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 distortion 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 uncertainly 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-

Received May 29, 1997; accepted for publication November 12, 1997; electronically published January 30, 1998.

Address for correspondence and reprints: Dr. Montgomery Slatkin, Department of Integrative Biology, VLSB 3060, University of California Berkeley, Berkeley, CA 94720-3140. E-mail: slatkin@socrates.berkeley.edu

0002-9297/98/6201-0023$02.00

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_0 of the i chromosomes carrying A_1. The number Y_0, 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_0, and information about the population from which the sample was drawn, to estimate either t_1, the time of origin of M; c, the recombination rate between M and the marker locus; or μ, 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 1*a*. 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 $t_2, \ldots, t_i$ 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_2, has a distribution that depends also on the number of chromosomes sampled from the population (see fig. 1*b*).

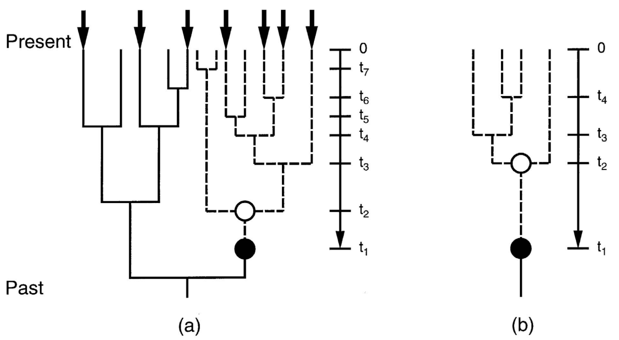

Figure 1 Gene genealogy of a population of chromosomes descended from a nonrecurrent mutation, M, that occurred at a time t_1 generations in the past. *a*, Genealogy of a population of 12 chromosomes. Mutant lineages are denoted by broken lines, and nonmutant lineages are denoted by unbroken lines. Arrows indicate the chromosomes that are sampled from the population. The time, t_1, at which the mutation occurred is denoted by a blackened circle. The time, t_2, at which mutant chromosomes first share a most recent common ancestor is denoted by an unblackened circle. The waiting times until mutant chromosomes coalesce to shared ancestral chromosomes are shown on the right, where t_7 is the waiting time until the seven chromosomes coalesce to six ancestral chromosomes, t_6 is the waiting time until they coalesce to five ancestral chromosomes, etc. *b*, Genealogy for a sample of four of the seven mutant chromosomes obtained by sampling from the population shown in *a*. The chromosomes that are sampled are indicated by arrows in *a*. The waiting times for the four sampled chromosomes to coalesce to common ancestral chromosomes are indicated on the right. Although t_1 remains fixed in the genealogy of this sample, t_2 is more recent than it is for mutant chromosomes from the whole population (compare the position of the unblackened circle in *a* with that of the unblackened circle in *b*). The waiting time t_2 until a sample of chromosomes first share a common ancestral chromosome is a random variable that is influenced by the time at which the mutation occurred, as well as by both the fraction of chromosomes sampled from the population and the population growth rate and/or selection coefficient.

A commonly used estimator of t_2, μ, 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 recombination (or mutation) event since t_2 is $Q = e^{-ut_2}$, where u is the rate of recombination (or mutation) per lineage per generation. In a sample of i M-bearing chromosomes, the expected value of Y_0, the number of chromosomes carrying A_1, is iQ. An estimator of either u or t_2 may then be obtained by equating the observed value of Y_0 with its expectation. We call this the "moments" estimator, because it is based on the standard method of moments for estimation of parameters. If u is known, the moments estimator of t_2 is

$$\hat{t}_2 = [\log(i) - \log(i - Y_0)]/u \ .$$

If t_2 is known, then u is estimated by exchanging u and t_2.

The moments estimator makes intuitive sense, but that intuition provides no guide to either the extent of bias or the confidence interval. The complete distribution of Y_0 in a sample—and, hence, the performance of the moments estimator—depends on the genealogy of M, which in turn depends on demographic parameters such as the population growth rate, the selection affecting M, and the fraction of M-bearing chromosomes in the population that are sampled. We can, however, consider an extreme case in which the moments estimator is also the maximum-likelihood estimator (MLE). If the coalescence times in the genealogy satisfy $t_i = t_{i-1} = ... = t_2$, which means that the gene genealogy is a "star" genealogy, then the moments estimator is also an MLE (see Appendix A). In general, a star genealogy provides a poor description of the genealogies expected when we consider the descendants of a particular mutant class (i.e., the intraallelic genealogy), although star genealogies may arise for population samples (i.e., chromosomes not restricted to a particular mutant class) in very rapidly growing populations (Slatkin and Hudson 1991).

