Theoretical Population Biology **56**, 183–201 (1999)
Article ID tpbi.1998.1411, available online at http://www.idealibrary.com on IDEAL®

Conditional Genealogies and the Age of a Neutral Mutant

Carsten Wiuf[1]

Institute of Biological Sciences, University of Aarhus, Ny Munkegade, Aarhus DK-8000, Denmark

and

Peter Donnelly

Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, England

Received May 23, 1998

This paper is concerned with the structure of the genealogy of a sample in which it is observed that some subset of chromosomes carries a particular mutation, assumed to have arisen uniquely in the history of the population. A rigorous theoretical study of this conditional genealogy is given using coalescent methods. Particular results include the mean, variance, and density of the age of the mutation conditional on its frequency in the sample. Most of the development relates to populations of constant size, but we discuss the extension to populations which have grown exponentially to their present size.

Key Words: **age of allele; coalescent model; conditional genealogies.**

1. INTRODUCTION

Suppose we examine a sample of chromosomes at a given locus and find that some subset of them shares a particular mutation, which is assumed to have arisen only once in the population's evolutionary history. This paper concerns the structure of the genealogical tree which describes the ancestral relationships among the subsample carrying the particular mutation, under the assumption that the mutation is neutral.

The relationship between genealogical structure and patterns of genetic variation is now well established. Indeed, so-called "coalescent methods," which exploit this relationship, are proving to be a powerful tool in the modelling and statistical analysis of molecular genetic data from within species. See, for example, Donnelly and Tavaré (1995) or Hudson (1990) for reviews. In the special setting considered here, aside from its inherent interest, an understanding of the appropriate genealogy is seen as an important first step in understanding the patterns of variability to be expected at loci linked to the one at which the mutation occurred, or in understanding the distribution of the size of the chromosomal regions around the mutation site shared, identical by descent, between the chromosomes in the subsample.

In addition to population modelling, two important classes of statistical questions arise from data of the sort we are considering. The first relates to the uses of variability at linked loci (so-called inter-allelic variability) to estimate the age of the mutation, or to examine hypotheses about the demographic history of the population in which it is found. The second (related) class of questions arises in studies of the patterns of linkage disequilibrium at nearby loci, and the extent to which these are

[1] Current address: Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, England.

0040-5809/99 $30.00

informative on aspects of population history. Our aim here is to provide a rigourous study of the genealogy of the alleles carrying the mutation. This should then subsequently provide a solid base for the development of efficient inference techniques for statistical questions such as those just described.

One further motivation for this study comes from the setting in which the mutation results in a predisposition to a particular disease. Here there are also applications to gene mapping, and subsequent positional cloning. After linkage to a particular region has been established, relevant data for fine mapping might consist of the lengths of chromosomal segments shared between affected individuals, with the goal of inference about the location of a posited "disease mutation" (under specific assumptions about population history). The results here, which assume that the mutation is neutral, may not be directly applicable to this problem, and we aim to consider the selected case elsewhere. Nonetheless, rigourous results in the neutral case may give useful intuition more generally. Further, approximate methods developed for the general setting might usefully be "tested" by comparing their predictions, in the special case of neutrality, with results given here.

One specific aspect of the genealogy of the subsample is the age of the mutation. We derive the distribution of this, conditional on the number of sampled alleles carrying the mutation. Kimura and Ohta (1973) used diffusion approximations to derive the mean and variance of the age in the case in which the sample coincided with the whole population. Slatkin and Rannala (1997) developed an approximate analysis in which they derived a maximum likelihood estimator for the age of the mutation on the basis of inter-allelic variability. One very simple special case of their analysis corresponds to a neutral mutant in which the estimation is based only on the fraction of alleles in the sample carrying the mutation. In the Discussion we compare their estimator in this setting with the formula (exact within the coalescent framework) for the mean age of the mutation. Particularly when the mutation is rare (frequency less than 0.2 say) in the sample, the Slatkin–Rannala estimator is smaller than the conditional mean age, with a relative underestimation by up to a factor of 4 or so for plausible frequencies of the mutant allele. Approximations related to those employed in Slatkin and Rannala (1997) are used in assessing likely levels of linkage disequilibrium in Thompson and Neel (1997), for example.

Throughout, we adopt the coalescent as an exact description of the genealogy of the population (or equivalently of samples from it). The approximation is good provided that the population is large (Kingman, 1982b). one view is that our analysis is exact for the Moran model of demography (Moran, 1958). Another is that the coalescent approximation is effectively equivalent to the classical diffusion approximations in population genetics. In the coalescent approximation time is measured in units of N generations, where N is the (variance) effective number of chromosomes in the population (Kingman, 1982b).

To fix some notation, we consider the setting in which we sample n chromosomes from the population. At a particular locus, the sample is divided into two subsamples $\mathscr{D}$ and $\mathscr{C}$ of size i and $k \equiv n - i$, respectively, with the property that all of the chromosomes in $\mathscr{D}$ (and none of those in $\mathscr{C}$) share a particular mutation. We assume in addition that the mutation is neutral, and that it has arisen only once in the history of the population. This would be reasonable for a mutation event (such as a single base change) for which the *a priori* mutation rate is very small. We consider this setting. In fact, formally, we examine the limit as the mutation rate tends to zero, conditional on a mutation having occurred.

We proceed in several stages. The event F that exactly those chromosomes in $\mathscr{D}$ share the mutation can be decomposed into two parts. The first of these, which we denote by E, is the event that all the chromosomes in $\mathscr{D}$ share a common ancestor before any of them shares a common ancestor with a chromosome from $\mathscr{C}$. The second (assuming E) is that a mutation event, giving rise to the mutation in question, occurred on the ancestral lineage common to all of $\mathscr{D}$ between the time of the most recent common ancestor (MRCA) of $\mathscr{D}$ and the time at which $\mathscr{D}$ first shares an ancestor with $\mathscr{C}$. We will denote this latter event by M. It is evident that the conditioning event F is equivalent to $E \cap M$.

Our approach is first to examine the effect of conditioning on E, and then to condition on M. The conditioning on E affects only the jump chain of the coalescent for the sample, and not the times between coalescent events. In the next section we derive the topological structure of the genealogy of $\mathscr{D}$ and $\mathscr{C}$, jointly, back to the time of the MRCA of $\mathscr{D}$. We subsequently obtain the joint distribution of the times τ and γ, where τ is the time back to the MRCA of $\mathscr{D}$ and γ is the time between this MRCA and the first time at which any of $\mathscr{D}$ and $\mathscr{C}$ share a common ancestor. The effect of the subsequent conditioning on M is effectively like size-biased sampling: the mutation had to occur during the time γ and the probability of this (for small mutation rate) is proportional to γ, so it is more likely to have arisen in a genealogy in which γ is large than in one in which it is small. For most of the analysis we assume a constant population size. This assumption can be weakened, and we treat explicitly the particular

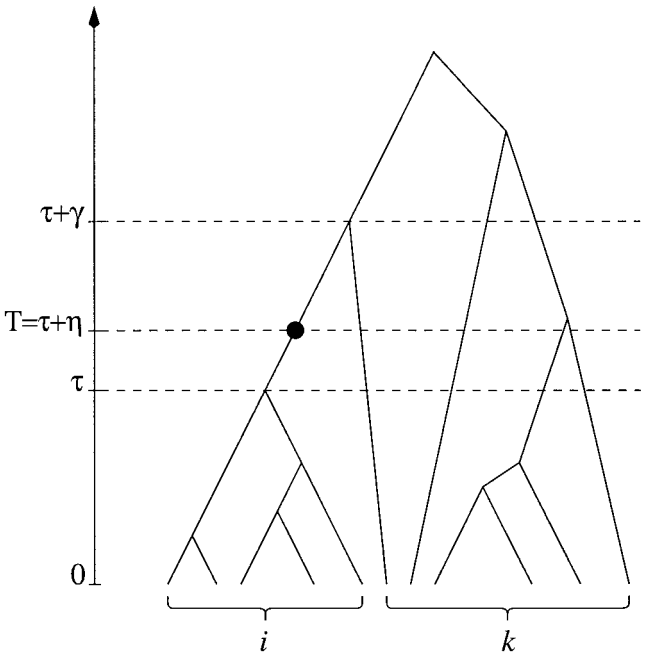

FIG. 1. Notation. In this example subsample $\mathscr{D}$ consists of $i=5$ individuals and subsample $\mathscr{C}$ of $k=n-i=6$ individuals. Time is measured backwards with zero denoting present time. Subsample $\mathscr{D}$ finds a MRCA at time τ, and there are five ancestors to the whole sample at that time, the ancestor to $\mathscr{D}$ and four ancestors to $\mathscr{C}$. At time $\tau+\gamma$, the last line of descent of subsample $\mathscr{D}$ coalesces with an ancestral line of $\mathscr{C}$. This is the second coalescence event further back in time from τ, i.e., $J=2$ (this and the rest of the text concerns notation that will be introduced later). Note that $\alpha=4+1=5$. The genealogy G_1 is below the dotted line at time τ, and G_2 is above the dotted line at time τ. The mutation shared in $\mathscr{D}$ is marked with a dot. The time from the MRCA of subsample $\mathscr{D}$ until the mutation event is denoted by η, and the total time from the present back to the mutation event is denoted by T.

case of a population which has grown exponentially in size to its current value.

Figure 1 illustrates the notation just described as well as some of that introduced in subsequent sections.

Griffiths and Tavaré (1998) also consider the problem of the distribution of the age of a mutant allele, amongst other things. Their approach is different from the one we adopt here. It cleverly exploits the marginal distribution of the jump chain of Kingman's coalescent and so applies, as do our results from Section 2, to any genealogical tree in which each coalescence involves exactly two existing lineages, with each possible choice of lineages to coalesce being equally likely. In the light of our eventual interest in statistical inference based on linked variation in this setting (for the age of the mutation, its location, or population demography) we focus instead on the conditional distribution of the (entire) genealogy of the mutant class. One specific application is to results on the age of the mutation (which of course agree with those of Griffiths and Tavaré (1998), when specialised to the current setting, and earlier of Kimura and Ohta (1973)). Stephens (1999) uses an approach similar to that of Griffiths and Tavaré (1998) in studying the age of the mutation for general mutation rate.

2. THE STRUCTURE OF THE JUMP CHAIN

In this section we examine the jump chain of the coalescent, specifically the joint topological structure of the ancestral trees of the subsamples $\mathscr{D}$ and $\mathscr{C}$, conditional on the event E. Saunders *et al.* (1984) discuss the ancestry of nested subsamples in the coalescent in the absence of specific conditioning. Slatkin (1996) also contains a discussion of subsamples in the coalescent, but there the analysis is not performed conditional on the subsample size, as in our approach.

Recall (Kingman, 1982a) that the probabilistic structure of the coalescent may be decomposed into two independent pieces. One of these is the jump chain of the coalescent, that is, the topological structure of the genealogical tree, which records which pairs of branches coalesce at each coalescence event. The second part is the sequence of times between coalescence events. The event E described above, that all the chromosomes in $\mathscr{D}$ share a common ancestor before any of them share a common ancestor with a chromosome from $\mathscr{C}$, depends only on the jump chain of the coalescent. It follows that conditioning on E affects only the jump chain.

