

# Linkage Disequilibrium Mapping With Genotype Data

Shuanglin Zhang and Hongyu Zhao\*

*Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, Connecticut*

Linkage disequilibrium mapping has proven a powerful tool for locating disease genes. Although all existing linkage disequilibrium mapping methods implicitly assume that individual haplotypes can be inferred, only genotypes are directly observable in practice, and haplotypes cannot always be uniquely resolved based on genotype data. In this article, we propose a likelihood-based linkage disequilibrium mapping approach to analyzing multilocus genotype data arising from case-control studies. Results from extensive simulation studies suggest that this approach may be a useful tool to fine map disease genes using case-control data. *Genet. Epidemiol.* 22:66–77, 2002. © 2002 Wiley-Liss, Inc.

**Key words:** linkage disequilibrium; genotype; haplotype; coalescent; complex traits

## INTRODUCTION

Motivated by the success and potential of linkage disequilibrium mapping (LDM), many statistical methods have been proposed for LDM in recent years. For detailed discussion and comparisons of various LDM methods, see recent reviews by Clayton [2000] and Lazeroni [2001]. Existing LDM methods rely on the availability of a set of individual “high-risk” chromosomes and a set of individual “normal” chromosomes. In practice, the genetic compositions of “high-risk” chromosomes, i.e., haplotypes across a set of genetic markers, are inferred using the genotypes of the affected individuals and the genotypes of their relatives. The haplotypes on the “normal” chromosomes can be constructed either using chromosomes not associated

Contract grant sponsor: National Institutes of Health; Contract grant number: GM59507.

\*Correspondence to: Hongyu Zhao, Ph.D., Department of Epidemiology and Public Health, 60 College Street, Yale University School of Medicine, New Haven, CT 06520-8034. E-mail: hongyu.zhao@yale.edu

Received for publication 20 November 2000; revision accepted 30 May 2001

© 2002 Wiley-Liss, Inc.

with the disease in the same set of pedigrees or unrelated normal chromosomes in the same population. The need for obtaining unambiguous haplotypes from individuals involved in the study poses potential problems in practice because the exact haplotypes for each individual cannot always be determined, especially when there is not enough information from the relatives to allow the reconstruction of haplotypes. One extreme case is the case-control study, where we only have genotypes from a set of affected individuals and genotypes from a set of normal individuals. Resolving haplotypes will become more important as well as more difficult when more markers are identified and genotyped in a candidate region.

In this article, we develop a statistical approach that is directly applicable to genotype data and does not require haplotypes be uniquely resolved. Our method is based on the decay of haplotype sharing (DHS) method for haplotype data developed by McPeck and Strahs [1999]. Among various existing LDM methods, the DHS method can be generally categorized as a haplotype method, because it analyzes haplotype data instead of treating each marker separately. Because haplotype methods can utilize the dependence across loci within an individual haplotype, which is expected to be high, such methods are likely more powerful than pairwise methods that examine each marker separately. In the following discussion, we first review the DHS approach for haplotype data, and then describe the model for genotype data and the statistical methods to estimate the location of the genetic variant and other parameters in the model. Throughout this article, we assume that strong linkage between the disease and the chromosomal region under study has been established and the locations of the genetic markers are precisely known.

## METHODS

### Haplotype Data

The DHS method [McPeck and Strahs, 1999] models the dependence across loci within a haplotype by considering LD as the extent of a region of shared haplotype around a variant. We first assume that the genetic variant that causes the disease was introduced  $\tau$  generations ago ( $\tau$  maybe unknown), and the chromosomes sharing the variant in the current sample are the descendents of this ancestral chromosome. Suppose that a sample of haplotypes with the variant and a sample of haplotypes without the variant are collected. Define the position of the variant to be locus 0, with  $1, 2, 3, \dots, l_{re}$  at increasing distance on one side of the variant and loci  $-1, -2, -3, \dots, -l_{le}$  at increasing distance on the other side of the variant. Let  $d_{ij}$  denote the genetic distance between loci  $i$  and  $j$ , and  $\mu$  denote the mutation rate. If a sampled haplotype  $h_{obs}$  is the  $\tau$ th generation descendent of the ancestral haplotype  $h_{anc}$ , the likelihood contribution of  $h_{obs}$  is [equation 4 in McPeck and Strahs, 1999]:

$$\begin{aligned} \Pr(h_{obs} | h_{anc}, \tau, \mu) &= \sum_{i=0}^{l_{re}} \sum_{j=0}^{l_{le}} g(\tau, -j, i) \prod_{k=-j}^i m[k, \tau, h_{anc}(k), h_{obs}(k)] \\ &\times P_{null}[h_{obs}(i+1), h_{obs}(i+2), \dots, h_{obs}(l_{re})] \\ &\times P_{null}[h_{obs}(-j-1), h_{obs}(-j-2), \dots, h(-l_{le})]. \end{aligned} \quad (1)$$

