The Annals of Applied Probability
1999, Vol. 9, No. 3, 567–590

THE AGES OF MUTATIONS IN GENE TREES

BY R. C. GRIFFITHS[1] AND SIMON TAVARÉ[2]

University of Oxford and University of Southern California

Under the infinitely many sites mutation model, the mutational history of a sample of DNA sequences can be described by a unique gene tree. We show how to find the conditional distribution of the ages of the mutations and the time to the most recent common ancestor of the sample, given this gene tree. Explicit expressions for such distributions seem impossible to find for the sample sizes of interest in practice. We resort to a Monte Carlo method to approximate these distributions. We use this method to study the effects of variable population size and variable mutation rates, the distribution of the time to the most recent common ancestor of the population and the distribution of other functionals of the underlying coalescent process, conditional on the sample gene tree.

1. Introduction. The seminal paper of Kimura and Ohta (1973) exploited diffusion theory to derive the expected age and the variance of the age of a neutral mutation observed to have frequency x in a *population*. This paper stimulated many authors to study the distribution of the age of an allele; the paper of Watterson (1996) describes some of the history and more of the biological context. The emergence of molecular techniques for assessing genetic variability in different regions of the genome in samples of individuals led directly to the development of a number of inference and estimation techniques for *sample* data. The Ewens sampling formula [Ewens (1972)] was among the first of these. The subsequent development of "coalescent methods" by Kingman (1982a), Tajima (1983) and Hudson (1983) changed the focus of the theory by forcing attention on the role of genealogy. For example, Griffiths and Tavaré (1998) put Kimura and Ohta's results in a coalescent context and obtain an analogous result for the expected age of a mutation observed z times in a sample of n genes.

Computer intensive estimation techniques for coalescent-based models have recently been devised for a number of mutation processes [cf. Griffiths and Tavaré (1994a, c) and Kuhner, Yamato and Felsenstein (1995)]. In such problems, the pattern of mutations in the observed DNA sequences forms the data from which estimates of parameters such as mutation rates are made. Another focus of this research has been ancestral inference, defined broadly as the evaluation (either theoretically or computationally) of the conditional dis-

Received February 1998; revised November 1998.

[1] Supported in part by an Australian Research Council grant at Monash University.

[2] Supported in part by NSF Grant BIR 95-04393.

AMS 1991 *subject classifications*. 60J70, 62M05, 65U05, 92D10, 92D20.

Key words and phrases. Ages of mutations, ancestral inference, coalescent process, gene trees, population genetics, samples of DNA.

tribution of various functionals of the coalescent process, conditional on the observed pattern of mutations, or some summary thereof. A central theme in the theory has concerned the distribution of the time to the most recent common ancestor (TMRCA) of the sample sequences. See, for example, Griffiths and Tavaré (1994b), Fu and Li (1997), Tavaré, Balding, Griffiths and Donnelly (1997). Griffiths and Tavaré exploit a relative of Markov chain Monte Carlo, having roots dating back to papers of Forsythe and Leibler (1950) and Halton (1970), to attack such problems.

In this paper we exploit this method to study the joint conditional distribution of the ages of mutations under a particular model for the mutation process, conditional on the pattern of mutations observed in the data. We use as examples of the approach molecular data taken from the Y chromosome [Whitfield, Sulston and Goodfellow (1995)], and the nuclear gene β-globin [Fullerton, Harding, Boyce and Clegg 1994, Harding, Fullerton, Griffiths, Bond, Cox, Schneider, Moulin and Clegg (1997) and Harding, Fullerton, Griffiths and Clegg (1997)]. We remark that in our approach to this problem, the ages are unobservable random variables; it is then natural to report their conditional distribution given the data.

We begin the paper with a brief description of the type of data we consider and the basic genetic terms used in the sequel.

1.1. *A Melanesian data set.* We use a data set comprising DNA from part of the β-globin locus from a sample of $n = 57$ sequences from a Melanesian population [Fullerton, Harding, Boyce and Clegg (1994)]. Each sequence is $\ell = 2320$ base pairs in length. The data are part of a larger world data set of 326 sequences described in Harding, Fullerton, Griffiths, Bond, Cox, Schnei-

TABLE 1
Melanesian β-globin sequences

	2	1		2	2			2		1	2	2	1	
Site	9	4	5	7	0	9	5	6	3	3	6	5	4	
position*	4	1	3	9	0	0	0	3	7	5	3	5	2	
	5	6	2	2	8	6	8	6	9	8	4	4	3	
Site #	1	2	3	4	5	6	7	8	9	10	11	12	13	
Root	T	T	T	A	T	C	T	C	T	C	G	G	C	
allele														freq
a	G	G	T	A	T	C	T	C	T	C	G	G	C	25
b	G	G	C	A	T	C	T	C	T	C	G	G	C	1
c	T	T	T	T	T	C	T	C	T	C	G	G	C	16
d	T	T	T	A	C	T	C	A	T	C	G	G	C	3
e	T	T	T	A	C	T	C	A	C	G	G	G	C	7
f	T	T	T	A	C	T	C	A	T	G	A	G	C	1
g	T	T	T	A	C	T	C	A	T	G	G	C	T	4

*Site positions from Table 1 of Harding, Fullerton, Griffiths and Clegg (1997).

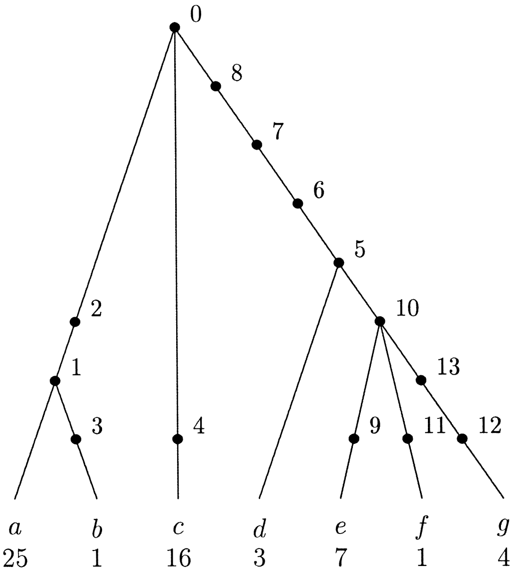


FIG. 1. *Melanesian β-globin tree.*

der, Moulin and Clegg (1997). In data such as these, there are some sites (i.e., positions in the DNA sequence) at which each sequence in the sample is identical, and some sites, called *segregating* sites, at which there is variability. Of the 2320 sites, 2307 were not segregating, and 13 were. In Table 1 a summary of the segregating sites is given.

There are $d = 7$ distinct sequences observed among the 57 sequences in the data; these *alleles* are labeled a–g in Table 1. In the last column the frequencies of the seven alleles in the sample are also given. The type of the ancestral base at each of the segregating sites, inferred from comparison with more distantly related species, is given in the row of the table labeled "Root." For example, site 5 was a T in the ancestral sequence; the alleles a–c have this ancestral base, whereas alleles d–g have the mutant base C. Notice that sites 5–8 have the same mutation structure: the ancestral base appears in alleles a–c, the mutant one in alleles d–g.

The data in Table 1 are equivalent to the rooted gene tree shown in Figure 1 [cf. Griffiths and Tavaré (1995)]. This rooted tree may be constructed as a phylogeny with mutations as characters, using the algorithm of Gusfield (1991), for example. The tree is unique up to permutations of mutations along edges (e.g., the mutations at sites 5, 6, 7 and 8). We use the notation $(T, \mathbf{n})$ for such a tree; T denotes the topology of the tree and $\mathbf{n}$ the multiplicities of its tips.

The matrix of segregating sites, with the ancestral bases replaced by 0 and mutant bases by 1, is the incidence matrix of mutations on lineages. The gene tree $(T, \mathbf{n})$ represented as mutation paths to the root is given in Table 2.

TABLE 2
Paths to root (0) *for Melanesian data*

25	:	1	2	0					
1	:	3	1	2	0				
16	:	4	0						
3	:	5	6	7	8	0			
7	:	9	10	5	6	7	8	0	
1	:	11	10	5	6	7	8	0	
4	:	12	13	10	5	6	7	8	0

2. Mutation and the coalescent.

In this paper we develop methods for inferring quantities of interest concerning the underlying stochastic process that models the evolution of DNA sequences such as those described in the last section. Under such models the tree $(T, \mathbf{n})$ is random and its probability distribution, and that of related quantities, is of interest. In this section we describe the stochastic process we use to model the data and show how the distribution of $(T, \mathbf{n})$ can be computed.

