A discrete approach to the external branches of a Kingman coalescent tree. Theoretical results and practical applications

Filippo Disanto*, Thomas Wiehe†

Abstract

The Kingman coalescent process is a classical model of gene genealogies in population genetics. It generates Yule-distributed, binary ranked tree topologies—also called *histories*—with a finite number of n leaves, together with $n-1$ exponentially distributed time lengths: one for each each layer of the history. Using a discrete approach, we study the lengths of the external branches of Yule distributed histories, where the length of an external branch is defined as the rank of its parent node. We study the multiplicity of external branches of given length in a random history of n leaves. A correspondence between the external branches of the ordered histories of size n and the non-peak entries of the permutations of size $n-1$ provides easy access to the length distributions of the first and second longest external branch in a random Yule history and coalescent tree of size n. The length of the longest external branch is also studied in dependence of root balance of a random tree. As a practical application, we compare the observed and expected number of mutations on the longest external branches in samples from natural populations.

Keywords Yule histories · Coalescent trees · External branches · Branch length · Combinatorics

1 Introduction

The Kingman coalescent is a fundamental model for population genetic analyses. In its original version [15, 16] it is the backward-in-time analogue of a pure birth process where each existing external branch is chosen uniformly to give rise to the next split into two offspring branches. As such, the involved trees are binary with internal nodes linearly ordered by time. Disregarding branch length and keeping track only of the ranking of the internal nodes, such trees are called (unlabeled) ranked trees [18] or histories [17], where the probability of a history of size n to be the underlying ranked tree topology of a random coalscent tree follows the Yule distribution [11, 22].

The stochastic, combinatoric, topological and population genetic properties of coalescent trees have been subject of numerous investigations. One prominent application in population genetics is to analyze and interpete the frequency spectrum of mutations in light of tree topology and of the length distribution of tree branches. In particular, singletons in the mutation frequency spectrum relate to the length of the external branches of the tree. Blum and François [3], as well as Caliebe *et al.* [5], have studied the length distribution of a randomly chosen external branch from a Kingman coalescent tree and derived also the limiting distribution for large n. The same topic has been investigated by Freund and Möhle [9] for the Bolthausen-Sznitman coalescent. These results have been generalized to the comprehensive class of Λ-coalescents by Diehl and Kersting [6], who also examined the asymptotic distribution of the external branch lengths ordered by size.

Here, we study the external branches of coalescent trees from a combinatorial point of view. We distinguish the time length of an external branch from its discrete length, the latter being defined as the rank (looking backward in time) of the parent node of the considered branch in the underlying history. An external branch of

*Dipartimento di Matematica, Università di Pisa, Italy. Corresponding author. Email: filippo.disanto@unipi.it
†Institut für Genetik, Universität zu Köln, Germany. Email: twiehe@uni-koeln.de

discrete length s is thus divided into s segments spanning the last s layers of the tree. When a random history of size n is selected under the Yule distribution, that is, it is the history underlying a random coalescent tree of n leaves, we derive several probabilistic properties of the length of its external branches. We focus on the probability of a given number of external branches of given length, on the probability that external branches of given length are absent, and on the probability of the length of the first and second longest external branches. Importantly, from the discrete length of an external branch, we can recover the probability density of its time length measured in coalescent units by summing exponentially distributed independent random variables.

Our study is also motivated by the practical question whether the observation of a certain number of singleton mutations in one single chromosome is compatible, or not, with the neutral infinite sites model [7, 14] of constant population size and constant mutation rate. We apply our results on the length of the two longest external branches of a tree to two kind of data: the mitochondrial genomes of three human populations [1], and to a nuclear gene of *Danio rerio* [20]. Non-recombining chromosomes, such as mitochondria, or short genomic fragments should not show any homogenizing effect, due to recombination, on the length distribution of external branches. Therefore, in the examples studied, we expect to recover and estimate the lengths of the longest and second-longest external branches of a single coalescent tree.

The paper is organized as follows. We introduce terminology and some useful properties of histories and coalescent trees in Section 2, showing in particular that external branch lengths in random trees can also be analyzed in terms of peaks of random permutations. In Section 3, we subdivide branches into branch segments. Given a random history of size n, i.e. with n leaves, we derive the counts—either 0, 1 or 2—of external branches with a given number of segments and ask in Section 4 how often a history misses external branches with a certain number of segments. We then consider the number of segments in the longest and second longest external branches (Section 5). Using convolution of exponential distributions, segment numbers can be scaled-back to coalescent time-units and the results be applied to experimental data (Section 6). We conclude with an outlook on some open problems (Section 7).

2 Histories, peaks of permutations and external branch length

Following [17], a *history* of size n is a full binary rooted tree with a ranking of its $n-1$ internal nodes. If $[a, b]$ denotes the set $\{i \in \mathbb{Z} : a \leq i \leq b\}$, then the internal nodes of a history of size n are labeled by the integers in $[1, n-1]$, starting from the bottom (i.e., closest to the leaves) of the tree proceeding to the top (i.e., the root) in increasing order (Fig. 1). The root has thus label $n-1$. A branch of a history is an edge connecting two internal nodes or one internal node and one leaf. In the latter case, the branch is said to be *external*. The ranking of the internal nodes of a history of size n divides the tree into $n-1$ layers, with the ith layer intersecting exactly i branches. A branch *segment* is a part of a branch that crosses a given layer of the tree. For example, in the history depicted in Fig. 1 the two branches appended to the node labeled 6 consist of 1 and 4 branch segments. The external branch appended to node 3 consists instead of 3 branch segments. In general, the number of segments of an external branch of a history t is the label of the internal node of t from which the considered external branch descends.

An *ordered* history of size n is a planar embedding of a history of size n in which subtrees have a left-right orientation. The *Yule* branching process [11, 22] creates a random ordered history of size n in $n-1$ consecutive steps. Starting with a root branch, in the ith step of the process each one of the i present terminal nodes has the same probability to split into two new terminal nodes. After $n-1$ steps an ordered history is created with uniform probability among the $(n-1)!$ possible ordered histories of n leaves. By summing the probability $1/(n-1)!$ of each ordered history with the same underlying (un-ordered) history, the uniform distribution over the set of ordered histories of size n induces a probability distribution—the Yule distribution—over the set of histories of size n. In particular, the Yule probability of a history t of size n can be seen as a function of the number $|\mathrm{or}(t)|$ of its different left-right orientations. More precisely, if $|\mathrm{ch}(t)|$ is the number of cherries (i.e., subtrees