In the above likelihood function,  $g(\tau, -j, i) = e^{-\tau d_{-j,i}}(1 - e^{-\tau d_{-j-1,-j}})(1 - e^{-\tau d_{i,i+1}})$  is the probability that, during  $\tau$  generations, there are no recombinations between loci  $-j$  and  $i$ , at least one recombination between loci  $-j - 1$  and  $-j$ , and at least one recombination between loci  $i$  and  $i + 1$ . If either locus  $i$  or locus  $-j$  is on the edge of the observed haplotype, the term  $1 - e^{-\tau d_{i,i+1}}$  or  $1 - e^{-\tau d_{-j-1,-j}}$  is not in the expression. The factor  $m[k, \tau, h_{anc}(k), h_{obs}(k)]$  is the conditional probability that a  $\tau$ th generation descendant has allele  $h_{obs}(k)$  at marker  $k$  given that the ancestral haplotype has allele  $h_{anc}(k)$  and that there are no recombinations between the genetic variant and locus  $k$  during the  $\tau$  generations, and the detailed formula of  $m[k, \tau, h_{anc}(k), h_{obs}(k)]$  can be found in McPeck and Strahs [1999, page 862]. The factor  $P_{null}[h_{obs}(i + 1), h_{obs}(i + 2), \dots, h_{obs}(l_{re})]$  is the joint probability that the alleles  $h_{obs}(i + 1), h_{obs}(i + 2), \dots, h_{obs}(l_{re})$  occur in a nonancestral haplotype, and the factor  $P_{null}[h_{obs}(-j - 1), h_{obs}(-j - 2), \dots, h(-l_e)]$  is the joint probability that the alleles  $h_{obs}(-j - 1), h_{obs}(-j - 2), \dots, h(-l_e)$  occur in a nonancestral haplotype.

The DHS method proposed by McPeck and Strahs [1999] allows for mutation by assuming equal mutation rates from one marker allele to all other marker alleles. Zhang and Zhao [2000] generalized the DHS method to use the step-wise mutation model to describe the mutation process for microsatellite markers. If the observed haplotypes can be considered independent, the overall likelihood is simply the product of the likelihoods for individual haplotypes. However, the independence assumption is not valid in general, and the conditional coalescent models can be used to take such dependence into account in the statistical inference [McPeck and Strahs, 1999]. Zhang and Zhao [2000] considered a general conditional coalescent model to incorporate variable population size. To allow for the possibility that the “high-risk” chromosomes are the descendants of two or more ancestral haplotypes, a heterogeneity parameter  $p$  can be introduced to represent the proportion of the variant haplotypes in the population that are not descended from the ancestral haplotype [McPeck and Strahs, 1999]. With this heterogeneity parameter  $p$ , the likelihood contribution of an observed haplotype  $h_{obs}$  is

$$L(h_{obs} | h_{anc}, \tau, \mu, p) = (1 - p) \Pr(h_{obs} | h_{anc}, \tau, \mu) + p P_{null}(h_{obs}), \quad (2)$$

where  $\Pr(h_{obs} | h_{anc}, \tau, \mu)$  is given in equation (1), and the value of  $P_{null}(h_{obs})$  is the haplotype frequency for a non-ancestral haplotype. With the above setup, the likelihood for the observed haplotype data can be evaluated for each set of model parameters. Therefore, maximum likelihood estimates of the model parameters and their statistical inference can be made from the observed haplotype data.

## Genotype Data

As mentioned above, it is not always possible to determine haplotypes for each individual, even with the help of genotypes from the relatives. In this article, we assume that we have collected a random sample of affected individuals and an independent sample of normal individuals. We further assume that each individual is genotyped at a set of genetic markers in a candidate region.