2.1. *The coalescent.* We use Kingman's (1982a, b, c) coalescent to model the ancestral relationships among the n sampled sequences. This model arises in the limit of large population size from a discrete population of N sequences undergoing random mating in each generation. When time is measured in units of N generations, and the limit $N \to \infty$ taken, the times W_i during which the sample has i distinct ancestors have independent exponential distributions with mean

$$E(W_i) = 2/i(i-1), \qquad i = 2, 3, \dots, n.$$

The corresponding model for deterministic fluctuations in population size is given in Section 6.2. At times W_n, $W_n + W_{n-1}, \dots, W_n + \cdots + W_2$, two ancestors are chosen at random to coalesce, corresponding to those sequences having a common ancestor. One way to visualize the coalescent process is as a random bifurcating tree. Reviews and further background may be found in Hudson (1991) and Donnelly and Tavaré (1995).

Mutations are superimposed on this tree according to Poisson processes of rate $\theta/2$, independently in each branch of the tree. The parameter θ is a function of the original population size N and the mutation rate μ per sequence per generation: $\theta = \lim_{N \to \infty} 2N\mu$. In this paper, we assume that whenever a mutation arises on the tree, it gives rise to a new segregating site. The resulting mutation model is known as the *infinitely many sites* model; compare Watterson (1975). In Figure 2, a coalescent tree with mutations is given. The numbers beneath the allele labels give the multiplicity of each allele. In this tree, the pattern of mutations is consistent with the Melanesian data set in Section 1.1. We note that for simplicity the coalescences involving multiple copies of a given allele are not drawn.

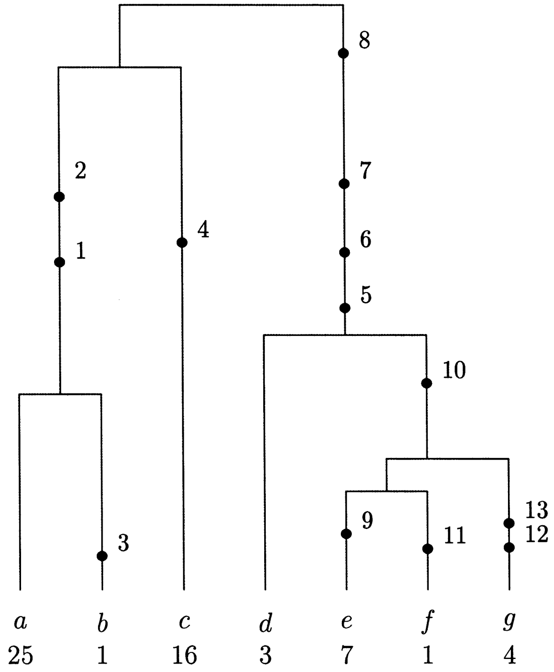


FIG. 2. *Possible Melanesian coalescent tree.*

We have seen that a gene tree T of d genes is constructed by labeling mutations in the coalescent tree of the d genes, then describing each gene by its mutation path $\mathbf{x}$ from current time back to the most recent common ancestor. $T = \{\mathbf{x}_1, \dots, \mathbf{x}_d\}$ is the collection of mutation paths $\mathbf{x}$ from leaves to the root of the tree, with multiplicities of the types $\mathbf{n} = (n_1, \dots, n_d)$. Mutations form vertices in a gene tree.

2.2. *Probabilities of gene trees.* In this section we review how to calculate the probability distribution of trees $(T, \mathbf{n})$ under the infinitely many sites model.

A fundamental recursion for the probability $p^0(T, \mathbf{n})$ of a tree $(T, \mathbf{n})$ is

$$\begin{aligned} p^0(T,\mathbf{n}) = {} & \frac{(n-1)}{(n-1+\theta)} \sum_{k: n_k \geq 2} \frac{(n_k-1)}{n-1} p^0(T, \mathbf{n} - \mathbf{e}_k) \\ & + \frac{\theta}{(n-1+\theta)} \sum_{\substack{k: n_k=1,\, x_{k0}, \text{distinct}, \\ \mathscr{S}\mathbf{x}_k \neq \mathbf{x}_j \,\forall\, j}} \frac{1}{n} p^0(\mathscr{S}_k T, \mathbf{n}) \qquad (2.1) \\ & + \frac{\theta}{(n-1+\theta)} \sum_{\substack{k: n_k=1, \\ x_{k0}\text{ distinct}}} \sum_{j: \mathscr{S}\mathbf{x}_k = \mathbf{x}_j} \frac{(n_j+1)}{n} p^0(\mathscr{R}_k T, \mathscr{R}_k(\mathbf{n} + \mathbf{e}_j)); \end{aligned}$$

see Ethier and Griffiths (1987), Griffiths (1989) and Griffiths and Tavaré (1994b, 1995). In (2.1), $\mathbf{e}_j$ is the jth unit vector, $\mathscr{S}$ is a shift operator which deletes the first coordinate of a path, $\mathscr{S}_k T$ deletes the first coordinate of the kth path of T, $\mathscr{R}_k T$ removes the kth path of T and "x_{k0} distinct" means that $x_{k0} \neq x_{ij}$ for all $(\mathbf{x}_1, \dots, \mathbf{x}_d)$ and $(i, j) \neq (k, 0)$. The boundary condition is $p^0(T_1, \mathbf{e}_1) = 1$. If we define the degree of $(T, \mathbf{n})$ as $\{n - 1 +$ the number of mutations in $T\}$, then the system (2.1) is recursive in the degree of $(T, \mathbf{n})$.

Recursions such as (2.1) can be derived directly from the structure of mutation in the coalescent by looking back to the first event that occurs in the history of the sample—either a mutation or a coalescence event. The first term on the right of (2.1) corresponds to this event being a coalescence [with probability $(n-1)/(n-1+\theta)$], the second and third to the event being a mutation [with probability $\theta/(n-1+\theta)$]. If this event was a mutation, the lineage with this mutation is necessarily a singleton in the sample. In the second term, removing the last mutation from a lineage leaves the lineage as a singleton in the data (e.g., mutation 11 in Figure 2). In the third term the lineage with the mutation removed is identical to another in the sample (e.g., mutation 3 in Figure 2). For each singleton path in T with a distinct first coordinate there is exactly one nonzero term in the second and third summations. A more detailed discussion and derivation appears in Griffiths and Tavaré (1994b, 1995). The result of (2.1) can also be derived from Ethier and Griffiths' (1987) measure-valued diffusion representation of the infinitely many sites model. The notation p^0 conforms with that in Griffiths and Tavaré (1994b). It is the probability of observing a labeled tree. If the tree is unlabeled, then the probability is a combinatorial multiple of p^0.

The program `ptree` implements (2.1) exactly. There are a large number of terms in the recursion, so it only runs effectively for small trees with up to about 15–20 sequences, depending on computer memory and speed. The implementation is recursive, combined with a storage and lookup scheme for probabilities of subtrees [Griffiths (1989)]. It is also possible to find a recurrence relationship similar to (2.1) when mutations are specified to have a particular age ordering. This too is implemented in `ptree`.

Griffiths and Tavaré (1994b) developed a Monte Carlo algorithm based on (2.1) for simulating $p^0(T, \mathbf{n}; \theta)$ as a function of θ. The simulated curve depends on a generating value θ_0 in a similar way to (4.13) below. The algorithm is implemented in the program `genetree` and works effectively for larger sample sizes.

3. Ages of mutations. We may associate with each mutation its age, measured back in time to when it arose. We extend the definition of T to a tree $T_{\mathbf{a}} = \{\mathbf{x}_1^{\mathbf{a}}, \dots, \mathbf{x}_d^{\mathbf{a}}\}$ which contains age information in the mutation paths to the root. Coordinates of the paths $\mathbf{x}^{\mathbf{a}}$ have the form (k, a_k), where a_k is the age of mutation k, measured back from the current time. For convenience the root of the tree is labeled 0, and a_0 is the age of the most recent common ancestor of the tree. The tree, the ages of mutations and the multiplicities of types are all random variables. Recall that in describing a gene tree such as that in Table 2 and Figure 1, we have chosen a particular ordering of equivalent sites. For example, in the Melanesian data sites 5, 6, 7 and 8 are ordered in such a way that site 5 is the youngest and site 8 the oldest. In what follows a given fixed labeling is assumed.

A gene tree with age information is illustrated in Figure 3, with the time axis to the right of the vertical line. The tree is drawn to scale with the expected age of mutations and the expected TMRCA conditional on the gene tree structure $(T, \mathbf{n})$. Numbers to the left of the vertical line are the expected number of ancestors conditional on the gene tree structure.

It is easy to work out the distribution of ages of mutations in a sample of $n = 2$ sequences. If there are two sequences with a and b mutations on them, then the conditional distribution of W, the time to the ancestor, is Gamma with power parameter $a + b + 1$ and scale parameter $1 + \theta$ [Tajima (1983)], and ages are uniformly distributed as order statistics along respective edges of the tree in $(0, W)$. For example, the age of the kth mutation on the edge with a sites has mean $k(1 + a + b)/(1 + \theta)(1 + a)$. For larger sample sizes, a computational approach is required to find these distributions. We develop this method in the next sections.

