Likelihood Analysis of Disequilibrium Mapping, and Related Problems

Bruce Rannala and Montgomery Slatkin
Department of Integrative Biology, University of California Berkeley, Berkeley

Summary
In this paper a theory is developed that provides the sampling distribution of alleles at a diallelic marker locus closely linked to a low-frequency allele that arose as a single mutant. The sampling distribution provides a basis for maximum-likelihood estimation of either the recombination rate, the mutation rate, or the age of the allele, provided that the two other parameters are known. This theory is applied to (1) the data of Hästbacka et al., to estimate the recombination rate between a locus associated with diastrophic dysplasia and a linked RFLP marker; (2) the data of Risch et al., to estimate the age of a presumptive allele causing idiopathic torsion dystonia in Ashkenazi Jews; and (3) the data of Tishkoff et al., to estimate the date at which, at the CD4 locus, non-African lineages diverged from African lineages. We conclude that the extent of linkage disequilibrium can lead to relatively accurate estimates of recombination and mutation rates and that those estimates are not very sensitive to parameters, such as the population age, whose values are not known with certainty. In contrast, we also conclude that, in many cases, linkage disequilibrium may not lead to useful estimates of allele age, because of the relatively large degree of uncertainty in those estimates.

Introduction
The presence of a nonrandom association either among alleles at two loci or between a disease phenotype and alleles at one or more marker loci indicates a recent shared history and can be used to estimate either the recombination rate or allele age (Lander and Botstein 1986). The principle underlying this approach is simple and relies on the fact that the extent of linkage disequilibrium decays exponentially with time, at a rate proportional to the recombination rate. If either the time span or the recombination rate is known, the other parameter can be estimated from the coefficient of linkage disequilibrium. This approach has been used several times. For example, Serre et al. (1990) used the extent of disequilibrium between the ΔF508 allele at the cystic fibrosis (CFTR) locus and a linked marker to estimate the age of that allele in European populations. Hästbacka et al. (1992) estimated the recombination rate between a microsatellite marker and a locus responsible for diastrophic dysplasia (DTD) in a Finnish population. Risch et al. (1995) used the extent of linkage disequilibrium between marker loci associated with idiopathic torsion dystonia (ITD) to estimate the age of a putative mutation causing that disease in Ashkenazi Jews. Tishkoff et al. (1996) applied similar reasoning to markers within the CD4 locus, in order to estimate the time at which the ancestors of non-Africans diverged from Africans.

The estimation of parameters on the basis of disequilibrium data raises several questions, in particular whether estimates are biased and what the associated confidence intervals are. To answer these questions, a statistical model is needed. Hästbacka et al. (1992) assumed that the Luria-Delbrück theory, originally developed to estimate mutation rates in bacteria, could be applied to estimate recombination rates in human populations. This application was questioned by Kaplan et al. (1995) and Kaplan and Weir (1995), who argued that “evolutionary variability” is not adequately accounted for by the Luria-Delbrück theory. They showed that accounting for that variability results in a confidence interval wider than that obtained by Hästbacka et al. (1992). Pritchard and Feldman (1996) made a similar point about the Tishkoff et al. (1996) conclusion, arguing that variability in the evolutionary process will make the confidence interval for the estimated time of origin of non-Africans much wider than had been suggested by Tishkoff et al. (1996).

In this paper, we address the same statistical issues but use a model that more adequately accounts for all sources of variation. This approach will allow us to make precise both the notion of evolutionary variability and its role in parameter estimation and will allow us to find the approximate confidence intervals for estimates of parameter values. We will show that, under many conditions, including those that arise in practice, relatively accurate estimates of recombination (or mu-
tation) rates can be obtained and that those estimates are often nearly independent of the time of origin of the population sampled. We will also show that in some cases there is little information about the time of origin of an allele. Although high levels of linkage disequilibrium are usually interpreted as evidence of recent origin, we find that it may be impossible to reject the hypothesis that the allele arose at a time indefinitely long ago.

We begin by reviewing several existing estimators and their statistical properties, then develop the theory underlying our likelihood method, and finally apply our method to three cases, DTD (Hästbacka et al. 1992), ITD (Risch et al. 1995), and the CD4 locus (Tishkoff et al. 1996).

Existing Methods of Parameter Estimation

Throughout, we will be concerned with a locus at which a mutant allele $M$ arose $t_1$ generations ago, a locus that is linked to a marker locus with two alleles, $A_1$ and $A_2$. The data consist of a sample of $i$ chromosomes carrying $M$, with $Y_i$ of the $i$ chromosomes carrying $A_1$. The number $Y_i$, which can take values from 0 to $i$, is the observed configuration of haplotypes in the sample. The statistical problem is to use $i$, $Y_i$, and information about the population from which the sample was drawn, to estimate either $t_i$, the time of origin of $M$; $c$, the recombination rate between $M$ and the marker locus; or $\mu$, the mutation rate at the marker locus.

The history of the $i$ $M$-bearing chromosomes can be represented by a gene genealogy, as shown in figure 1a. The age of the most recent common ancestor is $t_2$, also called the “age of the root” of the genealogy. The genealogy is characterized by the ancestor-descendent (branching) relationships among chromosomes and by the times of the branching events (nodes), which we will denote by $t_1$, . . . , $t_i$ and will call the “coalescence times.” If a continuous time approximation is used to describe the genealogy of the mutant class (see Slatkin and Rannala 1997), then only two branches arise at each node, and there are exactly $i - 1$ nodes.

The definition of the age of an allele is problematic. In our notation, the mutation $M$ arose $t_1$ generations ago but was present in only one ancestral lineage during the time interval $(t_1, t_2)$. Therefore, $t_1$ is the age of the allele, whereas $t_2$ is the age of the root of the allelic-gene genealogy. Any recombination or mutation events during the interval $(t_1, t_2)$ only change the identity of the marker allele that is on the $M$-bearing chromosome immediately before $t_2$, and that allele is the one that is initially in perfect linkage disequilibrium with $M$. Therefore, it would appear that using the configuration of haplotypes in a sample would necessarily result in an estimate of $t_2$; but that is true only if the estimate is based on the observed configuration alone, as is the case with the method-of-moments estimator discussed below. In contrast, if an estimate is based on an explicit population-genetic model, then it is possible to estimate either $t_1$ or $t_2$, because the length of the interval $(t_1, t_2)$ is a random variable whose distribution is determined by the model. Because $t_1$ is the time at which the mutation first arose, it is probably the parameter of greater interest. The age of the root, $t_1$, has a distribution that depends also on the number of chromosomes sampled from the population (see fig. 1b).

A commonly used estimator of $t_2$, $\mu$, or $c$ is what we will call the “moments estimator.” It is based on the expected exponential decay of linkage disequilibrium over time, as a result of mutation and recombination (Lander and Botstein 1986). If the recombination (or mutation process) is irreversible, which, in the case of recombination, implies that $A_1$ occurs at very low frequency on non-$M$ chromosomes, then the probability that any particular lineage has not experienced a recom-
successive doublings are the same. The method thus does not adequately account for stochastic variations in the coalescence times. In their application of the Luria-Delbrück theory to the Finnish population, Hästbacka et al. (1992) assumed that the time of origin of the population was 100 generations in the past and that a single copy of a disease mutation existed at that time, so that \( t = 100 \). The recombination rate between the disease-associated locus and a marker locus was the unknown parameter to be estimated.