Consider tracing, jointly, the number of ancestors of the two subsamples, $\mathscr{D}$ and $\mathscr{C}$, in the sample. Initially these have i and k ancestors. Write $q_{ik,\,jl}$ for the probability that when the number of ancestors changes from i and k, there will be j and l ancestors, respectively, of the two subsamples. It follows from the structure of the coalescent that the conditional probability that a coalescence will involve two ancestors from $\mathscr{D}$, respectively $\mathscr{C}$, is

$$q_{ik,\,(i-1)k}=\frac{i(i-1)}{n(n-1)},$$
$$q_{ik,\,i(k-1)}=\frac{k(k-1)}{n(n-1)}, \tag{1}$$

and

$$1-q_{ik,\,(i-1)k}-q_{ik,\,i(k-1)}=\frac{2ki}{n(n-1)}$$

is the probability that one individual from each subsample shares a common ancestor.

From (1) we can derive the probability $Q(i, k)$ of the event E. The quantity $Q(i, k)$ is given by the following recursion formula

$$(k+1)(k+i-1)\,Q(i,k) = i(i-1)\,Q(i-1,k) + k(k-1)\,Q(i,k-1)$$

with boundary condition $Q(1, k) = 1$, and $k, i \geqslant 1$. This recursion eventually leads to

$$Q(i,k) = \frac{2(i-1)!}{(i+1)(k+1)_{(i-1)}}, \tag{2}$$

where we have used the notation $a_{(x)} = a(a+1)\cdots(a+x-1)$.

Since there are a total of $\tau(n) = n!(n-1)!/2^{n-1}$ different topologies of a sample of size n, we have $\tau(n)\,Q(i,k) = (k+i)!\,k!(i-1)!/(i+1)\,2^{k+i-2}$ different topologies fulfilling the constraint which defines E.

Now consider the numbers of ancestors of the two subsamples $\mathscr{D}$ and $\mathscr{C}$ conditional on the event E. Let us first consider the jump chain until $\mathscr{D}$ has found a most recent common ancestor. This jump chain has transition probabilities $r_{jl,\,j'l'}$ given by

$$\begin{aligned} r_{jl,\,(j-1)l} &= \frac{j(j-1)\,Q(j-1,l)}{R(j,l)} = \frac{j+1}{l+j}, \\ r_{jl,\,j(l-1)} &= \frac{l(l-1)\,Q(j,l-1)}{R(j,l)} = \frac{l-1}{l+j}, \end{aligned} \tag{3}$$

with

$$R(j,l) = j(j-1)\,Q(j-1,l) + l(l-1)\,Q(j,l-1)$$

and $j > 1$ and $l > 1$. The transition probabilities are asymmetric in j and l due to the asymmetry in this conditional model: the chromosomes in $\mathscr{D}$ must find a common ancestor before coalescing with any chromosome ancestral to $\mathscr{C}$, whereas no such constraint applies to $\mathscr{C}$.

Consider now the jump chain further back in time: prior to the time of the MRCA of the chromosomes in $\mathscr{D}$ the process evolves according to an unconditional coalescent model. When there is a jump, the probability that the common ancestor to $\mathscr{D}$ is involved in the coalescence event is given by $2/(l+1)$, if there are $l+1$ ancestral individuals in total. Hence (3) can be extended to all $j \geqslant 1$ and $l > 1$, and describes the jumps in the history of the subsample $\mathscr{D}$ until the MRCA to $\mathscr{D}$ is absorbed into the rest of the ancestral sample. Prior to this time we put $r_{0l,\,0(l-1)} = 1$, for $l > 1$, with (0,1) being absorbing.

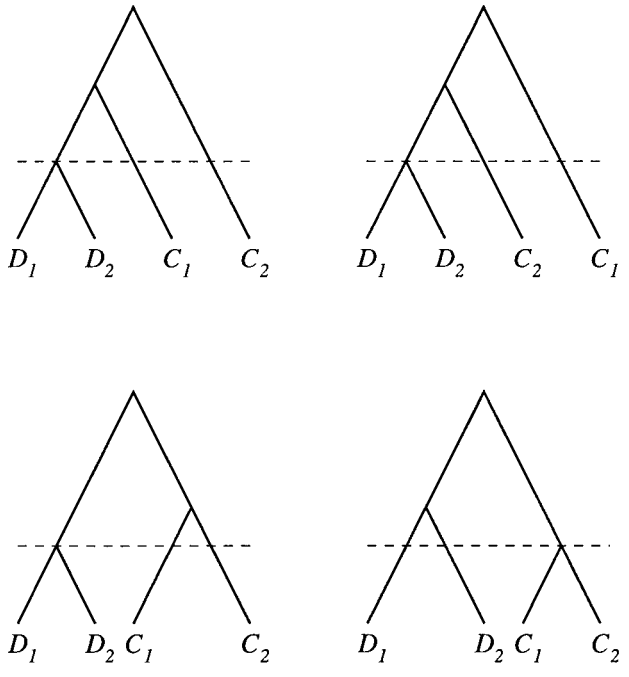

FIG. 2. Conditional genealogies. Here $i = 2$ and $k = n - i = 2$. Subsample $\mathscr{D}$ is labeled D_1 and D_2, and subsample $\mathscr{C}$ is labeled C_1 and C_2. There are four different topologies where subsample $\mathscr{D}$ finds a most recent common ancestor before any ancestor to subsample $\mathscr{D}$ coalesces with any ancestor to subsample $\mathscr{C}$. In three of these, D_1 and D_2 share a common ancestor before C_1 and C_2 share one. Hence conditional on E, the probability of this event is 3/4.

An illustration of the jump chain conditional on the event E is found in Fig. 2.

Until further notice, we condition on E. Let us introduce the following notation:

$A(t)$ = number of ancestors to the entire sample t generations before present,

$A_{\mathscr{D}}(t)$ = number of ancestors to subsample $\mathscr{D}$, t generations before present,

$$T(m) = \min\{t \mid A(t) = m\},$$

$$T_{\mathscr{D}}(j) = \min\{t \mid A_{\mathscr{D}}(t) = j\},$$

$$A^*(j) = A(T_{\mathscr{D}}(j)),$$

$$A^*_{\mathscr{D}}(m) = A_{\mathscr{D}}(T(m)).$$

(This notation is similar to that of Saunders *et al.* 1984.) The Markov chain $(A_{\mathscr{D}}(t), A(t))$ counts the number of ancestors to $\mathscr{D}$ and the whole sample at time t conditional on E. The time $T_{\mathscr{D}}(j)$ is the first time there are j ancestors of $\mathscr{D}$, and $A^*(j)$ is the number of ancestors of

the entire sample at this time. Similarly $T(m)$ is the first time there are m ancestors of the entire sample, and $A^*_{\mathscr{D}}(m)$ is the number of ancestors of the subsample $\mathscr{D}$ at this time. The independence of the jump chain and the times between events in the coalescent means that the $A^*_{\mathscr{D}}$'s are independent of the times $T(\cdot)$. The random variables A^* and $A^*_{\mathscr{D}}$ are related through

$$A^*(j)=m \Leftrightarrow A^*_{\mathscr{D}}(m)=j \qquad \text{and} \qquad A^*_{\mathscr{D}}(m+1)=j+1. \tag{4}$$

All results below are derived conditional on $A_{\mathscr{D}}(0)=i$, $A(0)=n=k+i$ (in addition to the conditioning on E), and henceforth this will be suppressed in the notation. The first lemma is effectively just a restatement of (3).

LEMMA 1. *The sequence* $A^*_{\mathscr{D}}(m)$, $m=n, \ldots, 1$, *is Markov, with transition probabilities given by*

$$\begin{aligned} P(A^*_{\mathscr{D}}(m-1)=j-1 \mid A^*_{\mathscr{D}}(m)=j) &= 1-P(A^*_{\mathscr{D}}(m-1)=j \mid A^*_{\mathscr{D}}(m)=j) \\ &= \frac{j+1}{m} \end{aligned}$$

for $(m-1)\wedge i \geqslant j \geqslant 1$, *and*

$$P(A^*_{\mathscr{D}}(m-1)=0 \mid A^*_{\mathscr{D}}(m)=0)=1$$

for $j=0$.

Proof. Note that the jump chain in (3) is given by $(A_{\mathscr{D}}(T(m)), A(T(m)))=(A^*_{\mathscr{D}}(m), m)$, so that the result follows from (3). ∎

The transition probabilities are thus independent of the initial subsample sizes, and dependent on the present sample only.

Lemma 1 implies that the probability of a path $A^*_{\mathscr{D}}(n)$, $A^*_{\mathscr{D}}(n-1)$, ..., $A^*_{\mathscr{D}}(m)$ with $A^*_{\mathscr{D}}(n)=i$ and $A^*_{\mathscr{D}}(m)=j$, $j\geqslant 0$, fixed is independent of the path. This has as a consequence that all subtopologies describing the history of a sample until time $T(m)$ and conditional on $A^*_{\mathscr{D}}(m)=j$ have an equal chance to occur. This is not true for the part of the topology further back in time than $T(m)$ (unless $j=0$).

The next lemma shows the distribution of $A^*_{\mathscr{D}}(m)+1$ to be almost hypergeometric except that the outcomes 0 and 1 are pooled.

LEMMA 2. *The distribution of* $A^*_{\mathscr{D}}(m)$, $m=1, \ldots, n$, *is given by*

$$P(A^*_{\mathscr{D}}(m)=j)=\frac{\binom{n-m}{i-j}\binom{m}{j+1}}{\binom{n}{i+1}} \tag{5}$$

for $j=1, \ldots, (m-1)\wedge i$, *and*

$$P(A^*_{\mathscr{D}}(m)=0)=P(0)+P(-1),$$

with $P(j)$ *being the probability expression in* (5) *evaluated at* j.

Proof. If $m=n$ then $A^*_{\mathscr{D}}(n)=i$, which can be written in the form above. Consider $j\geqslant 1$ and $m<n$. From Lemma 1 we have that $A^*_{\mathscr{D}}(n-1), \ldots, A^*_{\mathscr{D}}(m)$ can be considered as $n-m$ draws without replacement from an urn with initially $i+1$ white balls, n balls in total. Each white ball drawn reduces by one the number of ancestors of $\mathscr{D}$, so that we require exactly $i-j$ white balls and $n-m-i+j$ other balls to be drawn for $A^*_{\mathscr{D}}(m)=j$. The probability of this is straightforward, for example, from the hypergeometric distribution, and rearrangement then gives the desired probability, (5). Since the sum of the probabilities of all outcomes is one, it must be that the probability for $j=0$ is given by $P(0)+P(-1)$. ∎

Now define $a_{[x]}\equiv a(a-1)\cdots(a-x+1)$. Then we have:

COROLLARY 1. *The sequence* $A^*(j)$, $j=i, i-1, \ldots, 0$, *is Markov, with transition probabilities given by*

$$P(A^*(j-1)=l \mid A^*(j)=m)=\binom{l}{j}\binom{m}{j+1}^{-1},$$

for $l=j, j+1, \ldots, m-1$.

Proof. That $A^*(j)$ is Markov follows from (4) and Lemma 1, and using (4) and the Markov property we have

$$\begin{aligned} P(A^*(j-1)=l \mid A^*(j)=m) &= P(A^*_{\mathscr{D}}(l)=j-1 \mid A^*_{\mathscr{D}}(l+1)=j)\, P(A^*_{\mathscr{D}}(l+1) \\ &= j \mid A^*_{\mathscr{D}}(m)=j). \end{aligned}$$

Applying Lemma 1 gives the desired result. ∎

COROLLARY 2. *The distribution of* $A^*(j)$, $j=0, ..., i-1$, *is given by*

$$P(A^*(j)=m)=\frac{\binom{n-m-1}{i-j-1}\binom{m}{j+1}}{\binom{n}{i+1}}$$

for $m=j+1, ..., j+k$. *For* $j=i$ *we have* $A^*(i)=n$.