Suppose that there are two alleles at the variant locus: the disease causing allele  $D$  and the normal allele  $d$ . The basic idea of our approach is to treat the chromo-

somes in the affected individuals as if they were a random sample of individual chromosomes from a chromosome population that consists of both chromosomes with the disease causing allele  $D$  and chromosomes with the normal allele  $d$ . For simplicity, we call this population the disease chromosome population in our following discussion. Similarly, we treat the chromosomes in the normal individuals as a random sample of chromosomes from a chromosome population consisting of only chromosomes without the disease causing allele  $D$ . We call this chromosome population the normal chromosome population. Under this assumption, we can form the likelihood for the observed genotype data and estimate the genetic parameters of interest through the maximum likelihood approach as described in detail below. It is apparent that this model assumption may not hold for general disease models. However, no likelihood functions can be formed in the absence of disease models. Therefore, we use this assumption as an approximation and the simulation results in the following show that this assumption works well for a variety of disease models investigated.

In this article, we define an individual's genotype as the set of genotypes across all of the markers, and let  $G$  denote the number of distinct genotypes. For genotype  $g_i$ , let  $c_i$  denote the number of haplotype pairs that are compatible with  $g_i$ . If there are  $m$  ( $\geq 1$ ) heterozygous sites in  $g_i$ , then  $c_i = 2^{m-1}$ . The likelihood function for  $g_i$  for an affected individual is:

$$\Pr(g_i | h_{anc}, \tau, \mu, p) = \sum_{k=1}^{c_i} P\left[(h_{ik}, h_{ik_c}) | h_{anc}, \tau, \mu, p\right],$$

where  $h_{ik}$  and  $h_{ik_c}$  are the  $k$ th haplotype pair that are compatible with  $g_i$ , parameters  $h_{anc}$ ,  $\tau$ ,  $\mu$ , and  $p$  were defined in our discussion on haplotype data. Under the assumption that each individual haplotype is a random sample from a disease chromosome population consisting of descendants of the ancestral haplotype and normal chromosomes,

$$\Pr(g_i | h_{anc}, \tau, \mu, p) = \kappa_i \sum_{k=1}^{c_i} P(h_{ik} | h_{anc}, \tau, \mu, p) P(h_{ik_c} | h_{anc}, \tau, \mu, p),$$

where  $\kappa_i = 1$  if all markers are homozygous for  $g_i$  and  $\kappa_i = 2$  otherwise. The probability of  $P(h_{ik} | h_{anc}, \tau, \mu, p)$  or  $P(h_{ik_c} | h_{anc}, \tau, \mu, p)$  can be evaluated using equation (2).

With this model, we can use the Expectation-Maximization (EM) algorithm [Dempster et al., 1977] to obtain maximum likelihood estimates of genetic parameters of interest. Number all possible haplotypes from 1 to  $H$ , where  $H$  is the total number of possible haplotypes. The unobserved "complete" data are the counts for haplotypes  $h_j$  from the disease chromosome population and the counts for haplotypes  $h_j$  from the normal chromosome population, where  $j = 1, \dots, H$ . Let  $y_j^A$  denote the number of chromosomes with  $h_j$  in the disease population, and  $y_j^U$  denote the number of chromosomes with  $h_j$  in the normal population. Note that when only genotype data are available,  $y_j^A$  and  $y_j^U$  are not observable. The EM algorithm proceeds as follows. At the E-step, we first calculate the conditional probability that an affected individual carries haplotype  $h_j$  given that his/her genotype is  $g_i$ :

$$P^A(h_j | g_i) = \frac{P^A(h_j, g_i)}{P^A(g_i)} = \frac{P^A(h_j)P^A(h_{j_c})}{\sum_{k=1}^{c_i} P^A(h_{ik})P^A(h_{ik_c})},$$

where haplotype pairs  $\{h_j, h_{j_c}\}$  are compatible with  $g_i$ , and the sum is over all haplotype pairs compatible with  $g_i$ . Therefore, the expected number of  $h_j$  among the affected individuals conditional on the observed genotype data and the current parameters  $\theta = \{h_{anc}, \tau, \mu, p\}$  is

$$E(y_j^A | \theta) = \sum_{i=1}^G x_i^A P^A(h_j | g_i),$$

where  $x_i^A$  denotes the number of affected individuals with genotype  $g_i$ . Similarly, we can calculate  $E(y_j^U | \theta)$ , the expected number of haplotype  $h_j$  among the normal individuals. At the M-step, we find the maximum likelihood estimates of the model parameters by treating the  $E(y_j^A | \theta)$  and  $E(y_j^U | \theta)$  as if they were the observed counts of haplotype  $h_j$ . We then repeat the E-step and M-step until parameter estimates converge. Note that the EM algorithm has also been used to estimate haplotype frequencies [Hawley and Kidd, 1995; Long et al., 1995; Slatkin and Excoffier, 1995]. The difference between our procedure and the previous one is that only a homogeneous population is assumed and analyzed in the previous studies, whereas we are explicitly modeling two populations, a disease population and a normal population, with the goal of identifying the mutation location.

