The genealogy of a neutral mutation

R.C. Griffiths [1] **and Simon Tavaré** [2]

University of Oxford and University of Southern California

In *Highly Structured Stochastic Systems.*
Eds. Green P, Hjort N and Richardson S, pp. 393-412, 2003.
Oxford Statistical Science Series, Vol. 27
Oxford University Press.

Running head. Coalescent trees

[1] Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK. [2] Departments of Biological Sciences, Mathematics, and Preventive Medicine, University of Southern California, Los Angeles, CA 90089-1340, USA. ST was supported in part by NSF grant DBI95-04393 and NIH grant GM58897.

1 Introduction

Technological advances in molecular biology have made it possible to survey DNA sequence variation in natural populations. These data include restriction fragment length polymorphisms, microsatellite repeats, single nucleotide polymorphisms and complete DNA sequences of particular loci; cf. Chapter 2 of Hartl and Jones (2001). The analysis and interpretation of the patterns of variation seen in such data is complicated by the fact that the sampled chromosomes share a common ancestry, thus making the data highly dependent, for example a mutation appearing in an ancestor is carried by all descendants of that ancestor. To make matters worse, the nature of this common ancestry is not known precisely and therefore needs to be modeled. Since the pioneering work of Kingman (1982), Tajima (1983) and Hudson (1983), population geneticists have used *coalescent models* as a stochastic description of the ancestry of a sample of chromosomes, and there is now an extensive literature on theory and inference for such models. See for example Hudson (1990), Donnelly and Tavaré (1995), Nordborg (2001) and Stephens (2001).

One aspect of the theory that has received a lot of attention concerns the age of mutations. For historical overviews, see Watterson (1996) and Slatkin and Rannala (2000). The ages of mutations are of interest to human geneticists trying to map disease mutations using linkage disequilibrium methods. More on this aspect can be found in Nordborg and Tavaré (2002). Kimura and Ohta (1973) studied the age of a mutation known to have a certain frequency in a population using diffusion methods, and these results were recast in the coalescent framework by Griffiths and Tavaré (1998, 1999), Wiuf and Donnelly (1999), Stephens (2000) and Wiuf (2002). These authors also studied the age of a mutation observed to have a given frequency in a sample of chromosomes. Slatkin and Rannala (1997) addressed the problem of estimating the age of a mutation when given not just its frequency in a sample but also an estimate of the number of mutations occurring in a completely linked region of DNA. Their approach models the age of the mutation as a parameter, and so differs from the coalescent-based approach in which the age is an unobservable random variable; the natural quantity to report is the conditional distribution of the age, given the available data.

In this paper we study aspects of the age of a mutation from the coalescent perspective. After a brief introduction to the general coalescent tree (in which coalescence times can have any given continuous distribution), we describe in Section 2 an urn model that can be used to study the combinatorics of coalescent trees having a mutation of a given frequency. Section 3 describes the infinitely-many-sites model of mutation. Sections 4–6 give various properties of the age of a mutation having a given frequency in a sample. Section 7 discusses simulation algorithms and illustrates them by studying Slatkin and Rannala's (1997) problem as well as the distribution of Tajima's D in a subtree. Section 8 studies the coalescent subtree of that part of the population known to carry a given mutation, and Section 9 exploits these results to study the age of a mutation in a sample taken at random from chromosomes carrying a given mutation (the disease registry model), or together with data from chromosomes not carrying that mutation (the case-control model).

Figure 1 illustrates a coalescent tree of a sample of ten genes with mutations occurring in the ancestry of the sample. Our interest focuses on the descendants of one single given mutation. For example, the mutation on the far right of the tree subtends five descendants.

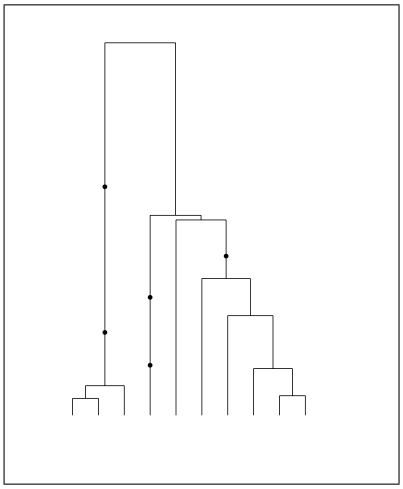

Figure 1. Coalescent tree with mutations.

1.1 Coalescent trees

In the absence of recombination, the ancestry of a sample of n genes from a large population can be described by a coalescent tree (Kingman 1982). Let $T_n, T_{n-1}, \ldots, T_2$ denote the lengths of time for which the sample has $n, n-1, \ldots, 2$ distinct ancestors back in time to its most recent common ancestor (MRCA). In the usual coalescent process, corresponding to a constant population size, the number of distinct ancestors $A_n(t)$ time t ago is a time-homogeneous death process with death rate μ_j from state j given by

$$\mu_j = \binom{j}{2}, \ j = n, n-1, \ldots, 2. \tag{1.1}$$

The times T_j are therefore distributed as independent exponential random variables with means μ_j^{-1}. This paper discusses results about the structure of coalescent trees under a general joint distribution for the times $(T_n, \ldots, T_2)$. As in Griffiths and Tavaré (1998) we assume that:

(A1) $T_n, \ldots, T_2$ are continuous random variables.

(A2) The ancestral tree is binary, and such that when there are k ancestral lines each pair has probability $\binom{k}{2}^{-1}$ of being the next pair to coalesce.

In this paper we use coalescent time units. In the population genetics setting, these may be converted to generations by letting one coalescent time unit correspond to $2N$ generations, appropriate for the coalescent approximation of a population of size N diploid individuals at the time of sampling.

1.2 Variable population size

The motivation for considering a general tree comes from a coalescent model with variable population size, and from other models such as the birth-and-death process generated forward in time.

The coalescent for a population undergoing deterministic population size fluctuations is described in Slatkin and Hudson (1991) and Griffiths and Tavaré (1994). For the Wright-Fisher model, let $\lambda(t)$ denote the ratio of the population size time t ago and the population size at the time of sampling. Let $\{A_n(t), t \geq 0\}$ be the death process described by (1.1), and let $\{A_n^\lambda(t), t \geq 0\}$ denote the corresponding process in the variable population size case. Then

$$A_n^\lambda(t) = A_n\left(\int_0^t \lambda(u)^{-1} du\right), \quad t \geq 0. \tag{1.2}$$

A formula for the distribution of $A_n(t)$ in the constant population size case is well known (Tavaré 1984, Griffiths 1980), and it follows from (1.2) that

$$\mathbb{P}(A_n^\lambda(t) = k) = \sum_{j=k}^{n} \rho_j(t) \frac{(-1)^{j-k}(2j-1)k_{(j-1)}n_{[j]}}{k!(j-k)!n_{(j)}}, k = 1, \ldots, n, \tag{1.3}$$

where $\rho_j(t) = \exp\left(-\binom{j}{2}\int_0^t \lambda(u)^{-1} du\right)$ and

$$a_{(0)} = 1, \qquad a_{(j)} = a(a+1)\cdots(a+j-1),\ j \geq 1; \tag{1.4}$$

$$a_{[0]} = 1, \qquad a_{[j]} = a(a-1)\cdots(a-j+1),\ j \geq 1. \tag{1.5}$$

The mean waiting time in state j is given by

$$\mathbb{E}(T_j) = \int_0^\infty \mathbb{P}(A_n^\lambda(t) = j)dt\,, j = 2, \ldots, n.$$

