Genetic Epidemiology 19(Suppl 1):S99-S105 (2000)

Linkage Disequilibrium Mapping in Populations of Variable Size Using the Decay of Haplotype Sharing and a Stepwise-Mutation Model

Shuanglin Zhang and Hongyu Zhao*

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

Linkage-disequilibrium (LD) mapping is a powerful tool for fine-mapping disease genes. Recently, McPeek and Strahs [(1999) Am J Hum Genet 65:858-875] proposed a multilocus model for LD mapping based on the decay of haplotype sharing. Here we extend their approach in two ways. First, instead of assuming each marker allele has an equal chance to mutate to one of the other marker alleles, we use the stepwise-mutation model to describe the mutation process for microsatellite markers. Second, in addition to the independence model and the constant population size model they considered, we model the dependence among observed haplotypes due to population structure by using a general conditional-coalescent model with variable population size. Through simulation studies, we study the effects of the stepwise-mutation model and variable population size on the estimates of disease gene location, mutation rate, and time to the most recent common ancestor of the sampled haplotypes. We then use this method to analyze progressive myoclonus epilepsy data. Genet. Epidemiol. 19(Suppl 1):S99-S105, 2000.

Key words: linkage disequilibrium; stepwise-mutation model; coalescent model

*Correspondence to: Dr. Hongyu Zhao, Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520. E-mail: hongyu.zhao@yale.edu

INTRODUCTION

Linkage disequilibrium (LD) mapping has been successfully applied to map many disease genes, and various statistical methods for its implementation have been proposed. McPeek and Strahs [1999] proposed a multilocus model based on the decay of haplotype sharing that has been found to work well for fine mapping. This method models dependence across loci within a haplotype by considering LD as the extent of a region of a shared haplotype around a variant. It allows for mutation and dependence among the observed haplotypes due to population structure. For mutations, McPeek and Strahs [1999] assumed that each allele of a given marker has equal chance to mutate to any other allele (the equal-mutation model (EMM) discussed below). The conditional-coalescent model they considered assumed a constant population size.

Microsatellite markers are widely used in genetic linkage studies. They evolve predominantly by the gain or loss of a single repeat unit, or a small number of repeat units. The stepwise-mutation model (SMM) [Ohta and Kimura, 1973] has been found to model well the mutation process for microsatellite markers [Goldstein, 1997]. Here we use SMM to describe the mutation process for microsatellite markers, and use the general conditional-coalescent model with variable population size (CCV) [Kingman, 1982; Griffiths and Tavaré, 1994, 1997] to model the population history. We incorporate SMM and CVV to extend McPeek and Strahs' [1999] decay-of-haplotype-sharing method, and study the performance of our modified approach for locating the disease gene and for estimating the mutation rate and the age of the most recent common ancestor for the sampled haplotypes. We then apply our method to analyze a real data set.

METHODS

Decay of Haplotype Sharing

Suppose a sample of haplotypes sharing a specific genetic variant are descendents of the same ancestral haplotype h_{anc}, with the genetic variant introduced τ generations ago. We further assume that genetic markers with known locations are observed surrounding the genetic variant. In our analysis, the position of the variant and the ancestral haplotype are unknown parameters. Define the position of the variant to be at 0, with markers 1, 2, 3, ..., l_{re} at increasing distance on one side of the variant and loci -1, -2, -3,..., - l_{le} at increasing distance on the other side of the variant. The likelihood function for the observed haplotype h_{obs} (equation (4) in McPeek and Strahs [1999]) is

$$L(h_{obs} \mid \tau, h_{anc}, \mu) = \sum_{i=0}^{l_{re}} \sum_{j=0}^{l_{le}} \{[g(\tau, -j, i) \prod_{k=-j}^{i} m(k, \tau, h_{anc}(k), h_{obs}(k))$$
$$\times P_{null}[h_{obs}(i+1), h_{obs}(i+2), ..., h_{obs}(l_{re})] \times P_{null}[h_{obs}(-j-1), h_{obs}(-j-2), ..., h_{obs}(-l_{le})]\}.$$

