African American individual NA19836 with posterior probability $> .9$. Using this calling threshold, the concordance between our assignments and those of a diploid analysis in HAPMIX was 98.5%. While most regions of African or European ancestry spanned many windows (80% of the tracts were over 1 Mb in length), some parts of the genome exhibited rapid switching of ancestry (Figure 4.8). Further investigation of these short segments is warranted; those which persist across many values of $\hat{G}$ and many calling thresholds, and where HAPMIX and PCAdmix agree, are likely to represent real features of the data, which may indicate recombination hotspots. In contrast, ancestry segments with only intermediate posterior probability which disappear under analysis with lower values of $\hat{G}$ (and therefore, lower transition probabilities) are more apt to be artifacts of the analysis, due to the fact that the maximum marginal posterior probability of ancestry for each window is not necessarily concordant with the most likely ancestry "path" through the chromosome.

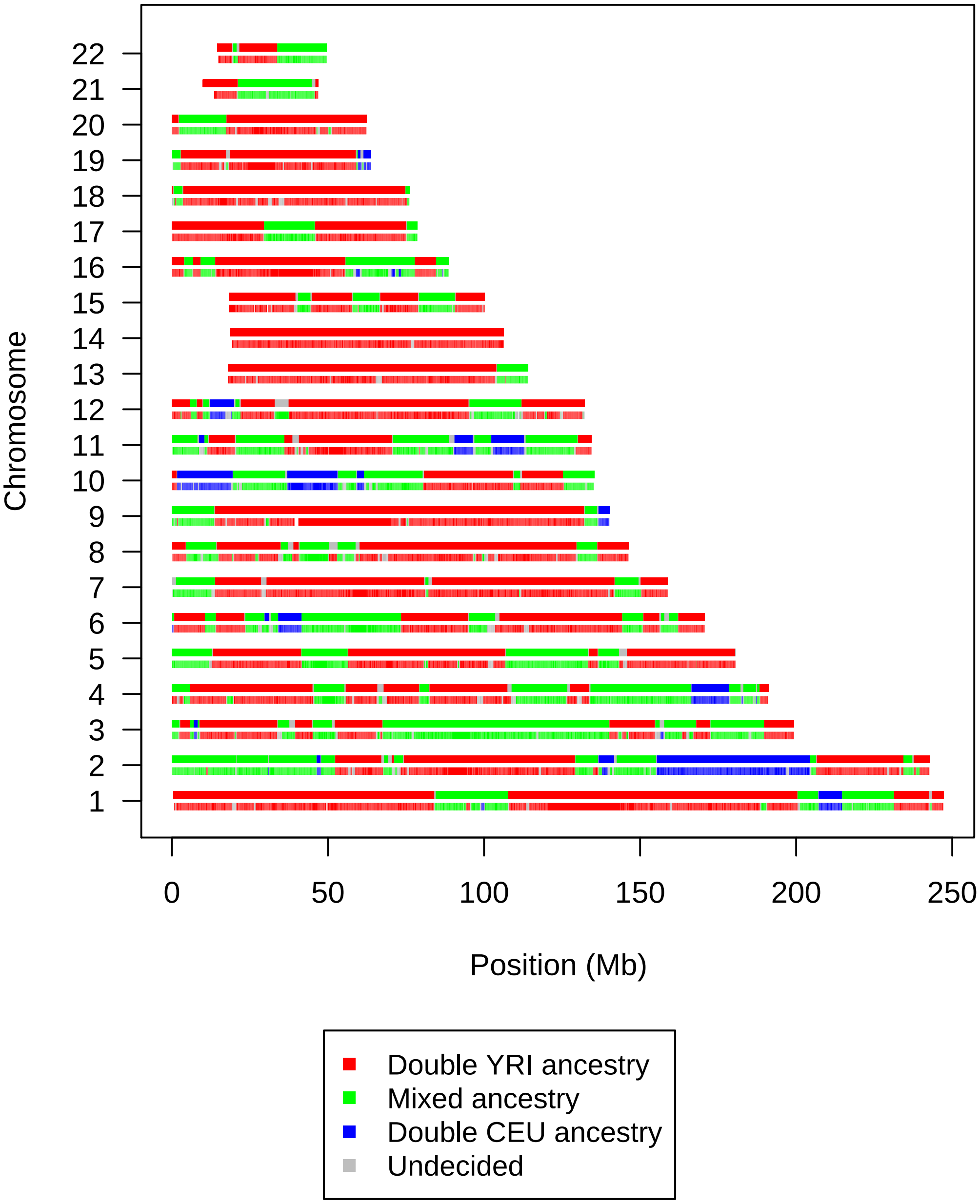

Figure 4.8: Analysis of an African American individual (NA19836) using HAPMIX and PCAdmix.

The bottom line of each chromosome is our method's diploid ancestry assignment of that chromosome; the top line is the assignment by HAPMIX. We used a calling threshold of .9 for both assignments.

Our analysis of the Mexican individual NA19730 revealed a large proportion of Native American ancestry (Figure 4.9), which agreed with PCA results on the unphased genotypes. The results using different Native American ancestral representatives were similar (Figure 4.9), reflecting the robustness to ancestral population misspecification we observed in our simulations. When we used the Mao *et al.* Native Americans as ancestral representatives, we observed fewer short regions of ancestry than with the HGDP Native Americans; however, it is not clear whether this is due to the Mao *et al.* Native Americans' being a better ancestral proxy, or to the decreased resolution obtained due to the lower number of SNPs in the combined HapMap3-Mao *et al.* data set.

