Estimating the Age of Alleles by Use of Intraallelic Variability

Montgomery Slatkin and Bruce Rannala

Department of Integrative Biology, University of California at Berkeley, Berkeley

Summary

A method is presented for estimating the age of an allele by use of its frequency and the extent of variation among different copies. The method uses the joint distribution of the number of copies in a population sample and the coalescence times of the intraallelic gene genealogy conditioned on the number of copies. The linear birth-death process is used to approximate the dynamics of a rare allele in a finite population. A maximum-likelihood estimate of the age of the allele is obtained by Monte Carlo integration over the coalescence times. The method is applied to two alleles at the cystic fibrosis (*CFTR*) locus, ΔF508 and G542X, for which intraallelic variability at three intronic microsatellite loci has been examined. Our results indicate that G542X is somewhat older than ΔF508. Although absolute estimates depend on the mutation rates at the microsatellite loci, our results support the hypothesis that ΔF508 arose <500 generations (≈10,000 years) ago.

Introduction

Both the extent of variability within an allelic class and the allele frequency contain information about the age of an allele, provided some assumptions can be made about the allele and the population in which it is found. Estimates of allele age are interesting in their own right, and, if other information restricts the range of possible ages, estimates of the age may provide information about what selection the allele has experienced. Here we develop a maximum-likelihood method for estimating allele age based on the extent of genetic variation within an allelic class and the allele frequency.

There is a well-developed theory to estimate the age of a neutral allele in a population of constant size by use of its frequency. Much of this theory assumes an infinite-alleles model of mutation and treats the age of an allele as a random variable (Watterson 1976; Sawyer 1979; Donnelly and Tavaré 1986). Thompson (1976) used a different approach and assumed that an allele was the result of a unique mutation event at some time in the past; its age is then a fixed parameter, rather than a random variable. Thompson used the theory of branching processes (with a fractional linear distribution of offspring numbers) to find the likelihood function for the time of occurrence of the mutation, given the number of copies found.

The extent of linkage disequilibrium between closely linked marker loci has also been used to estimate allele age. On the basis of theory that assumes an exponential decay of linkage disequilibrium with time, Serre et al. (1990) estimated the age of the ΔF508 allele of the *CFTR* locus to be between 3,000 and 6,000 years, and Risch et al. (1995) estimated the age of the allele causing idiopathic torsion dystonia in Ashkenazi Jews to be ∼350 years. The demographic model we consider here may also be applied to study the decay of linkage disequilibrium between a mutant allele and associated markers, but that approach will not be pursued in this paper. Instead, we focus on the information provided by the allele frequency and the intraallelic genealogy.

Before proceeding, we need to emphasize what is being estimated by our method. The parameter t_1, which we estimate, is the generation in the past during which the mutation defining the allelic class was created by mutation. Immediately following the mutation, there was a single allelic lineage representing the ancestors of the allelic class found in the sample at the present time. At a later time, t_2 generations in the past, the allelic lineage divided to produce two lineages both represented by present-day descendent lineages. The time t_2 therefore represents the age of the root of the intraallelic gene genealogy. The times t_1 and t_2 are illustrated graphically in figure 1. Our results show that the difference between t_1 and t_2 can be substantial, especially when there is a high rate of population growth or positive selection acting on the mutant allele. The extent of intraallelic variability depends on t_2, not on t_1, because any mutations that occur between t_1 and t_2 will be inherited by all descendents. Nevertheless, our method estimates t_1 because the population genetics model provides the distribution of t_2 and the other coalescence times as functions of t_1. It would be possible to estimate t_2 directly from the pattern of intraallelic variability, treating the problem in purely phylogenetic terms, but that approach would

Received August 6, 1996; accepted for publication November 21, 1996.

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

0002-9297/97/6002-0025$02.00

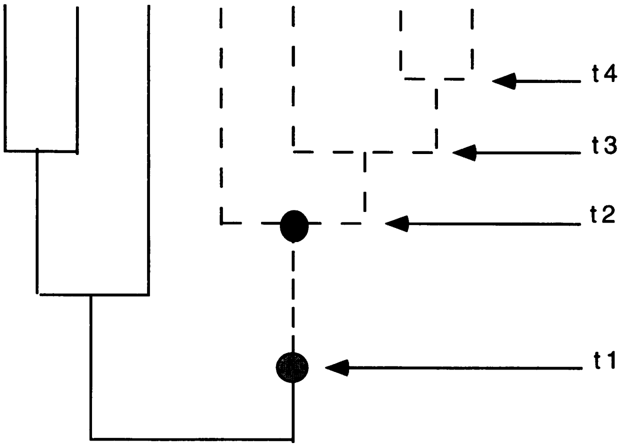

Figure 1 Example genealogy of a mutant allelic class within a larger population. The mutant lineages are indicated with dashed lines and the nonmutant lineages with solid lines. The mutation occurs at time t_1, indicated by a gray circle. Mutations occurring in the mutant class prior to time t_2 (indicated by a black circle) provide no additional information regarding the genealogical structure, since they are represented in all descendent lineages. The four mutant lineages shown in the figure coalesce to three ancestral lineages at time t_4, to two ancestral lineages at time t_3, and to a single ancestral lineage at time t_2.

require sufficiently extensive intraallelic variability to estimate all internal branch lengths of the intraallelic genealogy and would make no use of the fact that variation is being generated by population-genetic processes. Our method provides an estimate of t_1 by use of a relatively small amount of data because we use the model to provide a prior distribution of coalescence times. Estimates of allele age based on the extent of linkage disequilibrium (Serre et al. 1990; Risch et al. 1995) are estimates of t_2.

We begin by summarizing several results for linear birth-death processes. Then we present a likelihood method for estimating the age of an allele when intraallelic genetic variability is not taken into account. This method is similar to that of Thompson (1976). Next we develop the theory of intraallelic variation based on the distribution of coalescence times in the intraallelic gene genealogy. We then show how that theory leads to a likelihood function that may be used to derive an estimator of allele age with information from both the allele frequency and the intraallelic variation. Finally, we apply this method to estimate the ages of two alleles of the cystic fibrosis (CF) locus, ΔF508 and G542X. In the appendix, we show the relationship between coalescent theory for low-frequency alleles and the genealogical structure of a linear birth-death process.

The Linear Birth-Death Process

The general theory of linear birth-death processes, developed by Kendall (1948), predicts the distribution of the number of present-day descendents of a single lineage existing at time t_1 in the past. The key assumption of the birth-death process is that each lineage reproduces independently. For a time-homogeneous birth-death process, the individual birth rate is B, and the individual death rate is D per unit time. The probability that a birth occurs in a population of size j during an infinitesimal interval δt is $jB\delta t$, and the probability that a death occurs is $jD\delta t$. The probability of two or more events is of order $o(\delta t)$, where $o(\delta t)$ denotes terms satisfying $\lim_{\delta t \to 0} o(\delta t)/\delta t = 0$. The probability that a lineage alive at time t in the past has exactly $i \geqslant 1$ present-day descendents is

$$p_i(t) = P(0,\, t)(1 - v_t)v_t^{i-1}\,, \tag{1}$$

where $P(0, t)$ is the probability that a lineage alive at time t in the past has one or more present-day descendents and

$$v_t = 1 - P(0,\, t)\exp\{(D - B)t\}\,. \tag{2}$$

We show in the appendix that, under the assumptions leading to a coalescent process, with either constant population size or exponential population growth, the dynamics of a rare allele can be described using a time-homogeneous linear birth-death process with rates $B = \frac{1}{2}$ and $D = \frac{1}{2} - \xi$ (time is scaled in units of generations), where $\xi = s + r$, s is the selection coefficient for heterozygotes, and r is the exponential growth rate parameter. From the point of view of a rare allele, selection in favor of heterozygotes is equivalent to the growth of the whole population. The effects of selection against homozygotes are ignored. The present theory may be extended to cases with population size as an arbitrary function of time and a temporally varying selection coefficient using a generalized birth-death process with time-dependent rates.