Proof. From (4) and Lemma 1 we have

$$P(A^*(j)=m)=P(A^*_{\mathscr{D}}(m)=j, A^*_{\mathscr{D}}(m+1)=j+1)$$
$$=\frac{j+2}{m+1}P(A^*_{\mathscr{D}}(m+1)=j+1),$$

and the result follows from Lemma 2. ▮

Two of the distributions in Corollary 2 are of particular interest: The first is the distribution of the number of ancestors when the subsample $\mathscr{D}$ has found its MRCA ($j=1$), and the other is the distribution of the number of ancestors when this last ancestral line of $\mathscr{D}$ is absorbed into the rest of the ancestral sample ($j=0$).

COROLLARY 3. *The mean and variance of* $A^*(j)$, $j=0, ..., i$, *are given by*

$$E(A^*(j))=\frac{(k-1)(j+2)}{i+2}+j+1,$$

and

$$\mathrm{Var}(A^*(j))=\frac{(k-1)(k+i+1)(i-j)(j+2)}{(i+2)^2(i+3)}.$$

Proof. Simply evaluate the mean and variance of the distributions in Corollary 2. ▮

LEMMA 3. *Conditional on the event* $\{A^*(1)=\alpha\}$, $A^*_{\mathscr{D}}$ *forms a Markov chain, with transition probabilities*

$$P(A^*_{\mathscr{D}}(l-1)=j-1 \mid A^*_{\mathscr{D}}(l)=j, A^*(1)=\alpha)=\frac{j-2}{l-\alpha-1}$$

for $l=\alpha+2, ..., n$ *and* $j=2, ..., (l-\alpha+1)\wedge i$. *If* $l=\alpha+1$ *then* j *must be* 2, *and then with probability one,* $A^*_{\mathscr{D}}(l-1)=A^*_{\mathscr{D}}(\alpha)=1$.

Proof. Let $\beta(j, l)$ be the number of paths of $A^*_{\mathscr{D}}$ with initial subsample sizes j and $l-j$, respectively, that fulfil $A^*(1)=\alpha$. Then we have $\beta(j, l)=\beta(j-1, l-1)+\beta(j, l-1)$, and moreover $\beta(j, l)$ is given by the number of ways $l-\alpha-1$ coalescent events can be arranged into two groups, one of size $j-1$. The final coalescent event is fixed, reducing the subsample size from 2 to 1. Since

$$P(A^*_{\mathscr{D}}(l-1)=j-1 \mid A^*_{\mathscr{D}}(l)=j, A^*(1)=\alpha)$$
$$=\beta(j-1, l-1)/\beta(j, l)$$

(the remark below Lemma 1), the result follows by evaluating the quotient. ▮

In a close analogy with Lemma 2, the conditional distribution of $A^*_{\mathscr{D}}(l)$ given $A^*(1)$ can also be derived explicitly:

COROLLARY 4. *The conditional probability of* $A^*_{\mathscr{D}}(l)=j$ *given* $A^*(1)=\alpha$ *is given by*

$$P(A^*_{\mathscr{D}}(l)=j \mid A^*(1)=\alpha)=\frac{\binom{n-l}{i-j}\binom{l-\alpha-1}{j-2}}{\binom{n-\alpha-1}{i-2}}$$

for $l=\alpha+2, ..., n$ *and* $j=2, ..., (l-\alpha+1)\wedge i$.

Proof. Similar to the proof of Lemma 2, with $n-l$ draws from an urn with initially $i-2$ white balls and $n-\alpha-1$ balls in total. ▮

3. THE SAMPLE COALESCENT CONDITIONED ON *E*

As noted above, the event E depends only on the jump chain of the coalescent, and is independent of the respective waiting times $W_n, W_{n-1}, ..., W_2$ between coalescent events in the ancestry of the sample (where W_k is the time during which the sample has exactly k ancestors). It follows that conditional on E, these waiting times are still independent exponential random variables with $E(W_k)=2/(k(k-1))$. In particular they are independent of $A^*_{\mathscr{D}}$.

To handle the waiting times, we introduce the following notation, which is illustrated in Fig. 1. Let G be the genealogy of the whole sample until its most recent common ancestor. Write G in the form $G=(G_1, G_2)$, where G_1 is the genealogy of the sample until time $\tau\equiv T_{\mathscr{D}}(1)$, i.e., until the subsample $\mathscr{D}$ has found a most recent common ancestor. Write $G_2\equiv G-G_1$ for the genealogy of the entire sample further back in time than τ. Let γ be the branch length of the branch from the MRCA of $\mathscr{D}$ until this line is absorbed in the rest of the ancestral sample, that is, $\gamma=T_{\mathscr{D}}(0)-T_{\mathscr{D}}(1)=T_{\mathscr{D}}(0)-\tau$. Finally let $\alpha=A^*(1)$ be the number of ancestors to the whole sample at time τ, and J the number of coalescent events in the time

interval $(\tau, \tau+\gamma]$, that is, the number of coalescence events in the ancestry of $\mathscr{C}$ between the MRCA of $\mathscr{D}$ and the first time at which $\mathscr{D}$ and $\mathscr{C}$ share an ancestor (including the coalescence event which causes them to share an ancestor).

The following list of basic distributional results now follows from the structure of the coalescent and the conditional independences described above. We have written " $\sim$ " for "has the same distribution as," and "|" for "conditional on":

G_1 and G_2 are conditionally independent, given α. (6)

$$\tau \mid \alpha \sim W_n + \cdots W_{\alpha+1}. \tag{7}$$

$$\gamma \mid (\alpha, J) \sim W_\alpha + \cdots + W_{\alpha-J+1}. \tag{8}$$

The distribution of J can be derived using Corollary 2 with $n=\alpha$, $i=1$, and $j=0$:

$$P(J=x \mid \alpha) = \frac{2(\alpha-x)}{\alpha(\alpha-1)}, \quad x=1, ..., \alpha-1. \tag{9}$$

There are two special cases: For $i=1$ we see that $\tau=0$, and that γ is the time until a single line of descent is absorbed. And for $i=n-1$ we see that $\alpha=2$ with certainty, and γ is the waiting time until there are first two ancestors of the entire sample. In that case all coalescence events (up to τ) happen in the subsample $\mathscr{D}$, but with rate $(j+1)\,j/2$, if the subsample size is j.

Below we list a number of results on moments that will be used in the following. All expressions are derived by conditioning on α and (possibly) on J, and then using well-known results on moments of waiting times from the coalescent model. Proofs are straightforward, but cumbersome, and we omit them.

Expressions concerning τ:

$$\begin{aligned} E(\tau \mid \alpha) &= 2\left(\frac{1}{\alpha} - \frac{1}{n}\right), \\ E(\tau^2 \mid \alpha) &= 8 \sum_{j=\alpha}^{n} \frac{1}{j^2} + \frac{8}{n} - \frac{8}{\alpha} - \frac{8}{n\alpha}, \\ E(\tau) &= \frac{i-1}{n}. \end{aligned} \tag{10}$$

Expressions concerning γ:

$$E(\gamma) = \frac{i+1}{n},$$

$$E(\gamma^2) = 4\binom{n}{i+1}^{-1} \sum_{m=2}^{k+1} \frac{1}{m}\binom{n-m}{i-1},$$

$$E(\gamma^3) = 24\frac{(i+1)\,i}{n(n-1)} - 24\binom{n}{i+1}^{-1} \sum_{m=2}^{k+1} \frac{1}{m}\binom{n-m-1}{i-2} \sum_{j=1}^{m} \frac{1}{j}. \tag{11}$$

The expression for the third moment of γ is valid for $i>1$ only. The result for $i=1$ is obtained by replacing the double summation by $(1/n)\sum_{j=1}^{n}(1/j)$. Moments of γ conditional on α can be derived from (11) by letting $i=1$, $k=\alpha-1$, and $n=\alpha$.

4. THE EFFECT OF CONDITIONING ON THE MUTATION

In the previous section we described the genealogy of a subsample $\mathscr{D}$ in a larger sample with the constraint, E, that the chromosomes in $\mathscr{D}$ were forced to find a most recent common ancestor before coalescing with any chromosome ancestral to the rest of the sample. We will now impose the condition that a single mutation occurred on the branch γ, and that no other mutations have occurred (at the particular locus in question, in the history of the sample). In doing so, we will assume that the mutation rate is very small. Formally, we will consider the limiting case in which the mutation rate tends to zero.

In the coalescent approximation, mutations occur on the branches of the tree at the points of a Poisson process with rate $\theta/2$: conditional on a branch length w, the number of mutations on that branch has a Poisson distribution, $\mathrm{Po}(w\theta/2)$ with mean $w\theta/2$, independent of the locations of mutations elsewhere on the tree. In approximating a real population with N chromosomes at the locus in question, $\theta=2N\mu$, where μ is the chance of a mutation at the locus per chromosome per generation.

Denote by M the event that there is a single mutation along γ, and by η the time measured from τ until the mutation arises (see Fig. 1). Then we have

$$P(M \mid G) = P(M \mid \gamma) = \frac{\theta}{2}\gamma \exp\{-\gamma\theta/2\}, \tag{12}$$

and

G_1 and the triple (G_2, η, M) are conditionally independent given α, (13)

G_1 and G_2 are conditionally independent given (α, M). (14)

In particular, (13) and (14) are valid with G_1 and G_2 replaced by τ and γ, respectively. The properties (13) and

(14) are both consequences of (6) and the assumption that mutations follow a Poisson process on the tree.

We will derive results about the waiting time until the mutant arises in the sample as well as results related to the number of ancestors to the subsample $\mathscr{D}$ conditional on M. The method of proof consists of conditioning on the number of ancestors α when subsample $\mathscr{D}$ finds a most recent common ancestor, using (13) and (14), and applying the results from the previous two sections. Most parts of proofs consist of evaluations of elementary and finite sums and details of these will be omitted. Moreover we have dropped $\lim_{\theta \to 0}$ in front of expressions, where it formally should be, and merely indicated by $\to$ when the limit for $\theta \to 0$ is taken. Finally we have adopted the following notation for densities: all density functions are denoted by f and distinguished by the variable for which f is a density; that is, $f(\gamma)$ is the density of γ and $f(\tau)$ the density of τ.

We will first consider the jump chain $A^*_{\mathscr{D}}$, conditional on M. From Corollary 4 we know that conditional on $A^*(1) = \alpha$, $A^*_{\mathscr{D}}$ forms a Markov chain, and this turns out to be of importance.

LEMMA 4. *Conditional on M, the sequence $A^*_{\mathscr{D}}(l)$, $l = n, ..., 2$ is Markov with transition probabilities given by*

$$P(A^*_{\mathscr{D}}(l-1) = j-1 \mid A^*_{\mathscr{D}}(l) = j, M) = \frac{j}{l-1} \qquad (15)$$

for $l = 3, ..., n$ and $j = 2, ..., (l-1) \wedge i$, and with the convention that state $j = 1$ is absorbing.