In principle, we need to maximize the likelihood with respect to other parameters for each possible ancestral haplotype. However, the large number of possible ancestral haplotypes makes this approach infeasible in practice. Instead, we only consider those haplotypes whose observed frequency is higher than some pre-specified threshold value as candidate ancestral haplotypes. As for the estimation of haplotype frequencies in the normal sample group, generally there is not enough information to estimate the frequencies for all possible haplotypes, although the EM algorithm can be applied in principle. In our simulations, there are far more haplotypes than the number of individuals in the normal group. Therefore, as an approximation, we use the Markov model to calculate each haplotype frequency. To obtain a confidence interval for the location of the variant, we can follow the procedure of McPeck and Strahs [1999] by inverting the likelihood ratio test.

The above discussion treats each haplotype as independent of other haplotypes, and such a procedure is referred to as a type I composite likelihood method by Rannala and Slatkin [2000]. They noted that this approach may substantially underestimate the uncertainties in the parameter estimates. To account for the dependence due to population structure, the quasi-score estimating equation [McCullagh and Nelder, 1989; McPeck and Strahs, 1999] can be used. Assume that the variance of the score function for every individual is equal and the correlation (denoted by  $\rho$ ) between any two individuals is also equal. Then, the quasi-score estimators of the parameters are equivalent to the maximum likelihood estimates of the parameters in the case of

independence, but the standard errors are inflated by a factor of  $\sqrt{1 + (n - 1)\rho}$ , where  $n$  is the sample size [McPeck and Strahs, 1999]. In this case, the log-likelihood that we use is the same as that in the independent case but is multiplied by the factor  $[1 + (n - 1)\rho]^{-1}$ . Thus, the quasi-likelihood estimates are equal to the maximum-likelihood estimates in the case of independence, but the standard errors of the estimates are inflated by a factor of  $\sqrt{1 + (n - 1)\rho}$ . To evaluate the value of correlation  $\rho$  between two haplotypes, which depends on population structure, we can use the conditional coalescence model with variable population size (CCV) to evaluate  $\rho_h$  [Zhang and Zhao, 2000]. Here, for a pair of genotypes  $G_i$  and  $G_j$  each containing two haplotypes, i.e.,  $G_i = \{H_{i1}, H_{i2}\}$  and  $G_j = \{H_{j1}, H_{j2}\}$ , we use the canonical correlation between the two groups of the haplotypes  $\{H_{i1}, H_{i2}\}$  and  $\{H_{j1}, H_{j2}\}$  as the correlation between  $G_i$  and  $G_j$ . The canonical correlation of  $\{H_{i1}, H_{i2}\}$  and  $\{H_{j1}, H_{j2}\}$  is the maximum value of the correlation between the linear combinations of  $H_{i1}$  and  $H_{i2}$  and the linear combinations of  $H_{j1}$  and  $H_{j2}$ , which is  $\rho = 2\rho_h\sqrt{1 + \rho_h}$ .

### Simulation Setup

We assessed the performance of our proposed method through simulation studies by varying the disease model, the mutation rates of the microsatellite markers, the inter-marker distance, and the sample size. The time  $\tau$  to the most recent common ancestor was fixed at 100 generations for all simulation models. The specifications of these parameters are summarized as follows.

1. Four disease models: recessive, multiplicative, additive, and dominant. Let  $f_{DD}$ ,  $f_{Dd}$ , and  $f_{dd}$  be the penetrances of genotypes  $DD$ ,  $Dd$ , and  $dd$ , respectively, with  $f_{DD} \geq f_{Dd} \geq f_{dd}$ . The penetrances under these disease models are: (1) recessive model:  $f_{dd} = \alpha$  and  $f_{Dd} = f_{DD} = \beta$ ; (2) multiplicative model:  $f_{DD} = \alpha$ ,  $f_{Dd} = \sqrt{\alpha\beta}$ , and  $f_{dd} = \beta$ ; (3) additive model:  $f_{DD} = \alpha$ ,  $f_{Dd} = (\alpha + \beta)/2$ , and  $f_{dd} = \beta$ ; and (4) dominant model:  $f_{DD} = f_{Dd} = \alpha$  and  $f_{dd} = \beta$ . The penetrances for the three genotypes for each disease model, and the frequency of the disease-causing allele  $D$  are summarized in Table I.
2. Genetic markers: We considered six microsatellite markers, each having five alleles. For most simulation settings, these six markers were evenly distributed across a one centi-Morgan region with the disease-causing variant located in the middle of this region. We always defined the location of the disease-causing variant at 0 cM. Therefore, the six markers were located at  $-0.5$ ,  $-0.3$ ,  $-0.1$ ,  $0.1$ ,  $0.3$ , and  $0.5$  cM, respectively. The only exception was when we considered the effects of marker spacings, where the distance between the two outside markers was varied from 1, 2, 3, to 4 cM region.
3. Mutation rate: The mutation rate was fixed at  $2 \times 10^{-4}$  per locus per meiosis except for the results in Table III where we discuss the effects of mutation rates on our disease gene location estimate.
4. Sample size: The sample sizes for the two groups were the same and the number of individuals in each group was varied from 100, 200, to 300.