The theory presented above focuses on the distribution of the number of copies of a rare allele in the population as a whole. In most cases, the number of copies in a sample from a population is the quantity of interest. In this section, we consider the sampling theory for a birth-death process. Population sampling is most easily incorporated into the present theory by taking an approach similar to that of Nee et al. (1994), who studied the genealogical properties of a birth-death process. Let a fraction f of a total of $2N$ lineages in the present-day population be sampled at random. Each lineage then has a probability f of being sampled, and i is the number of sampled lineages belonging to the allelic class of interest. The sampling process may be modeled as a mass-killing event that occurs precisely at time present, where f is the probability that a lineage alive at present survives the killing event. Let $g(t, T)$ be the Dirac delta function defined by

$$1(\tau > T) = \int_{-\infty}^{\tau} g(t, T)dt\,. \qquad (3)$$

The death rate is then defined to be $D(t, T) = \frac{1}{2} - \xi - g(t, T)\ln f$ to obtain

$$P(0, t) = \frac{2f\xi}{f - (f - 2\xi)\exp\{-\xi t\}}\,. \qquad (4)$$

It follows that

$$v_t = 1 - \frac{1}{f} P(0, t)e^{-\xi t}\,, \qquad (5)$$

and

$$p_1(t) = \frac{1}{f} P(0, t)^2 e^{-\xi t}\,, \qquad (6)$$

where n is the number of sequences sampled, $f = n/2N$, and $p_1(t)$ is the probability that a lineage alive at time t in the past leaves exactly one present-day descendent (Nee et al. 1994).

Allele Age Inferred from the Frequency Alone

In this section we present a method for estimating the age of an allele based on the number of copies of the allele in a sample from a population. The probability that i copies of a mutant allele that arose at time t_1 in the past are observed in a sample of n sequences from a present-day population of size N (conditional on the allele having at least one copy in the sample) is

$$P(i|t_1, N, n, s; i > 0) = (1 - v_{t_1})v_{t_1}^{i-1}\,. \qquad (7)$$

A maximum likelihood estimator (MLE) of t_1 (in units of generations) is then

$$\hat{t}_1 = \frac{1}{\xi} \log\left\{\frac{4N}{n}\xi(i - 1) + 1\right\}. \qquad (8)$$

The estimate of allele age obtained using (8) is biased for small i but is asymptotically efficient and consistent (Cox and Hinkley 1974). The asymptotic variance of the estimator is easily obtained by determining the support interval of the likelihood (Edwards 1972). If $\xi = 0$, the MLE simplifies to $\hat{t}_1 = 4N(i - 1)/n$, which is an unbiased estimator of t_1.

This result is similar to that of Thompson (1976). Thompson used a discrete-generation branching process to model a rare allele in a population and assumed a fractional linear distribution of offspring, which implies a modified geometric distribution of the number of copies at any time. The results for the linear birth-death process are equivalent because they also lead to a modified geometric distribution. The birth-death process can be made formally identical to a branching process with a fractional linear distribution of offspring number (for large t) by censusing the population at evenly spaced time intervals. Our model differs from Thompson's in allowing for sampling.

Figure 2*a* shows the log-likelihood (l) as a function of t_1, using parameter values appropriate for the ΔF508 mutation of the CF gene (see below). The resulting estimate is $\hat{t}_1 = 2339.1$ with the 95% confidence interval (2037.8, 2928.6). Figure 2*b* shows l as a function of t_1 if ξ is at the upper limit of its reasonable range, $\xi = 0.02$. In that case, we instead obtain $\hat{t}_1 = 654.1$ with the confidence interval (578.1, 801.5). Figure 2*c* shows l as a function of t_1 with parameters appropriate for a second allele of cystic fibrosis, G542X. The MLE is $\hat{t}_1 = 1508.8$ with the confidence interval (1202.4, 2099.8). Figure 2*d* shows l as a function of t_1 if ξ is at the upper limit of its reasonable range, $\xi = 0.02$. In that case, we instead obtain $\hat{t}_1 = 446.5$ with confidence interval (369.8, 594.3). Solely on the basis of on the observed frequencies of the two alleles, ΔF508 then appears to be somewhat older than G542X. The estimated population frequency of ΔF508 is higher than that of G542X, and this conclusion agrees with the general view that more frequent alleles are likely to be older.

Gene Genealogy of a Mutant Allelic Class

The allele we are concerned with is assumed to be distinguished by a nonrecurrent mutation, but other sites within the allele continue to mutate. For example, the ΔF508 mutation of the CF gene has a deletion of the 508th amino acid, and it is thought that all copies of the allele are descended from a single ancestral copy (Morral et al. 1994). Different copies of that mutation carry different alleles at three microsatellite loci that are found within introns. At the present time, there is little information available about intraallelic variation, so we will concentrate here on the total number of mutations that have occurred. However, the likelihood method we present is readily extended to other models of mutation (see Discussion).

Under the infinite sites model, all mutations can be detected, so the number of segregating sites within an allelic class, S, indicates the number of mutations that have occurred. With mutation rate μ, the number of segregating sites within the allelic class, S is a Poisson distributed random variable with mean μT, where T is the total length of the intraallelic gene genealogy (see,