Kaplan et al. (1995) developed a likelihood method for the estimation of recombination rates in the context of disequilibrium mapping. Their approach is similar to ours, in that it assumes that the M-bearing chromosomes all descend from a single ancestral chromosome and that their numbers can be modeled by a stochastic process. Kaplan et al. used a discrete-time branching process, whereas we use a continuous-time birth-death process, but the difference is minor, apart from the fact that the continuous-time model is more easily studied by use of analytical methods and leads to more efficient computations. The important difference between our approach and that of Kaplan et al. is the way in which sampling is accounted for. In any study, the \( i \) M-bearing chromosomes in the sample come from a larger number of mutant chromosomes in the whole population. The total number of mutants can be estimated from the frequency of \( M \) and the size of the population from which the sample was drawn. Kaplan et al. accounted for sampling by selecting, among simulated histories of M-bearing chromosomes, those for which the total number of copies in the simulated data were in the range of the estimated total number of copies in the population. From this subset of simulations, they could estimate the probability of finding the observed proportion of \( MA \) chromosomes and, hence, the likelihood. Kaplan et al. (1995) noted that their results were not sensitive to the estimated number of copies in the population. In our analysis, we explicitly model the sampling of M chromosomes so that we are able to obtain the likelihood of the observed configuration of haplotypes directly.

Thompson and Neel (1997) used another approach for modeling the genealogy of a mutant allele. They assumed that the total number of mutant chromosomes is equal to the expected number under a branching process, conditioned on nonextinction. The genealogy of a sample of mutant chromosomes was then modeled by use of a standard coalescent process for a population of variable size. This method provides an approximation to the gene genealogy of the mutants and could provide a potential way to approximate the likelihood of a configuration at a marker locus, although Thompson and Neel (1997) did not address that problem.

Xiong and Guo (1997) also adopt a likelihood approach, but they base it on a different approximation. They assume that the likelihood can be approximated by a quadratic function of the recombination rate, and they use a diffusion method for determining the coefficients of the quadratic function. Xiong and Guo emphasize the computational efficiency of their method and point out that it can be easily generalized to cases with more than two alleles at a marker locus and more than one marker locus. They applied their method to several...
published data sets and found a good agreement between their results and actual map locations of loci that have since been found. One potential weakness of the Xiong and Guo method is that they assumed that the frequency of the mutant chromosome is constant. In most cases, however, that assumption is not valid.

All likelihood methods, including the method of Kaplan et al. (1995), the method of Xiong and Guo (1997), and the one developed here, have similar goals. They provide ways to compute the likelihood of the observed data under different parameter values. In the cases that we have examined, the different methods yield similar results when applied to the problem of disequilibrium mapping. It is less clear whether they would give similar results when applied to the problem of estimating the allele age. Although Kaplan et al.’s (1995) method could, in principle, be used to estimate allele age, computational problems would likely arise, because the time required for each simulation would increase linearly with the assumed allele age, making it difficult to obtain accurate estimates for older alleles. It is not clear whether the approximations made by Thompson and Neel (1997) or Xiong and Guo (1997) would remain valid for arbitrarily large assumed ages. Our method allows us to vary allele age, with no increase in computation time.

Theoretical Analysis

A unique nonrecurrent mutation, \( M \), occurs on a chromosome at a time \( t_1 \) generations in the past. The problem is to use the observed configuration of marker alleles on \( M \)-bearing chromosomes to estimate either the time of origin of the mutation, \( t_1 \), or other parameters, such as the recombination rate and the mutation rate. A complete description of the process that generated the observed configuration has two components: (i) a model of the genealogical process, which describes both the coalescence times and the ancestor-descendent relationships among sampled \( M \)-bearing chromosomes, and (ii) a model of the process of recombination between the mutation \( M \) and the linked markers and of the process of mutation at the linked markers. The genealogy relating a sample of chromosomes descended from a particular nonrecurrent mutant chromosome \( M \) arising at a time \( t_1 \) generations in the past is illustrated in figure 1. The processes of recombination and mutation at linked markers are illustrated in figure 2.

Intraallelic Coalescent

The distribution of the coalescence times for a sample of chromosomes descended from a nonrecurrent mutation, \( M \), is derived in Appendix B and depends on three parameters: (1) \( \xi \), which incorporates the combined effects of population growth and selection, if any, affecting individuals heterozygous for \( M \); (2) \( f = n/(2N) \), the fraction of the population sampled, where \( n \) is the total number of chromosomes sampled and \( N \) is the (diploid) population size; and (3) \( i \), the number of \( M \)-bearing chromosomes in the sample. Let \( t = \{t_1, t_1, ..., t_1\} \), where \( t_1 \) is the waiting time until \( i \) sampled \( M \)-bearing chromosomes coalesce to \( j \) ancestral chromosomes. The joint probability density of \( t \) has been given by Slatkin and Rannala (1997), under the assumption that \( M \) is in sufficiently low frequency that its numbers can be modeled by a linear birth-death process.

Recombination and Mutation

A marker locus \( A \) is closely linked to \( M \). The marker has two alleles in the population: \( A_1 \) and \( A_2 \). Let \( \mu \) be the instantaneous mutation rate from \( A_1 \) to \( A_2 \), and let \( \nu \) be the instantaneous mutation rate from \( A_2 \) to \( A_1 \). The instantaneous rate of recombination is \( c \). The frequency of \( A_1 \) among nonmutant chromosomes is \( p \), which is assumed to have remained constant since \( M \) appeared. We consider a diploid model, but we ignore homozygous mutant individuals, since these will occur with negligible frequency (for a rare mutation), focusing exclusively on \( M \)-bearing chromosomes found in heterozygotes. Recombination of \( M \)-bearing chromosomes then always involves exchanges with nonmutant chromosomes.

The probability of a transition (by either mutation or recombination) from an \( MA_1 \) to an \( MA_2 \) chromosome during an interval \( dt \) is \( u dt \), where \( u = \mu + c(1 - p) \). The probability of a transition from an \( MA_1 \) to an \( MA_1 \) chromosome is \( v dt \), where \( v = \nu + cp \). The transition prob-
abilities during an interval of length $t$ may be calculated analytically (see Appendix C):

$$P(t)(qM_1 \to MA_2) = \frac{u}{u+v}[1-e^{-(u+v)t}]$$

and

$$P(t)(MA_2 \to MA_1) = \frac{v}{u+v}[1-e^{-(u+v)t}] .$$

For simplicity, in the derivation that follows, we will model the processes of recombination and mutation on the genealogy moving forward in time—rather than backward, as we have done previously. If we ascend the genealogy of the $M$-bearing chromosomes, moving forward in time from the initial mutant toward the sample of present-day descendants, all existing chromosomes are equally likely to have given rise to the additional chromosome that is generated at each coalescence event. If we look forward in time, a coalescence event is seen to be the birth of an additional mutant lineage that leaves descendants in the sample. By considering all the possible ways in which each coalescence event might have occurred and by calculating the probability of each, we can avoid dealing directly with the branching relationships of the genealogy, yet we can still calculate the probability distribution of ancestral haplotype configurations among mutant chromosomes after each coalescence event in the genealogy.