Time in units of 100,000 years

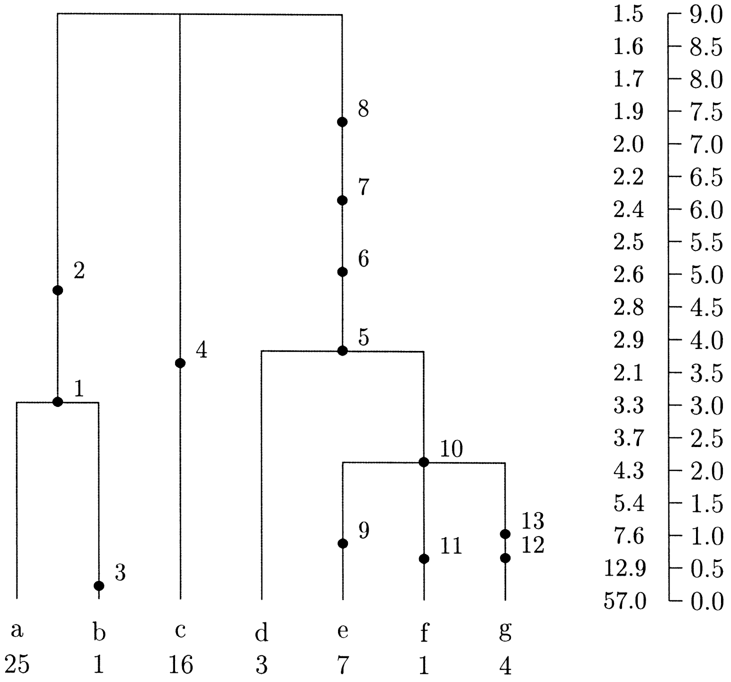


FIG. 3. *Melanesian β-globin tree. Time in units of* 100,000 *years.*

3.1. *Recursions for ages.* The probability of a tree $(T, \mathbf{n})$ can be found from the discrete recursion in (2.1). Keeping track of age information has the effect of changing the recurrence to an integro-recurrence. To see this, let $q^0(T_{\mathbf{A}}, \mathbf{n}) = P(T_{\mathbf{A}}, \mathbf{n}, A_0 \le a_0, \dots, A_s \le a_s)$, where the tree T has s segregating sites. A recursion for $q^0(T_{\mathbf{a}}, \mathbf{n})$ is

$$\begin{aligned} q^0(T_{\mathbf{a}}, \mathbf{n}) = \int_0^\infty \Bigg\{ & \sum_{k:\, n_k \ge 2} \frac{n_k - 1}{n - 1 + \theta} q^0(T_{\mathbf{a}-t}, \mathbf{n} - \mathbf{e}_k) \\ & + \sum_{\substack{k:\, n_k = 1,\ x_{k0}\text{ distinct},\\ \mathscr{S}\mathbf{x}_k \ne \mathbf{x}_j\ \forall\, j}} \frac{\theta}{n(n-1+\theta)} q^0(\mathscr{S}_k T_{\mathbf{a}'-t}, \mathbf{n}) \\ & + \sum_{\substack{k:\, n_k = 1,\\ x_{k0}\text{ distinct}}} \ \sum_{j:\, \mathscr{S}\mathbf{x}_k = \mathbf{x}_j} \frac{\theta(n_j + 1)}{n(n-1+\theta)} \\ & \qquad \times q^0\big(\mathscr{R}_k T_{\mathbf{a}''-t}, \mathscr{R}_k(\mathbf{n} + \mathbf{e}_j)\big) \Bigg\} g(t; n)\, dt, \end{aligned} \tag{3.1}$$

where $g(t; n)$ is the exponential density with rate $n(n + \theta - 1)/2$ and $\mathbf{a}'$ and $\mathbf{a}''$ denote appropriately relabeled ages after removal of the youngest mutation. The argument used to obtain (3.1) is similar to that used for (2.1), with t being the time of the first event back. Implicit in (3.1) is that $q^0(T_{\mathbf{a}-t}, \mathbf{n}) = 0$ if $t > \min_j\{a_j\}$. Ethier and Shiga (1993) study a measure-valued diffusion process with mutation history and age information. Since (2.1) can be derived from Ethier and Griffiths' (1987) measure-valued diffusion, it is likely that (3.1) can also be derived from Ethier and Shiga's process.

It is possible to obtain a recursive system for the conditional expected ages $\{E(A_i \mid (T, \mathbf{n})), i = 0, \dots, s\}$ and solve it in a similar way to `ptree`. Define $\mu_i(T, \mathbf{n}) = E(A_i I\{(T, \mathbf{n})\})$, where $I\{\cdot\}$ denotes the indicator function, with the convention that $\mu_i(T, \mathbf{n}) = 0$ if mutation labeled i does not belong to the tree. Then

$$\begin{aligned} \mu_i(T, \mathbf{n}) = {} & \frac{2}{n(n-1+\theta)} p^0(T, \mathbf{n}) \\ & + \frac{(n-1)}{(n-1+\theta)} \sum_{k:\, n_k \ge 2} \frac{(n_k - 1)}{n - 1} \mu_i(T, \mathbf{n} - \mathbf{e}_k) \\ & + \frac{\theta}{(n-1+\theta)} \sum_{\substack{k:\, n_k = 1,\ x_{k0}\text{ distinct},\\ \mathscr{S}\mathbf{x}_k \ne \mathbf{x}_j\ \forall\, j}} \frac{1}{n} \mu_i(\mathscr{S}_k T, \mathbf{n}) \\ & + \frac{\theta}{(n-1+\theta)} \sum_{\substack{k:\, n_k = 1,\\ x_{k0}\text{ distinct}}} \ \sum_{j:\, \mathscr{S}\mathbf{x}_k = \mathbf{x}_j} \frac{(n_j + 1)}{n} \mu_i\big(\mathscr{R}_k T, \mathscr{R}_k(\mathbf{n} + \mathbf{e}_j)\big). \end{aligned} \tag{3.2}$$

Together (2.1) and (3.2) allow $\mu_i(T, \mathbf{n})$ to be evaluated by recursion, $i = 0, \dots, s$.

Let $W = W_n + \cdots + W_2$ denote the time to the most recent common ancestor (MRCA) of the sample of sequences. Griffiths and Tavaré (1994b) obtain a representation of $P(W \le w, (T, \mathbf{n}))$ as the expected value of a functional of a stochastic process which has a tree state space and moves from an initial state T to a singleton sequence, the root of T. The reason for obtaining this representation is to provide a way to compute the conditional distribution of W given $(T, \mathbf{n})$ by repeated simulation. The aim here is to extend this representation to allow computation of the joint distribution of the ages $\mathbf{A}$, conditional on $(T, \mathbf{n})$.

4. A Monte Carlo method. To develop the Monte Carlo approach, we rescale the coefficients on the right of (3.1) to add to 1 and interpret the scaled coefficients as transition probabilities in a Markov chain. Let $f(T, \mathbf{n})$ be the scale factor, the sum of the coefficients. The rewritten recursion then has the form

$$\begin{aligned} q^0(T_{\mathbf{a}}, \mathbf{n}) &= f(T, \mathbf{n}) \\ &\quad \times \int_0^\infty \Biggl\{ \sum_{k:\, n_k \ge 2} p(T, \mathbf{n} - \mathbf{e}_k \mid T, \mathbf{n}) q^0(T_{\mathbf{a}-t}, \mathbf{n} - \mathbf{e}_k) \\ &\qquad + \sum_k p(T', \mathbf{n}' \mid T, \mathbf{n}) q^0(T'_{\mathbf{a}'-t}, \mathbf{n}') \\ &\qquad + \sum_{k \to j} p(T'', \mathbf{n}'' \mid T, \mathbf{n}) q^0(T''_{\mathbf{a}''-t}, \mathbf{n}'') \Biggr\} g(t; n)\, dt, \end{aligned} \tag{4.1}$$

where

$$\begin{aligned} f(T, \mathbf{n}) &= \frac{\theta}{n(n-1+\theta)} \Biggl(\bigl| k\colon n_k = 1,\ x_{k0} \text{ distinct},\ \mathscr{S}\mathbf{x}_k \ne \mathbf{x}_j\, \forall\, j \bigr| \\ &\qquad + \sum_{\substack{k:\, n_k = 1, \\ x_{k0} \text{ distinct}}} \ \sum_{j:\, \mathscr{S}\mathbf{x}_k = \mathbf{x}_j} (n_j + 1) \Biggr) \\ &\quad + \frac{n-d}{n-1+\theta} \end{aligned} \tag{4.2}$$

and the generic form of the transition probabilities is

$$\begin{aligned} p(T, \mathbf{n} - \mathbf{e}_k \mid T, \mathbf{n}) &= \frac{n_k - 1}{(n-1+\theta) f(T, \mathbf{n})}, \\ p(T', \mathbf{n}' \mid T, \mathbf{n}) &= \frac{\theta}{n(n-1+\theta) f(T, \mathbf{n})}, \\ p(T'', \mathbf{n}'' \mid T, \mathbf{n}) &= \frac{\theta(n_j + 1)}{n(n-1+\theta) f(T, \mathbf{n})}. \end{aligned} \tag{4.3}$$