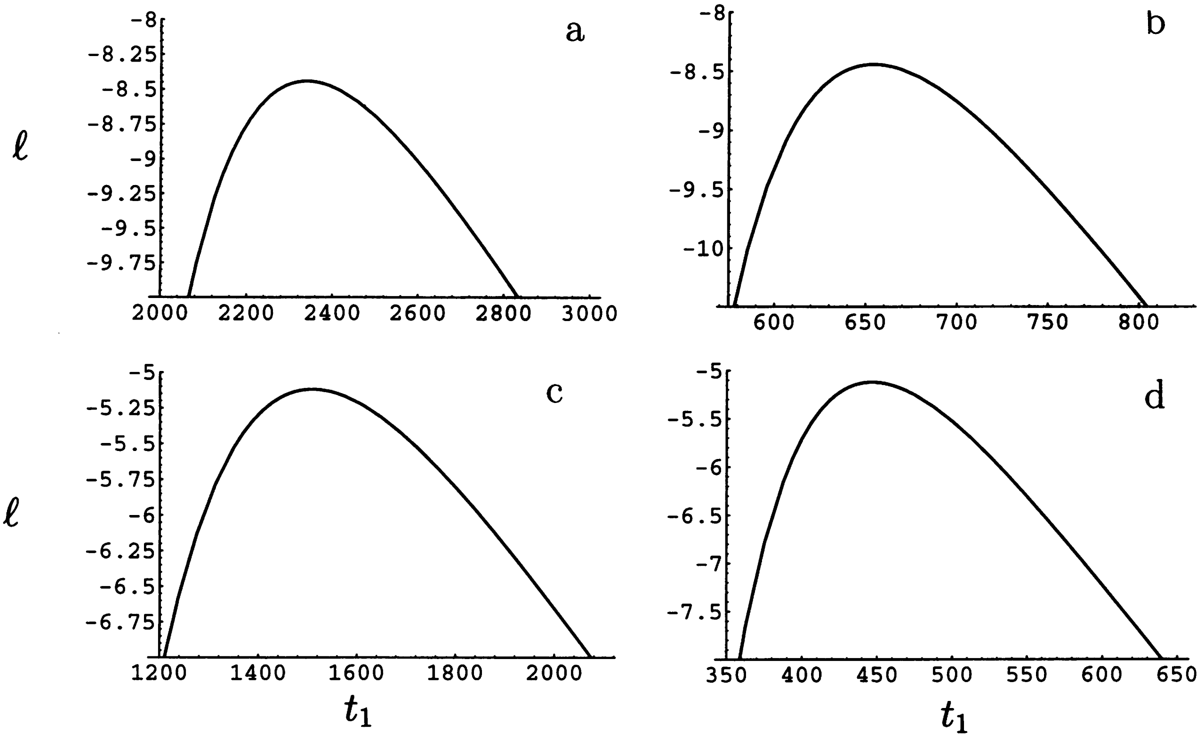

Figure 2 Log-likelihood (l) of the observed number of copies of the ΔF508 and G542X mutations of the CF gene as a function of the time of origin of each mutation (t_1) measured in units of generations. *Panel a* shows l as a function of t_1 for the ΔF508 allele with $\xi = 0.005$, $N = 3 \times 10^8$, and $f = 0.00014$. The MLE is then $\hat{t}_1 = 2{,}339.1$, with confidence interval (2,037.8, 2,928.6). *Panel b* shows l as a function of t_1 for the ΔF508 allele with $\xi = 0.02$, which corresponds to high population growth and/or selection, and the other parameters as given above. The MLE is $\hat{t}_1 = 654.1$ (578.8, 801.5). *Panel c* shows l as a function of t_1 for the G542X allele with $\xi = 0.005$, $N = 3 \times 10^8$, and $f = 0.00014$. The MLE is $\hat{t}_1 = 1{,}508.8$ (1,202.4, 2,099.8). *Panel d* shows l as a function of t_1 for the G542X allele with $\xi = 0.02$ and the other parameters as given above. The MLE is $\hat{t}_1 = 446.5$ (369.8, 594.3).

e.g., Hudson 1990). If the intraallelic variability is known only for microsatellite loci, which appear to experience recurrent mutations, it is still possible to infer a lower bound on the number of mutations (see below). The approach of Yang (in press) could be used to correct for multiple substitutions if sequence data were available; for extensive sequence data, methods analogous to those of Griffiths and Tavaré (1994) could be used to analyze the complete genealogical structure among intraallelic variants and provide a more refined estimate of allele age.

The problem is to find the distribution of T, the total tree length, for the intraallelic genealogy, given the allele age t_1, the number of copies of the allele in the sample, i, the population size, N, and the growth and selection rate, $\xi = r + s$. Slatkin's (1996) method for examining the genealogy of the mutant allelic class can be used to generate the moments of T, but it does not easily provide the entire distribution. We can find the distribution of T using a linear birth-death process that approximates the coalescent model used by Slatkin (1996). In the appendix, we show that the birth-death process provides an approximation to the coalescent model when the sample size is relatively large and t_1 is relatively small.

The Intraallelic Genealogical Process

In this section, we develop the theory needed to derive the distribution of times at which lineages within a mutant class coalesce using a linear birth-death process to model the dynamics of a rare allele. Thompson (1975) derived the joint density of the coalescence times under a linear birth-death process, conditional on the time of the final coalescence. This result has also been derived by Nee et al. (1994). Our derivation differs from these authors only in that we condition on the time the ultimate ancestor of a population of lineages first arose (i.e., the time a mutation first occurred) rather than the time of the final coalescence event. Let $t_i < t_{i-1} \ldots < t_2$ be the times (in the past) at which i individuals alive at time present ($t = 0$) coalesce to $i - 1$ ancestral lineages, $i - 2$ ancestral lineages, and so on (see fig. 1). The probability that an individual is born to one of the $j - 1$ lineages existing at time t_j that ultimately survive to time present and the new lineage also survives to time present and leaves a single descendent lineage is

$$(j - 1)Bp_1(t_j)\,, \qquad (9)$$

where $p_1(t_j)$ is the probability that a lineage represented by a single ancestral lineage at time t_j in the past leaves exactly one descendent at time present. Let a single ancestral lineage arise at time t_1 in the past. The joint probability of i descendent lineages at present, arising at times $\mathbf{t} = t_i < t_{i-1} < \ldots < t_2$ in the past, is

$$\begin{aligned} P(\mathbf{t}, i|B, D) &= p_1(t_1)Bp_1(t_2)2Bp_1(t_3) \\ &\quad \ldots (i-1)Bp_1(t_i) \qquad (10) \\ &= (i-1)!p_1(t_1)\prod_{j=2}^{i} Bp_1(t_j) . \end{aligned}$$

Conditioning on the observed number of descendent lineages at present (i), the joint probability of $\mathbf{t}$ is

$$\begin{aligned} P(\mathbf{t}|i; B, D) &= \frac{P(\mathbf{t}, i|B, D)}{P(i|B, D)} \\ &= (i-1)!\frac{p_1(t_1)}{p_i(t_1)}\prod_{j=2}^{i} Bp_1(t_j) \qquad (11) \\ &= (i-1)!\prod_{j=2}^{i}\frac{Bp_1(t_j)}{v_{t_1}} . \end{aligned}$$

In the special case of a mutant allele in low frequency in a diploid population of constant size N, or experiencing exponential growth with rate r, the birth and death rates are $B = 1/2$ and $D = 1/2 - \xi$ (see appendix). The joint density of (11) is then equivalent to the joint density of the order statistics of $i - 1$ random variables (see Rannala, in press) that are independent and identically distributed (iid) with density

$$h(x) = \frac{p_1(x)}{2v_{t_1}} . \qquad (12)$$

This result may be used to calculate the expected total length of the genealogy under a birth-death process and also simplifies the calculation of the likelihood using Monte Carlo integration (see below).

It follows that, if we consider a sample of n sequences from a population of N diploid individuals and define $f = n/2N$, we obtain the kernal density

$$\begin{aligned} h(x) &= \frac{p_1(x)}{2v_{t_1}} \qquad (13) \\ &= \frac{P(0, x)^2e^{-\xi x}}{2[f - P(0, t_1)e^{-\xi t_1}]} . \end{aligned}$$

where this corresponds to the density of (12) but allows for random sampling. If all individuals in the population are sampled so that $f = 1$, this density reduces to the result for the population coalescent under a birth-death process.