If we consider the interval between the $(j-1)$th and $j$th coalescence events (moving forward in time), there are $j - 1$ independent lineages, each undergoing a transition process as described above. At the $j$th coalescence event, an additional chromosome is formed by choosing one of the $j - 1$ chromosomes that exist immediately prior to the coalescent event, to duplicate with equal probability, so that

$$P(Y_j | Y_{j-1}) = \left(\frac{Y_j - 1}{j - 1}\right)P(Y_{j-1} | Y_{j-1})$$

$$+ \left(\frac{j - 1 - Y_j}{j - 1}\right)P(Y_j | Y_{j-1}) ,$$

where $Y_j$ denotes the number of $MA_1$ chromosomes immediately after the $j$th coalescence event, where $Y_{j-1}$ denotes the number immediately after the $(j-1)$th coalescence event, and where $P^*(Z|Y_{j-1})$ is the probability the $Z MA_1$ chromosomes exist immediately prior to the $j$th coalescence event, given that $Y_{j-1}$ existed immediately after the $(j-1)$th coalescence event (see below). Suppose that immediately after the $(j-1)$th coalescence event there are $Y_{j-1}$ $MA_1$ chromosomes. The probability that $k$ of these are replaced by $MA_1$ chromosomes immediately prior to the $j$th coalescence event is

$$P(k|Y_{j-1}) = \binom{Y_{j-1}}{k}u\Gamma(1-u\Gamma)^{Y_{j-1}-k} ,$$

where $\Gamma = [1 - e^{-(u+v)t}]/(u + v)$ and where $\Delta t = t_{j-1} - t_j$ is the waiting time between coalescence events $j$ and $j - 1$. If there are $j - 1 - Y_{j-1}$ $MA_1$ chromosomes immediately after the $(j-1)$th coalescence event, the probability that $l$ of these are replaced by $MA_1$ chromosomes immediately prior to the $j$th coalescence event is

$$P(l|Y_{j-1}) = (j - 1 - Y_{j-1})u\Gamma(1-u\Gamma)^{Y_{j-1}-l} .$$

The number of $MA_1$ chromosomes immediately before the $j$th coalescence event is then $Y_{j-1} - l + k$, and the probability that there are $Y_{j-1}$ $MA_1$ chromosomes is

$$P^*(Y_{j-1} | Y_{j-1}) = \sum_k \binom{Y_{j-1}}{k} u\Gamma(1-u\Gamma)^{Y_{j-1}-l}$$

$$\times \left(\frac{j - 1 - Y_{j-1}}{Y_{j-1} - l + k}\right)\times [u\Gamma]^l (1 - u\Gamma)^{Y_{j-1}-l-k}$$

$$\times [v\Gamma]^k (1 - v\Gamma)^{Y_{j-1}-l-k} ,$$

where the range of the sum over $k$ depends on the specific values of $Y_j$ and $Y_{j-1}$.

The probability of any particular configuration of haplotypes, $Y_0$, for the $i$ $M$-bearing chromosomes can be calculated directly by taking a sum over the probabilities for all possible ancestral configurations of haplotypes that might have resulted in the observed haplotype configuration and then integrating over the coalescence times, to obtain

$$P(Y_0 | \Theta_1)$$

$$= \sum_{Y_0} \sum_{Y_1} \cdots \sum_{Y_{i-1}} \sum_{t_1} \sum_{t_2} \cdots \sum_{t_{i-1}} P(Y_0 | Y_i, t_i, \Theta_2)$$

$$\times \prod_{l=2}^i P(Y_l | Y_{l-1}, t_i, \Theta_3)$$

$$\times P(Y_i | p) \times P(t | \Theta_4) dt_1 \cdots dt_2 ,$$

where $\Theta_1 = \{u, v, p, t_1, i, f, \xi\}$, $\Theta_2 = \{u, v, i\}$, $\Theta_3 = \{u, v, t_1, i\}$, and $\Theta_4 = \{i, f, \xi, t_i\}$. There are $l$ terms in the above sum, each of which includes an $(i - 2)$-dimensional integral, so it is not practical to evaluate the right-hand side of this equation explicitly for a sample of more than $\sim 10$ mutant chromosomes. An alternative method
for evaluating this equation is to use Monte Carlo integration.

**Monte Carlo Integration**

A Monte Carlo estimator of \(P(Y_0|\Theta_i)\) is obtained as

\[
P(Y_0|\Theta_i) \approx \frac{1}{R} \sum_{r=1}^{R} P(Y_0|\bar{Y}_r(k),\bar{h}(k);u,v) ,
\]

where the sum is evaluated over \(R\) replicate Monte Carlo realizations of the random variables \(\bar{Y}_r(k)\) and \(\bar{h}(k)\). Note that \(\bar{Y}_r(k)\) is the \(k\)th of \(R\) independent random variables generated from the distribution

\[
P(Y_r|t;\Theta_j,p) = \prod_{i=2}^{10} P(Y_i|Y_{i-1},t;\Theta_j)P(Y_1|p) .
\]

Random variables are generated from this distribution by sequential simulation from the conditional probability distributions, beginning with configuration \(Y_1\) and ending with configuration \(Y_n\) as described below. If \(p < 1\), then \(Y_2 = 2\) with probability \(pe^{-\xi + \eta/2} + 1 - e^{-\xi + \eta/2}\) and otherwise takes the value \(Y_2 = 0\). To simulate \(\bar{Y}(k)\) for \(2 < j \leq i\), we first simulate two independent binomial random variables, \(h\) and \(l\), from the binomial distributions of equations (1) and (2), respectively, conditional on the variate \(\bar{Y}_j(k)\) generated at the previous step. The random variable \(\bar{Y}_j(k)\) is then obtained as \(\bar{Y}_j(k) = \bar{Y}_{j-1}(k) + l - h\). The vector of random variables \(\bar{h}(k)\) is generated for each replicate of the Monte Carlo integration by means of the simulation method described by Slatkin and Rannala (1997).

**Genealogy and the Age of a Mutation**

There are limits on the allele ages that may be estimated by use of genealogical information. As the time \(t_1\) at which mutation \(M\) arose increases, the distribution of the coalescence times converges to a stationary distribution that no longer depends on \(t_1\) (Rannala 1997). If \(t_1\) is sufficiently large, the distribution of coalescence times cannot be distinguished from the limiting distribution (i.e., the density obtained in the limit as \(t_1 \to \infty\)). When this is the case, the data no longer contain any information about the parameter \(t_1\), even though there may still be considerable linkage disequilibrium. A likelihood-ratio test (LRT) can be used to detect such cases and to determine whether \(t_1\) differs significantly from infinity, where the likelihood ratio is

\[
\Lambda = |P(Y_0|t_1 \to \infty)|/|P(Y_0|\hat{t}_1)| .
\]

The hypotheses are nested so that \(-2\log\Lambda\) is approximately \(\chi^2\) distributed with 1 df. Because \(t_1 \to \infty\) is a boundary condition, it is not clear that the \(\chi^2\) approximation is justified in this case. An alternative approach is to simulate observations under the null hypothesis that \(t_1 \to \infty\) and to generate the distribution of the statistic. Given the demographic parameters \(\xi\) and \(f\), it may be possible to predict a priori the range of values for which \(t_1\) may be estimated, although we do not pursue that question in this paper. Numerical analyses suggest that, for a fixed sample size of \(i\) copies of the mutant allele, an increase in \(\xi\) or \(f\) reduces the maximum age that may be inferred for the mutant.