2 The relationship between coalescent trees and urn models

In the classical Pólya urn model, an urn contains balls of k distinct colours. At discrete time instants a ball is chosen at random from the urn and replaced with an additional ball of the same colour. If the initial configuration of colours is $\boldsymbol{c} = (c_1, \ldots, c_k)$, then after r draws the probability of a configuration $\boldsymbol{r} + \boldsymbol{c} = (r_1 + c_1, \ldots, r_k + c_k)$ is

$$\mathbb{P}(\boldsymbol{r} + \boldsymbol{c}) \quad = \quad \binom{r}{\boldsymbol{r}} \frac{(c_1)_{(r_1)} \cdots (c_k)_{(r_k)}}{(c_1 + \cdots + c_k)_{(r)}}$$

$$= \binom{c_1+\cdots+c_k+r-1}{r}^{-1}\prod_{j=1}^{k}\binom{c_j+r_j-1}{r_j}, \tag{2.1}$$

where $r = r_1 + \cdots + r_k$ and $c_{(r)}$ is defined by (1.5); cf. Feller (1968, Chapter V). This distribution is the Multinomial-Dirichlet distribution:

$$\mathbb{P}(\boldsymbol{r}+\boldsymbol{c}) = \int \binom{r}{\boldsymbol{r}} x_1^{r_1}\cdots x_k^{r_k} f(x_1,\dots,x_k)d\boldsymbol{x}, \tag{2.2}$$

where $f(x_1,\dots,x_k)$ is a Dirichlet$(c_1,\dots,c_k)$ density defined by

$$f(x_1,\dots,x_k) = \frac{\Gamma(c_1+\cdots+c_k)}{\Gamma(c_1)\cdots\Gamma(c_k)}x_1^{c_1-1}\cdots x_k^{c_k-1},\quad (x_1,\dots,x_k)\in\Delta,$$

with $\Delta = \{(x_1,\dots,x_k)\in\mathbb{R}_+^k : x_1+\cdots+x_k = 1\}$. As $n\to\infty$, the limit distribution of the relative proportions of the balls has this Dirichlet distribution. For $j\geq 1$, let $\boldsymbol{u}_j$ be an indicator vector whose lth component is 1 if the jth draw is colour l. Then $\boldsymbol{u}_j, j\geq 1$ are exchangeable random vectors, and the density f above is de Finetti's representing measure of the sequence; cf. Feller (1971, Chapter VII).

A coalescent tree can be generated either forward in time or backward in time. In forward time an edge of the tree is chosen at random to branch to increase the number of ancestors, corresponding to coalescence decreasing the number of ancestors backwards in time. The descendants of edges in a general coalescent tree generated forward in time can be identified with this classical urn model. Consider a cross section of a coalescent tree at a particular time in the past when there are k edges, and give each a distinct colour. Then at each branch point in forward time an additional edge is added, analogous to the urn model with $c_1 = \cdots = c_k = 1$. The probability of getting $n_1,\dots,n_k$ descendants of edges $1,2,\dots,k$ in a sample of size n is

$$\binom{n-1}{k-1}^{-1} \tag{2.3}$$

for $n_1+\cdots+n_k = n$. This follows from (2.1) by setting $r_i = n_i - 1, 1\leq i\leq k$ and $r = n-k$, and shows that the distribution is uniform on the collection of ordered non-empty k-subsets of a set with n elements. This result is derived in a different way in Kingman (1982).

2.1 The number of descendants of an edge

The probability $p_{nk}(b)$ that a particular edge among k ancestors subtends b particular descendants in a sample of n also follows from (2.1) by setting $k = 2, c_1 = 1, c_2 = k-1, r_1 = b-1, r_2 = n-b-k+1$:

$$p_{nk}(b) = \binom{n-b-1}{k-2}\binom{n-1}{k-1}^{-1},\quad 1\leq b\leq n-k+1, \tag{2.4}$$

the number of ways the $n-b$ descendants can be assigned to $k-1$ ancestors, divided by the total number of ways the n descendants can be assigned to their k ancestors.

2.2 Coalescence times in a subtree

A b-subtree of an n-tree is the subtree formed by the ancestral tree of a particular b genes from the sample of n. In an exchangeable model all b-subtrees are identically distributed. The coalescence times in a b-subtree are

$$T_i^{\#} = T_{M_i} + \cdots + T_{M_{i-1}+1},$$

where for $i = 1, 2, \ldots, b-1$, M_i is the number of edges in the n-tree at the time the b-subtree first has i edges (and we define $M_b \equiv n$). From Theorem 2 of Saunders *et al.* (1984), we have

$$\begin{aligned} &\phi(k, i; n, b) \\ &:= \quad \mathbb{P}(b\text{−subtree has } i \text{ edges when the } n\text{−tree has } k \text{ edges}) \\ &= \quad \frac{(n-b)!(n-k)!b!(b-1)!k!(k-1)!(n+i-1)!}{(b-i)!(k-i)!n!(n-1)!i!(i-1)!(k+b-1)!(n+i-k-b)!} \end{aligned} \qquad (2.5)$$

The limit as $n \to \infty$ is

$$\phi(k, i; \infty, b) = \frac{\binom{k}{i}\binom{b-1}{i-1}}{\binom{k+b-1}{k-1}}. \qquad (2.6)$$

Since T_k is included in the sum defining $T_i^{\#}$ if, and only if, there are i edges in the b-subtree when the n-tree has k edges, it follows that, for n finite or infinite,

$$\mathbb{E}\left(T_i^{\#}\right) = \sum_{k=i}^{n} \phi(k, i; n, b)\mathbb{E}(T_k). \qquad (2.7)$$

The reverse Markov chain $\{M_i, i = b-1, \ldots, 1\}$ can be simulated from $M_b = n$ by noting that the transition probabilities are, for $i-1 \leq \ell < k$,

$$\mathbb{P}(M_{i-1} = \ell \mid M_i = k) = \left(1 - \frac{i(i-1)}{k(k-1)}\right) \cdots \left(1 - \frac{i(i-1)}{(\ell+2)(\ell+1)}\right) \cdot \frac{i(i-1)}{(\ell+1)\ell}. \qquad (2.8)$$

As $n \to \infty$, the limiting marginal distribution of M_i is

$$\begin{aligned} \mathbb{P}(M_i = k) &= \frac{b!(b-1)!k!(k-1)!}{(b-i-1)!(k-i)!i!(i-1)!(b+k)!} \\ &= \binom{k-1}{i-1}\mathbb{E}\left\{\rho^i(1-\rho)^{k-i}\right\}, \ k \geq i, \end{aligned} \qquad (2.9)$$

where ρ has a Beta$(b-i, i+1)$ distribution. Thus M_i has a Negative Binomial–Beta mixture distribution. This is useful for simulating the $\{M_i\}$: generate M_{b-1} with a Negative Binomial–Beta$(1, b)$ mixture, then use the transition probabilities (2.8) to simulate $\{M_i, b-1 > i \geq 1\}$. See Saunders *et al.* (1994) for details of (2.8) and (2.9).

