Theoretical Population Biology **58**, 61–75 (2000)
doi:10.1006/tpbi.2000.1469, available online at http://www.idealibrary.com on IDEAL®

On the Genealogy of a Sample of Neutral Rare Alleles

Carsten Wiuf

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

Received August 24, 1999

This paper concerns the genealogical structure of a sample of chromosomes sharing a neutral rare allele. We suppose that the mutation giving rise to the allele has only happened once in the history of the entire population, and that the allele is of *known* frequency q in the population. Within a coalescent framework C. Wiuf and P. Donnelly (1999, *Theor. Popul. Biol.* 56, 183–201) derived an exact analysis of the conditional genealogy but it is inconvenient for applications. Here, we develop an approximation to the exact distribution of the conditional genealogy, including an approximation to the distribution of the time at which the mutation arose. The approximations are accurate for frequencies $q < 5$–10%. In addition, a simple and fast simulation scheme is constructed. We consider a demography parameterized by a d-dimensional vector $\alpha = (\alpha^1, ..., \alpha^d)$. It is shown that the conditional genealogy and the age of the mutation have distributions that depend on $a = q\alpha$ and q only, and that the effect of q is a linear scaling of times in the genealogy; if q is doubled, the lengths of all branches in the genealogy are doubled. The theory is exemplified in two different demographies of some interest in the study of human evolution: (1) a population of constant size and (2) a population of exponentially decreasing size (going backward in time).

Key Words: **age of mutation; coalescent theory; genealogy; rare allele; sampling scheme.**

INTRODUCTION

In this paper we study a sample from a subpopulation consisting of those chromosomes sharing a neutral rare allele. Assume that the frequency, q, of the allele in the entire population is small. We give, within a coalescent framework, a simple approximation to the exact conditional genealogy of a sample (or of the subpopulation) *given* q. It is assumed that the mutation giving rise to the allele is unique in the history of the entire population.

The problem of describing the genealogy of a sample of rare alleles has had considerable recent attention. Slatkin and Rannala (1997) developed an approximate method (combining a linear birth and death process with the coalescent) to study the genealogy of a rare allele. In their setup the time T, at which the mutation arose, is treated as a parameter. In contrast to this, T is in this paper (and in Wiuf and Donnelly, 1999) treated as a stochastic variable conditional on the frequency q. Wiuf and Donnelly (1999) showed, based on an exact coalescent analysis, that in the simple scenario of constant population size the difference between the approximation of Slatkin and Rannala (1997) and the exact results of Wiuf and Donnelly (1999) becomes substantial for small q. Similarly, Thompson and Neel (1997) use fractional linear branching processes to model the demography of a rare allele. Again the age T of the mutation is treated as a parameter.

Treating the age as a parameter has a serious drawback. We should condition both on the fact that the mutation is seen only in a fraction of the population, and the fact that the mutation arose at all. The mutation is

0040-5809/00 $35.00

more likely to have arisen in genealogical trees with a long branch between the most recent common ancestor (MRCA) of the subpopulation of rare alleles, and the ancestry of the rest of the population. As a consequence, conditioning on the mutation has the effect of stochastically increasing the length of this branch. This effect turns out to be important, especially for small q, and is not captured in Slatkin and Rannala (1997) and Thompson and Neel (1997).

Throughout this paper, we adopt the coalescent as an exact description of the genealogy of the entire population, or equivalently of samples from it. The effective number of chromosomes in the population is allowed to vary with time such that the effective number at time t in the past is $N(t)$. Time starts at the present, $t = 0$, and is measured in units of $N = N(0)$ generations. We consider demographies parameterized by a d-dimensional vector, $\alpha = (\alpha^1, ..., \alpha^d)$. Examples of this kind include a constant population size scenario, $\alpha = (\ )$, and a scenario with exponentially decreasing population size (going backward in time), $\alpha = (\beta)$, where β is the rate of decrease in population size per N generations.

In this setting, we consider a sample of n chromosomes taken from the population at the present time. At a particular locus, the sample is divided into two subsamples, $\mathscr{D}$ and $\mathscr{C}$ of size k and $n - k$, respectively, with the property that all of the chromosomes in $\mathscr{D}$ 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. If the mutation rate at the locus is very small this assumption is likely to be true; the chance of more than one mutation event at the locus becomes negligible. Formally, we examine the limit as the mutation rate tends to zero, conditional on the mutation having occurred. We are interested in the genealogy of $\mathscr{D}$ and, in particular, in the case where the sample sizes n and k both are large but such that the frequency $q \approx k/n$ is small; i.e., the mutants are rare. The notation to be introduced is illustrated in Fig. 1.

Let the event E be that all of the chromosomes in $\mathscr{D}$ share a common ancestor before any chromosome in $\mathscr{D}$ shares a common ancestor with a chromosome in $\mathscr{C}$ (Fig. 1). Assuming E, let M denote the event that a single mutation has occurred on the ancestral lineage common to all of $\mathscr{D}$ between the time of the MRCA of $\mathscr{D}$ and the time at which $\mathscr{D}$ first shares an ancestor with $\mathscr{C}$. The event that exactly those chromosomes in $\mathscr{D}$ share the mutation is given by $E \cap M$, and we proceed by first examining the genealogy of $\mathscr{D}$ conditional on E, and then additionally conditional on M. The event E affects the jump chain of the coalescent for the sample only, and not the times between coalescent events (Kingman, 1982). Throughout the paper, conditioning on E and the sample configuration (the number of chromosomes in $\mathscr{D}$ and $\mathscr{C}$) will be suppressed in the notation.

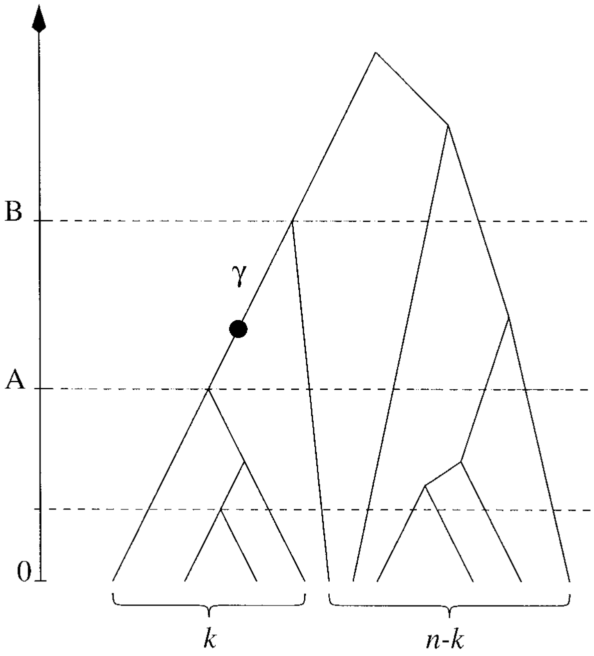

FIG. 1. Notation. In this example subsample $\mathscr{D}$ consists of $k = 4$ individuals and subsample $\mathscr{C}$ of $n - k = 6$ individuals. Time is measured backward with zero denoting present time. Subsample $\mathscr{D}$ finds a MRCA at time A, and there are five ancestors of the whole sample at that time, the ancestor of $\mathscr{D}$ and four ancestors of $\mathscr{C}$. At time B, 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 A. The variables $A(j), j = 0, 1, ..., 4$, count the number of ancestors of the entire sample the first time there are j ancestors of $\mathscr{D}$. For example, $A(1) = 5$ and $A(3) = 9$ (marked with a dotted line). The mutation shared in $\mathscr{D}$ is marked with a dot and the branch at which it arose is called γ.

This paper is organized into several sections. Section 1 concerns the jump chain of the genealogy of $\mathscr{D}$. In Sections 2 and 3 an approximation to the distribution of the genealogy of $\mathscr{D}$, conditional on E only, is derived. We show that for small values of q the number of ancestors, $A(j)$, of the entire sample when there are j ancestors of $\mathscr{D}$ can be assumed to be in a one-to-one correspondence with time, and that the distribution of $A(j), j \geqslant 0$, has a particular, simple form. Sections 4–6 extend these results to the genealogy conditional on both E and M and give examples. It is shown that q affects the distribution of the conditional genealogy essentially through a linear scaling of time. A scheme for simulating the genealogy of a finite sample taken from $\mathscr{D}$ is developed in Section 7 and, finally, in Section 8 we evaluate the theory presented in the paper and display a few consequences. All proofs are given in the Appendix.

1. THE ANCESTRAL CHAINS OF $\mathscr{D}$

The topological structure of the genealogy of $\mathscr{D}$ is described by two jump chains: $A(j)$, $j=0, 1, ..., k$, and $D(m)$, $m=1, 2, ..., n$. The variable $A(j)$ is the number of ancestors of the entire sample the first time there are j ancestors of $\mathscr{D}$ (Fig. 1). We say that there are zero ancestors of $\mathscr{D}$ when any ancestor of the sample is an ancestor of at least one chromosome in $\mathscr{C}$. In particular, there are zero ancestors of $\mathscr{D}$ when the last ancestral lineage common to all of $\mathscr{D}$ is absorbed into the rest of the genealogy, and $A(0)$ denotes the number of ancestors of the entire sample when this happens.

The variable $D(m)$, $m=1, 2, ..., n$, is the number of ancestors of $\mathscr{D}$ when there are m ancestors of the entire sample. By the convention described above $D(m)$ can be zero. In fact we always have $D(1)=0$ because if there is only one ancestor of the entire sample this ancestor is in particular an ancestor of all chromosomes of $\mathscr{C}$.