**Applications**

**DTD in Finland**

DTD is an autosomal recessive disorder that has disease-associated chromosomes present at a frequency of \(\sim 0.8\%\) in Finland. Hästbacka et al. (1992) measured the extent of linkage disequilibrium between disease-associated chromosomes and several markers located within the CSF1R locus on chromosome 5q, to estimate that the locus carrying alleles causing DTD is \(\sim 62\) kb from CSF1R. Subsequently, by positional cloning, Hästbacka et al. (1994) found the locus to be \(\sim 70\) kb from CSF1R. We can use the Hästbacka et al. (1992) data to illustrate our method.

In the Hästbacka et al. (1992) data set, there are 146 DTD chromosomes, and, of those, all but 7 had an EcoRI restriction site and a StyI restriction site (the 1–1 haplotype) within the CSF1R locus. Thus, in our notation, \(i = 146\) and \(Y_0 = 139\). We do not know \(n\), the total sample size, because the DTD chromosomes were not obtained in a population survey. This situation is typical for disease-associated alleles, because chromosomes are not sampled randomly from the population but instead are sampled disproportionately both from individuals treated for the disease and from their close relatives. By using the estimated frequency, 0.8%, we can estimate how large \(n\) would have to be for us to expect to obtain 146 disease chromosomes (on average) in a random sample of chromosomes. To find 146 DTD chromosomes in a sample of size \(n\), 146/\(n\) = 0.008, so on average, \(n = 18,250\). Given the current population size of Finland, 5 million (\(N = 5 \times 10^6\)), this hypothetical sample represents a proportion \(f = n/2N = 1.825 \times 10^{-3}\) of the population. Following Hästbacka et al. (1992), we assume that the Finnish population was founded by 1,000 individuals \(\sim 2,000\) years, or \(t_1 = 100\) generations ago, and that the population has grown exponentially since that time, at a rate of \(\xi = .085\). In the initial population, there was a single mutant copy of the DTD allele on a chromosome carrying the 1–1 haplotype at the marker locus. That haplotype occurs with a frequency of 3% on the nondisease chromosomes (\(p_1 = .03\)). In the Hästbacka et al. (1992) data set, there
is one 1–2 haplotype associated with the DTD chromosome, but we will ignore that haplotype and will treat
the two restriction sites as a single marker locus. We assume that the mutation rates at the marker locus are
0, and hence, in our notation, \( u = c(1 - p) \) and \( v = cp \), where \( c \) is the unknown recombination rate.

We have used these parameter values and the Monte Carlo integration method described above to compute
the likelihood of the observed haplotype configuration as a function of \( c \), the recombination rate between
the marker and disease locus. Figure 3 shows the log likelihood's dependence on \( c \). The maximum-likelihood estimate of \( c \) is \( 0.0008 \), and the 95% confidence interval is \( 0.0002-0.0016 \). This estimate of \( c \) is slightly higher than
\( 0.0006 \), the estimate given by Hästbacka et al. (1992), and our confidence interval is also larger than theirs
(which was \( 0.0005-0.0009 \)). For humans, a map distance of 1 cM (i.e., a recombination rate of \( 0.01 \)) corresponds
to a physical distance of \( \sim 1 \) Mb. The estimated recombination rate \( 0.0008 \) therefore corresponds to a physical
distance, between the marker and the mutation, that is \( \sim 80 \) kb, which is in close agreement with the actual
physical distance, now known to be \( \sim 70 \) kb. Kaplan et al. (1995) used a simulation approach to calculate the
likelihood for these data and also found a wider confidence interval than did Hästbacka et al. (1992). Kaplan et al.
computed the upper bound on the estimated recombination rate to be \( 0.022 \).

Because of the high growth rate, estimates of \( c \) are insensitive to \( t \). Numerical analyses suggest that, even
if the time of the founding event for the Finnish population was much more than 100 generations ago, the
genealogy and the resulting MLE of \( c \) would be roughly the same.

The frequency of the mutation \( M \) in the population also carries information about \( t \). (Slatkin and Rannala
1997). The MLE of \( t \) for the Hästbacka et al. (1992) data, on the basis of frequency, is \( \hat{t} = 111.9 \) generations,
with a 95% confidence interval of 94.1–146.7, when equation (8) of Slatkin and Rannala (1997) is used. This
is quite close to the Finnish population's estimated age based on our knowledge of its demographic history.

**ITD in Ashkenazi Jews**

ITD is a movement disorder with variable clinical effects. It is relatively frequent in Ashkenazi Jews and is
best modeled as being attributable to a dominant allele with \( \sim 30\% \) penetrance (Risch et al. 1995). The locus
carrying this allele has been mapped to 9q34. Risch et al. (1995) have shown that the disease is in very strong
disequilibrium with several microsatellite marker loci, including loci that are separated by as much as 4 cM.
These observations have suggested that most cases of ITD in Ashkenazi Jews are caused by a single allele of
relatively recent origin. Risch et al. (1995) used these observations to estimate the age of the mutation as being
8–22 generations.

In our analysis, we assume that the current population size of Ashkenazi Jews is \( N = 5 \times 10^4 \) and use Risch et al.’s (1995) estimate of the frequency of the allele causing ITD in Ashkenazi Jews to be \( 1/6000 \). We will analyze
54 ITD haplotypes reported by Risch et al. (1995). The estimated value of the total sample size \( n \), the total sample
size, \( 54 \times 6,000 = 324,000 \) and \( f = n/(2N) = 0.0324 \). Historical records suggest that the rate of pop-
ulation increase has been \( \sim 1.5 \) fold per generation (Risch et al. 1995), which implies a growth rate of \( \xi = \ln(1.5) = 0.4055 \).

Risch et al. (1995, table 4) have reported the haplotypes of 59 chromosomes associated with ITD. Of these,
54 have been identified as carrying the same mutant allele for ITD, because they have the same or nearly the
same haplotypes at three marker loci (D9S62a, D9S62b, and D9S63) that previous analysis showed are very
closely linked and that are only slightly distal to the presumptive disease locus (which they call “DYT1”).
Two other highly polymorphic markers (ASS and D9S64) are in the same region. The locus ASS is estimated
to have a recombination rate of \( c_1 = 0.018 \) with D9S63, with a confidence interval of \( 0.007–0.033 \), and a
recombination rate of \( c_2 = 0.023 \) with D9S64, with a

$$\begin{align*}
\text{log likelihood} & = -\ln(1.5) \\
& = -0.4055
\end{align*}$$

Figure 3 Log likelihood (\( \ell \)) of the data, analyzed by Hästbacka et al. (1992), for Styl marker haplotypes of chromosomes associated with ITD in the Finnish population, as a function of the recombi-
nation rate with the marker (see text). The maximum-likelihood estimate of the recombination rate \( c \), as well as the 95% confidence interval of the estimate (i.e., the values that are \( \leq 2 \) units of log like-
elihood below the maximum-likelihood estimate), are indicated by ar-
rows. The MLE is \( \hat{c} = 0.0008 \), with the 95% confidence interval
\( 0.0002–0.0016 \).
confidence interval of .012–.042 (Risch et al. 1995). For both markers, \(i = 54\). One allele of ASS, allele 12, is found on \(Y_0 = 47\) chromosomes, and its frequency on nondisease chromosomes, \(p_1\), is .086. (The subscripts for \(p\), \(u\), and \(v\) indicate the locus, ASS or D9S64.) For D9S64, two alleles are in relatively high frequency: allele 10 is found on 16 chromosomes, and allele 2 is found on 18 chromosomes; Risch et al. combined these into a single ancestral category. For ASS, \(u_1 = .018 \times (1 - .086) = .0164\) and \(v_1 = .018 \times .086 = .001548\), if we assume that mutation can be ignored. For D9S64, we assumed that there was no interference of recombination, so \(c_{ASS}\), the recombination rate with \(DYT1\), is \(.018 + .023 - .018 \times .023 = .0406\). Hence \(u_2 = .035366\) and \(v_2 = .0050034\), when allele 2 and allele 10 are combined, and \(u_2 = .03788\) and \(v_2 = .000272\), when allele 2 is assumed to be ancestral. The extensive polymorphism on many disease chromosomes at this locus suggests that the mutation rate might be high enough to be important. Assuming a mutation rate of 0 makes the values of \(u_2\) and \(v_2\) minimum values, and any increase in those parameters would tend to reduce estimates of \(t_1\) or \(t_2\).

