Minimal clade size and external branch length under the neutral coalescent

Michael G.B. Blum†‡ Olivier François†*

† Laboratory TIMC-TIMB, Institute for Information and Health Engineering, Faculty of Medicine, F38706 La Tronche cedex, France

‡ Laboratory of Ecology, Systematics and Evolution, Bat. 360, University Paris-Sud, 91405 Orsay Cedex, France

*Corresponding author e.mail: Olivier.Francois@imag.fr

Abstract

Given a sample of genes taken from a large population, we consider the neutral coalescent genealogy, and we study the theoretical and empirical distributions of the size of the smallest clade which contains a fixed gene. We show that the theoretical distribution is strongly related to a Yule distribution of parameter two, 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 we 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.

1 Introduction

Coalescent theory associated with molecular information has proven an invaluable tool for assessing the degree of relatedness between individuals. To this 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.* [6] estimated the time since the most recent common ancestor of modern humans from DNA sequences of the ZFY intron.

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, non recombinanting genes will be considered. 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 [9]. The approximation arises as a diffusion limit when time is measured in units of N generations. See [17, 5] 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 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 (or Zipf) distribution of parameter 2. Given the genealogy, we shall also stufy 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, we study the amount of relatedness of a gene with the rest of the sample. A natural measure for non-recombinating regions is given by the time since the most recent ancestor. This time τ_n corresponds to the length of an external branch of the coalescent tree [7]. We shall describe the distribution of the number of lineages present in the sample at the moment of coalescence, and establish formulas for the Laplace transform and the probability density of τ_n. Using the *infinitely-many sites model* of the DNA molecule, we shall also describe the conditional distribution of the coalescence time given the number of substitutions between the reference sequence and one of its closest relative. This distribution corresponds to an explicit mixture of gamma distributions, and extends results of [16] as well. An application to the human chromosome Y is presented at the end of the article.

2 Background and Notations

Consider a sample of n genes. In the coalescent, one wishes to record information about the number of ancestors at various times and also information about which genes share common ancestors. The *ancestral process* $A_n(t)$ records the number of distincts ancestors of the sample a time t in the past.

It can be described as a continous-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

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

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 denote

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

To understand the topology of the tree, one possibility is labelling genes in the sample from the set $[n]$ and defining a random equivalence relation. In this relation, the genes i and j are in the same class at time t iff they share a common ancestor at this time. Denote by $C(t)$ the random partition obtained from the former equivalence relation at time t. According to Kingman [9], 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

$$\text{for all } \alpha, \beta \in \mathcal{E}_n, \quad q_{\alpha\beta} = \begin{cases} 1 & \text{if } \alpha \prec \beta \\ 0 & \text{otherwise,} \end{cases}$$

where the symbol $\alpha \prec \beta$ means that α is a nested partition of β which may be obtained by merging two classes in α. 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

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

and has transition probabilities

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

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

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 S_{i+1} exactly. In these notations, 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. Now, we define the coalescence time of the lineage 1 with the rest of the sample as follows

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

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

By definition, the smallest clade which contains the gene 1 consists of the equivalence class α_1 in the partition $C(\tau_n)$ so that $\{1\} \subset \alpha_1$. 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

$$X_n = \#\alpha_1, \quad \alpha_1 \in C(\tau_n), \quad \{1\} \subset \alpha_1.$$

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 following probability distribution*

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

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 eg [3]). In the Markov linear growth model or *Yule* branching process, this number is indeed distributed according to a *Yule* law

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

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

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

and corresponds to the Yule distribution of parameter $\rho = 2$. This is noteworthy that the distribution of X_n coincide 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 created by the chance that the reference lineage connects at the root of the tree. As a corollary of Theorem 1, the expected size can be computed easily. We find that

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

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

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

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

$$f_n^x = \frac{\#\{i\,;\, X_{i,n} = x\}}{n}$$

where $X_{i,n} = \#\alpha_i$ and α_i is the equivalence class which contains the gene i in $C(\tau_{i,n})$. (Here the definition of $\tau_{i,n}$ is relative to i instead of 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*

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

where $\sigma^2 = 8/45$, if $x = 2$, and

$$\sigma^2 = \frac{4(11 + 4x^4 - 27x^2)}{x(2x+1)(2x-1)(x-1)^2(x+1)^2}, \quad \text{for all } x \geq 3.$$

The case $x = 2$ is a direct consequence of McKenzy and Steel's results [11] because nf_n^2 is twice the number of cherries in a coalescent or Yule tree. McKenzy and Steel's result was derived from the analogy with 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].