To simulate genotypes for individuals in the case group and those in the control group, we first generate haplotypes carrying normal allele  $d$  and those carrying disease mutation  $D$ . To generate the haplotypes carrying  $d$ , we assume that the markers

TABLE I. Disease Models Used in the Simulations\*

Model	$f_{DD}$	$f_{Dd}$	$f_{dd}$	Prevalence	$P_D$
Recessive	0.3	0.01	0.01	0.05	0.371
Multiplicative	0.3	0.055	0.01	0.05	0.276
Additive	0.3	0.15	0.01	0.05	0.142
Dominant	0.3	0.30	0.01	0.05	0.072

\*Two alleles are assumed at the disease locus,  $D$  and  $d$ . For each mode of inheritance, the penetrance for each genotype,  $DD$ ,  $Dd$ , or  $dd$ , is given together with the frequency of the disease causing allele  $D$ .

were in linkage equilibrium and all marker alleles had the same allele frequency. To generate the haplotypes carrying  $D$ , we use the CCV [Zhang and Zhao, 2000] under the assumption that the sampled haplotypes have a most recent common ancestor (MRCA) 100 generations ago; the population size (the number of haplotypes carrying disease mutation  $D$ ) is  $10^5$  at present time and 100 at the time 100 generations before the present time and the population has experienced an exponential growth. To pair two haplotypes to form an individual, we assumed a random mating population and derived the genotype distribution at the mutation site conditional on an individual being affected or not under specific disease models. Then we generated both affected and normal individuals independently according to this expected genotype distribution within each phenotype class, affected or normal. For each simulated data set with genotypes from the affected and normal individuals, we estimated the disease mutation location  $z$ , the mutation rate of the microsatellite markers  $\mu$ , the ancestral haplotype  $h_{anc}$ , the time to the MRCA  $\tau$ , and the heterogeneity parameter  $p$  through the EM algorithm described above. To obtain a confidence interval for location of the variant, we evaluated the correlation coefficient  $\rho$  based on the CCV using the estimated  $\tau$  as the time to the MRCA. The population size at the present time and at the time of the MRCA were assumed to be known. For each simulation scenario, the performance of the genetic variant location estimate, its variability, and the coverage probability of its confidence interval were assessed through 500 independent repetitions.

## RESULTS

### Simulation Results

In the first set of simulations, the disease-causing mutation was introduced 100 generations ago, the mutation rate was fixed at  $2 \times 10^{-4}$ , and the inter-marker distance was fixed at 0.2 cM. The results under different disease models and sample sizes are summarized in Table II. In Table II, the Sample Size column is the number of affected individuals studied, and the same number of normal individuals as studied in the control group. The Estimated Location column is the average of the estimated gene locations from 500 simulated samples. The Standard Deviation column is the standard deviation of the 500 estimated gene locations. For each type of confidence interval, 90 or 95%, we summarize the proportion of the intervals that covered the true gene location in the Coverage column and the average length of the confidence intervals in the Length column. The average estimate of the variant location from 500 independent replications was almost unbiased. The variability of the loca-

TABLE II. Performance of the Proposed Method to Map the Disease-Causing Variant

Model	Sample size	Estimated location	Standard deviation	90% CI		95% CI	
				Coverage (%)	Length	Coverage (%)	Length
Recessive	100	-0.000	0.104	91	0.288	96	0.325
	200	-0.002	0.081	92	0.243	97	0.288
	300	-0.001	0.084	91	0.200	96	0.253
Multiplicative	100	0.001	0.123	88	0.260	92	0.300
	200	-0.002	0.093	89	0.231	94	0.267
	300	0.000	0.081	89	0.189	94	0.241
Additive	100	0.003	0.145	93	0.396	98	0.610
	200	0.001	0.090	93	0.340	97	0.510
	300	-0.002	0.085	92	0.241	96	0.310
Dominant	100	-0.001	0.167	93	0.415	97	0.651
	200	0.000	0.096	92	0.381	97	0.567
	300	0.001	0.093	92	0.356	96	0.520

tion estimates increased in the order of the recessive model, the multiplicative model, the additive model, and the dominant model. The average length of the confidence intervals was less different across the disease models. The proportion of confidence intervals that covered the true variant location was close to the nominal level.