Several studies have tried to take genealogy into account. Hästbacka et al. (1992) used the Luria-Delbrück theory, originally developed for the analysis of mutations in bacteria, as a demographic model for the analysis of linkage disequilibrium at a disease locus in the Finnish population. The Luria-Delbrück theory assumes that the genealogy is a result of successive synchronized binary fissions, representing the continued doubling of bacterial cells in culture. In our notation, that implies that $t_3 = t_4$, $t_5 = t_6 = t_7 = t_8$, etc., and that the intervals between 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 $t_1 = 100$. 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_1 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 of haplotypes 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) ξ, 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 $\boldsymbol{t} = \{t_2, t_3, ..., t_i\}$, where t_j is the waiting time until i sampled M-bearing chromosomes coalesce to $j-1$ ancestral chromosomes. The joint probability density of $\boldsymbol{t}$ 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.

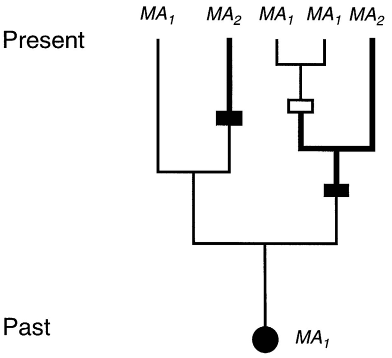

Figure 2 Intraallelic gene genealogy illustrating the effects of recombination and mutation involving a marker locus A linked to a nonrecurrent mutation M. There are two alleles for the linked marker: A_1 and A_2. The distribution of the two haplotypes MA_1 and MA_2 is shown for a sample of chromosomes descended from the ancestral haplotype, MA_1 (indicated by a blackened circle), on which the mutant first arose. Blackened boxes superimposed on the lineages denote mutation (or recombination) events, at the marker locus, of the form $A_1 \rightarrow A_2$, whereas unblackened boxes denote equivalent events of the form $A_2 \rightarrow A_1$.

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 μ be the instantaneous mutation rate from A_1 to A_2, and let ν 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 udt, where $u = \mu + c(1-p)$. The probability of a transition from an MA_2 to an MA_1 chromosome is vdt, where $v = \nu + cp$. The transition probabilities during an interval of length t may be calculated analytically (see Appendix C):

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

and

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

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 jth coalescence events (moving forward in time), there are $j-1$ independent lineages, each undergoing a transition process as described above. At the jth 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 jth 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 jth 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_2 chromosomes immediately prior to the jth coalescence event is

$$P(k|Y_{j-1}) = \binom{Y_{j-1}}{k}(uG)^k(1-uG)^{Y_{j-1}-k} \; , \qquad (1)$$

where $G = [1 - e^{-(u+v)\Delta t}]/(u+v)$ 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_2 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 jth coalescence event is

$$P(l|Y_{j-1}) = \binom{j-1-Y_{j-1}}{l}(vG)^l(1-vG)^{j-1-Y_{j-1}-l} \; . \qquad (2)$$

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

$$\begin{aligned} P^*(Y'_{j-1}|Y_{j-1}) = \sum_k &\binom{Y_{j-1}}{k} \\ &\times \binom{j-1-Y_{j-1}}{Y'_{j-1} - Y_{j-1} + k} \\ &\times (uG)^k(1-uG)^{Y_{j-1}-k} \\ &\times (vG)^{Y'_{j-1}-Y_{j-1}+k}(1-vG)^{j-1-Y'_{j-1}-k} , \end{aligned}$$

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

$$\begin{aligned} &P(Y_0|\Theta_1) \\ &= \sum_{Y_i=0}^{i} \sum_{Y_{i-1}=0}^{i-1} \ldots \sum_{Y_1=0}^{1} \int_{t_2=0}^{t_1} \int_{t_3=0}^{t_2} \ldots \int_{t_i=0}^{t_{i-1}} P(Y_0|Y_i, \mathbf{t}, \Theta_2) \\ &\times \textstyle\prod_{j=2}^{i} P(Y_j|Y_{j-1}, \mathbf{t}, \Theta_3) \\ &\times P(Y_1|p) \times P(\mathbf{t}|\Theta_4) dt_i \ldots dt_2 \; , \end{aligned}$$

where $\Theta_1 = \{u, v, p, t_1, i, f, \xi\}$, $\Theta_2 = \{u, v, i\}$, $\Theta_3 = \{u, v, t_1, i\}$, and $\Theta_4 = \{i, f, \xi, t_1\}$. There are $i!$ 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 ~10 mutant chromosomes. An alternative method for evaluating this equation is to use Monte Carlo integration.