In (4.1)–(4.3) T' and T'' denote the trees in the last two summations in (3.1). Here $\{p(\cdot \mid T, \mathbf{n})\}$ are transition probabilities in a Markov chain with a tree state space without age information. The chain is imbedded in a Markov process with a tree state space, including age information, with jump distribution $g(t;n)$. The process makes transitions from $(T_{\mathbf{a}}, \mathbf{n})$ to $(T_{\mathbf{a}+t}, \mathbf{n}-\mathbf{e}_k)$, $(T'_{\mathbf{a}'+t}, \mathbf{n}')$ or $(T''_{\mathbf{a}''+t}, \mathbf{n}'')$. The final age of a mutation is determined by the time at which it is removed in the second or third type of transition. There is a single absorbing state at a singleton tree $T_{\mathbf{A}} = \{(0, A_0)\}$, the MRCA of the sample. A functional representation argued directly from (4.1) is

$$\begin{aligned} &P\big(T_{\mathbf{A}}, \mathbf{n}, A_0 \le a_0, \dots, A_s \le a_s\big) \\ &\quad = E_{(T_{\mathbf{0}}, \mathbf{n})}\left[\prod_{\ell=0}^{\tau-1} f\big(T(\ell), \mathbf{n}(\ell)\big) I\big\{A_0(\xi) \le a_0, \dots, A_s(\xi) \le a_s\big\}\right], \end{aligned} \tag{4.4}$$

where the Markov chain passes through states $\{(T(\ell), \mathbf{n}(\ell)),\ \ell = 0, 1, \dots, \tau\}$ and is absorbed at transition τ at time ξ at a singleton tree. The expectation E is in the full Markov process which includes age information, beginning with all ages equal to zero. Letting $a_0, \dots, a_s \to \infty$ in (4.4), we see that

$$p^0(T, \mathbf{n}) = E_{(T_{\mathbf{0}}, \mathbf{n})}\left[\prod_{\ell=0}^{\tau-1} f\big(T(\ell), \mathbf{n}(\ell)\big)\right], \tag{4.5}$$

recovering the representation of Griffiths and Tavaré (1994b).

More generally, (4.4) may be replaced by

$$\begin{aligned} &E\Big(h(A_0, A_1, \dots, A_s) I\big\{(T_{\mathbf{A}}, \mathbf{n})\big\}\Big) \\ &\quad = E_{(T_{\mathbf{0}}, \mathbf{n})}\left[h(A_0, A_1, \dots, A_s) \prod_{\ell=0}^{\tau-1} f\big(T(\ell), \mathbf{n}(\ell)\big)\right]. \end{aligned} \tag{4.6}$$

Writing $F = \prod_{\ell=0}^{\tau-1} f(T(\ell), \mathbf{n}(\ell))$, we then have

$$E\big(h(A_0, A_1, \dots, A_s) | (T, \mathbf{n})\big) = \frac{E_{(T_{\mathbf{0}}, \mathbf{n})}\big(h(A_0, A_1, \dots, A_s) F\big)}{E_{(T_{\mathbf{0}}, \mathbf{n})}(F)}. \tag{4.7}$$

The quantity in (4.7) can be estimated by repeated simulation of the process. Letting $F_j, \mathbf{a}_j = (A_0^j, \dots, A_s^j)$ denote the values of F and $\mathbf{a}$ on the jth of r simulation runs, we can use as an approximation to the right side of (4.7) the ratio

$$\frac{r^{-1} \sum_{j=1}^r h(A_0^j, A_1^j, \dots, A_s^j) F_j}{r^{-1} \sum_{j=1}^r F_j}.$$

Thus an estimate of the tree probability $p^0(T, \mathbf{n})$ is

$$\hat{p}^0(T, \mathbf{n}) = r^{-1} \sum_{j=1}^r F_j, \tag{4.8}$$

and the empirical distribution of the ages, conditional on the tree $(T, \mathbf{n})$, is given by the discrete distribution

$$\big\{(\mathbf{a}_1, p_1), \dots, (\mathbf{a}_r, p_r)\big\}, \tag{4.9}$$

where

$$p_i = \frac{F_i}{\sum_{j=1}^r F_j}, \qquad i = 1, \ldots, r. \tag{4.10}$$

Characteristics of the joint conditional distribution of $\mathbf{A}$ given $(T, \mathbf{n})$, such as mean ages, can be calculated from this empirical distribution. If $\alpha_i^{(j)}$ is the simulated age of the ith site on the jth run, then

$$\hat{E}(A_i \mid T, \mathbf{n}) = \frac{\sum_{j=1}^r \alpha_i^{(j)} F_j}{\sum_{j=1}^r F_j}. \tag{4.11}$$

Note that $\hat{p}^0(T, \mathbf{n})$, being an average of independent identically distributed random variables, is asymptotically normal with mean $p^0(T, \mathbf{n})$. Ratio estimates such as (4.7) have bias of order r^{-1} and, by the strong law of large numbers, are asymptotically unbiased; typically r is chosen to be very large. We note that the discrete distribution (4.9) can also be used to generate (approximately) i.i.d. observations from the required conditional distribution.

The computer program `genetree` implements the algorithm based on (4.4). Output are mean ages and standard deviations of ages of mutations, an empirical TMRCA distribution and the mean number of ancestors at times back in the tree. A gene tree such as in Figure 3 can be drawn to scale in an automated way using the program `treepic`.

4.1. *Importance sampling.* The algorithm using (4.1)–(4.11) can be modified by using importance sampling to simulate a family of empirical age distributions $(\mathbf{a}_1, p_1(\theta)), \ldots, (\mathbf{a}_r, p_r(\theta))$ indexed by θ. Making θ explicit in the notation, (4.1) can be manipulated into the form

$$q_\theta^0(T_{\mathbf{a}}, \mathbf{n}) = \int_0^\infty \sum^* h(T, \mathbf{n}; T^*, \mathbf{n}^*) q_\theta^0(T^*_{\mathbf{a}-t}, \mathbf{n}^*) p_{\theta_0}(T^*, \mathbf{n}^* \mid T, \mathbf{n}) g(t; n)\, dt,$$

where $(T^*, \mathbf{n}^*)$ denotes a state reached from $(T, \mathbf{n})$, and

$$\begin{aligned} h(T, \mathbf{n}; T^*, \mathbf{n}^*) &= f_\theta(T, \mathbf{n}) \frac{p_\theta(T^*, \mathbf{n}^* \mid T, \mathbf{n})}{p_{\theta_0}(T^*, \mathbf{n}^* \mid T, \mathbf{n})} \\ &= \begin{cases} f_{\theta_0}(T, \mathbf{n})(n-1+\theta_0)/(n-1+\theta), \\ \qquad \text{if } (T^*, \mathbf{n}^*) = (T, \mathbf{n} - \mathbf{e}_k), \\ f_{\theta_0}(T, \mathbf{n})\big(\theta(n-1+\theta_0)/\big(\theta_0(n-1+\theta)\big)\big), \\ \qquad \text{if } (T^*, \mathbf{n}^*) = (T', \mathbf{n}'), \text{ or } (T'', \mathbf{n}''). \end{cases} \end{aligned} \tag{4.12}$$

Then

$$\begin{aligned} &P_\theta\big(T_{\mathbf{A}}, \mathbf{n}, A_0 \le a_0, \ldots, A_s \le a_s\big) \\ &\quad = E^{\theta_0}_{(T_{\mathbf{0}}, \mathbf{n})}\Bigg[\prod_{\ell=0}^{\tau-1} h\big(T(\ell), \mathbf{n}(\ell); T(\ell+1), \mathbf{n}(\ell+1)\big) \\ &\qquad\qquad \times I\big\{A_0(\xi) \le a_0, \ldots, A_s(\xi) \le a_s\big\}\Bigg]. \end{aligned} \tag{4.13}$$

Thus a family of age distributions indexed by θ is returned in the simulation run with a single generating θ_0. In practice these distributions will be accurate only in the vicinity of θ_0.

4.2. *Melanesian data set.* To illustrate the algorithms of this section, six realizations of ages in the Melanesian gene tree with their relative likelihoods are shown in Figure 4. While individual trees are not likely to be informative, they do serve to illustrate the variability inherent in each run.

