Adv. Appl. Prob. **37,** 647–662 (2005)
Printed in Northern Ireland

MINIMAL CLADE SIZE AND EXTERNAL BRANCH LENGTH UNDER THE NEUTRAL COALESCENT

MICHAEL G. B. BLUM * AND
OLIVIER FRANÇOIS,* ** *Institut National Polytechnique de Grenoble*

Abstract

Given a sample of genes taken from a large population, we consider the neutral coalescent genealogy and study the theoretical and empirical distributions of the size of the smallest clade containing a fixed gene. We show that the theoretical distribution is strongly related to a Yule distribution of parameter 2, and that the empirical count statistics are asymptotically Gaussian as the number of genes grows to infinity. Then we consider external branches of the coalescent tree, and describe their lengths. Using the infinitely many sites model of mutation, we also describe the conditional distribution of the external branch lengths, given the number of pairwise differences between a reference DNA sequence and the sequence of one closest relative in the sample.

Keywords: Coalescent genealogies; tree shape statistics; statistical genetics

2000 Mathematics Subject Classification: Primary 92D10; 92D20
Secondary 60J70

1. Introduction

Coalescent theory associated with molecular information has proven to be an invaluable tool for assessing the degree of relatedness between individuals. To date, these tools have been successively applied by human geneticists to infer very deep relationships, of the order of hundreds of generations [12]. For example, Donnelly *et al.* [5] estimated the time since the most recent common ancestor of modern humans from DNA sequences of the ZFY intron (the intron of the zinc finger protein on the Y chromosome).

In this article, we are concerned with the estimation of intermediate ancestry and, more specifically, the relatedness of a given gene (or individual) to a sample of $n-1$ weakly related genes. The precise details of population history or geographic structure will not enter the analysis, and the evolution of the population will be assumed to be selectively neutral. In addition, we will consider genes that do not recombine. In other words, genetic drift will be the only factor responsible for the allelic variation within the population. Under these assumptions, the genealogy of n genes is well approximated by the coalescent model when the population size N grows to infinity [10]. The approximation arises as a diffusion limit when time is measured in units of N generations. See [17] and [6] for recent reviews on the subject.

The article is structured in two parts. In the first part, we give results about the numbers of relatives of the reference gene, i.e. the smallest number of genes (minus one) that share an ancestor with the reference gene. This subset of genes will be called a *minimal clade* of

Received 4 October 2004; revision received 1 April 2005.
* Postal address: Laboratoire TIMC-TIMB, Institute for Health and Information Engineering, Faculty of Medicine, F38706 La Tronche cedex, France.
** Email address: olivier.francois@imag.fr

the coalescent tree. A minimal clade contains the reference gene plus the subset of its closest relatives. We shall show that the number of closest relatives follows a Yule distribution of parameter 2 (see [3]). Given the genealogy, we shall also study the limit probability distributions of count statistics as n grows to infinity. The studied statistics correspond to the number of genes with k relatives, $k \geq 2$. We shall prove their asymptotic normality and, hence, provide new statistical tests of neutral evolution based on the shape of trees [11].

In the second part of the article, we study the length of an external branch of the coalescent tree [7]. This length corresponds to the time since the coalescence of an arbitrary lineage with the other lineages, and constitutes a natural measure of the degree of relatedness of a gene with the rest of the sample. Using the *infinitely many sites model* of the DNA molecule, we also describe the conditional distribution of the coalescence time given the number of substitutions between a reference DNA sequence and one of its closest relatives. This distribution corresponds to an explicit mixture of gamma distributions, and extends results of Tajima [16]. We conclude the article with an application to the human Y chromosome.

2. Background and notation

Consider a sample of n genes. In the coalescent, one wishes to record information both about the number of ancestors at various times and about which genes share common ancestors. The *ancestral process* $A_n(t)$ records the number of distinct ancestors of the sample at a time t in the past. It can be described as a continuous-time Markov chain on $[n] := \{1, \ldots, n\}$, such that

$$A_n(0) = n,$$

1 is an absorbing state, and the rate of transition from k to $k-1$ is equal to

$$\lambda_k = \tfrac{1}{2}k(k-1), \qquad k = n, \ldots, 2.$$

This means that the times (T_k), $k = n, \ldots, 2$, separating coalescence events are independent and exponentially distributed with mean $2/k(k-1)$. In the sequel, we shall also write

$$S_k = T_n + \cdots + T_k.$$

To understand the topology of the tree, one possibility is to label genes in the sample from the set $[n]$ and define a random equivalence relation. In this relation, the genes i and j are in the same class at time t if and only if they share a common ancestor at this time. Denote by $C(t)$ the random partition obtained from this equivalence relation at time t. According to Kingman [10], the process $C(t)$ is a continuous-time Markov chain on the set of all partitions of $[n]$ (denoted $\mathcal{E}_n$) for which

$$C(0) \equiv \{\{1\}, \ldots, \{n\}\}.$$

The transition rates of this Markov chain can be described as follows. For all $\alpha, \beta \in \mathcal{E}_n$,

$$q_{\alpha\beta} = \begin{cases} 1 & \text{if } \alpha \prec \beta, \\ 0 & \text{otherwise,} \end{cases}$$

where the expression $\alpha \prec \beta$ means that α and β are nested partitions such that β may be obtained from α by merging two classes. The observation that 1 is an absorbing state of $A_n(t)$ means that $C(t)$ converges almost surely to $\{[n]\}$, the final state in which a single class remains. The embedded discrete-time Markov chain $\{\mathcal{C}_k\}$, $k = n, \ldots, 1$, moves through the sequence

$$\mathcal{C}_n \equiv C(0) \prec \mathcal{C}_{n-1} \prec \mathcal{C}_{n-2} \prec \cdots \prec \mathcal{C}_1 \equiv \{[n]\}$$

https://doi.org/10.1239/aap/1127483740 Published online by Cambridge University Press

and has transition probabilities

$$\mathrm{P}(\mathcal{C}_{k-1} = \alpha_{k-1} \mid \mathcal{C}_k = \alpha_k) = \frac{2}{k(k-1)}, \qquad k = n, \ldots, 2.$$

Transitions happen if $\alpha_k \prec \alpha_{k-1}$ and α_k has exactly k classes; otherwise the transition probability is 0.