Monte Carlo Integration

A Monte Carlo estimator of $P(Y_0|\Theta_1)$ is obtained as

$$P(Y_0|\Theta_1) \approx \frac{1}{R}\sum_{k=1}^{R} P\left(Y_0|\,\tilde{Y}_i(k), \tilde{\mathbf{t}}(k); u, v\right) \ ,$$

where the sum is evaluated over R replicate Monte Carlo realizations of the random variables $\tilde{Y}_i(k)$ and $\tilde{\mathbf{t}}(k)$. Note that $\tilde{Y}_i(k)$ is the kth of R independent random variables generated from the distribution

$$P(Y_i|\mathbf{t};\Theta_3,p) = \prod\nolimits_{k=2}^{i} P(Y_k|Y_{k-1},\mathbf{t};\Theta_3)P(Y_1|p) \ .$$

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_i, as described below. If $p < 1$, then $Y_2 = 2$ with probability $pe^{-(u+v)(t_1-t_2)} + v[1 - e^{-(u+v)(t_1-t_2)}]$ and otherwise takes the value $Y_2 = 0$. To simulate $\tilde{Y}_j(k)$ for $2 < j \leqslant i$, 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 $\tilde{Y}_j(k)$ generated at the previous step. The random variable $\tilde{Y}_j(k)$ is then obtained as $\tilde{Y}_j(k) = \tilde{Y}_{j-1}(k) + l - h$. The vector of random variables $\tilde{\mathbf{t}}(k)$ 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 $t_1 \rightarrow \infty$). 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

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

The hypotheses are nested so that $-2\log\Lambda$ is approximately χ^2 distributed with 1 df. Because $t_1 \rightarrow \infty$ is a boundary condition, it is not clear that the χ^2 approximation is justified in this case. An alternative approach is to simulate observations under the null hypothesis that $t_1 \rightarrow \infty$ and to generate the distribution of the statistic.

Given the demographic parameters ξ 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 ξ 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 ~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 ~62 kb from *CSF1R*. Subsequently, by positional cloning, Hästbacka et al. (1994) found the locus to be ~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 *Eco*RI restriction site and a *Sty*I 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, $146/n = .008$, 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 ~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, $u = c(1 - p_1)$ and $v = cp_1$, 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 .0008, and the 95% confidence interval is .0002–.0016. This estimate of c is slightly higher than .0006, the estimate given by Hästbacka et al. (1992), and our confidence interval is also larger than theirs (which was .0005–.0009). For humans, a map distance of 1 cM (i.e., a recombination rate of .01) corresponds to a physical distance of ~1 Mb. The estimated recombination rate .0008 therefore corresponds to a physical distance, between the marker and the mutation, that is ~80 kb, which is in close agreement with the actual physical distance, now known to be ~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 .0022.

Because of the high growth rate, estimates of c are insensitive to t_1. 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_1 (Slatkin and Rannala 1997). The MLE of t_1 for the Hästbacka et al. (1992) data, on the basis of frequency, is $\hat{t}_1 = 111.9$ 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 ~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.