The two chains are called *the ancestral chains of* $\mathscr{D}$. We have the following relationship between $A(j)$ and $D(m)$ (see Wiuf and Donnelly, 1999):

$$A(j)=m \Leftrightarrow D(m)=j \quad \text{and} \quad D(m+1)=j+1, \tag{1}$$

and

$$D(m)=j \Leftrightarrow A(j)\geqslant m \quad \text{and} \quad A(j-1)\leqslant m-1. \tag{2}$$

Thus, the distribution of either chain can be found from the distribution of the other chain. We will in particular focus on properties of the chain $A(j)$, $j=0, 1, ..., k$. It tells us the number of ancestors of the entire sample at each coalescent event in subsample $\mathscr{D}$, and thus also the time from present until there are j ancestors of $\mathscr{D}$. For instance, under an assumption of constant effective population size this time is distributed like a sum $V_n+V_{n-1}+\cdots+V_{A(j)+1}$ of exponential variables, where $V_h \sim \mathrm{Exp}(h(h-1)/2)$ and V_h is the time while there are h ancestors of the entire sample. This follows from the fact that the waiting times, V_h, are independent of the jump chain $A(j), j\geqslant 0$ (Kingman, 1982).

Let P_n denote the probability measure associated with a sample of size n. Applying results in Wiuf and Donnelly (1999) we find the distribution of the chain $A(j), j\geqslant 0$.

LEMMA 1. *The number of ancestors of the whole sample the first time there are j ancestors of* $\mathscr{D}$, $A(j)$, $j=0, 1, ..., k$, *forms a Markov chain with marginal distributions given by*

$$P_n(A(j)=m)=\frac{\binom{n-m-1}{k-j-1}\binom{m}{j+1}}{\binom{n}{k+1}}, \tag{3}$$

for $j=0, 1, ..., k-1$, *and* $m=j+1, ..., n-1$. *If* $j=k$ *we have* $P_n(A(k)=m)=1$, *if* $m=n$ *and zero otherwise. The transition probabilities are given by*

$$P_n(A(j)=m \mid A(j-1)=l)=\frac{\binom{n-m-1}{k-j-1}}{\binom{n-l-1}{k-j}}, \tag{4}$$

with $j=1, 2, ..., k-1$, $l=j, j+1, ..., n-2$, *and* $m=l+1, l+2, ..., n-1$. *If* $j=k$ *we have* $P_n(A(k)=n \mid A(k-1)=l)=1$ *for all* $l=k, k+1, ..., n-1$, *and zero otherwise.*

From (3) and (4) and the Markov chain property of $A(j)$, $j=0, 1, ..., k$, one can easily find the joint distribution of any vector $A(0), A(1), ..., A(j)$, with $0\leqslant j\leqslant k$. The distribution of $D(m)$, $j=1, 2, ..., m$, can be derived from (1) and Lemma 1.

2. CONVERGENCE OF THE ANCESTRAL CHAINS

In this section we discuss convergence properties of the ancestral chains, $A(j)$, $j=0, 1, ..., k$, and $D(m)$, $m=1, 2, ..., n$, as the size of the entire sample becomes large and the frequency q becomes small. The conditioning on the event M (in different scenarios) will be postponed to subsequent sections; i.e., the results given in this section are derived conditional on E, the topological structure, and the sample configuration only.

We show that as q decreases and n increases $qA(j)$ tends to a continuous variable. Thus, the number of ancestors of the entire population is naturally measured in units of $1/q$. In the next two sections we show that the distribution of the genealogy of $\mathscr{D}$ can be derived from that of $qA(j)$.

Formally, we consider a series of samples of size n, $n=2, 3, ...,$ and subsamples $\mathscr{D}_n$ of size k_n such that $q_n=k_n/n\to 0$ and $k_n\to\infty$. We let $\mathscr{D}_\infty$ refer to the model that emerges as $n\to\infty$. The benefit of these assumptions is that the fraction, q_n, of the rare allele becomes small while both the entire sample (of size n) and the subsample of the rare allele become very large (as $k_n\to\infty$.)

In the limit this corresponds to sampling the whole population. Subscript n is suppressed in the notation of $A(j)$ and $D(m)$.

Let $X \sim Y$ denote that the variable X is distributed like the variable Y, and let $X \sim g(x)$ denote that X has density $g(x)$ (wrt Lebesgue measure). The conditional variable X given Y is denoted $X \mid Y$.

Theorem 1. *Assume that* $q_n \to 0$ *and* $k_n \to \infty$. *The chain* $q_n A(j)$, $j = 0, 1, \ldots$, *converges in distribution to a Markov process* $A_\infty(j), j = 0, 1, \ldots$, *with marginal distributions given by*

$$A_\infty(j) \sim \frac{1}{(j+1)!} x^{j+1} \exp(-x) \tag{5}$$

for $x > 0$; *i.e.*, $A_\infty(j)$ *is gamma distributed* $\Gamma(j+2, 1)$. *The increments* $A_\infty(j) - A_\infty(j-1)$, $j \geq 1$, *form a series of independent variables with*

$$A_\infty(j) - A_\infty(j-1) \sim \text{Exp}(1). \tag{6}$$

Here Exp(λ) *denotes an exponential variable with intensity* λ. *In particular*, $A_\infty(j) - A_\infty(j-1)$ *is independent of* $A(j'), j' = 0, 1, \ldots, j-1$.

It is of interest to note that the distribution of $A_\infty(j)$ does not depend on how q_n approaches 0, as long as $k_n \to \infty$. Equations (5) and (6) imply that $A_\infty(0)$, $A_\infty(1), \ldots$, can be simulated by a series of exponential variables $X_j \sim \text{Exp}(1), j \geq 0$,

$$A_\infty(j) \sim X_0 + X_1 + \cdots + X_{j+1}. \tag{7}$$

Here, the ancestral chain $A_\infty(j)$, $j \geq 0$, is constructed from $A_\infty(0)$ and then "moving" toward the present with increasing j. The present time corresponds to $\sum_j X_j = \infty$ as the size of $\mathscr{D}_\infty$ at the present time is infinity. Further, the process is almost a Poisson process with intensity 1 apart from the distribution of the first point "0" which arrives according to a $\Gamma(2, 1)$ distribution. This is effectively the increase in ancestral lineages from the MRCA of the entire population until the lineage ancestral to all $\mathscr{D}_\infty$ is introduced.

Denote by $D(x/q_n)$ the variable $D(\lfloor x/q_n \rfloor)$, where $\lfloor u \rfloor$ is the integer part of u.

Theorem 2. *With the assumptions of Theorem* 1, *the Markov process* $D(x/q_n)$, $x > 0$, *converges to a Markov process* $D_\infty(x)$, $x > 0$, *with marginal distributions*

$$P(D_\infty(x) = j) = \frac{1}{(j+1)!} x^{j+1} \exp(-x), \tag{8}$$

if $j \geq 1$ *and*

$$P(D_\infty(x) = 0) = (1 + x) \exp(-x) \tag{9}$$

if $j = 0$. *Let* $x > y > 0$. *The transition probabilities of the process* $D_\infty(x)$, $x > 0$, *are given by*

$$P(D_\infty(x) = j \mid D_\infty(y) = i) = \frac{1}{(j-i)!} (x-y)^{j-i} \exp\{-(x-y)\} \tag{10}$$

if $j \geq i \geq 1$,

$$P(D_\infty(x) = j \mid D_\infty(y) = 0) = \frac{x + jy}{(j+1)!\,(y+1)} (x-y)^j \exp\{-(x-y)\} \tag{11}$$

if $j > 0$ *and* $i = 0$, *and finally*

$$P(D_\infty(x) = 0 \mid D_\infty(y) = 0) = \frac{x+1}{y+1} \exp\{-(x-y)\} \tag{12}$$

if $j = i = 0$.

The variable $D_\infty(x)$, $x > 0$, is Poisson distributed with intensity x, except that the outcomes 0 and 1 are pooled. Note that whenever $i > 0$, the increment $D_\infty(x) - D_\infty(y)$ is independent of $D_\infty(y) = i$, and its distribution is Poisson with intensity $\delta = x - y$. This is in agreement with (6). If $i = 0$ the distribution of $D_\infty(x) - D_\infty(y)$ given $D_\infty(y) = 0$ depends on both $\delta = x - y$ and y (see Eqs. (11) and (12)). We find that $P(D_\infty(\delta + y) = 0 \mid D_\infty(y) = 0)$ decreases in y for all δ. Equations (8)–(12) are in agreement with Eq. (7). New ancestral lines to subsample $\mathscr{D}_\infty$ arrive almost like points in a Poisson process with intensity 1.

Here, we will briefly discuss two other scenarios: (2) $q_n \to 0$ but k_n remains fixed, $k_n = k$ for all n, and (3) $k_n \to \infty$ but $k_n/n \to q > 0$. Both cases are natural extensions of the previously discussed case, (1) $k_n/n \to 0$ and $k_n \to \infty$, but both have limitations wrt applications. Case (2) cannot handle infinite size samples, but results similar to those given in case (1) can be proven. Case (3) can handle large sample sizes, but a good approximation does not exist.