We can use our method to find the likelihoods for any values of \(t_1\), the time of origin of the mutation, or \(t_2\), the age of the root of the mutant genealogy, by treating data at each marker locus separately. In the first case, we use the intraallelic genealogy obtained by conditioning on \(t_1\) and in the second case we use the intraallelic genealogy obtained by conditioning on \(t_2\). Figure 4 shows the results for the ASS locus. We can see that, for shorter times, the likelihood functions for \(t_1\) and \(t_2\) are nearly the same. For both, the MLE is \(\sim10\) generations, and the lower limit of the 95% confidence interval is \(\sim2\) generations. The similarity of these two curves reflects the fact that, when \(t_1\) is small, the first branching of the mutant-gene genealogy will occur very soon after the mutant arises. For larger values of \(t_1\), the difference between \(t_2\) and \(t_1\) increases, indicating that there is a longer time during which there is a single lineage ancestral to all copies of the mutant in the sample. As a consequence, the difference between the likelihoods will increase with \(t_1\). We can see, though, that, whether the goal is to estimate \(t_1\) or \(t_2\), the upper limit of the confidence interval is relatively large, \(\sim72\) generations for \(t_1\) and \(\sim131\) generations for \(t_2\). We found similar results for the data from D9S64. When alleles 2 and 10 are combined, as in the Risch et al. (1995) analysis, the MLE of both \(t_1\) and \(t_2\) is 14 generations, with a lower limit, for the confidence interval, of 5 generations. The upper limit is 51 generations for \(t_2\) and 71 for \(t_1\). For allele 2 alone, the MLE of both \(t_1\) and \(t_2\) is \(\sim25\) generations, with upper confidence limits of \(\sim100\) and \(\sim80\) generations, respectively. On the basis of the observed frequency of mutant chromosomes in the population and according to equation (8) of Slatkin and Rannala (1997), with \(\xi = .4\), the MLE of \(t_1\) is 17.8 (13.9–25.1).

Our estimates of allele age are similar to those reported by Risch et al. (1995), but we find that the data are consistent with a much wider range of allele ages than was suggested by Risch et al. (1995). It is illustrative to examine in more detail the reasons that we obtained such broad confidence intervals. There are potential two sources of error for the moments estimator of \(t_2\). First, the time at which the original mutant \(M\) arose may be far enough in the past that the genealogy is nearly independent of the parameter \(t_1\), as appears to be the case for \(DYT1\). Second, there may be large errors associated with estimates of \(t_1\) or \(t_2\), because of the large variance among the observed sample configurations of haplotypes even when \(t_1\) is recent. As mentioned above, the moments estimator is an MLE of \(t_2\) if the intraallelic genealogy is starlike, but that will not generally be the case for realistic demographic models. We have examined the effects that genealogy has on the distribution of the moments estimator, by using simulations. The large effect that genealogy can have on the distribution of the mo-
ments estimator and its mean-square error is illustrated in figure 5.

To generate the frequency distribution of the moments estimator of $t_2$, shown in figure 5a, we have assumed a star genealogy and have simulated 1,000 artificial configurations of haplotypes, using $t_2 = 8.4$, which is the value that Risch et al. (1995) obtained from the moments estimator. The other parameters for these simulations were the same as those used by Risch et al. (1995). To generate the frequency distribution of the moments estimator shown in figure 5b, we simulated 1,000 artificial configurations of haplotypes, using $t_1 = 8.4$ generations, but we instead used the distribution of genealogies and coalescence times expected on the basis of the inferred demography for this population and used the population-growth rate of $\xi = .4$ suggested by Risch et al. (1995). The distribution for the simulated data when a star genealogy is assumed is unimodal and is closely concentrated around the average, as shown in figure 5a. In such cases, quite accurate estimates are possible: the average of the moments estimator is 7.70, and the mean-square error (MSE; the average squared difference between the inferred value of the parameter and its true value, 8.4) is 9.09. Allowing the genealogy to be determined by the population demography in the simulations results in a distinctly bimodal distribution of the statistic and in a large associated error for the estimates: the average is 12.0, and the MSE is 1,365.3, indicating that the moments estimator of $t_2$ is not very accurate.

**Age of Non-African Lineages at CD4**

Tishkoff et al. (1996) surveyed several populations from different geographic regions for two very closely linked polymorphic markers within the CD4 locus on chromosome 12: an Alu element, in the first intron, that is polymorphic for a 265-bp deletion, and a polymorphic pentanucleotide short tandem repeat polymorphism (STRP), or microsatellite, locus 5′ to exon 1 and 9.8 kb from the Alu polymorphism. The Alu deletion (called the “Alu(−) allele”) is not found in gorillas and chimpanzees and thus appears to have arisen after the separation of the lineage leading to modern humans, ~5 million years ago. Tishkoff et al. (1996) argue that Alu(−) arose only once, on a chromosome carrying a 90-bp allele at the STRP marker, and that variability at the STRP marker on the Alu(−) chromosomes is the result of later mutation and recombination.

Tishkoff et al. (1996) divided their study populations into two groups, sub-Saharan Africans (population A) and non-Africans (population B). They found that 97.8% of the 270 non-African Alu(−) chromosomes carried the 90-bp allele, whereas only 25.8% of the 132 African chromosomes carried that allele. Both the mutation rate at the STRP locus and the recombination rate between the two markers are unknown. Tishkoff et al. (1996, p. 1384) assumed a mutation rate $\mu$ from the 90-bp allele to any other allele, no back mutation, and a 0 rate of recombination between the STRP and Alu mark-
ers. Furthermore, they excluded from their analysis STRP alleles $\geq 110$ bp, because those alleles represent the descendants of an ancient recombination event with an Alu(+) chromosome. There was one such allele (size 110 bp) among the non-Africans, and there were 47 (sizes $\geq 115$ bp) among the Africans. Tishkoff et al. (1996) then assumed a Poisson mutation process, with mutation rate $\mu$, and used the formula $Q = e^{-\mu t}$ to express the proportion of ancestral (90-bp) alleles in a sample, as a function of $t$, the time of origin of each group. Tishkoff et al. (1996) used “N” for the number of generations, but we use $t$ here, to be consistent with our notation. There are two values of $Q$: $Q_A = 34/85 = 0.40$ is the proportion of 90-bp allele on the Alu(−) chromosomes in Africans, and $Q_n = 264/269 = 0.9814$ is the proportion in non-Africans, where we use “Q” instead of the “P” of Tishkoff et al. (1996). Tishkoff et al. (1996) assumed (1) that Alu(−) chromosomes in non-Africans are descended from a single ancestral lineage that diverged from the African lineages at a time $t_n$ that is approximately the time at which ancestors of non-Africans left Africa and (2) that the Alu(−) allele arose by mutation $\leq t_A = 5$ million years ago. Under the assumption of exponential decay of the $Q$'s, the ratio $\ln(Q_A)/\ln(Q_n) = t_A/t_n = 48.8$ gives a value of $t_n$ that is $\sim 100,000$ years before present if $t_A = 5$ million years. Note that this calculation does not depend on either the generation time or the mutation rate, as long as each has remained constant during the time of interest. In this calculation, equating the expected values of $Q$ with those observed is equivalent to using the moments estimator of $t$ for each sample separately.