2.3 Branches in the subtree below a mutation

Later on we study the subtree of a general coalescent tree formed by considering the individuals who carry a particular mutation that arises just once in the history of the sample. If there are b copies of the mutation in the sample and the mutation arose while there were $J_0 = j$ ancestors, then the subtree describing the ancestry of the individuals carrying the mutation has $b - 1$ coalescence events. The first of these results in $J_{b-1} \leq n - 1$ ancestors of the sample, the second in J_{b-2} ancestors, and so on until the most recent common ancestor of the individuals carrying the mutation occurs when the sample has $J_1 \geq j$ ancestors. Notice that individuals not carrying the mutation cannot share common ancestors with those that do until the mutant individuals have coalesced to their most recent common ancestor. Wiuf and Donnelly (1999) study properties of the coalescent conditional on having this property. Here we use the urn model to derive the basic results we need.

Consider the urn model starting from $j-1$ red balls and one black ball. For $k = j, \ldots, n-1$, let U_k be the indicator function of the event that a black ball is added when there are k balls in the urn. According to the de Finetti urn representation, the $\{U_k\}$ are conditionally independent Bernoulli trials, given success probability Z having a Beta$(1, j-1)$ distribution with density $(j-1)(1-z)^{j-2}, 0 < z < 1$. There will be b black balls in the urn when there are n balls altogether if $U_j + \cdots + U_{n-1} = b-1$. It follows that conditioning the urn on having b black balls out of n is equivalent to conditioning on $U_j + \cdots + U_{n-1} = b - 1$.

For $i = 1, 2, \ldots, b-1$ let J_i be the number of balls already in the urn when the ith additional black ball is added. Given that there are b black balls in the urn containing n balls, we obtain the joint distribution of $(J_1, \ldots, J_{b-1})$ by exchangeability:

$$\begin{aligned}
&\mathbb{P}(J_1 = j_1, \ldots, J_{b-1} = j_{b-1} \mid J_0 = j_0) = \\
&\quad \mathbb{P}\left(U_{j_1} = 1, \ldots, U_{j_{b-1}} = 1 \mid \sum_{l=j}^{n-1} U_l = b-1\right) = \\
&\quad \binom{n-j}{b-1}^{-1}, \ (j_1, \ldots, j_{b-1}) \in \mathbf{I}(j, n),
\end{aligned} \tag{2.10}$$

where $\mathbf{I}(j, n) = \{(j_1, \ldots, j_{b-1}) : j \leq j_1 < j_2 \cdots < j_{b-1} \leq n-1\}$.

We can identify J_i as the number of ancestors of the sample at the time the subtree of size b has i distinct ancestors, for $i = b-1, b-2, \ldots, 1$. The result in (2.10) says that, conditional on $J_0 = j_0$, $J_1, \ldots, J_{b-1}$ are uniformly distributed over $\mathbf{I}(j, n)$.

3 Mutations in the tree

We are interested in the effects of mutation on general coalescent trees. We assume that for some $\theta \in (0, \infty)$, (A3) Conditional on the edge lengths of the tree, mutations occur according to independent Poisson processes of rate $\theta/2$ along the edges of the tree.

In the population genetics setting, the compound parameter θ is given by $\theta = 4Nu$, where u is the mutation rate per sequence per generation. To model the effects of each mutation, we use the *infinitely-many-sites model* (cf. Watterson 1975), under which a new mutation in the population is assumed to occur at a site in an infinitely-long DNA sequence where there has never previously been a mutation. Thus in this model the number of mutations in the ancestral tree of n sample genes is the number of *segregating sites* in these n sequences, that is, sites which contain two distinct types of base. A mutation on an edge of the tree at a site occurs at that site in all leaves subtended by that edge, while other leaves contain the ancestral base. Note that each mutation that has arisen in the history of the sample back to its most recent common ancestor is represented in that sample. In a finite-sites model of mutation there is a fixed number of sites in a gene and mutation at the same site more than once is possible. The infinitely-many-sites model can be obtained as a limit from the finite-sites model as the number of sites tends to infinity and the mutation rate per gene is kept fixed.

It is often convenient to think of the DNA sequences as being represented by unit intervals, and to label the locations of new mutations that arise in the sample using a sequence of independent and identically distributed random variables having an arbitrary distribution on (0,1). Thus for any set $M \subset (0,1)$, mutations with locations in M arise in a branch of the coalescent tree at rate $\theta\lambda(M)/2$ where $\lambda(M)$ is the probability of M under the mutation distribution. When mutation is uniform in (0,1) $\lambda(M) = |M|$, and non-uniform choices for λ correspond to mutational hotspots.

3.1 The distribution of the number of segregating sites

Let S_n be the number of segregating sites in a sample of n sequences under the infinitely-many-sites model. Since mutations occur as a Poisson process on the edges of the tree, S_n has a compound Poisson distribution with mean $\theta \sum_{j=2}^{n} jT_j/2$, and it follows immediately that

$$\mathbb{E}(S_n) \quad = \quad \frac{\theta}{2}\sum_{j=2}^{n} j\mathbb{E}(T_j), \qquad \mathrm{var}(S_n) = \frac{\theta^2}{4}\mathrm{var}\left(\sum_{j=2}^{n} jT_j\right) + \frac{\theta}{2}\sum_{j=2}^{n} j\mathbb{E}(T_j).$$

Watterson (1975) showed that in the standard coalescent process $\mathbb{E}(S_n) = \theta\sum_{j=1}^{n-1} j^{-1}$, $\mathrm{var}(S_n) = \sum_{j=1}^{n-1}\left(j^{-2}\theta^2 + j^{-1}\theta\right)$, and suggested the now commonly used moment estimator of θ given by $S_n/\sum_{j=1}^{n-1} j^{-1}$.

3.2 The distribution of pairwise differences

Let Π_n be the average number of pairwise differences between the $\binom{n}{2}$ sequences in a sample of n sequences in a general coalescent tree. If the coalescence time for two randomly chosen genes is T'_{n2}, then $\mathbb{E}(\Pi_n) = \theta\mathbb{E}(T'_{n2})$. Letting η_n be the number of ancestral lines in the coalescent when the two genes coalesce, we see that

$$\mathbb{P}(\eta_n \leq k) = \prod_{j=k+1}^{n} \left(1 - \binom{j}{2}^{-1}\right) = \frac{(n+1)(k-1)}{(n-1)(k+1)}.$$

Since

$$T'_{n2} = \sum_{j=\eta_n}^{n} T_j,$$

it follows that

$$\mathbb{E}(T'_{n2}) = \sum_{k=2}^{n} \mathbb{P}(\eta_n \leq k)\mathbb{E}(T_k) = \frac{(n+1)}{(n-1)} \sum_{k=2}^{n} \frac{(k-1)}{(k+1)} \mathbb{E}(T_k). \qquad (3.1)$$

Note that if there is a set of coalescent trees for $n = 2, 3, \ldots$ with coalescence times $\{T_{nj}, j = n, \ldots, 2\}$ on the same probability space such that, for $2 \leq k \leq n$, $\{T_{nj}, j = k, \ldots, 2\}$ is distributed as $\{T_{kj}, j = k, \ldots, 2\}$, then T'_{n2} is distributed as T_{22}. This is true for a coalescent process modeling constant or varying population size, but not necessarily true for a general coalescent tree.