To study the effects of the mutation rate of the microsatellite markers, we varied the mutation rate at  $10^{-4}$ ,  $5 \times 10^{-4}$ , and  $10^{-3}$ , and the results are summarized in Table III. Genotype data were available from 200 affected individuals and 200 normal individuals in the target population, and the inter-marker distance was fixed at 0.2 cM. The mutation was assumed to be introduced 100 generations ago. In Table III, the Mutation Rate column is the assumed mutation rate for the microsatellite markers. The Estimated Mutation, Age, and Location columns are the average of the estimated mutation rates, ages of the MRCA, and gene locations from 500 simulated samples, respectively. It can be seen that higher mutation rate did not result in bias in the genetic variant location estimate but did increase the variability of both the location estimate and the average length of the confidence intervals. The coverage probability was close to the nominal level in almost all of the cases. The estimates of the

TABLE III. Effects of Mutation Rates on Locating the Disease-Causing Variant Under Different Disease Models for the Common Disease

Model	Mutation rate	Estimated mutation	Estimated age	Estimated location	Standard deviation	90% CI	
						Coverage (%)	Length
Recessive	$1 \times 10^{-4}$	$8 \times 10^{-5}$	90	-0.002	0.092	92	0.243
	$5 \times 10^{-4}$	$2 \times 10^{-4}$	93	-0.001	0.117	93	0.265
	$1 \times 10^{-3}$	$5 \times 10^{-4}$	101	0.002	0.140	91	0.288
Multiplicative	$1 \times 10^{-4}$	$8 \times 10^{-5}$	109	-0.002	0.093	89	0.230
	$5 \times 10^{-4}$	$2 \times 10^{-4}$	104	0.001	0.123	91	0.251
	$1 \times 10^{-3}$	$7 \times 10^{-4}$	110	-0.008	0.144	88	0.280
Additive	$1 \times 10^{-4}$	$5 \times 10^{-4}$	110	0.001	0.090	93	0.330
	$5 \times 10^{-4}$	$7 \times 10^{-4}$	105	0.000	0.165	92	0.370
	$1 \times 10^{-3}$	$1.5 \times 10^{-3}$	110	0.008	0.148	95	0.420
Dominant	$1 \times 10^{-4}$	$6 \times 10^{-4}$	105	0.000	0.096	93	0.370
	$5 \times 10^{-4}$	$9 \times 10^{-4}$	107	0.004	0.142	92	0.396
	$1 \times 10^{-3}$	$2 \times 10^{-3}$	109	0.005	0.158	95	0.430

time of the MRCA were close to the true value. However, the estimates of the mutation rate had some bias.

In the last set of simulations, we studied the effects of inter-marker distance on the location estimate and summarize the results in Table IV. Genotype data were available from 200 affected individuals and 200 normal individuals in the target population. In Table IV, the Inter-marker Distance column is the assumed distance between adjacent microsatellite markers. Larger inter-marker distance did not lead to bias but did increase the variation of the location estimate. The effect was most pronounced for the average length of the confidence intervals. The length almost increased linearly with the inter-marker distance. For almost all of the cases, the coverage probability of the confidence intervals was very close to the nominal level.

In summary, these simulation results suggest that the proposed estimate of genetic mutation location from genotype data work well under the simulation models studied, even with a relatively small sample size ( $n = 100$ ) and a relatively high mutation rate ( $\mu = 10^{-3}$ ). Note that the proposed procedure works well despite the fact that the analysis models may be different from the simulation models in these simulations. In addition, the estimation of the time to the MRCA was close to the true value.