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 τ generations there are no crossovers between markers $-j$ and i, at least one crossover between markers $-j$ - 1 and $-j$ and at least one crossover between markers i and i + 1, where d_{kl} is the genetic distance between loci k and l. If either marker $-j$ or i is on the edge of the observed haplotype, the term $(1 - e^{-\tau d_{-j-1,-j}})$ or $(1 - e^{-\tau d_{i,i+1}})$ is not

in the expression. The factor $m(k,\tau,h_{anc}(k),h_{obs}(k))$ is the conditional probability that a τth generation descendant has allele $h_{obs}(k)$ at marker k, given that the ancestral haplotype has allele $h_{anc}(k)$ at marker k, and given that there are no crossovers between the genetic variant and marker k during the τ generations, i.e., m is the mutation rate. The factor $P_{null}[h_{obs}(i+1),h_{obs}(i+2),...,h_{obs}(l_{re})]$ is the joint probability that the alleles $h_{obs}(i+1),h_{obs}(i+2),...,h_{obs}(l_{re})$ occur in a nonancestral haplotype, and the factor $P_{null}[h_{obs}(-j-1),h_{obs}(-j-2),...,h_{obs}(-l_{le})]$ is defined similarly.

Extension 1: The Stepwise-Mutation Model (SMM)

Here the factor $m(k,\tau,h_{anc}(k),h_{obs}(k))$ is evaluated under SMM. For a given marker, SMM assumes that allele M_i, is indexed according to the number of repeats. SMM also assumes that allele M_i, can only mutate to the next larger allelic state M_{i+1} (expansion) with probability μ and to the next smaller allelic state M_{i-1} (contraction) with probability ν. Let M_1 denote the allele with the smallest number of repeats and M_G denote the allele with the largest number. Assume that allele M_1 only can mutate to allele M_2 and that allele M_G only can mutate to allele M_{G-1}. We also assume $\mu = \nu$. Let $\mathbf{P} = (p_{ij})$, where $p_{ij} = 1-2\mu$ when $i = j$, $i \neq 1$, $i \neq G$; $p_{ij} = 1-\mu$ when $i = j = 1$ or $i = j = G$; $p_{ij} = \mu$ when $|i-j| = 1$; and $p_{ij} = 0$ otherwise. Then, the mutation probability from allele M_i to allele M_j after τ generations at locus l, denoted by $m(l,\tau,i,j)$ is the ijth entry in the matrix exponential $\mathbf{P}^{\tau}$. While we assume the same mutation rate for all markers, it is a straightforward extension to allow different marker mutation rates.

Extension 2: Conditional-Coalescent Model for Variable Population Size (CCV)

If the observed haplotypes can be considered to be independent, the overall likelihood is simply the product of the likelihoods for individual haplotypes. In general, the independence assumption is not valid, especially for isolated populations that underwent rapid population growth. Assume that the variance of the score function for every individual is equal, and the correlation (denoted by ρ) between any two individuals is also equal. Then, the quasi-score estimators of parameters are equivalent to the MLEs of the parameters for the independence case, but the standard errors are inflated by a factor $\sqrt{1+(n-1)\rho}$, where n is the sample size. McPeek and Strahs [1999] showed that

$$\rho = \sum_{j=1}^{n-1} \frac{2(n+1)}{(n-1)(n-j+1)(n-j+2)} \int f_j(t) \frac{\tau - t}{\tau + t} dt,$$

where $f_j(t)$ is the density function of the jth coalescent time conditional on time τ to the most recent common ancestor. We model the population structure by CCV conditional on time τ to the most recent common ancestor and use the correlation of the ancestral segment length of two individuals as an approximation of ρ. Our conditional-coalescent model is a variation on the coalescent model of variable population size [Kingman, 1982;