In the coalescent tree, a clade is an equivalence class $\alpha_\star$ for some $\mathcal{C}_i$, $i = n-1, \ldots, 1$, such that the time since the most recent common ancestor of the genes in $\alpha_\star$ is exactly S_{i+1}. In this notation, i corresponds to the number of ancestors present in the sample at the instant of coalescence.

3. Main results

In this article, we consider a specific realization of the random genealogy $C(t)$. This realization can hence be represented as a rooted tree with the genes at the tips and the most recent common ancestor of the sample at the root. Consider an arbitrary gene in the sample, and give to this gene the label 1. We then define the coalescence time of the lineage of gene 1 with the rest of the sample as

$$\tau_n = \sup\{t \geq 0 \colon \{1\} \text{ is an element of } C(t)\}.$$

The random variable τ_n corresponds to the length of a so-called external branch of the genealogy [7].

By definition, the smallest clade containing the gene 1 consists of the equivalence class α_1 in the partition $C(\tau_n)$, meaning that $\{1\} \subset \alpha_1$. In turn, this means that α_1 contains the reference lineage 1 at the instant of coalescence with the rest of the sample. In the sequel, α_1 is sometimes called the *minimal clade*. Our interest is in the size X_n of the minimal clade. Formally, X_n is then defined as

$$X_n = \mathrm{card}(\alpha_1), \qquad \alpha_1 \in C(\tau_n),\ \{1\} \subset \alpha_1.$$

The first result in this section gives the distribution of the random variable X_n.

Theorem 1. *Let $n \geq 2$. The random variable X_n has the probability distribution*

$$\mathrm{P}(X_n = x) = \frac{4}{(x-1)x(x+1)}, \qquad x = 2, \ldots, n-1,$$

and

$$\mathrm{P}(X_n = n) = \frac{2}{n(n-1)}.$$

The limiting distribution of X_n has mode 2 and is long tailed. This is an expected result because this behaviour is a typical feature of the number of species in a genus in traditional hierarchical phylogenetic taxonomy (see, e.g. [3]). In the Markov linear growth model or *Yule branching process*, this number, denoted X', is indeed distributed according to a *Yule law*,

$$\mathrm{P}(X' = x) = \rho\Gamma(1+\rho)\frac{\Gamma(x)}{\Gamma(x+1+\rho)}, \qquad x \geq 1,$$

https://doi.org/10.1239/aap/1127483740 Published online by Cambridge University Press

for some parameter $\rho > 0$ [23]. As n grows to infinity, the distribution of $M_n = X_n - 1$ converges to

$$\mathrm{P}(M = m) = \frac{4}{m(m+1)(m+2)}, \qquad m \geq 1,$$

and corresponds to the Yule distribution of parameter $\rho = 2$. It is noteworthy that the distribution of X_n coincides with the minimum of $M+1$ and n, and does not depend on n except through the event $(X_n = n)$. The small discrepancy between the finite and asymptotic probability values is due to the possibility that the reference lineage might connect at the root of the tree. As a corollary of Theorem 1, the expected size can be computed easily. We find that

$$\mathrm{E}[X_n] = 3 - \frac{2}{n}, \qquad n \geq 2,$$

and that the expected value converges to 3. In addition, the variance is equal to

$$\mathrm{var}[X_n] = 4 \sum_{i=3}^{n} \frac{1}{i} - 6 + o(1), \qquad n \geq 3,$$

which is of order $\log n$ and converges to infinity rather slowly.

Given a coalescent tree with n genes, we now consider the empirical frequencies of minimal clades of fixed size

$$f_n^x = \frac{\mathrm{card}(i : X_{i,n} = x)}{n},$$

where $X_{i,n} = \mathrm{card}(\alpha_i)$ and α_i is the equivalence class containing the gene i in $C(\tau_{i,n})$. (Here, the definition of $\tau_{i,n}$ is relative to gene i instead of gene 1.) The following result states that the distribution of f_n^x is approximatively Gaussian for large n.

Theorem 2. *When n goes to infinity, we have*

$$\sqrt{n}\left(f_n^x - \frac{4}{(x-1)x(x+1)}\right) \to \mathcal{N}(0, \sigma^2),$$

where $\sigma^2 = \frac{8}{45}$ if $x = 2$ and

$$\sigma^2 = \frac{4(11 + 4x^4 - 27x^2)}{x(2x+1)(2x-1)(x-1)^2(x+1)^2},$$

for all $x \geq 3$.

The case $x = 2$ is a direct consequence of McKenzie and Steel's results [11] because nf_n^2 is twice the number of cherries in a coalescent or Yule tree. (A cherry is a pair of genes whose lineages coalesce before sharing any ancestors with other genes.) McKenzie and Steel's result was derived from the analogy to extended Polya urns. Theorem 2 is based on a link to recent results in theoretical computer science regarding binary search trees. Binary search trees appear as formal representations for *divide-and-conquer* algorithms [13], [14]. The proof will exploit the one-to-one correspondence between binary search trees and coalescent trees, and use the stochastic recurrence equations involved in these data structures [2].