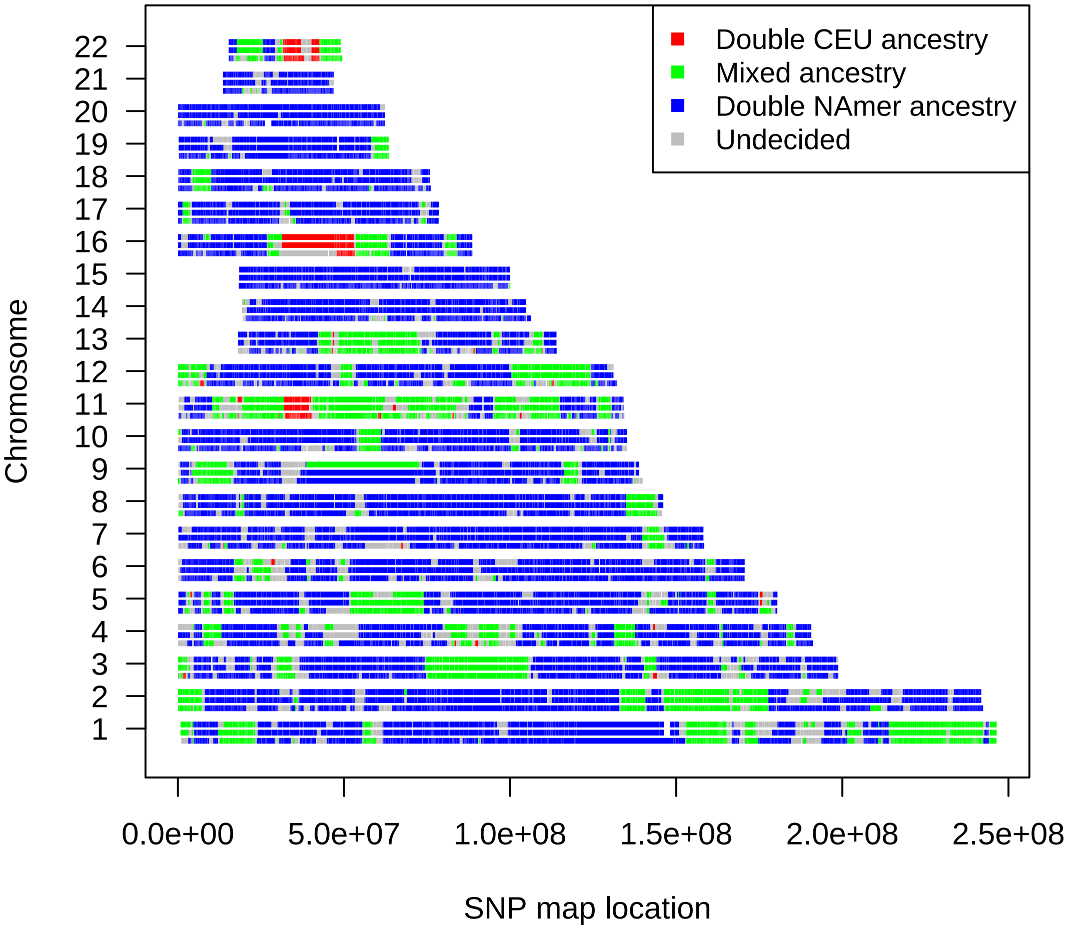

Figure 4.9: Analysis of a Mexican individual (NA19730) using different ancestral representative groups.

The bottom line of each chromosome uses HGDP Native Americans as the Native American ancestral representatives; the middle line uses Native Americans from the Mao *et al.* dataset; the top line uses only the Nahua from the Mao *et al.* dataset.

Our analysis of the NYULatino data confirmed our expectations about mean ancestry proportions (Figure 4.10): Dominicans, followed by Puerto Ricans, showed the largest proportion of African ancestry, and Ecuadorians, followed by Colombians, showed the largest proportion of Native American ancestry. We identified 12 regions having extreme levels of ancestry in more than one population (Table 4.5). In particular, regions on chromosomes 2, 6, and 8 showed elevated levels of ancestry in three of the four Latino populations (Figures 4.11, 4.12, 4.13). In addition, we identified several other regions showing extreme levels of ancestry in one population which warrant further investigation, perhaps with larger sample sizes.

These regions may have reached their extreme levels of ancestry due to selection during or after the initiation of admixture; it would be valuable to pursue further investigation in these regions, including simulations of admixture with and without selection, and a more detailed examination of haplotype diversity. The regions on chromosome 6 are especially intriguing, as Tang *et al.* [66] also found a region centered at 28.8 Mb to have elevated African ancestry in Puerto Ricans, and these regions are close to the human leukocyte antigen (HLA) loci (around 30-32 Mb). The HLA plays an important role in immunity, and may have undergone balancing selection favoring more-diverse African haplotypes. Another potential explanation for the extreme levels of ancestry is that the chromosome phasing was of lower quality in these regions, and the apparently greater haplotype diversity due to poor phasing was attributed to greater African ancestry; in particular, this may be a concern around the HLA loci, where high levels of diversity could complicate phasing. This concern is somewhat mitigated by the agreement between our findings and those of Tang *et al*, whose method, SABER [67], computes ancestry tracts from unphased data.

Nevertheless, it would be valuable to repeat this analysis on Latino individuals from genotyped family trios, where the phasing can be more certain.

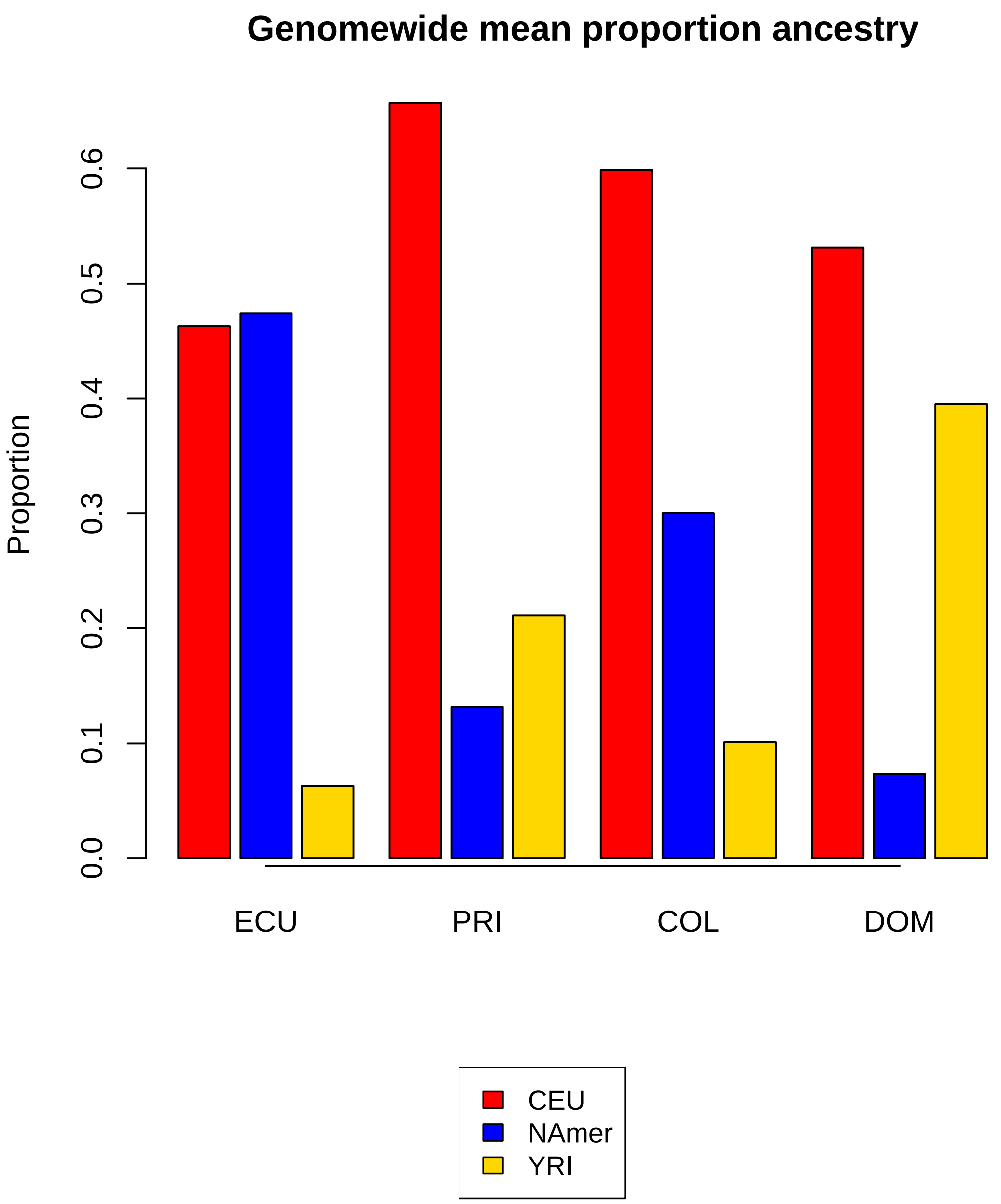

Figure 4.10: Genomewide autosomal mean ancestry proportions in four Latino populations.