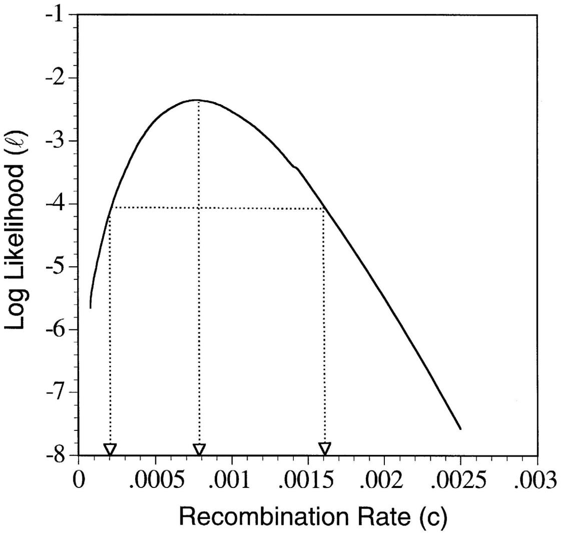

Figure 3 Log likelihood (ℓ) of the data, analyzed by Hästbacka et al. (1992), for *Sty*I marker haplotypes of chromosomes associated with DTD in the Finnish population, as a function of the recombination rate with the marker (see text). The maximum-likelihood estimate of the recombination rate c, as well as the 95% confidence interval of the estimate (i.e., the values that are $\leqslant 2$ units of log likelihood below the maximum-likelihood estimate), are indicated by arrows. The MLE is $\hat{c} = .0008$, with the 95% confidence interval .0002–.0016.

In our analysis, we assume that the current population size of Ashkenazi Jews is $N = 5 \times 10^6$ 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 $f = n/(2N) = .0324$. Historical records suggest that the rate of population increase has been ~1.5 fold per generation (Risch et al. 1995), which implies a growth rate of $\xi = \ln(1.5) = .4055$.

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 $c_1 = .018$ with D9S63, with a confidence interval of .007–.033, and a recombination rate of $c_2 = .023$ with D9S64, with a 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_2, the recombination rate with *DYT1*, is $.018 + .023 - .018 \times .023 = .0406$. Hence $u_2 = .035566$ 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 gene 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 ~10 generations, and the lower limit of the 95% confidence interval is ~2 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, ~72 generations for t_2 and ~131 generations for t_1. 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 ~25 generations, with upper confidence limits of ~100 and ~80 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).

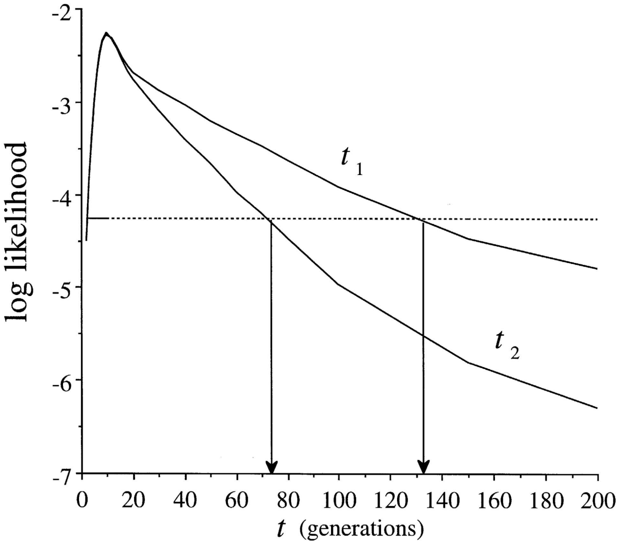

Figure 4 Log likelihood (ℓ) curves for the estimation of the age of the ITD allele found in high frequency in Ashkenazi Jews. Time, t, is measured in generations, and both curves are based on the data at the *ASS* locus, as described in the text; $n = 54$ and $Y_0 = 47$. The parameters for both curves were as follows: $\xi = .4055$, $u = .0164$, $v = .001548$, and $f = .0324$. A total of 100,000 replicates of the Monte Carlo integration described in the text were used to obtain the likelihood for each point. The upper curve is the log likelihood of t_1, the time of origin of the mutant, and the lower curve is the log likelihood of t_2, the age of the root of the mutant-gene genealogy. The broken line shows values of log(L) that are 2 units less than the maximum and that hence provide an approximate graphical way to determine the upper and lower 95% confidence limits on estimates of t_1 and t_2.

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 star like, 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 moments 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.