We will also superimpose a mutation process on the coalescent tree, and assume that DNA sequences are observed at the tips of the genealogical tree. The times at which the mutations occur are modelled as a Poisson process of constant rate $\frac{1}{2}\theta$, for some $\theta > 0$. If a branch of

https://doi.org/10.1239/aap/1127483740 Published online by Cambridge University Press

the tree has length t, then the number of mutations has a Poisson distribution with mean $\frac{1}{2}\theta t$, independently of the other branches. Among the various models that describe the mutation types, the infinitely many sites model may be one of the most appropriate [20]. In this model, each DNA sequence consists of completely linked sites (i.e. no recombination occurs). Each mutation occurs at a site of the DNA sequence that has not mutated previously, meaning that a new segregating site is produced. The number of segregating sites corresponds to the number of substitutions of ancestral bases since the most recent common ancestor in the sample.

Under the infinitely many sites assumption, Tajima [16] studied a sample of size 2. After observing ℓ substitutions, it follows from Bayes' theorem that the conditional distribution of the coalescence time is a gamma distribution gamma$(1+\ell, 1/(1+\theta))$ of shape $(1+\ell)$ and scale $1/(1+\theta)$, where

$$\text{gamma}(a,\lambda)(t) = \frac{\lambda^a}{\Gamma(a)} t^{a-1} \mathrm{e}^{-\lambda t}, \qquad t \geq 0.$$

In this article, we shall describe the conditional distribution of the coalescence time τ_n given that $\Delta = \ell$ substitutions are observed. In this notation, Δ is the number of substitutions found when comparing the DNA sequence of the gene 1 to that of a closest parent in the coalescent tree. The conditional distribution can be formulated as a mixture of gamma distributions. For $k = 2, \ldots, n-1$ and $k \leq j \leq n$, let us write

$$c(j,k) = \prod_{j \leq \ell \neq k \leq n} (\ell(\ell-1) - k(k-1)).$$

In addition, we set $c(n,n) = 1$. Define

$$a_k = \frac{(n-1)!\,(n-2)!}{\lambda_k} \sum_{j=2}^{k} \frac{c(j,k)^{-1}}{((j-2)!)^2} \qquad \text{for } k = 2, \ldots, n.$$

Let $\mathcal{G}_{\mathrm{s}}(p,\cdot)$ be the shifted geometric distribution defined as follows:

$$\mathcal{G}_{\mathrm{s}}(p,\ell) = p(1-p)^{\ell}, \qquad \ell = 0, 1, \ldots.$$

Then, for the conditional probability density function of the coalescence time, we have

$$f_{\tau_n \mid \Delta=\ell}(t) = \sum_{k=2}^{n} a(k,\ell)\,\text{gamma}(1+\ell, \lambda_k + \theta)(t), \qquad t \geq 0,$$

where

$$a(k,\ell) = \frac{a_k \mathcal{G}_{\mathrm{s}}(p_k,\ell)}{\mathrm{P}(\Delta=\ell)}, \qquad \ell = 0, 1, \ldots,$$

and

$$\mathrm{P}(\Delta=\ell) = \sum_{k=2}^{n} a_k \mathcal{G}_{\mathrm{s}}(p_k,\ell), \qquad \ell = 0, 1, \ldots,$$

with $\mathcal{G}_{\mathrm{s}}(p_k,\cdot)$ the shifted geometric distribution of parameter

$$p_k = \frac{1}{1+\theta/\lambda_k}, \qquad k = 2, \ldots, n.$$

https://doi.org/10.1239/aap/1127483740 Published online by Cambridge University Press

4. The size of the minimal clade

4.1. Level of coalescence with the rest of the sample

The coalescence level K of the gene 1 with the rest of the sample is defined as a random variable

$$K = k \quad \text{if and only if} \quad \tau_n = S_k,$$

with $k = n, n-1, \ldots, 2$. More specifically, K is related to the ancestral process as follows:

$$K = 1 + A_n(\tau_n).$$

We give the distribution of K below.

Proposition 1. *For $n \geq 2$, we have*

$$\mathrm{P}(K = k) = \frac{2(k-1)}{n(n-1)} \quad \textit{for all } k = 2, \ldots, n.$$

Proof. Recall that, for all $k = 2, \ldots, n$, we have

$$\mathrm{P}(\mathcal{C}_k = \alpha) = \frac{(n-k)!\,k!\,(k-1)!}{n!\,(n-1)!} n_1! \cdots n_k!, \tag{1}$$

where $\alpha \in \mathcal{E}_n$ is a partition in k classes, such that each class has cardinality $\mathrm{card}(\alpha_i) = n_i$ and $n_1 + \cdots + n_k = n$. In the sequel, we shall use the notation $|\alpha| = k$ for the number of classes of a partition. For a proof of this classical result, see [10], [9], or [17, p. 39, Proposition 2.2.2].

For $k = 2, \ldots, n$, the probability that the coalescence occurs at a level of the genealogy lower than k is equal to

$$\mathrm{P}(K \leq k) = \mathrm{P}(\{1\} \in \mathcal{C}_k).$$

To compute the probability of this event, we can write

$$\mathrm{P}(K \leq k) = \sum_{\alpha\,:\,|\alpha|=k-1} \mathrm{P}(\mathcal{C}_k = \{1\} \cup \alpha),$$

where the sum runs over all partitions of $\{2, \ldots, n\}$ into $(k-1)$ classes. Using (1), we find that

$$\mathrm{P}(K \leq k) = \sum_{\alpha\,:\,|\alpha|=k-1} \frac{(n-k)!\,k!\,(k-1)!}{n!\,(n-1)!} n_1! \cdots n_{k-1}!$$

and, since $n_1 + \cdots + n_{k-1} = n-1$, we have

$$\sum_{\alpha\,:\,|\alpha|=k-1} \frac{(n-k)!\,(k-1)!\,(k-2)!}{(n-1)!\,(n-2)!} n_1! \cdots n_{k-1}! = 1.$$

Therefore, we find that

$$\mathrm{P}(K \leq k) = \frac{k(k-1)}{n(n-1)}, \qquad k = 2, \ldots, n,$$

which yields the desired result.

https://doi.org/10.1239/aap/1127483740 Published online by Cambridge University Press

Note that the distribution of K could also be obtained in a less direct way using [22, p. 188, Corollary 2]. The mean and the variance of K can be computed from elementary algebra. We find that the expectation is equal to

$$\mathrm{E}[K] = \tfrac{2}{3}(n+1), \qquad n \geq 2.$$

In order to compute the variance, recall that

$$\sum_{k=2}^{n} k^3 = \tfrac{1}{4}((n+1)^4 - 2(n+1)^3 + (n+1)^2) - 1.$$

We then have

$$\mathrm{var}[K] = \tfrac{1}{18}(n^2 - n - 2).$$

The interpretation is that the average level at which the coalescence occurs is closer to the tips of the tree than to the root. However, the variance is relatively large with respect to the mean for large sample sizes.