Proof. Consider the case $l > 3$ and $2 > j$:

$$\begin{aligned} &P(A^*_{\mathscr{D}}(l-1) = j-1 \mid A^*_{\mathscr{D}}(l) \\ &\quad = j, A^*_{\mathscr{D}}(l+1) = j', ..., A^*_{\mathscr{D}}(n) = i, M) \\ &\quad = \frac{\begin{array}{c} P(A^*_{\mathscr{D}}(l-1) = j-1, A^*_{\mathscr{D}}(l) = j, \\ A^*_{\mathscr{D}}(l+1) = j', ..., A^*_{\mathscr{D}}(n) = i, M) \end{array}}{P(A^*_{\mathscr{D}}(l) = j, A^*_{\mathscr{D}}(l+1) = j', ..., A^*_{\mathscr{D}}(n) = i, M)} \\ &\quad \equiv \frac{P_1}{P_2}, \end{aligned}$$

say. Conditioning on α, using (13) and Lemma 3 we get

$$\begin{aligned} P_1 = &\sum_{\alpha=2}^{l-j+1} P(A^*_{\mathscr{D}}(l-1) = j-1 \mid A^*_{\mathscr{D}}(l) = j, \alpha) \cdots \\ &P(A^*_{\mathscr{D}}(l) = j \mid A^*_{\mathscr{D}}(l+1) = j', \alpha) \\ &\quad \times P(A^*_{\mathscr{D}}(n) = i \mid \alpha)\, P(M \mid \alpha)\, P(\alpha) \\ &= \sum_{\alpha=2}^{l-j+1} \frac{\begin{array}{c} (j-2)(j-1)\cdots(i-2)(l-\alpha-j+2) \\ \times (l-\alpha-j+3)\cdots(k-\alpha+1) \end{array}}{(l-\alpha-1)(l-\alpha)\cdots(n-\alpha-1)} \\ &\quad \times P(M \mid \alpha)\, P(\alpha). \end{aligned}$$

A similar equation holds for P_2, and note that P_1 and P_2 depend on j and l only, not on j', etc. If $l \geqslant 3$ and $j = 2$ the calculations are similar, but P_1 turns out to consist of just one term. If $l = 3$ and $j = 2$ then it is easily seen that $P_1/P_2 = 1$. Hence the conditional jump chain is Markovian. Either by simplifying P_1 and P_2 or by straightforward calculation of $P(A^*_{\mathscr{D}}(l-1) = j-1 \mid A^*_{\mathscr{D}}(l) = j, M)$ using (13), (11), Lemma 3, and Corollary 4 we get (here $l > 3$ and $j > 2$)

$$P_1 = C(n, i) \sum_{m=2}^{l-j+1} (m-1) \binom{l-m-2}{j-3} = C(n, i) \binom{l-2}{j-1},$$

and

$$P_2 = C(n, i) \sum_{m=2}^{l-j+1} (m-1) \binom{l-m-1}{j-2} = C(n, i) \binom{l-1}{j},$$

where $C(n, i)$ is a constant dependent on n and i only. Hence $P_1/P_2 = j/(l-1)$. Similarly for $l \geqslant 3$ and $j \geqslant 2$, and if $l = 3$ and $j = 2$ then $P_1/P_2 = 1 = 2/(3-1)$. This completes the proof. ∎

If we compare Lemma 1 and Lemma 4 we see that conditional on M, the probability that the next coalescence event is among ancestors to $\mathscr{D}$ is less than the probability of the equivalent event in the chain conditioned only on E. Again, this is to be expected, since the probability of M is increasing in α.

The transition probabilities in Lemma 4 are very similar to those obtained earlier in Lemma 1. Exactly as in Corollary 2, one can then obtain the marginal distribution of A^* conditional on M. We will not do so in general, but focus instead on two special cases of interest. The derivations are analogous to those given earlier and are omitted. The first concerns the distribution, conditional on M, of $\alpha \equiv A^*(1)$, the number of ancestors to the sample when $\mathscr{D}$ finds a most recent common ancestor (at time $\tau = T_{\mathscr{D}}(1)$). We have

$$P(\alpha \mid M) = (\alpha - 1) \binom{n-\alpha-1}{i-2} \binom{n-1}{i}^{-1} \qquad (16)$$

for $i = 2, 3, ..., n-1$, and $\alpha = 2, 3, ..., k+1$. Further, conditional on M, the mean of α is (Corollary 3)

$$E(\alpha \mid M) = \frac{2n}{i+1}.$$

Hence $(\alpha - 1) \mid M$ is distributed like $A^*(0)$ in a model with $i-1$ individuals in $\mathscr{D}$ and $n-1$ in total. Comparison with Corollary 3 shows that the mean of α condi-

tional on M is about two-thirds of its unconditional mean. This decrease is not surprising, since the probability of M is increasing in γ, and the distribution of γ is stochastically increasing as α decreases.

Next, consider α^*, defined to be the number of ancestors of the entire sample at the time the mutation occurs. Suppose $\alpha \equiv A^*(1) = l$. Then the probability that the next event is a mutation on the lineage ancestral to $\mathscr{D}$, rather than any of the possible coalescences, is

$$\frac{\theta}{2}\left(\frac{\theta}{2}+\frac{l(l-1)}{2}\right)^{-1}.$$

Also,

$$P(M \mid \alpha = l) \approx \frac{\theta}{2} E(\gamma \mid \alpha = l) = \frac{\theta}{l}$$

from (11) (with $i=1, n=l$). Thus, the probability, conditional on M, that the next event is a mutation on the lineage ancestral to $\mathscr{D}$, rather than a coalescence, is

$$\frac{\theta}{2}\left\{\left(\frac{\theta}{2}+\frac{l(l-1)}{2}\right) P(M \mid \alpha = l)\right\}^{-1} \to \frac{1}{l-1}. \quad (17)$$

Now, for the moment, redefine $A^*_{\mathscr{D}}$ so that $A^*_{\mathscr{D}}$ takes the value 0 when the mutation shared by $\mathscr{D}$ occurs. It follows from (17) that the transition probabilities (15) still apply, with $j=1, l>1$. Then, again by analogy with the derivation of Corollary 2,

$$P(\alpha^* \mid M) = \binom{n-\alpha^*}{i-1}\binom{n-1}{i}^{-1} \quad (18)$$

for $i = 1, 2, ..., n-1$, and $\alpha^* = 2, 3, ..., k$. Note that $\binom{n-\alpha^*}{i-1}$ is the number of ways $n-\alpha^*$ coalescent events can be partitioned into $i-1$ subsets, so that since the extension of Lemma 4 to the case $j=1$ just given ensures that all allowable sample paths of the jump chain back to the mutation have equal probability, the result (18) would also follow by a counting argument. The distribution (18) was originally due to Stephens (1999).

The distribution (18) has mean (Corollary 3)

$$E(\alpha^* \mid M) = 1 + \frac{n}{i+1}, \quad (19)$$

and variance

$$\mathrm{Var}(\alpha^* \mid M) = \frac{ni(n-i-1)}{(i+1)^2(i+2)}. \quad (20)$$

Let α_0 be the number of ancestors when the last ancestral line of subsample $\mathscr{D}$ is absorbed into the rest of the sample at time $\tau+\gamma = T_{\mathscr{D}}(0)$. The distribution of α_0 conditional on M is more complex,

$$P(\alpha_0 \mid M) \to \frac{\sum_{\alpha=2}^{k+1} E(\gamma \mid \alpha, \alpha_0)\, P(\alpha \mid \alpha_0)}{E(\gamma)} P(\alpha_0)$$

as in the argument above for the conditional distribution of α, with an additional conditioning on α. Note that

$$P(\alpha_0 \mid \alpha) = P(J = \alpha - \alpha_0 \mid \alpha),$$

and so by using (8) and (9) we have

$$P(\alpha_0 \mid M) = 2\binom{n-1}{i}^{-1}\left\{\binom{n-\alpha_0-1}{i-1} - \sum_{m=\alpha_0+1}^{k+1} \frac{\alpha_0}{m}\binom{n-m-1}{i-2}\right\} \quad (21)$$

for $\alpha_0 = 1, ..., k$. The mean of α_0, conditional on M, is

$$E(\alpha_0 \mid M) = \frac{2n}{3(i+1)} + \frac{1}{3}.$$

Consider now the distribution of J conditional on α and M. By applying Bayes' formula, and conditioning on γ, we have

$$P(J \mid M, \alpha) = \int_\gamma \frac{P(M \mid \alpha, J, \gamma)}{P(M \mid \alpha)}\, dP(\gamma \mid \alpha, J)\, P(J \mid \alpha) \to \frac{E(\gamma \mid \alpha, J)}{E(\gamma \mid \alpha)} P(J \mid \alpha).$$

Using (11) and (9) we get

$$P(J = x \mid M, \alpha) = \frac{2x}{\alpha(\alpha-1)} \quad (22)$$

for $x = 1, ..., \alpha-1$. That is, $J \mid M, \alpha \sim \alpha - J \mid \alpha$. Thus the effect of also conditioning on M is to increase stochastically the distribution of J relative to its distribution conditional only on α. Again, this is to be expected.

Consider now the time, denoted by η (see Fig. 1), from τ back until the mutation event giving rise to the mutation shared in $\mathscr{D}$. Conditional on M and γ we have as a consequence of the mutation process being Poisson that η is uniformly distributed on γ:

$$\eta \mid (M, G) \sim \eta \mid (M, \gamma) \sim U(0, \gamma). \quad (23)$$

The density function of η can be derived as

$$\begin{aligned} f(\eta \mid M) &= \int_{\eta}^{\infty} f(\eta \mid M, \gamma)\, dP(\gamma \mid M) \\ &= \frac{1}{P(M)} \int_{\eta}^{\infty} P(M \mid \gamma) \frac{dP(\gamma)}{\gamma} \\ &\to \frac{1}{E(\gamma)} P(\gamma > \eta) \\ &= \frac{n}{i+1} P(\gamma > \eta), \end{aligned} \tag{24}$$

where we have used (23) and (11). The density of η is strictly decreasing from the value $n/(i+1)$ at $\eta = 0$, and tends towards zero for $\eta \to \infty$. The probability $P(\gamma > \eta)$ can be found by conditioning on α and J and using results on tree heights from the coalescent (Tavaré 1984, p. 131ff). Here we will not be concerned with the exact form of the distribution but report the first two moments only,

$$\begin{aligned} E(\eta \mid M) &= \int_{\gamma} E(\eta \mid M, \gamma)\, dP(\gamma \mid M) \\ &= \frac{1}{2} E(\gamma \mid M) \to \frac{1}{2} \frac{E(\gamma^2)}{E(\gamma)} \\ &= 2 \binom{n-1}{i}^{-1} \sum_{m=2}^{k+1} \frac{1}{m} \binom{n-m}{i-1}, \end{aligned} \tag{25}$$

where we have used (23), (12), and (11). Similarly

$$\begin{aligned} E(\eta^2 \mid M) &\to \frac{1}{3} \frac{E(\gamma^3)}{E(\gamma)} \\ &= \frac{8i}{n-1} - 8 \binom{n-1}{i}^{-1} \sum_{m=2}^{k+1} \frac{1}{m} \binom{n-m-1}{i-2} \sum_{j=1}^{m} \frac{1}{j}. \end{aligned} \tag{26}$$

For $i = 1$ the second moment is obtained by replacing the double summation in (26) by $(1/n) \sum_{j=1}^{n} (1/j)$.