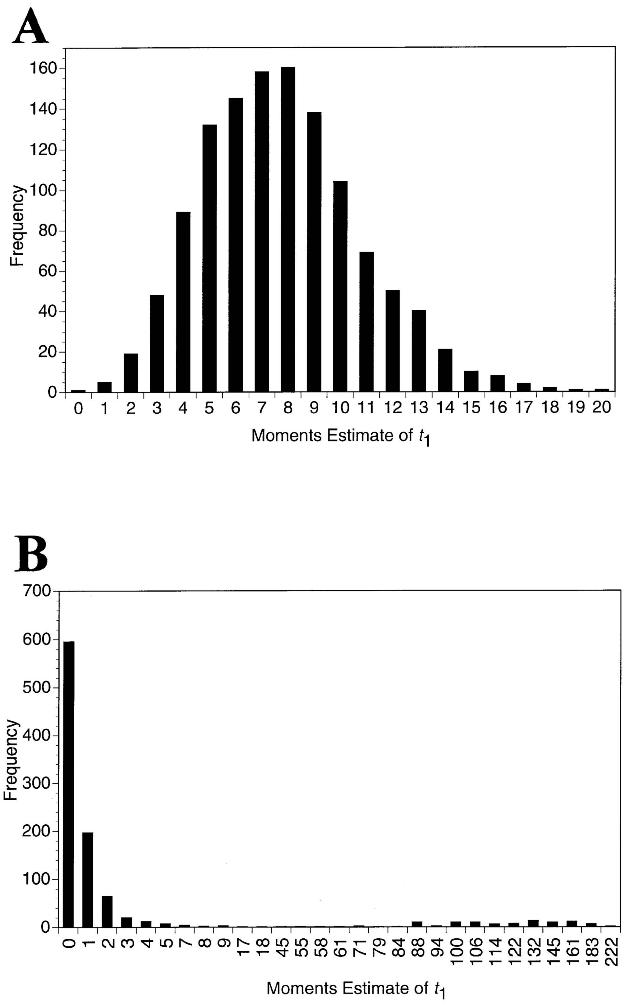

Figure 5 Frequency distributions of estimates of t_2, obtained by use of the moments estimator of allele age, for artificial data simulated on the basis of parameters for data on ITD that were collected from the population of Ashkenazi Jews by Risch et al. (1995). For these simulations, it was assumed that the actual time of the shared common ancestor for the sampled chromosomes was $t_2 = 8.4$, which is the estimate obtained from the original data by the moments estimator. *A,* Distribution of the estimator for simulated data when the population genealogy is assumed to be a star genealogy (i.e., $t_2 = t_3 = \ldots = t_i = 8.4$). *B,* Distribution of the estimator for simulated data when the distribution of genealogies is as expected on the basis of the inferred demography of the population of Ashkenazi Jews. The average value of the estimator for the simulated data, when a star genealogy is assumed, is 7.70, and the MSE (the average squared difference between the inferred value of the parameter and its true value 8.4) is 9.09. The average value of the moments estimator for the data simulated by use of the distribution of genealogies expected for the demographic parameters of the population of Ashkenazi Jews is 12.0, and the MSE is 1,365.3.

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 μ from the 90-bp allele to any other allele, no back mutation, and a 0 rate of recombination between the STRP and Alu markers. Furthermore, they excluded from their analysis STRP alleles $\geqslant$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 $\geqslant$115 bp) among the Africans. Tishkoff et al. (1996) then assumed a Poisson mutation process, with mutation rate μ, 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: $Q_A = 34/85 = .40$ is the proportion of 90-bp allele on the Alu(−) chromosomes in Africans, and $Q_B = 264/269 = .9814$ 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_B that is approximately the time at which ancestors of non-Africans left Africa and (2) that the Alu(−) allele arose by mutation $\leqslant t_A = 5$ million years ago. Under the assumption of exponential decay of the Q's, the ratio $\ln(Q_A)/\ln(Q_B) = t_A/t_B = 48.8$ gives a value of t_B that is ~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_2. For both populations, the values of t_1 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 μ 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, μ 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 μ or whether μ 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 $N_B = 4 \times 10^9$. 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 $\xi_A = 6.447 \times 10^{-5}$. We found that the exact value of ξ did not matter, and, in fact, almost the same estimates of μ 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 $\xi_B = .002$ and $\xi_B = .005$ 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 ξ_B allows for a later initiation of exponential growth. If the STRP alleles $\geqslant$110 bp are ignored, $n_A = 349$, $i_A = 85$, $Y_{0,A} = 34$, $n_B = 1424$, $i_B = 269$, and $Y_{0,B} = 264$. These numbers are obtained from table 3 of Tishkoff et al. (1996), by removal of all data with allele sizes $\geqslant$110 bp at the STRP locus. The sampling fractions are $f_A = 349/(2 \times 10^9) = 1.745 \times 10^{-7}$ and $f_B = 1,424/(8 \times 10^9) = 1.78 \times 10^{-7}$.