The Genealogy Length

The total length of the genealogy of individuals in the sample is obtained using the waiting times between coalescence events in a birth-death process as described above and applying the following linear transformation:

$$T = \sum_{j=2}^{N} t_j + t_2 . \qquad (14)$$

The exact probability distribution of T under a birth-death process appears difficult to derive for samples of more than two individuals, but may be approximated using Monte Carlo simulation methods (see below). The preceding theory may be applied to derive an MLE of the age of the ultimate ancestor (t_1) of the class of mutant alleles, and we now consider this problem in some detail.

MLE of Allele Age

In this section, we develop a MLE of t_1 that takes into account both the frequency of the mutant class in a sample of DNA sequences and the observed intraallelic variability as measured by the number of segregating nucleotide sites (S). The joint probability of observing S segregating nucleotide sites among i mutants in a sample of n sequences from a population is

$$\begin{aligned} &P(S, i|t_1, N, n, \xi, \mu) \\ &\quad = P(S|t_1, N, n, \xi, \mu; i)P(i|t_1, N, n, \xi; i > 0) \\ &\quad = \int_{t_2=0}^{t_1} \cdots \int_{t_i=0}^{t_{i-1}} P(S|\mathbf{t}; \mu)P(\mathbf{t}|t_1, N, n, \xi; i) \qquad (15) \\ &\qquad \times P(i|t_1, N, n, \xi; i > 0)dt_i \ldots dt_2 , \end{aligned}$$

where the density $P(i|t_1, N, n, \xi; i > 0)$ is given by equation (7) above. The density of (15) is the likelihood function for t_1 based on the random variables S and i, which we will denote as $L(t_1|S, i; N, n, \xi, \mu)$, indicating that the likelihood is a function of the population size N, the sample size n, the joint growth and selection parameter ξ, and the mutation rate μ, as well as the observed random variables S and i.

Monte Carlo Integration of Likelihood

For samples of more than a few individuals, the multiple integral of (15) cannot be evaluated explicitly. A Monte Carlo estimator of the likelihood function is

$$\begin{aligned} &L(t_1|S, i; N, n, \xi, \mu) \qquad (16) \\ &\quad \approx P(i|t_1, N, n, \xi; i > 0)\frac{1}{R}\sum_{j=1}^{R} P(S|\mu; T_j) , \end{aligned}$$

where T_j is the jth random observation of T generated by Monte Carlo simulation methods (see below), conditional on the observation of a total of i alleles of the mutant class in the sample, and R replicate simulations are used to obtain each estimate. The probability of S segregating sites, conditional on T, under an infinite sites model is

$$P(S|\mu;\, T) = \frac{e^{-\mu T}(\mu T)^S}{S!}\,. \tag{17}$$

The Monte Carlo estimate of the likelihood is unbiased and consistent (see Fishman 1996).

It is straightforward to simulate from the joint density t by taking advantage of the property that it is equivalent to the density of the order statistics of $i - 1$ iid random variables with common density $h(z)$. The procedure is to generate, for each replicate of the Monte Carlo integration, a set of $i - 1$ pseudorandom variables $\mathbf{z} = z_1, z_2, \ldots, z_{i-1}$ from the density $h(x)$ and then label these in order of decreasing magnitude (i.e., $z_{[i]} < z_{[i-1]} < \ldots < z_{[2]}$). A random observation from T is then obtained as $T = \Sigma_{j=2}^{i} z_{(j)} + z_{(2)}$. The pseudorandom observations from $h(z)$ may be generated using the inverse transformation method according to the following procedure: (1) generate a uniform (0,1) random variable y; (2) obtain an observation from $h(z)$ using the following transformation (for $\xi > 0$):

$$z = \frac{\ln[\phi - yf/2 + y(\xi)] - \ln[\phi - yf/2]}{\xi}\,, \tag{18}$$

where,

$$\phi = \frac{(f/2)(e^{t_1\xi} - 1) + \xi}{e^{t_1\xi} - 1}\,. \tag{19}$$

For the case $\xi = 0$, we can instead use the transformation:

$$z = \frac{y}{t_1^{-1} + (f/2)(1 - y)}\,. \tag{20}$$

The likelihood function may be evaluated numerically at any point by use of the Monte Carlo procedure described above and can therefore be numerically maximized under the constraint $t_1 \geqslant 0$. For the analyses presented in this paper, the log-likelihood was evaluated at a set of points in the region of the maximum and the program *Mathematica* (Wolfram Research 1992) was then used to fit the points to a polynomial from which a maxima and support interval of the log-likelihood were obtained by maximizing the interpolated function.

Applications

To apply our methods to specific cases, we need the values of several quantities: i, the number of copies of the mutant allele found in a sample, S, the number of segregating sites within the allelic class (or at least an estimate of the number of mutations that have occurred within the class), f, the fraction of the total population that is sampled, μ, the mutation rate, and ξ, the parameter that indicates the combined effects of population growth and selection on the allele of interest. Given these quantities, our method yields the MLE and the support interval of t_1, the time in the past at which the allele arose by mutation. In general, some of these quantities are unknown, but we can estimate t_1 when the unknown parameters lie within reasonable ranges.

In our applications, the values of i and S are known and values of f can be inferred from estimates of the allele frequencies and population sizes. The values of μ and ξ are, in general, unknown. Other genetic data can provide rough estimates of μ. For samples from European populations, it is reasonable to assume some population growth in the recent past, and it is important to consider a range of possible values of ξ to allow for the possibility of selection.

CF

CF is the most common severe genetic disorder in Caucasians, affecting between 1/2,000 and 1/4,000 newborns. The disorder is caused by the recessive effect of mutations at a single locus, the cystic fibrosis transmembrane conductance regulator (*CFTR*) cloned in 1989 (Kerem et al. 1989). This gene spans ~230 kb on chromosome 7 and consists of 27 exons that code for a polypeptide of 1,480 amino acids. Approximately 70% of CF cases are caused by a mutation with a deletion of the 508th codon for phenylananine. This class of mutations, called "ΔF508," is thought to have arisen by mutation only once (Morral et al. 1993) and hence is a single allelic class of the kind we have modeled. Morral et al. (1993, 1994) have examined the extent of variability among different copies of ΔF508 at three microsatellite loci in introns of *CFTR* and used that information to predict that ΔF508 arose ~2,600 generations, or 52,000 years, ago. Kaplan et al. (1994) questioned that conclusion and argued that the extent of disequilibrium with closely linked markers supports the estimate of Serre et al. (1990) of an origin between 3,000 and 6000 years ago, an order of magnitude smaller than the estimate of Morral et al. (1994). The application of our method leads to conclusions similar to those of Kaplan et al. (1994).

Morral et al. (1994) examined 1,705 copies (i = 1,705) of ΔF508 from individuals seeking treatment at clinics throughout Europe. There were 54 different haplotypes, 4 of which accounted for 88.4% of the sample. On the basis of a parsimony analysis of these haplotypes, the three-locus haplotype on which ΔF508 arose could be identified. The parsimony analysis also indicated that there were ~46 mutational events on the intraallelic genealogy. Because the extent of variation at the three microsatellite loci was measured using only allele repeat lengths, the possibility of more than one mutation to alleles of the same size cannot be ignored. In our analysis, we will assume that the number of mutations is $S = 46$ and then consider the possibility that the actual number is higher.