The standard deviation of ages of mutations in the tree in Figure 3 are shown in Table 3. The standard deviations are typically half the means. For illustration, the empirical density of the time to the most recent common ancestor and of the age of mutation 5 are shown in Figure 5.

The maximum likelihood estimate $\hat{\theta} = 2.55$ found by Harding, Fullerton, Griffiths and Clegg (1997) is used. Assuming a generation time of 20 years together with the mutation rate of $v = 1.34 \times 10^{-9}$ per site per year used by

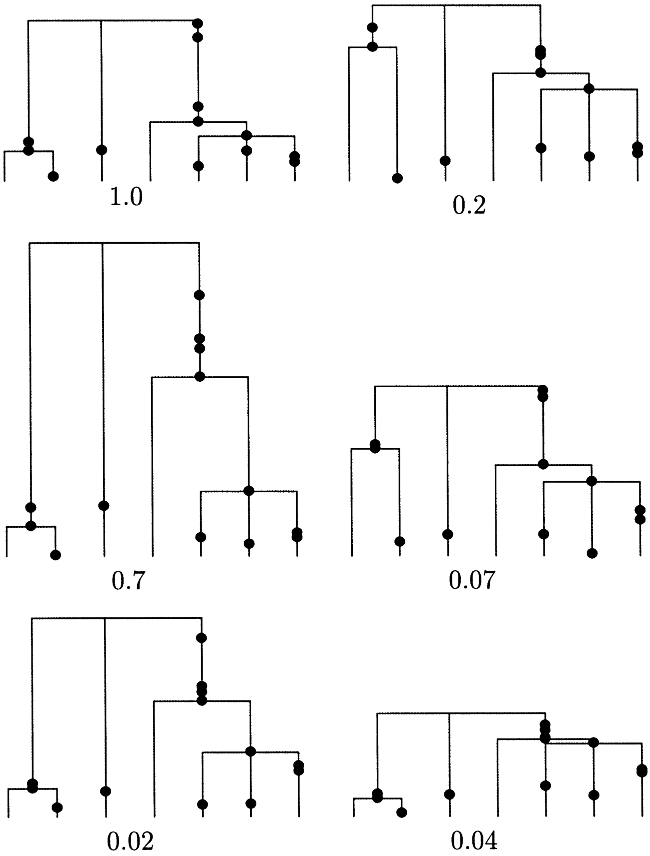


FIG. 4. *Simulated Melanesian trees with relative likelihoods, conditional on observed data.*

TABLE 3
*Ages of mutations**

site	3	11	12	9	13	10	5	1	2	6	4	7	8
sequences	1	1	4	7	4	12	15	26	26	15	16	15	15
mean age | tree	0.2	0.6	0.6	0.8	1.0	2.1	3.8	3.0	4.7	5.0	3.6	6.1	7.3
s.d. age | tree	0.2	0.5	0.4	0.5	0.5	0.9	1.5	1.5	2.0	1.9	1.9	2.2	2.4
mean age (4.14)	0.5	0.5	1.6	2.3	1.6	3.3	3.8	5.3	5.3	3.8	4.0	3.8	3.8

*Time in units of 100,000 years.

Harding, Fullerton, Griffiths, Bond, Cox, Schneider, Moulin and Clegg (1997), this gives an implied effective population size of approximately 10,250 diploid individuals. Therefore 1 unit of coalescent time corresponds to about 410,000 years. The simulations used $r = 500{,}000$. The mean and standard deviation of the age of site 5 are 390,000 and 150,000 years with a 95% interpercentile interval of 170,000–1,030,000 years. (Here, and in the remainder of the paper, such an interval is determined by the 2.5 and 97.5 percentiles of the distribution of interest.) The mean time to the most recent common ancestor (TMRCA) is around 900,000 years. The effective population size for a nuclear gene such as β-globin, being diploid and carried by both sexes, is four times that of mitochondrial DNA or Y-chromosome DNA, implying roughly a four times longer ancestry.

Griffiths and Tavaré (1998) show that the expected age of a mutation observed to be in z sequences in a sample of n $(1 \le z < n)$ is

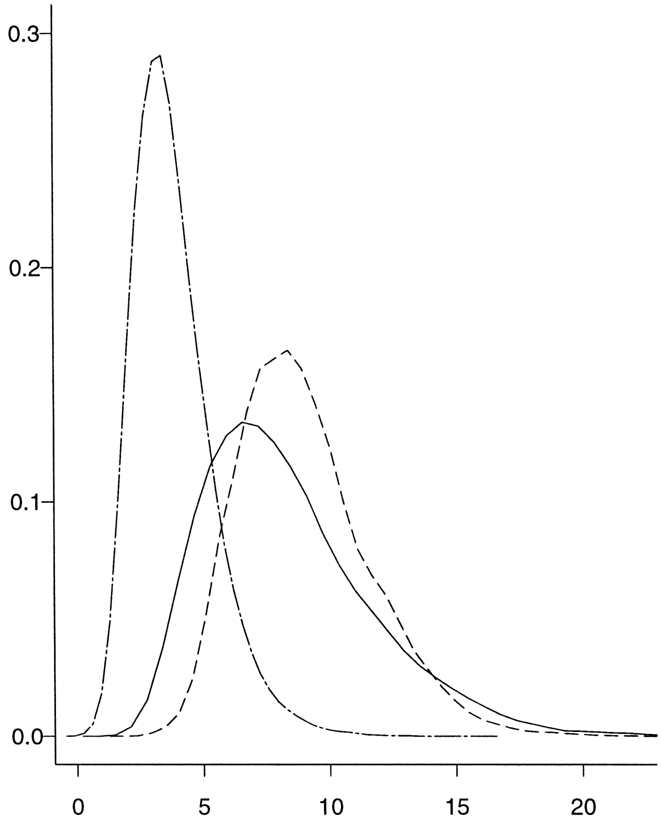
$$\frac{2z}{n-z}\sum_{j=1}^{n-z}\frac{1}{j}\,\frac{(n-z)\cdots(n-z-j+1)}{n\cdots(n-j+1)}. \tag{4.14}$$

It is of interest to compare the expected ages of mutations using (4.14) with the expected ages conditional on the gene tree, which are shown in Table 3. The mean age of a typical one of sites 5, 6, 7 and 8 is 5.6, compared to the estimate of 3.8 from (4.14). Clearly there is much more information in the gene tree; the ages of particular sites are constrained by where they occur in the tree.

5. TMRCA of the sample and population. In a sample of n sequences the probability that the sample and the population share the same most recent common ancestor is $(n-1)/(n+1)$ [Saunders, Tavaré and Watterson (1984)]. However the conditional probability given a gene tree $(T, \mathbf{n})$ may be quite different, depending on whether the tree suggests a short time to its ancestor or not. It is of interest to develop an algorithm to compute the joint distribution of the TMRCA in both sample and population, conditional on $(T, \mathbf{n})$. Let Y_0 be the TMRCA in a sample of n sequences and Y_1 the TMRCA in a larger sample of $m+n$ sequences representing the population and containing the n.

Let $T_{\mathbf{Y}}$ denote a tree with information additional to T about $\mathbf{Y} = (Y_0, Y_1)$ and

$$q(T_{\mathbf{y}}, \mathbf{n}; m) = P\big(T_{\mathbf{Y}}, \mathbf{n}, Y_0 \le y_0, Y_1 \le y_1\big).$$



Densities beginning from left are
age of site 5 | gene tree as data,
TMRCA of sample | number of segregating sites as data,
TMRCA of sample | gene tree as data.

FIG. 5. *Age of site* 5 *and TMRCA densities. Time in units of* 100,000 *years.*

A recursive equation is

$$
\begin{aligned}
q(T_{\mathbf{y}}, \mathbf{n}; m) = \int_0^\infty \biggl\{ & \frac{n(n-1)}{r(n, m, \theta)} \sum_{n_k > 1} \frac{n_k - 1}{n - 1} q(T_{\mathbf{y}-t}, \mathbf{n} - \mathbf{e}_k; m) \\
& + \frac{(m+n)(m+n-1) - n(n-1)}{r(n, m, \theta)} q(T_{\mathbf{y}-t}, \mathbf{n}; m-1) \\
& + \frac{\theta}{r(n, m, \theta)} \sum_k q(T'_{\mathbf{y}-t}, \mathbf{n}'; m) \\
& + \frac{\theta}{r(n, m, \theta)} \sum_{k \to j} q(T''_{\mathbf{y}-t}, \mathbf{n}''; m) \biggr\} g(t; n, m)\, dt,
\end{aligned}
\tag{5.1}
$$

where $r(n, m, \theta) = (m+n)(m+n-1) + n\theta$, and $g(t; n, m)$ is the exponential density with rate $r(n, m, \theta)/2$.