There are two special cases we will mention. The first one is when $i = n-1$. Here we have that $E(\eta \mid M) = 1$ and $E(\eta^2 \mid M) = 2$. This is as it should be, because the case reduces to considering the waiting time until a mutation occurs in a sample of size 2 (given that it occurs). The other case is $i = 1$, and here we have

$$E(\eta \mid M) \approx \frac{2}{n-1} \log(n-1) \quad \text{and} \quad E(\eta^2 \mid M) \approx \frac{8}{n-1}.$$

In this case, the variable η is the time until a mutation arises in a specific line conditional on its occurrence.

The distribution of τ conditional on M can also be expressed as a weighted sum of waiting time distributions from the coalescent model,

$$f(\tau \mid M) = \sum_{\alpha=2}^{k+1} f(\tau \mid \alpha)\, P(\alpha \mid M) \tag{27}$$

according to (14) and with $P(\alpha \mid M)$ given by (16). An expression for $f(\tau \mid \alpha)$ can be found from Tavaré (1984, p. 131ff), leading to an explicit expression for $f(\tau \mid M)$. Again we will be more concerned with moments of $\tau \mid M$ than of the exact distribution. Using (27), (16), and (10) we find that

$$\begin{aligned} E(\tau \mid M) &= \sum_{\alpha=2}^{k+1} E(\tau \mid \alpha)\, P(\alpha \mid M) \\ &= \frac{2i}{n-1} - \frac{2}{n} - 2 \binom{n-1}{i}^{-1} \sum_{m=2}^{k+1} \frac{1}{m} \binom{n-m-1}{i-2}, \end{aligned} \tag{28}$$

and

$$\begin{aligned} E(\tau^2 \mid M) &= \frac{8(k-1)}{n(n-1)} - \frac{8i}{n-1} \\ &\quad + 8 \binom{n-1}{i}^{-1} \left\{ \frac{n+1}{n} \sum_{m=2}^{k+1} \frac{1}{m} \binom{n-m-1}{i-2} \right. \\ &\quad \left. - \sum_{m=2}^{k} \frac{1}{m} \binom{n-m-1}{i-1} \right\} \\ &\quad + 8 \sum_{m=2}^{n} \frac{1}{m^2} - 8 \binom{n-1}{i}^{-1} \sum_{m=2}^{k} \frac{1}{m^2} \binom{n-m-1}{i}. \end{aligned} \tag{29}$$

Moreover, using (13) and conditioning on α gives

$$\begin{aligned} E(\tau\eta \mid M) &= \sum_{\alpha=2}^{k+1} E(\tau \mid \alpha)\, E(\alpha \mid M, \alpha)\, P(\alpha \mid M) \\ &= \frac{1}{2} \sum_{\alpha=2}^{k+1} E(\tau \mid \alpha) \frac{E(\gamma^2 \mid \alpha)}{E(\gamma \mid \alpha)} P(\alpha \mid M). \end{aligned}$$

Applying (11) with $i = 1$, $k = \alpha - 1$, and $n = \alpha$ and (10) we get

$$\begin{aligned} E(\tau\eta \mid M) &= 4 \binom{n-1}{i}^{-1} \sum_{m=2}^{k+1} \frac{1}{m} \binom{n-m-1}{i-2} \\ &\quad \times \sum_{j=2}^{m} \frac{1}{j} - \frac{2}{n} E(\eta \mid M). \end{aligned} \tag{30}$$

For $i=1$, $\tau=0$ and (28), (29), and (30) reduce to zero, and for $i=n-1$ we have $E(\tau \mid M)=1-2/n$ and $E(\tau\eta)=1-2/n$. The mean and variance of $T=\tau+\eta$ (Fig. 1) can be derived using (25), (26), (28), (29), and (30). For example,

$$E(T \mid M)=\frac{2i}{n-1}-\frac{2}{n}+2\binom{n-1}{i}^{-1}\sum_{m=2}^{k+1}\frac{1}{m}\binom{n-m-1}{i-1}.$$

This is also in Griffiths and Tavaré (1998) and Stephens (1999). Later we will give the limiting values of $E(T \mid M)$ and $\mathrm{Var}(T \mid M)$ for $n\to\infty$ and $i/n\to f$.

The joint density of (τ, γ) conditional on M can be obtained by conditioning on α and applying Bayes' formula (analogously to calculations above):

$$f(\tau, \gamma \mid M)\to\frac{\gamma}{E(\gamma)}\sum_{\alpha=2}^{k+1} f(\gamma \mid \alpha)\, f(\tau \mid \alpha)\, P(\alpha). \tag{31}$$

The density $f(\gamma \mid \alpha)$ can be obtained by conditioning on J and using Tavaré (1984, p. 131ff). From (31) we can derive the joint density of (τ, η) if we apply (23):

$$f(\tau, \eta \mid M)=\int_{\eta}^{\infty} f(\tau, \gamma \mid M)\frac{d\gamma}{\gamma}.$$

Inserting (31) then gives

$$f(\tau, \eta \mid M)=\frac{1}{E(\gamma)}\sum_{\alpha=2}^{k+1} P(\gamma>\eta \mid \alpha)\, f(\tau \mid \alpha)\, P(\alpha). \tag{32}$$

Finally we will derive the density function for $T=\tau+\eta$ from (32) by convolution. We have

$$f(T \mid M)=\sum_{\alpha=2}^{k+1}\int_{0}^{T} P(\gamma>T-\tau \mid \alpha)\, f(\tau \mid \alpha)\, d\tau\,\frac{P(\alpha)}{E(\gamma)}. \tag{33}$$

Since $P(\gamma>t \mid \alpha)$ can be written as a weighted sum of exponentials, an explicit expression for the density of T,

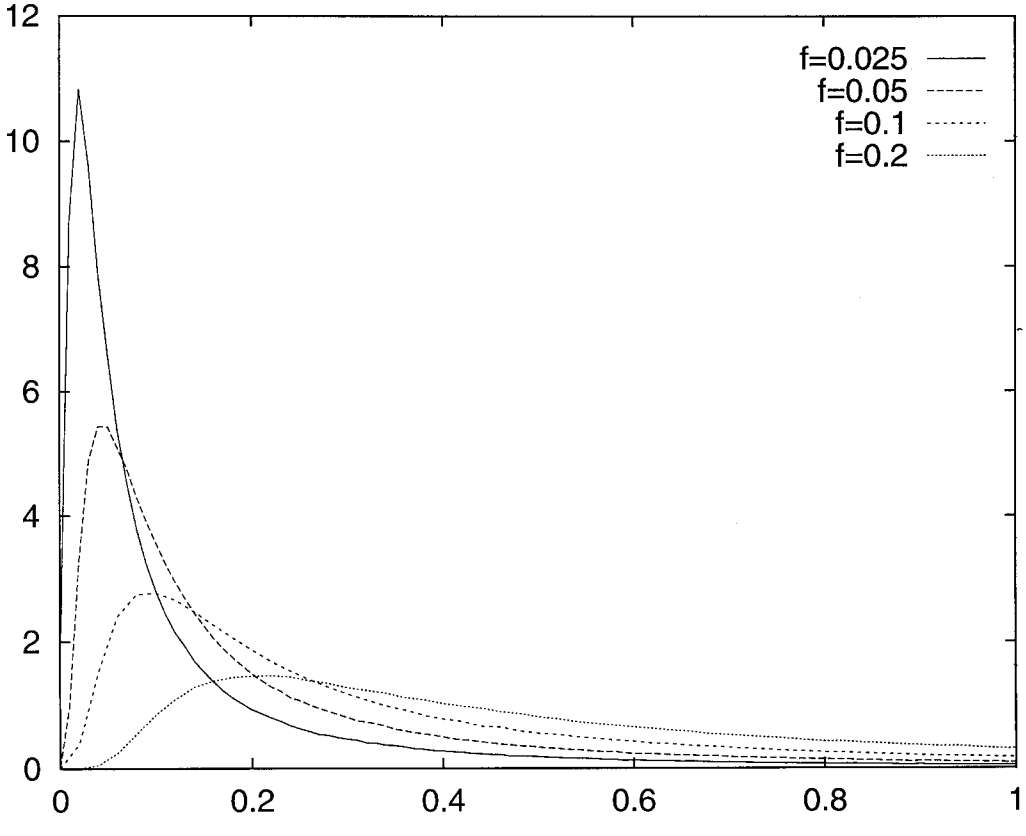

FIG. 3. Density of T in units of N_e (the effective number of chromosomes) in a population of constant size. The sample size is $n=1000$ and the fraction $f=i/n$ takes the values 0.025, 0.05, 0.1, and 0.2.

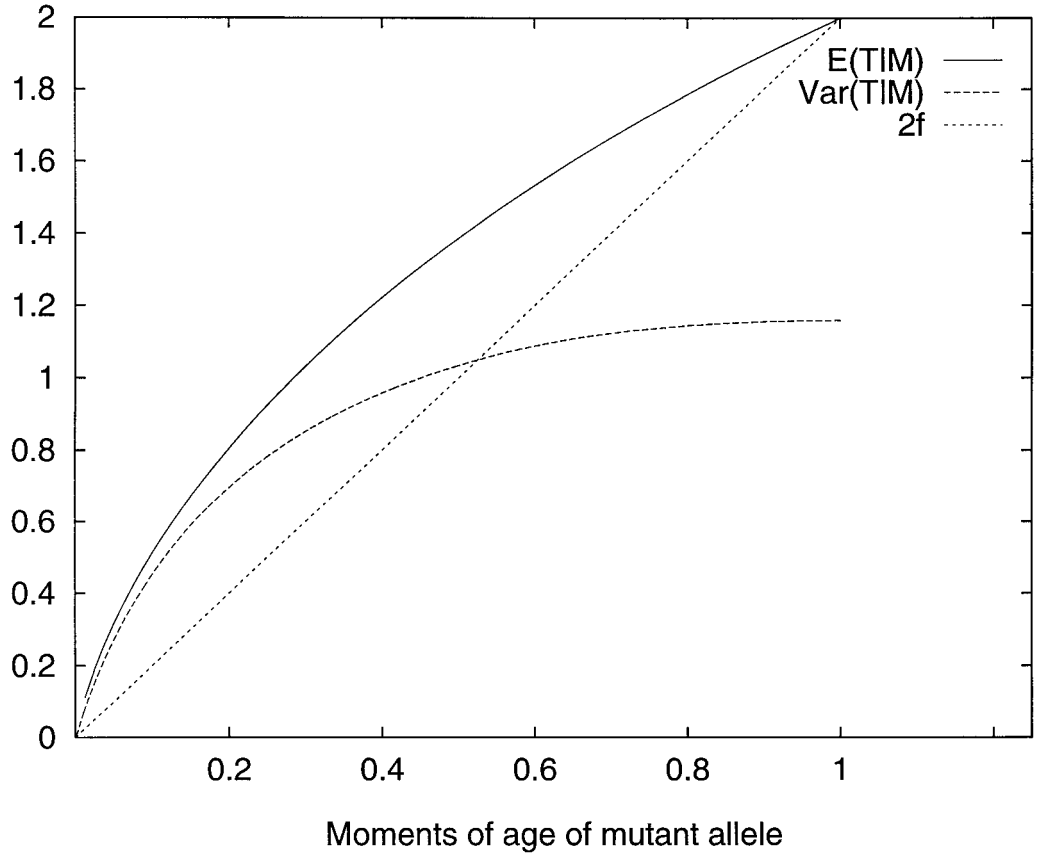

FIG. 4. Mean and variance of the age of the mutant allele. For $f > 0.5$ the variance is almost constant, tending to $4\pi^2/3 - 12 \approx 1.16$ as f approaches 1. Also shown is the mean $E(\tau + \gamma) = 2f$ where the conditioning is on the event E, but not on M.

the time since the mutation event, can be obtained from Tavaré (1984, p. 131ff). The density is zero for $T = 0$ and tends to zero for large T. Figure 3 plots the density for a large sample size, $n = 1000$, and various values of i. (Recall that time is measured in units of N_e generations, where N_e is the (variance) effective number of chromosomes in the population.) Figure 4 gives the mean and variance of the age of the mutation, in the limit of large sample size, as a function of the proportion of chromosomes in the sample carrying the mutation.