COL = Colombian; DOM = Dominican; ECU = Ecuadorian; PRI = Puerto Rican.

Table 4.5: Regions showing extreme ancestry proportions in multiple Latino populations.

All regions shown here exhibited ancestry proportions more than 3 standard deviations above the genomewide mean for that population. YRI = Yoruba (African); NAmer = Native American; COL = Colombian; DOM = Dominican; ECU = Ecuadorian; PRI = Puerto Rican.

Chromosome	Position (Mb)	Ancestry	Latino Populations
2	136.8-136.9	NAmer	COL, DOM, PRI
6	27.3-28.8	YRI	COL, ECU, PRI
6	31.4-31.5	YRI	COL, ECU, PRI
8	10.8-10.9	NAmer	COL, DOM, PRI
2	134.9-135.5	NAmer	DOM, PRI
5	30.5-30.9	YRI	COL, ECU
8	8.4-8.8	NAmer	DOM, PRI
11	87.5-87.6	YRI	COL, PRI
13	58.3-58.5	NAmer	DOM, PRI
15	59.7-59.8	YRI	ECU, PRI
15	60.8-61.0	YRI	ECU, PRI
15	66.8-67.5	YRI	COL, ECU

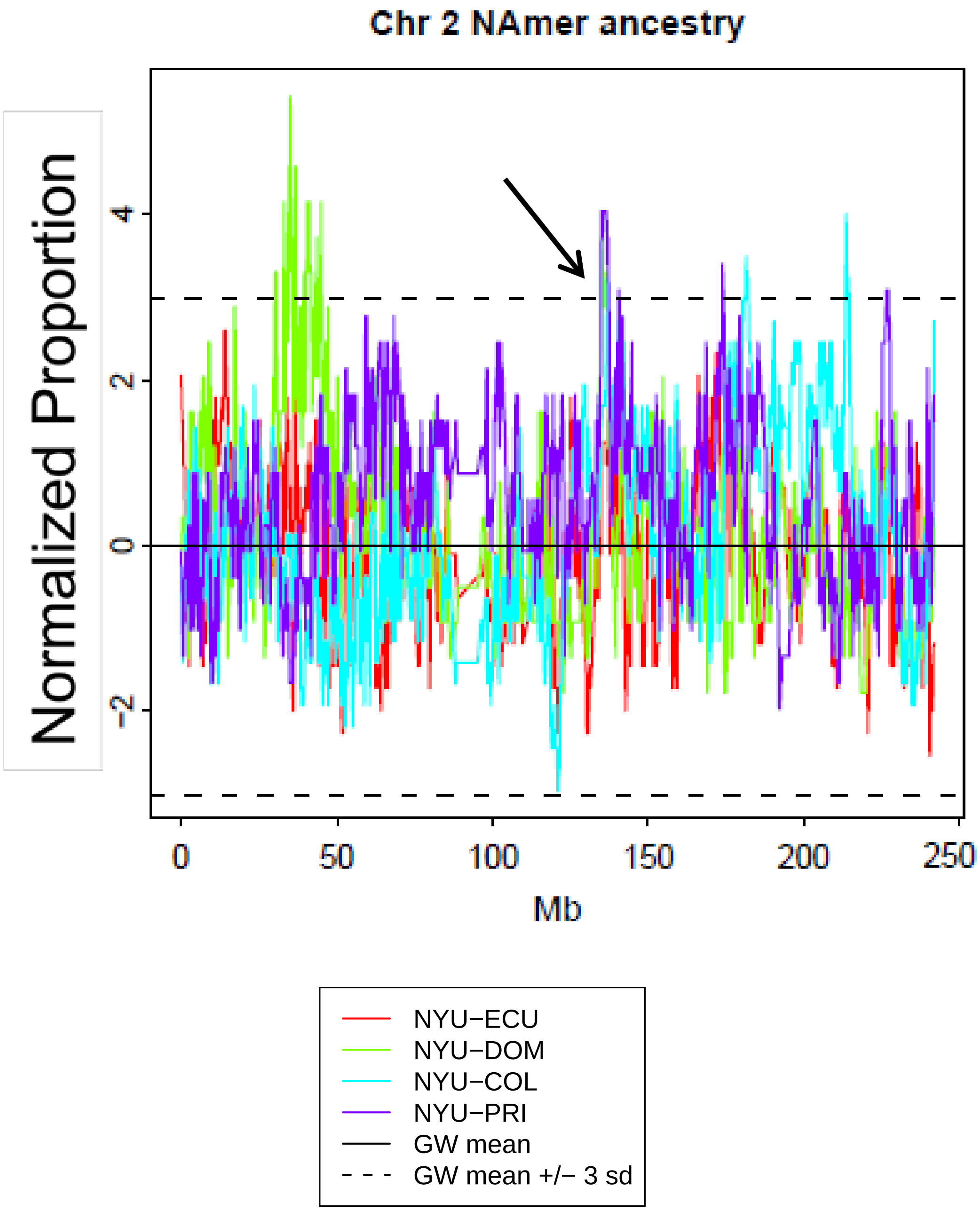

Figure 4.11: Normalized proportions of Native American ancestry on chromosome 2 in Latino populations.

The dashed lines indicate the values 3 standard deviations from the mean. The black arrow indicates a region where Colombians, Dominicans, and Puerto Ricans have extremely high proportions of Native American ancestry.

