Title

A Hidden Markov Modeling Approach for Admixture Mapping Based on Case-Control Haplotype Data

Authors

Zhang, Chun
Chen, Kun
Seldin, Michael F
et al.

Publication Date

2003-12-01

A Hidden Markov Modeling Approach for Admixture Mapping Based on Case-Control Haplotype Data

Chun Zhang, Kun Chen, Michael F. Seldin and Hongzhe Li

Departments of Statistics and Medicine and Rowe Program in Human Genetics

University of California, Davis, CA 95616

Running title: Hidden Markov Model for Admixture Mapping

Address for correspondence:

Hongzhe Li, Ph.D.

Rowe Program in Human Genetics

University of California Davis School of Medicine

Davis, CA 95616-8500, USA

Tel: (530) 754-9234; Fax: (530) 754-6015

E-mail: hli@ucdavis.edu

Summary

Admixture mapping is potentially a powerful method for mapping genes for complex human diseases, when the disease frequency due to a particular disease susceptible gene is different between founding populations of different ethnicity. The method tests for genetic linkage by detecting association of the allele ancestry with the disease. Since the markers used to define ancestral populations are not fully informative for the ancestry status, direct test of such association is not possible. In this paper, we develop a hidden Markov model (HMM) framework for estimating the unobserved ancestry haplotypes across a chromosomal region based on marker haplotypes. The HMM efficiently utilizes all the marker data to infer the latent ancestry states at the putative disease locus. In this modelling framework, we consider a likelihood based approach for detecting genetic linkage based on case-control data. We evaluate by simulations how several factors affect the power of admixture mapping, including sample size, ethnicity relative risk, marker density and the different admixture dynamics. Our simulation results indicate correct type 1 error rates of the proposed likelihood ratio test and great impact of marker density on the power. In addition, simulation results indicate that the methods work well for the admixed populations derived from both hybrid-isolation and continuous gene-flowing models.

KEYWORDS: Admixture Mapping, Hidden Markov Model, Linkage Disequilibrium, Haplotype.

1 Introduction

When there has been recent admixture between two populations with different prevalence for a certain disease and markers with large allele frequency differences between the two founding populations have been identified throughout the genome, it has been suggested that such an admixed population can be utilized for mapping susceptible genes for complex diseases (Chakroborty and Weiss, 1988; Stephens *et al.*, 1994; Brisoe *et al.*, 1994; McKeigue, 1998). Examples of the recently admixed populations include the African-American (AA) population and the Mexican-American population, in which the founding populations are European American (EA) and African (AF) or EA and Native American populations. Example of the complex diseases which show different risks in different populations include multiple sclerosis, primarily a Caucasian disease, especially affecting those of northern European ancestry (e.g. Scandanavians, English, Irish). Its prevalence is low in African blacks, Asians and other ethnic groups with little Caucasian admixture (Sadovnick 1994). The markers with large allele frequency differences between the two founding populations, called ancestry informative markers (AIMs), have been gradually identified throughout the genome (Shriver *et al.*, 1997; Smith *et al.*, 2001; Collins-Schramm *et al.* 2002). The mapping methods which use the particular genetic structure of the admixed population are called mapping by admixture linkage disequilibrium or admixture mapping. When admixture occurs between two populations, linkage disequilibrium (LD) is created between loci with large allele frequencies in the two founding populations. The LD between unlinked markers rapidly decays with successive generations while the LD between linked markers persists for many more generations. Efficiently utilizing such LD can provide evidence of genetic linkage.

Most of recent efforts have focused on studying the LD patterns in the admixture populations (Briscoe *et al.*, 1994; Stephens *et al.*, 1994; Pfaff *et al.*, 2001; Para *et al.*, 2001; Rybicki *et al.* 2002; Collins-Schramm *et al.* 2003)). Pfaff *et al.* (2001) observed by empirical simulation study that the LD pattern is highly dependent on admixture dynamics: populations that follow a continuous-gene-flow (CGF) history of admixture have greater LD over large chromosomal region than do populations that more closely fit a hybrid-isolation (HI) admixture model. While studying LD patterns in the admixed populations is important, statistical methods for actually testing for linkage based on admixed populations are less developed. Other efforts have been on developing methods for inference of population structure using multilocus genotype data (Pritchard *et al.* 2000: Fulash *et al.* 2003).

Admixture mapping can be viewed as examining the linkage between a susceptibility gene and marker alleles that distinguish between the ancestry of the founding populations. This association between the susceptibility gene and the marker alleles in the same chromosomal region is maintained for many generations but the linkage does not depend on LD per se between the susceptibility gene and any specific marker allele. Thus, instead of testing for allelic association, markers are typed for obtaining ancestry information and this information is used to examine linkage of the trait. Therefore, admixture mapping can be regarded as testing for linkage by using association or by using admixture LD. Based on this view of admixture mapping, McKeigue (1998) and McKeigue *et al* (2000) developed several interesting statistical tests for linkage by testing for unequal transmission of ancestry alleles conditional on parental admixture using only affected individuals. The difficulty is that the markers are not always fully informative for inferring the ancestry states. However, since AIMs are often linked, it is possible to utilize all the marker data to infer the ethnicity states for a given locus in a candidate region by using a hidden Markov model (HMM) (MacDonald and Zucchini, 1997) formulation. This idea was briefly discussed in McKeigue (1998). However, McKeigue (1998) did not provide a clear formulation of the HMM and the methods by McKeigue do not account for any uncertainty associated with the estimation of most probable ancestry states. In addition, McKeigue (1998) only considered case-only design by examining transmissions of alleles of different ancestries to probands.

Recently, Falusch *et al.* (2003) has developed an algorithm based on Morkov Chain Monte Carlo and HMM for inference of population structure using multicolus genotpe data. While the algorithm can be used to infer the population origin of chromosomal regions, it is not specifically designed for the purpose of admixture mapping. In this paper, we develop methods for testing genetic linkage using case-control data based on HMM by formally defining the latent states, the transition probabilities and the probabilities of observed data given latent states. We develop an iterative procedure utilizing the posterior decoding algorithm (Rabiner, 1989) to estimate the latent ancestry haplotype together with the allele frequencies conditioning on the latent states and other model parameters. We then consider case control design for admixture mapping and propose a likelihood ratio test for testing the linkage between the disease and a candidate locus in a test chromosomal region. Factors that affect the power of such test are fully investigated by using simulations.

The rest of the paper is organized as follows: we first define the HMM for the haplotype data and present methods for estimating the parameters such as the conditional allele fre-

quencies. We then present a likelihood ratio test for linkage between the test chromosomal region and the disease. Following the methods sections, we present results from simulation studies. Finally we give a brief discussion of the methods and results.