4.2. Minimal clade size

This section deals with the size X_n of the minimal family of an arbitrary individual in the sample. In it, we give a proof that the random variable X_n has a power law distribution

$$\mathrm{P}(X_n = x) = \frac{4}{(x-1)x(x+1)}, \qquad x = 2, \ldots, n-1,$$

with

$$\mathrm{P}(X_n = n) = \frac{2}{n(n-1)}.$$

Before giving the proof of Theorem 1, we establish a useful combinatorial identity in the next lemma.

Lemma 1. *Let $n \geq 4$ and let x be an integer such that $n > x \geq 3$. We then have*

$$\sum_{k=x-2}^{n-3} k(k-1)\cdots(k-x+3)(n-k-1)(n-k-2) = 2\frac{n(n-1)\cdots(n-x)}{(x-1)x(x+1)}.$$

Proof. First, use induction to prove that

$$\sum_{k=x-2}^{n-1} k(k-1)\cdots(k-x+3) = \frac{n(n-1)\cdots(n-x+2)}{x-1},$$

and then use a similar recursion argument to prove that

$$\sum_{k=x-2}^{n-2} k(k-1)\cdots(k-x+3)(n-k-1) = \frac{n(n-1)\cdots(n-x+1)}{x(x-1)}.$$

By applying the recursion again, we obtain the lemma from the above equations.

https://doi.org/10.1239/aap/1127483740 Published online by Cambridge University Press

Proof of Theorem 1. Consider the coalescent starting with n lineages. A standard result in coalescent theory states that if we pick one lineage at random when there are $k \leq n$ lineages, then the probability it will contain m of the n starting lineages is

$$\mathrm{P}(M_k^n = m) = \binom{n-m-1}{k-2}\binom{n-1}{k-1}^{-1}, \qquad 1 \leq m \leq n-k+1.$$

For a proof of this result, see, e.g. [6, Chapter 1, Equation (3.14)].

At the moment of the coalescence with the rest of the sample, the lineage of individual 1 coalesces with a random subset of size M^{n-1}. Conditional on $K = k$, this means that X_n has the same distribution as $1 + M_{k-1}^{n-1}$, i.e.

$$X_n \sim 1 + M_{k-1}^{n-1},$$

where the coalescent starts with $(n-1)$ lineages in M_{k-1}^{n-1}. Then, for $k = 3, \ldots, n$, we have

$$\mathrm{P}(X_n = 1 + m \mid K = k) = \binom{n-m-2}{k-3}\binom{n-2}{k-2}^{-1}, \qquad m = 1, \ldots, n-k+1,$$

and, for $k = 2$, we have

$$\mathrm{P}(X_n = n \mid K = 2) = 1.$$

Then we have

$$\mathrm{P}(X_n = n) = \frac{2}{n(n-1)}$$

and, for all $m = 1, \ldots, n-2$,

$$\mathrm{P}(X_n = 1 + m) = \sum_{k=3}^{n-m+1} \frac{2(k-1)}{n(n-1)}\binom{n-m-2}{k-3}\binom{n-2}{k-2}^{-1}.$$

For $m = 1$, we obtain

$$\mathrm{P}(X_n = 2) = \sum_{k=3}^{n} \frac{2(k-1)(k-2)}{n(n-1)(n-2)},$$

which is equal to $\frac{2}{3}$. Similarly, for $n > x \geq 3$, we obtain

$$\mathrm{P}(X_n = x) = \sum_{k=x-2}^{n-3} \frac{2k(k-1)\cdots(k-x+3)(n-k-1)(n-k-2)}{n(n-1)(n-2)\cdots(n-x)}.$$

Using Lemma 1, we find that

$$\mathrm{P}(X_n = x) = \frac{4}{(x-1)x(x+1)}$$

for all $x = 2, \ldots, n-1$.

We now turn to the proof of Theorem 2. Given a coalescent tree, we compute the number of subtrees with x leaves (genes) and a lineage that connects to the root of the subtree. If we

https://doi.org/10.1239/aap/1127483740 Published online by Cambridge University Press

divide it by n, this leads to f_n^x, which is an unbiased estimate of the probability $\mathrm{P}(X_n = x)$ for all $x = 2, \ldots, n$. When n goes to infinity, we obtain a Gaussian central limit theorem

$$\sqrt{n}\left(f_n^x - \frac{4}{(x-1)x(x+1)}\right) \to \mathcal{N}(0, \sigma^2),$$

where $\sigma^2 = \frac{8}{45}$ if $x = 2$ and

$$\sigma^2 = \frac{4(11 + 4x^4 - 27x^2)}{x(2x+1)(2x-1)(x-1)^2(x+1)^2} \tag{2}$$

for all $x \geq 3$.