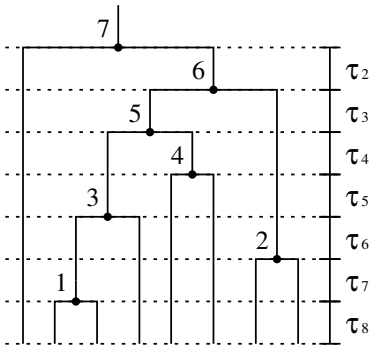

Figure 1: A history of size $n = 8$. Internal nodes are ranked and labeled by the integers in $[1, n-1]$, from bottom to top. The ranking divides the history into $n-1$ layers, with the ith layer intersecting exactly i branches. A segment is a part of a branch that extends across a given layer. The number of segments of an external branch corresponds to the label of the internal node from which the branch descends. In a coalescent tree, an exponentially distributed variable τ_i assigns a time length to the ith layer of the history underlying the tree.

of size 2) in t, then $|\mathrm{or}(t)| = 2^{n-1-|\mathrm{ch}(t)|}$—by flipping subtrees stemming from the $n-1-|\mathrm{ch}(t)|$ non-cherry internal nodes of t, we obtain the possible left-right orientations of t—and the Yule probability of the history t is given by $|\mathrm{or}(t)|/(n-1)!$ [17]. In this manuscript, all random histories of fixed size n are selected under the Yule distribution, whereas ordered histories of size n are uniformly distributed.

The fact that the Yule distribution over the set of histories of size n is induced by the uniform distribution over the set of ordered histories of size n allows us to derive probabilistic properties of the number of branch segments present in the external branches of random histories by studying the number of branch segments in the external branches of random ordered histories of the same size. In particular, in the study of the number of segments of the external branches, combinatorial properties of ordered histories can be derived through a series of known enumerative results [4] on the number of permutations of fixed size with a given set of peak entries. Indeed, by bijectively encoding the $(n-1)!$ ordered histories of size n as permutations of $n-1$ integers, the external branches of an ordered history t can be seen to correspond to the non-peak entries of the associated permutation π_t, where the number of segments in each external branch of t is the value of the associated non-peak entry in the permutation π_t. The mapping $t \to \pi_t$ is well known [10] and can be described as follows.

Given an ordered history t of size n with internal nodes labeled from bottom to top in increasing order (Fig. 1), let us consider an internal node of t labeled by the integer k. The permutation $\pi_t[k]$ associated with the subtree of t rooted at k is constructed recursively as $\pi_t[k] = (\pi_t[k_\ell], k, \pi_t[k_r])$, where $\pi_t[k_\ell], \pi_t[k_r]$ are the permutations associated with the subtrees of t rooted at the left and right children nodes k_ℓ and k_r of k, respectively. For example, if t is the ordered history of size $n = 8$ depicted in Fig. 1, then $\pi_t[7] = (7, 1, 3, 5, 4, 6, 2)$, where $\pi_t[7_\ell] = \emptyset$ is the empty permutation and $\pi_t[7_r] = \pi_t[6] = (1, 3, 5, 4, 6, 2)$. Similarly, the permutation $\pi_t[6]$ has been constructed as $\pi_t[6] = (\pi_t[6_\ell], 6, \pi_t[6_r])$, where $\pi_t[6_\ell] = \pi_t[5] = (1, 3, 5, 4)$ and $\pi_t[6_r] = \pi_t[2] = (2)$. In particular, the permutation $\pi_t = \pi_t[n-1]$ of size $n-1$ is by definition the permutation associated with the considered ordered history t of size n. Denoting by $\pi_t(i)$ the ith entry of the permutation π_t, we say that $\pi_t(i)$ is a *peak* when $i \neq 1, i \neq n-1$ and $\pi_t(i-1) < \pi_t(i) > \pi_t(i+1)$, and we observe the following property of the mapping $t \to \pi_t$: the entry k in the permutation π_t is a peak if and only if the node of t labeled by k has both its left and right child that are internal nodes of t. In other words, k is an *external* node of t, that is, k has *at least* one descending external branch, if and only if the entry k in the permutation π_t is not a peak. For instance, the peak entries of the permutation $\pi_t = (7, 1, 3, 5, 4, 6, 2)$ associated with the ordered history t depicted in Fig. 1 are 5 and 6, which correspond to the internal nodes of t without a descending external branch.

The study of probabilistic properties of the number of segments of the external branches of a random history can assist in the analysis of the time length of the external branches of a Kingman coalescent tree [12, 15, 21]. A coalescent tree of size n can be modeled as a random history t of n leaves and a sequence $(\tau_2, \ldots, \tau_n)$ of independent exponentially distributed random variables assigning a time length to the different layers of t (Fig. 1) [19, 23]. Under the model, the variable τ_i has mean $\mathbb{E}[\tau_i] = 1/\lambda_i$, with $\lambda_i = \binom{i}{2}$. From $\mathbb{E}[\tau_i]$ we can easily recover the expected value of the time length $\tau(s)$ of an external branch of t containing exactly s segments. The expectation of $\tau(s)$ is the sum of the expectations of the time length of the last s layers of the history t, that is,

$$\sum_{i=n+1-s}^{n} 1/\lambda_i = \frac{2}{n-s} - \frac{2}{n}. \tag{1}$$

More generally, the probability density function $f_s(x)$ of the time length $\tau(s)$ is the density of the sum $\sum_{i=n+1-s}^{n} \tau_i$, which is given [2, 8] by

$$f_s(x) = \prod_{i=1}^{s} \lambda_{i+n-s} \times \sum_{j=1}^{s} \frac{e^{-\lambda_{j+n-s}x}}{\prod_{k\in[1,s]\setminus\{j\}}(\lambda_{k+n-s} - \lambda_{j+n-s})}. \tag{2}$$

The latter formula enables the calculation of the probability of the time length of an external branch of a coalescent tree given the number of segments possessed by that branch in the underlying history.

3 The number of external branches with a given number of segments

In this section, we study the number of ordered histories of size n with $\mu \in \{0, 1, 2\}$ external branches consisting of exactly s segments. In particular, dividing this number by $(n-1)!$—i.e., by the total number of ordered histories of n taxa—we obtain the probability that a random history of size n selected under the Yule distribution has μ_s external branches of s segments. This calculation is then extended to the conditional probability that a Yule-distributed history of size n has μ external branches of s segments given that it has μ_r external branches of r segments.