Methods

Hidden Markov Model for Ancestry Haplotypes

Consider a map of AIMs with known locations in a candidate chromosomal region and assume within this region an arbitrary location for the disease locus, 0. Given this location 0, the region is divided into two regions with L markers present to the left of the disease locus and R markers to the right for a total of $L+R$ markers, denoted by $1, \cdots, L+R$. Denote "d_{i+1}" as the distance between the adjacent marker loci i and $(i+1)$. In this paper, we assume that we know the marker haplotypes over these $L+R$ markers in this region for all individuals and all the AIMs are diallelic with two alleles, 1 and 2. Let $M_i = \{M_i^{(1)}, M_i^{(2)}\}$ denote the marker alleles at locus i of the two haplotypes. Then for a given haplotype, each of these $L+R$ marker loci may take one of two possible ancestral states: identity by descent (IBD) with an X founder chromosome, or Y founder chromosome. However, these ancestral states are not observable. Let X denote the status IBD with X founder chromosome, Y denote the status IBD with Y founder chromosome. Let $S_i = \{S_i^{(1)}, S_i^{(2)}\}$ denote the ancestral states at locus i on the two haplotypes, which takes values X or Y.

We assume that each observed haplotype is generated by a hidden Markov model, in which the unobserved ancestry haplotype serves as the latent path of ancestry states. We consider here the HI model, in which the admixture occurs in a single generation and is followed by recombination and genetic drift, with no further genetic contribution from either parental population. Let $\pi = Pr(X)$ be the prior probability that a randomly chosen allele in the admixed population is from the X founding population, which can also be interpreted as the contribution of population X to the admixture in the current admixed population. Given the chromosome's ancestral states at the locus i, we can calculate the probability of the two states at the locus j on the kth haplotype,

$$\Pr(S_j^{(k)} = X | S_i^{(k)} = X) = \Pr(\text{NR}) + [1 - \Pr(\text{NR})]\Pr(\text{MRR}=X),$$

where NR denotes that no recombination has occurred between the loci, MRR=X is used to indicate that the most recent recombination event occurred, at locus j, with a chromosome

IBD with the X population"; similarly,

$$\Pr(S_j^{(k)} = X | S_i^{(k)} = Y) = [1\text{-}\Pr(\text{NR})]\Pr(\text{MRR}=X)),$$

since a recombination event must have occurred between two loci of different ancestral states. We further assume that the probability that, at locus j, a chromosome IBD with the X population has remained constant over time, i.e., $Pr(MRR = X) = \pi$. This assumption holds for the HI model, but not the CGF model. Based on this general formula, the transition probabilities from $S_i \rightarrow S_{i+1}$ are defined as

$$\begin{aligned}
\tau_{XX}^{i+1} &= \exp(-\gamma d_{i+1}) + [1 - \exp(-\gamma d_{i+1})]\pi, \\
\tau_{XY}^{i+1} &= [1 - \exp(-\gamma d_{i+1})](1 - \pi), \\
\tau_{YX}^{i+1} &= [1 - \exp(-\gamma d_{i+1})]\pi, \\
\tau_{YY}^{i+1} &= \exp(-\gamma d_{i+1}) + [1 - \exp(-\gamma d_{i+1})](1 - \pi). \qquad (1)
\end{aligned}$$

Under the assumption of no interference, $exp(-\gamma d_{i+1})$ is the probability of no recombination events in generations since the admixture, where γ is the expected frequency of recombination events for each cM of a chromosomal region since admixture. Here 100γ can be regarded as the number of generations since admixture. Note that the parameters which are associated with these transition probabilities are π and γ. This formulation of the latent states and transition probabilities are similar to those used by Morris *et al* (2000) for fine-scale mapping of disease loci. It is important to note that these transition probabilities are derived by assuming a hybrid-isolation model. However, as shown in a later section, the methods developed by this model work well for other population models such as the continuous gene-flow model.

We next introduce the probability models to relate the observed data with the latent ancestry states. The observed data are the observed alleles, $M_i = \{M_i^{(1)}, M_i^{(2)}\}$, at each of the $R + L$ AIM loci on the two haplotypes and the disease status for each individual. Let $p_{iS}(a)$ be the probability of observing allele a at the locus i if the latent ancestry state is S, for $i = 1, \cdots R+L$, $S = X, Y$, and $a = 1, 2$. We call these $p_{iS}(a)$ the conditional marker allele frequencies. In practice, these conditional allele frequencies are unknown and are estimated from the data.

Finally, for the putative disease locus 0, we define the probability of phenotype, $Z = 1$ or 0 for case or control, given the latent ancestry states at the putative disease locus which is

given in the following logistic penetrance function,

$$Pr(Z=1|S_0^{(1)},S_0^{(2)}) = \frac{\exp\{\beta_0+\beta_1 I(S_0^{(1)}S_0^{(2)}=XY)+\beta_2 I(S_0^{(1)}S_0^{(2)}=YY)\}}{1+\exp\{\beta_0+\beta_1 I(S_0^{(1)}S_0^{(2)}=XY)+\beta_2 I(S_0^{(1)}S_0^{(2)}=YY)\}}, \quad (2)$$

where $I(.)$ is the indicator function. Under this model, the null hypothesis of no linkage of the putative disease locus 0 with the disease can be formulated as testing $H_0 : \beta_1 = \beta_2 = 0$. Other penetrance functions can also be assumed.

Genotypic Relative Risk, Ethnicity Relative Risk and Population Risk Ratio

In the previous penetrance model (2), the risk of disease is defined in term of the ancestry statuses of the two alleles at the putative disease locus 0. However, in practice, the genetic effect of a disease is often measured in term of genotypic relative risks and disease prevalence in the study population, in this case, the admixed population. It is important to establish the relationships between these different measurements of the disease risk.

Suppose that there are two variants at the disease locus, 0 and 1, where 1 is the high risk allele in both founding populations. Let p_X (or p_Y) be the frequency of allele 0 in the X (or Y) population. For each founding population, we assume a multiplicative model for the disease risk with genotypic relative risk ratio (GRR) of r_X and r_Y, then the prevalence of the disease in the founding population X and Y can be written as

$$K_X = f_{0X}[p_X^2 + 2p_X(1-p_X)r_X + (1-p_X)^2 r_X^2],$$

$$K_Y = f_{0Y}[p_Y^2 + 2p_Y(1-p_Y)r_Y + (1-p_Y)^2 r_Y^2],$$

where $f_{0X} = Pr(Z = 1|00, XX)$ is the baseline risk of disease in the X population and $f_{0Y} = Pr(Z = 1|00, YY)$ the baseline risk of disease in the Y population corresponding to the genotype 00. The population risk ratio RR_{XY} can be written as K_Y/K_X.