The individuals in the sample analyzed by Morral et al. (1994) were not from a random sample of Europeans, so we do not know f. The frequency of CF-causing mutants is ~.03, and ~70 percent of those mutants are ΔF508, giving a population frequency of ~.02 for ΔF508. Therefore, we would expect to find 1,705 copies of ΔF508 in a sample of total size $n = 1{,}705/0.02 = 85{,}250$. Given a total population size of N individuals, we can define $f = n/(2N)$. The current population of western Europe is ~300 million ($N = 3 \times 10^8$), but individuals in this population were not equally likely to be included in the sample, and many Caucasians are found in other geographic areas, so we considered smaller and larger values of N as well.

We do not know the mutation rates at these three microsatellite loci. Morral et al. (1994) observed no mutations at these loci in 3,000 meioses, which implies that the upper 95% confidence limit on the mutation rate per locus is 3.33×10^{-4}, assuming the rates at the three loci are equal. This upper bound is somewhat lower than the average rate estimated by Weber and Wong (1993) of 1.2×10^{-3}. As Kaplan et al. (1994) point out, in this context only the combined rate at all three sites $\mu = 9.99 \times 10^{-4}$ is relevant to the analysis. Because that is only an upper bound, we consider lower rates as well.

The value of ξ is unknown, but it is almost certainly positive. European populations have grown relatively rapidly, at least in the past 10,000 years, since the spread of agriculture. The mtDNA sequence data of DiRienzo and Wilson (1991) suggests that the period of rapid growth began very recently, possibly only 50,000 years ago (Slatkin and Hudson 1991). The growth rate over this period does not have to be very high in order for the European population to increase to its current size. For example, with an initial population of 5,000 individuals, the population size would increase to 10^9 in 50,000 years, with a growth rate of only 0.00488 per generation if the generation length is 20 years (Slatkin and Hudson 1991). If heterozygotes for ΔF508 had the same fitness as the wild-type homozygotes, then $\xi = r = 0.00488$ would be a reasonable estimate. There has long been a controversy, however, concerning whether CF heterozygotes have a fitness advantage over non-CF homozygotes, thus accounting for the high frequency of CF in Europeans (reviewed by Romero et al. 1989). Recent evidence suggests that CF heterozygotes may have increased resistance to bronchial asthma (Schroeder et al. 1995) and cholera (Gabriel et al. 1994). If the selective advantage of heterozygotes is even as large as 1%, then selection would outweigh the effects of population growth. We consider a range of values of ξ, including $\xi = 0$, to allow for different levels of selection and rates of population growth. In our analysis, we ignore the fact that ΔF508 was until recently effectively a recessive lethal. Because the allele has probably always been in low frequency, the vast majority of copies are found in heterozygous carriers, so the fate of the heterozygotes is much more important than that of the homozygotes.

Figure 3 shows the log-likelihood curves for t_1 under several different combinations of μ, N, and ξ. These four combinations illustrate the general patterns we found in a much larger number of likelihood curves. Figure 3*a* can be regarded as a reference set. The values of the unknown parameters are $\mu = 10^{-4}$, $\xi = 0.005$, and $N = 3 \times 10^8$. The estimated age is $\hat{t}_1 = 146.0$ with a 95% confidence interval of (116.5, 178.2), where time is measured in generations. The value of ξ used in these calculations assumes no selective advantage of the heterozygotes, and the mutation rate is lower than is thought to be typical for microsatellite loci in humans (Weber and Wong 1993). Even with this relatively low mutation rate, the estimated age is on the order of 3,000 years if we assume 20 years per generation. That is lower than the estimate of ~6,000 years by Serre et al. (1990).

Selection in favor of the heterozygotes reduces $\hat{t}_1$ substantially. For example, if ξ is increased to 0.02 (fig. 3*b*), $\hat{t}_1$ is reduced to 80.10 generations. Reducing the population size has a similar effect. For example, if N is reduced to 10^8, $\hat{t}_1$ decreases to 83.9. And, of course, increasing the mutation rate also decreases $\hat{t}_1$. If $\mu = 10^{-3}$, $\hat{t}_1 = 46.1$, a value far too small to be reasonable.

Although our method does not give a precise estimate of the age of ΔF508, because the relevant parameters are not known with sufficient precision, our analysis does show that, unless the number of mutations creating intraallelic variability is vastly greater than is indicated by the analysis of Morral et al. (1994), those data do not support a very early date for the origin of ΔF508. The reason is that there are relatively few mutational events, given the large number of copies of the allele that were examined. Given the relatively high rate of mutation of microsatellites in humans, much more intraallelic variability would be expected than is observed if ΔF508 were actually ≥50,000 years old.

Our conclusion about the age of ΔF508 is not changed if we assume larger values of S, the numbers of mutations. For example, with the parameter values in figure 3*a*, if S is doubled from 46 to 92, $\hat{t}_1$ would only increase to 221.2,