Events that may occur when there is a configuration of (n, m) sequences are coalescence in the n subgroup with probability $n(n-1)/r(n, m, \theta)$; coalescence in the $m+n$ sequences, but not in the n subgroup with probability $\big((m+n)(m+n-1) - n(n-1)\big)/r(n, m, \theta)$; and mutation in the n subgroup with probability $n\theta/r(n, m, \theta)$. If $n = 1$, (5.1) is just an obvious recurrence for Y_1 based on convolution of waiting time distributions in states

$$P(Y_1 \le y_1; m) = \int_0^\infty P\big(Y_1 \le y_1 - t; m-1\big) g(t; 1, m)\, dt.$$

Construct a Markov process with a state space $(T_{\mathbf{y}}, \mathbf{n}, m)$ by rescaling the coefficients in (5.1). Then a representation is

$$\begin{aligned} q(T_{\mathbf{y}}, \mathbf{n}; m) = EE_{(T_0, \mathbf{n}, m)}\Bigg[&\prod_{\ell=0}^{\tau-1} f\big(T(\ell), \mathbf{n}(\ell), m(\ell)\big) \\ &\times I\big\{Y_0(\xi) \le y_0, Y_1 \le y_1\big\} \mid m(\xi), Y_1(\xi)\Bigg], \end{aligned} \tag{5.2}$$

where the imbedded Markov chain passes through states $\{\big(T(\ell), \mathbf{n}(\ell), m(\ell)\big), \ell = 0, \dots, \tau\}$ and τ is the step at the hitting time when first $n = 1$. The outside expectation is taken over the tail of the waiting time from when first $n = 1$ until $m = 0$. The simulation rule then generates replicates (y_0, y_1, p) similarly to (4.5) leading to an empirical distribution of Y_0, Y_1 conditional on $(T, \mathbf{n})$. The functional values are $F = \prod_{\ell=0}^{\tau-1} f(T(\ell), \mathbf{n}(\ell), m(\ell))$, y_0 is the waiting time until the ancestor of the subgroup of n, and y_1 is the total time until the ancestor of the $m+n$. The algorithm is implemented in a program popsim.

5.1. *A Y chromosome data set.* As an illustrative example, we use data having three segregating sites arising in five Y-chromosome sequences given in Whitfield, Sulston and Goodfellow (1995). Tavaré, Balding, Griffiths and Donnelly (1997) also discuss these data. The major point is that the TMRCA of the population and the sample can be quite different for some parameter values. Using $\theta = 3.52$ (calculated from 15,680 bases at a rate of 1.123×10^{-9} per base per year and an effective population size of $N = 5{,}000$), the expected number of segregating sites is 7.3, so a short ancestry of the sample is suggested. The tree with expected ages and the TMRCA of the population as the height of the box is shown in Figure 6.

If θ is really too large, because of the mutation rate or the effective population size being incorrect, then a short ancestry is not suggested. The maximum likelihood estimate given the gene tree as data is $\hat{\theta} = 1.65$. The likelihood curve is shown in Figure 7. This is an exact, rather than simulated, curve

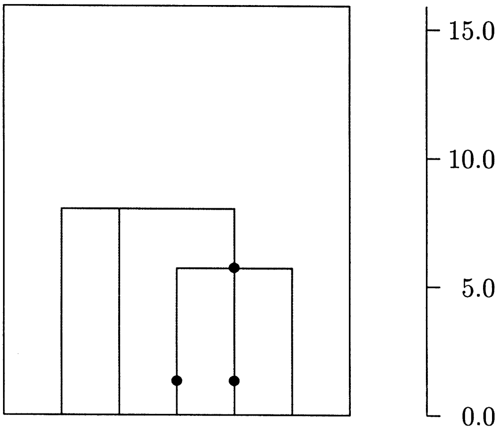


FIG. 6. *Whitfield's Y-chromosome tree. Time in units of* 10,000 *years.*

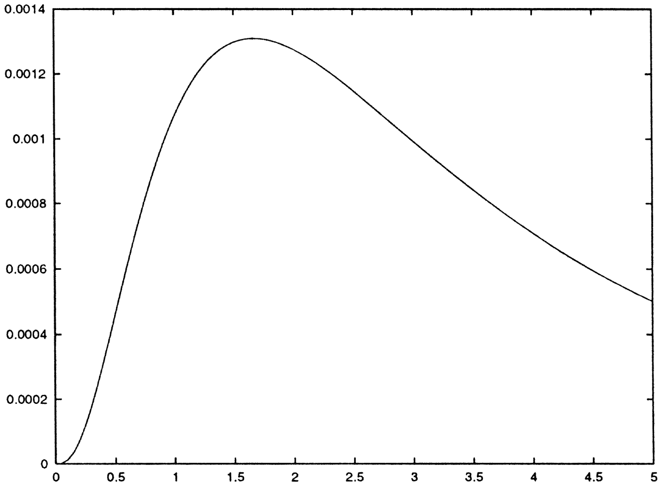


FIG. 7. *Likelihood curve for Whitfield data.*

TABLE 4
Ages of mutations in Whitfield's tree

θ	**1.0**	**1.65**	**3.52**	**5.0**
Age 1	2.1	1.8	1.3	1.0
Age 2	2.1	1.8	1.3	1.0
Age 3	10.8	8.7	5.7	4.4
TMRCA (S)	16.0	12.8	8.1	6.3
TMRCA (P)	18.6	16.4	15.8	16.5
Prob (S=P)	0.77	0.66	0.37	0.22
Likelihood	0.8	1.0	0.6	0.4

computed from the recursion (2.1) using ptree. Table 4 shows characteristics of the tree for various values of θ. S and P denote sample and population MRCA's. Units are in 10,000 years, and the likelihood is relative to the maximum when $\theta = 1.65$. Intuitively there cannot be a large amount of information in a small tree like this.

Whitfield, Sulston and Goodfellow (1995) actually use the particular topology in Figure 8 for their tree. The other possibility is that one of the younger mutations occurs on the rightmost edge. In a small tree it is possible, though tedious, to enumerate all possible sequences of mutation and coalescence events to the ancestor, with their probability and times of occurrence. It can be shown that the topology in Figure 8 has a very high probability of being the correct one by comparing the likelihood curve with a likelihood curve produced by ptree for the corresponding gene tree with no assumption as to the coalescence order. The TMRCA density can also be found exactly by a combinatorial argument. This was done for the particular topology. There are essentially six cases to consider, depending on whether the coalescence in the left of the

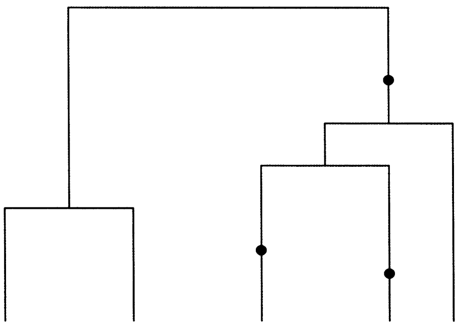


FIG. 8. *Tree topology.*

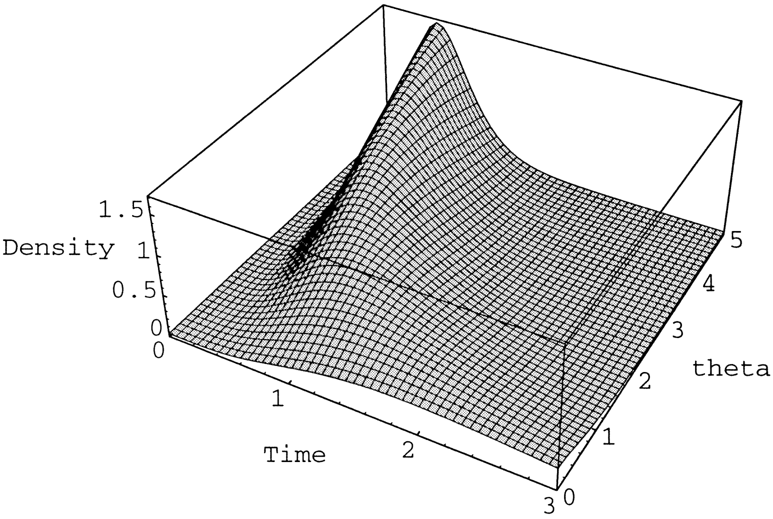


FIG. 9. *Family of TMRCA densities as θ varies.*

diagram in Figure 8 is below the last two mutations, between the last two mutations, above the last two mutations and before the coalescence of their lines, above the coalescence of their lines and below the topmost mutations, or above the topmost mutation.