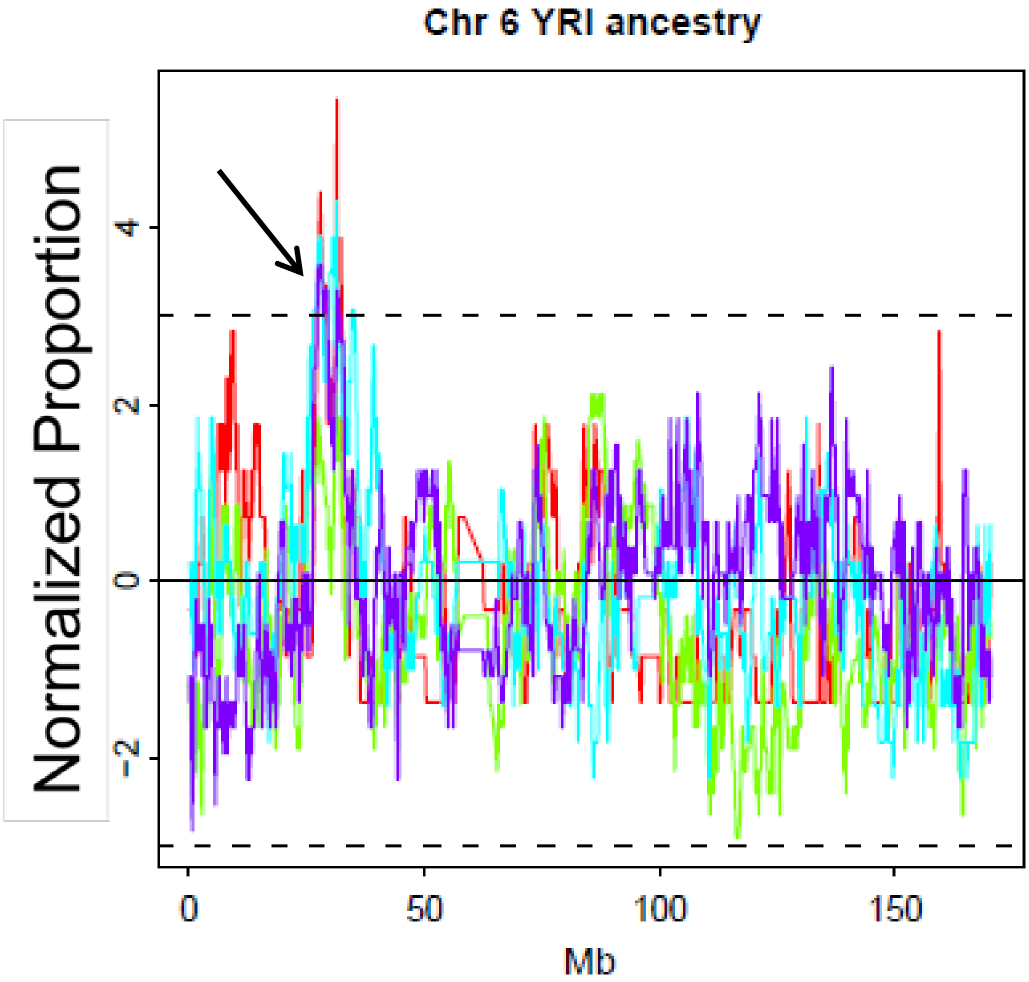

Figure 4.12: Normalized proportions of African ancestry on chromosome 6 in Latino populations.

The black arrow indicates a pair of regions where Colombians, Ecuadorians, and Puerto Ricans have extremely high proportions of African ancestry. See Figure 4.11 for legend.

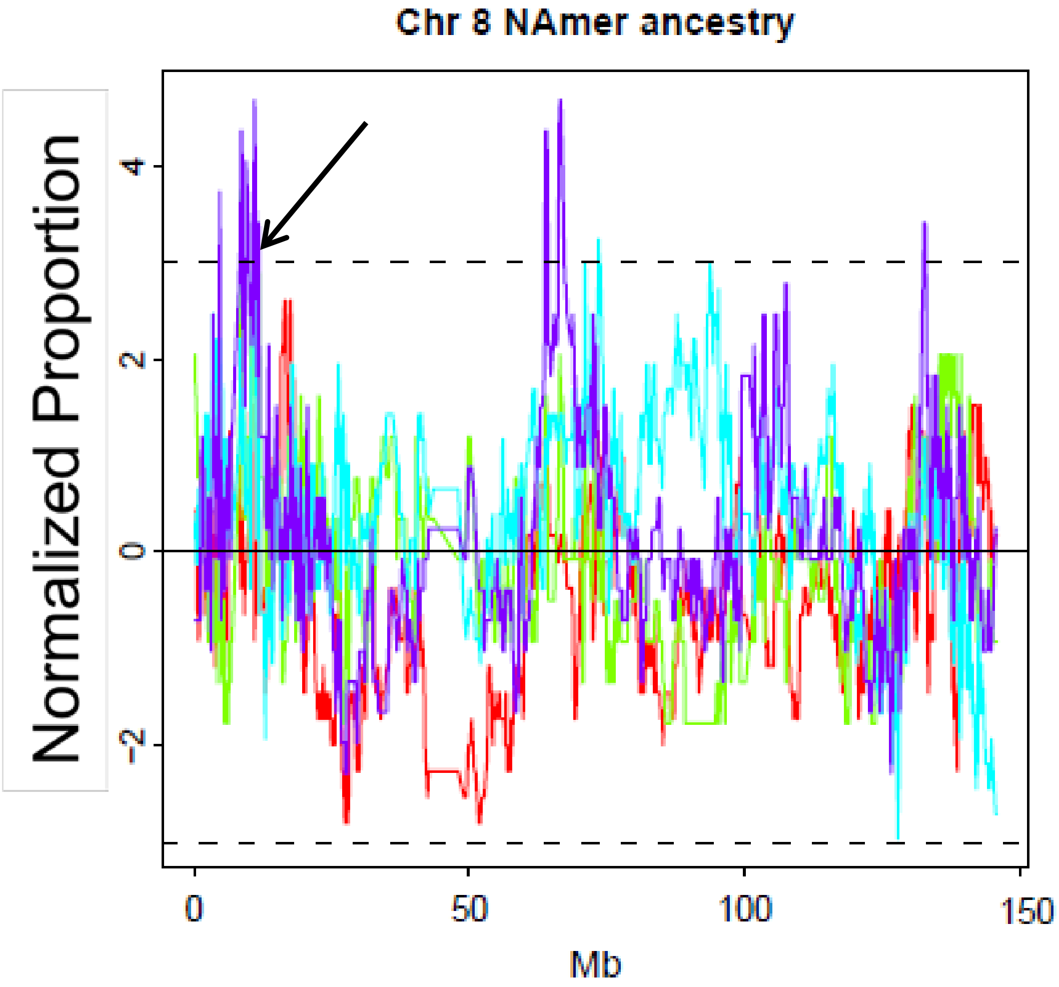

Figure 4.13: Normalized proportions of Native American ancestry on chromosome 8 in Latino populations.

The black arrow indicates a region where Colombians, Dominicans, and Puerto Ricans have extremely high proportions of Native American ancestry. See Figure 4.11 for legend.

4.5 Discussion

In this paper, we have presented a Principal Components-based approach to assigning ancestry along the genome in admixed individuals. Our approach is highly accurate at assigning ancestry, and is applicable to admixture of 2 or more populations. The approach is robust to the choice of window size, to misspecification of ancestral populations, and to the estimation of time since admixture. We have implemented our method in the software PCAdmix, available upon request.