3.3 Tajima's D

An important problem in the interpretation of genomic polymorphism data is the detection of regions of a chromosome that have undergone selection. One of the statistics most widely used for this purpose is Tajima's D, a standardized version of $S_{\text{scaled}} - \Pi_{\text{scaled}}$ (Tajima, 1989), where S_n and Π_n are scaled to be unbiassed estimates of θ. Departures from the assumptions of the usual neutral coalescent model of Kingman can be detected using the null distribution of D.

In the general coalescent tree

$$S_{\text{scaled}} = S_n / \sum_{j=2}^{n} \frac{1}{2} j \mathbb{E}(T_j), \qquad \Pi_{\text{scaled}} = \Pi_n / \mathbb{E}(T'_{n2}). \qquad (3.2)$$

In the usual coalescent process Π_n is an unbiased estimate of θ because $\mathbb{E}(T'_{n2}) = \mathbb{E}(T_2) = 1$, so $\Pi_{\text{Scaled}} = \Pi_n$. The analogue of D in the general coalescent tree, derived from equation (3.2), is $S_{\text{scaled}} - \Pi_{\text{scaled}}$. The distribution of D is discussed later in the paper.

4 Frequency spectra

The distribution of the number of mutant genes arising from a single mutation in an ancestor is of considerable interest. We assume that the mutation is segregating in the sample, so that it arose between the present and the time of the most recent common ancestor of the sample. Here we derive this distribution under the general conditions (A1), (A2) and (A3).

Let $\boldsymbol{T}$ denote the sequence of waiting times $T_2, \ldots, T_n$ in the coalescent tree of the sample. Consider a mutation arising at rate $\mu/2$, and let C denote the event that this mutation arises just once. Let $C_b \subseteq C$ denote the event that this mutation has b copies in the sample, and let I_k denote the event that the mutation arises when the sample has k ancestors. First we calculate $\mathbb{P}(C_b \cap I_k)$ using the Poisson nature of the mutation process.

$$\mathbb{P}(C_b \cap I_k|\boldsymbol{T}) \quad = \quad p_{n,k}(b)\left(kT_k\frac{\mu}{2}e^{-kT_k\mu/2} \times e^{-(L_n-kT_k)\mu/2}\right),$$

where $L_n = \sum_k jT_j$ is the length of the tree. Averaging over the distribution of $\boldsymbol{T}$ gives

$$\mathbb{P}(C_b \cap I_k) = kp_{n,k}(b)\mathbb{E}\left(T_k\frac{\mu}{2}e^{-L_n\mu/2}\right). \tag{4.1}$$

Summing (4.1) over $b = 1, \ldots, n-1, k = 2, \ldots, n-b+1$ gives

$$\mathbb{P}(C) = \sum_{k=2}^{n} k\mathbb{E}\left(T_k\frac{\mu}{2}e^{-L_n\mu/2}\right). \tag{4.2}$$

Dividing (4.1) by (4.2) shows that

$$\mathbb{P}(C_b \cap I_k \mid C) =: q_{n,b;k} = \frac{kp_{n,k}(b)\mathbb{E}\left(T_k e^{-L_n\mu/2}\right)}{\sum_{k=2}^{n} k\mathbb{E}\left(T_k e^{-L_n\mu/2}\right)}, \tag{4.3}$$

for $0 < b < n,\ k = 2, \ldots, n-b+1$. Letting $\mu \to 0$ we obtain

$$q_{n,b;k} = \frac{kp_{n,k}(b)\mathbb{E}(T_k)}{\sum_{k=2}^{n} k\mathbb{E}(T_k)},\ 0 < b < n,\ k = 2, \ldots, n-b+1. \tag{4.4}$$

The *frequency spectrum* is the probability distribution $q_{n,b}, b = 1, \ldots, n-1$ of the number of times the mutation is represented in the sample. Since $q_{n,b} = \sum_{k=2}^{n-b+1} q_{n,b;k}$, we see that

$$q_{n,b} = \frac{\sum_{k=2}^{n-b+1} kp_{n,k}(b)\mathbb{E}(T_k)}{\sum_{k=2}^{n} k\mathbb{E}(T_k)},\ 0 < b < n. \tag{4.5}$$

as given in Griffiths and Tavaré (1998). Equation (4.4) provides the frequency spectrum for a *particular* segregating site in the infinitely-many-sites model. In the case of a constant population size, we see that $\sum_{k=2}^{n-b+1} kp_{nk}(b)\mathbb{E}(T_k) = 2/b$, so that

$$q_{nb} = \frac{1}{b}\left(\sum_{k=1}^{n-1}\frac{1}{k}\right)^{-1}.$$

Stephens (2000) derived the result analogous to (4.5) for arbitrary μ. We note that this is a particular case of (4.5) obtained by modifying the distribution of

$\boldsymbol{T}$ to $\boldsymbol{T}'$ such that the Laplace transform of $\boldsymbol{T}'$ is

$$\mathbb{E}\left(e^{-s_nT'_n-\cdots-s_2T'_2}\right)=\frac{\mathbb{E}\left(e^{-s_nT_n-\cdots-s_2T_2}e^{-\frac{1}{2}\mu L_n}\right)}{\mathbb{E}\left(e^{-\frac{1}{2}\mu L_n}\right)}.$$

Therefore

$$\mathbb{E}\left(T'_k\right)=\frac{\mathbb{E}\left(T_ke^{-\frac{1}{2}\mu L_n}\right)}{\mathbb{E}\left(e^{-\frac{1}{2}\mu L_n}\right)},$$

which evaluates to

$$\mathbb{E}\left(T'_k\right)=\frac{2}{k(k-1+\mu)}$$

in the coalescent process with constant population size.