The results on the density and the moment of T are independently obtained by Griffiths and Tavaré (1998, Eqs. (5.6) and (5.8)).

5. LIMITING BEHAVIOUR FOR LARGE SAMPLES

In this section we will discuss the behaviour of various quantities derived above in the limit of large sample size: $n \to \infty$ with f denoting the limiting proportion of chromosomes in the sample carrying the mutation: $i/n \to f$. Numerical evaluations show the approximations derived here to be quite good even for small sample sizes.

Let us start by noting that

$$\binom{n-x}{i-j}\binom{n}{i}^{-1} = \frac{i_{[j]}(n-x)_{[i-j]}}{n_{[j]}(n-j)_{[i-j]}} \to f^j(1-f)^{x-j}, \tag{34}$$

where we have defined $a_{[x]} \equiv a(a-1)\cdots(a-x+1)$.

The technique of proof will in all cases be to pass to the limit using (34), and then evaluate the infinite series that result. The limit operation is interchangeable with summation in all cases, because all terms in summations are positive and moreover all summations are uniformly bounded in n and i.

Equations (34), (16), and (21) immediately give

$$P(\alpha \mid M) \to (\alpha - 1)\, f^2(1-f)^{\alpha-2}, \tag{35}$$

for $\alpha = 2, 3, \ldots$. That is, $(\alpha - 2) \mid M$ has a negative binomial distribution with parameters 2 and $1 - f$. Further,

$$P(\alpha_0 \mid M) \to 2f(1-f)^{\alpha_0-1} + \frac{2\alpha_0 f^2}{(1-f)^2}\left(\log(f) + \sum_{j=1}^{a_0} \frac{1}{j}(1-f)^j\right) = 2f(1-f)^{a_0-1} - \frac{2\alpha_0 f^2}{(1-f)^2}\int_0^{1-f} \frac{x^{\alpha_0}}{1-x}\,dx,$$

for $\alpha_0 = 1, 2, \ldots$. Some moments of $\alpha \mid M$ and $\alpha_0 \mid M$ are

$$E(\alpha \mid M) \to \frac{2}{f}, \qquad \mathrm{Var}(\alpha \mid M) \to \frac{2}{f}\left(\frac{1}{f} - 1\right),$$

$$\text{and} \qquad E(\alpha_0 \mid M) \to \frac{2}{3f} + \frac{1}{3}.$$

The limiting distribution of α^* conditional on M is

$$P(\alpha^* \mid M) \to f(1-f)^{\alpha^*-2},$$

that is, $\alpha^* - 1$ is geometrical with parameter $1-f$. The mean and variance are given by

$$E(\alpha^* \mid M) \to 1 + \frac{1}{f} \quad \text{and} \quad \mathrm{Var}(\alpha^* \mid M) \to \frac{1}{f}\left(\frac{1}{f} - 1\right).$$

Of more interest is the distribution of T. From (25), (26), (28), (29), and (30) we derive

$$E(\eta \mid M) \to -\frac{2f}{(1-f)^2}(\log(f) + 1 - f),$$

$$E(\eta^2 \mid M) \to \frac{8f}{1-f} + \frac{8f^2}{(1-f)^2}\int_0^{1-f} \frac{\log(1-x)}{x(1-x)}\,dx,$$

$$E(\tau \mid M) \to 2f + \frac{2f^2}{(1-f)^2}(\log(f) + 1 - f),$$

$$E(\tau^2 \mid M) \to \frac{4}{3}\pi^2 - \frac{8f}{1-f} - \frac{8f(2f-1)}{(1-f)^2}\log(f) + 8\int_0^{1-f} \frac{\log(1-x)}{x}\,dx,$$

$$E(\tau\eta \mid M) \to \frac{4f^2}{(1-f)^2}\log(f) - \frac{4f^2}{(1-f)^2}\int_0^{1-f} \frac{\log(1-x)}{x(1-x)}\,dx,$$

which in turn give us the mean and variance of T:

$$E(T \mid M) \to -\frac{2f}{1-f}\log(f), \tag{36}$$

and

$$\mathrm{Var}(T \mid M) \to \frac{4}{3}\pi^2 + \frac{8f}{1-f}\log(f) - \frac{4f^2}{(1-f)^2}\log^2(f) + 8\int_0^{1-f} \frac{\log(1-x)}{x}\,dx. \tag{37}$$

The moments (36) and (37) were first derived by Kimura and Ohta (1973) using the diffusion approximation and recently the first moment (36) was derived by Griffiths and Tavaré (1998). Figure 4 shows plots of $E(T \mid M)$ and $\mathrm{Var}(T \mid M)$, and the mean is compared with $E(\tau + \gamma)$, which is the time until the sample $\mathscr{D}$ first shares an ancestor with $\mathscr{C}$ (evaluated conditional on E but not on M). From (10) and (11) we have $E(\tau + \gamma) \to 2f$.

6. EXPONENTIALLY INCREASING POPULATION SIZE

As noted in a previous section, the results on the jump chain are independent of the population size. However, conditioning on the existence of the mutation will have effects that depend on the changes in the population size. In this section we will derive some results in a framework where the population size is exponentially increasing forwards in time. Mutation will still be assumed to be Poisson with constant rate $\theta/2$, with θ very small. We also assume that the demography of the population is such that the effect of variation in population size can be accounted for by the "usual" non-linear timescaling, effectively by the inverse of the population size. (See, e.g., Donnelly and Tavaré, 1995.)

In allowing for variation in population size in this framework, one can either rescale the coalescence rates, keeping the mutation rate constant over time, or keep the coalescence rates constant over time and rescale the mutation rate (e.g., Griffiths and Tavaré, 1994). Here it is more convenient to adopt the second approach, and keep the coalescence rate constant. This implies that the times between coalescence events are described by the usual coalescent model (Kingman, 1982a), and that the relation between time t in this framework and real time $\tilde{t}$, measured in generations per the effective (haploid) population size at time $t = 0$, is given by

$$t = \frac{1}{\beta}(e^{\beta\tilde{t}} - 1) \qquad \text{or} \quad \tilde{t} = \frac{1}{\beta}\log(\beta t + 1)$$

(e.g., Griffiths and Tavaré, 1994) with $\beta = N\xi$, where ξ is the rate of decrease in the population size (viewed backwards in time) each generation. Moreover in this formulation, the mutation rate is given as a function of time by

$$\theta(t) = \theta e^{-\beta t}.$$

We will take τ and γ to denote the same quantities as before. Formulas (6), (7), (8), and (9) are all valid in this setting, as well as (10) and (11). However, the last two are difficult to relate to real time $\tilde{\tau}$ and $\tilde{\gamma}$, and hence of less importance.

The probability of M conditional on τ and γ is given by

$$\begin{aligned} P(M \mid \tau, \gamma) &\approx 1 - \exp\left\{ -\int_{\tau}^{\tau+\gamma} \theta(t)/2 \, dt \right\} \\ &\approx \frac{\theta}{2\beta} e^{-\beta\tau}(1 - e^{-\beta\gamma}), \end{aligned} \tag{38}$$

where $\approx$ indicates that only first-order terms in θ are taken into account. The probability of M is then given by

$$\begin{aligned} P(M) &= \int_{\tau,\gamma} P(M \mid \tau, \gamma) \, dP(\tau, \gamma) \\ &= \sum_{\alpha=2}^{k+1} \int_{\tau,\gamma} P(M \mid \tau, \gamma) \, dP(\tau \mid \alpha) \, dP(\gamma \mid \alpha) \, P(\alpha). \end{aligned} \tag{39}$$

Substituting (38) then gives

$$P(M) \approx \frac{\theta}{2\beta} \sum_{\alpha=2}^{k+1} E(e^{-\beta\tau} \mid \alpha) \, E(1 - e^{-\beta\gamma} \mid \alpha) \, P(\alpha).$$

The distribution $P(\cdot)$ is given in Corollary 2 and the two expected values can be found by exploiting the fact that τ and γ are sums of exponential random variables:

$$E(e^{-\beta\tau} \mid \alpha) = \prod_{j=\alpha+1}^{n} \frac{j(j-1)}{j(j-1) + 2\beta}, \tag{40}$$

and

$$\begin{aligned} E(1 - e^{-\beta\gamma} \mid \alpha) &= 1 - \sum_{x=1}^{\alpha-1} E(e^{-\beta\gamma} \mid J = x, \alpha) \, P(J = x \mid \alpha) \\ &= 1 - \sum_{x=1}^{\alpha-1} \frac{2(\alpha - x)}{\alpha(\alpha - 1)} \prod_{j=\alpha-x+1}^{\alpha} \frac{j(j-1)}{j(j-1) + 2\beta}. \end{aligned} \tag{41}$$

Finally we have

$$P(\eta \leqslant t \mid M, \tau, \gamma) = \frac{P(\eta \leqslant t, M \mid \tau, \gamma)}{P(M \mid \tau, \gamma)} \approx \frac{1 - e^{-\beta t}}{1 - e^{-\beta\gamma}} \tag{42}$$

for $t \in (0, \gamma)$. That is, $\eta \mid M, \tau, \gamma$ follows a truncated exponential distribution.

With (38), (39), and (42) in hand and

$$f(\tau, \gamma \mid M) = \frac{P(M \mid \tau, \gamma)}{P(M)} f(\tau, \gamma),$$

we can derive expressions for distributions of waiting times τ, γ, η, and T conditional on M. For application, these must then in turn be transformed into distributions of similar waiting times expressed in the timescale for a constant population size coalescent. The relation between (τ, γ) and $(\tilde{\tau}, \tilde{\gamma})$ is given by

$$(\tilde{\tau}, \tilde{\gamma}) = \frac{1}{\beta}(\log(\beta\tau + 1), \log(\beta(\tau + \gamma) + 1) - \log(\beta\tau + 1)).$$

Table 1 gives approximate expected values of waiting times for different β's, evaluated numerically. These are in time units of N_e degenerations, where N_e is now the variance effective number of chromosomes in the population in the present.

The first line of Table 1 effectively corresponds to the constant population size case. The effect of a decrease in population size in the past is to hasten coalescent events. It is not surprising then that all times are decreasing in the parameter β.

Plots of the density of the age of the mutation (again in time units of N_e generations) can be obtained numerically. Some of these are exhibited in Figs. 5 and 6.