Let $a_{n,s,\mu}$ denote the number of ordered histories of size n with *exactly* μ external branches of $1 \leq s \leq n-1$ segments. Constructing ordered histories of size $n \geq 3$ by splitting a leaf of an ordered history of size $n-1$ yields for $1 < s \leq n-1$

$$\begin{aligned} a_{n,s,2} &= a_{n-1,s-1,2}(n-3) \\ a_{n,s,1} &= a_{n-1,s-1,1}(n-2) + 2\,a_{n-1,s-1,2} \\ a_{n,s,0} &= (n-1)! - a_{n,s,2} - a_{n,s,1}, \end{aligned}$$

where $a_{n,1,2} = (n-1)!$ and $a_{n,1,1} = 0$ for every $n \geq 2$. For example, the recurrence for $a_{n,s,1}$ generates a tree of size n with exactly one external branch of s segments either from a tree of size $n-1$ with exactly one external branch of $s-1$ segments by splitting one of its $n-2$ external branches of length different from $s-1$ or from a tree of size $n-1$ with 2 external branches of $s-1$ segments by splitting one of these two branches. Note, that when we split a branch, all the remaining branches increase their number of segments by one.

Solving the recurrences above we find

$$\begin{aligned} a_{n,s,2} &= (n-3)!\,(n-s)\,(n-s-1) \\ a_{n,s,1} &= (n-2)!\,2\sum_{k=1}^{s-1} \frac{a_{n-k,s-k,2}}{(n-k-1)!} = (n-3)!\,2(n-s)(s-1) \\ a_{n,s,0} &= (n-1)! - (n-3)!\,(n-s)\,(n+s-3), \end{aligned}$$

and the probability $p_s(\mu)$ of a history of size n with μ external branches of s segments is given by

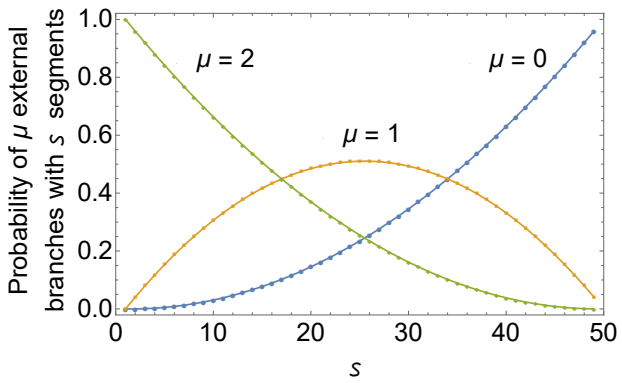

Figure 2: Probability that a random Yule-distributed history of size $n = 50$ has $\mu \in \{0, 1, 2\}$ external branches with $s \in [1, n-1]$ segments. The probability increases (resp. decreases) with s, when $\mu = 0$ (resp. $\mu = 2$).

$$p_s(\mu) = \frac{a_{n,s,\mu}}{(n-1)!} = \begin{cases} \frac{(n-s)(n-s-1)}{(n-1)(n-2)}, & \text{if } \mu = 2; \\ \frac{2(n-s)(s-1)}{(n-1)(n-2)}, & \text{if } \mu = 1; \\ \frac{(s-1)(s-2)}{(n-1)(n-2)}, & \text{if } \mu = 0. \end{cases} \tag{3}$$

As shown in Fig. 2, we have the symmetries $p_s(0) = p_{n-s+1}(2), p_s(1) = p_{n-s+1}(1)$.

The expected number of external branches with s segments in a random history of size n decreases linearly with s as given by $\mathbb{E}_s[\mu] = 2(n-s)/(n-1)$. Furthermore, we can calculate the value $s^* = s^*(n)$ such that $p_s(0) \leq 1/2$ for every $s \leq s^*$. We find

$$s^* = \frac{1}{2}(3 + \sqrt{5 + 2\,n(n-3)}) \sim \frac{n}{\sqrt{2}} \approx 0.7 \cdot n.$$

In other words, for a random history of size n, the probability of having at least one external branch with s segments is larger or smaller than 50% depending on whether $s < s^*$ or $s > s^*$, respectively. Because $p_s(0) = p_{n-s+1}(2)$, we also have that $p_s(2)$ is larger or smaller than 50% when $s < n - s^* + 1$ or $s > n - s^* + 1$, respectively, where $n - s^* + 1 \sim n(1 - 1/\sqrt{2}) \approx 0.3 \cdot n$.

Conditional probability calculation. The calculation above can be extended to the number a_{n,s,μ,r,μ_r} of ordered histories of size n in which there are $\mu \in \{0, 1, 2\}$ external branches of s segments and $\mu_r \in \{0, 1, 2\}$ external branches of r segments. When $1 < s, r \leq n-1$, we have the following recurrences

$$\begin{aligned}
a_{n,s,2,r,2} &= a_{n-1,s-1,2,r-1,2}(n-5) \\
a_{n,s,2,r,1} &= 2a_{n-1,s-1,2,r-1,2} + a_{n-1,s-1,2,r-1,1}(n-4) \\
a_{n,s,2,r,0} &= a_{n-1,s-1,2,r-1,1} + a_{n-1,s-1,2,r-1,0}(n-3) \\
a_{n,s,1,r,1} &= 2a_{n-1,s-1,2,r-1,1} + 2a_{n-1,s-1,1,r-1,2} + a_{n-1,s-1,1,r-1,1}(n-3) \\
a_{n,s,1,r,0} &= 2a_{n-1,s-1,2,r-1,0} + a_{n-1,s-1,1,r-1,1} + a_{n-1,s-1,1,r-1,0}(n-2) \\
a_{n,s,0,r,0} &= a_{n-1,s-1,1,r-1,0} + a_{n-1,s-1,0,r-1,1} + a_{n-1,s-1,0,r-1,0}(n-1)
\end{aligned}$$

where $a_{n,s,\mu,1,\mu_r} = a_{n,s,\mu}$ if $\mu_r = 2$, $a_{n,s,\mu,1,\mu_r} = 0$ if $\mu_r \neq 2$, $a_{n,1,\mu,r,\mu_r} = a_{n,r,\mu_r}$ if $\mu = 2$, and $a_{n,1,\mu,r,\mu_r} = 0$ if $\mu \neq 2$. For instance, the recurrence for $a_{n,s,1,r,1}$ yields a tree of size n with exactly one external branch of s segments and exactly one external branch of r segments either from a tree of size $n-1$ with exactly one external

branch of $s-1$ segments and one external branch of $r-1$ segments by splitting one of its $n-3$ external branches of length different from $s-1$ and $r-1$ or from a tree of size $n-1$ with 2 external branches of $s-1$ (or $r-1$) segments and one external branch of $r-1$ (or $s-1$) segments by splitting one of the two branches with $s-1$ (or $r-1$) segments.