Because Tishkoff et al. (1996) used the moments estimator for allele ages in both populations, their estimates are necessarily of $t$. For both populations, the values of $t_i$ will be older by an amount that depends on the demographic history of each population. Assuming, as they did, that $t_2$ for the age of the Alu(−) allele in Africans is 5 million years implies that this allele actually arose in the common ancestor of chimpanzees and humans but survived only in the lineage leading to humans.

We can use our maximum-likelihood method to estimate $t_B$, the time of origin of non-Africans, by jointly estimating $\mu$ and $t_B$, using the configurations of both African and non-African Alu(−) chromosomes. We assume that Africans and non-Africans have evolved independently since the dispersal of the ancestors of non-Africans from Africa. We can then generate the likelihood functions for each and then multiply these to obtain the overall likelihood of the observations, as a function of the unknown parameters, $\mu$ and $t_B$. For all combinations of parameters that we considered, the same MLE of $t_B$ was obtained whether the likelihood was maximized jointly for $t_B$ and $\mu$ or whether $\mu$ was first estimated by use of only the African samples and then that estimate was substituted into the likelihood to estimate $t_B$ for non-Africans. For ease of discussion, we will treat the two estimation problems separately.

For our analysis, we assume that the current population size for Africans is $N_A = 10^9$ and that the population size for non-Africans is $N_n = 4 \times 10^9$. It is reasonable to assume a much lower rate of population growth for Africans than for non-Africans. We assumed initially that the Africans have descended from a population of 100 individuals that grew exponentially to the current size, which implies $t_A = 6.447 \times 10^5$. We found that the exact value of $\xi$ did not matter, and, in fact, almost the same estimates of $\mu$ were obtained when we assumed constant population sizes of $10^7$, $10^8$, and $10^9$. A later start for growth of the African population, followed by a consequently higher growth rate, did matter, as we shall show. For the non-Africans, we used both $\xi = 0.002$ and $\xi = 0.005$ in our analysis. The first value is the growth rate if the non-Africans did, in fact, arise from a single small population that grew exponentially for 200,000 years (10,000 generations); the larger value of $\xi$ allows for a later initiation of exponential growth.

If the STRP alleles $\geq 110$ bp are ignored, $n_A = 349$, $n_n = 1424$, $i_A = 85$, $i_n = 269$, and $Y_{A,A} = 264$. These numbers are obtained from table 3 of Tishkoff et al. (1996), by removal of all data with allele sizes $\geq 110$ bp at the STRP locus. The sampling fractions are $f_A = 349/(2 \times 10^4) = 1.745 \times 10^{-7}$ and $f_n = 1.424/(8 \times 10^3) = 1.78 \times 10^{-7}$.

Figure 6 shows the log likelihood as a function of $\mu$ for the African samples. The MLE of $\mu$ is $5.88 \times 10^{-6}$ ($2.13 \times 10^{-6}$, $9.63 \times 10^{-6}$) if we assume that $t_1$ is $2.5 \times 10^5$ generations. If we instead assume that $t_1 = 2.5 \times 10^5$ generations, the results are almost the same; the MLE of $\mu$ is $4.91 \times 10^{-6}$ ($1.20 \times 10^{-5}$, $8.63 \times 10^{-6}$). Figure 7 shows the log likelihood as a function of either $t_1$ or $t_2$ for the non-African samples, with growth rate $\xi_n = 0.002$ and mutation rate $\mu = 5.88 \times 10^{-6}$. Note that the results are similar to those for ITD that are shown in figure 4, in that, for shorter time periods, the likelihoods are virtually identical functions of $t_1$ or $t_2$. The MLE is $\sim 4,100$ generations (82,000 years), and the lower confidence limit is slightly less than 1,000 generations. The upper confidence limit depends on whether $t_1$ or $t_2$ is being estimated. There is no upper confidence limit for $t_1$, whereas the upper confidence limit for $t_2$ is slightly more than 100,000 generations, or 2 million years. It is arguable whether $t_1$ or $t_2$ is more relevant in this case, but which time is estimated does not affect the main conclusion—namely, that these data provide little reason to reject even very early dates for the divergence of non-Africans from Africans. Similar results were obtained for the cases with $\xi_n = 0.005$. In general, larger growth rates make it even more difficult to estimate allele ages confidently.
The observed frequency of the Alu(−) allele also contains information about \( t_m \), if the population growth rate is known. With \( \xi_A = .002 \) the MLE based on the allele frequency, when equation (8) of Slatkin and Rannala (1997) is used, is \( t_m = 7,139.2 \) \( (6,882.1, 10,019.2) \) generations (or 142,785 years).

It is reasonable to ask how sensitive our results are to changes in the various parameters that we have assumed. The assumed population sizes of African and non-Africans, which determined the values of \( f_A \) and \( f_B \), make essentially no difference. For example, if we increase \( f_A \) by a factor of 10, meaning that the current population size of Africans is \( 10^8 \) instead of \( 10^9 \), the maximum-likelihood estimate of \( \mu \) increases only slightly, to \( 6.9 \times 10^{-6} \), with the confidence interval \( 1.7 \times 10^{-6} \sim 1.2 \times 10^{-5} \). The growth rate of the African population is more important. For example, if we assume that the population ancestral to the modern sub-Saharan African populations was small until 500,000 years ago (i.e., \( t_A = 2.5 \times 10^4 \), rather than \( 2.5 \times 10^5 \), generations) and then began to grow exponentially to a current size of \( 10^9 \), \( \xi_A \) would then be \( 6.447 \times 10^{-4} \), 10 times larger than the value used to generate figure 7. The resulting MLE of \( \mu \) is \( 4.63 \times 10^{-5} \), almost 10 times larger than the estimate obtained from figure 6. The effect of a larger mutation rate on the estimated age of the non-African populations is dramatic. With \( \mu = 4.63 \times 10^{-5} \), the maximum-likelihood estimate of \( t_m \) is \( \sim 600 \) generations (or 12,000 years), with a confidence interval of \( \sim 100\text{--}1,000 \) generations, for \( \xi_A = .002 \). This estimate is much too recent to be consistent with the fossil and archaeological record of modern humans.

Our conclusion about the Tishkoff et al. (1996) data agrees with that of Pritchard and Feldman (1996), who argued that their simulation study showed that the data indicated a much larger upper bound for the confidence interval on \( t_B \). Our reasons, however, are quite different. Pritchard and Feldman (1996) argue that uncertainly in the estimate of \( \mu \) results in greater uncertainty in the estimate of \( t_B \). We agree that there is some uncertainty in the estimate of \( \mu \) (see fig. 6), although we do not agree with their method for finding the confidence interval on the estimate of \( \mu \). Our point is that, even if the mutation rate were known, the variability in gene genealogies alone would result in reduced confidence in estimates of \( t_B \). Our method can be easily extended to take account of the uncertainty in \( \mu, f_A, \xi_A \) or other parameters, by integrating over assumed prior distributions. All that would be necessary would be to draw random parameter values from the appropriate distributions for each replicate of the Monte Carlo integration described above.