TABLE 1

Simulated Expected Values of Waiting Times, $f = 0.2$

	Conditional on M and E		Conditional on E	
β	η	T	γ	$\tau + \gamma$
10^{-2}	5.05×10^{-1}	8.02×10^{-1}	1.98×10^{-1}	3.97×10^{-1}
1	1.94×10^{-1}	4.13×10^{-1}	1.26×10^{-1}	3.00×10^{-1}
10^{3}	5.09×10^{-4}	5.60×10^{-3}	5.33×10^{-4}	5.63×10^{-3}
10^{6}	5.10×10^{-7}	1.25×10^{-5}	5.35×10^{-7}	1.25×10^{-5}

Note. Simulated expected values of waiting times for different β's are given in units of N_e; 1,000,000 simulations were performed for each entry using a importance sampling scheme.

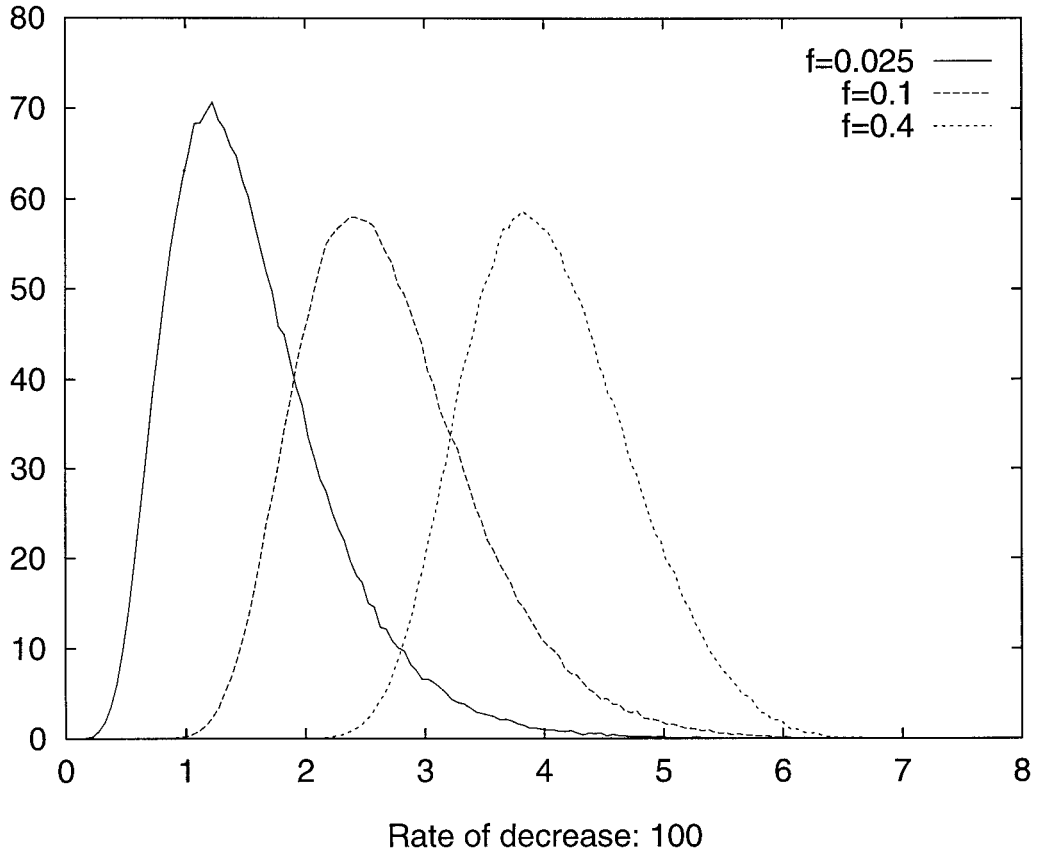

FIG. 5. Density of T in an exponentially decreasing population. The sample size is $n = 1000$, and the scaled rate of decrease β takes the value 100. The density of T is shown for three values of the fraction $f = i/n$: 0.025, 0.1, and 0.4. The x-axis is given in units of $10^2 N_e$. Here N_e is the effective number of chromosomes in the population at the time the sample was taken. Compare also with Fig. 3 showing the density of T in a population of constant size.

Recall that in the constant population size case, the effect of the additional conditioning on M is to increase γ, the time to the MRCA of $\mathscr{D}$. The intuition is that conditioning on M stochastically decreases α, hence increasing γ.

There is a second effect in the variable population size case, for large enough β, which works in the opposite direction. Conditioning on M will stochastically increase the real time between the MRCA of $\mathscr{D}$, and the first time this lineage shares an ancestor with $\mathscr{C}$. In a population whose size decreases monotonically into the past, coalescence rates increase as we go into the past. Thus, the longer into the past the MRCA of $\mathscr{D}$ is found, the faster are all coalescence rates at that time, and hence (other things being equal) the sooner the lineage from this MRCA will coalesce with an ancestor of the rest of the sample. It follows that (other things being equal) conditioning on M would have the effect of moving the MRCA of $\mathscr{D}$ closer to the present, and hence shortening γ. We now formalise this last heuristic argument.

Here we will only be concerned with the density $f(\tau \mid M)$ of τ conditional on M in proportion to $f(\tau)$. Let $Q(M) = 2\beta P(M)/\theta$. Then we have

$$\frac{f(\tau \mid M)}{f(\tau)} \to \frac{\int_\gamma P(M \mid \tau, \gamma)\, dP(\gamma \mid \tau)}{Q(M)} = \frac{e^{-\beta\tau} E(1 - e^{-\beta\gamma} \mid \tau)}{Q(M)}$$

after substitution from (38). The density $f(\tau)$ does not depend on β whereas $f(\tau \mid M)$ does. Note that for $0 < \varepsilon_1 < 1$ we can choose β large such that $E(1 - e^{\beta\gamma} \mid \alpha)$ is bounded uniformly from below in α by ε_1. This gives us

$$\frac{f(\tau \mid M)}{f(\tau)} \leqslant \frac{e^{-\beta\tau}}{E(e^{-\beta\tau})\, \varepsilon_1} \leqslant \frac{e^{-\beta\tau}}{\varepsilon_2^\beta \varepsilon_1}$$

for large β and $0 < \varepsilon_i < 1$, $i = 1, 2$. The last inequality follows the fact that $\{E(e^{-\beta\tau})\}^{1/\beta} \uparrow \| e^{-\tau} \|_\infty = 1$ for $\beta \to \infty$ (Hoffmann-Jørgensen, 1994). Hence we have that for all τ there exists a β such that

$$f(\tau' \mid M) \ll f(\tau')$$

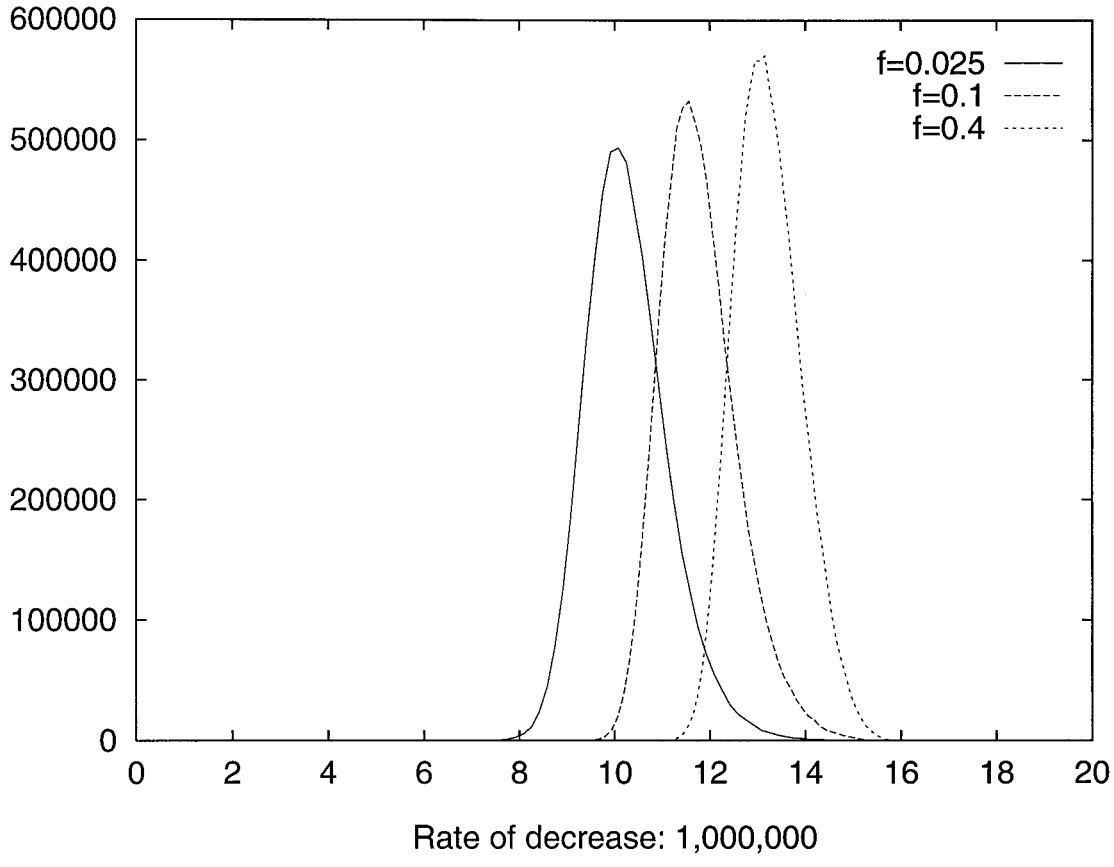

FIG. 6. Density of T in an exponentially decreasing population. The sample size is $n = 1000$, and the scaled rate of decrease β takes the value 1,000,000. The density of T is shown for three values of the fraction $f = i/n$: 0.025, 0.1, and 0.4. The x-axis is given in units of $10^6 N_e$, where N_e is the effective number of chromosomes in the population at the time the sample was taken.

for all $\tau' > \tau$ and $\beta' > \beta$. This indicates, although does not prove, that for large β the waiting time $\tilde{\tau}$ will tend to be smaller in mean than the unconditional waiting time. It proves that $E(\tau \mid M) < E(\tau)$ for large β but not the similar inequality for the real time $\tilde{\tau}$. This is in contrast to the constant population size case where $E(\tau) < E(\tau \mid M)$ (here there is no difference between τ and $\tilde{\tau}$).

The effect just described is visible in Table 1 and Table 2.

7. DISCUSSION

We have given a rigourous treatment of the structure of the genealogy of a sample when part of that sample, $\mathscr{D}$,

TABLE 2

Simulated Expected Values of τ

	$f = 0.025$		$f = 0.1$		$f = 0.4$	
β	M, E	E	M, E	E	M, E	E
10^{-2}	4.45×10^{-2}	2.30×10^{-2}	1.63×10^{-1}	9.79×10^{-2}	5.13×10^{-1}	3.97×10^{-1}
1	3.82×10^{-2}	2.25×10^{-2}	1.27×10^{-1}	9.09×10^{-2}	3.61×10^{-1}	3.20×10^{-1}
10^2	1.12×10^{-2}	1.07×10^{-2}	2.22×10^{-2}	2.20×10^{-2}	3.54×10^{-2}	3.55×10^{-2}
10^4	5.14×10^{-4}	5.17×10^{-4}	6.66×10^{-4}	6.67×10^{-4}	8.11×10^{-4}	8.12×10^{-4}
10^6	9.74×10^{-6}	9.77×10^{-6}	1.13×10^{-5}	1.13×10^{-5}	1.27×10^{-5}	1.27×10^{-5}

Note. Simulated expected values of τ conditional on M, E and E, respectively, are given in units of N_e; 1,000,000 simulations were performed for each entry using an importance sampling scheme.

shares a mutation, assumed to have arisen uniquely in the population's history. In particular we have derived expressions for the distribution and the moments of the age of the mutation, conditional on the size of the subsample which carries it.