Proof of Theorem 2. Let $X_n^x = nf_n^x$ denote the number of minimal clades of size x. The case $x = 2$ is a direct consequence of McKenzie and Steel's results [11], because X_n^2 is twice the number of cherries in a coalescent or Yule tree. For $x \geq 3$, the proof follows from the fact that the random variable X_n^x has a quicksort-like recurrence equation [8]

$$X_n^x = X_{I_n}^x + X_{n-I_n}^{x*} + t_n^x, \tag{3}$$

where I_n is uniform over the set $\{1, \ldots, n-1\}$ and the *toll function* t_n^x is equal to

$$t_n^x = \delta_{n,x}(\delta_{I_n,1} + \delta_{I_n,n-1}),$$

where δ denotes the Kronecker symbol. The expression for σ^2 is found by taking the variance of both sides of (3). The induction part follows from an analytic lemma [8, Lemma 1] and a symbolic algebra computer package. For $n \geq 2x + 1$, we find that $\mathrm{var}[X_n^x] = \sigma^2 n$, with σ^2 given by (2). The final result is a consequence of Hwang and Neininger's classification of toll functions [8].

Remarks. Note that the proof of Theorem 2 contains an implicit proof of Theorem 1. Taking expectations in (3) leads to recursive identities that can also be solved using [8, Lemma 1]. After short calculations, the results of Theorem 1 can again be recovered.

5. Some comparisons

5.1. Two random genes

For the sake of comparison, we will also describe the distribution of the level of coalescence for two arbitrarily chosen genes. The two genes can be labelled 1 and 2. Let us denote their coalescence time by $\tilde{\tau}_n$, and set $\tilde{K} = 1 + A_n(\tilde{\tau}_n)$. We wish to compare $\tilde{K}$ to K. A well-known result in coalescence theory is that the coalescence time $\tilde{\tau}_n$ of two lineages has exponential distribution Exp(1). In Section 6, we will describe the result for τ_n. As far as the coalescence of two random lineages in the sample is concerned, we have the following probability distribution: for $n \geq 2$,

$$\mathrm{P}(\tilde{K} = k) = \frac{n+1}{n-1}\frac{2}{k(k+1)} \qquad \text{for all } k = 2, \ldots, n.$$

The argument is based on the bivariate coalescent (see, e.g. [18, p. 87]) and a result of Saunders *et al.* [15] that describes the joint distribution of

$$B(t) = (A_m(t), A_n(t)), \qquad t \geq 0,$$

https://doi.org/10.1239/aap/1127483740 Published online by Cambridge University Press

where $A_n(t)$ is the ancestral process at time t and $A_m(t)$ is the ancestral process of a subsample of size $m \leq n$. Using [15], we find that

$$\mathrm{P}(A_2(t) = 1,\ A_n(t) = k-1) = \mathrm{P}(A_n(t) = k-1)\,\frac{2(n-k+1)}{k(n-1)}, \qquad n \geq k \geq 2.$$

For $2 \leq k \leq n-1$, we have

$$\mathrm{P}(\tilde{\tau}_n \leq S_k) = \mathrm{P}(A_2(S_k) = 1) =: F(k) = \frac{2}{k}\frac{n-k+1}{n-1}$$

and, so,

$$\mathrm{P}(\tilde{K} = k) = \mathrm{P}(\tilde{\tau}_n = S_k) = F(k) - F(k+1) = \frac{n+1}{n-1}\frac{2}{k(k+1)}.$$

For $k = n$, we have

$$\mathrm{P}(\tilde{K} = n) = \mathrm{P}(\tilde{\tau}_n \leq S_n) = F(n) = \frac{2}{n(n-1)}.$$

Note that the probability that two individuals share their most recent ancestor with the whole sample was calculated by Watterson [21], whose result agrees with the fact that

$$\mathrm{P}(\tilde{K} = 2) = \frac{1}{3}\frac{n+1}{n-1}.$$

In conclusion, the distribution of $\tilde{K}$ is very different from that of K. The nodes close to the root are given more important weights in the distribution of K. This can also be seen from the average level $\mathrm{E}[\tilde{K}]$, which is $O(\log n)$ in contrast with the $O(n)$ result obtained for $\mathrm{E}[K]$.

5.2. A random clade

A useful comparison to Theorem 1 may be given by the distribution of the number of individuals in a random clade of the coalescent tree. Choose $I = i$ from the uniform distribution on $[n-1]$ and consider the number of genes Y_n in the (unique) clade of $\mathcal{C}_I$. We have the following result.

Proposition 2. *Let $n \geq 2$ and Y_n be the number of individuals in a random clade of the coalescent. We have*

$$\mathrm{P}(Y_n = y) = \frac{n}{n-1}\frac{2}{y(y+1)} \quad \textit{for all } y = 2, \ldots, n-1,$$

with

$$\mathrm{P}(Y_n = n) = \frac{1}{n-1}.$$

Proof. Using arguments similar to those of Theorem 1, we deduce the conditional distribution of Y_n given that $I = i$, for $i = 2, \ldots, n-1$. We obtain

$$\mathrm{P}(Y_n = y \mid I = i) = (y-1)\binom{n-y-1}{i-2}\binom{n-1}{i}^{-1}, \qquad y = 2, \ldots, n-i+1.$$

Hence, for $2 \leq y \leq n-2$, we have

$$\mathrm{P}(Y_n = y) = \frac{n(y-1)}{(n-1)}\sum_{j=y-1}^{n-2}\frac{(j-1)(j-2)\cdots(j-y+2)(n-j)(n-j-1)}{n(n-1)\cdots(n-y)},$$

which yields the result.

https://doi.org/10.1239/aap/1127483740 Published online by Cambridge University Press

This distribution can be found in a different manner by using results on binary search trees introduced in theoretical computer science. Aldous [1] and Devroye [4] described the number of occurrences of subtrees of a given size. The above proposition is strongly connected to their results.