Solving the recurrences yields for $n \geq 5$ and $1 \leq s, r \leq n-1$ the following formulas

$$a_{n,s,2,r,2} = \begin{cases} (n-5)!\,(n-r-2)(n-r-3)(n-s)(n-s-1), & \text{if } r < s; \\ (n-5)!\,(n-s-2)(n-s-3)(n-r)(n-r-1), & \text{if } r > s. \end{cases} \tag{4}$$

$$a_{n,s,2,r,1} = \begin{cases} 2(n-5)!\,(n-r-2)(n-s)(n-s-1)(r-1), & \text{if } r < s; \\ 2(n-5)!\,(n-r)(n-s-2)[n(r-1)+3s-r(3+s)+1], & \text{if } r > s. \end{cases} \tag{5}$$

$$a_{n,s,2,r,0} = \begin{cases} (n-5)!\,(n-s)(n-s-1)(2-3r+r^2), & \text{if } r < s; \\ (n-5)!\,[n(2r-2-s)-r(s+6)+7s+2](n-r)(s-1)+(n-3)!\,(s-r)(s-r+1), & \text{if } r > s. \end{cases} \tag{6}$$

$$a_{n,s,1,r,1} = \begin{cases} 4(n-5)!\,(n-s)(r-1)[2+3r+n(s-2)-s(r+2)], & \text{if } r < s; \\ 4(n-5)!\,(n-r)(s-1)[2+3s+n(r-2)-r(s+2)], & \text{if } r > s. \end{cases} \tag{7}$$

$$a_{n,s,1,r,0} = \begin{cases} 2(n-5)!\,(r-2)(r-1)(n-s)(s-3), & \text{if } r < s; \\ 2(n-5)!\,(r-3)(s-1)[n(r-2)+4s-r(2+s)], & \text{if } r > s. \end{cases} \tag{8}$$

$$a_{n,s,0,r,0} = \begin{cases} (n-5)!\,(r-2)(r-1)(s-4)(s-3), & \text{if } r < s; \\ (n-5)!\,(s-2)(s-1)(r-4)(r-3), & \text{if } r > s. \end{cases} \tag{9}$$

Therefore, for a random history with $n \geq 5$ taxa, the conditional probability $p_s(\mu|r,\mu_r)$ of $\mu \in \{0,1,2\}$ external branches of s segments given $\mu_r \in \{0,1,2\}$ external branches of r segments can be computed as

$$p_s(\mu|r,\mu_r) = \begin{cases} \frac{a_{n,s,\mu,r,\mu_r}}{a_{n,r,\mu_r}} & \text{if } \mu \geq \mu_r; \\ \frac{a_{n,r,\mu_r,s,\mu}}{a_{n,r,\mu_r}} & \text{if } \mu < \mu_r. \end{cases} \tag{10}$$

In Fig. 3, we plot $p_s(\mu|r,\mu_r)$ (solid line) and $p_s(\mu)$ (boxes) for $n = 50$. When $r = 10$ and $\mu_r = 0$ (left column), we see that $p_s(0|r,\mu_r) \leq p_s(0)$, $p_s(1|r,\mu_r) \leq p_s(1)$ if $s \leq n/2$, and $p_s(2|r,\mu_r) \geq p_s(2)$. Thus, a random history that misses a short ($r = 10$) external branch, has a slightly smaller probability to miss an external branch of s segments than a random unconstrained history of the same size. Interestingly, the missing short external branch of r segments does not increase, for s close to r, the probability of having one external branch of s segments—in fact, $p_s(1|r,\mu_r) < p_s(1)$—but it increases the probability of having two external branches of s segments. When instead $r = 40$ and $\mu_r = 2$ (right column of Fig. 3), we see that the existence of two long ($r = 40$) external branches decreases, for s close to r, both the probability of having two external branches of s segments and the probability of having one external branch of s segments.

4 The probability that only the external branches of length s are absent

Here, we provide a procedure for evaluating the probability that a random history of size n does not have external branches of s segments if and only if s belongs to a given integer set. Results of this section are derived by considering ordered histories and the correspondence between their external branches and the non-peak entries of the associated permutations (Section 2).