In future, this method could be enhanced by the development of a wrapper HMM to estimate the time since admixture based on length of ancestry tracts, and by an investigation of “bootstrapping” results when ancestral representatives are not available; as pointed out in [48], admixture proportions are detectable even without source populations for up to 15 generations after admixture. The method would also benefit from an implementation of the PCA portion of the code in C, which would enable the simultaneous analysis of genome-wide, rather than chromosome-wide, data sets.

We have demonstrated that our method is useful in identifying regions of extreme ancestry proportions within populations, which may indicate sites of selection during or after the process of admixture. Our method will also be valuable for admixture mapping on dense genomewide data and for understanding the population genetic history of admixed populations.

4.6 Acknowledgements

This research was supported by NSF grant number 0516310R01.

APPENDIX A

SUPPLEMENTARY INFORMATION FOR CHAPTER 2

A.1 Equations used in variable updates

The update for disease locus alleles Q_{fi} and Q_{mi}, jointly with selector variables $sel_{Q,fi}$ and $sel_{Q,mi}$, is analogous to that for M_{fi} and M_{mi} (equation 2.2), with the substitution of $P(d_i|Q_{fi}, Q_{mi}, penetrance)$ for $P(M_{i,obs}|M_{fi}, M_{mi})$:

$$
\begin{aligned}
(Q_{fi}, Q_{mi}, sel_{Q,fi}, sel_{Q,mi} \quad | & \quad \text{Markov Blanket}) \propto \\
& \quad P(Q_{fi} \mid \mathbf{Q}_f, sel_{Q,fi}) \cdot P(Q_{mi} \mid \mathbf{Q}_m, sel_{Q,mi}) \\
\cdot & \quad P(d_i \mid Q_{fi}, Q_{mi}, penetrance) \\
\cdot & \quad P(sel_{Q,fi} \mid sel_{marker,fi}) \cdot P(sel_{Q,mi} \mid sel_{marker,mi}) \\
\cdot & \quad \Pi_{offspring=j} P(Q_{ij} \mid Q_{fi}, Q_{mi}, sel_{Q,ij})
\end{aligned}
$$

Here,

$$
P(sel_{Q,fi}|sel_{marker,fi}) = \begin{cases} 1-\theta & \text{for } sel_{Q,fi} = sel_{marker,fi} \\ \theta & \text{for } sel_{Q,fi} \neq sel_{marker,fi} \end{cases}
$$

where θ is the probability of recombination between the marker and the disease locus; that is, individual i's disease locus and marker alleles come from different haplotypes with probability θ.

For founders, $P(Q_{fi}|\mathbf{Q}_f, sel_{Q,fi})$ is replaced by

$$
P(Q_{fi}) = \begin{cases} a & \text{if } Q_{fi} = Q \\ 1-a & \text{if } Q_{fi} = q \end{cases}
$$

where a is a constant describing the frequency of the disease allele in the founder population.

If the unphased marker genotype $M_{i,obs}$ is unobserved, it is updated according to the distribution $P(M_{i,obs}|M_{fi}, M_{mi})$ (equation 2.3). If the phenotype d_i is unobserved, it is updated according to the distribution $P(d_i|Q_{fi}, Q_{mi}, penetrances)$, determined by the penetrance matrix.

A.2 Simulated Tempering

In our chain, at $\lambda = 0$, the penetrances, recombination rate, mutation rate, and frequency of the disease allele are assigned their desired values (recombination rate=θ, mutation rate=0, freq(Q) as set by user, penetrances as described in the user-specified matrix). At $\lambda = 1$, all parameters are relaxed to uniform probabilities to allow faster mixing (recombination rate=.5, disease locus mutation rate=.5, marker mutation rate=$\frac{m-1}{m}$, where m is the number of possible marker alleles; freq(Q)=.5, $P(d_i = j \mid g = k) = 1/n$, where n is the number of levels of the trait). At intermediate λs, each parameter p_λ is a linear combination: $p_\lambda = (1 - \lambda) * p_{\lambda=0} + \lambda * p_{\lambda=1}$.

At each iteration, the temperature of the chain is updated according to a Metropolis-Hastings algorithm. The first 50,000 iterations of each sampler run are used to fine-tune the rate of temperature transitions according to the Robbins-Munro method [27]. After this fine-tuning, the chain is sampled whenever $\lambda = 0$, when its stationary distribution coincides with the desired posterior distribution P($Y \mid X, \theta$).

To assess whether simulated tempering was effective in improving the mixing, we examined the lag-k autocorrelation of $P(X, Y \mid \theta)$ for runs of the Gibbs sampler with and without simulated tempering, starting from the same initial configuration. Whenever the tempered chain visited $\lambda = 0$, we recorded $P(X, Y)$ for both chains. Figure 2.2 shows the correlation between $P(X, Y_i \mid \theta = .10)$ and $P(X, Y_{i+k} \mid \theta = .10)$ for visits i and $i + k$ to $\lambda = 0$, for $1 \leq k \leq 100$. The autocorrelation with simulated tempering (with 7 temperatures) quickly drops to below .05, "near-independence" levels, while the autocorrelation for a run of the sampler without simulated tempering remains above .3 even for $k = 100$. This demonstrates that simulated tempering effectively improved the mixing of our Gibbs sampler.

A.3 Application to Data

The three additional trichotomous models we tested are shown in Table A.1.

Table A.1: Additional trichotomous penetrance models used to analyze Panic Disorder data.

We tested each of these models on the 96 subfamilies discussed in the Application to Data section of chapter 2, in addition to the selected model (model A) in Table 2.3.

Model	Phenotype	qq	Qq	QQ
B	$d = 1$	.99	.3	.2
	$d = 2$	.005	.4	.3
	$d = 3$	.005	.3	.5
C	$d = 1$	.9	.2	.05
	$d = 2$	.05	.6	.15
	$d = 3$	.05	.2	.8
D	$d = 1$	.9	.05	.05
	$d = 2$	.05	.9	.05
	$d = 3$	.05	.05	.9

APPENDIX B

SUPPLEMENTARY INFORMATION FOR CHAPTER 3

B.1 Elimination Algorithm Example

As an example of the elimination algorithm reducing the number of terms required to compute $P(X \mid \theta)$, consider a pedigree consisting of mother (A), father (B), and two offspring (C and D). Letting Z_i be the observed phenotype of individual i, and g_i be the joint genotype at the marker and disease locus, the joint probability of the family's phenotypes is

$$\begin{aligned} P(Z_A, Z_B, Z_C, Z_D) &= \Sigma_{g_A}\Sigma_{g_B}\Sigma_{g_C}\Sigma_{g_D} P(Z_A \mid g_A)P(g_A \mid HWE) \\ &\cdot \quad P(Z_B \mid g_B)P(g_B \mid HWE) \\ &\cdot \quad P(Z_C \mid g_C)P(g_C \mid g_A, g_B)P(Z_D \mid g_D)P(g_D \mid g_A, g_B) \end{aligned} \tag{B.1}$$

requiring $3^4 = 81$ terms, because there are 3 possible genotypes to consider for each individual's portion of the summation, assuming the disease locus is diallelic. (Here, $P(g_A|HWE)$ is the probability of founder A's genotype under Hardy-Weinberg equilibrium, conditional on allele frequencies in the population, which are assumed to be known.)