Figure 6 shows the log likelihood as a function of μ for the African samples. The MLE of μ is 5.88×10^{-6} (2.13×10^{-6}, 9.63×10^{-6}) if we assume that t_1 is 2.5×10^5 generations. If we instead assume that $t_2 = 2.5 \times 10^5$ generations, the results are almost the same; the MLE of μ is 4.91×10^{-6} (1.20×10^{-6}, 8.63×10^{-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 $\xi_B = .002$ 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 ~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 $\xi_B = .005$. In general, larger growth rates make it even more difficult to estimate allele ages confidently.

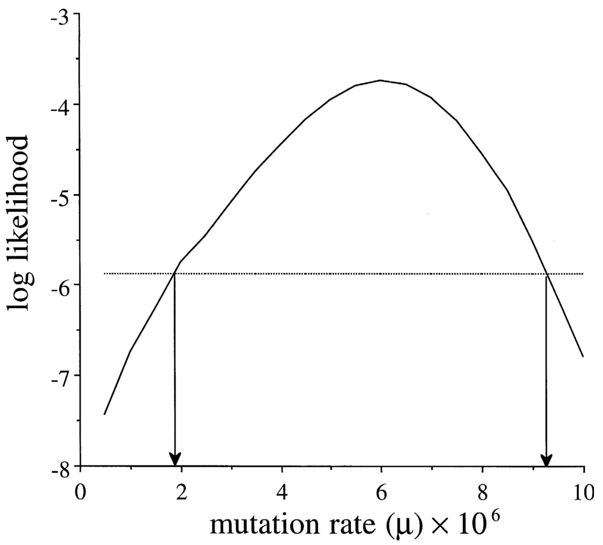

Figure 6 Log likelihood (ℓ) of the mutation rate, μ, at the STRP locus, for sub-Saharan African samples in the Tishkoff et al. (1996) data set. Only Alu(−) alleles $\leqslant 110$ bp (see text) as a function of μ measured in units of 10^{-6} mutations/chromosome/generation are shown. Arrows indicate the position of the bounds for the 95% confidence interval of this estimate (i.e., the values that are $\leqslant 2$ log likelihoods below the MLE). The parameter values used to generate this curve were as follows: $i_A = 85$, $Y_{0,A} = 34$, $\xi_A = 6.447 \times 10^{-5}$, $f_A = 1.745 \times 10^{-7}$, and $t_1 = 250{,}000$ generations, where the subscript A denotes African populations.

The observed frequency of the Alu(−) allele also contains information about t_1, if the population growth rate is known. With $\xi_B = .002$ the MLE based on the allele frequency, when equation (8) of Slatkin and Rannala (1997) is used, is $\hat{t}_1 = 7{,}139.2$ (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 μ increases only slightly, to 6.9×10^{-6}, with the confidence interval 1.7×10^{-6}–1.2×10^{-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., $t_1 = 2.5 \times 10^4$, rather than 2.5×10^5, generations) and then began to grow exponentially to a current size of 10^9, ξ_A would then be 6.447×10^{-4}, 10 times larger than the value used to generate figure 7. The resulting MLE of μ is 4.63×10^{-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_1 is ~600 generations (or 12,000 years), with a confidence interval of ~100–1,000 generations, for $\xi_B = .002$. 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 μ results in greater uncertainty in the estimate of t_B. We agree that there is some uncertainly in the estimate of μ (see fig. 6), although we do not agree with their method for finding the confidence interval on the estimate of μ. 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 μ, f, ξ, 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_0, 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)