5 Distribution of the age of a mutation

Let $\xi_{n,b}$ denote the age of a mutant having b copies in a sample of size n, for $0<b<n$. Griffiths and Tavaré (1998) showed that the density of $\xi_{n,b}$ is given by

$$g_{n,b}(t)=\frac{\sum_{k=2}^n kp_{n,k}(b)\mathbb{P}(A_n(t)=k)}{\sum_{k=2}^n kp_{n,k}(b)\mathbb{E}(T_k)},\ t>0, \tag{5.1}$$

where $A_n(t)$ denotes the number of ancestors of the sample of n time t ago. Furthermore, the moments of $\xi_{n,b}$ are given by

$$\mathbb{E}(\xi_{n,b}^j)=\frac{\sum_{k=2}^n k(k-1)\binom{n-k}{b-1}\frac{1}{j+1}\mathbb{E}\Big(W_k^{j+1}-W_{k+1}^{j+1}\Big)}{\sum_{k=2}^n k(k-1)\binom{n-k}{b-1}\mathbb{E}(T_k)},j=1,2\dots, \tag{5.2}$$

where for $k=n,n-1,\dots,2,$

$$W_k=T_n+\cdots+T_k \tag{5.3}$$

is the time taken to reach state $k-1$, with $W_{n+1}\equiv 0$. The mean and variance of $\xi_{n,b}$ can be derived from (5.2).

The population versions of (5.1) and (5.2) were also studied in Griffiths and Tavaré (1998). If we assume that $\{A_n(t),t\geq 0\}$ converges in distribution to a process $\{A(t),t\geq 0\}$ as $n\to\infty$, and that the time taken for $A(\cdot)$ to reach 1 is finite with probability one, then as $n\to\infty$, and $b/n\to x$, $0<x<1$, we obtain the density of the age ξ_x as

$$g_x(t)=\frac{\sum_{k=2}^\infty k(k-1)(1-x)^{k-2}\mathbb{P}(A(t)=k)}{\sum_{k=2}^\infty k(k-1)(1-x)^{k-2}\mathbb{E}(T_k)}, \tag{5.4}$$

and the moments of ξ_x are given by

$$\mathbb{E}(\xi_x^j)=\frac{\sum_{k=2}^\infty k(k-1)(1-x)^{k-2}\frac{1}{j+1}\mathbb{E}\Big(W_k^{j+1}-W_{k+1}^{j+1}\Big)}{\sum_{k=2}^\infty k(k-1)(1-x)^{k-2}\mathbb{E}(T_k)},j=1,2\dots. \tag{5.5}$$

Related results appear in Wiuf and Donnelly (1999).

6 Coalescence times in a subtree

In this section, we study properties of the subtree that relates a sample of chromosomal regions carrying a particular mutation. In the b-subtree under a mutation in an n-tree, the coalescence times are

$$T_i' = T_{J_{i-1}+1} + \cdots + T_{J_i} = W_{J_{i-1}+1} - W_{J_i+1}\,, i = 2, \ldots, b, \tag{6.1}$$

where W_k is defined in (5.3) and we define $J_b \equiv n$. Note that $W_i' := T_b' + \cdots + T_i' = W_{J_{i-1}+1}$.

We can use the results of Section 4 to find the distribution of J_0, the number of ancestors at the time the mutation arose. ¿From (4.4) and (4.5) we have

$$\mathbb{P}(J_0 = j) = \frac{q_{n,b;j}}{q_{n,b}} = \frac{jp_{nj}(b)\mathbb{E}(T_j)}{\sum_{l=2}^{n-b+1} lp_{nl}(b)\mathbb{E}(T_l)},\ j = 2, \ldots, n-b+1. \tag{6.2}$$

Conditional on $J_0 = j$, we saw in (2.10) that $J_1, \ldots, J_{b-1}$ are uniform in $\mathbf{I}(j, n) = \{j \le j_1 < j_2 < \ldots < j_{b-1} \le n-1\}$. By considering the $k-j$ available indices less than k for $J_1, \ldots, J_{i-1}$ and the $n-k-1$ available indices for $J_{i+1}, \ldots, J_{b-1}$, it follows that

$$\mathbb{P}(J_i = k \mid J_0 = j) = \frac{\binom{k-j}{i-1}\binom{n-k-1}{b-i-1}}{\binom{n-j}{b-1}},\ k = j+i-1, \ldots, n-b+i. \tag{6.3}$$

The unconditional mean waiting times may be computed from the formula

$$\mathbb{E}(W_i') = \frac{\sum_{j=2}^{n-b+1} jp_{nj}(b) \sum_{k=j+i-1}^{n-b+i} \mathbb{E}(W_k T_j) \mathbb{P}\Big(J_{i-1} + 1 = k \mid J_0 = j\Big)}{\sum_{j=2}^{n-b+1} jp_{nj}(b)\mathbb{E}(T_j)} \tag{6.4}$$

and $\mathbb{E}(T_i') = \mathbb{E}(W_i') - \mathbb{E}(W_{i+1}')$.

The length of the subtree is $L_{nb} = \sum_{l=2}^{b} lT_l'$, and its mean can be computed from (6.4) and the fact that

$$\mathbb{E}\sum_{l=2}^{b} lT_l' = \sum_{l=2}^{b} l\mathbb{E}(W_l' - W_{l+1}') = \mathbb{E}W_2' + \sum_{l=2}^{b} \mathbb{E}W_l'.$$

However, we can also exploit the urn representation from Section 2.3.

Suppose the mutation occurs when there are $J_0 = j$ ancestors of the sample. Let $U_k, k = j, \ldots, n-1$ be the indicator of the event that the subtree branches while the sample has k ancestors. We saw earlier that conditioning on b descendants in the subtree is equivalent to conditioning on $\sum_{i=j}^{n-1} U_i = b-1$, and we define J_1 to be the number of ancestors of the sample when the subtree reaches its most recent common ancestor. In terms of $\{U_i\}$, we have

$$L_{nb} = \sum_{k=J_1+1}^{n} \Big(1 + \sum_{i=J_1}^{k-1} U_i\Big) T_k,$$

with $U_i = 0, j \le i \le J_1 - 1$ and $U_{J_1} = 1$. Consider the conditional distribution of L_{nb} given $J_1 = j_1, \sum_{i=j_1+1}^{n-1} U_i = b-2$. By exchangeability, the conditional distribution of $\{U_i, j_1+1 \le i \le n-1\}$ is uniform on $\binom{n-j_1-1}{b-2}$ positions for which $b-2$ of the indicator variables are 1.

It follows from this approach after some algebra that the mean edge length, conditional on a mutation subtending b descendants is

$$\mathbb{E}(L_{nb}) = \frac{\sum_{j=2}^{n-b+1} j p_{nj}(b) \sum_{k=j+1}^{n} c_{jk} \mathbb{E}(T_j T_k)}{\sum_{j=2}^{n-b+1} j p_{nj}(b) \mathbb{E}(T_j)}, \tag{6.5}$$

where

$$c_{jk} = b - (b-1)\frac{n-k}{n-j} - \frac{(n-k)!(n-j-b+1)!}{(n-j)!(n-k-b+1)!}.$$

In the usual constant-size coalescent, (6.2) reduces to

$$\mathbb{P}(J_0 = j) = \frac{\binom{n-j}{b-1}}{\binom{n-1}{b}}, \tag{6.6}$$

and it follows that

$$\mathbb{E}(L_{nb}) = \binom{n-1}{b}^{-1} \sum_{j=2}^{n-b+1} \binom{n-j}{b-1} \sum_{k=j+1}^{n} \frac{2}{k(k-1)} c_{jk}. \tag{6.7}$$

7 Further mutations in subtrees

In the remainder of this paper, we discuss the theoretical issues relating to frequency spectra and properties of ages of mutations. To this end, recall that the number of additional mutations falling in the subtree determined by the given mutation has, under an infinitely-many-sites assumption for these additional mutations, a Compound Poisson($\theta L_{nb}/2$) distribution, where L_{nb} is the total edge length of the subtree up to its MRCA and θ is the mutation parameter appropriate for the additional mutations. In particular, the expected number of segregating sites in the subtree is just $\theta \mathbb{E}(L_{nb})/2$, which can be found from (6.5).