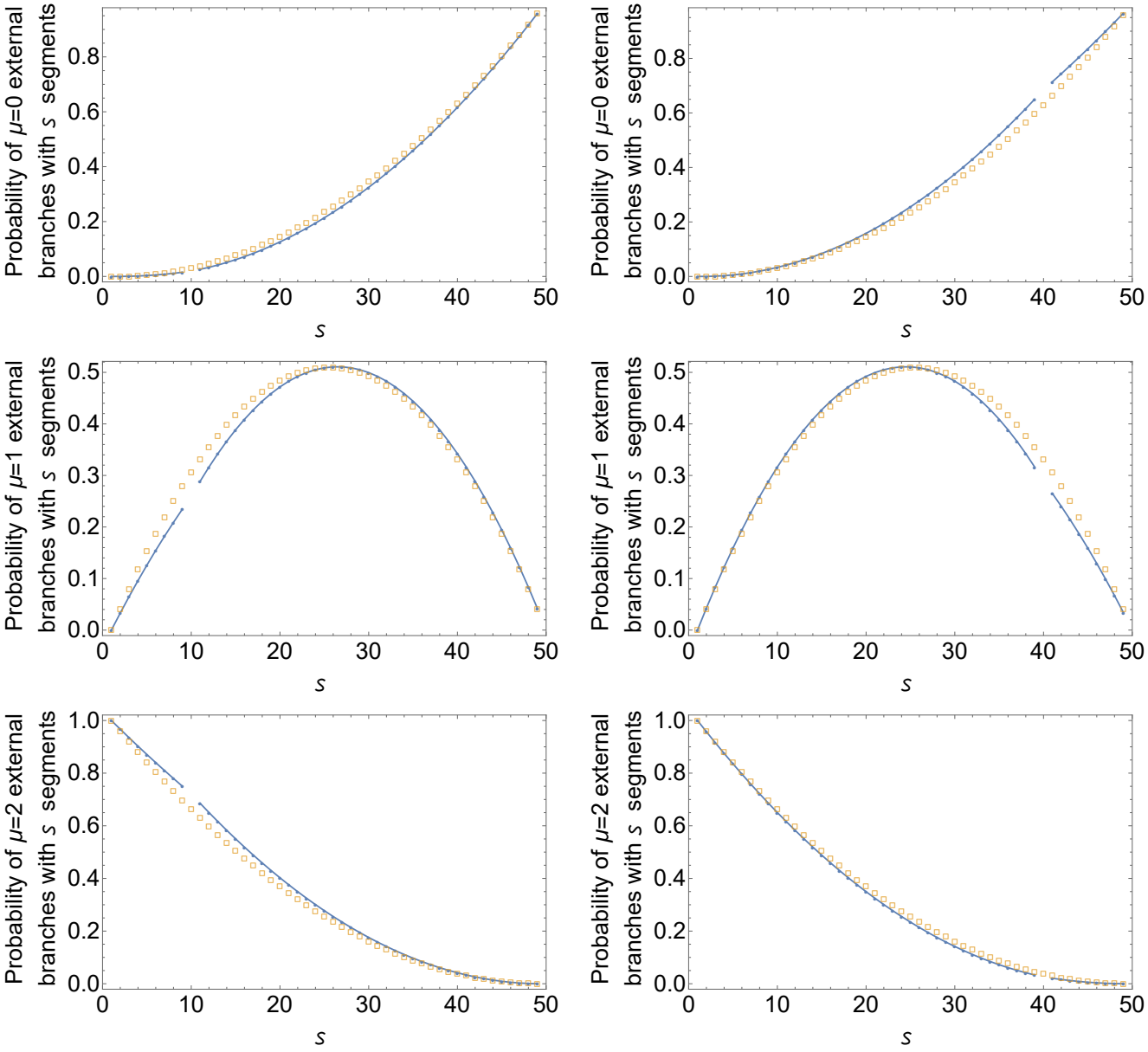

Figure 3: Conditional probability $p_s(\mu|r,\mu_r)$ of a random Yule-distributed history of size $n = 50$ having $\mu \in \{0,1,2\}$ external branches of $s \in [1, n-1]$ segments. In the left column, $r = 10$ and $\mu_r = 0$. In the right column, $r = 40$ and $\mu_r = 2$. Boxes give the probability $p_s(\mu)$ in an unconstrained history.

Histories with few missing external branches. We recall from Section 2 that a node of a history is said to be *external* when it has at least one child that is a leaf. The external branches of a history are therefore those branches that are appended to an external node. In particular, when we label the nodes of a history of size n from 1 to $n-1$ moving upwards in the tree (Fig. 1), the number of segments of an external branch is given by the integer label of the external node to which the branch is appended. For a given history t with $\lceil n/2 \rceil \leq k \leq n-1$ external nodes, let $\mathcal{E}_t \equiv \{e_1, ..., e_k\}$ be the set of labels of the external nodes of t. Thus, t has at least one external branch with s segments if and only if $s \in \mathcal{E}_t$, and we say that t misses an external branch of s segments when $s \notin \mathcal{E}_t$.

When $\mathcal{E} \subseteq [1, n-1]$ is a set of $k \in \{n-1, n-2, n-3\}$ positive integers, the probability that a random history of size n has its set of missing external branches given by $\overline{\mathcal{E}} = [1, n-1] \setminus \mathcal{E}$ can be calculated as

$$p_n(\overline{\mathcal{E}}) = \begin{cases} \frac{2^{n-2}}{(n-1)!}, & \text{if } \overline{\mathcal{E}} = \emptyset; \\ \\ \frac{2^{n-3}(2^{i-2}-1)}{(n-1)!}, & \text{if } \overline{\mathcal{E}} = \{i\} \subseteq [3, n-1]; \\ \\ \frac{2^{n-4}(2^{i-2}-1)(2^{j-i-1}-1)+3(3^{i-2}-2^{i-1}+1)2^{n+j-i-6}}{(n-1)!}, & \text{if } \overline{\mathcal{E}} = \{i, j\} \subseteq [3, n-1] \text{ and } i < j. \end{cases} \tag{11}$$

In particular, as shown in Theorem 3.1 of [4], the enumerator in each formula counts the number of permutations of size $n-1$—i.e., the ordered histories of size n—in which the peak entries—i.e., the non-external nodes—are exactly those belonging to the set $\overline{\mathcal{E}}$. We remark that the probability $p_n(\overline{\mathcal{E}})$ is different from the type of probabilities analyzed in Section 3. For instance, when $\overline{\mathcal{E}} = \{i\}$ the probability $p_n(\overline{\mathcal{E}})$ in (11) considers the histories of size n in which the only non-external node is i. The probability $p_i(0)$ of Eq. (3) is instead the probability of an history in which at least the node i is not external—or, equivalently, in which there are 0 external branches of i segments.

Existence and probability of a history with a given set of missing external branches. We first give a necessary and sufficient condition for the existence of at least one history of size n with a given set of missing external branches. If $\mathcal{E} \subseteq [1, n-1]$ and $\overline{\mathcal{E}} = \{i_1, i_2, ..., i_w\}$ $(i_j < i_{j+1})$, then there exists at least one history t of size $n \geq 4$ such that $\mathcal{E}_t = \mathcal{E}$ if and only if either $w = 0$—i.e. there are no missing external branches—or $i_j \geq 2j+1$ for every $1 \leq j \leq w$—i.e. the number of segments of the jth smallest missing external branch is larger than or equal to $2j+1$. This characterization is given in Theorem 2.1 of [4] in terms of allowed peak sets of a permutation of given size.

When $n \geq 4$ and $\overline{\mathcal{E}} \subseteq [3, n-1]$ satisfies the condition above, the probability $p_n(\overline{\mathcal{E}})$ that a random history of size n has its set of missing external branches given by $\overline{\mathcal{E}}$ can be calculated recursively as follows

$$p_n(\overline{\mathcal{E}}) = \begin{cases} \frac{2\, p_{n-1}(\overline{\mathcal{E}})}{n-1}, & \text{if } n-1 \notin \overline{\mathcal{E}}; \\ \\ \frac{(n-1-2\,|\overline{\mathcal{E}}|) \cdot p_{n-1}(\overline{\mathcal{E}} \setminus \{n-1\}) + 2 \sum_{j \in ([3, n-2] \setminus \overline{\mathcal{E}})} p_{n-1}\left((\overline{\mathcal{E}} \setminus \{n-1\}) \cup \{j\}\right)}{n-1}, & \text{if } n-1 \in \overline{\mathcal{E}}, \end{cases} \tag{12}$$

where $p_4(\emptyset) = 2/3, p_4(\{3\}) = 1/3$, and $p_4(\overline{\mathcal{E}}) = 0$ otherwise. The second formula in (12) follows from Lemma 3.2 of [4]. The first formula is instead a direct consequence of the fact that an ordered history of size n having an external branch of $n-1$ segments and a set $\overline{\mathcal{E}} \subseteq [3, n-2]$ of missing external branches (Fig. 4) must have as left or right root subtree an ordered history of size $n-1$ whose set of missing external branches is $\overline{\mathcal{E}}$. In Fig. 4 (right), we plot for $n = 8$ the probability $p_n(\overline{\mathcal{E}})$ for all the 20 admissibe sets $\overline{\mathcal{E}}$ of missing external branches. Among sets $\overline{\mathcal{E}}$ of the same cardinality, we observe a correlation between the probability of each $\overline{\mathcal{E}}$ and its ranking in the lexicographic order. This correlation is weaker if we compare the probabilities of sets $\overline{\mathcal{E}}$ with a different number of elements. For example, $\{5, 7\}$ is lexicographically smaller than $\{6\}$ but it has a larger probability. Similarly, the probability of $\{6, 7\}$ is larger than the probability of $\{7\}$.