Theorem 3. *Assume that* $q_n \to 0$ *and* $k_n = k$ *for all n. The chain* $A(j)/n, j = 0, 1, \ldots$, *converges in distribution to a Markov process* $A_\infty(j)/k$, $j = 0, 1, \ldots, k$ (*this notation is* *consistent with the notation adopted in Theorem* 1), *with marginal distributions given by*

$$\frac{A_\infty(j)}{k} \sim \frac{\Gamma(k+2)}{\Gamma(j+2)\,\Gamma(k-j)}\, x^{j+1}(1-x)^{k-j-1}, \qquad (13)$$

with $0<x<1$. *That is,* $A_\infty(j)/k$ *is beta distributed,* $\mathrm{Beta}(j+2, k-j)$. *The ratios* $R_j = \{A_\infty(j) - A_\infty(j-1)\}/\{k - A_\infty(j-1)\}$, $j \geqslant 1$, *form a series of independent variables with*

$$R_j \sim \mathrm{Beta}(1, k-j). \qquad (14)$$

In particular, R_j *is independent of* $A_\infty(j')$, $j'=0, 1, ..., j-1$.

As k increases the distribution of the process $A_\infty(j)$, $j \geqslant 0$, in Theorem 3 tends to that of $A_\infty(j)$, $j \geqslant 0$, in Theorem 1.

Theorem 4. *Assume that* $k_n/n \to q > 0$ *and* $k_n \to \infty$. *The chain* $A(j)$, $j=0, 1, ...$, *converges in distribution to a Markov process* $A_\infty(j)$, $j=0, 1, ...$, *with marginal distributions given by*

$$P(A_\infty(j)=m) = \binom{m}{j+1} q^{j+2}(1-q)^{m-j-1} \qquad (15)$$

for $j=0, 1, ...,$ *and* $m=j+1, j+2, ...,$ *or*

$$A_\infty(j) - j - 1 \sim \mathrm{NB}(j+2, q), \qquad (16)$$

where $\mathrm{NB}(\alpha, \kappa)$ *denotes a negative binomially distributed variable with parameters* α *and* κ (*in the notation of Feller,* 1950, *p.* 165). *Further, the increments* $A_\infty(j) - A_\infty(j-1)$, $j \geqslant 1$, *form a series of independent variables with*

$$A_\infty(j) - A_\infty(j-1) \sim \mathrm{Geo}(q). \qquad (17)$$

Here $\mathrm{Geo}(\lambda)$ *denotes a geometrical variable with parameter* λ. *In particular,* $A_\infty(j) - A_\infty(j-1)$ *is independent of* $A_\infty(j')$, $j'=0, 1, ..., j-1$.

As q decreases to zero the distribution of the process $qA_\infty(j)$, $j \geqslant 0$, in Theorem 4 tends to that of the process described in Theorem 1. Also Eqs. (13) and (14), respectively Eqs. (16) and (17), provide simple schemes analogous to (7) to simulate an approximate distribution of $A(j)$, $j \geqslant 0$, if k_n is fixed, respectively if $k_n/n \to q > 0$ applies.

When $k_n/n \to q > 0$, the chain converges to a process in which the distribution depends on the frequency q. A similar remark applies to the previous case, k_n fixed, where the limiting process depends on k, the fixed number of chromosomes in subsample $\mathscr{D}$.

Further, there is a striking difference between on the one hand cases (1) and (2), $k_n/n \to 0$, and on the other hand case (3), $k_n/n \to q > 0$. In the former cases the distribution of $A_\infty(0), A_\infty(1), ..., A_\infty(j)$ has a continuous density (wrt Lebesgue measure on $\mathscr{R}^j$,) whereas in the latter case the distribution is discrete. Also, under (1) and (2) the number of ancestors of the entire sample while there are j ancestors of $\mathscr{D}$ is very large, about $A_\infty(j)/q$, a number that becomes infinity as $q \to 0$. This fact has important consequences for the distribution of waiting times between coalescent events in subsample $\mathscr{D}$ and will be discussed in the next section.

3. TIME IN THE GENEALOGY OF $\mathscr{D}$

In this section, and here only, we assume that the population is of constant size. Let V_h denote the waiting time while there are h ancestors of the entire sample. The variable V_h is exponentially distributed with parameter $h(h-1)/2$. Define $U(j)$, $j \geqslant 1$, to be the time while there are at least j ancestors of subsample $\mathscr{D}_n$; $U(j+1) = V_n + \cdots + V_{A(j)+1}$, $j \geqslant 0$. We have $U(j) > U(j+1)$, for all $j \geqslant 1$. The process $U(j)$, $j \geqslant 1$, is not Markov because the value of the chain $A(j)$, $j \geqslant 0$, is not known. Subscript n is suppressed in the notation of $U(j)$. If q is small, $A(j)$ is large (Theorem 1) and $U(j)$ is a sum of variables with almost negligible variances, $\mathrm{Var}(V_h) = 4/h^2(h-1)^2$, thereby indicating that $U(j+1)$ is in an almost deterministic relation to $A(j)$. The next theorem states this relation.

Theorem 5. *Assume that* $q_n \to 0$ *and* $k_n \to \infty$. *The process* $(U(j+1)/q_n, q_n A(j))$, $j \geqslant 0$, *is Markov and converges in distribution to a Markov process* $(U_\infty(j+1), A_\infty(j))$, $j \geqslant 0$, *that fulfills*

$$U_\infty(j+1) = \frac{2}{A_\infty(j)}. \qquad (18)$$

The distribution of the process $U_\infty(j+1)$, $j \geqslant 0$, can be found from the distribution of the process $A_\infty(j)$, $j \geqslant 0$, and a simple transformation taking $A_\infty(j)$ into $U_\infty(j+1)$. It follows that $U_\infty(j)$, $j \geqslant 1$, forms a Markov chain because $A_\infty(j)$, $j \geqslant 0$, is Markov (Theorem 1). We find that it is natural to measure time in units of qN generations as the variable $U(j)/q$, converges for $q_n \to 0$ and as time in the coalescent is measured in units of N generations.

Note that if $k_n/n \to q > 0$, $A_\infty(j)$ has a discrete distribution. Thus, the time from the present until there are j ancestors is given by a sum of exponential variables $U_\infty(j+1) = V_{A_\infty(j)+1} + V_{A_\infty(j)} + \cdots$ and there is no deterministic relation between $A_\infty(j)$ and $U_\infty(j)$.

Let $W_\infty(j) = U_\infty(j) - U_\infty(j+1)$ be the time while there are j ancestors of $\mathscr{D}_\infty$. We have from (5),

$$U_\infty(j) \sim \frac{2^{j+1}}{j!\, u^{j+2}} \exp(-2/u), \tag{19}$$

with $u > 0$, and $j \geqslant 1$. The variable $U_\infty(j)$ follows a generalized inverse Gaussian distribution, GIG$(-(j+1), 4, 0)$ (in the notation of Seshadri, 1993, p. 27). Further, by evaluation of the joint distribution of $(W_\infty(j), U_\infty(j+1))$ (Theorems 1 and 5) we find

$$W_\infty(j) \mid U_\infty(j+1) = u \sim (j+1)\left(\frac{u}{w+u}\right)^{j+1} \frac{1}{w+u}, \tag{20}$$

with $w, u > 0$. Especially, we see that $W_\infty(j)$ is *not* independent of $U_\infty(j+1)$ as is the case in the ordinary coalescent setup where V_h is independent of all times $V_{h'}$, $h' > h$. However, it follows from (20) that the relative increment $\Delta_j = W_\infty(j)/U_\infty(j+1), j \geqslant 1$, is independent of $U_\infty(j+1)$ and has distribution $\Delta_j \sim (j+1)(1+x)^{-(j+2)}$, $x > 0$. Further, because $U(j+1)$, $j \geqslant 1$, is Markov, the relative increments, $\Delta_j, j \geqslant 1$, form a series of independent variables.

For reasons of comparison we note the following moments of $U_\infty(j)$:

$$E(U_\infty(j)) = \frac{2}{j} \quad \text{and} \quad \operatorname{Var}(U_\infty(j)) = \frac{4}{j^2(j-1)}. \tag{21}$$

In particular, $U_\infty(1)$ has expectation 2 and infinite variance, and $U_\infty(2)$ has expectation 1 and variance 1. This is in agreement with results in Wiuf and Donnelly (1999). The times $W_\infty(j), j \geqslant 2$, have moments

$$E(W_\infty(j)) = \frac{2}{j(j+1)},$$
$$\operatorname{Var}(W_\infty(j)) = \frac{4(j+3)}{(j-1)\, j^2 (j+1)^2}, \tag{22}$$

and

$$\operatorname{Cov}(W_\infty(j), W_\infty(j+1)) = \frac{4(j+4)}{j^2 (j+1)^2\, (j+2)},$$

where Var and Cov denote variance and covariance, respectively.

Theorem 5 also holds if $q_n \to 0$ and k_n is fixed; $k_n = k$. However, the form of the densities of $U_\infty(j)$, $j \geqslant 1$, depends on whether $k_n \to \infty$ or $k_n = k$ (Theorems 1 and 2).

4. CONDITIONING ON THE MUTATION

We now impose the condition that a single mutation occurred at the locus on the branch γ of length $W(1) \approx q_n(U_\infty(1) - U_\infty(2))$, and that no other mutations have occurred at the locus in the history of the sample. The branch γ spans the time from the MRCA of $\mathscr{D}$ until $\mathscr{D}$ shares an ancestor with $\mathscr{C}$ (see Fig. 1). Recall that M denotes the event that there is a single mutation along γ. We assume that the mutation process at the locus is Poisson with rate $\theta/2$. The parameter θ is related to the effective size $N = N(0)$ of the population by $\theta = 2N\mu$, where μ is the chance of a mutation at the locus per chromosome per generation. To ensure that at most one mutation occurs, we consider the limiting case in which the mutation rate tends to zero. Formally, for n fixed we let $\theta \to 0$ and then $q_n \to 0$ and $k_n \to \infty$.