**Discussion**

We have presented a stochastic model that predicts the distribution of coalescence times in the intraallelic gene genealogy of an allele that arose once by mutation. The key assumption is that the allele of interest is and has been sufficiently rare that each copy in the population reproduces independently of each other copy (see Appendix B). We have shown that the distribution of coalescence times depends on both the population growth rate and the fraction of the population represented in the sample. From the distribution of coalescence times, a model of recombination and mutation at a linked marker locus, and the assumption that all copies of the mutant are equivalent, we obtain the likelihood of a configuration, \( Y_m \), of alleles at a marker locus. That likelihood function, which we can estimate to an arbitrary degree of accuracy by using a Monte Carlo procedure, provides the basis for estimating the recombination rate between a marker and a novel mutant locus, the mutation rate at a marker locus, or the time of origin of the mutant.

For the purpose of disequilibrium mapping, our results are similar to those of Kaplan et al. (1995) and Xiong and Guo (1997), in that we compute the likelihood of an observed configuration of linked markers. Our method has the advantage that, unlike the Kaplan et al. (1995) method, it does not require extensive simulations and that, unlike the Xiong and Guo (1977)
Figure 7  Log likelihood ($\ell$) of the age of the Alu(−) allele in non-Africans in the Tishkoff et al. (1996) data set. The upper curve is for $t_1$, the earliest time at which the allele ancestral to all copies of Alu(−) in non-Africans formed a distinct lineage, and the lower curve is for $t_2$, the age of the most recent common ancestor in the gene genealogy of Alu(−) alleles in non-Africans. For these curves, $t_0 = 269$, $Y_{0} = 264$, $\xi_0 = 0.002$, $\Delta = 1.78 \times 10^{-7}$, and $\mu = 5.88 \times 10^{-6}$ (the MLE from fig. 6).

method, it does not rely on approximations that are difficult to verify independently. The only approximation that we use is to assume that the number of copies of a low-frequency allele can be modeled by a linear birth-death process. Nevertheless, all the likelihood methods produce results that are similar and that, for practical purposes, may well be equivalent. All the methods also require assumptions about the demographic history of the sample population; and those assumptions are probably not true. Different likelihood methods will give slightly different results for the same data set and demographic parameter values, but uncertainty about the sample population is probably a greater source of error than are any of the mathematical assumptions or approximations that are made.

Because of the uncertainty about demographic parameters, methods that do not seem to require as many assumptions—such as the method of moments and the Hästbacka et al. (1992) method, which relies on the Luria-Delbrück theory—may seem preferable, on intuitive grounds. Those methods, however, make implicit assumptions about the gene genealogy of the mutant allele class; and those assumptions are more restrictive than the assumptions made in the likelihood methods. The method of moments implicitly assumes a star genealogy, and the Luria-Delbrück theory assumes a genealogy with synchronous symmetrical branching. Those methods may provide adequate answers in some cases, as the Luria-Delbrück theory did for the mapping of DTD in the Finnish population, but likelihood methods should work in a much wider variety of situations. Freely distributed computer programs, including ours, make the use of likelihood methods for disequilibrium mapping relatively easy. Even when demographic parameters are not known with confidence, likelihood methods provide a framework within which to explore the sensitivity of the results to different parameter values.

Because the likelihood method that we have developed uses Monte Carlo integration to evaluate the probability of the observed data, it is easy to take account of uncertainties in estimates of parameters such as population size or mutation rate, by assigning to these parameters a prior distribution that reflects these uncertainties and then integrating over this prior during the Monte Carlo integration.

For the purpose of estimating the ages of alleles, there may be greater differences between the likelihood methods discussed, although most of the other methods have not been applied to the problem of the estimation of allele age (but see Guo and Xiong 1997). The approach of Kaplan et al. (1995) is similar to ours but does not lend itself to the development of analytic theory pointing to the intrinsic limitations of estimating allele age, and it would lead to computational problems for old alleles. The method of Xiong and Guo (1997) is based on an approximation that may not remain valid when the time of appearance of the mutant is allowed to vary.

Population Subdivision

All of the likelihood methods discussed here, including ours, assume a single randomly mating population. There is probably some subdivision even in relatively homogeneous populations, including the population of Finland. For other populations, including Ashkenazi Jews and sub-Saharan Africans, the assumption of no subdivision is even less likely to be true. Models based on either a branching process or a birth-death process are somewhat insensitive to population subdivision, because they assume that each copy of the mutant allele, M, reproduces independently. If there is population subdivision, however, our method and those of Kaplan et al. (1995) and Xiong and Guo (1997) require the implicit assumption that each subpopulation grows at the same rate and that the same fraction, $f$, of each subpopulation is sampled. Further work is needed to quantify the effects of population subdivision when these implicit assumptions are invalid.

Conclusions

Although it is difficult to generalize from our results, because of the many parameters in our model, the three examples that we have analyzed, as well as other cases
that we have examined in the course of this study, suggest that likelihood methods are quite robust for the estimation of recombination rates, which is the goal of disequilibrium mapping (Lander and Botstein 1986). Estimates of $c$, such as those for the data of Hästbacka et al. (1992), are relatively insensitive to values of $f$ (i.e., the sampling fraction), $\xi$ (i.e., the population growth rate), and, most important, to $t_1$ (the time of founding of the population). For the Hästbacka et al. (1992) dataset, we found that it made little difference, in the estimate of $c$, whether we assumed $t_1 = 100$ or much larger values of $t_1$. In contrast, it can be difficult to accurately estimate $t_1$, the time at which the mutant arose, when $t_1$ is moderately large, particularly if the population has been growing rapidly. When there has been rapid growth, the coalescence times will be relatively recent, even though the mutant may have arisen in the distant past. As a consequence, substantial disequilibrium is expected in the data, regardless of allele age, and the confidence interval on estimates of allele age can therefore be quite broad. Surprisingly, allele frequency alone, when the recombination rate, $\theta$, and, most important, to $t_1$ are used, can solve for the limitations of disequilibrium analysis for estimation of allele age nor its robustness for estimation of recombination rates appears to result from special simplifying assumptions in our analysis. Rather, these results appear to arise from intrinsic properties of intraallelic genealogies in growing populations.

**Program Availability**

The method described in this paper for calculation of the log likelihoods of different haplotype configurations has been implemented as a computer program written in the C language. The program DMLE is available either by anonymous ftp at mw511.biol.berkeley.edu or on the World Wide Web at http://mw511.biol.berkeley.edu/software.html.

**Acknowledgments**

This research was supported in part by a Natural Sciences and Engineering Research Council of Canada postdoctoral fellowship to B.R. and by National Institutes of Health grant GM-40282 to M.S. We thank J. Pritchard for providing details of his simulation program, and we thank A. G. Clark, J. Pritchard, N. Risch, S. A. Tishkoff, E. A. Thompson, and Z. Yang for helpful comments on an earlier version of this paper.