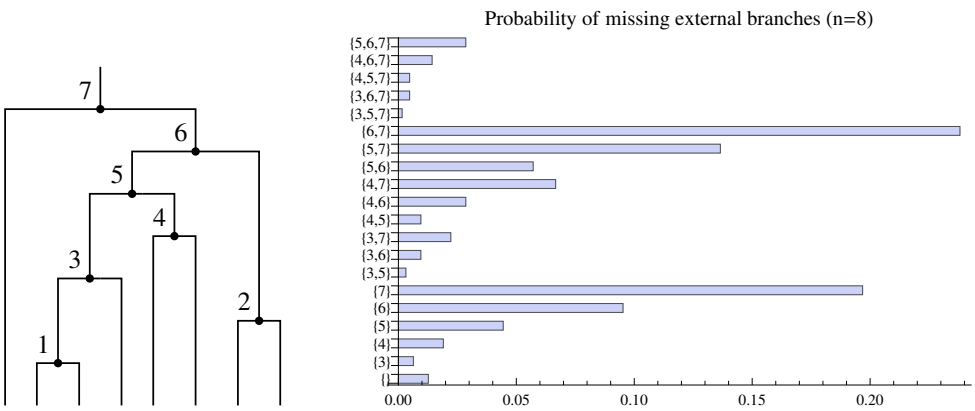

Figure 4: Left: a history of size $n = 8$ whose set of missing external branches is $\overline{\mathcal{E}} = \{5, 6\}$. In the tree, there is at least one external branch of s segments if and only if $s \in [1, n-1] \setminus \overline{\mathcal{E}} = \{1, 2, 3, 4, 7\}$. Right: probability of all the possible admissible sets of missing external branches for a random Yule-distributed history of size $n = 8$.

The recursive procedure described in (12) can be used to extend the calculation performed in Section 3 of the joint probability $a_{n,s,0,r,0}/(n-1)!$ of missing two external branches of s and r segments in a random history of size n. More precisely, given a set $S \subseteq [3, n-1]$, the probability that a random history of size n misses *at least* those branches with discrete length listed in S can be evaluated as $\sum_{\overline{\mathcal{E}} \supseteq S} p_n(\overline{\mathcal{E}})$, that is, by summing the probabilities of the admissible sets $\overline{\mathcal{E}}$ containing S.

5 The number of segments in the first and second longest external branches

In this section, we study the random variables s' and s'' counting, respectively, the number of branch segments in the longest and second longest external branch of a random history of given size. In Section 5.1, we provide an exact formula for the probability of a given value of $s' \in [\lceil n/2 \rceil, n-1]$. The probability of s' to be smaller than a certain threshold is studied in Section 5.1.1. Relationships between s' and tree imbalance are analyzed in Section 5.1.2. In Section 5.2, we calculate the probability of a given value of s'' and the conditional probability of s'' given s' for a random history of fixed size.

5.1 The number of segments in the longest external branch

In this section, we calculate the probability that the longest external branch in a random history of given size has $s' = s$ segments. As in the previous sections, we use the equivalence with the uniform distribution over ordered histories of size n or permutations of size $n-1$.

Let $\Pi_n(X)$ denote the number of permutations of size n with peak entries matching the elements of the set $X \subseteq [3, n]$, and choose $s \in [\lceil (n+1)/2 \rceil, n]$. For a fixed set $S \subseteq [3, s-1]$, by setting $k = n - s$ in Lemma 3.3 of [4], for $n \geq 2$ we find

$$\Pi_n(S \cup [s+1, n]) = 2(n-s+1)\,\Pi_{n-1}(S \cup [s, n-1]) + (n-s)(n-s+1)\,\Pi_{n-2}(S \cup [s, n-2]).$$

The latter equation relates the number of permutations of size n with peak entries given by the elements of the set $S \cup [s+1, n]$ to the number of permutations of size $n-1$ and $n-2$ with peak entries given by $S \cup [s, n-1]$ and $S \cup [s, n-2]$, respectively. Note that if $s = n$ and $n-1 \in S$, then $\Pi_{n-2}(S \cup [s, n-2]) = 0$.

If we sum both sides of the latter equation over the possible subsets S of $[3, s-1]$, then we obtain

$$\sum_S \Pi_n(S \cup [s+1, n]) = 2(n-s+1) \sum_S \Pi_{n-1}(S \cup [s, n-1]) + (n-s)(n-s+1) \sum_S \Pi_{n-2}(S \cup [s, n-2]), \quad (13)$$

where the sum $\sum_S \Pi_n(S \cup [s+1, n])$ counts the permutations of size n in which the largest non-peak entry is s, and the sums $\sum_S \Pi_{n-1}(S \cup [s, n-1])$ and $\sum_S \Pi_{n-2}(S \cup [s, n-2])$ count respectively the permutations of size $n-1$ and $n-2$ in which the largest non-peak entry is strictly smaller than s. For instance, set $n = 5$. For $s = 3$, Table 2 of [4]—which reports the number of permutations of size $n \leq 8$ with a fixed set of peak entries—gives $\sum_S \Pi_n(S \cup [s+1, n]) = 12$, $\sum_S \Pi_{n-1}(S \cup [s, n-1]) = 0$ and $\sum_S \Pi_{n-2}(S \cup [s, n-2]) = 2$, where $12 = 2 \cdot 3 \cdot 0 + 2 \cdot 3 \cdot 2$ in agreement with Eq. (13). For $s = 4$, we have $\sum_S \Pi_n(S \cup [s+1, n]) = 60$, $\sum_S \Pi_{n-1}(S \cup [s, n-1]) = 12$ and $\sum_S \Pi_{n-2}(S \cup [s, n-2]) = 6$, where $60 = 2 \cdot 2 \cdot 12 + 1 \cdot 2 \cdot 6$. Finally, for $s = 5$ we have $\sum_S \Pi_n(S \cup [s+1, n]) = 48$, $\sum_S \Pi_{n-1}(S \cup [s, n-1]) = 24$ and $\sum_S \Pi_{n-2}(S \cup [s, n-2]) = 6$, where $48 = 2 \cdot 1 \cdot 24 + 0 \cdot 1 \cdot 6$.