Next, a mutation process is superimposed on the coalescent tree, and we assume that DNA sequences are observed at the tips of the genealogical tree. The times at which the mutations occur are modeled as a Poisson process of constant rate $\theta/2$, for some $\theta > 0$. If a branch of the tree has length t, then the number of mutations has a Poisson distribution with mean $\theta t/2$, 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 (ie, no recombination occurs). Each mutation occurs at a site of the DNA sequence that had not been mutated before, so that a new segregating site arises. 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 two. After observing ℓ substitutions, it follows from Bayes Theorem that the conditional distribution of the coalescence time is a Gamma distribution of shape $(1+\ell)$ and scale $1/(1+\theta)$, $G(1+\ell, 1/(1+\theta))$, where

$$G(a,\lambda)(t) = \frac{\lambda^a}{\Gamma(a)} t^{a-1} e^{-\lambda t}, \quad t \geq 0.$$

In this article, we shall describe the conditional distribution of the coalescence time τ_n given that $\Delta = \ell$ substitutions are observed. In this notations, Δ is the number of substitutions found when comparing the DNA sequence of the gene 1 and the one 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 denote

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

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

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

Then, we have

$$f_{\tau_n|\Delta=\ell}(t) = \sum_{k=2}^{n} a(k,\ell) G(1+\ell, \lambda_k + \theta)(t), \quad t \geq 0,$$

where

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

and

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

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

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

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 the random variable taking values $k = n, n-1, \dots, 2$ so that

$$K = k \quad \text{iff} \quad \tau_n = S_k.$$

More specifically, K is related to the ancestral process as follows

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

We give the distributions of K below.

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

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

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

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

where $\alpha \in \mathcal{E}_n$ is a partition in k classes such that each class has cardinality $\#\alpha_i = n_i$ and $n_1 + \cdots + n_k = n$. In the sequel, we shall use the notation $|\alpha| = k$. For a proof of this classical result, see [9, 10], 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 equation (1), we find that

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

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

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

Therefore, we obtain that

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

which yields the desired result. □

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

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

In order to compute the variance, recall that

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

Then we have

$$\mathrm{Var}[K] = \frac{1}{18}(n^2 + 31n - 2).$$

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. It gives a proof of that the random variable X_n has a power law distribution

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

and

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

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

Lemma 1 *Let $n \geq 4$, and x an integer so that $n > x \geq 3$. We 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

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

Applying the recursion again, the lemma follows from the above equations.

□

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

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

For a proof of this result, see eg [5] chap 1., eq. (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 to $K = k$, this means that X_n has the same distribution as

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

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

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

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 $m = 1, \ldots, n-2$,

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

For $x = 2$, 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 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 $x = 2, \ldots, n-1$. $\square$

We turn now 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. Dividing 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 $\sigma^2 = 8/45$ if $x = 2$, and

$$\sigma^2 = \frac{4(11 + 4x^4 - 27x^2)}{x(2x+1)(2x-1)(x-1)^2(x+1)^2}, \quad \text{for all } x \geq 3. \tag{2}$$

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 McKenzy 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 can be involved into a quicksort-like recurrence equation [8]

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

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

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

where δ denotes the Kronecker symbol. The expression of σ^2 was found by taking the variance in both size of equation 3. The induction can be solved using an analytic lemma (Lemma 1, [8] p. 1691) and a symbolic algebra package (eg MAPLE). For $n \geq 2x+1$, we obtained that $\mathrm{Var}[X_n^x] = \sigma^2 n$ with σ^2 given by the equation (2). The final result is a consequence of Hwang and Neininger's classification of toll functions ([8] p. 1701). $\square$

5 Some comparisons

5.1 Two random genes

For sake of comparison, we will also describe the distribution of the level of coalescence for two arbitrarily chosen genes. The two genes can be labeled 1 and 2. Then let us denote their coalescence time as $\tilde{\tau}_n$. We set

$$\tilde{K} = 1 + A_n(\tilde{\tau}_n),$$

and 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 $\mathcal{E}(1)$. A next section 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 $n \geq 2$, we have

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

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

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

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

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

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

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

and then

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

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 K. The nodes close to the root are given more important weights in the distribution of K. This can also be seen with 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*

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

and

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

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

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

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

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

which yields the result. $\square$

This distribution can be found in a different manner using results on binary search trees introduced in theoretical computer science. Aldous [1] and Devroye [4] described the number of occurrence 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 $\log n$

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

where H_n is the 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

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

which corresponds to the Yule distribution of parameter $\rho = 1$ given that the variable 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 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 that

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

6 External branch lengths

6.1 Unconditional distribution

The main result in 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 follows from proposition 1. Fu and Li [7] provided a different proof for these results. Using the fact that

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

we obtain the following results.