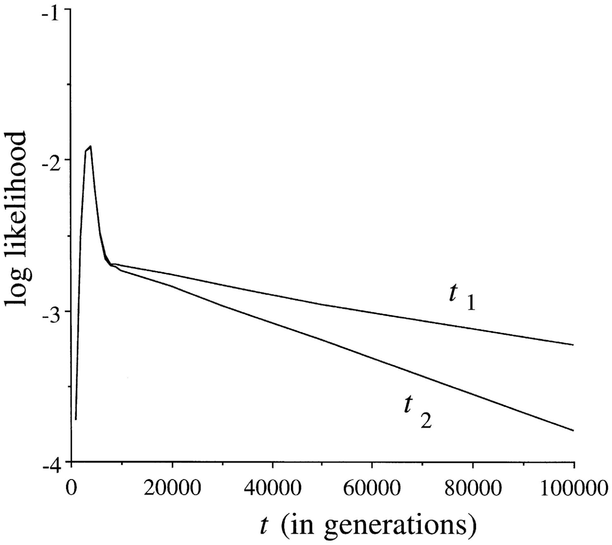

Figure 7 Log likelihood (ℓ) of the age of the Alu(−) allele in non-Africans in the Tishkoff et al. (1996) data set. The upper curve is for t_1, the earliest time at which the allele ancestral to all copies of Alu(−) in non-Africans formed a distinct lineage, and the lower curve is for t_2, the age of the most recent common ancestor in the gene genealogy of Alu(−) alleles in non-Africans. For these curves, i_B = 269, $Y_{0,B}$ = 264, ξ_B = .002, $f_B = 1.78 \times 10^{-7}$, 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 uncertainly 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), ξ (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) data set, 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 results from Slatkin and Rannala (1997) are used, can provide narrower bounds on estimates of allele age. Neither 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 gene 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

$$\hat{t}_2 = \frac{\log(i) - \log(i - Y_0)}{u} \ , \qquad \text{(A1)}$$

obtained by setting the observed fraction of nonmutated markers $(i - Y_0)/i$ equal to the expected proportion $\exp\{-ut_2\}$ 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,

$$\mathrm{L}(q|\mathrm{Y}_0) = \binom{i}{\mathrm{Y}_0} q^{\mathrm{Y}_0}(1-q)^{i-\mathrm{Y}_0} \ ,$$

where $\mathrm{L}(q|\mathrm{Y}_0)$ denotes the likelihood function and where $q = 1 - e^{-ut_2}$. The MLE of q is then $\hat{q} = j/i$. 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 μ is known. The approximate asymptotic variance for the MLE of t_2 is

$$\mathrm{Var}(\hat{t}_2) \approx \frac{\left\{\frac{\partial}{\partial q}\left[\frac{-\log(1-q)}{u}\right]\right\}^2 \Big|_{q=\hat{q}}}{-\frac{\partial^2}{\partial q^2}\log L(q|\mathrm{Y}_0)\big|_{q=\hat{q}}} = \frac{\mathrm{Y}_0}{i(i-\mathrm{Y}_0)u} \ .$$

Appendix B

Distribution of Intraallelic Coalescence Times

In this appendix, we derive the distribution of coalescence times for a sample of chromosomes descended from a nonrecurrent mutant ancestor M that arose at generation t_1 in the past. It is assumed that the genealogy of a sample of chromosomes from the population as a whole can be described by the coalescent process of Kingman (1982). A continuous-time approximation is used, with time measured in generations. Let $n(0)$ be the total number of chromosomes sampled, and let $2N$ be the total number of chromosomes in the population (i.e., N diploid individuals) with $n(0) \ll 2N$. The population size may vary over time, in which case we let $N(t)$ be the (diploid) population size at time t in the past, with $N(0)$ as the current population size. Let $j(0)$ be the number of M-bearing chromosomes in the sample, and let $j(t) \leqslant j(0)$ be the number of ancestors of the sampled M-bearing chromosomes at time t in the past. Let $n(t)$ be the total number of chromosomes existing at time t in the past that are ancestral to the $n(0)$ sampled chromosomes, where $n(t) \geqslant j(t)$. Let $i(t)$ be the number of nonmutant chromosomes existing at time t that are ancestral to the sample, so that $n(t) = i(t) + j(t)$. If $j(t)$ ancestral M-bearing chromosomes exist at time t,then the probability that one additional M-bearing chromosome arises during the interval $(t, t + dt)$ and leaves descendants in the sample is

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

where

$$\begin{aligned} &P[j(t-dt)=j(t)+1|i(t)] \qquad \text{(B1)}\\ &=\frac{1}{4N(t)}[i(t)+j(t)][i(t)+j(t)-1]\left[\frac{j(t)}{i(t)+j(t)}\right]dt\\ &=\frac{1}{4N(t)}[i(t)+j(t)-1]j(t)dt \ . \end{aligned}$$

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

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

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

$$\begin{aligned} P[j(t-dt)=j(t)+1] &\cong \frac{j(t)}{4N(t)}\sum_{i(t)=1}^{n(0)-j(0)} i(t)P[i(t)]dt\\ &= \frac{j(t)}{4N(t)}\mathrm{E}[i(t)]dt\\ &\cong \frac{j(t)}{4N(t)}\mathrm{E}[n(t)]dt \ . \end{aligned}$$