Let us remark that the average size of a random clade grows as $\log n$, i.e.

$$\mathrm{E}[Y_n] = \frac{2n}{n-1}(H_n - 1), \qquad n \geq 3,$$

where H_n is the nth harmonic number, while $\mathrm{E}[X_n]$ remains bounded by 3. The variance of Y_n grows as n. The asymptotic distribution of Y_n is given by

$$\mathrm{P}(Y = y) = \frac{2}{y(y+1)}, \qquad y \geq 2,$$

which corresponds to the conditional Yule distribution of parameter $\rho = 1$, given that the number of species is greater than 2. According to Devroye's result and the correspondence with binary search trees, the frequencies of subtrees in a coalescent tree are asymptotically Gaussian for large n. Denoting by f_n^y the frequency of subtrees of size y, $2 \leq y \leq n$, we have

$$\sqrt{n}\left(f_n^y - \frac{2}{y(y+1)}\right) \to \mathcal{N}(0, \sigma^2),$$

where the convergence holds in distribution. Modifying Devroye's result, we obtain

$$\sigma^2 = \frac{2(y-1)(4y^2 - 3y - 4)}{y(2y+1)(2y-1)(y+1)^2}, \qquad y \geq 2.$$

6. External branch lengths

6.1. Unconditional distribution

The main result of this section is the description of the distribution of the coalescence time τ_n. This random variable corresponds to the length of an external branch, in the terminology of Fu and Li [7]. Before giving the distribution of τ_n, we remark that the mean and variance of this random variable follow from Proposition 1. Fu and Li [7] provided a different proof for these results. Here, we use the fact that

$$\tau_n = T_n + \cdots + T_K.$$

Proposition 3. *Let $n \geq 2$. Consider the coalescence time τ_n. We have*

$$\mathrm{E}[\tau_n] = \frac{2}{n} \quad \textit{and} \quad \mathrm{var}[\tau_n] = \frac{4}{n^2}.$$

Now recall that, for $k = 2, \ldots, n-1$ and $k \leq j \leq n$,

$$c(j,k) = \prod_{j \leq \ell \neq k \leq n} (\ell(\ell-1) - k(k-1)), \qquad c(n,n) = 1.$$

We have the following result.

https://doi.org/10.1239/aap/1127483740 Published online by Cambridge University Press

Theorem 3. *Let $n \geq 2$. The Laplace transform of τ_n is given by*

$$L_{\tau_n}(s) = \mathrm{E}[\mathrm{e}^{-s\tau_n}] = \sum_{k=2}^{n} a_k \frac{\lambda_k}{s+\lambda_k}, \qquad s \geq 0,$$

where

$$a_k = \frac{(n-1)!\,(n-2)!}{\lambda_k} \sum_{j=2}^{k} \frac{c(j,k)^{-1}}{((j-2)!)^2}. \tag{4}$$

The probability density function can be described as a mixture of exponential distributions:

$$f_{\tau_n}(t) = \sum_{k=2}^{n} a_k \lambda_k \mathrm{e}^{-\lambda_k t}, \qquad t \geq 0.$$

Proof. According to Proposition 1, we have

$$\mathrm{E}[\mathrm{e}^{-s\tau_n}] = \sum_{k=2}^{n} \frac{2(k-1)}{n(n-1)} \prod_{j=k}^{n} \frac{\lambda_j}{s+\lambda_j}.$$

Using fractional decomposition, we have

$$\prod_{j=k}^{n} \frac{1}{s+\lambda_j} = \sum_{j=k}^{n} \frac{2^{n-k-1} c(k,j)^{-1}}{s+\lambda_j}.$$

Let

$$b(k,j) = c(k,j)^{-1} \frac{k-1}{n(n-1)} \prod_{\ell=k}^{n} \ell(\ell-1).$$

Reordering the sums, we find that

$$\sum_{k=2}^{n} \sum_{j=k}^{n} \frac{b(k,j)}{s+\lambda_j} = \sum_{k=2}^{n} \left(\sum_{j=2}^{k} b(j,k) \right) \frac{1}{s+\lambda_k}$$

and

$$\lambda_k a_k = \sum_{j=2}^{k} b(j,k).$$

6.2. Conditional distributions

We now study the number of substitutions Δ in the DNA sequence of an arbitrary gene, compared with the sequence of a closest relative. Recall that, for two arbitrary genes, the number of pairwise differences has a shifted geometric distribution of parameter $p = 1/(1+\theta)$. The mean is θ and the variance is $\theta + \theta^2$. We obtain the following result.

Proposition 4. *Let $n \geq 2$ and assume that the infinitely many sites model of mutation in the coalescent is being used. Then the number of substitutions Δ between a gene and a closest relative is distributed as a mixture of shifted geometric distributions, i.e.*

$$\mathrm{P}(\Delta = \ell) = \sum_{k=2}^{n} a_k \mathcal{G}_{\mathrm{s}}(p_k, \ell), \qquad \ell = 0, 1, \ldots,$$

https://doi.org/10.1239/aap/1127483740 Published online by Cambridge University Press

where a_k is as given in (4),

$$\mathcal{G}_{\mathrm{s}}(p_k, \ell) = (1 - p_k)^{\ell} p_k, \qquad \ell = 0, 1, \ldots,$$

and

$$p_k = \frac{1}{1 + \theta/\lambda_k}, \qquad k = 2, \ldots, n.$$

Proof. Let $\theta > 0$. Conditional on $\tau_n = t$, $t \geq 0$, we have

$$\mathrm{P}(\Delta = \ell \mid \tau_n = t) = \frac{\theta^{\ell} t^{\ell}}{\ell!} \mathrm{e}^{-\theta t}, \qquad \ell = 0, 1, \ldots,$$

and, so,

$$\mathrm{P}(\Delta = \ell) = \frac{(-1)^{\ell} \theta^{\ell}}{\ell!} L_{\tau_n}^{(\ell)}(\theta),$$

where $L_{\tau_n}^{(\ell)}$ is the ℓth derivative of the Laplace transform L_{τ_n}. The proof then follows from Theorem 3.