The effect of conditioning on the mutation is studied in a general demography and exemplified in two scenarios; see Sections 5 and 6. The results, obtained in Section 2, on the ancestral chains $A(j)$, $j = 0, 1, \ldots$, and $D(m)$, $m = 1, 2, \ldots$, conditional on E only, apply here as well because the jump chains are stochastically independent of times between coalescent events (Kingman, 1982). However, the conditioning on the mutation has effects on the tines between coalescence events, effects that depend on changes in the population size.

Assume that $\nu(t; \alpha) = N(t)/N$ depends on a vector of parameters $\alpha = (\alpha^1, \ldots, \alpha^d)$ describing the demography. For example if $N(t)$ is constant then $\nu(t; \alpha) = 1$ and $\alpha = (\;)$, and if $N(t) = N\exp(-\beta t)$ then $\nu(t; \alpha) = \exp(-\beta t)$ and $\alpha = (\beta)$. Following Griffiths and Tavaré (1994), we define the population size intensity function by

$$\Lambda(t; \alpha) = \int_0^t \frac{1}{\nu(u; \alpha)}\, du. \tag{23}$$

Put $\lambda(t'; \alpha) = \nu(\Lambda^{-1}(t'; \alpha); \alpha)$, where $\Lambda^{-1}(t'; \alpha)$ denotes the inverse of (23).

Consider a sequence of models $\mathscr{D}_n$ each with a given α_n and q_n. Assume that α_n and q_n are related such that $\alpha_n q_n \to a = (a^1, \ldots, a^d)$ as n tends to infinity.

THEOREM 6. *Assume that $q_n \to 0$ and $k_n \to \infty$. Put $s = t/q$, $s' = t'/q$, and $a = q\alpha$. Further, assume that v expressed as a function of s, a, and q satisfies*

$$v(s;\, a,\, q) = v(s;\, a), \tag{24}$$

i.e., depends on s and a only, and that $v(s;\, a)$ is a continuous function in s and a. Then also λ expressed as a function of s', a, and q satisfies $\lambda(s;\, a,\, q) = \lambda(s;\, a)$. Further, the process $(U(j+1)/q_n,\, q_n A(j))$, $j \geqslant 0$, conditional on M, is Markov and converges in distribution to a Markov process $(U_\infty(j+1),\, A_\infty(j))$, $j \geqslant 0$, that fulfills

$$\int_0^{U_\infty(j+1)} \frac{1}{v(z;\, a)}\, dz = \frac{2}{A_\infty(j)}. \tag{25}$$

Essentially, the proof of Theorem 6 relies on Theorem 5 obtained under the assumption of a population of constant size (see Appendix). If the size of the population is constant, v is constant and we retrieve the relation between $U(j+1)$ and $A(j)$ from Theorem 5; $U(j+1) = 2/A(j)$.

THEOREM 7. *With the assumptions of Theorem 6, the chain $A_\infty(j)$, $j \geqslant 0$, conditional on M, is Markov and fulfills*

$$(A_\infty(0),\, A_\infty(1)) \mid M \sim \frac{1}{Q(M;\, a)}\, x_0 \exp(-x_1) \int_{2/x_1}^{2/x_0} \lambda(z;\, a)\, dz, \tag{26}$$

with $x_1 > x_0 > 0$, and

$$Q(M;\, a) \equiv E\left[\int_{2/A_\infty(1)}^{2/A_\infty(0)} \lambda(z;\, a)\, dz\right]. \tag{27}$$

Further,

$$A_\infty(j+1) \mid (A_\infty(j),\, M) \sim A_\infty(j+1) \mid A_\infty(j) \tag{28}$$

for $j \geqslant 1$. The distribution of $A_\infty(1)$ given $A_\infty(0)$ and M is not independent of M.

It is interesting to note that $A(j) \mid M$, $j \geqslant 0$, is not in general Markov, though $A(j)$, $j \geqslant 0$, is. The process $A_\infty(j) \mid M, j \geqslant 0$, becomes Markov because of the one-to-one correspondence with time and the Markov property of $(U_\infty(j+1),\, A_\infty(j))$, $j \geqslant 0$ (see Theorem 6).

COROLLARY 1. *With the assumptions of Theorem 6, the processes $A_\infty(j)$, $j \geqslant 0$, and $U_\infty(j+1)$, $j \geqslant 0$, conditional on M, have distributions that depend on q and α through a only.*

Theorems 6 and 7 allow us to find the distribution of the time, T_∞, at which the mutation arose.

COROLLARY 2. *The age of the mutation, T_∞, has distribution given by*

$$T_\infty \mid (M,\, U_\infty(1),\, U_\infty(2)) \sim Z W_\infty(1) + U_\infty(2), \tag{29}$$

where Z is uniform on $(0, 1)$ and independent of $U_\infty(j)$, $j \geqslant 1$. The distribution depends on q and α through a only.

Corollaries 1 and 2 are somewhat remarkable results; if a is fixed the frequency q affects time only through a linear scaling. Again we stress that the natural measure of time in the genealogy of $\mathscr{D}$ is in units of qN generations.

In some instances simulations from the process $A_\infty(j)$, $j \geqslant 0$, given M can successfully be performed using an acceptance–rejection scheme; simulate $A_\infty(j)$ from the unconditional process, and accept the outcome with a probability $p(A_\infty(0),\, A_\infty(1))$ proportional to $\int_{2/A_\infty(1)}^{2/A_\infty(0)} \lambda(z;\, a)\, dz$ (see, e.g., Ripley, 1987). Values of T_∞ can easily be obtained from those of $A_\infty(0)$ and $A_\infty(1)$.

Theorem 6 and Corollaries 1 and 2 also hold if $k_n = k$ and $q_n \to \infty$, but the distribution of $A(j)$, $j \geqslant 0$, conditional on M, will take a different form (compare Theorems 1 and 3).

5. CONSTANT POPULATION SIZE

The first example concerns the scenario where the effective size N of the number of chromosomes remains constant through time; i.e., $N(t) = N$. The chains $A(j)$, $j = 0, 1, ...,$ and $D(m)$, $m = 1, 2, ...,$ conditional on M, are discussed in some length in Wiuf and Donnelly (1999), and closed expressions for the distributions of $A(j)$ and $D(m)$ conditional on M are obtained. These allow us to prove very similar results to the results derived in Sections 1 and 2. Here, however, we apply the general results from Section 4.

Applying the theory in the preceding section with $v(t;\, \alpha) = \lambda(t';\, \alpha) = 1$ and $\alpha = (\;)$ we find

$$(A_\infty(0),\, A_\infty(1)) \mid M \sim 2\left(1 - \frac{x_0}{x_1}\right) \exp(-x_1), \tag{30}$$

with $x_1 > x_0 > 0$. In particular,

$$A_\infty(1) \mid M \sim \Gamma(2, 1), \tag{31}$$

and

$$A_\infty(0) \mid M \sim 2\exp(-x_0) - 2x_0 \int_1^\infty \frac{1}{u} \exp(-x_0 u)\, du. \tag{32}$$

Based on (31) and Theorems 1 and 7 the distribution of $A_\infty(j), j \geqslant 1$, given M can be found,

$$A_\infty(j) \mid M \sim \Gamma(j+1, 1), \tag{33}$$

and there exist exponential variables $X_j, j \geqslant 0$, such that $X_j \sim \text{Exp}(1)$, and

$$A_\infty(j) \mid M \sim X_0 + X_1 + \cdots + X_j. \tag{34}$$

Comparing with the distribution of $A_\infty(j)$ conditional on E only, we find $A_\infty(j) \sim \Gamma(j+2, 1)$, and the effect of conditioning on the mutation is the "removal" of an exponential variable. The number of ancestors goes down when conditioning on M. This is not a surprise because if $A(1)$ is small the branch γ tends to be longer and there is a higher chance of a mutation along γ (see also Wiuf and Donnelly, 1999, for a discussion of this).

From Theorem 6 we find $U_\infty(j+1) = 2/A_\infty(j)$ and conclude from (33) that

$$U_\infty(j) \mid M \sim \frac{2^j}{(j-1)!\, u^{j+1}} \exp(-2/u), \tag{35}$$

with $x > 0$, and $j \geqslant 2$. The variable $U_\infty(j) \mid M$ follows a generalized inverse Gaussian distribution, GIG$(-j, 4, 0)$. Further, the joint distribution of $(W_\infty(j), U_\infty(j+1))$ given M is given by

$$W_\infty(j) \mid (M, U_\infty(j+1) = u) \sim j\left(\frac{u}{w+u}\right)^{j+1} \frac{1}{w+u}, \tag{36}$$

with $u, w > 0$ and $j \geqslant 2$. Similarly to the case conditional on E only, the relative increment, $\Delta_j = W_\infty(j)/U_\infty(j+1)$, is independent of $U_\infty(j+1)$ and has distribution $\Delta_j \sim j(1+x)^{-(j+1)}$, $x > 0$. Also here the relative increments form a series of independent variables.