Griffiths and Tavaré, 1994, 1997]. Suppose that the Wright-Fisher model holds and the population size is $N(0) \equiv N$ (haplotypes) at the time of sampling, and is size $N(r)$ at the rth generation back from the present day. We assume that there is a relative size function v such that for all $x \geq 0$,

$$v(x) = \lim_{N\to\infty} \frac{N([Nx])}{N} > 0.$$

Define the population size intensity function $\Lambda(t) = \int_0^t v(x)^{-1} dx$, where T_j is the jth coalescent time with unit of N generations. Then, $E_j = \Lambda(T_j) - \Lambda(T_{j-1})$ are independent exponential random variables with parameter $(n-j)(n-j+1)/2$. The density of $\Lambda(T_j)$ under the condition $T_{n-1}=\frac{\tau}{N}=a$ (or $\Lambda(T_{n-1}) = \Lambda(a) = b$) is

$$f_j^*(x) = -\sum_{i=n-j}^{n-1} \prod_{k=n-j,k\neq i}^{n-1} (\lambda_i - \lambda_k)^{-1} e^{-x\lambda_i} \bullet \sum_{h=1}^{n-j-1} \prod_{l=1,l\neq h}^{n-j-1} (\lambda_h - \lambda_l)^{-1} e^{-(b-x)\lambda_h}$$
$$\div \sum_{i'=1}^{n-1} \prod_{j'=1,j'\neq i'}^{n-1} (\lambda_{i'} - \lambda_{j'})^{-1} e^{-b\lambda_{i'}},$$

where $\lambda_i = i(i+1)/2$. Then the correlation ρ can be calculated as

$$\rho = \sum_{j=1}^{n-1} \frac{2(n+1)}{(n-1)(n-j+1)(n-j+2)} \int_0^b f_j^*(t) \frac{a - \Lambda^{-1}(t)}{a + \Lambda^{-1}(t)} dt.$$

When the population underwent exponential growth, $N(t) = Ne^{-rt}$ and $\Lambda(x) = (e^{rNx} - 1)/(rN)$.

RESULTS

Analysis of Simulated Data

In our simulations, we considered six microsatellite markers. Each had five alleles with equal allele frequency in the normal population. These six markers were evenly distributed across a 1cM region (0.2cM between each pair of markers), with the variant located at the middle of the six markers (0.5cM from the leftmost marker). We assumed that the ancestral haplotype was introduced 100 generations ago (τ = 100). To obtain a confidence interval for the location of the variant, we inverted the likelihood ratio test [McPeek and Strahs, 1999]. In the first set of simulations, we assessed the performance of the method under different mutation rates for SMM. In addition, we compared the results with those obtained under EMM studied by McPeek and Strahs [1999], and a no-mutation model (NMM) [Service et al., 1999]. The results are summarized in Table I. For the genetic variant location estimation, the three models yielded similar results when the

TABLE I. Statistical Estimates of Genetic Variant Location, Mutation Rate, and Age of Mutation Under Different Mutation Models (50 Independent Haplotypes)

	True μ	Estimate d μ	Location Estimate	Coverage for 95% CI	SD for Location Estimate	Mean Length of 95% CI	Mean τ Estimate
SMM	1×10^{-4}	9.7×10^{-5}	0.50	95%	0.052	0.17	101
	6×10^{-4}	8.5×10^{-4}	0.50	94%	0.059	0.19	99
	1×10^{-3}	1.5×10^{-3}	0.50	94%	0.065	0.20	99
	3×10^{-3}	3.7×10^{-3}	0.51	92%	0.110	0.26	104
EMM	1×10^{-4}	*	0.50	94%	0.054	0.17	98
	6×10^{-4}	*	0.50	95%	0.065	0.22	104
	1×10^{-3}	*	0.52	91%	0.070	0.26	110
	3×10^{-3}	*	0.57	75%	0.215	0.49	125
NMM	1×10^{-4}	0	0.50	94%	0.041	0.16	109
	6×10^{-4}	0	0.50	96%	0.042	0.18	168
	1×10^{-3}	0	0.51	91%	0.048	0.20	220
	3×10^{-3}	0	0.56	65%	0.200	0.30	424