#### Application to the EPM1 Data

The EPM1 gene involved in progressive myoclonus epilepsy was mapped to chromosome 21q22.3 by Virtaneva et al. [1996] using 88 haplotypes with five microsatellite markers spanning a 900-kb region (D21S1885–D21S2040–D21S1259–D21S1912–PFKL), and it was then cloned between D21S2040 and D21S1259, with approximately 30 kb from marker D21S2040. Besides the 88 haplotypes with EPM1 gene, there are allele frequencies for the normal individuals reported in Virtaneva et al. [1996]. Because this is a recessive disease, we assumed the following penetrances

**TABLE IV. Effects of Inter-Marker Distances on Locating Disease Causing Variant Under Different Disease Models for the Common Disease**

Model	Inter-marker distance	Estimated location	Standard deviation	90% CI		95% CI	
				Coverage (%)	Length	Coverage (%)	Length
Recessive	0.2	-0.001	0.092	92	0.243	97	0.288
	0.4	0.001	0.145	93	0.415	96	0.471
	0.6	-0.002	0.213	92	0.660	97	0.781
	0.8	-0.003	0.275	93	0.915	96	1.131
Multiplicative	0.2	-0.002	0.093	89	0.263	93	0.310
	0.4	-0.002	0.166	89	0.410	92	0.510
	0.6	-0.003	0.202	91	0.682	96	0.812
	0.8	-0.002	0.320	91	0.101	93	1.350
Additive	0.2	0.001	0.090	93	0.280	97	0.361
	0.4	0.002	0.170	92	0.450	97	0.580
	0.6	0.000	0.201	92	0.720	96	0.900
	0.8	-0.003	0.382	89	1.115	93	1.480
Dominant	0.2	0.000	0.096	93	0.288	97	0.350
	0.4	-0.003	0.176	92	0.500	97	0.590
	0.6	-0.007	0.230	89	0.702	93	0.900
	0.8	0.008	0.390	88	1.215	92	1.560

$f_{DD} = 0.8$ ,  $f_{Dd} = f_{dd} = 0.001$  and the prevalence of 0.005 and calculated the genotype distribution conditional on an individual being affected or not in a homogeneous population. We then generated both affected and normal individuals independently according to this expected genotype distribution. To generate an individual carrying two copies of the disease allele  $D$ , we randomly sampled two haplotypes from the 88 haplotypes with the EPM1 mutation. To generate an individual carrying one disease allele  $D$  and one normal allele  $d$ , we randomly sampled one haplotype from the 88 haplotypes and another haplotype generated according to the allele frequencies in the normal population by assuming that the markers are in linkage equilibrium. To generate an individual with two copies of the normal allele  $d$ , we randomly generated two haplotypes according to the allele frequencies in the normal population by assuming that the markers are in linkage equilibrium. In our simulations, we generated 500 data sets with each data set consisting of 100 cases and 100 controls and another 500 data sets with each data set consisting of 200 cases and 200 controls.

When we applied our method to the data sets with 100 cases and 100 controls, the mean estimated location was in the correct marker interval and approximately 81 kb from marker D21S2040, the mean estimated mutation rate was  $1.2 \times 10^{-4}$ , and the mean estimated time to the MRCA was 49 generations from present day. For the data sets with 200 cases and 200 controls, the mean estimated location was also in the correct marker interval and approximately 85 kb from marker D21S2040, the mean estimated mutation rate was  $1.4 \times 10^{-4}$ , and the mean estimated time to the MRCA was 47 generations from present day. These results are very similar to the results by using haplotype data directly [Xiong and Guo, 1997; McPeck and Strahs, 1999; Zhang and Zhao, 2000]. To construct confidence interval by using the CCV, we used the estimated time to the MRCA and assumed an exponential population growth model with current disease population size  $10^4$  at present day and 20 at the time of the MRCA [Zhang and Zhao, 2000]. For the 95% confidence interval, the proportion of the intervals that covered the true gene location was 96% when the sample consisted of 100 cases and 100 controls and the coverage probability was 99% when the sample consisted of 200 cases and 200 controls.

## DISCUSSION

We have proposed a multilocus linkage disequilibrium mapping method to directly analyze genotype data. One advantage of the proposed method is that it is designed to analyze genotype data without the need to resolve haplotypes. Therefore, our method is particularly suited to case-control data where no relative information is available. Another advantage of the proposed method is that it does not specify a particular disease model. Therefore, this approach is applicable in the absence of exact knowledge on the underlying disease model. In our approach, we make the assumption that the two chromosomes in the affected individuals are random samples of chromosomes from a disease chromosome population. Although this assumption will not be strictly correct for an arbitrary disease model, simulation results show that this model assumption leads to excellent estimation of the genetic variant location. The heterogeneity parameter in the model,  $\rho$  in equation (2), can be thought as a parameter that adapts the model to different modes of inheritance.