It is possible to find the joint distribution of the TMRCA and the tree. Considering the six possible trees leads to the distribution of the TMRCA given the tree topology in Figure 8. A family of TMRCA densities, with parameter θ, is shown as a surface in Figure 9. The density is quite flat for small θ, but concentrated around the mode for larger values, at small TMRCA values.

6. Further applications. In this section we collect together some results that show how the basic approach can be applied to several related problems.

6.1. *Infinitely many alleles model.* One useful summary statistic of the set of sequences is the counts $\mathbf{n}$ of allele frequencies (recall that many distinct trees can have the same allele counts). These allele frequencies evolve according to the *infinitely many alleles* model; compare Ewens (1972). The age of an allele is defined as the age of the youngest mutation in that allele. It is simple to derive a recursion that allows us to compute the joint age distribution of *alleles* in a sample, conditional on their frequencies $n_1, \dots, n_d$. Let $A_1, \dots, A_d$ be the ages of alleles $1, \dots, d$, A_0 the TMRCA of $n_1, \dots, n_d$ and define

$$r_{\mathbf{a}}(\mathbf{n}) = Eh(A_0, \dots, A_d)I\{\mathbf{n}\}.$$

Then, for $n > 1$,

$$
\begin{aligned}
n(n-1+\theta)r_{\mathbf{a}}(\mathbf{n}) \\
&= \int_0^\infty \Bigg\{ \sum_{k:\, n_k \ge 2} n(n_k - 1) r_{\mathbf{a}-t}(\mathbf{n} - \mathbf{e}_k) \\
&\qquad + \sum_{k:\, n_k = 1} \; \sum_{j:\, j \ne k,\, n_j > 0} \theta(n_j + 1) r_{\mathbf{a}_k - t}(\mathbf{n} - \mathbf{e}_k + \mathbf{e}_j) \\
&\qquad\qquad + \sum_{k:\, n_k = 1} \theta r_{\mathbf{a}_k - t}(\mathbf{n}) \Bigg\} g(t; n)\, dt,
\end{aligned} \tag{6.1}
$$

where $\mathbf{a}_k = (a_0,\ a_1, \ldots, a_{k-1},\ \infty,\ a_{k+1}, \ldots, a_d)$. The recursion follows lines back in time, replacing an age a_k by ∞ once it has been determined. It can happen that an allele has an age greater than the TMRCA if that type is the same as the type of the MRCA; if $n = 1$, with one type A_1 and $a_0 < a_1$, one needs to consider the time to the mutation after the MRCA, which is exponential with rate $\theta/2$. The recursion is similar to that obtained for the Ewens' (1972) labeled sampling formula and the marginal distribution of the allele configuration without ages is

$$\frac{n!\theta^d}{n_1 \cdots n_d \cdot \theta(\theta+1)\cdots(\theta+n-1)}.$$

A Monte Carlo method similar in construction to that in Section 4 can then be used to find the age distribution of a sample of n; see Griffiths and Tavaré (1994c) for related material. The algorithm used in the construction can be changed by scaling the terms in the recursion. The implementation here seems to have a lower simulation variance when based on $b(\mathbf{n}) = r(\mathbf{n})\theta_{(n)}/n!\theta^d$. The mean and standard deviation of the ages of the alleles in the Melanesian data set are shown in Table 5, these values being estimated from three million runs of the simulation algorithm. The standard deviation of the younger ages are large relative to their mean.

6.2. *Variable population size.* The algorithm in Section 4 for computing age distributions can be extended to a population with deterministic variable

TABLE 5
*Ages of Melanesian alleles**

frequency	25	16	7	4	3	1	1	TMRCA
mean age	4.4	3.4	2.0	1.4	1.1	0.5	0.5	8.0
s.d. age	3.3	3.0	2.4	2.0	1.8	1.2	1.2	4.3

*Time in units of 100,000 years.

population size. We assume that in coalescent units, the population size a time t back from the present time satisfies $N(t) = N(0)\nu(t)$, $t \geq 0$. For example, exponential growth of a population (forward in time) at rate ρ can be modeled by taking $\nu(t) = \exp(-\rho t)$, $t > 0$.

The analogous representation is to consider a Markov process with a jump density from t to an event at $s > t$ when there are n ancestors of

$$\binom{n}{2}\lambda(s)\exp\left(-\binom{n}{2}\int_t^s \lambda(u)\,du\right), \qquad s > t,$$

where $\lambda(t) = 1/\nu(t)$. The transition probabilities are derived similarly to before; however if a transition is made at time s, then the coalescence coefficient in (3.1) becomes $(n_k - 1)\lambda(s)/((n-1)\lambda(s) + \theta)$, so the type of a transition made now has a probability depending on the time at which it occurs. The functional constructed by scaling coefficients depends on s also, and in (4.5) $F = \prod_{\ell=0}^{\tau-1} f(T(\ell), \mathbf{n}(\ell), s(\ell))$, where $\{s(\ell), \ell = 0, \dots, \tau-1\}$ are the event times. More detail about variable population size appears in Slatkin and Hudson (1991) and Griffiths and Tavaré (1994c).

6.3. *Varying mutation rates.* In this section we show that when there are different mutation rates in different regions of the sequence, the joint distribution of ages depends only on the total mutation rate. We incorporate varying mutation rates as follows: suppose there are k different regions in the sequence with mutation rates $\theta_1, \dots, \theta_k$, and overall mutation rate $\theta = \theta_1 + \cdots + \theta_k$. We interpret this to mean that, given a mutation has occurred, it is of type i with probability θ_i/θ. The extension of (3.1) is to replace $\theta/(n-1+\theta)$ by $\theta_i/(n-1+\theta)$ for mutations of type i. Then $q^0(T_{\mathbf{a}}, \mathbf{n})$ is the probability when sites are arranged into these k types. Let $q^0_\theta(T_{\mathbf{a}}, \mathbf{n})$ denote the probability when all sites have an equal rate. If there are $s_1, \dots, s_k$ sites of the k types, then

$$q^0(T_{\mathbf{a}}, \mathbf{n}) = \left(\prod_1^k \left(\frac{\theta_i}{\theta}\right)^{s_i}\right) q^0_\theta(T_{\mathbf{a}}, \mathbf{n}),$$

and it follows that the conditional distribution of ages given the tree $(T, \mathbf{n})$ depends on $\theta_1, \dots, \theta_k$ only through θ. An extension of the argument implies that if rates are variable, and the types of the sites are unspecified, then the conditional age distribution still only depends on the total rate θ. The same result holds if the population size varies deterministically.

6.4. *The number of ancestors time t ago.* Let $Z_n, Z_{n-1}, \dots, Z_2$ be the coalescence times in a sample of n sequences. Our interest is in the conditional distribution of these times, given the gene tree $(T, \mathbf{n})$. Of course the unconditional times are distributed as points in a death process of rate $\mu_k = \binom{k}{2}$, but the conditional distribution is much more complex. A representation analogous to (4.4) holds,

$$\begin{aligned} &P(T_{\mathbf{A}}, \mathbf{n}, Z_n \le z_n, \dots, Z_2 \le z_2) \\ &\quad = E_{(T,\mathbf{n})}\left[\prod_{\ell=0}^{\tau-1} f(T(\ell), \mathbf{n}(\ell)) I\{Z_n' \le z_n, \dots, Z_2' \le z_2\}\right], \end{aligned} \tag{6.2}$$

where $Z_n', \dots, Z_2'$ are the "coalescence" times in the Markov process.

The conditional distribution of $\{A_n(t), t \ge 0\}$, the number of ancestors of the sample time t ago, is easily related to the conditional distribution of coalescence times. $E(A_n(t) \mid T, \mathbf{n})$ is shown in Figure 3 for various values of t for the β-globin data. Since the tree is drawn with mutations separating lines, the number of ancestors at a particular time is less than or equal to the number of edges across the tree.