**Appendix A**

**Statistical Properties of the Moments Estimator of Allele Age**

In this appendix, we show that the estimator

$$\hat{t}_2 = \frac{\log(i) - \log(i - Y_0)}{u},$$

obtained by setting the observed fraction of nonmutated markers $(i - Y_0)/i$ equal to the expected proportion $\exp(-u t_2)$ and solving for $t_2$, is an MLE if the genealogy of the sample is a star genealogy. We use the fact that, for a rare mutation, essentially all recombination events involve individuals heterozygous for $M$. Because recombination events on each branch are independent in a star genealogy,

$$L(q|Y_0) = \left(\frac{i}{Y_0}\right) q^{Y_0} (1 - q)^{i - Y_0},$$

where $L(q|Y_0)$ denotes the likelihood function and where $q = 1 - e^{-\mu t}$. The MLE of $q$ is then $\hat{q} = \hat{q}$. By the invariance property of MLEs (see Casella and Berger 1990), equation (A1) provides an MLE of $t_2$, since we can solve for $t_2$ as a function of the MLE $\hat{q}$ when $\mu$ is known. The approximate asymptotic variance for the MLE of $t_2$ is

$$\text{Var}(\hat{t}_2) \approx \frac{\left(\frac{\hat{q}^2}{\hat{q}^{2 - 1}} \frac{-\log(1 - \hat{q})}{\hat{q}^{2 - 1}}\right)^2}{\hat{q}^{2 - 1} \log L(q|Y_0)_{q=\hat{q}}} 
= \frac{Y_0}{i(i - Y_0)u}.$$
probability that one additional M-bearing chromosome arises during the interval \((t, t + dt)\) and leaves descendants in the sample is

\[
P[j(t - dt) = j(t) + 1] = \sum_{i=1}^{n(0)-j(0)} P[j(t - dt)] = j(t) + 1 | i(t) | P[i(t)] ,
\]

where

\[
P[j(t - dt) = j(t) + 1 | i(t)] = \frac{1}{4N(t)} | i(t) + j(t) | i(t) - 1 | i(t)P[i(t)]dt.
\]

The first three terms of the product on the right-hand side of equation (B1) give the probability that a coalescence event occurs during \(dt\), whereas the last term gives the probability that the coalescence event involves an M-bearing chromosome. The unconditional probability is then

\[
P[j(t - dt) = j(t) + 1] = \sum_{i=1}^{n(0)-j(0)} \frac{1}{4N(t)} | i(t) + j(t) | i(t) - 1 | i(t)P[i(t)]dt.
\]

If the mutant M has remained rare, so that \(i(t) \gg j(t)\) and \(i(t) \equiv n(t)\), then

\[
P[j(t - dt) = j(t) + 1] \approx \frac{j(t)}{4N(t)} \sum_{i=1}^{n(0)-j(0)} i(t)P[i(t)]dt
\]

\[
= \frac{j(t)}{4N(t)} E[i(t)]dt
\]

\[
= \frac{j(t)}{4N(t)} E[n(t)]dt .
\]

For a population of deterministically variable size, an approximate expression for \(E[n(t)]\), when this value is large is (according to Slatkin and Rannala 1997)

\[
E[n(t)] = \frac{n(0)}{1 + n(0)r(t)/2} ,
\]

where \(r(t)\) is time, rescaled to allow for variable population size,

\[
r(t) = \int_0^t dt' / 2N(t').
\]

With a constant population size, the transition probability is then

\[
P[j(t - dt) = j(t) + 1] = \frac{n(0)}{1 + n(0)(r/2)/4N(0)} j(t)dt
\]

\[
= \frac{f/2}{1 + fr} j(t)dt ,
\]

where \(f = n(0)/2N(0)\) is the fraction of chromosomes in the population that are sampled. This is the instantaneous birth rate for the “reconstructed” linear birthdeath process (Nee et al. 1994), which describes the growth of the lineages that ultimately leave descendants in the sample, with \(B = D = 1/2\), where \(B\) is the birth rate, and where \(D\) is the death rate, per lineage (see Slatkin and Rannala 1997).

If the population has experienced exponential growth at rate \(r\), then we use the time transformation \(\tau(t) = (e^r - 1)/[2N(0)r]\) to obtain

\[
P[j(t - dt) = j(t) + 1] = \frac{n(0)}{1 + n(0)(e^r - 1)/[4N(0)r]}
\]

\[
- \frac{1}{4N(t)} j(t)dt
\]

\[
= \frac{fr}{f - (f - 2r)e^{-rt}} j(t)dt .
\]

This is the instantaneous birth rate for a “reconstructed” linear birth-death process with \(B = 1/2\) and \(D = 1/2 - r\).

**Appendix C**

**Transition Probabilities**

In this appendix, we derive the transition probabilities between chromosomal haplotypes for a model with recombination between a nonrecurrent mutation M linked to a single marker locus A with two alleles, \(A_1\) and \(A_2\), which experience reversible mutations. We use a continuous-time Markov process to model mutation and recombination. During an infinitesimal time interval \(\Delta t\), the probability that a transition occurs from haplotype MA\(_1\) to haplotype MA\(_2\) is \(n\Delta t + o(\Delta t)\), where \(n = \mu + cp\), \(\mu\) is the instantaneous mutation rate from \(A_1\) to \(A_2\), \(c\) is the recombination rate between \(M\) and marker \(A\), \(p\) is the frequency of \(A_2\), and \(1 - p\) is the frequency of \(A_1\) among normal chromosomes (which is assumed to be constant). We focus our attention on haplotypes from individuals heterozygous for mutation \(M\), so that re-
combination always occurs between a normal chromosome and an M chromosome (see main text). The probability of a transition from haplotype MA\(_i\) to haplotype MA\(_2\) during \(\Delta t\) is \(v \Delta t + o(\Delta t)\), where \(v = v + c(1 - p)\) and \(v\) is the instantaneous mutation rate from \(A_2\) to \(A_1\). The Kolmogorov forward equations are

\[
\frac{dP_i(t)}{dt} = -uP_i(t) + vP_2(t)
\]

and

\[
\frac{dP_2(t)}{dt} = uP_1(t) - vP_2(t),
\]

where \(i = 1, 2\), \(P_i(t)\) is the probability that a lineage undergoes the transition \(i \rightarrow j\) in time \(t\), where a “1” in the subscript denotes haplotype MA\(_1\) and where a “2” denotes haplotype MA\(_2\). Note that \(P_{11}(t) + P_{12}(t) = P_{21}(t) + P_{22}(t) = 1\) and that, by substitution, we have

\[
\frac{dP_{11}(t)}{dt} + (u + v)P_{11}(t) = v
\]

and

\[
\frac{dP_{22}(t)}{dt} - (u + v)P_{22}(t) = u.
\]

Solving the above equations with the initial conditions \(P_{11}(0) = 1\) and \(P_{22}(0) = 1\) gives

\[
P_{11}(t) = \frac{v}{u + v} + \frac{u}{u + v} e^{-\frac{u+u+v}{u+v}},
\]

\[
P_{12}(t) = 1 - P_{11}(t) = \frac{u}{u + v} [1 - e^{-\frac{u+u+v}{u+v}}],
\]

\[
P_{22}(t) = \frac{u}{u + v} + \frac{v}{u + v} e^{-\frac{u+u+v}{u+v}},
\]

and

\[
P_{21}(t) = 1 - P_{22}(t) = \frac{v}{u + v} [1 - e^{-\frac{u+u+v}{u+v}}].
\]

References