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

$$\mathrm{E}[\tau_n] = \frac{2}{n},$$

and

$$\mathrm{Var}[\tau_n] = \frac{4}{n^2}.$$

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

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

We have the following result.

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

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

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

$$f(t) = \sum_{k=2}^{n} a_k \, \lambda_k \exp(-\lambda_k t), \quad t \geq 0.$$

Proof. According to proposition 1, we have

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

Using decomposition in simple elements, 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 one 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 the infinitely-many sites model of mutation in the coalescent. Then, the number of substitutions Δ between a gene and one closest relative is distributed as a mixture of shifted geometric distributions*

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

where a_k is given in Theorem 3 equation (4),

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

and

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

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

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

and then

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

where $L_{\tau_n}^{(\ell)}$ is the ℓth derivative of the Laplace transform L_{τ_n}. Using Theorem 3, we can conclude the proof. $\square$

The moments of Δ were found by Fu and Li following a different method [7]. We have

$$\mathrm{E}[\Delta] = \frac{2}{n}\theta$$

and

$$\mathrm{Var}[\Delta] = \frac{2}{n}\theta + \frac{4}{n^2}\theta^2.$$

(We use the fact that $\mathrm{E}[\tau_n] = \sum_{k=2}^{n} \frac{a_k}{\lambda_k} = \frac{2}{n}$ and that $\mathrm{E}[\tau_n^2] = 2\sum_{k=2}^{n} \frac{a_k}{\lambda_k^2} = \frac{8}{n^2}$.)

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

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

Using proposition 4, the conditional density can be reformulated as a mixture of Gamma distributions

$$f_{\tau_n|\Delta=\ell}(t) = \sum_{k=2}^{n} a(k,\ell) G(1+\ell, \lambda_k+\theta)(t), \quad t \geq 0,$$

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

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

Table 1 reports the values of $\mathrm{P}(\Delta = \ell)$ for $n = 10, 30$, $\ell = 0, \ldots, 9$ and $\theta = 1, 10$. Figure 1 displays the curves of f_{τ_n} and $f_{\tau_n|\Delta=\ell}$ for $\ell = 0, 2, 5$ and $\theta = 10$. Exact computations of the conditional expectation are reported in

Table 2 reports for $\theta = 1, 10$ and $\ell = 0, \dots, 10$. In order to provide numerical values, we used the following formula

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

We remark that the distribution of Δ is concentrated on small integers, with a rapid decrease as the number of substitutions increases. As well, larger sample sizes lead to more concentrated distributions. On the other hand, the conditional expectations become rather independent on 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 (MRCA) for a non recombinating 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 iff no mutation occurred since the MRCA. For $\ell = 0, \dots, m$, the conditional probability of observing ℓ matches out of m is then binomial

$$\text{for } t > 0, \quad p(\ell \mid \tau_n = t) = \frac{n!}{\ell!(m-\ell)!} e^{-\theta \ell t} \left(1 - e^{-\theta t}\right)^{m-\ell} .$$

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

$$\text{for } t > 0, \quad p_{\tau_n}(t \mid \ell) \propto e^{-\theta \ell t} \left(1 - e^{-\theta t}\right)^{m-\ell} f_{\tau_n}(t).$$

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

$$p(\tau_n = t \mid \ell) = \frac{\sum_{k=2}^{n} a_k \lambda_k \left(\begin{array}{c} m \\ \ell \end{array} \right) e^{-(\theta \ell + \lambda_k) t} \left(1 - e^{-\theta t}\right)^{m-\ell}}{\sum_{k=2}^{n} a_k \lambda_k \left(\begin{array}{c} m \\ \ell \end{array} \right) I(m, \ell, \theta, k)}$$

where

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

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

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

Walsh [19] used a mutation rate per generation equal to $\mu = 1/500$ motivated by estimates on the Y human chromosome, and the effective size was estimated as $N_e \approx 5,000$. These values lead to an estimate of $\theta = 20$ using that $\theta = 2N_e\mu$. Considering a perfect match at $m = 20$ markers and $n = 2$ individuals, the 95% Bayesian credible region was computed as $(6\text{e}{-05}, 0.00922)$ which corresponded to an interval of $(0.3, 46.1)$ generations for the MRCA. One conclusion was that the forensic use of the Y chromosome is rather limited. Here, we reexamine the upper bound of the 95% Bayesian credible region given $n = 40, 60, 80, 100$ individuals. From the equation (5), we find that the upper bound decreases to $\lfloor tN_e + 1 \rfloor = 40, 30, 8, 7$ generations respectively. Using $m = 100$ markers, these numbers reduce to $\lfloor tN_e + 1 \rfloor = 9, 7, 5, 4$ generations.