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

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

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

$$\tau(t) = \int_0^t \frac{dt'}{2N(t')} \ .$$

With a constant population size, the transition probability is then

$$\begin{aligned} P[j(t-dt)=j(t)+1] &= \frac{n(0)}{(1+n(0)t/2)4N(0)}j(t)dt\\ &= \frac{f/2}{1+ft/2}j(t)dt \ , \end{aligned}$$

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 birth-death 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 $\tau(t) = (e^{rt} - 1)/[2N(0)r]$ to obtain

$$\begin{aligned} &P[j(t-dt)=j(t)+1]\\ &= \frac{n(0)}{1+n(0)(e^{rt}-1)/[4N(0)r]}\\ &\quad \frac{1}{4N(t)}j(t)dt\\ &= \frac{fr}{f-(f-2r)e^{-rt}}j(t)dt \ . \end{aligned}$$

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 Δt, the probability that a transition occurs from haplotype MA_1 to haplotype MA_2 is $u\Delta t + o(\Delta t)$, where $u = \mu + cp$, μ 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 recombination always occurs between a normal chromosome and an M chromosome (see main text). The probability of a transition from haplotype MA_2 to haplotype MA_1 during Δt is $v\Delta t + o\ (\Delta t)$, where $\nu = v + c(1 - p)$ and v is the instantaneous mutation rate from A_2 to A_1. The Kolmogorov forward equations are

$$\frac{dP_{i1}(t)}{dt} = -uP_{i1}(t) + vP_{i2}(t)$$

and

$$\frac{dP_{i2}(t)}{dt} = uP_{i1}(t) - vP_{i2}(t)\ ,$$

where $i = 1, 2, P_{ij}(t)$ 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

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

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

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

and

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

References

Casella G, Berger RL (1990) Statistical interference. Duxbury Press, Belmont

Guo S-W, Xiong M (1997) Estimating the age of mutant disease alleles based on linkage disequilibrium. Hum Hered 47: 315–337

Hästbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver W, Lander ES (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat Genet 2:204–211

Hästbacka J, de la Chapelle A, Mahtani MM, Clines G, Reeve-Daly MP, Daly M, Hamilton BA, et al (1994) The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. Cell 78:1073–1087

Kaplan NL, Hill WG, Weir BS (1995) Likelihood methods for locating disease genes in nonequilibrium populations. Am J Hum Genet 56:18–32

Kaplan NL, Weir BS (1995) Are moment bounds on the recombination fraction between a marker and disease locus too good to be true? allelic association mapping revisited for simple genetic diseases in the Finnish population. Am J Hum Genet 57:1486–1498

Kingman JFC (1982) The coalescent. Stochastic Proc Appl 13: 235–248

Lander ES, Botstein D (1986) Mapping complex genetic traits in humans: new methods using a complete RFLP linkage map. Cold Spring Harb Symp Quant Biol 51:49–62

Nee S, May RM, Harvey PH (1994) The reconstructed evolutionary process. Philos Trans R Soc Lond Biol 344: 305–311

Pritchard JK, Feldman MW (1996) Genetic data and the African origin of humans. Science 274:1548

Rannala B (1997) Gene genealogy in a population of variable size. Heredity 78:417–423

Risch N, de Leon D, Ozelius L, Kramer P, Almasy L, Singer B, Fahn S, et al (1995) Genetic analysis of idiopathic torsion dystonia in Ashkenazi Jews and their recent descent from a small founder population. Nat Genet 9:152–159

Serre JL, Simon-Bouy B, Mornet E, Jaume-Roig B, Balassopoulou A, Schwartz M, Taillandier A, et al (1990) Studies of RFLP closely linked to the cystic fibrosis locus throughout Europe lead to new considerations in populations genetics. Hum Genet 84:449–454

Slatkin M, Hudson RR (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129:555–562

Slatkin M, Rannala B (1997) Estimating the age of alleles by use of intraallelic variability. Am J Hum Genet 60:447–458

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

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, Bonne-Tamir B, et al (1996) Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271:1380–1387

Xiong M, Guo S-W (1997) Fine-scale genetic mapping based on linkage disequilibrium: theory and applications. Am J Hum Genet 60:1513–1531