In the admixed population between X and Y founding populations, the frequency of allele 0 is $p_{XY} = \pi p_X + (1-\pi)p_Y$. If we assume a multiplicative model, we can define the ethnicity relative risk (ERR) as

$$ERR = \frac{Pr(Z=1|YY)}{Pr(Z=1|XY)} = \frac{Pr(Z=1|XY)}{Pr(Z=1|XX)},$$

which measures the increased risk when an allele at the disease locus in the admixed population is from the Y population. Then the prevalence of the disease in the admixed population is

$$K_p = Pr(Z=1|XX)Pr(XX) + 2Pr(Z=1|XY)Pr(XY) + Pr(Z=1|YY)Pr(YY)$$

$$
\begin{aligned}
&= Pr(Z=1|XX)\pi^2 + 2Pr(Z=1|XY)\pi(1-\pi) + Pr(Z=1|YY)(1-\pi)^2 \\
&= Pr(Z=1|XX)[\pi^2 + 2\pi(1-\pi)ERR + (1-\pi)^2 ERR^2],
\end{aligned}
$$

where $Pr(Z|XX)$ is risk of disease in the X population. In addition, it can be shown that the ethnicity relative risk ratio can be written in terms of the relative risks in the founding populations as

$$ERR = \sqrt{\frac{f_{0Y}}{f_{0X}}} \times \frac{p_Y + r_Y(1-p_Y)}{p_X + r_X(1-p_X)}.$$

If we further assume that the genotype-specific penetrance functions are the same in both founding populations and therefore $r_X = r_Y = r$, then the ethnicity relative risk can be further reduced to

$$ERR = \frac{p_Y + r(1-p_Y)}{p_X + r(1-p_X)}.$$

In this case, the genotypic relative risk ratio in the admixed population is also r, and the disease prevalence can be written as

$$K_p = f_0[p_{XY} + 2p_{XY}(1-p_{XY})r + (1-p_{XY})^2 r^2],$$

where f_0 is the probability of disease in non-carriers of the disease allele. If $r_X \neq r_Y$, then the the genotypic relative risk ratio in the admixed population depends on understanding the epistatic interaction in the admixed population.

Parameter Estimation and Likelihood Ratio Test for Linkage

Suppose we have n_A affected and n_U unaffected individuals in an admixed population from X and Y founding populations. We then have $2n_A$ haplotypes from cases and $2n_U$ haplotypes from controls. Let $MH_j^{(k)} = \{M_{j1}^{(k)}, \cdots, M_{jR+L}^{(k)}\}$ denote the kth marker haplotype for the jth individual, and similarly, let $SH_j^{(k)} = \{S_{j1}^{(k)}, \cdots, S_{jL}^{(k)}, S_{j0}^{(k)}, S_{jL+1}^{(k)}, \cdots, S_{jR+L}^{(k)}\}$ denote the kth ancestry haplotype for the jth individual, for $j = 1, \cdots, N$ and $k = 1, 2$. The likelihood of observing n_A disease cases and n_U controls (let $N = n_A + n_U$) and the $2N$ marker haplotypes can be written as

$$
\begin{aligned}
L &= \prod_{j=1}^{N}\prod_{k=1}^{2} Pr(Z_j, MH_j^{(k)}) \\
&= \prod_{j=1}^{N}\prod_{k=1}^{2}\{\sum_{SH_j^{(k)}} [Pr(Z_j, MH_j^{(k)}|SH_j^{(k)}) \times Pr(SH_j^{(k)})]\}
\end{aligned}
$$

$$= \prod_{j=1}^{N}\prod_{k=1}^{2}\left\{\sum_{SH_j^{(k)}}\left[Pr(Z_j|S_{j0}^{(k)})\times\prod_{i=1}^{R+L}(Pr(M_{ji}^{(k)}|S_{ji}^{(k)})\times Pr(SH_j^{(k)})\right]\right\}, \quad (3)$$

where $Pr(M_{ji}^{(k)}|S_{ji}^{(k)})$ depends on the conditional allele frequencies $p_{iS}(a)$, and $Pr(SH_j^{(k)})$ depends on the transition probabilities (1). In addition,

$$Pr(Z_j|S_{j0}^{k}) = \sum_{A\in\{X,Y\}} Pr(Z_j|S_{j0}^{(k)}, S_{j0}^{(|k-1|)})Pr(S_{j0}^{(|k-1|)} = A),$$

as defined in equation (2). Finally we denote the parameters associated with the HMM by $\theta = \{\theta_1, \theta_2\}$, where $\theta_1 = \{p_{iX}, i = 1, 2, \cdots, L+R\}$, and $\theta_2 = \{\pi, \gamma, \beta_0, \beta_1, \beta_2\}$.

We propose the following iterative procedure for estimating the parameters in θ and for calculating the likelihood function (3), which involves iteration between the following steps:

Step 1: For given parameters in θ_2, for each allele, we estimate the probabilities of ancestral states by the posterior decoding algorithm and then estimate the allele frequencies conditional on ancestry by counting the number of alleles with the given ancestry at each marker locus across all the individuals. For example, for locus i,

$$\begin{aligned} p_{iX}(a) &= \frac{\sum_{j=1}^{N}\sum_{k=1}^{2}E[I(S_{ji}^{(k)} = X, M_{ji}^{(k)} = a)]}{\sum_{j=1}^{N}\sum_{k=1}^{2}E[I(S_{ji}^{(k)} = X)]} \\ &= \frac{\sum_{j=1}^{N}\sum_{k=1}^{2}Pr(S_{ji}^{(k)} = X, M_{ji}^{(k)} = a)}{\sum_{j=1}^{N}\sum_{k=1}^{2}Pr(S_{ji}^{(k)} = X)}, \end{aligned}$$

where I is the indicator function.

Step 2: For given θ_2, we estimate the parameters by maximizing the likelihood function over parameter θ_1 by using a Quasi-Newton optimization algorithm. The forward algorithm (Rabiner, 1989)is used for calculating the likelihood function given the conditional allele frequencies.

We iterate between the Step 1 and Step 2 till convergence. The allele frequencies in the founding populations, if available, can be used as the initial values for the conditional allele frequencies. In order to address the issue of local maxima of the general Quasi-Newton optimization algorithm, we first use a grid search that roughly scans the whole range of the parameters to find the initial values for the parameters so that the procedure will converge

at an approximately most likely point. After the convergence of the algorithm, we can obtain the maximum likelihood function $maxL$.

Similarly, we can obtain the maximum likelihood function $maxL_0$ under $H_0 : \beta_1 = \beta_2 = 0$. We can then define the likelihood ratio test statistic as LR=2log $maxL/maxL_0$.

Results

In this section, we present simulation studies to evaluate the proposed methods and to study how different parameters such as sample size, ethnicity disease relative risk, number of AIMs and marker density affect the power of the proposed LR test. We consider simulated data from both the true models and from two different admixture population models, including the HI and CGF models.

In the following simulation studies, we used the actual allele frequencies of a total of 31 indel or SNP AIMs in a 60.6cM segment of chromosome 5. Table 1 gives the conditional allele frequencies for these markers and the corresponding δ values, which is defined as sum of absolute value of the allele-frequency difference between two populations divided by two.