The moments of Δ were found by Fu and Li using a different method [7]: we find that

$$\mathrm{E}[\Delta] = \frac{2}{n}\theta \quad \text{and} \quad \mathrm{var}[\Delta] = \frac{2}{n}\theta + \frac{4}{n^2}\theta^2,$$

using the facts that

$$\mathrm{E}[\tau_n] = \sum_{k=2}^{n} \frac{a_k}{\lambda_k} = \frac{2}{n} \quad \text{and} \quad \mathrm{E}[\tau_n^2] = 2 \sum_{k=2}^{n} \frac{a_k}{\lambda_k^2} = \frac{8}{n^2}.$$

The conditional distribution of the coalescence time τ_n, given that ℓ substitutions are observed, can be deduced from the Bayes formula as follows:

$$f_{\tau_n \mid \Delta = \ell}(t) = \frac{\theta^{\ell}}{\ell!\, \mathrm{P}(\Delta = \ell)} t^{\ell} \mathrm{e}^{-\theta t} f_{\tau_n}(t), \qquad t \geq 0,$$

Using Proposition 4, the conditional density can be reformulated as a mixture of gamma distributions, i.e.

$$f_{\tau_n \mid \Delta = \ell}(t) = \sum_{k=2}^{n} a(k, \ell)\, \mathrm{gamma}(1 + \ell, \lambda_k + \theta)(t), \qquad t \geq 0,$$

where, for $k = 2, \ldots, n$,

$$a(k, \ell) = \frac{a_k \mathcal{G}_{\mathrm{s}}(p_k, \ell)}{\mathrm{P}(\Delta = \ell)}, \quad \ell = 0, 1, \ldots.$$

In Figure 1, we display the curves of f_{τ_n} and $f_{\tau_n \mid \Delta = \ell}$ for $\ell = 0, 1, 2, 5$ and $\theta = 10$. In Table 1, we report the values of $\mathrm{P}(\Delta = \ell)$ for $n = 10, 30$, $\ell = 0, \ldots, 9$ and $\theta = 1, 10$. Exact computations of the conditional expectation are reported in Table 2 for $\theta = 1, 10$ and $\ell = 0, \ldots, 10$. In order to provide numerical values, we used the following formula:

$$\mathrm{E}[\tau_n \mid \Delta = \ell] = \frac{(1 + \ell)}{\theta} \frac{\mathrm{P}(\Delta = \ell + 1)}{\mathrm{P}(\Delta = \ell)}, \qquad \ell = 0, 1, \ldots.$$

https://doi.org/10.1239/aap/1127483740 Published online by Cambridge University Press

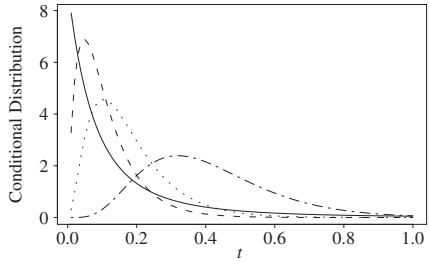

FIGURE 1: The unconditional coalescence time distribution (*solid*) and the conditional distributions, given that $\Delta = \ell = 1, 2, 5$ segregating sites are observed. The mode of the conditional distributions increases with Δ. The sample size is $n = 10$ and the mutation rate is $\theta = 10$.

TABLE 1: Probability distribution of the number of segregating sites Δ. Where not shown, values are below 0.001. The last row gives the sum of the ten probabilities.

	$\theta = 1$		$\theta = 10$	
Δ	$n = 10$	$n = 30$	$n = 10$	$n = 30$
0	0.852	0.942	0.426	0.673
1	0.114	0.050	0.218	0.193
2	0.022	0.005	0.120	0.067
3	0.006	0.001	0.071	0.028
4	0.001	—	0.044	0.013
5	—	—	0.029	0.007
6	—	—	0.020	0.004
7	—	—	0.014	0.002
8	—	—	0.010	0.001
9	—	—	0.007	0.001
Sum	0.999	0.999	0.963	0.994

TABLE 2: Conditional expectations of the coalescence time, given the number of segregating sites $\Delta = 0, \ldots, 10$.

	$\theta = 1$		$\theta = 10$	
Δ	$n = 10$	$n = 30$	$n = 10$	$n = 30$
0	0.13	0.05	0.05	0.02
1	0.39	0.20	0.11	0.07
2	0.84	0.59	0.17	0.12
3	1.46	1.23	0.25	0.19
4	2.12	1.98	0.32	0.27
5	2.76	2.68	0.41	0.35
6	3.36	3.31	0.49	0.44
7	3.92	3.89	0.58	0.53
8	4.45	4.44	0.67	0.63
9	4.97	4.96	0.76	0.72
10	5.48	5.48	0.86	0.82

https://doi.org/10.1239/aap/1127483740 Published online by Cambridge University Press

We remark that the distribution of Δ is concentrated on small integers, with a rapid decrease as the number of substitutions increases. Also, larger sample sizes lead to more concentrated distributions. On the other hand, the conditional expectations become rather independent of the sample size as the number of substitutions increases.

6.3. Application: multilocus haplotypes

In this section, we compute posterior distributions for the time to the most recent common ancestor for a non-recombining segment of DNA and its closest relative in a sample of $n-1$ other sequences of the same segment, given that they match at ℓ out of m scored markers. The results presented here extend those of Walsh for two segments [19].