Note that the number $\sum_S \Pi_n(S \cup [s+1, n])$ of permutations of size n in which the largest non-peak entry is s can be seen as the difference between the number of permutations of size n whose largest non-peak entry is strictly smaller than $s+1$ and the number of permutations of size n whose largest non-peak entry is strictly smaller than s. Hence, by using the correspondence between non-peak entries of permutations of size n and external branches of ordered histories of size $n+1$, from (13) we have the following equation for the number $a_{n,s}$ of ordered histories of size n in which the longest external branch has a number of segments *strictly* smaller than s:

$$a_{n+1,s+1} - a_{n+1,s} = 2(n-s+1)\, a_{n,s} + (n-s)(n-s+1)\, a_{n-1,s}.$$

Replacing $n+1$ by n and $s+1$ by s in the latter equation, for $n \geq 3$ we obtain the recurrence

$$a_{n,s} = a_{n,s-1} + 2(n-s+1)\, a_{n-1,s-1} + (n-s)(n-s+1)\, a_{n-2,s-1}, \tag{14}$$

where $a_{n,s} = 0$ if $s \leq \lceil n/2 \rceil$, and $a_{n,s} = (n-1)!$ if $s = n$. By iteratively setting $s = n, s = n-1, s = n-2, \ldots$ in Eq. (14) and extracting the term $a_{n,s-1}$, we can recursively calculate a formula for $a_{n,n-i}$. For the first values of $i \in [0, 5]$, we find

$$\begin{aligned}
a_{n,n} &= (n-1)! \\
a_{n,n-1} &= -2(n-2)! + (n-1)! \\
a_{n,n-2} &= 6(n-3)! - 6(n-2)! + (n-1)! \\
a_{n,n-3} &= -24(n-4)! + 36(n-3)! - 12(n-2)! + (n-1)! \\
a_{n,n-4} &= 120(n-5)! - 240(n-4)! + 120(n-3)! - 20(n-2)! + (n-1)! \\
a_{n,n-5} &= -720(n-6)! + 1800(n-5)! - 1200(n-4)! + 300(n-3)! - 30(n-2)! + (n-1)!,
\end{aligned}$$

and more in general, as shown in the Appendix, we have

$$a_{n,n-i} = i! \sum_{k=1}^{i+1} (-1)^{k+1} \frac{(n-k)!}{(i+1-k)!} \binom{i+1}{i+2-k}. \tag{15}$$

Setting $s = n - i$, the latter formula can be rewritten as

$$a_{n,s} = (n-s)! \sum_{k=1}^{n-s+1} (-1)^{k+1} \frac{(n-k)!}{(n-s+1-k)!} \binom{n-s+1}{n-s+2-k}, \tag{16}$$

and for $n \geq 3$ the probability $p_n(s)$ that a random history of size n has its longest external branch containing exactly $s' = s$ segments (Fig. 5, left) can be computed as

$$p_n(s) = \frac{a_{n,s+1} - a_{n,s}}{(n-1)!}. \tag{17}$$

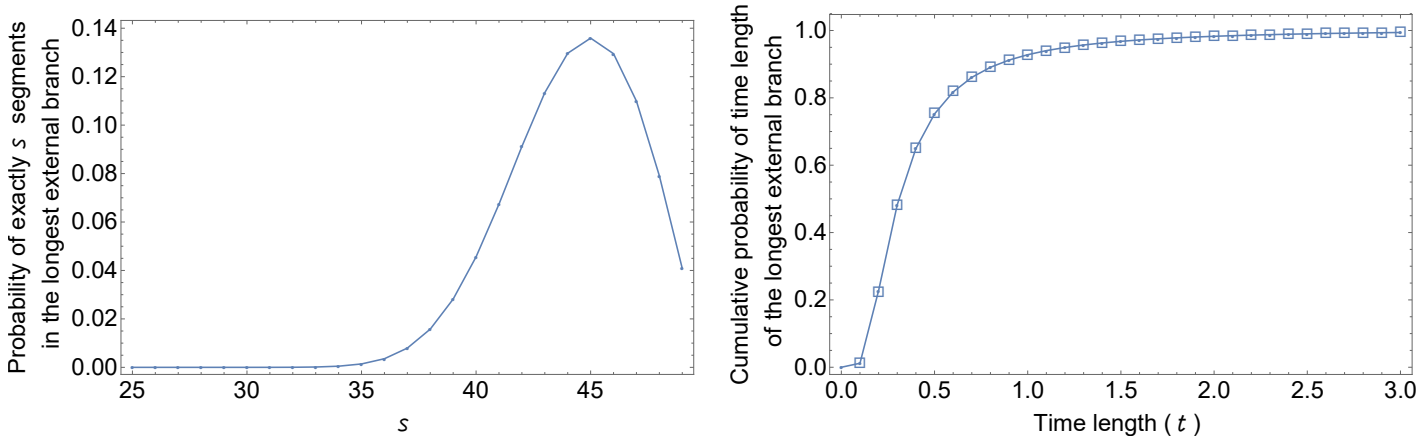

Figure 5: Left: Probability that a Yule-distributed history of size $n = 50$ has exactly $s \in [\lceil n/2 \rceil, n-1]$ segments in its longest external branch. Right: Cumulative probability that the longest external branch of a history of size $n = 50$ has length at most t (in coalescent time units). Theoretical probabilities (solid line) are calculated by Eq. (17), using Eq. (2) with $\lambda_i = \binom{i}{2}$. Simulated data (boxes) have been obtained through `ms` [13].

By using Eqs. (2) and (17), the probability that the time length τ of the longest external branch in random history of size n lies in the interval $a < \tau \leq b$ can be calculated as

$$p_n(a < \tau \leq b) = \sum_{s=\lceil n/2 \rceil}^{n-1} p_n(s) \int_a^b f_s(x) dx.$$

In Fig. 5 (right), the cumulative probability $p_n(\tau \leq t)$ is plotted setting $n = 50$ and letting t range over the interval $[0, 3]$ in steps of 0.1. The theoretical line is in perfect agreement with rescaled data obtained through the `ms` coalescent simulator [13]. Note that in the `ms` setting, the mean length of the ith layer of a coalescent tree is $1/\left(2\binom{i}{2}\right)$, while in our theoretical calculations the mean is $1/\lambda_i$, with $\lambda_i = \binom{i}{2}$.

5.1.1 The longest external branch has a large number of segments

The probability that the longest external branch of a random Yule-distributed history of size n has exactly s segments has been calculated in the previous section. The plot given in Fig. 5 (left) shows that for a large fraction of the ordered histories of size n the longest external branch has a number of segments quite close to the maximum value $n-1$. To better understand this observation, we derive in this section an approximation for the value d_α such that with probability α a random history of size n has less than $n - d_\alpha$ segments in its longest external branch.