Simulation 1: generating data based on the HMM

In order to exam how well the proposed methods estimate the parameters, especially the conditional allele frequencies, we first simulated the case control data based on our proposed HMM. We chose $\pi = 0.20$ and $\gamma = 0.10$, which implies that the X population contributes 20% to admixture and there has been about 10 generations since admixture. These parameters were chosen to approximate the AA population, where X is the EA founding population, and Y is the AF founding population. We assumed that marker CV8844618 (map position 137.1 cM) is the true disease locus with allele 2 being the high risk allele. The allele frequency of the high risk allele is 0.25 and 0.00 in the African and EA population, respectively. This marker was used for simulating the disease status data but was removed for all subsequent analyses. In addition, we assume the same genotypic relative risks in both population and therefore the disease risk is higher in the Y (African) population. To generate cases and controls and the corresponding marker haplotypes, we first generate the ancestral state sequence as a Markov chain with marginal distribution $P(S_i = X) = \pi = 0.20$ and transition probability defined by equations (1) with $\gamma = 0.10$. We then generated the marker data according to the allele

frequencies conditional on ancestral states given as in Table 1. We then simulate the disease status Z according to the ancestry-specific penetrance functions $P(Z = 1|S_0^{(1)}, S_0^{(2)})$. We assume $K_p = 0.0072$ and $ERR = 1.75$ which corresponds to $Pr(Z = 1|XX) = 0.0028$.

To examine how well our proposed methods estimate the conditional allele frequencies, we first simulated 200 cases and controls with the haplotypes over all 29 markers for an ERR=1.75. Figure 1 (a) shows the log odds ratio of disease for African ancestry at each of the 29 markers for 50 replications using the true simulated ancestry statuses for each individual. A clear peak can be observed around the disease locus. As a comparison, Figure 1 (b) shows the log odds ratio of disease for marker allele 2 for each of the 34 markers around the true disease locus. Although there are some peaks, the signal is erratic, and important location information is obscured by the properties of the markers. We also noted large variations among the 50 replications and that many of the markers close to the disease locus did not show any significant association with the disease. These results indicated that comparing the marker allele difference between cases and controls can often result in false association or no association. This demonstrates the importance of comparing the latent ancestry states between cases and controls in admixture mapping. Figure 1 (c) presents the plots of the estimated and true conditional allele frequencies for all markers and all 50 replications. Clearly, our methods provide a reasonable estimate of these conditional allele frequencies.

To evaluate the behavior of the LR test statistic, Figure 1 (d) shows the likelihood ratio statistic calculated for each location in the region for 50 replications. A clear peak of the LR statistic at the disease locus and higher LR values around the true disease locus are observed, indicating that the proposed LR test can indeed be applied to localize the disease region. Compared to the marker based test of association (Figure 1 (b)), we obtained a much clearer picture and the LR statistic decays almost monotonically. We performed similar comparisons for an unlinked region with the same set of markers. No large differences are observed between the cases and controls in ancestry estimates in the unlinked region and the LR statistics are small (results not shown).

Simulation 2: generating data based on population models

To further evaluate the performance of the proposed test, we performed simulation studies by generating data based on two distinct patterns of admixture dynamics, the HI model and

the CGF model. Our CGF model assumes that admixture occurs at a steady but reduced rate in every generation, such that the cumulative amount of admixture is equal to that in the HI model. This allows comparison of the two models. To simulate the data, we assume a constant population size of 10,000 and linkage equilibrium between markers in both founding populations. We simulated the data based on a forward branching process. Based on the population-specific allele frequencies we first obtained the haplotype frequencies in the founding populations. Under the HI model, the X population (European) and the Y population (African) are initially admixed with proportions of π and $1-\pi$. This is then followed by random mating for a total of t generations to generate the current admixed population (African-American population). Under the CGF model, the X population and the Y population are initially admixed with proportions of $\pi^{1/t}$ and $1-\pi^{1/t}$. For each of the next generations, a proportion of $\pi^{1/t}$ from offspring population and a proportion of $1-\pi^{1/t}$ from Y population enter next generation and randomly mate with each other within the population. The recombination fraction is calculated based on the map distance.

The same disease model as in Simulation 1 was used to simulate cases and controls based on the ethnicity-specific penetrance functions for each individual in the population. We also assumed that $K_P = 0.0072$ and $ERR = 1.75$ which corresponds to $Pr(Z = 1|XX) = 0.0028$. At the end of t generations, we collect 200 case and 200 control from the final admixture population according to ethnicity relative risk.

Figures 2 (a) and (c) show the log odds ratio values of 50 replications for the disease associated allele for each of the 29 markers around the true disease locus. Similar to the previous results, the analysis of LD between markers and disease does not show a clear identification of true simulated disease gene. It shows that some of the markers close to the true disease locus do not show any significant association with the disease. This is due to the fact that these markers are all in linkage equilibrium with the disease locus in the original founding populations. In contrast, Figures 2 (b) and (d) show the likelihood ratio statistics of 50 replications for location in the test chromosomal region. We observed that the statistics obtained their highest values at the loci close to the true disease locus and became smaller for more distant markers. It is also interesting to note that the region around the true disease locus that shows statistical significance is larger for the CGF model than for the HI model. These two plots indicate that LR statistic for testing linkage indeed behaves as we would expect.

We next performed simulation studies to evaluate the type 1 error rate and the power of the proposed LR test for $K_p = 0.0072$ and $ERR = 1, 1.375$ and 1.75, which correspond to GRR of

1, 2.5 and 4.0. The corresponding ethnicity-specific penetrance function is $Pr(Z = 1|XX) = 0.0072, 0.0043$ and 0.0028, respectively. We assume that the test locus is the true disease locus. Table 2 shows the power of the LR test for the true disease location for various combinations of relative risk, marker number and density and sample size based on 200 simulations (1000 simulations when the ERR is 1). First, we observed that the simulated error rates are close to the true type 1 error rates of 0.01 or 0.05, indicating that the proposed LR test has correct type 1 error rate. Second, we observed that the marker density in the test region can greatly affect the power to detect linkage by admixture mapping. Dense markers result in higher power. However, for markers with the same density, doubling the width of the testing region only marginally increase the power. Third, for the same genetic model, data generated from the CGF model shows similar power that generated from the HI model.

We lastly considered how mixture of individuals with different admixture proportions and different times since admixture affect the power of our proposed test. In particular, we considered the scenario that both cases and controls include equal mixture of four different populations with different π and γ parameters (see legend of Table 3). Note that the overall proportion of admixture is still 20%. For this set of simulations, the panel of 29 markers were simulated. Table 3 shows the power for different population models, different relative risks and different sample sizes for an α level of 0.01 and 0.05. First, we noticed that the test still has correct type 1 error rates. Second, we observed that the power is only slightly lower than in the scenario when all individuals were simulated under $\pi = 0.20, \gamma = 0.15$ (see Table 2). These results indicate that even when some of the assumptions are violated in real data, the proposed test still have correct type 1 error rate and only decreases the the power marginally.