7.1 A simulation algorithm

While a number of explicit results are available for properties of subtrees, it is useful to have a simulation algorithm that produces them. Here we focus on subtrees arising below a mutation having frequency b in a sample of size n. One approach is provided by Wiuf and Donnelly (1999). Another method is:

(B1) Choose j_0 according to the distribution of J_0 in (6.2).

(B2) Choose $j_1 < \cdots < j_{b-1}$ from the conditional distribution of $J_1, \ldots, J_{b-1}$ given $J_0 = j_0$; this is uniform over $\mathbf{I}(j_0, n)$, as in (2.10).

(B3) Join edges at random to form the subtree.

If, in addition, coalescence times $T_i', i = b, b-1, \ldots, 2$ in the subtree are required, we need only add:

(B4) Simulate an observation from the joint distribution of $T_n, \ldots, T_{j_0+1}$.

(B5) Compute the times T_i' via (6.1).

This allows us to calculate summary statistics about the subtree, such as its height (that is, the time to the MRCA of the subtree), and its length L_{nb}.

7.2 The age of the mutation

To simulate from the age of the mutation, we add:

(B6) Conditional on the results of (B4), simulate from the random variable Z having the size-biased distribution of T_{j_0} and set $T^* = UZ$, where U is an independent U(0,1) random variable.

The time $A = T_n + \cdots + T_{j_0+1} + T^*$ is the required age of the mutation. Note that the random variable Z has density proportional to $xf(x)$, where f is the density of T_{j_0}. Hence if T_{j_0} has an exponential distribution with parameter τ, then so too does UZ.

If one wants to simulate observations from the posterior distribution of trees and times conditional on the number k of segregating sites appearing *in the b individuals carrying the mutation* in a region completely linked to the mutation, then one can add a rejection step (cf. Tavaré *et al.* 1997):

(B7) Accept the results of (B1) – (B6) with probability $\text{Po}(\theta L_{nb}/2)\{k\}/\text{Po}(k)\{k\}$ where θ is the mutation parameter appropriate for the linked region, L_{nb} is the subtree length, and we use the notation $\text{Po}(\mu)\{k\} = \mu^k e^{-\mu}/k!$. Otherwise, go to (B1).

Note that in this case step (B3) is not needed.

Slatkin and Rannala (1997) discussed the problem of estimating the age of a mutation given its frequency in the sample together with (an estimate of) the number of mutations that had arisen in a completely linked region among the chromosomes carrying the mutation. Under an infinitely-many-sites model for these extra mutations, the algorithm in (B7) provides one approach to Slatkin and Rannala's problem in the coalescent setting. An example appears in the next section. When the additional data are complete DNA sequences from the linked region, this algorithm no longer works, essentially because the acceptance probability is far too small. In this case, a Markov chain Monte Carlo approach can be implemented, as in Markovtsova *et al.* (2000).

7.3 An example of simulation of ages

The distribution of the age of a mutation and the height of the subtree were simulated using 50,000 runs of algorithm (B7) for the case $n = 200, b = 30, \theta = 4.0$ and 5 segregating sites. The mean age was 1.01 with standard deviation 0.91, while the mean subtree height was 0.40 with a standard deviation of 0.25. Percentiles of the distributions are given below, together with the estimated densities.

	2.5%	25%	50%	75%	97.5%
age	0.156	0.412	0.721	1.289	3.544
subtree height	0.099	0.218	0.334	0.514	1.056

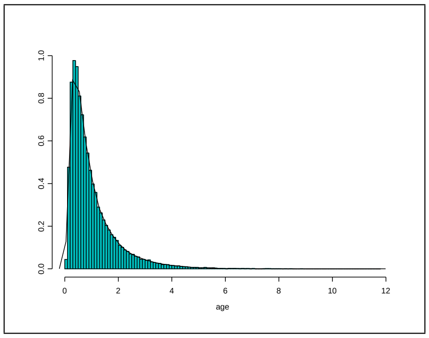

Figure 2. Density of age of mutation.

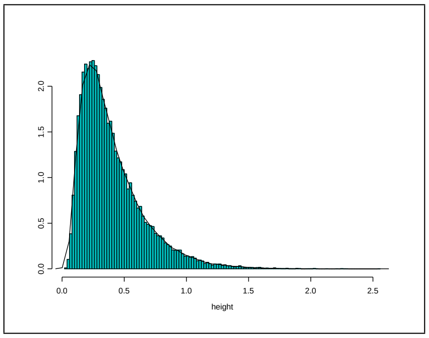

Figure 3. Density of height of subtree.

7.4 Simulation of Tajima's D in a subtree

The formula (6.5) can be used together with (3.1) and (3.2) to find the expected number of segregating sites and the expected pairwise difference in a subtree, and to study the analogue of Tajima's D in the subtree. This last provides a way to test for neutrality of the region around the mutation of interest; see Innan and Tajima (1997) for related material. It is straightforward to use the simulation algorithm outlined above to simulate from the distribution of $S_{\text{scaled}} - \Pi_{\text{scaled}}$: use steps (B1) – (B3) to produce the subtree, and then simulate mutations (with parameter θ) on that tree. Once done, the observed values of S_n and Π_n can be recovered.

We used this approach to simulate 50000 observations for the case $n = 200$, $b = 30$, $\theta = 4.0$. The mean of $S_{\text{scaled}} = 4.0143$ and the mean of $\Pi_{\text{scaled}} = 4.0243$. The percentage points of $S_{\text{scaled}} - \Pi_{\text{scaled}}$ are given in the following table:

2.5%	5.0%	10.0%	25.0%	50.0%	75.0%	90.0%	95.0%	97.5%
-5.3	-4.0	-2.4	-0.8	0.0	1.4	2.3	2.9	3.8

In a situation such as this the percentage points can be used to test for departures from neutrality for a subsample of b genes under a mutation. For example, if genes with the mutation were under positive selection $S_{\text{scaled}} - \Pi_{\text{scaled}}$ would have a heavier distribution on the negative side.