Note that the moments of $U_\infty(j), j \geqslant 1$, take the form

$$E(U_\infty(j)) = \frac{2}{j-1} \quad \text{and}$$

$$\text{Var}(U_\infty(j)) = \frac{4}{(j-1)^2 (j-2)}. \tag{37}$$

In particular, $U_\infty(1)$ has infinite expectation variance, and $U_\infty(2)$ has expectation 2 and infinite variance. This is in agreement with results in Wiuf and Donnelly (1999). The times $W_\infty(j)$ while there are j, $j \geqslant 1$, ancestors of subsample $\mathscr{D}_\infty$ have moments

$$E(W_\infty(j)) = \frac{2}{(j-1)\, j},$$

$$\text{Var}(W_\infty(j)) = \frac{4(j+2)}{(j-2)(j-1)^2\, j^2}, \tag{38}$$

and

$$\text{Cov}(W_\infty(j), W_\infty(j+1)) = \frac{4(j+3)}{(j-1)^2\, j^2 (j+1)}.$$

The variables $W_\infty(j)$, $j \geqslant 1$, conditional on E only, have first and second moments given by $2/j(j+1)$ and $4(j+3)/[(j-1)\, j^2 (j+1)^2]$ (see (22)). Thus, the conditioning on M increases both the expectations and the variances.

Comparing with the ordinary coalescent model we see that the first moment of $W_\infty(j)$ given M is identical to the first moment of the "same" variable in the ordinary coalescent model. But the expressions of variances as well as covariances are not shared in the two models.

If $j = 1$ the joint distribution of $(U_\infty(j), U_\infty(j+1))$ given M takes the form (see (30))

$$(U_\infty(1), U_\infty(2)) \mid M \sim \frac{8}{u_1^2 u_2^2}\left(1 - \frac{u_2}{u_1}\right) \exp(-2/u_2), \tag{39}$$

and in combination with (29) this gives in turn the distribution of the age, T_∞, of the mutation:

$$T_\infty \mid M \sim \frac{2}{t^2} \exp(-2/t) \sim \text{GIG}(-1, 4, 0). \tag{40}$$

Note the strong similarity between the distributions of $U_\infty(j)$, $j \geqslant 2$, and T_∞. By inspection, it can be seen that Eqs. (35)–(38) hold as well for $j=1$ if $U_\infty(1)$ is replaced by T_∞. These facts are also explored and elaborated on in Wiuf and Donnelly (1999). The expected age of the mutation is infinite in agreement with Kimura and Ohta (1973) among others.

6. EXPONENTIAL DECREASING POPULATION SIZE

Consider now a population where the effective number of chromosomes decreases exponentially going backward in time, $N(t)=N(0)\exp(-\beta t)$. The parameter β is related to $N=N(0)$ by $\beta = Nr$, where r is the rate of decrease in the population size from generation to generation. Both Griffiths and Tavaré (1998) and Wiuf and Donnelly (1999) discuss this setting in some detail, but only a few analytical results have been derived.

Here $v(t;\alpha)=v(t;\beta)=\exp(-\beta t)$ and $\lambda(t';\beta)=1/(\beta t'+1)$. Let the limit of $\beta_n q_n$ be b. We find (Theorem 7) that

$$(A_\infty(0), A_\infty(1)) \mid M \sim \frac{x_0\exp(-x_1)}{C(b)}\{\log(2b/x_0+1)-\log(2b/x_1+1)\} \tag{41}$$

with $x_1 > x_0 > 0$, and $C(b)$ being a norming constant. In particular, by integration over x_0 in (41)

$$A_\infty(1) \mid M \sim \frac{b\exp(-x_1)}{C(b)}\{x_1 - 2b\log(1+x_1/2b)\}. \tag{42}$$

If b is large, $A_\infty(1) \mid M$ almost follows a gamma distribution, $\Gamma(3, 1)$, and simulations of $A_\infty(1) \mid M$ can in general be performed using an acceptance–rejection scheme with acceptance probability $1-2b\log(1+X/2b)/X$ and proposal $X \sim \Gamma(2, 1)$.

Simulation results (not shown) show that the number, $A_\infty(1)$, of ancestors increases (in expectation) with increasing b. Remember that in the constant population size case the expectation of $A_\infty(1)$ given M is less than the expectation of $A_\infty(1)$, i.e., $E(A_\infty(1) \mid M) < E(A_\infty(1))$, to allow for longer branches γ. A similar effect is expected for $b>0$, but as b increases the population size decreases fast and branches between events become tiny. In order to allow for the mutation to occur at all, the MRCA of $\mathscr{D}_\infty$ must be put forward in time; that is, $A_\infty(1)$ is likely to increase with increasing b.

From Theorem 6 and (41) we can find the joint distribution of $(U_\infty(1), U_\infty(2))$ given M,

$$(U_\infty(1), U_\infty(2)) \mid M \sim \frac{8b^6(u_1-u_2)\,e^{b(u_1+u_2)}}{C(b)(e^{bu_1}-1)^3\,(e^{bu_2}-1)^2}\exp\left\{-\frac{2b}{e^{bu_2}-1}\right\}, \tag{43}$$

with $u_1 > u_2 > 0$. The distribution until the MRCA of $\mathscr{D}_\infty$ can be found from (43) by integration over u_1, $u_1 > u_2$, and the distribution of the age of the mutation, T_∞, can be found using Corollary 2 and (43). This results in

$$T_\infty \mid M \sim \frac{2b^3}{C(b)(e^{bt}-1)^2}\exp\left\{-\frac{2b}{e^{bt}-1}\right\}. \tag{44}$$

It is easily seen that the expectation of (44) is finite for any $b>0$ in contrast to the case of a population of constant size where the age of the mutation has infinite expectation.

7. A SIMULATION ALGORITHM

In this section we review some results obtained by Saunders *et al.* (1984) on nested subsamples and relate them to the results found in the previous sections. The results found by Saunders *et al.* (1984) apply to binary coalescent trees (Griffiths and Tavaré, 1998), trees where each pair of genes has an equal chance of forming the next coalescence, independently of the times between events. Thus, in particular, they apply to the trees discussed in this paper.

Consider two samples $\mathscr{D}_0$ and $\mathscr{D}$ such that $\mathscr{D}_0$ is a subsample of $\mathscr{D}$. Assume that the sample size of $\mathscr{D}$ is infinite and that of $\mathscr{D}_0$ is k_0, with k_0 finite. Saunders *et al.* (1984) found the probability that there are j ancestors of $\mathscr{D}$ at the time just before the (k_0-1)th coalescent event, going backward in time, among ancestors of $\mathscr{D}_0$,

$$p_{k_0-1}(j \mid k_0) = \frac{k_0!\,(k_0-1)}{j(j+1)\cdots(j+k_0-1)} \tag{45}$$

for $j \geqslant 2$. Further, the conditional probability that there are j ancestors of $\mathscr{D}$ at the time just before the ith coalescent event among ancestors of $\mathscr{D}_0$ given that there are j' ancestors just before the $(i+1)$th coalescent event is

$$p_i(j \mid j', k_0) = \frac{p_i(j, j' \mid k_0)}{p_{(i+1)}(j' \mid k_0)} = \frac{i(j'+k_0-i)(j'+k_0-i+1)\cdots(j'+k_0-1)}{(j+k_0-i-1)(j+k_0-i)\cdots(j+k_0-1)} \tag{46}$$

for $j \geqslant j'+1$. Equation (45) is of the same form as (46); in fact, if we put $j'=1$ for $i=k_0-1$ then (46) reduces to (43). It can be shown that the expectation of (46) and (45) is finite unless $i=1$ or $k_0=2$. Moreover, the variance is finite unless $i=1, 2$ or $k_0=2, 3$.

The conditional distributions in (46) can easily be simulated: simulate a uniform variable Z on $(0, 1)$ and find the number j such that $p_i(j-1 \mid j', k_0) < Z \leqslant p_i(j \mid j', k_0)$. If the expectation or the variance is infinite, it is useful to adopt another approach. For example, if $i=1$ the variable $\lfloor (j'+k_0-1)/Z \rfloor - k_0 + 1$ is distributed like (46).

Consider now a sample $\mathscr{D}_0$ taken from $\mathscr{D}$, the population of rare alleles. The genealogy of $\mathscr{D}$ is described in Section 4 and the genealogical structure of $\mathscr{D}_0$ can be simulated as follows:

1. Simulate, according to (46), the number of ancestors j_1, j_2, ... and j_{k_0-1} of $\mathscr{D}$ the first time there are 1, 2, ... and k_0-1 ancestors of $\mathscr{D}_0$, respectively.

2. Put $j_0=1$, and perform the following according to Theorems 5 and 6.

3. Simulate $A_\infty(1) \mid M$.

4. Simulate $A_\infty(j_i) - A_\infty(j_{i-1}) \sim \Gamma(j_i - j_{i-1}, 1)$, $1 \leqslant i \leqslant k_0-1$.

5. Transform $A_\infty(j_i)$ into $U_\infty(j_i)$, $1 \leqslant i \leqslant k_0-1$.

Note that j_1 can be 1; that is, $j_1 = j_0 = 1$. In that case we take $\Gamma(0, 1)$ to be constantly zero.

In the special case with a population of constant size the distribution of the time until there are k_0-i, $1 \leqslant i \leqslant k_0-1$, ancestors of $\mathscr{D}_0$ can be found. We state without proof (the proof is cumbersome and consists in evaluating a summation) the density function $f_i(t)$ of the time until there are k_0-i ancestors

$$f_i(t) = \frac{k_0!(k_0-1)!}{(i-1)!\,(k_0-i)!\,(k_0-i-1)!}\,\frac{t^{k_0-2}}{2^{k_0}} \cdot \left\{ \sum_{j=0}^{k_0-i-1} \binom{k_0-i-1}{j} (-1)^j (k_0+1)(k_0+2) \cdots (k_0+j)\,\frac{2^{k_0-i-1-j}}{t^{k_0-i-1-j}} - e^{-2/t} \sum_{j=0}^{k_0} \frac{2^j}{j!t^j} (-1)^j (k_0+1-j)(k_0+2-j) \cdots (2k_0-i-j-1) \right\}. \tag{47}$$

The distributions of other waiting times as well as the distribution of the genealogy of the sample $\mathscr{D}_0$ can be found, but the expressions are even more complicated than (47), and some do not reduce to finite sums. The expectation of the time while there are i ancestors, $W_i = U_\infty(j_i) - U_\infty(j_{i-1})$, can be found from (37), (45), and (46). We find that $E(W_i) = 2/i(i-1)$.