*Values not shown, since these parameter estimates have different meaning from those in SMM.

mutation rate was small ($\mu \le 6\times10^{-4}$). When μ was larger, there was little bias in location estimation using SMM and the approximate correct coverage probability for the 95% CI was obtained. However, the results from EMM and NMM yielded larger bias and the constructed 95% CIs had poor coverage probability. The estimated τ was more sensitive to the mutation model used in the analysis. When SMM was used, the estimated τ was very close to the true value (100 generations). The bias was larger when EMM was assumed, and the bias was substantial under NMM.

We then studied the effects of marker distance, sample size, and the presence of non-ancestral haplotypes on the variant location estimate. Larger inter-marker distance did not result in bias but did increase the CI. When the sample size was varied, there was no bias even for samples of size 25, but larger sample size lead to smaller CIs. In the circumstance where the variant may lie on two or more ancestral haplotypes, McPeek and Strahs [1999] introduced a parameter p to represent the proportion of the variant haplotypes in the population that are not descended from the ancestral haplotype. Using this parameter p, the likelihood for the observed haplotype h_{obs} can be written as $(1-p)L(h_{obs}\,|\,\tau, h_{anc}, \mu) + pP_{null}(h_{obs})$. The presence of non-ancestral haplotypes did not cause bias, but resulted in more uncertainty in the location estimate.

We further studied the performance of our method under the exponential population growth model. We assumed that the current disease population size is 10^5 and the founding population size was 10, with the most recent common ancestor at 100 generations from the present day. The results are summarized in Table II. The location estimate for the genetic variant was almost unbiased and the 95% CIs had approximately the correct coverage probability. In addition, the estimated τ was almost unbiased. The increase in the size of the CIs due to non-independence among observed haplotypes depended on the population growth model specified. For example, in the above simulations, the factor $\sqrt{1+(n-1)\rho}$ used in the CI construction was 1.98. This factor would be 2.38, 1.70, 1.60, 1.45, and 1.37 if the current disease population size were 10^4, 10^6, 10^7, 10^8, and 10^9, respectively.

TABLE II. Statistical Estimates of Genetic Variant Location, Mutation Rate, and Age of Mutation Under the Exponential Population Growth Model (50 Haplotypes)

True μ	Estimated μ	Location Estimate	Coverage for 95% CI	SD for Location Estimate	Mean Length of 95% CI	Mean τ Estimate
$1x10^{-4}$	$1.3x10^{-4}$	0.50	95%	0.087	0.38	97
$6x10^{-4}$	$9.5x10^{-4}$	0.50	96%	0.096	0.43	98
$1x10^{-3}$	$1.7x10^{-3}$	0.50	94%	0.121	0.46	98
$3x10^{-3}$	$3.7x10^{-3}$	0.51	93%	0.180	0.54	104

Analysis of Progressive Myoclonus Epilepsy Data

The EPM1 gene involved in progressive myoclonus epilepsy was mapped to chromosome 21q22.3 [Vitaneva 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 (~30-kb from marker D21S2040). We applied our method to this data set to estimate the location of the EPM1 gene, time to the most recent common ancestor τ, mutation rate μ, and the heterogeneity parameter p. Because the estimated heterogeneity parameter p was not significantly different from zero, we assume $p = 0$ in the following discussion. Figure 1 shows the log-likelihood curve. The estimated location (triangle in the figure) was in the correct marker interval. The 95% CI assuming independent haplotypes (top horizontal bar) and the 95% CI based on CCV assuming an exponential population growth model (bottom horizontal bar) both contain the true gene (vertical bar). The estimated mutation rate was 10^{-4} and the estimated time to the most recent common ancestor of the sampled haplotypes was $\tau = 50$. In the exponential population growth model, we assumed that the present day population with the disease mutation is 10^4, and the founding population size was 20. Compared to McPeek and Strah's [1999] results, our estimated EPM1 location was closer to the true location, and the 95% CI based on CCV was smaller (0.33cM vs. 0.50cM).