8 Sampling under a mutation in the population

Consider a mutation which has frequency $x \in (0, 1)$ in the population. Our interest is in characteristics of a sample taken from this proportion of the population. To obtain the distribution of coalescence times under the mutation suppose, as previously in the paper, that a sample of n has b copies of a mutation and then let $n \to \infty$, $b \to \infty$ in such a way that $b/n \to x$. Recall that conditional on $J_0 = j_0$, $J_1, \ldots, J_{b-1}$ have a uniform distribution on $\mathbf{I}(j_0, n)$. The joint conditional distribution of $J_1 < J_2 < \cdots < J_i$ is thus

$$\mathbb{P}(j_1,\dots,j_i) = \frac{\binom{n-1-j_i}{b-i-1}}{\binom{n-j_0}{b-1}},\ j_1 < \cdots < j_i. \tag{8.1}$$

The limit distribution of (8.1) as $n \to \infty, b/n \to x$ is

$$\mathbb{P}(j_1,\dots,j_i) = x^i(1-x)^{j_i-i-j_0+1},\ j_1 < \cdots < j_i. \tag{8.2}$$

We denote the limit random variables by $J_0^x, J_1^x, \dots$ Informally, in this limit, when the whole tree branches there is probability x that the branch is in the subtree, since the subtree has a total proportion x of the branching points. $J_1^x < J_2^x < \cdots$ are thus distributed, from (8.2), as success epochs in a sequence of Bernoulli trials, shifted by J_0. If G is a geometric random variable with distribution

$$x(1-x)^g,\ g = 0,1,\dots$$

then $Q_i = J_i^x - J_{i-1}^x$, $i = 1,2,\dots$, are independent random variables with Q_1 distributed as G, and Q_i, $i > 1$ distributed as $G+1$. $L_i = J_i^x - J_0^x - i + 1$ has a negative binomial distribution

$$\binom{\ell+i-1}{i-1} x^i(1-x)^\ell,\ \ell = 0,1,\dots$$

The distribution of J_0^x depends on the coalescence times in the population. Assuming convergence of $\{T_{nj}\}$ to a proper collection of coalescence times $\{T_j\}$, with the means also converging, the distribution of J_0^x converges to

$$\frac{j(j-1)(1-x)^{j-2}\mathbb{E}(T_j)}{\sum_{i=2}^\infty i(i-1)(1-x)^{i-2}\mathbb{E}(T_i)},\ j \geq 2, \tag{8.3}$$

and the population then has coalescence times

$$T_i^x = T_{J_{i-1}^x+1} + \cdots + T_{J_i^x},\ i = 2,3,\dots \tag{8.4}$$

The collection of coalescence times $\{T_i^x\}$ can then be used in formulae for characteristics of samples under the mutation in the population for general coalescent trees. It is straightforward to simulate $\{T_i^x\}$ by simulation of J_0^x from the distribution (8.3), $\{J_i^x; i \geq 1\}$ from their geometric structure, and $\{T_i^x\}$ from (8.4).

8.1 Results for the standard coalescent

In the usual coalescent process it is possible to find explicit formulae for the means and second moments of the coalescence times $\{T_j^x\}$ and then use these in applications. In particular,

$$\mathbb{P}(J_0^x = j) = x(1-x)^{j-2},\ j \geq 2.$$

It follows that

$$\mathbb{E}(T_i^x) \quad = \quad 2\int_0^1 z^{-1}f^i\Big(1-f\Big)dz, \tag{8.5}$$

and

$$\begin{aligned}\mathbb{E}\big(W_i^x\big) &= 2\int_0^1 z^{-1}f^i dz, \\ \mathbb{E}\big((W_i^x)^2\big) &= 8\int_0^1 z^{-1}f^i\Big((-\log z)(1-z)^{-1}-1\Big)dz. \end{aligned} \tag{8.6}$$

where

$$f(z;x) = zx(1-z(1-x))^{-1}$$

is the pgf of a geometric random variable shifted by 1. It follows from (8.6) the mean time to coalescence of the subtree under the mutation is

$$\mathbb{E}(W_2^x) = 2x(1-x)^{-2}(x\log(x)-x+1). \tag{8.7}$$

Substituting into (5.5) with $j=1$ and simplifying, we see that the mean age of another mutation occurring in the subtree under the mutation which subtends a frequency $0<y<1$ (relative to x) is given by

$$A(x,y) = -\frac{4xy\big((1+xy)\log(xy)-2xy+2\big)}{2xy\log(xy)-x^2y^2+1}. \tag{8.8}$$

Note that $A(x,y)$ is a function of xy, the proportion of the second mutant in the total population.

Alternative forms for integrals in this section can be found by changing the variable of integration from z to f, noting that $0\le f\le 1$, and that

$$\frac{dz}{df} = x^{-1}\Big(1+\frac{1-x}{x}f\Big)^{-2}.$$

9 Other sampling schemes

In this section, we develop the theory required for studying two different sampling schemes. Motivated by the problem of sampling from a disease registry, we study the genealogy of a random sample from the population carrying a particular mutation known to have frequency x in the whole population. For related material see Wiuf (2000). The second scheme we address is motivated by the case-control design, in which one considers in addition a sample of the same size from that part of the population not carrying the mutation. We note that the results (9.2), (9.4)–(9.5) and (9.7) apply to the standard coalescent.

9.1 Sampling from the disease population

Suppose then that a random sample of n genes is taken from a mutant class whose frequency is x in the population. The coalescence times in the sample are distributed as $\{T_i^{\#}\}$ of Section 2.2, where

$$T_i^{\#} = T_{M_i}^x + \cdots + T_{M_{i-1}+1}^x,\ i = n, \dots, 2. \tag{9.1}$$

Thus the subpopulation of mutant genes of frequency x plays the role of the population, with $\{T_i^{\#}\}$ the population coalescence times. The formula (2.7) for $\mathbb{E}\Big(T_i^{\#}\Big)$ holds with n there set equal to ∞, and substituting we obtain

$$\mathbb{E}\Big(T_i^{\#}\Big) = \sum_{k=i}^{\infty} \frac{\binom{k}{i}\binom{n-1}{i-1}}{\binom{k+n-1}{k-1}} \mathbb{E}\Big(T_k^x\Big) =$$

$$2n\binom{n-1}{i-1}\int_0^1\int_0^1 z^{-1} f^i w^{i-1}(1-fw)^{-(i+1)}(1-w)^{n-1}(1-f)dwdz, \tag{9.2}$$

where $f \equiv f(z;x)$.