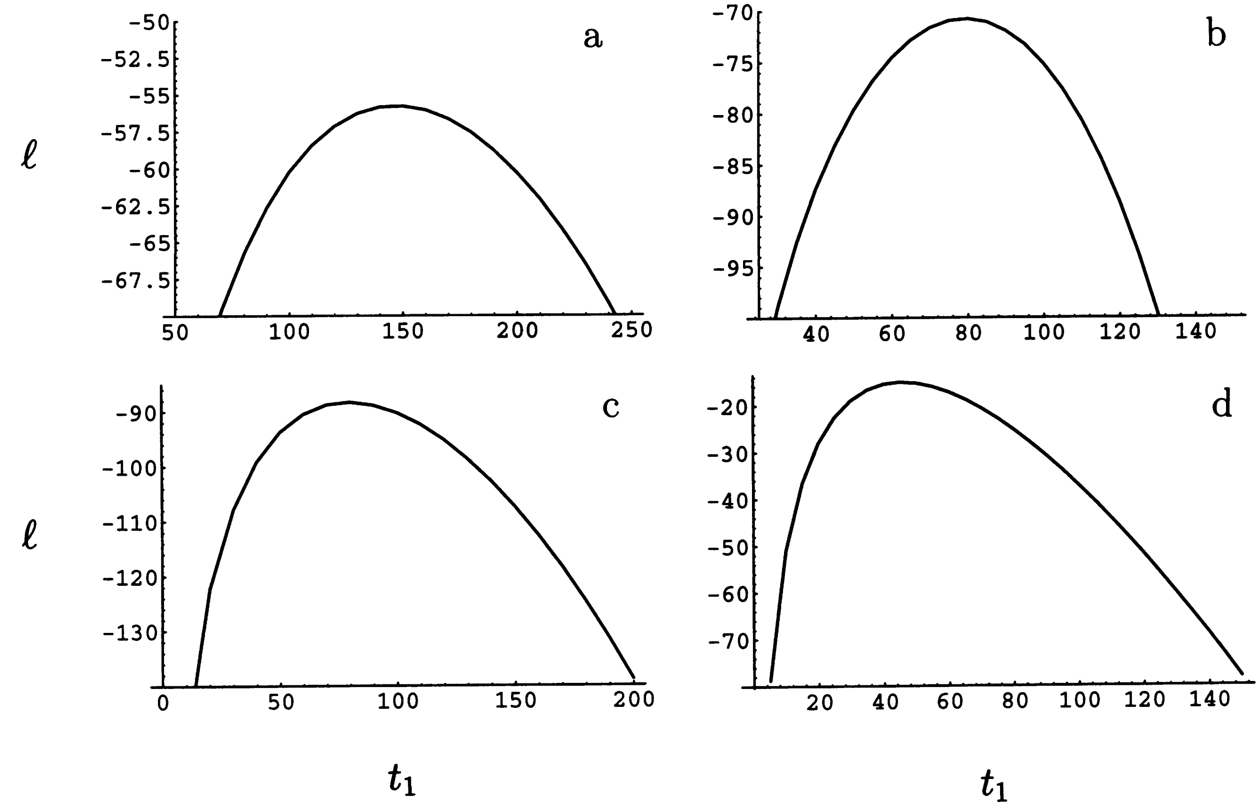

Figure 3 Log-likelihood (l) of the observed number of copies (i = 1,705) and segregating sites (S = 46) for the ΔF508 allele of CF. *Panel a* shows l as a function of t_1 with $\xi = 0.005$, $N = 3 \times 10^8$, $f = 0.00014$, and $\mu = 10^{-4}$. The MLE is $\hat{t}_1 = 146.0$ with the 95% confidence interval (116.5, 178.2). *Panel b* shows l as a function of t_1 for $\xi = 0.02$ and the other parameters as given above. The MLE is $\hat{t}_1 = 80.1$ (67.1, 93.0). *Panel c* shows l as a function of t_1 for $N = 1 \times 10^8$, $f = 0.00043$, and the other parameters as given above. The MLE is $\hat{t}_1 = 83.9$ (68.3, 100.6). *Panel d* shows l as a function of t_1 for $\mu = 10^{-3}$ and other parameters as given above. The MLE is $\hat{t}_1 = 46.1$ (36.7, 56.8).

or ~4,500 years. The reason is that, as t_1 increases, the total tree length, T, increases very rapidly because each of a large number of lineages must become longer. The value of S would have to increase very dramatically in order for $\hat{t}_1$ to increase by even an order of magnitude. Furthermore, any population subdivision, or isolation of subpopulations within Europe, would make our conclusion stronger because subdivision would tend to increase the total tree length for any given value of t_1.

Our results suggest there was probably no strong selection in favor of the heterozygous carriers of ΔF508. With prolonged selection of moderate intensity, on the order of 1%–2% advantage, $\hat{t}_1$ becomes too small to be consistent with its widespread geographic distribution and its relatively high frequency in Basques, who are though to be descended from early inhabitants of Europe (Casals et al. 1993).

We can apply our method to another mutation of *CFTR,* G542X, a nonsense mutation in exon 11, that has been assessed by Casals et al. (1993). The data set is much smaller: i = 62 copies were examined at the same three microsatellite loci used by Morral et al. (1994) with ΔF508. The data of Casals et al. (1993, table 4) indicate that at least $S = 5$ mutations occurred at the three microsatellite loci. The frequency of G542X is much smaller than ΔF508, making up ~8% of the CF alleles found in a survey of Spain. The frequency of G542X varies with location in Spain (Casals et al. 1993), but for the purposes of our analysis we will assume a frequency of .03 × .08 = .0024 in a random sample of the Spanish population. The choice of a population size for our analysis is problematic. The population of Spain is ~37 million (3.7 × 10^7), but Casals et al. (1993) suggest the geographic distribution supports the introduction of G542X from North Africa between 2,000 and 3,000 years ago, which implies that the appropriate population for analysis may be much larger. We used values of N between 10^7 and 10^8. Although we still do not know the mutation rates, the same microsatellite loci were surveyed for both alleles so we can obtain reasonable estimates of the relative ages of the two alleles.

Figure 4 shows the likelihood curves for four combinations of parameter values for G542X. These figures, and others not shown, indicate that G542X is somewhat