Our motivation stems from the fact that an understanding of the genealogy will lead to properties of various related types of data, including inter-allelic variability (patterns of variability within $\mathscr{D}$ at linked loci), and patterns of linkage disequilibrium, and lengths of shared segments, around the site of the mutation. This will, in turn, provide a foundation for statistical methods which aim to infer aspects of the genealogy, or of population history, from such data. These issues will be pursued elsewhere.

One statistical question which we can address is the estimation of the age of the mutation, conditional (only) on its frequency within the sample. The age of the mutation, T in our notation, is a random quantity within the coalescent model, not a parameter in the conventional sense. Inference (for *either* a Bayesian or a Frequentist statistician) should thus consist of reporting the conditional distribution of this age given the data, or perhaps summaries of this distribution, such as its mean and variance. Expressions for these, and their limits for large samples, are given in earlier sections.

Slatkin and Rannala (1997) developed a method for estimating the age of the mutation. Their setting is considerably more general than ours in two respects. First, they give a procedure for estimation based on information on variability at linked loci, in addition to the frequency of the mutation. Second, they allow for the mutation not to be neutral. On the other hand, their method involves several heuristic approximations. It may thus be of interest to compare their estimator with our exact results in the special setting of this paper.

In the setting here, for large sample size and constant population size, Slatkin and Rannala's estimator of the age of the mutation, $\hat{T}_{\mathrm{SR}}$ say, is $2f$, where f is the frequency of chromosomes in the sample carrying the mutation. Their approach is to treat the age as a parameter and derive its maximum likelihood estimator under an approximate model. Our approach derives the conditional distribution of the age. For the purposes of com-

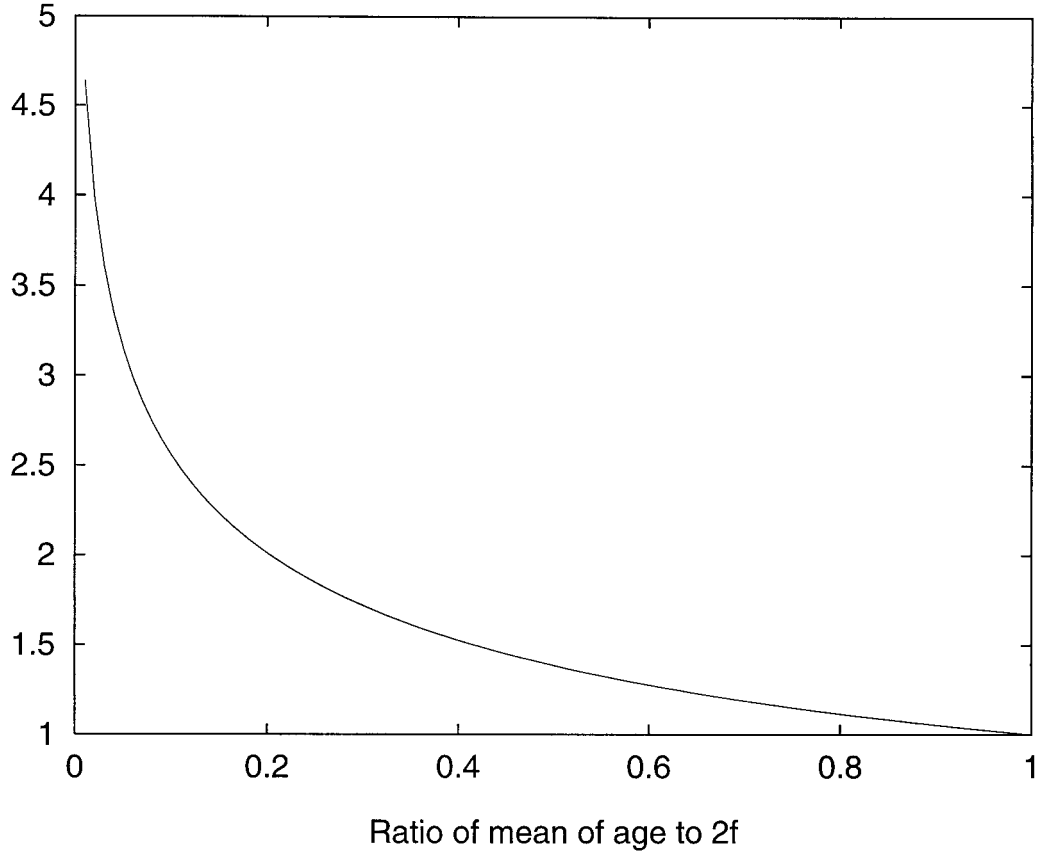

FIG. 7. The ratio of $E(T \mid M)$ to $\hat{T}_{\mathrm{SR}}$. The curve shows the ratio of $E(T \mid M)$ to $\hat{T}_{\mathrm{SR}}$, which is the estimator of the age of the mutant allele given in Slatkin and Rannala (1997). The ratio increases towards infinity for f tending to 0, and tends to 1 as f approaches 1.

parison we have compared the mean of this conditional distribution, $E(T \mid M) = -2f \log(f,)/(1-f)$, with $\hat{T}_{SR}$. The Slatkin and Rannala estimator is always smaller than the conditional mean, and Fig. 7 plots the ratio of the latter to the former as a function of f. It can be seen that, especially for small f, the relative underestimation of $\hat{T}_{SR}$ can be quite marked. Some caution may thus be appropriate in applying their estimator in more complicated settings.

There would seem to be several factors which may affect the comparison of the estimators. The first is the fact that Slatkin and Rannala rely on various approximations. The second stems from their treatment of the age of the mutation as a parameter. As we mentioned earlier, this seems to us inappropriate as a matter of statistical principle: the age is a random variable, not a parameter in the usual sense. Aside from issues of principle, treating the age as a parameter has a serious practical drawback. We should condition on both the fact that the mutation is seen only in $\mathscr{D}$ *and* the fact that the mutation arose at all. (Recall our conditioning on first E, and then M, above.) The mutation is more likely to have arisen in genealogical trees with a long branch between the MRCA of $\mathscr{D}$ and the ancestry of the remainder of the sample. As a consequence, as we have seen, conditioning on M has the effect of stochastically increasing the length of this branch. This effect works directly, but it also has an indirect effect in pushing the MRCA of $\mathscr{D}$ further into the past. Both effects increase the age of the mutation. (We know the age must be less than $\tau + \gamma$, but note from Fig. 4 that $E(T \mid M)$ is substantially larger than the mean of $\tau + \gamma$ without conditioning on M.) Informally then, conditioning on the fact that the mutation arose at all increases its age. This effect is real, and should be allowed for. If one treats the age as a parameter, it seems difficult to allow for this.

Table 3 relates to the setting of a population whose size has increased exponentially through time. It compares the Slatkin and Rannala estimator in this context (here $\hat{T}_{SR} = \log(2\beta f + 1)/\beta$) with $E(T \mid M)$. The relative discrepancy between the estimators decreases with the growth rate β, though we note that for large β, $\hat{T}_{SR}$ is slightly larger than the conditional mean age.

ACKNOWLEDGMENTS

The first author was supported by the Faculty of Natural Sciences, University of Aarhus. Part of the work was completed while he was visiting the Department of Statistics at the University of Oxford. The second author was supported in part by NSF Grant DMS 9505129 and UK EPSRC Grants B/AF/1255 and GR/M14197. Both authors also received partial support from UK BBSRC Grant E09788.

REFERENCES

Donnelly, P., and Tavaré, S. 1995. Coalescents and genealogical structure under neutrality, *Annu. Rev. Genet.* **29**, 401–422.

Griffiths, R. C., and Tavaré, S. 1994. Sampling theory for neutral alleles in a varying environment, *Philos. Trans. R. Soc. London B* **344**, 403–410.

Griffiths, R. C., and Tavaré, S. 1998. The age of a mutant in a general coalescent tree, *Stochastic Models* **14**, 273–295.

Hoffmann-Jørgensen, J. 1994. "Probability with a View Towards Statistics," Chapmann & Hall, New York.

Hudson, R. R. 1990. Gene genealogies and the coalescent process, *Oxford Surv. Evol. Biol.* **7**, 1–44.

TABLE 3

Comparison of $E(T \mid M)$ with $\hat{T}_{SR}$

f	β	0	10^{-2}	1	10^{2}	10^{4}	10^{6}
0.025	$E(T \mid M)$	1.89×10^{-1}	1.82×10^{-1}	1.10×10^{-1}	1.55×10^{-2}	5.67×10^{-4}	1.03×10^{-5}
	$\hat{T}_{SR}$	5.00×10^{-2}	5.00×10^{-2}	4.88×10^{-2}	1.79×10^{-2}	6.22×10^{-4}	1.08×10^{-5}
	Ratio	3.78	3.64	2.25	0.87	0.91	0.95
0.1	$E(T \mid M)$	5.12×10^{-1}	5.03×10^{-1}	2.74×10^{-1}	2.70×10^{-2}	7.18×10^{-4}	1.18×10^{-5}
	$\hat{T}_{SR}$	2.00×10^{-1}	2.00×10^{-1}	1.82×10^{-1}	3.04×10^{-2}	7.60×10^{-4}	1.22×10^{-5}
	Ratio	2.56	2.52	1.51	0.89	0.94	0.97
0.4	$E(T \mid M)$	1.22	1.20	6.06×10^{-1}	4.04×10^{-2}	8.62×10^{-4}	1.32×10^{-5}
	$\hat{T}_{SR}$	8.00×10^{-1}	7.97×10^{-1}	5.88×10^{-1}	4.39×10^{-2}	8.99×10^{-4}	1.36×10^{-5}
	Ratio	1.53	1.51	1.03	0.92	0.96	0.97

Note. The expected value of T is compared to $\hat{T}_{SR} = \log(2\beta f + 1)/\beta$, the estimator given in Slatkin and Rannala (1997). Both are in units of N_e. The expected value of T is very close to the mode of the density (Figs. 5 and 6). Also the ratio between $E(T \mid M)$ and $\hat{T}_{SR}$ is given. Surprisingly, this is not a monotone function in β.

Kimura, M., and Ohta, T. 1973. The age of a neutral mutant persisting in a finite population, *Genetics* **75**, 199–212.

Kingman, J. F. C. 1982a. The coalescent, *Stochastic Proc. Appl.* **13**, 235–248.

Kingman, J. F. C. 1982b. Exchangeability and the evolution of large populations, *in* "Exchangeability in Probability and Statistics" (G. Koch and F. Spizzichino, Eds.), pp. 97–112, North-Holland, Amsterdam.

Moran, P. A. P. 1958. Random processes in genetics, *Proc. Cambridge Philos. Soc.* **54**, 60–71.

Saunders, I. W., Tavaré, S., and Watterson, G. A. 1984. On the genealogy of nested subsamples from a haploid population, *Adv. Appl. Probab.* **16**, 471–491.

Slatkin, M., 1996. Gene genealogies within mutant allelic classes, *Genetics* **143**, 579–587.

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

Stephens, M. 1999. Times on trees, and the age of an allele, submitted for publication.

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

Thompson, E. A., and Neel, J. V. 1997. Allelic disequilibrium and allele frequency distribution as a function of social and demographic history, *Am. J. Hum. Genet.* **60**, 197–204.