Simulation 3: generating data based on real genotype data

In order to further evaluate the applicability of our methods to admixture mapping, cases and controls were simulated from the 268 African American subjects with known genotypes on 31 indel or SNP markers in a 60.6cM segment of chromosome 5. The median distance between adjacent markers was 1.2 cM and 1.5 Mb. The simulations were performed using the typing results of two diallelic markers to model susceptibility genes originating in the two different parental populations (AF and EA). These markers, CV8844618 (AF allele 1, 0.25; EA allele 1, 0.00; map position, 137.1 cM) and MID-990 (AF allele 1, 0.04; EA allele1, 0.31; map position; 131.5 cM), were used to model AF (AF model) and EA (EA model) susceptibility

genes, respectively. Setting genotypic relative risk ratios of 2.5 and 1.0 in the AA population, 500 cases and 500 controls were selected from the 268 AA subjects that had been genotyped. The disease status was simulated based on the individual's genotype at the chosen disease locus assuming the probability of disease for non-carrier to be 0.000625 and a multiplicative model. Haplotypes for the cases and controls for each GRR and model were then separately estimated using PHASE (Stephens *et al.*, 2001). The genotyping data that was used for haplotype estimation or subsequent HMM analysis did not include either of the markers used for the models. The haplotypes were then analyzed using our proposed HMM methods. We randomly generated 50 data sets for each model. For both EA and AF susceptibility models there were peaks in the respective likelihood ratio statistics around the chromosomal location of the modeled markers (see Figure 3 plots (a) and (c)). Strong evidence linkage for both models were observed. The peaks were close to the modeled loci.

As a comparison, we also simulated 500 cases and 500 controls that were selected from the 268 AA subjects but with GRR=1.0 (null model) for both the AF and EA models. We generated 50 replications. None of the markers in this region showed any significance (see Figure 3 (b) and (c)), further indicating the correct type 1 error rates of the proposed LR test.

Discussion

We have presented a hidden Markov modelling framework for admixture mapping that utilizes ancestry informative markers to infer the ancestry states at the putative disease locus. Within this framework, we have developed a likelihood ratio based approach for testing genetic linkage between a candidate chromosomal region and the disease. We evaluated the proposed methods and tested these by simulation studies. Results indicate a correct type 1 error rate for the proposed test and the effects of marker density on the power can be quite large. These simulation results have practical implications when designing a study for admixture mapping. In addition, the simulation results indicate that our methods are not restricted to any particular admixture dynamic pattern and are applicable to any recently admixed population as long as there is a different disease prevalence in two founding populations and set of AIMs are identified. Our simulation results also indicate that populations with a CGF admixture dynamic pattern may result in modestly better power than that with HI admixture dynamic pattern.

The methods developed in this paper can be applied to both testing a candidate chromosomal region and genome-wide scanning. Our simulations were conducted for testing a chromosomal region. For genome-wide scans, we can calculate the likelihood ratio statistics or the corresponding Lod scores along evenly-spaced locations on a given chromosome. The cutoff value of the LR statistic for genome-wide significance developed for genome-wide family-based linkage analysis (e.g., Lander and Kruglyak, 1995; Morton, 1998) can directly applied to such genome-wide admixture mapping.

In developing the proposed methods for admixture mapping, we made several key assumptions. First, we assume that the marker haplotype data are available for the test chromosomal region for all the sampled individuals. In practice, the haplotype may not be known, in which case we can estimate them using available algorithms such as the EM algorithm (Long *et al.* 1995) or the Bayesian algorithm PHASE (Stephens *et al.* 2002). We can choose the most likely haplotypes and apply our proposed methods directly, as we did for our simulations. Our simulated data demonstrate that the proposed methods work well by using the PHASE-based estimation of the marker haplotypes when the uncertainty of the haplotype estimation is low. However, if the marker haplotypes cannot be estimated with high certainty, the test procedure needs to account for such uncertainty. We are currently investigating an unified approach to account for the uncertainly of the marker haplotype estimates in our likelihood formation. Second, we assume that the γ parameter, which is a parameter for historical recombination since admixture is the same for all the individuals. However, for populations with a continuous-gene-flow history of admixture, the value of γ could be different. Although this assumption might be violated, it should not affect the type 1 error of the proposed test, as indicated in our simulations. Our limited simulations (See Tables 2 and 3) also indicated that the power of the proposed test is only marginally affected by violation of constant γ assumption. One possibility to relax this assumption is to assume that γ is a random variable following some distribution. Lastly, the methods are developed for the admixed populations derived from two founding populations. It is however possible to extend the HMM to mixture of multiple populations by expanding the latent states and the transition probabilities (Falush *et al.*, 2003).

The power of admixture mapping is a complicated function of many factors. As our simulations implied, the number of markers, marker densities and informativeness of markers have impact on how well we can infer the chromosomal ancestry at the putative disease locus and therefore on the power of the proposed test. The disease characteristics such as disease

allele frequencies and penetrance functions in the founding populations will also determine the power of the proposed test. In addition, the population admixture dynamics and admixture proportions will have affect the power. Our simulations only examined the effects on test power for some of these factors. We are currently examining how these factors affect the power of admixture mapping by both analytical calculations and large-scale simulations.

Finally, in this paper we considered the case control design for admixture mapping. An interesting alternative is case-only design. For such design, our goal is to identify the chromosomal regions which have different ancestry haplotype makeups than the other regions throughout the genome. This approach is in spirit similar to the genomic control for detecting excess-haplotype sharing (Devlin *et al.* 2000). It should however be noted that this genome-wide case only design is very different from that of McKeigue (1998), in which unequal transmissions of alleles of different ancestry origins are tested. We are currently investigating such approach and comparing the power of case-only design and case-control design for admixture mapping. It is also possible to combine both case only test and case-control test in order to increase the power of detecting genetic association by admixture mapping.

Acknowledgements

This research was supported by NIH grant R01-ES09911 and U01-DK57249.

References

Briscoe D, Stephens JC, and O'Brien SJ (1994): Linkage disequilibrium in admixed populations: applications in gene mapping. *The Journal of Heredity*, 85(1):59-63.

Collins-Schramm HE, Phillips CM, Operario DJ, Lee JS, Weber JL, Hanson RA, Knowler WC, CoopermR, Li H, and Seldin MF (2002): Ethnic-difference markers for use in mapping by admixture linkage disequilibrium. *American Journal of Human Genetics*, 70:737-750.

Collins-Schramm HE, Chima B, Operario DJ, Criswell L, and Seldin MF (2002): Markers informative for ancestry demontrate consistent megabase-length linkage disequilibrium in the African American population. *Human Genetics*, 113: 211-219.

Devlin B, Roeder K, Wasserman L (2000): Genomic control for association studies: a semiparametric test to detect excess-haplotype sharing. *Biostatistics*, 1(4),369-387.

Falush D, Stephens M, Pritchard JK (2003): Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164, 1567-1587.