8. DISCUSSION

In the previous sections we described an approximation to the distribution of the genealogy of $\mathscr{D}$, a subpopulation of neutral rare alleles in a general demography. The restrictions put on the demography are very mild, and are fulfilled in all cases of interest known to the author. Assume that the demography is described by a numberof parameters $\alpha = (\alpha^1, ..., \alpha^d)$ in a coalescent framework. The model is reparameterized such that time is measured in units of qN and such that the parameters are $a = q\alpha = (q\alpha^1, ..., q\alpha^d)$. Here N is the effective size of the population at the present time and q the frequency of the rare allele. The dependence of q upon the distribution of the genealogy of $\mathscr{D}$ is only through a linear scaling of the time; if q is doubled, the lengths of all branches in the genealogy are doubled. For example, if the size of the population is decreasing exponentially at rate $\beta = rN$ per N generations, the genealogy has a distribution that depends on $b = q\beta$ and qN only.

The convergence in distribution of the variables $A(j)$, $j \geqslant 0$, is slowest for $j=0$ because the number of ancestors of the entire population becomes smaller with decreasing number of ancestors of $\mathscr{D}$. When the number of ancestors of the entire population is small it is less likely to be accurately approximated by a continuous variable.

Therefore, the rate of convergence of $A(0)$ or $A(1)$ (which relates to the time of the MRCA of $\mathscr{D}$) measures the rate of convergence of the whole process.

The approximation seems to be fairly accurate, even for frequencies of the neutral rare allele as high as 10%. For example, the approximated density of the time until a MRCA of $\mathscr{D}$ (assuming constant population size) is almost indistinguishable from the densities in the exact coalescent framework (Figs. 2 and 3). But the expectation of the time until a MRCA differs under the approximation and the exact coalescent. The approximated value is 2 whereas the exact values are 1.65 if $q=10\%$ and 1.93 if $q=1\%$ (Wiuf and Donnelly, 1999). The discrepancies in the expectations are more profound for the distribution of the age of the mutation; under the approximation the expectation is infinity whereas in the exact coalescent the expectation is $-2\log(q)/(1-q)$ (Kimura and Ohta, 1973). For example, if $q=10\%$ the expected age is 5.1 and if $q=1\%$ the expected age is 9.3. In contrast to this, it is seen that the mode of the approximated distribution and that of the exact distribution are close to each other for both $q=1\%$ and $q=10\%$.

Several variables have particular interest: (1) The age of the mutation. We found an expression for the distribution of the age of the mutation, T_∞, in a general demography, and we showed that the expected age of the mutation is finite for any $b>0$ in the example discussed above. In contrast, the expected age is infinity under an assumption of constant population size (corresponding to $b=0$). This is illustrated in Fig. 4. (2) The time, $U_\infty(2)$, until a MRCA of subpopulation $\mathscr{D}$. The expected age was shown to be finite for all values of $b \geqslant 0$ (Fig. 4), and the ratio of the expectation of T_∞ to that of $U_\infty(2)$

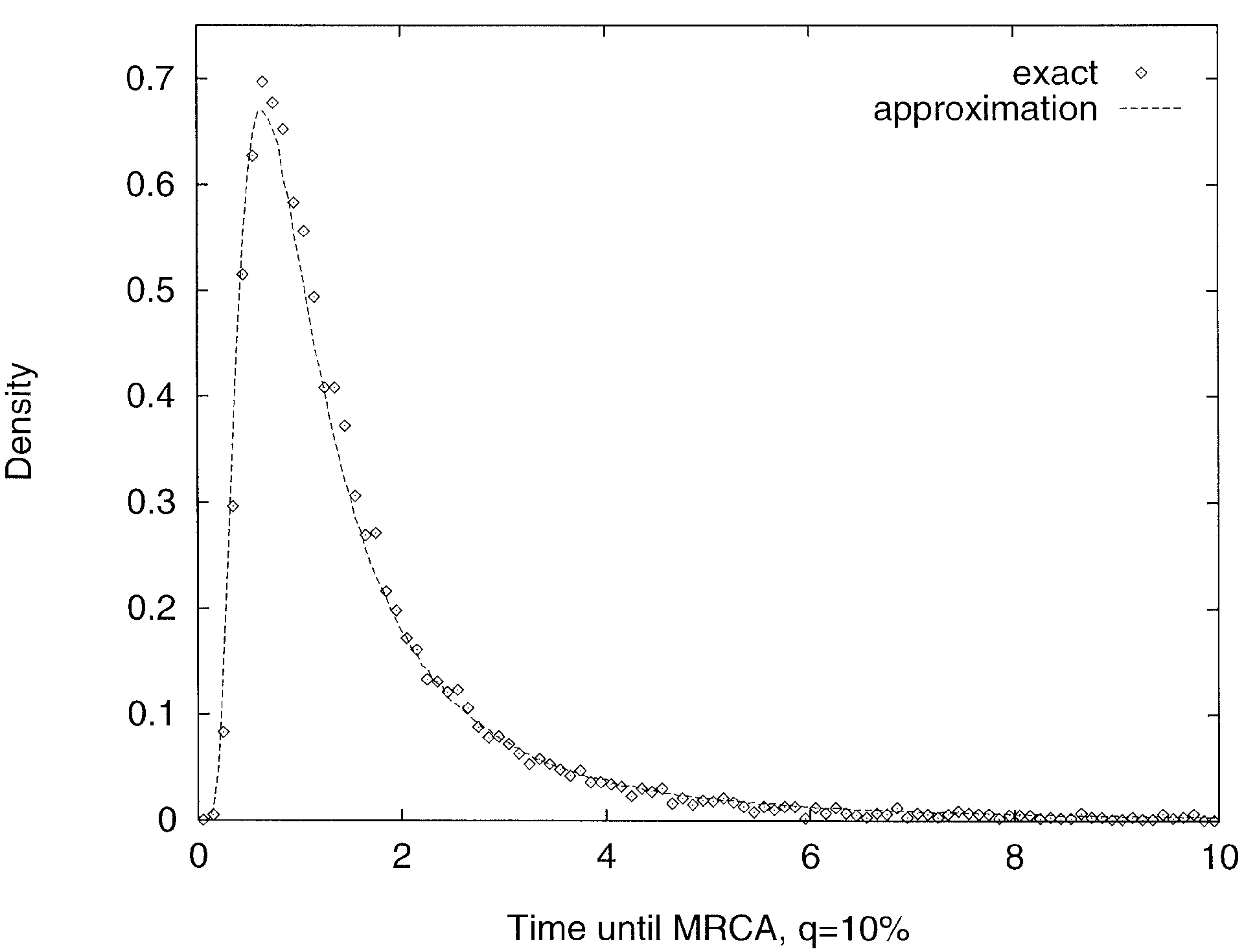

FIG. 2. The approximation versus the exact coalescent, $q=10\%$. The figure shows the density of the time, in units of qN generations, until a MRCA of the population of the rare alleles under the approximation and under the exact coalescent. Although the frequency is as high as 10% the two curves follow each other very closely. The exact curve has a slightly higher peak than the approximated curve. 10^4 simulations were performed to obtain the exact density and the approximated density was obtained using (35).