However, the above equation can be rewritten as

$$\begin{aligned} P(Z_A, Z_B, Z_C, Z_D) &= \Sigma_{g_A} P(Z_A \mid g_A)P(g_A \mid HWE)\Sigma_{g_B} P(Z_B \mid g_B)P(g_B \mid HWE) \\ &\cdot \quad \Sigma_{g_C} P(Z_C \mid g_C)P(g_C \mid g_A, g_B)\Sigma_{g_D} P(Z_D \mid g_D)P(g_D \mid g_A, g_B) \end{aligned}$$

"Eliminating" individual D corresponds to computing

$$f_1(g_A, g_B) = \Sigma_{g_D} P(Z_D \mid g_D)P(g_D \mid g_A, g_B),$$

which requires 27 terms, because each of A, B, and D have 3 possible genotypes. It is convenient to think of f_1 as a 3x3 table which describes $P(Z_D \mid g_A, g_B)$ as a function of A and B's genotypes. We are left with the simplified formula

$$\begin{aligned} P(Z_A, Z_B, Z_C, Z_D) &= \Sigma_{g_A} P(Z_A \mid g_A) P(g_A \mid HWE) \Sigma_{g_B} P(Z_B \mid g_B) P(g_B \mid HWE) \\ &\cdot f_1(g_A, g_B) \Sigma_{g_C} P(Z_C \mid g_C) P(g_C \mid g_A, g_B). \end{aligned}$$

Eliminating C similarly requires 27 terms, leaving us with the formula

$$\begin{aligned} P(Z_A, Z_B, Z_C, Z_D) &= \Sigma_{g_A} P(Z_A \mid g_A) P(g_A \mid HWE) \Sigma_{g_B} P(Z_B \mid g_B) P(g_B \mid HWE) \\ &\cdot f_1(g_A, g_B) f_2(g_A, g_B). \end{aligned}$$

Eliminating B involves 9 terms, computing

$$f_3(g_A) = \Sigma_{g_B} P(Z_B \mid g_B) P(g_B \mid HWE) f_1(g_A, g_B) f_2(g_A, g_B)$$

to obtain

$$P(Z_A, Z_B, Z_C, Z_D) = \Sigma_{g_A} P(Z_A \mid g_A) P(g_A \mid HWE) f_3(g_A).$$

Finally, we eliminate A by summing 3 terms and giving the desired solution to $P(Z_A, Z_B, Z_C, Z_D)$. The total number of terms involved in this calculation is $27 + 27 + 9 + 3 = 66$, compared to 81 terms for the brute-force summation of equation B.1. The elimination algorithm gains even larger computational savings over the brute-force method in larger pedigrees, in which greater numbers of individuals are conditionally independent, as the two offspring were in this example.

B.2 Simulations

We used PyPedal's options simulate_n=45, simulate_ns=10, simulate_nd=10, simulate_g=5, simulate_ir=0, simulate_mp=0, simulate_fs=1, simulate_po=1.

These options create a pedigree of 45 individuals, of which 10 are founder sires and 10 are founder dams. Simulate_g controls the number of generations in the simulated pedigree; however, changes to this parameter did not have much effect on the structure of the pedigree. The ir and mp parameters disallow immigration and missing parents among individuals not counted in the founders set, and the fs and po parameters allow full-sib and parent-offspring matings. After simulation, we removed any individuals which were disconnected from the pedigree. We then added "leaf" individuals to bring the pedigree size up to 100 individuals, distributing the individuals as offspring of existing matings randomly according to a uniform distribution.

BIBLIOGRAPHY

[1] G.R. Abecasis, S.S. Cherny, W.O. Cookson, and L.R. Cardon. Merlin - rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet*, 30:97–101, 2002.

[2] L. Almasy and J. Blangero. Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet*, 62(5):1198–1211, 1998.

[3] S. Apostolakis, V. Amanatidou, E.G. Papadakis, and D.A. Spandidos. Genetic diversity of CX3CR1 gene and coronary artery disease: New insights through a meta-analysis. *Atherosclerosis*, 207(1):8–15, 2009.

[4] L.E. Baum. An equality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, 3:1–8, 1972.

[5] HL Bodlaender, JR Gilbert, H. Hafsteinsson, and T. Kloks. Approximating treewidth, pathwidth, frontsize, and shortest elimination tree. *Journal of Algorithms*, 18(2):238–255, 1995.

[6] A. Brisbin, M.M. Weissman, A.J. Fyer, S.P. Hamilton, J.A. Knowles, C.D. Bustamante, and J.G. Mezey. Bayesian linkage analysis of categorical traits for arbitrary pedigree designs. *Submitted to PLoS One. In revision.*

[7] K. Bryc, A. Auton, M.R. Nelson, J.R. Oksenberg, S.L. Hauser, S. Williams, A. Froment, J.M. Bodo, C. Wambebe, S.A. Tishkoff, et al. Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proceedings of the National Academy of Sciences*, 107(2):786, 2010.

[8] K. Bryc, C. Velez, T. Karafet, A. Moreno-Estrada, A. Reynolds, A. Auton, M. Hammer, C.D. Bustamante, and H. Ostrer. Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proc Natl Acad Sci USA, in press*, 2010.

[9] R. Chakraborty and K.M. Weiss. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proceedings of the National Academy of Sciences*, 85(23):9119, 1988.

[10] J.B. Cole. PyPedal: A computer program for pedigree analysis. *Computers and Electronics in Agriculture*, 57(1):107–113, 2007.

[11] The International HapMap 3 Consortium. An integrated haplotype map of rare and common genetic variation in diverse human populations. In revision. 2010.

[12] J. Corbett, C.C. Gu, J.P. Rice, T. Reich, M.A. Province, and D.C. Rao. Power loss for linkage analysis due to the dichotomization of trichotomous phenotypes. *Hum Hered*, 57(1):21–27, 2004.

[13] J. Cruickshank, R.L. Quaas, J. Li, S. Hemsley, T.M. Gunn, and N.S. Moïse. Genetic analysis of ventricular arrhythmia in young German Shepherd dogs. *Journal of Veterinary Internal Medicine*, 23(2):264–270, 2009.

[14] A.P. Dempster, N.M. Laird, D.B. Rubin, et al. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[15] G. Diao and D.Y. Lin. Variance-components methods for linkage and association analysis of ordinal traits in general pedigrees. *Genetic Epidemiology*, 2009.