6.5. *Unrooted trees.* When the ancestral labeling of the sites is unknown, the data may be represented by an unrooted tree [cf. Griffiths and Tavaré (1995)]. The unrooted tree of the Melanesian data appears in Figure 10. Circles

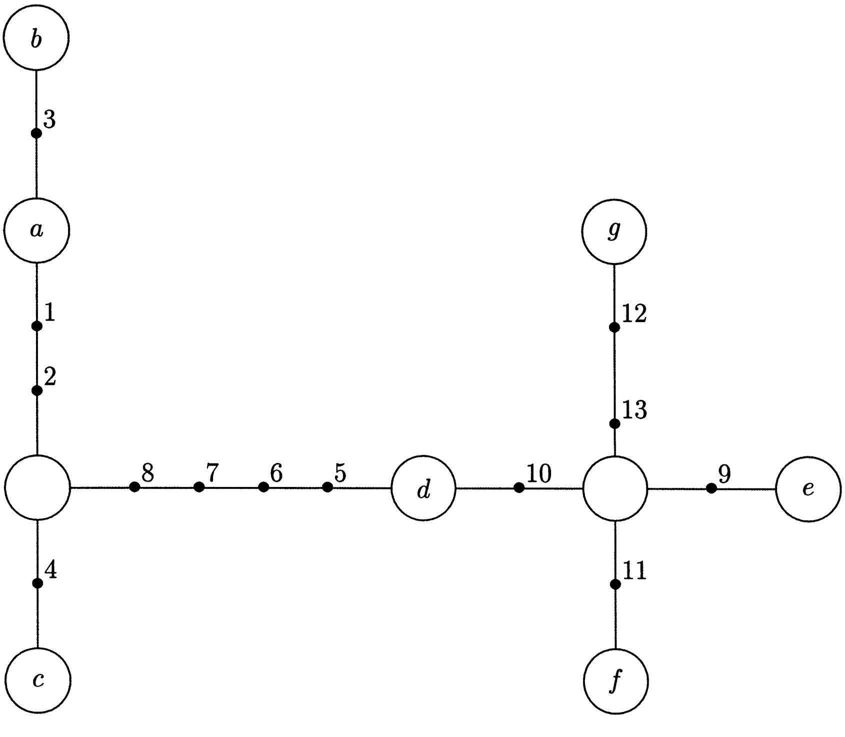


FIG. 10. *Unrooted Melanesian tree.*

without labels are inferred individuals in the genealogy. Sites such as 5, 6, 7, 8 are now ordered between individuals for convenience, rather than age-ordered as in rooted trees. There are $s+1$ rooted trees $T_0, \dots, T_s$ corresponding to an unrooted tree with s mutations, and the likelihood of the unrooted tree Q, say, is

$$p^0(Q, \mathbf{n}) = \sum_{j=0}^{s} p^0(T_j, \mathbf{n}). \tag{6.3}$$

The conditional distribution of ages of sites is

$$\frac{\sum_{j=0}^{s} P_j(A_0 \le a_0, \dots, A_s \le s \mid T_j, \mathbf{n}) p^0(T_j, \mathbf{n})}{p^0(Q, \mathbf{n})}. \tag{6.4}$$

An empirical distribution of ages conditional on $(Q,\ \mathbf{n})$ is a mixture weighted by the simulated functionals. For example, if $\alpha_{ji}^{(k)}$ is the simulated age in the tree T_j of site i on the kth of r_j replicates, with functional F_{jk}, then

$$\hat{E}(A_i \mid Q, \mathbf{n}) = \frac{\sum_{j=0}^{s} \sum_{k=1}^{r_j} \alpha_{ji}^{(k)} F_{jk}}{\sum_{j=0}^{s} \sum_{k=1}^{r_j} F_{jk}}. \tag{6.5}$$

7. Discussion.

In this paper we have developed a technique for approximating the distribution of the ages of mutations in a gene tree under a particular model of DNA sequence evolution. It is possible to extend this model to allow for recombination along the sequences [Griffiths and Marjoram (1996)]. In this case, there can be multiple ancestors of a given sequence. The computational approach used here may also be used to study similar problems when the population of interest is subdivided, and migration is allowed between "islands" [Bahlo and Griffiths (2000)]. Software is available at the mathematical genetics web site at `http://www.stats.ox.ac.uk/`. Finally, while we have focused on the ages of mutations, the same approach may be exploited to study conditional properties of other functionals of the coalescent process, given sample data.

Acknowledgment. We thank Rosalind Harding for helpful comments.

REFERENCES

Bahlo, M. and Griffiths, R. C. (2000). Gene trees in subdivided populations. *Theoret. Population Biol.* To appear.

Donnelly, P. and Tavaré, S. (1995). Coalescents and genealogical structure under neutrality. *Ann. Rev. Genet.* **29** 401–421.

Ethier, S. and Griffiths, R. C. (1987). The infinitely-many-sites-model as a measure valued diffusion. *Ann. Probab.* **15** 515–545.

Ethier, S. and Shiga, T. (1994). Neutral allelic genealogy. In *Measure-valued Processes, Stochastic PDEs, and Interacting Systems* 87–97. Amer. Math. Soc., Providence, RI.

EWENS, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoret. Population Biol.* **3** 87–112.

FORSYTHE, G. E. and LEIBLER, R. A. (1950). Matrix inversion by the Monte Carlo method. *Math. Comp.* **26** 127–129.

FU, Y.-X. and LI, W.-H. (1997). Estimating the age of the common ancestor of a sample of DNA sequences. *Mol. Biol. Evol.* **14** 195–199.

FULLERTON, S. M., HARDING, R. M., BOYCE, A. J. and CLEGG, J. B. (1994). Molecular and population genetic analysis of allelic sequence diversity at the human β-globin locus. *Proc. Nat. Acad. Sci. U.S.A.* **91** 1805–1809.

GRIFFITHS, R. C. (1989). Genealogical-tree probabilities in the infinitely-many-site model. *J. Math. Biol.* **27** 667–680.

GRIFFITHS, R. C. and MARJORAM, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *J. Comp. Biol.* **3** 479–502.

GRIFFITHS, R. C. and TAVARÉ, S. (1994a). Simulating probability distributions in the coalescent. *Theoret. Population Biol.* **46** 131–159.

GRIFFITHS, R. C. and TAVARÉ, S. (1994b). Ancestral inference in population genetics. *Statist. Sci.* **9** 307–319.

GRIFFITHS, R. C. and TAVARÉ, S. (1994c). Sampling theory for neutral alleles in a varying environment. *Proc. Roy. Soc. London Ser. B* **344** 403–410.

GRIFFITHS, R. C. and TAVARÉ, S. (1995). Unrooted genealogical tree probabilities in the infinitely many-sites model. *Math. Biosci.* **127** 77–98.

GRIFFITHS, R. C. and TAVARÉ, S. (1998). The age of a mutation in a general coalescent tree. *Stoch. Models* **14** 273–295.

GUSFIELD, D. (1991). Efficient algorithms for inferring evolutionary trees. *Networks* **21** 19–28.

HALTON, J. H. (1970). A retrospective and prospective study of the Monte Carlo method. *SIAM Rev.* **12** 1–63.

HARDING, R. M., FULLERTON, S. M., GRIFFITHS, R. C., BOND, J., COX, M. J., SCHNEIDER, J. A., MOULIN, D. and CLEGG, J. B. (1997). Archaic African and Asian lineages in the genetic ancestry of modern humans. *Amer. J. Hum. Genet.* **60** 772–789.

HARDING, R. M., FULLERTON, S. M., GRIFFITHS, R. C. and CLEGG, J. B. (1997). A gene tree for β-globin sequences from Melanesia. *J. Mol. Evol.* **44** S133–S138.

HUDSON, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoret. Population Biol.* **23** 183–201.

HUDSON, R. R. (1991). Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology* (D. Futuyma and J. Antonovics, eds.) **7** 1–44. Oxford Univ. Press.

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). On the genealogy of large populations. *J. Appl. Probab.* **19A** 27–43.

KINGMAN, J. F. C. (1982b). The coalescent. *Stochastic Process. Appl.* **13** 235–248.

KINGMAN, J. F. C. (1982c). Exchangeability and the evolution of large populations. In *Exchangeability in Probability and Statistics* (G. Koch and F. Spizzichino, eds.) 97–112. North-Holland, Amsterdam.

KUHNER, M. K., YAMATO, J. and FELSENSTEIN, J. (1995). Estimating effective population size and mutation rate from sequence data using Metropolis–Hastings sampling. *Genetics* **140** 1421–1430.

SAUNDERS, I. W., TAVARÉ, S. and WATTERSON, G. A. (1984). On the genealogy of nested subsamples from a haploid population. *Adv. in Appl. Probab.* **16** 471–491.

SLATKIN, M. and HUDSON, R. R. (1991). Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129** 555–562.

TAJIMA, F. (1983). Evolutionary relationships of DNA sequences in finite populations. *Genetics* **105** 437–460.

TAVARÉ, S., BALDING, D., GRIFFITHS R. C. and DONNELLY, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics* **145** 505–518.

WATTERSON, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoret. Population Biol.* **7** 256–276.

WATTERSON, G. A. (1996). Motoo Kimura's use of diffusion theory in population genetics. *Theoret. Population Biol.* **49** 154–188.

WHITFIELD, L. S., SULSTON, J. E. AND GOODFELLOW, P. N. (1995). Sequence variation of the human Y chromosome. *Nature* **378** 379–380.

DEPARTMENT OF STATISTICS
UNIVERSITY OF OXFORD
1 SOUTH PARKS ROAD
OXFORD OX1 3TG
ENGLAND

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF SOUTHERN CALIFORNIA
LOS ANGELES, CALIFORNIA 90089-1113