In this article, we use quasi-score estimator and the CCV [Zhang and Zhao, 2000] to incorporate the dependence due to population structure and use canonical correlation of the haplotypes as an approximation of the correlation of individual genotypes. For most genetic models considered in our simulations, the coverage probability of the confidence intervals was very close to the nominal level. It is possible to extend our model to incorporate uncertainties in ancestral haplotype inference using the Bayesian approach developed by Morris et al. [2000].

Recombination and mutation that directly influence LD are explicitly incorporated in the DHS model. However, LD is also affected by many other factors, e.g., demographic history [Slatkin, 1994; Thompson and Neel, 1997; Laan and Pääbo, 1997]. In small populations with no demographic expansion, LD is likely to be caused by genetic drift [Terwilliger et al., 1998]. Such situations pose serious challenges to the DHS method and other LDM methods to make rigorous statistical inference. More heuristic and descriptive methods may be needed to help locate the disease genes.

The approach developed in this article is readily applicable to case-control data where only genotypes from the independent affected and normal individuals are available. If some relatives of these affected and normal individuals are also genotyped, such additional information can be used to reduce the uncertainty in the unobservable haplotypes, e.g., Excoffier and Slatkin [1998]. More generally, if both phenotype and genotype information is available from many individuals in a pedigree, we should use each pedigree as the study unit. Extensions of our approach in this article to general pedigrees are not straightforward and further research is needed to develop general statistical and numerical methods to analyze general pedigree data. The method described in this article has been implemented in a computer program and the source code written in C is available at <http://zhao.med.yale.edu>.

## ACKNOWLEDGMENTS

We thank two referees for their constructive comments.

## REFERENCES

- Clayton D. 2000. Linkage disequilibrium mapping of disease susceptibility genes in human populations. *Int Stat Rev* 68:23–43.
- Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via EM algorithm (with discussion). *J R Stat Soc B* 39:1–38.
- Excoffier L, Slatkin M. 1998. Incorporating genotypes of relatives into a test of linkage disequilibrium. *Am J Hum Genet* 62:171–80.
- Hawley ME, Kidd KK. 1995. HAPLO: a program using the EM algorithm to estimate frequencies of multi-site haplotype. *J Hered* 86:409–11.
- Laan M, Pääbo S. 1997. Demographic history and linkage disequilibrium in human populations. *Nat Genet* 17:435–8.
- Lazzeroni LC. 2001. A chronology of fine-scale gene mapping by linkage disequilibrium. *Stat Methods Med Res* 10:57–76.
- Long JC, Williams RC, Urbanek M. 1995. An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am J Hum Genet* 56:799–810.
- McCullagh P, Nelder JA. 1989. *Generalized linear models*. London: Chapman and Hall.
- McPeck MS, Strahs A. 1999. Assessment of linkage disequilibrium by the decay of haplotype sharing with application to fine-scale genetic mapping. *Am J Hum Genet* 65:858–75.

- Morris AP, Whittaker JC, Balding DJ. 2000. Bayesian fine-scale mapping of disease loci, by hidden Markov models. *Am J Hum Genet* 67:155–69.
- Rannala B, Slatkin M. 2000. Method for multipoint disease mapping using linkage disequilibrium. *Genet Epidemiol* 19:S71–7.
- Slatkin M. 1994. Linkage disequilibrium mapping in growing and stable populations. *Genetics* 137:331–6.
- Slatkin M, Excoffier L. 1995. Maximum likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 12:921–7.
- Terwilliger JD, Zöllner S, Laan M, Pääbo S. 1998. Mapping genes through the use of linkage disequilibrium generated by genetic drift: ‘drift mapping’ in small populations with no demographic expansion. *Human Hered* 48:138–54.
- Thompson EA, Neel JV. 1997. Allelic disequilibrium and allele frequency distribution as a function of social and demographic history. *Am J Hum Genet* 60:197–204.
- Virtaneva K, Miao J, Träskelin A-L, Stone N, Warrington JA, Weissenbach J, Meyers RM, et al. 1996. Progressive myoclonus epilepsy EPM1 locus maps to a 175-kb interval in distal 21q. *Am J Hum Genet* 61:1247–53.
- Xiong M, Guo SW. 1997. Fine-scale genetic mapping based on linkage disequilibrium. Theory and applications. *Am J Hum Genet* 60:1513–31.
- Zhang S, Zhao H. 2000. Linkage disequilibrium mapping using the decay of haplotype sharing method with the step-wise mutation model and variable population size. *Genet Epidemiol* 19:S99–S105.