We consider m completely linked markers, and score their allelic states, assuming a perfect match if and only if no mutation has occurred since the most recent common ancestor. For $\ell = 0, \ldots, m$, the conditional probability of observing ℓ matches out of m is then binomial, i.e. for $t > 0$,

$$p(\ell \mid \tau_n = t) = \frac{n!}{\ell!\,(m-\ell)!} \mathrm{e}^{-\theta \ell t} (1 - \mathrm{e}^{-\theta t})^{m-\ell},$$

where $\mathrm{e}^{-\theta t}$ is the probability of a perfect match at one locus. Using the Bayes formula, we obtain the conditional density of τ_n given that ℓ matches are observed: for $t > 0$,

$$p_{\tau_n}(t \mid \ell) \propto \mathrm{e}^{-\theta \ell t} (1 - \mathrm{e}^{-\theta t})^{m-\ell} f_{\tau_n}(t).$$

The normalizing constant can be computed using a symbolic algebra package, or deduced from [19]. We find that

$$p(\tau_n = t \mid \ell) = \frac{\sum_{k=2}^{n} a_k \lambda_k \binom{m}{\ell} \mathrm{e}^{-(\theta \ell + \lambda_k)t} (1 - \mathrm{e}^{-\theta t})^{m-\ell}}{\sum_{k=2}^{n} a_k \lambda_k \binom{m}{\ell} I(m, \ell, \theta, k)},$$

where

$$I(m, \ell, \theta, k) = \frac{\theta^{m-\ell}(m-\ell)!}{\prod_{i=0}^{m-\ell} (\lambda_k + \theta(m-i))^{-1}}.$$

Given a perfect match at $\ell = m$ markers, we find that

$$p_{\tau_n}(t \mid m) = \frac{\sum_{k=2}^{n} a_k \lambda_k \mathrm{e}^{-(\theta n + \lambda_k)t}}{\sum_{k=2}^{n} a_k \lambda_k (\lambda_k + \theta n)^{-1}}. \tag{5}$$

Walsh [19] used a mutation rate per generation equal to $\mu = \frac{1}{500}$, motivated by estimates on the human Y chromosome, and the effective size was estimated as $N_\mathrm{e} \approx 5000$. These values led to an estimate of $\theta = 20$, using the fact that $\theta = 2N_\mathrm{e}\mu$. Considering a perfect match at $m = 20$ markers and $n = 2$ individuals, the 95% Bayesian credible region was computed to be $(6 \times 10^{-5}, 0.009\,22)$, which corresponded to an interval of (0.3, 46.1) generations for the most recent common ancestor. One conclusion was that the forensic use of the Y chromosome is rather limited [19]. Here, we reexamine the upper bound of the 95% Bayesian credible region, given $n = 40, 60, 80, 100$ individuals. From (5), we find in each case that the upper bound decreases to $\lfloor t N_\mathrm{e} + 1 \rfloor = 40, 30, 8, 7$ generations, respectively. Using $m = 100$ markers, these numbers further reduce to $\lfloor t N_\mathrm{e} + 1 \rfloor = 9, 7, 5, 4$ generations, respectively.

https://doi.org/10.1239/aap/1127483740 Published online by Cambridge University Press

References

[1] Aldous, D. J. (1991). Asymptotic fringe distributions for general families of random trees. *Ann. Appl. Prob.* **1,** 228–266.

[2] Aldous, D. J. (1996). Probability distributions on cladograms. In *Random Discrete Structures*, eds D. J. Aldous and R. Pemantle, Springer, Berlin, pp. 1–18.

[3] Aldous, D. J. (2001). Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statist. Sci.* **16,** 23–34.

[4] Devroye, L. (1991). Limit laws for local counters in random binary search trees. *Random Structures Algorithms* **2,** 303–315.

[5] Donnelly, P., Tavaré, S., Balding, D. J. and Griffiths, R. C. (1996). Estimating the age of the common ancestor of men from the ZFY intron. *Science* **272,** 1357–1359.

[6] Durrett, R. (2003). *Probabilistic Models of DNA Sequences*. Springer, New York.

[7] Fu, Y. X. and Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics* **133,** 693–709.

[8] Hwang, H.-K. and Neininger, R. (2002). Phase change of limit laws in the quicksort recurrence under varying toll functions. *SIAM J. Comput.* **31,** 1687–1722.

[9] Kingman, J. F. C. (1982). On the genealogy of large populations. In *Essays in Statistical Science* (J. Appl. Prob. Spec. Vol. **19A**), Applied Probability Trust, Sheffield, pp. 27–43.

[10] Kingman, J. F. C. (1982). The coalescent. *Stoch. Process. Appl.* **13,** 235–248.

[11] McKenzie, A. and Steel, M. (2000). Distributions of cherries for two models of trees. *Math. Biosci.* **164,** 81–92.

[12] Nordborg, M. (2001). Coalescent theory. In *Handbook of Statistical Genetics*, eds D. J. Balding *et al*., John Wiley, New York, pp. 179–208.

[13] Régnier, M. (1989). A limiting distribution for quicksort. *RAIRO Inf. Théor. Appl.* **23,** 335–343.

[14] Rösler, U. (1992). A limit theorem for quicksort. *RAIRO Inf. Théor. Appl.* **25,** 85–100.

[15] 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.

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

[17] Tavaré, S. (2004). Ancestral inference in population genetics. In *Lectures on Probability Theory and Statistics* (Lecture Notes Math. **1837**), Springer, Berlin, pp. 1–188.

[18] Tavaré, S. (1997). Ancestral inference from DNA sequence data. In *Case Studies in Mathematical Modeling in Ecology, Physiology and Cell Biology*, eds H. G. Othmer *et al*., Prentice Hall, Upper Saddle River, NJ, pp. 81–96.

[19] Walsh, B. (2001). Estimating the time to the most recent common ancestor for the Y chromosome or mitochontrial DNA for a pair of individuals. *Genetics* **158,** 897–912.

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

[21] Watterson G. A. (1982). Mutant substitutions at linked nucleotide sites. *Adv. Appl. Prob.* **14,** 206–224.

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

[23] Yule, G. U. (1924). A mathematical theory of evolution, based on the conclusions of Dr J. C. Willis. *Philos. Trans. R. Soc. London B* **213,** 21–87.