[16] J. Dupuis, J. Shi, A.K. Manning, E.J. Benjamin, J.B. Meigs, L.A. Cupples, and D. Siegmund. Mapping quantitative traits in unselected families: algorithms and examples. *Genetic Epidemiology*, 33(7), 2009.

[17] A.O. Edwards, R. Ritter III, K.J. Abel, A. Manning, C. Panhuysen, and L.A. Farrer. Complement factor H polymorphism and age-related macular degeneration. *Science*, 308(5720):421, 2005.

[18] R.C. Elston and J. Stewart. A general model for the genetic analysis of pedigree data. *Hum Hered*, 21(6):523–542, 1971.

[19] D. Falush, M. Stephens, and J.K. Pritchard. Inference of population structure using multilocus genotype data linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587, 2003.

[20] R. Feng, J.F. Leckman, and H. Zhang. Linkage analysis of ordinal traits for pedigree data. *Proc Natl Acad Sci USA*, 101(48):16739–16744, 2004.

[21] M. Fishelson and D. Geiger. Exact genetic linkage computations for general pedigrees. *Bioinformatics*, 18(Suppl 1):S189–S198, 2002.

[22] K.A. Frazer, D.G. Ballinger, D.R. Cox, D.A. Hinds, L.L. Stuve, R.A. Gibbs,

J.W. Belmont, A. Boudreau, P. Hardenbol, S.M. Leal, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861, 2007.

[23] A.J. Fyer, S.P. Hamilton, M. Durner, F. Haghighi, G.A. Heiman, R. Costa, O. Evgrafov, P. Adams, A.B. de Leon, N. Taveras, D.F. Klein, S.E. Hodge, M.M. Weissman, and J.A. Knowles. A third-pass genome scan in panic disorder: Evidence for multiple susceptibility loci. *Biol Psychiatry*, 60(4):388–401, 2006.

[24] E.R. Gansner and S.C. North. An open graph visualization system and its applications to software engineering. *Software Practice and Experience*, 30(11):1203–1233, 2000.

[25] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell*, 6:721–741, 1984.

[26] C.J. Geyer. Reweighting Monte Carlo mixtures. Technical Report 568, School of Statistics, University of Minnesota, 1991.

[27] C.J. Geyer and E.A. Thompson. Annealing Markov chain Monte Carlo with applications to ancestral inference. *J Am Stat Assoc*, pages 909–920, 1995.

[28] W.R. Gilks and G.O. Roberts. Strategies for improving MCMC. In W.R. Gilks, S. Richardson, and D. Spiegelhalter, editors, *Markov chain Monte Carlo in Practice*, pages 89–114. Chapman & Hall/CRC, 1996.

[29] S. Girirajan, J.A. Rosenfeld, G.M. Cooper, F. Antonacci, P. Siswara, A. Itsara, et al. A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nature Genetics*, 42(3):203–209, 2010.

[30] Huntington's Disease Collaborative Research Group. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, 72:971–983, 1993.

[31] C.A. Hackett and J.I. Weller. Genetic mapping of quantitative trait loci for traits with ordinal distributions. *Biometrics*, 51:1252–1263, 1995.

[32] J.M. Hall, M.K. Lee, B. Newman, J.E. Morrow, L.A. Anderson, B. Huey, and M.C. King. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*, 250(4988):1684, 1990.

[33] W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, pages 97–109, 1970.

[34] T. Hayashi and T. Awata. Interval mapping for loci affecting unordered categorical traits. *Heredity*, 96(2):185–194, 2006.

[35] S.C. Heath. Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am J Hum Genet*, 61(3):748–760, 1997.

[36] C.J. Hoggart, E.J. Parra, M.D. Shriver, C. Bonilla, R.A. Kittles, D.G. Clayton, and P.M. McKeigue. Control of confounding of genetic associations in stratified populations. *The American Journal of Human Genetics*, 72(6):1492–1504, 2003.

[37] B.N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5(6), 2009.

[38] V.M. Ingram. Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. *Nature*, 180(4581):326–328, 1957.

[39] M.I. Jordan. Graphical models. *Stat Sci*, 19:140–155, 2004.

[40] S.A. Knott and C.S. Haley. Maximum likelihood mapping of quantitative trait loci using full-sib families. *Genetics*, 132(4):1211, 1992.

[41] L. Kruglyak, M.J. Daly, M.P. Reeve-Daly, and E.S. Lander. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet*, 58(6):1347–1363, 1996.

[42] J. Li, S. Wang, and Z.B. Zeng. Multiple-interval mapping for ordinal traits. *Genetics*, 173(3):1649, 2006.

[43] F. Liu, A. Kirichenko, T.I. Axenovich, C.M. van Duijn, and Y.S. Aulchenko. An approach for cutting large and complex pedigrees for linkage analysis. *European Journal of Human Genetics*, 16(7):854–860, 2008.

[44] E.J. Lyons, W. Amos, J.A. Berkley, I. Mwangi, M. Shafi, T.N. Williams, C.R. Newton, N. Peshu, K. Marsh, J.A.G. Scott, et al. Homozygosity and risk of childhood death due to invasive bacterial disease. *BMC Medical Genetics*, 10(1):55, 2009.

[45] D.J.C. MacKay. Introduction to Monte Carlo methods. In M.I. Jordan, editor, *Learning in graphical models*, pages 175–204. Kluwer Academic Publishers, Boston, 1998.

[46] T.A. Manolio, F.S. Collins, N.J. Cox, D.B. Goldstein, L.A. Hindorff, D.J. Hunter, M.I. McCarthy, E.M. Ramos, L.R. Cardon, A. Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[47] X. Mao, A.W. Bigham, R. Mei, G. Gutierrez, K.M. Weiss, T.D. Brutsaert, F. Leon-Velarde, L.G. Moore, E. Vargas, P.M. McKeigue, et al. A genomewide admixture mapping panel for Hispanic/Latino populations. *The American Journal of Human Genetics*, 80(6):1171–1178, 2007.

[48] G. McVean. A genealogical interpretation of Principal Components Analysis. *PLoS Genetics*, 5(10):1–10, 2009.

[49] C.S. Mellersh, L. Pettitt, O.P. Forman, M. Vaudin, and K.C. Barnett. Identification of mutations in HSF4 in dogs of three different breeds with hereditary cataracts. *Veterinary ophthalmology*, 9(5):369–378, 2006.

[50] J.F. Monahan. *Numerical methods of statistics*. Cambridge Univ Pr, 2001.

[51] J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, A.R. Boyko, A. Auton, A. Indap, K.S. King, S. Bergmann, M.R. Nelson, et al. Genes mirror geography within Europe. *Nature*, 456(7218):98, 2008.

[52] N. Patterson, N. Hattangadi, B. Lane, K.E. Lohmueller, D.A. Hafler, J.R. Oksenberg, S.L. Hauser, M.W. Smith, S.J. OBrien, D. Altshuler, et al. Methods for high-density admixture mapping of disease genes. *The American Journal of Human Genetics*, 74(5):979–1000, 2004.