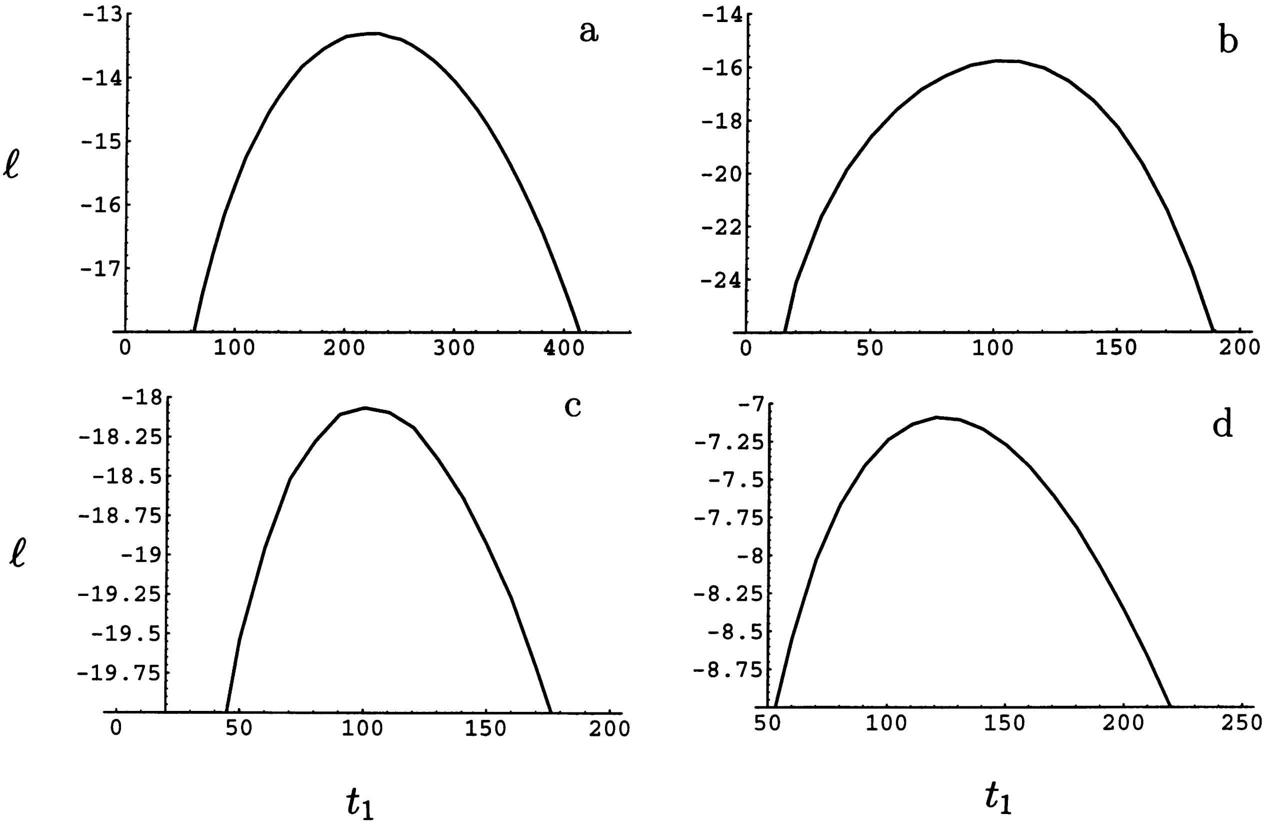

Figure 4 Log-likelihood (l) of the observed number of copies (i = 62) and segregating sites (S = 5) for the G542X allele of CF. *Panel a* shows l as a function of t_1 with $\xi = 0.005$, $N = 4 \times 10^7$, $f = 0.00032$, and $\mu = 10^{-4}$. The MLE is $\hat{t}_1 = 216.9$ with the 95% confidence interval (109.9, 347.7). *Panel b* shows l as a function of t_1 for $\xi = 0.02$ and the other parameters as given above. The MLE is $\hat{t}_1 = 98.7$ (55.7, 144.9). *Panel c* shows l as a function of t_1 for $N = 1 \times 10^7$, $f = 0.0013$, and the other parameters as given above. The MLE is $\hat{t}_1 = 104.5$ (41.5, 178.2). *Panel d* shows l as a function of t_1 for $\mu = 10^{-3}$ and other parameters as given above. The MLE is $\hat{t}_1 = 121.4$ (51.7, 244.7).

older than ΔF508. Although many fewer copies were examined for the G542X allele, the relative amount of intraallelic variability is slightly larger than that found for the ΔF508 allele. The ratio of estimated ages depends on the values of unknown parameters, but, if we assume the same value of ξ for both alleles, then G542X appears to be ~1.5–3 times older than ΔF508. If we assume that the value of ξ for G542X is lower than for ΔF508, reflecting the possibility that ΔF508 may have a heterozygote advantage that G542X lacks, then the ratio of the ages would be even greater.

The inclusion of intraallelic variability leads to a different conclusion about the relative ages of these two mutations than we obtained by using the allele frequency alone. Because G542X is in much lower frequency than ΔF508, (8) implies that it is younger. But our analysis of the extent of intraallelic variability implies that it is somewhat older, and possibly much older.

Discussion

We have shown how to use the theory of birth-death processes to provide an estimate of the age of an allele, using both its observed frequency and the extent of intraallelic variability. The method assumes that all of the intraallelic variation arose by mutation after the mutation that defines the allelic class occurred. That assumption excludes the possibility of intraallelic recombination, which is reasonable for alleles that have arisen recently. Although some parameters of the model are not known with certainty, at present, our results show what parameters strongly affect estimates of allele age. In particular, estimates are very sensitive to the mutation rate, about which some information can be obtained from existing estimates of mutation rates at microsatellite loci.

Our method, of course, relies on simplifying assumptions about the population from which samples are taken and about dynamics of the alleles of interest. We have ignored population subdivision, historical fragmentation of populations, and the possibility that selection affecting the allele may have varied with time. Other methods for estimating allele age, including those of Serre et al. (1990) and Risch et al. (1995), make the same simplifying assumptions. The simplicity of birth and death models suggests that many of these complicating factors could be taken into account.

In this paper, we have concentrated on the number of mutations occurring within an allelic class, but the same approach could be used to predict the probability of any observed configuration of alleles, provided that the mutation model were known. The theory of birth-death processes provides the joint distribution of coalescence times, which in turn may be used to generate the distribution of S as a function of the demographic and mutational parameters. To find the distribution of allelic configurations, we would have to add, at each coalescence, all descendent configurations as equally likely. The resulting analysis would be essentially the same as that of Griffiths and Tavaré (1994), with the primary difference being the distribution of the coalescence times.

Acknowledgments

This project was supported in part by an NIH grant GM40282 (to M.S.) and by a Natural Sciences and Engineering Research Council of Canada postdoctoral fellowship (to B.R.).

Appendix

Relationship between the Birth-Death Process and the Conditional Coalescent

Our method for estimating the age of an allele assumes that the allele of interest is found in low frequency in a sample from a large population. For a neutral allele in a large population, the ancestry of the entire sample can be described by the coalescent process defined by Kingman (1982). Slatkin (1996) showed how to find properties of the genealogy of a neutral allelic class that arose by a single mutation at a known time in the past, conditioned on the number of copies of that allele found in the sample. We will call the genealogical process for the allelic class the "conditional coalescent" and show how the conditional coalescent process can be approximated by a linear birth-death process with appropriate parameter values.

Using the notation of Tavaré (1984), the number of ancestors i of a sample of n copies of a locus at time t in the past is a random variable described by a continuous time Markov chain with state space $\{1, 2, \ldots, n\}$:

$$\frac{dp_i(t)}{dt} = -\frac{i(i-1)}{4N(t)}p_i(t) + \frac{(i+1)i}{4N(t)}p_{i+1}(t)\,, \quad \text{(A1)}$$

for $i = 1, 2, \ldots, n$, where $p_i(t)$ is the probability of being in state i at time t in the past, $p_{n+1}(t) = 0$, and $N(t)$ is the population size at time t in the past. The population size is assumed to be large enough that only one coalescence event can occur in any generation. We allow for arbitrary deterministic changes in population size because time can be rescaled using (from Griffiths and Tavaré 1994)

$$\tau(t) = \int_0^t \frac{dt'}{2N(t')}\,. \quad \text{(A2)}$$

With the initial condition, $p_n(0) = 1$ and $p_i(0) = 0$ for $i < n$, a complete solution can be obtained (Tavaré 1984). From this solution, $x(t)$, the expected value of i at t, can be expressed as an infinite sum involving ratios of factorials (Tavaré 1984, eq. 6.7).

The general formula for $x(t)$ is not useful here, but there is a simple approximation that can be obtained directly from (A1). By multiplying each side of (A1) by i and summing over i, we obtain

$$\frac{dx(\tau)}{d\tau} = -\frac{x(\tau)[x(\tau)-1]}{2} - \frac{\sigma_i^2(\tau)}{2}\,, \quad \text{(A3)}$$

where $\sigma_i^2(\tau)$ is the variance of i at $\tau(t)$. If we assume that σ_i^2 is small relative to x during the times of interest, then we can ignore the second term in (A3). By further assuming that $x \gg 1$, we obtain

$$\frac{dx(\tau)}{d\tau} = -\frac{x^2(\tau)}{2}\,, \quad \text{(A4)}$$

which has the solution

$$x(t) = \frac{n}{1 + n\tau(t)/2} \quad \text{(A5)}$$

for the initial condition $x(0) = n$. Numerical analysis shows that (A5) provides an excellent approximation to the exact value provided that n is large and that t is sufficiently small that $x(t)$ remains large.