Lander ES, Kruglyak L (1995): Genetic dissection of complex traits: guidelines for interpreting and reporting linkage rsults. *Nature Genetics*, 11, 241-247.

Long JC, Williams RC, Urbanek M (1995): An EM algorithm and testing strategy for multiple locus haplotypes. *American Journal of Human Genetics*, 56:799-810.

MacDonald IL and Zucchini W(1997): *Hidden markov and other models for discrete-valued time series*, Chapman and Hall, 1997.

McKeigue PM (1998): Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. *American Journal of Human Genetics*, 63:241-251.

McKeigue PM, Carpenter JR, Parra EJ and Shriver MD (2000): Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. *Annnals of Human Genetics*, 64:171-186.

Morton NE(1998): Significance levels in complex inheritance. *American Journal of Human Genetics*, 7:217-318.

Morris AP, Whittaker JC, Balding DJ (2000): Bayesian Fine-scale mapping of disease loci, by hidden markov models. *American Journal of Human Genetics*, 67:155-169.

Parra EJ, Kittles RA, Argyropoulos G, Pfaff CL, Hiester K, Bonilla C, Sylvester N, Parrish-Gause D, Garvey WT, Jin L, McKeigue PM, Kamboh MI, Ferrell RE, Pollitzer WS, Shriver MD (2001): Ancestral proportions and admxiture dynamics in geographically defined African Amricans living in South Coralina. *American Journal of Physiological Anthropology*, 114: 18-29.

Pfaff CL, Parra EJ, Bonilla C, Hiester K, McKeigue PM, Kamboh MI, Hutchinson RG, Ferrell RE, Boerwinkle E, and Shriver MD (2001): Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. *American Journal of Human Genetics*, 68:198-207.

Prichard JK, Stephens M, Donnelly (2000): Inference of population structure using multilocus genotype data. *Genetics*, 155: 945-959.

Rabiner LR (1989): A Tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of The IEEE*, 77(2):257:285.

Rybicki BA, Iyengar SK, Harris T, Elston RC, Sheffer R, Chen KM, Major M, Maliarik MJ, Iannuzzi (2002): The distribution of long range admxiture linkage disequilibrium in an African-American population. *Human Heredity*, 53: 187-196.

Sadovnick AD (1994): Genetic epidemiology of multiple sclerosis: A survey. *Annals of Neurology*,36(Suppl 2) :S194-203

Stephens JC, Briscoe D, O'Brien SJ (1994): Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. *American Journal of Human Genetics*, 55:809-824.

Stephens M, Smith NJ, Donnelly P(2001): A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, 68:978-989.

Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, R Deka, Ferrell RE (1997): Ethnic-affiliation estimation by use of population-specific DNA markers. *American Journal of Human Genetics*, 60:957-964.

Smith MW, Lautenberger JA, Doo SH, Chretien J, Shrestha S, Gilbert DA, O'Brien SJ (2001): Markers for mapping by admixture linkage disequilbrium in African American and Hispanic populations. *American Journal of Human Genetics*, 69: 1080-1094.

Figure Legends

Figure 1 Simulation results based on case-control data generated from the proposed HMM model for 50 replications of 200 case/control pairs for $\pi = 0.2, \gamma = 0.10, ERR = 1.75$. Under this model, the probability of getting disease in the low risk populaton is 0.0028. (a) log odds ratio of disease for true African ethnicity; (b) log odds ratio of disease for true marker genotypes; (c) estimated and true conditional allele frequencies; (d) 2LR based on the full likelihood ratio tests for 50 replications. For plots (a), (b) and (d), the vertical line indicates the true disease locus, and the solid curve is the average across 50 replications.

Figure 2: Simulation results based on hybrid (HI) (plots (a) and (b)) and continuous gene-flowing model (CGF) (plots (c) and (d)) for 50 replications of 200 case/control pairs for $\pi = 0.2, \gamma = 0.10, ERR = 1.75$. Under this model, the probability of getting disease in the low risk populaton is 0.0028. (a) and (c): log odds ratio of disease for true marker genotypes shown for 50 replications of 200 case control pairs; (b) and (d): 2LR values for 50 replications of 200 case control pairs based on the full likelihood ratio tests for each markers in the 54 cM region. The vertical line indicates the true disease locus, the horizontal line indicate the significance value for α level of 0.01 and the solid curve is the average across 50 replications.

Figure 3: Simulation results based on re-sampling real genotype data. (a) and (b): LR statistics for the AF model for GRR=2.5 and GRR=1 based on 500 cases and 500 controls; (c) and (d): LR statistics for the EA model for GRR=2.5 and GRR=1 based on 500 cases and 500 controls. For each plot, the vertical line indicates the true disease locus, the horizontal line corresponds to critical value of 0.01, and 50 curves represent 50 replications with the solid curve being the averge of the curves.

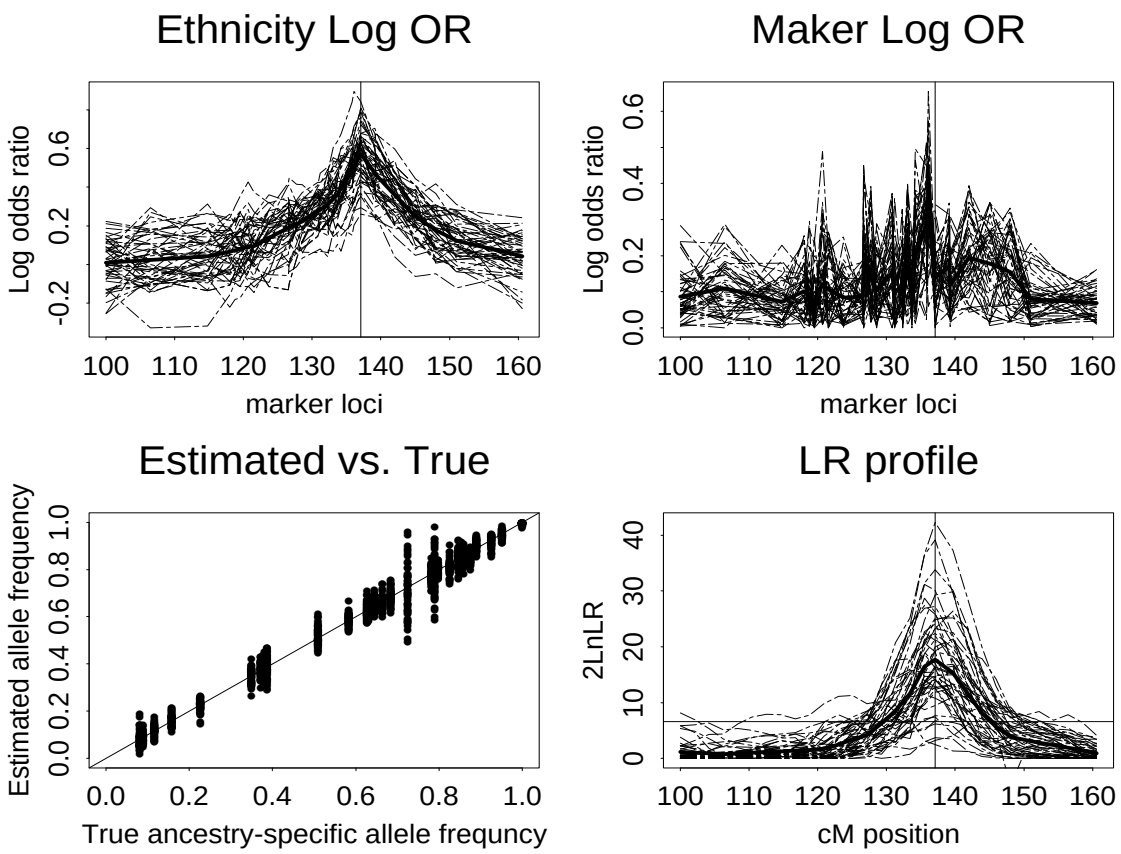
Ethnicity Log OR
Log odds ratio
-0.2
0.2
0.6
100 110 120 130 140 150 160
marker loci
Maker Log OR
Log odds ratio
0.0
0.2
0.4
0.6
100 110 120 130 140 150 160
marker loci
Estimated vs. True
Estimated allele frequency
0.0 0.2 0.4 0.6 0.8 1.0
0.0 0.2 0.4 0.6 0.8 1.0
True ancestry-specific allele frequncy
LR profile
2LnLR
0 10 20 30 40
100 110 120 130 140 150 160
cM position