References

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

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

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

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

[5] Durrett, R. *Probabilistic models of DNA sequences.* Springer-Verlag, New-York, 2003.

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

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

[8] Hwang H-K., R. Neininger. Phase change of limit laws in the quicksort recurrence under varying toll functions. SIAM J. Comput. Vol. 31, 6 (2002), pp. 1687-1722.

[9] Kingman, J.F.C. The coalescent. *Stochastic Processes and their Applications*, **13** (1982), 235-248.

[10] Kingman, J.F.C. On the genealogies of large populations. *Journal of Applied Probability*,**19A** (1982), 27-43.

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

[12] Nordborg, M. Coalescent theory. In *Handbook of statistical genetics*, D.J. Balding et al eds, Wiley and sons, inc, New-York, 2001, 179-208.

[13] Regnier, M. A limiting distribution for quicksort. *RAIRO Inform. Theor. Appl.*, **23** (1989), 335-343.

[14] Rösler U. A limit theorem for quicksort. *RAIRO Inform. Theor. Appl.*, **25** (1992), 85-100.

[15] Saunders, I.W., S. Tavaré and G.A. Watterson. On the genealogy of nested subsamples from a haploid population. *Advances in Applied probability*, **16** (1984), 471-491.

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

[17] Tavaré, S. *Ancestral inference in population genetics*, Lectures on Probability Theory and Statistics. Ecole d'Et de Probabilit de Saint-Flour XXXI – 2001. (Ed. Picard J.) Lecture Notes in Mathematics, 1837, Springer Verlag, 2004.

[18] Tavaré, S. Ancestral inference from DNA sequence data. In *Mathematical modeling in Ecology, Physiology and cell biology*, chap 5, Othmer at al eds, 1997, 81-96.

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

[20] Watterson, G.A. On the number of segregating sites in genetical models without recombination, *Theoretical Population Biology*, **7** (1975), 256-276.

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

[22] Wiuf, C. and P. Donnelly. Conditional genalogies and the age of a neutral mutant, *Theoretical Population Biology*, **56** (1999), 183-201.

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

Δ	$\theta = 1$ $n = 10$	$\theta = 1$ $n = 30$	$\theta = 10$ $n = 10$	$\theta = 10$ $n = 30$
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 1: *Probability distribution of the number of segregating sites* Δ. *The symbol "–" indicates values below* 0.001. *The last row gives the sum of the ten probabilities.*

Δ	$\theta = 1$ $n = 10$	$\theta = 1$ $n = 30$	$\theta = 10$ $n = 10$	$\theta = 10$ $n = 30$
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

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

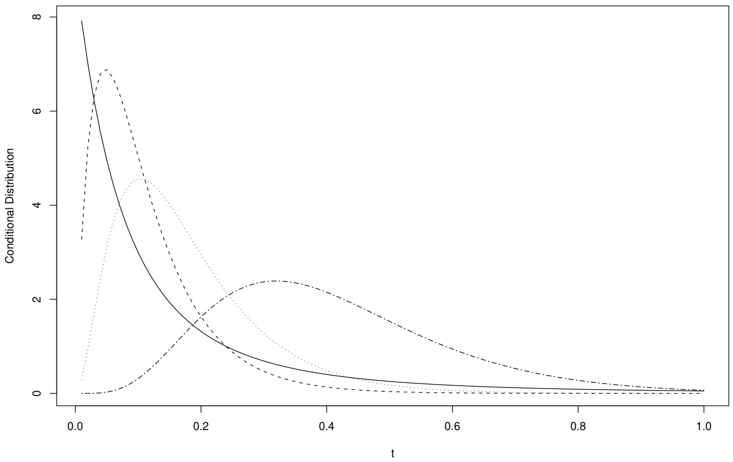

Figure 1: *The unconditional coalescence time distribution (solid line) and the conditional distributions given that* $\Delta = 1, 2, 5$ *segregating sites are observed. The sample size is* $n = 10$ *and the mutation rate is* $\theta = 10$.