The genealogy of an allelic class that arises by mutation at time t_1 in the past is obtained by randomly choosing one the of i ancestral lineages at t_1 and focusing on the gene genealogy of all descendents (Slatkin 1996). Considering only the number of descendents at time t between t_1 and 0, the number of mutants follows a pure birth process beginning with one copy at t_1. There is no death, because every lineage leaves one or more descendents at $t = 0$, the present. We can find the birth rate for this process by considering the time interval (t, $t - \delta t$). Assume that there are j mutant lineages of the i lineages at t. We assume that i is equal to its expectation, $x(t)$, given by (A5). After a small time δt, $x(t)$ will increase to $x(t - \delta t) \approx x(t) + x^2(t)\delta t/[4N(t)]$, and the fraction of those the lineages that carry the mutant is

$j/x(t)$. Therefore, the increase in the number of mutant lineages is approximately

$$\frac{x(t)}{4N(t)} j\delta t , \tag{A6}$$

and the instantaneous birth rate for the process describing the number of mutant lineages is $b(t) = x(t)/[4N(t)]$, where $x(t)$ is given by (A5).

We will be concerned with two special cases, a population of constant size and a population that has grown exponentially in the past. For a population of constant size, $x(t) = n/[1 + nt/(4N)]$ and

$$b(t) = \frac{f/2}{1 + ft/2} , \tag{A7}$$

where $f = n/(2N)$ is the fraction of the population sampled. For exponential growth at rate r in the past, $N(t) = N_0e^{-rt}$, $\tau(t) = (e^{rt} - 1)/(2N_0r)$, and

$$b(t) = \frac{fr}{f - (f - 2r)e^{-rt}} . \tag{A8}$$

where $f = n/(2N_0)$.

We now show that the birth process for the numbers of mutant lineages is equivalent to the reconstructed birth-death process. Nee et al. (1994) summarize the theory of linear birth-death processes developed by Kendall (1948) and others and extend that theory to apply to the case of interest here, where a fraction f of the total population is sampled. We assume a linear birth-death process with birth rate B and death rate D. To predict the numbers of mutant lineages that leave one or more descendents in the sample, the appropriate model is a pure birth process with birth rate $BP(0, t)$, where $P(0, t)$ is the probability that an individual present at time zero leaves one or more descendents t generations later, when the sample is taken (Nee et al. 1994). The function $P(0, t)$ is derived by Nee et al. (1994) for the case when a fraction of the population is sampled:

$$BP(0, t) = \frac{fB}{1 + fBt} , \tag{A9}$$

when $B = D$ and

$$BP(0, t) = \frac{f\xi}{f - (f - 2\xi)e^{-B\xi t}} , \tag{A10}$$

when $D = B(1 - \xi/2)$. The factor 2 is used in the definition of ξ in order to make later notation simpler.

Equation (A9) is appropriate for a population of constant size, and we can see by comparison with (A7) that the birth rate for the reconstructed process corresponds to that for the number of mutant lineages provided that $B = 1/2$. Similarly, (A10) is appropriate for a population that has been growing exponentially in size, provided $\xi > 0$. Equation (A8) shows that the birth rate for the two models are equivalent when $B = 1/2$ and $\xi = r$.

Thus, we have demonstrated that the numbers of mutant lineages in a coalescent model can be approximated by a linear birth-death process, provided that n is large, t is relatively small, and the parameters of the birth-death process are scaled appropriately. The factor of $1/2$ enters for the same reason that there is a factor of $1/2$ difference between many properties of the Moran model and the Wright-Fisher model (Ewens 1979). That factor is necessary to make the time scâle of the birth and death model correspond to generations in the coalescent model.

References

Casals T, Nunes V, Palacio A, Gimenez J, Gaona A, Ibanez N, Morral N (1993) Cystic fibrosis in Spain: high frequency of the mutation G542X in the Mediterranean coastal area. Hum Genet 91:66–70

Cox DR, Hinkley DV (1974) Theoretical statistics. Cambridge University Press, Cambridge

DiRienzo A, Wilson AC (1991) The pattern of mitochondrial DNA variation is consistent with an early expansion of the human population. Proc Nat Acad Sci USA 88:1597–1601

Donnelly P, Tavaré S (1986) The age of alleles and a coalescent. Adv Appl Prob 18:1–19

Edwards AWF (1972) Likelihood. Cambridge University Press, Cambridge

Ewens WJ (1979) Mathematical population genetics. Springer, New York

Fishman GS (1996) Monte Carlo: concepts, algorithms, and applications. Springer, New York

Gabriel SE, Brigman KN, Koller BH, Boucher RC, Stutts MJ (1994) Cystic fibrosis heterozygote resistance to cholera toxin in the cystic fibrosis mouse model. Science 266:107–109

Griffiths RC, Tavaré S (1994) Sampling theory for neutral alleles in a varying environment. Phil Trans R Soc Lond B 344:403–410

Hudson RR (1990) Gene genealogies and the coalescent process. Oxf Surv Evol Biol 7:1–44

Kaplan NL, Lewis PO, Weir BS (1994) Age of the Δ-F508 cystic fibrosis mutation. Nat Genet 8:216–218

Kendall DG (1948) On the generalized birth-and-death process. Ann Math Stat 19:1–15

Kerem B, Rommens JM, Buchanan JA, Narkiewicz D, Cox TK, Chakravarti A, Buchwald M, et al (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245: 1073–1080

Kingman JC (1982) On the genealogy of large populations. J Appl Prob 19A:27–43

Morral N, Bertranpetit J, Estivill X, Nunez V, Casals T, Giménez J, Reis A, et al (1994) The origin of the major cystic fibrosis mutation (Δ-*F508*) in European populations. Nat Genet 7:169–175

Morral N, Nunez V, Casals T, Chillon M, Giménez J, Bertranpetit J, Estivill X (1993) Microsatellite haplotypes for cystic fibrosis: mutation frameworks and evolutionary tracers. Hum Mol Genet 2:1015–1022

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

Rannala B. Gene genealogy in a population of variable size. Heredity (in press)

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

Romero G, Devoto M, Balletta LJG (1989) Why is the cystic fibrosis gene so frequent? Hum Genet 84:1–5

Sawyer S (1979) On the past history of an allele now known to have frequency *p*. J Appl Prob 14:439–450

Schroeder SA, Gaughan DM, Swift M (1995) Protection against bronchial asthma by *CFTR* Δ-F508 mutation: a heterozygote advantage in cystic fibrosis. Nat Med 1:703–705

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 population genetics. Hum Genet 84:449–454

Slatkin M (1996) Gene genealogies within mutant allelic classes. Genetics 143:579–587

Slatkin M, Hudson RR (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and growing populations. Genetics 129:355–362

Tavaré S (1984) Line-of-descent and genealogical processes, and their applications in population genetics models. Theor Popul Biol 26:119–164

Thompson EA (1975) Human evolutionary trees. Cambridge University Press, Cambridge

——— (1976) Estimation of age and rate of increase of rare variants. Am J Hum Genet 28:442–452

Watterson GA (1976) Reversibility and the age of an allele. Theor Popul Biol 10:239–253

Weber JL, Wong C (1993) Mutation of human short tandem repeats. Hum Mol Genet 2:1123–1128

Wolfram Research (1992) Mathematica, version 2.2. Wolfram Research, Champaign, IL

Yang Z. Statistical properties of a DNA sample under the finite-sites model. Genetics (in press)