Figure 1:

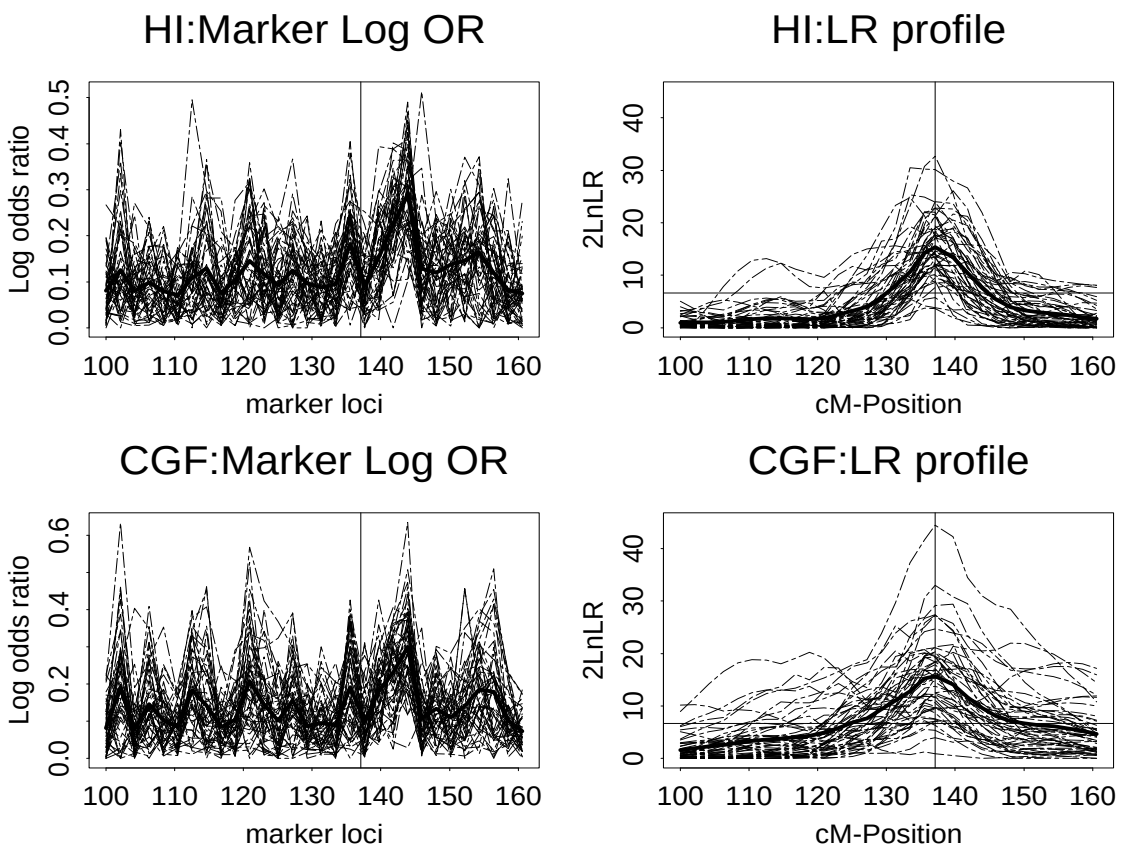
HI:Marker Log OR
Log odds ratio
0.0 0.1 0.2 0.3 0.4 0.5
100 110 120 130 140 150 160
marker loci
HI:LR profile
2LnLR
0 10 20 30 40
100 110 120 130 140 150 160
cM-Position
CGF:Marker Log OR
Log odds ratio
0.0 0.2 0.4 0.6
100 110 120 130 140 150 160
marker loci
CGF:LR profile
2LnLR
0 10 20 30 40
100 110 120 130 140 150 160
cM-Position

Figure 2:

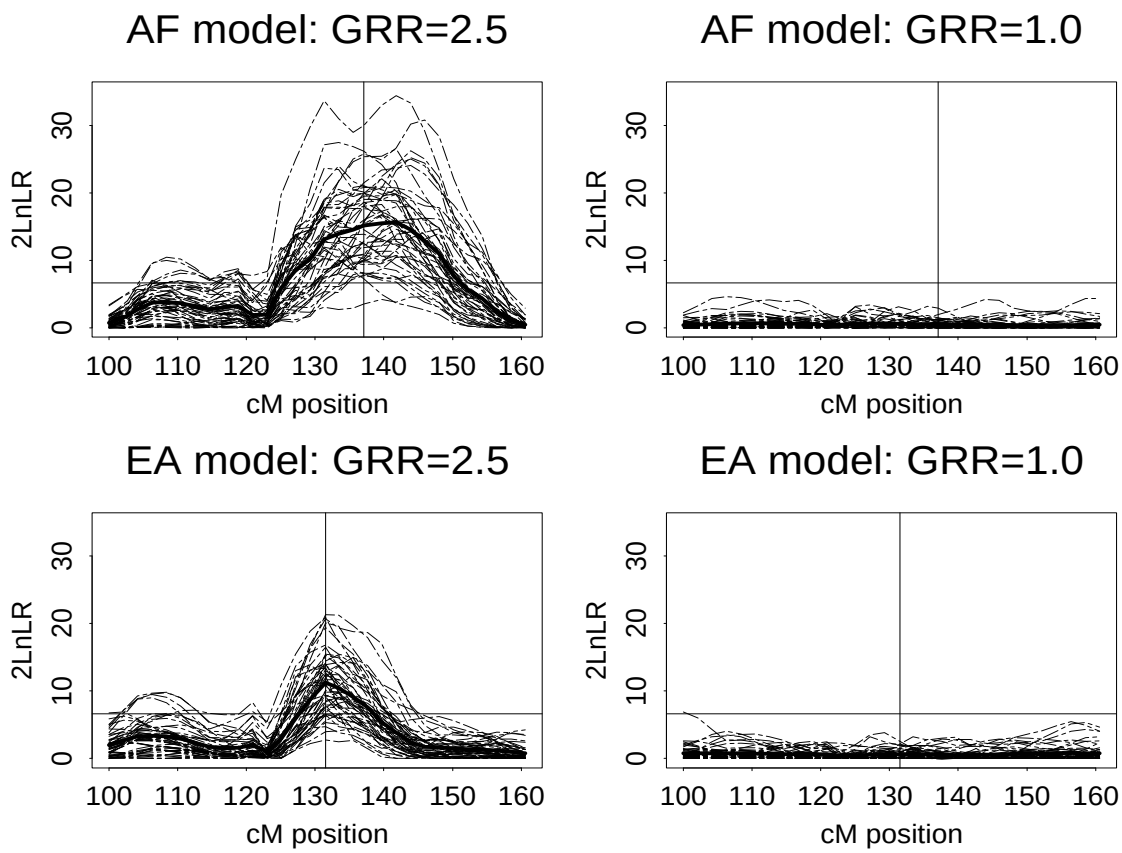
AF model: GRR=2.5
AF model: GRR=1.0
EA model: GRR=2.5
EA model: GRR=1.0
2LnLR
0
10
20
30
100
110
120
130
140
150
160
cM position

Figure 3:

Table 1: Conditional allele frequencies for 31 AIMS markers identified in a 60.6 cM region on chromosome 5, where X and Y denote European and African populations and 1 and 2 represent allele 1 and allele 2. These markers are used for simulations, where markers marked with * are used as the disease loci in the simulations.

marker	cM	$Pr(1\|X)$	$Pr(1\|Y)$	$Pr(2\|X)$	$Pr(2\|Y)$	δ
MID-1683	100.0	0.790	0.120	0.210	0.880	0.67
MID-737	106.5	0.387	0.000	0.613	1.000	0.387
CV3163022	114.8	0.846	0.231	0.154	0.769	0.615
MID-1272	118.0	0.89	0.061	0.11	0.939	0.829
MID-883	118.8	0.226	0.873	0.774	0.127	0.647
MID-1848	119.5	0.158	0.646	0.842	0.354	0.488
MID-879	120.7	0.583	0.017	0.417	0.983	0.566
CV118646	121.5	0.684	0.054	0.316	0.946	0.63
TSC0232289	123.7	1.000	0.697	0.000	0.303	0.303
MID-739	126.5	0.645	0.191	0.355	0.809	0.455
MID-1191	126.7	0.510	0.03	0.490	0.970	0.48
TSC0569173	127.7	0.627	0.057	0.373	0.943	0.57
MID-1937	128.6	0.664	0.144	0.336	0.856	0.52
CV2060865	130.8	0.630	0.100	0.370	0.900	0.53
CV159565557	131.2	0.798	0.382	0.202	0.618	0.416
MID-990(*)	131.5	0.690	0.960	0.310	0.040	0.268
TSC0237153	132.3	0.951	0.404	0.049	0.596	0.547
CV3167763	132.4	0.09	0.536	0.91	0.464	0.447
CV11532818	133.1	0.859	0.104	0.141	0.896	0.755
MID-1030	133.7	0.782	0.357	0.218	0.643	0.425
CV1561700	134.2	0.372	0.901	0.628	0.099	0.529
MID-768	135.8	0.826	0.102	0.174	0.898	0.724
MID-1102	136.1	0.876	0.066	0.124	0.934	0.811
CV8844618(*)	137.1	1.000	0.750	0.000	0.250	0.250
MID-719	139.1	0.854	0.313	0.146	0.687	0.541
CV8958376	139.3	0.116	0.593	0.884	0.407	0.477
CV2083528	142.0	0.35	0.906	0.65	0.094	0.557
CV1989090	145.0	0.926	0.071	0.074	0.929	0.854
CV1675518	148.0	0.381	0.972	0.619	0.028	0.59
CV3220692	151.0	0.081	0.731	0.919	0.269	0.65
MID-1348	160.6	0.724	0.211	0.276	0.789	0.513

Table 2: Power table for $\pi = 0.2, \gamma = 0.10$ ($\alpha = 0.01$ (0.05)) based on 1000 replications when RR=1 and 200 replications otherwise. *: all 29 markers in the 60.6 cM region are used; **: 18 markers (nine from each side of the disease locus) in the 21.5 cM region are used; ***: 15 markers in the 60.6 cM region are used.

			Population model(case-control pairs)			
Marker	ERR	GRR	HI(300)	CGF(300)	HI(500)	CGF(500)
29 marker	1	1	0.01(0.04)	0.01(0.06)	0.01(0.04)	0.01(0.05)
60.6cM*	1.375	2.5	0.26(0.46)	0.24(0.50)	0.39(0.59)	0.41(0.61)
	1.75	4.0	0.69(0.88)	0.68(0.88)	0.92(0.98)	0.90(0.98)
18 marker	1	1	0.01(0.04)	0.01(0.05)	0.01(0.04)	0.01(0.05)
21.5cM**	1.375	2.5	0.26(0.46)	0.23(0.51)	0.38(0.60)	0.42(0.62)
	1.75	4.0	0.68(0.88)	0.68(0.88)	0.92(0.98)	0.90(0.97)
15 marker	1	1	0.01(0.05)	0.01(0.06)	0.01(0.04)	0.01(0.05)
60.6cM***	1.375	2.5	0.140(0.32)	0.20(0.39)	0.25(0.52)	0.28(0.53)
	1.75	4.0	0.52(0.74)	0.53(0.77)	0.79(0.90)	0.80(0.94)

Table 3: Power table for data of mixed 4 groups ($\alpha = 0.01(0.05)$):$\pi = 0.25, \gamma = 0.08; \pi = 0.15, \gamma = 0.10; \pi = 0.10, \gamma = 0.12; \pi = 0.30, \gamma = 0.15$ based on 1000 replications when ERR=1 and 200 replications otherwise. 29 markers on 60.6cM chromosomal region were used.

		Population model(case-control pairs)			
ERR	GRR	HI(300)	CGF(300)	HI(500)	CGF(500)
1	1	0.01(0.04)	0.01(0.05)	0.01(0.05)	0.01(0.02)
1.375	2.5	0.17(0.42)	0.19(0.39)	0.37(0.61)	0.37(0.61)
1.75	4.0	0.62(0.84)	0.67(0.84)	0.87(0.95)	0.90(0.98)