[53] L. Pauling, H.A. Itano, S.J. Singer, and I.C. Wells. Sickle cell anemia. *Science*, 110:543–8, 1949.

[54] D. Pitocco, G. Zelano, G. Gioffrè, E. Di Stasio, F. Zaccardi, F. Martini, T. Musella, G. Scavone, M. Galli, S. Caputo, et al. Association Between Osteoprotegerin G1181C and T245G Polymorphisms and Diabetic Charcot Neuroarthropathy. *Diabetes Care*, 32(9):1694, 2009.

[55] J.E. Pool and R. Nielsen. Inference of historical changes in migration rate from the lengths of migrant tracts. *Genetics*, 181(2):711, 2009.

[56] A.L. Price, A. Tandon, N. Patterson, K.C. Barnes, N. Rafaels, I. Ruczinski, T.H. Beaty, R. Mathias, D. Reich, and S. Myers. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics*, 5(6):e1000519, 2009.

[57] J.K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[58] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006.

[59] S. Rao and X. Li. Strategies for genetic mapping of categorical traits. *Genetica*, 109(3):183–197, 2000.

[60] S. Rao and S. Xu. Mapping quantitative trait loci for ordered categorical traits in four-way crosses. *Heredity*, 81(2):214–224, 1998.

[61] G.O. Roberts. Markov chain concepts related to sampling algorithms. In W.R. Gilks and DJ Spiegelhalter, editors, *Markov chain Monte Carlo in Practice*, pages 45–58. Chapman & Hall/CRC, 1996.

[62] N.A. Rosenberg, J.K. Pritchard, J.L. Weber, H.M. Cann, K.K. Kidd, L.A. Zhivotovsky, and M.W. Feldman. Genetic structure of human populations. *Science*, 298(5602):2381, 2002.

[63] S. Sankararaman, S. Sridhar, G. Kimmel, and E. Halperin. Estimating local ancestry in admixed populations. *The American Journal of Human Genetics*, 82(2):290–303, 2008.

[64] A. Sundquist, E. Fratkin, C.B. Do, and S. Batzoglou. Effect of genetic divergence in identifying ancestral origin using HAPAA. *Genome Research*, 18(4):676, 2008.

[65] N.B. Sutter, C.D. Bustamante, K. Chase, M.M. Gray, K. Zhao, L. Zhu, B. Padhukasahasram, E. Karlins, S. Davis, P.G. Jones, et al. A single IGF1 allele is a major determinant of small size in dogs. *Science*, 316(5821):112, 2007.

[66] H. Tang, S. Choudhry, R. Mei, M. Morgan, W. Rodriguez-Cintron, E.G. Burchard, and N.J. Risch. Recent genetic selection in the ancestral admixture of Puerto Ricans. *The American Journal of Human Genetics*, 81(3):626–633, 2007.

[67] H. Tang, M. Coram, P. Wang, X. Zhu, and N. Risch. Reconstructing genetic ancestry blocks in admixed individuals. *The American Journal of Human Genetics*, 79(1):1–12, 2006.

[68] H. Tang, J. Peng, P. Wang, and N.J. Risch. Estimation of individual admixture: analytical and study design considerations. *Genetic epidemiology*, 28(4):289–301, 2005.

[69] H.K. Tang and D. Siegmund. Mapping quantitative trait loci in oligogenic models. *Biostatistics*, 2(2):147, 2001.

[70] S. Tavaré, D.J. Balding, R.C. Griffiths, and P. Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518, 1997.

[71] A. Thomas. Linkage analysis on complex pedigrees by simulation. *IMA J Math Appl Med Biol*, 11(2):79–93, 1994.

[72] D.C. Thomas and V. Cortessis. A Gibbs sampling approach to linkage analysis. *Hum Hered*, 42:63–76, 1992.

[73] D.C. Thomas and W.J. Gauderman. Gibbs sampling methods in genetics. In W.R. Gilks, S. Richardson, and D. Spiegelhalter, editors, *Markov chain Monte Carlo in Practice*, pages 419–440. Chapman & Hall/CRC, Boca Raton, 1996.

[74] X. Wang, Y. Ye, and H. Zhang. Family-based association tests for ordinal traits adjusting for covariates. *Genet Epidemiol*, 30(8), 2006.

[75] R. Wooster, S.L. Neuhausen, J. Mangion, Y. Quirk, D. Ford, N. Collins, K. Nguyen, S. Seal, T. Tran, D. Averill, et al. Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science*, 265(5181):2088, 1994.

[76] C. Xie, D.D.G. Gessler, and S. Xu. Combining different line crosses for mapping quantitative trait loci using the identical by descent-based variance component method. *Genetics*, 149(2):1139, 1998.

[77] C. Xu, Y.M. Zhang, and S. Xu. An EM algorithm for mapping quantitative resistance loci. *Heredity*, 94(1):119–128, 2005.

[78] S. Xu and C. Xu. A multivariate model for ordinal trait analysis. *Heredity*, 97(6):409–417, 2006.

[79] B.S. Yandell, T. Mehta, S. Banerjee, D. Shriner, R. Venkataraman, J.Y. Moon, W.W. Neely, H. Wu, R. Von Smith, and N. Yi. R/qtlbim: QTL with Bayesian interval mapping in experimental crosses. *Bioinformatics*, 23(5):641, 2007.

[80] N. Yi, S. Banerjee, D. Pomp, and B.S. Yandell. Bayesian mapping of genomewide interacting quantitative trait loci for ordinal traits. *Genetics*, 176(3):1855, 2007.

[81] N. Yi, S. Xu, V. George, and D.B. Allison. Mapping multiple quantitative trait loci for ordinal traits. *Behav Genet*, 34(1):3–15, 2004.

[82] H. Zhang, X. Wang, and Y. Ye. Detection of genes for ordinal traits in nuclear families and a unified approach for association studies. *Genetics*, 172(1):693–699, 2006.

[83] M. Zhang, R. Feng, X. Chen, B. Hu, and H. Zhang. LOT: a tool for linkage analysis of ordinal traits for pedigree data. *Bioinformatics*, 24(15):1737–1739, 2008.

[84] T. Zhao, Y. Liu, P. Wang, S. Li, D. Zhou, D. Zhang, Z. Chen, T. Wang, H. Xu, G. Feng, et al. Positive association between the PDLIM5 gene and bipolar disorder in the Chinese Han population. *Journal of Psychiatry & Neuroscience: JPN*, 34(3):199, 2009.

[85] L. Zhu, Z. Zhang, F. Feng, P. Schweitzer, J. Phavaphutanon, M. Vernier-Singer, E. Corey, S. Friedenberg, R. Mateescu, A. Williams, et al. Single nucleotide polymorphisms refine QTL intervals for hip joint laxity in dogs. *Animal Genetics*, 39(2):141–146, 2008.