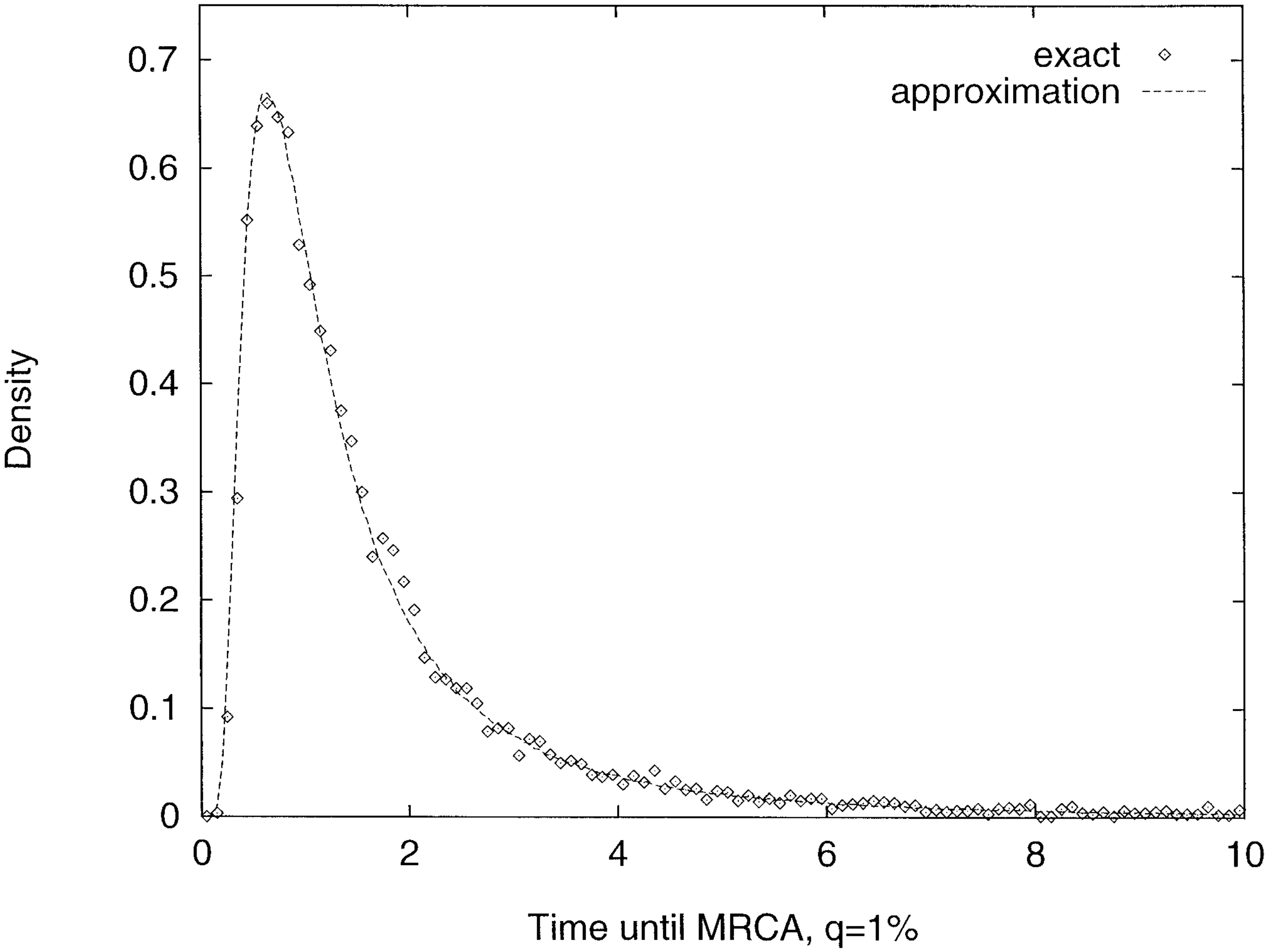

FIG. 3. The approximation versus the exact coalescent, $q = 1\%$. The figure shows the density of the time, in units of qN generations, until a MRCA of the population of the rare alleles under the approximation and as well under the exact coalescent. There are hardly any visual differences. 10^4 simulations were performed to obtain the exact density and the approximated density was obtained using (35).

tends to one as b tends to infinity. For small values of b the ratio becomes arbitrarily large.

APPENDIX

Proof of Lemma 1. The Markov property of $A(j)$, $j = 0, 1, ..., k$, follows from the Markov property of the reversed chain $A(j), j = k, k-1, ..., 0$ (indices decreasing) proven in Wiuf and Donnelly (1999). Corollary 1 in Wiuf and Donnelly (1999) gives (3) and the transition probabilities can be found from Lemma 2 and Corollary 4.

Proof of Theorem 1. Consider m_n such that $m_n \leqslant n$ and $q_n m_n \to x$. It follows that $m_n \to \infty$ because $q_n \to 0$, and $m_n/n \to 0$ because $k_n \to \infty$. From here on subscript n is dropped at m_n and k_n for notational convenience. Consider (3). It can be rewritten as

$$P_n(A(j) = m)$$
$$= \frac{1}{(j+1)!} \frac{(n-m-1)(n-m-2)\cdots(n-m-k+j+1)}{(n-1)(n-2)\cdots(n-k+j+1)}$$
$$\cdot \frac{k(k-1)\cdots(k-j)\, m(m-1)\cdots(m-j)}{(n-k+j)(n-k+j-1)\cdots(n-k)} \frac{(k+1)}{n}$$
$$= \frac{1}{(j+1)!} \frac{(k+1)}{n} F_1 F_2.$$

There are $k-j-1$ terms in both the denominator and the numerator in the first fraction, F_1, and there are $j+1$ terms involving k, m, and n, respectively, in the second fraction, F_2.

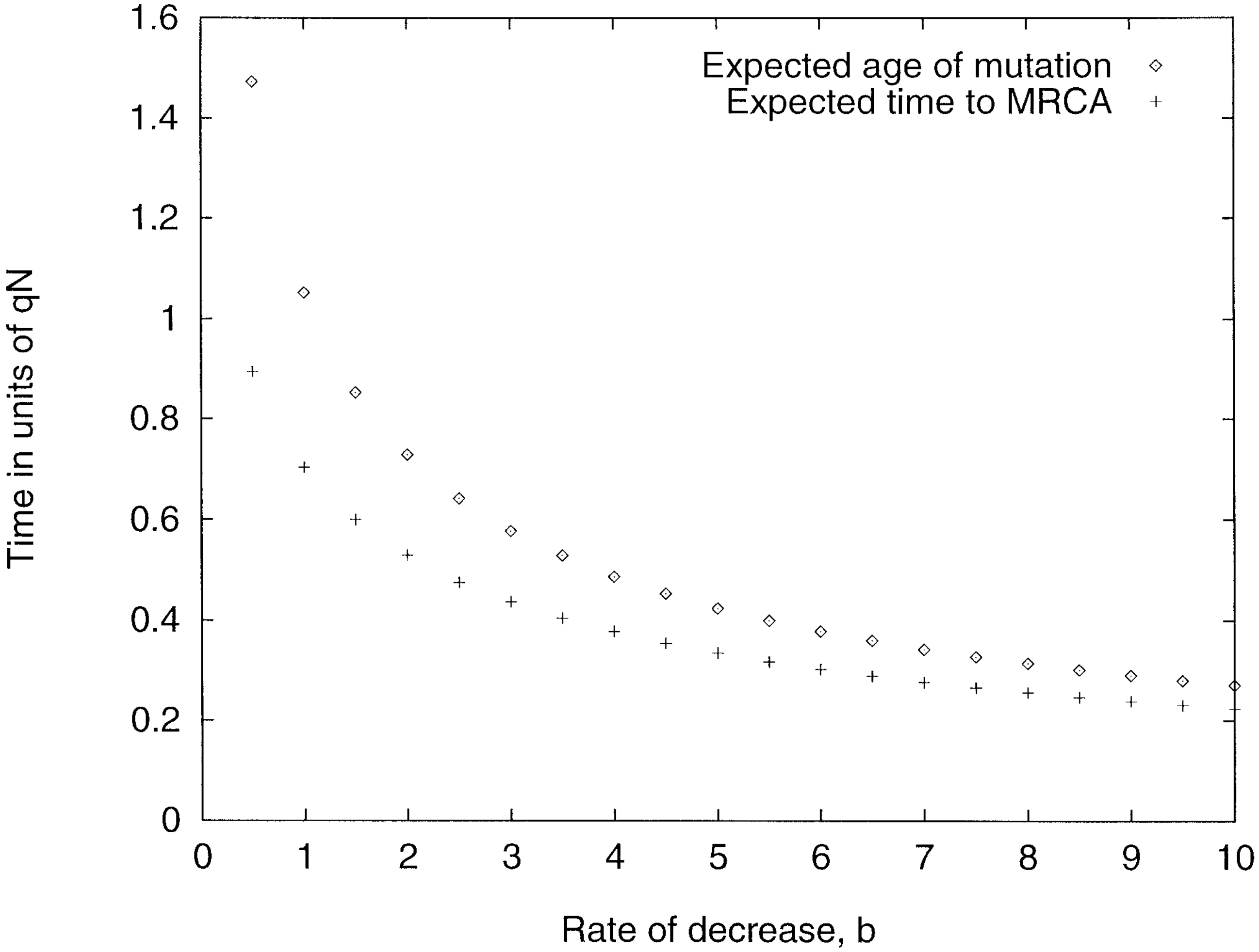

FIG. 4. Age of the mutation and time until a MRCA. The figure shows the expectation of the age of the mutation and the expectation of the time until a MRCA for various values of $b=\beta q$. Time is measured in units of qN generations. When $b=0$, the expected time to a MRCA is 2 while the expected age of the mutation is infinity. As b increases the ratio between the two expectations decreases toward a limit of one.

We have $km/n \to x$ so that $F_2 \to x^{j+1}$. Concerning F_1 we find

$$\log(F_1) = -\sum_{i=m+1}^{m+k-j-1} \frac{i}{n} + \sum_{i=1}^{k-j-1} \frac{i}{n} - \sum_{i=m+1}^{m+k-j-1} \sum_{r=2}^{\infty} \frac{1}{r}\left(\frac{i}{n}\right)^r + \sum_{i=1}^{k-j-1} \sum_{r=2}^{\infty} \frac{1}{r}\left(\frac{i}{n}\right)^r,$$

using the series expansion of the logarithm. The sum of the first two terms converges to $-x$. The absolute value of the third term is easily seen to be bounded by

$$-(k-j-1)\frac{(m+k-j-1)}{n}\log\left(1-\frac{m+k-j-1}{n}\right) \leqslant \frac{k^3}{n^2}(1+f_1(m,k,n)), \tag{48}$$

and the fourth by

$$-(k-j-1)\frac{(k-j-1)}{n}\log\left(1-\frac{k-j-1}{n}\right) \leqslant \frac{k^3}{n^2}(1+f_2(k,n)), \tag{49}$$

where f_i, $i=1,2$, are functions that tend to zero under the above conditions. Note that m and $k-j-1$ in $\log(F_1)$ are symmetric in the sense that

$$-\sum_{i=m+1}^{m+k-j-1} a_i + \sum_{i=1}^{k-j-1} a_i = -\sum_{i=k-j}^{m+k-j-1} a_i + \sum_{i=1}^{m} a_i,$$

where a_i denotes the inner sum over r. Thus, Eqs. (48) and (49) also apply with $k-j-1$ replaced by m and vice versa. It follows that the two terms are dominated by m^3/n^2 also.