The site frequency spectrum in a random sample of size n under a mutation of frequency x in the population is, from (4.5),

$$q_{n,j} = \frac{(n-j-1)!(j-1)!\sum_{k=2}^n k(k-1)\binom{n-k}{j-1}\mathbb{E}(T_k^{\#})}{(n-1)!\sum_{k=2}^n k\mathbb{E}(T_k^{\#})} \tag{9.3}$$

The properties of Tajima's D in this setting may be studied using the following results. From Section 3.1, the mean number of mutations in a sample of n is

$$\begin{aligned}
\mathbb{E}(S_n) &= \frac{\theta}{2}\sum_{i=2}^n i\mathbb{E}\Big(T_i^{\#}\Big) \\
&= n\theta\int_0^1\int_0^1 z^{-1}(1-w)^{n-1}f(1-f)(1-fw)^{-2} \\
&\qquad \cdot \Big((1+\phi)^{n-1} + (n-1)\phi(1+\phi)^{n-2} - 1\Big)dwdz, \qquad (9.4)
\end{aligned}$$

where $\phi = fw/(1-fw)$. From Section 3.2, the expected number of pairwise differences between two of the n sequences is

$$\begin{aligned}
\mathbb{E}(\Pi_n) &= \theta\mathbb{E}(T_2^{\#}) = \theta\sum_{k=2}^{\infty}\frac{k-1}{k+1}\mathbb{E}(T_k^x) \\
&= 2\theta\int_0^1 xf^{-1}(x+(1-x)f)^{-1}\Big(2-f+\frac{2(1-f)\log(1-f)}{f}\Big)df (9.5)
\end{aligned}$$

because a random sample of two genes from a random sample of n is distributed as a random sample of two from under the subtree. Tabulated below are mean coalescence times for a sample of two from the subtree.

x	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$\mathbb{E}(T_2^{\#})$	0.00	0.09	0.16	0.22	0.28	0.34	0.39	0.44	0.49	0.53	0.58

Further properties are perhaps most simply studied by a simulation approach. First, simulate $\{T_i^{\#}\}$ and the coalescence pattern in the tree, and add mutations along the edges according to a Poisson process of rate $\theta/2$. Simulation of $\{T_i^{\#}\}$ involves simulation of $\{J_i^x\}$, $\{T_j^x\}$, $\{M_i\}$, $\{T_i^{\#}\}$ in order. It is of interest to consider simulated percentage points of Tajima's D analogue $S_{\text{scaled}} - \Pi_{\text{scaled}}$ for a test of the standard neutral model under a mutation, against various alternatives such as selection or growth.

9.2 Case-control sampling

In a case-control model a sample is taken from chromosomes carrying a particular mutation and another sample taken from genes without that mutation. Denote the coalescence times in the population of genes without the mutation as $\{T_k^{1-x}\}$ and the number of ancestors of the population when the subtree branches as $\{J_i^{1-x}\}$. If $J_0^x = j$, then the two coalescent trees under and not under the mutation are coupled, with branching occurring in the respective trees with probabilities x and $1-x$ while there are greater than or equal to j ancestors of the total population. Let H be a geometric random variable with distribution $(1-x)x^h, h = 0, 1, \ldots$ Then $\{J_{j+r}^{1-x} - J_{j+r-1}^{1-x}, r = 0, 2, \ldots .\}$ are independent with $J_j^{1-x} - j$ is distributed as H, and $J_{j+r}^{1-x} - J_{j+r-1}^{1-x}$ distributed as $H+1, r \geq 1$. Coalescence times are

$$T_k^{1-x} = \begin{cases} T_k, & k < j-1 \\ T_{J_{k-1}^{1-x}+1} + \cdots + T_{J_k^{1-x}}, & k \geq j-1. \end{cases} \tag{9.6}$$

It can be shown that the mean coalescence time in the subtree is

$$\mathbb{E}(W_2^{1-x}) = (1-x)^{-3}(1-2x)^{-1}\Big(2x^2(x^4 - 6x^3 + 15x^2 - 12x + 3)\log(x)$$
$$-2x(1-x)^5\log(1-x) + (1-x)(1-2x)(x^4 - x^3 - 4x^2 + 2)\Big) \tag{9.7}$$

The table below gives the mean coalescence times in the two subtrees by numerically evaluating (8.7) and (9.7).

x	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$\mathbb{E}(W_2^x)$	0.00	0.17	0.30	0.41	0.52	0.61	0.70	0.78	0.86	0.93	1.00
$\mathbb{E}(W_2^{1-x})$	2.00	2.29	2.48	2.63	2.77	2.90	3.05	3.20	3.36	3.52	3.67

There is no doubt that coalescent methods have revolutionized the way in which molecular variation data are studied. They provide a way to model the ancestral relationships among gene regions, which in turn provides the basis for inference and estimation for such data. There have been two basic applications of the coalescent approach: *"forward" methods* that are used to study the properties of typical samples, and *"backward" methods* that are used to infer the features of a coalescent consistent with a given set of data. Ancestral inference, an example of the second type, has provided some challenging statistical problems, some of which are discussed here. As more genome-wide data are collected, these inference questions become more challenging. For example, it should be possible to use genome-wide data to understand better the fluctuations that have occurred in population sizes through time. Weiss (2002) provides an illustrative example in this volume.

10 References

Donnelly, P. and Tavaré, S. (1995) Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* 29, 410–421.

Feller, W. (1968) *An Introduction to Probability Theory and its Applications.* Vol 1, 3rd ed. Wiley, New York.

Feller, W. (1971) *An Introduction to Probability Theory and its Applications.* Vol 2, 2nd ed. Wiley, New York.

Griffiths, R. C. (1980) Lines of descent in the diffusion approximation of neutral Wright-Fisher models. *Theoret. Popul. Biol.* 17, 37–50.

Griffiths, R. C. and Tavaré, S. (1994) Sampling theory for neutral alleles in a varying environment. *Proc. R. Soc. Lond. B* 344, 403–410.

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

Griffiths, R. C. and Tavaré, S. (1999) The ages of mutations in gene trees. *Ann. Appl. Prob.* 9, 567–590.

Hartl, D. L. and Jones, E. W. (2001) *Genetics. Analysis of genes and genomes.* Fifth edition. Jones and Bartlett Publishers Inc., Sudbury, Massachusetts.

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

Hudson, R. R. (1990) Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology,* eds. D Futuyma, J Antonovics, 7, 1–44. Oxford University Press.

Innan, H. and Tajima, F. (1997) The amounts of nucleotide variation within and between allelic classes and the reconstruction of the common ancestral sequence in a population. *Genetics* 147, 1431-1444.

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

Markovtsova, L., Marjoram, P. and Tavaré, S. (2000) The age of a unique event polymorphism. *Genetics* 156, 401–409.

Nordborg, M. (2001) Coalescent theory. In *Handbook of Statistical Genetics*, eds. Balding, D. J., Bishop M., and Cannings, C., 179–208. Wiley, Chichester.

Nordborg, M. and Tavaré, S. (2002) Linkage disequilibrium: what history has to tell us. *Trends Genet.* 18, 83-90.

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

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

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

Slatkin, M. and Rannala, B. (2000) Estimating allele age. *Annu. Rev. Genomics Hum. Genet.* 1, 225–249.

Stephens, M. (2000) Times on trees and the age of an allele. *Theoret. Popul. Biol.* 57, 109–119.

Stephens, M. (2001) Inference under the coalescent. In *Handbook of Statistical Genetics*, eds. Balding, D. J., Bishop M., and Cannings, C., 213–238. Wiley, Chichester.

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

Tajima, F. (1989) Statistical methods for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595.

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

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

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

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

Weiss, G. (2002) Linked versus unlinked DNA data – a comparison based on ancestral inference. This volume.

Wiuf, C. (2000) On the genealogy of a sample of neutral rare alleles. Theoret. Popul. Biol. 58, 61–75.

Wiuf, C. (2002) The age of a rare mutation. This volume.

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