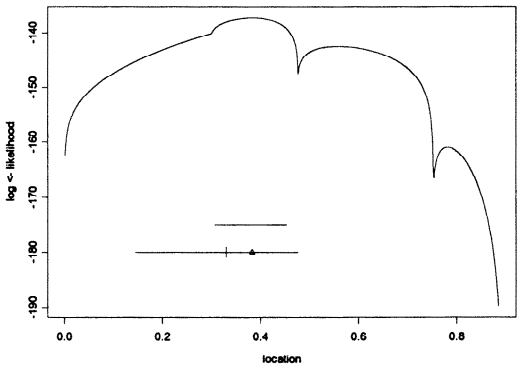

Fig. 1. Log-likelihood versus location for the EPM1 data set.

DISCUSSION

We have extended McPeek and Strahs' [1999] model for LD based on the decay of haplotype sharing to incorporate SMM for microsatellite markers and CCV to model dependence among the sampled haplotypes. Our simulation results showed that the method performed well for small sample sizes, a relatively high mutation rate, and the presence of non-ancestral haplotypes. For genetic variant location estimation, EMM and NMM yielded similar results when the mutation rate was low, but may have larger bias when the mutation rate was high if the mutation mechanism follows SMM. The mutation rate estimate was not as good, but the estimation of time to the most recent common ancestor was nearly unbiased under the correct mutation model. Note that SMM is most applicable for microsatellite markers. For single nucleotide polymorphisms (SNPs) that will be increasingly used in the future, all three mutation models will likely yield similar results because the mutation rate for SNPs is believed to be very low.

When the population structure was taken into account, the CI based on a conditional-coalescent model for constant population size larger than the CI based on the independence assumption. The CI based on the conditional-coalescent model for constant population size model is conservative when there is population expansion. When the exponential growth model was assumed in the analysis of the EPM1 data set, the CI based on CCV was smaller than the CI based on the coalescent model for constant population size, and it contained the true gene. This CCV model will provide more accurate information on the disease gene location when there is good knowledge on population history, which is the case for most isolated populations currently under study.

ACKNOWLEDGEMENTS

This research was supported by NIH (grants GM59507 and HD36834) and the March of Dimes (FY98-0752). Software for the statistical methods described in this paper (including source code written in C), as well as other programs developed by our group, can be found on the World Wide Web at http://zhao.med.yale.edu.

REFERENCES

Goldstein DB, Pollock DD. 1997. Launching microsatellites: a review of mutation processes and methods of phylogenetic inference. J Hered 88:335-342.

Griffiths RC, Tavaré S. 1994. Ancestral inference in population genetics. Stat Sci 9:307-319.

Griffiths RC, Tavaré S. 1997. Computational methods for the coalescent. In: Tavaré S, Donnelly P, editors. Progress in population genetics and human evolution. Berlin: Springer-Verlag. pp 165-182.

Kingman JFC. 1982. The coalescent. Stoch Proc Appl 13:235-248.

McPeek 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 65:858-875.

Ohta T, Kimura M. 1973. The model of mutation appropriate to estimate the number of electrophoretically detectable allele in a genetic population. Genet Res 22:201-204.

Service SK, Temple Lang DW, Freimer NB, Sandkuijl LA. 1999. Linkage-disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. Am J Hum Genet 64:1728-1738.

Virtaneva K, Miao J, Träskelin A-L, Stone N, Warrington JA, Weissenbach J, Meyers RM. 1996. Progressive myoclonus epilepsy EPM1 locus maps to a 175-kb interval in distal 21q. Am J Hum Genet 61:1247-1253.