The following suffices to prove that $\log(F_1)$ converges to $-x$: For all ε, there exists an N such that either $m^3/n^2 < \varepsilon$ or $k^3/n^2 < \varepsilon$ for each $n > N$ (not necessarily the same inequality for all n). To prove it choose N such that $k^3m^3/n^4 < \varepsilon^2$ for all $n > N$. This can be done because $km/n \to x$ and, thus, $k^3m^3/n^4 \approx x^3/n$. But then either $k^3/n^2 < \varepsilon$ or $m^3/n^2 < \varepsilon$ which completes the proof.

Combining the results gives

$$\frac{1}{q} P_n(A(j) = m) \to \frac{1}{(j+1)!} x^{j+1} \exp(-x)$$

which proves (5). Equation (6) is obtained similarly using (4).

From the Markov property of the chain $q_n A(j)$ and Eqs. (5) and (6) it follows that any finite vector $q_n A(0), q_n A(1), ..., q_n A(j')$ converges in distribution to a vector $A_\infty(0), A_\infty(1), ..., A_\infty(j')$ which is Markov. Convergence of finite dimensional marginals ensures convergence of the whole process (see, e.g., Pollard, 1984).

Proof of Theorem 2. Equations (8) and (9) are proven similarly to (5). From (1), (2), and (7) the conditional probabilities in (10), (11), and (12) can easily be derived. Together this proves convergence in distribution of any finite dimensional vector, $D(x_1/q_n), D(x_2/q_n), ..., D(x_j/q_n)$, to $D_\infty(x_1), D_\infty(x_2), ..., D_\infty(x_j)$, and convergence of the whole process $D(x/q_n)$, $x > 0$, to $D_\infty(x)$, $x > 0$, follows thereupon (see, e.g., Pollard, 1984). The Markov property of $D_\infty(x)$, $x > 0$, follows from the Markov property of $D(x/q_n)$, $x > 0$.

Proof of Theorem 3. The number of terms in F_1 and F_2 (see the proof of Theorem 1 are independent of n and the limit probability is easily evaluated. The rest follows similarly to the proof of Theorem 1.

Proof of Theorem 4. Similarly to considerations in Wiuf and Donnelly (1999).

Proof of Theorem 5. The Markov property of $(U(j+1)/q_n, q_n A(j))$, $j \geqslant 0$, follows easily from the Markov property of $A(j), j \geqslant 0$, and that $A(j), j \geqslant 0$, and V_h, $h \geqslant 2$, are independent. To prove (18) we consider the finite dimensional conditional distributions of $U(j+1)/q_n, j \geqslant 0$, given the $q_n A(j), j \geqslant 0$. For reasons of simplicity we consider the one-dimensional case only. Let m and l be such that $l < m \leqslant n$, $q_n m \to x$ and $q_n l \to y$. It follows that $m, l \to \infty$ because $q_n \to 0$, and $m/n, l/n \to 0$ because $k_n \to \infty$. On the event $\{A(j) = \lfloor x/q_n \rfloor\}$, $U(j+1) = V_n + \cdots + V_{\lfloor x/q_n \rfloor + 1} \equiv T_{\lfloor x/q_n \rfloor}$. Further, $T_{\lfloor x/q_n \rfloor}$ is independent of $q_n A(j)$. Next we prove that $T_{\lfloor x/q_n \rfloor}$ converges to a degenerate variable. Note that $E_n(T_{\lfloor x/q_n \rfloor}/q_n) \to 2/x$ and $\mathrm{Var}(T_{\lfloor x/q_n \rfloor}/q_n) \to 0$. The last statement follows from

$$\mathrm{Var}_n(V_m + \cdots + V_{l+1}) = 8 \sum_{h=l}^{m} \frac{1}{h^2} - \frac{4}{m^2} - \frac{4}{l^2} + \frac{8}{m} - \frac{8}{l}$$

(Tavaré *et al.*, 1997), and

$$\frac{1}{q_n^2} \left\{ \sum_{h=l}^{m} \frac{2}{h^2} - \frac{1}{m^2} - \frac{1}{l^2} + \frac{2}{m} - \frac{2}{l} \right\}$$
$$= \frac{n}{k^2} \int_{l/n}^{m/n} \frac{2}{z^2}\, dz + \frac{2n^2}{k^2 m} - \frac{2n^2}{k^2 l} + O(1/l) = O(1/l).$$

The variables $T_{\lfloor x/q_n \rfloor}$, $n \geqslant 2$, can be constructed such that all of them are defined on the same probability space. If so done the above equation proves that the series $T_{\lfloor x/q_n \rfloor}$, $n \geqslant 2$, converges to $2/x$ in L^2-norm, hence also in distribution. In conclusion, $P(U_\infty(j+1) = 2/x \mid A_\infty(j) = x) = 1$ which proves (18). The Markov property of $(U_\infty(j+1), \ A_\infty(j))$, $j \geqslant 0$, follows from that of $(U(j+1)/q_n, q_n A(j)), j \geqslant 0$.

Proof of Theorem 6. The equation $\lambda(s; a, q) = \lambda(s; a)$ follows from (24). In allowing for variation in population size, one can either rescale the coalescent rates, and keep the mutation rate constant over time, or keep the coalescent rates constant and rescale the mutation rate (Griffiths and Tavaré, 1994). In this proof we rescale the mutation rate and keep coalescent rates constant. The times between coalescent events (conditional on the event E only, not M) are thus described by the usual coalescent model, and the results obtained in Section 3 apply. The relation between time, t', in setup (A) with constant coalescent rates and real time, t, in setup (B) with variable coalescent rates is given by

$$A(t; \alpha) = t'. \tag{50}$$

Time variables in setup (A) are marked. For simplicity we consider only the two-dimensional marginals $(A(0), A(1))$ and $(U'(1), U'(2))$. Let U' and A be short for $(U'(1), U'(2)) = (q_n u'_1, q_n u'_2)$ and $(A(0), A(1)) = (\lfloor x_0/q_n \rfloor, \lfloor x_1/q_n \rfloor)$, respectively. Apply Bayes' theorem to obtain

$$P_n(U' \mid M, A) = \frac{P_n(M \mid A, U')}{P_n(M \mid A)} P_n(U' \mid A).$$

According to Griffiths and Tavaré (1994), $\theta(t') = \theta\lambda(t'; \alpha)$ and the probability of M conditional on A and U' is given by

$$P_n(M \mid A, U') \approx 1 - \exp\left\{ -\int_{q_n u'_2}^{q_n u'_1} \theta(t)/2 \, dt \right\}$$

$$\approx \frac{\theta}{2} \int_{q_n u'_2}^{q_n u'_1} \lambda(z; \alpha) \, dz, \qquad (51)$$

where $\approx$ indicates that only first-order terms in θ are taken into account. Applying (51) and the continuity of λ (which follows from that of v)

$$P_n(U' \mid M, A) = = \frac{P_n(M \mid A, U')}{E_n[P_n(M \mid A, U') \mid A]} \times P_n(U' \mid A) \to P_n(U' \mid A).$$

The relation between $A(j)$ and $U'(j+1)$ follows from Theorem 5. Using (50) and (24) we find

$$\int_0^{U_\infty(j+1)} \frac{1}{v(z; a)} \, dz = U'(j+1) = \frac{2}{A_\infty(j)}$$

as desired. The Markov property follows from that of $(U(j+1)/q_n, q_n A(j)), j \geqslant 0$, given M. The proof that the last process is Markov is proven similarly to the proof given in Theorem 5 that $(U(j+1)/q_n, q_n A(j))$, $j \geqslant 0$ (conditional on E only), is Markov.

Proof of Theorem 7. Similarly to Theorem 6. The Markov property of $A_\infty(j)$, $j \geqslant 0$, follows from the Markov property of $(U_\infty(j+1), A_\infty(j)), j \geqslant 0$, and the relation between $U_\infty(j+1)$ and $A_\infty(j)$ given in Theorem 6.

Proof of Corollary 1. Follows from Theorems 6 and 7.

Proof of Corollary 2. Conditional on $U_\infty(1)$ and $U_\infty(2)$, T_∞ is uniform on $U_\infty(1) - U_\infty(2)$ because the mutation process is Poisson with constant rate.

ACKNOWLEDGMENTS

M. Stephens and P. Donnelly are thanked for reading and commenting on the first version of the manuscript. Their suggestions improved the presentation of the results considerably. An anonymous reviewer is thanked for his patience in dealing with later versions. The author was supported by Grant BBSRC 43/MMI09788 and by the Carlsberg Foundation, Denmark. Part of this work was carried out while the author visited the Isaac Newton Institute, University of Cambridge.

REFERENCES

Feller, W. 1950. "An Introduction to Probability Theory and Its Applications," Wiley, New York.

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.

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. 1982. The coalescent, *Stochastic Proc. Appl.* **13**, 235–248.

Pollard, D. 1984. "Convergence of Stochastic Processes," Springer-Verlag, New York.

Ripley, B. D. 1987. "Stochastic Simulation," Wiley, New York.

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.

Seshadri, V. 1993. "The Inverse Gaussian Distribution," Clarendon Press, Oxford.

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

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

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. 1997. Inferring the coalescence times from DNA sequences, *Genetics* **145**, 505–518.

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.

Wiuf, C., and Donnelly, P. 1999. Conditional genealogies and the age of a neutral mutant, *Theor. Popul. Biol.* **56**, 183–201.