From the left-top plot of Fig. 3, we see that missing an external branch of size r has only a small effect on the probability of missing an external branch of size $s \neq r$. For a given value of $d \geq 1$, we approximate the probability $\text{Prob}(\mu_{n-1} = ... = \mu_{n-d} = 0)$ that for all $i \in [1, d]$ a random history of size n has $\mu_{n-i} = 0$ external branches of length (number of segments) $n - i$ as the product $\prod_{i=1}^{d} p_{n-i}(0)$ of the probabilities $p_{n-i}(0)$ given in Eq. (3). That is,

$$\begin{aligned} \text{Prob}(\mu_{n-1} = ... = \mu_{n-d} = 0) &\approx \prod_{i=1}^{d} p_{n-i}(0) \\ &= \frac{(n-2)!\,(n-3)!}{(n-1)^d\,(n-2)^d\,(n-d-2)!\,(n-d-3)!}. \end{aligned} \tag{18}$$

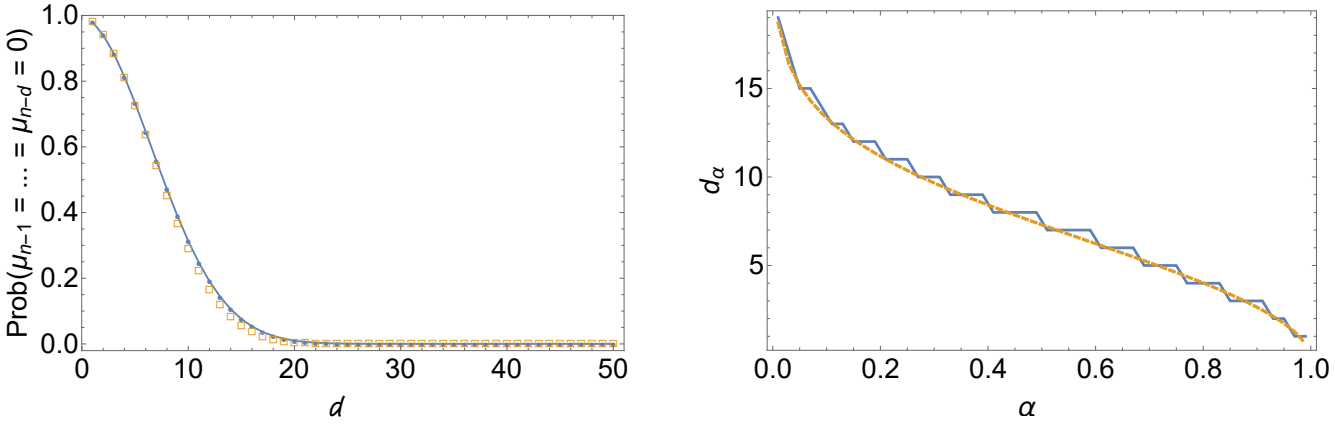

Figure 6: Left: The probability of 0 external branches of size larger than or equal to $n-d$—or the probability of the longest external branch with less than $n-d$ segments—in a random Yule-distributed history of size $n = 100$. Boxes give the exact probability evaluated as $a_{n,n-d}/(n-1)!$ by using Eq. (16), the solid line is the approximation given in Eq. (18). Right: Plot of d_α for $n = 100$ and $0.01 \leq \alpha \leq 0.99$ (in steps of 0.02). The zigzag line is d_α computed as the integer d for which the probability $\text{Prob}(\mu_{n-1} = ... = \mu_{n-d} = 0)$ is closest (not necessarily equal) to α. The smooth line is the approximation of d_α given in Eq. (21).

Note that $\text{Prob}(\mu_{n-1} = ... = \mu_{n-d} = 0)$ can be seen as the probability of a random history of size n with its longest external branch containing less than $n-d$ segments, and its exact value can be calculated from Eq. (16) as $\text{Prob}(\mu_{n-1} = ... = \mu_{n-d} = 0) = a_{n,n-d}/(n-1)!$. The accuracy of the approximation in (18) can be verified for $n = 100$ in Fig. 6 (left). The figure also shows that for a fixed value $0 \leq \alpha \leq 1$, the probability $\text{Prob}(\mu_{n-1} = ... = \mu_{n-d} = 0)$ is equal to α for a value of d much smaller than n. For instance, if we set $\alpha = 0.2$ with $n = 100$, then $\text{Prob}(\mu_{n-1} = ... = \mu_{n-d} = 0) = \alpha$ for $d \approx 12$. In other words, there is an 80% probability that a random Yule-distributed history of size $n = 100$ has at least $100 - 12 = 88$ segments in its longest external branch. To measure the probability that the longest external branch of a random history is shorter than a certain threshold, we use the approximation in Eq. (18) for studying the value d_α of d such that

$$\text{Prob}(\mu_{n-1} = ... = \mu_{n-d_\alpha} = 0) = \alpha. \tag{19}$$

We find that d_α grows roughly like a constant multiple of $\sqrt{n}$, with the constant depending on the chosen α.

Assuming $d_\alpha/n \approx 0$ for a fixed α and n sufficiently large, we apply Stirling's approximation $n! \approx \sqrt{2\pi n}(n/e)^n$ to Eq. (18) and, from (19), we obtain

$$e^{-2-2d_\alpha}(n-2)^{n-d_\alpha-5/2}(n-1)^{n-d_\alpha-3/2}(n-d_\alpha-3)^{d_\alpha+5/2-n}(n-d_\alpha-2)^{d_\alpha+3/2-n} = \alpha.$$

Taking the logarithm of the latter expression gives the equation

$$\frac{{d_\alpha}^3}{n^2} + \frac{7{d_\alpha}^2}{n^2} + \frac{{d_\alpha}^2}{n} + \frac{29d_\alpha}{2n^2} + \frac{d_\alpha}{n} + \frac{17}{2n^2} + \log(\alpha) = 0, \tag{20}$$

where we have used the second order approximations $\log(n-2) \approx \log(n) - 2/n - (-2/n)^2/2, \log(n-1) \approx \log(n) - 1/n - (-1/n)^2/2, \log(n-d_\alpha-3) \approx \log(n) - (d_\alpha+3)/n - [-(d_\alpha+3)/n]^2/2, \log(n-d_\alpha-2) \approx \log(n) - (d_\alpha+2)/n - [-(d_\alpha+2)/n]^2/2$. Note that $-{d_\alpha}^3/n^2 - {d_\alpha}^2/n = -{d_\alpha}^2/n(d_\alpha/n+1) \approx -{d_\alpha}^2/n$ is the leading term in the left-hand side of Eq. (20), where all the remaining terms of that side of the equation are close to 0 if $d_\alpha/n \approx 0$. Thus, in order to satisfy the latter equation, $-{d_\alpha}^2/n$ must be close to $\log(\alpha)$, that is, $d_\alpha \approx \sqrt{\log(1/\alpha)} \cdot \sqrt{n}$ for

n large. Approximating d_α as a sum of powers of $\sqrt{n}$, we find

$$d_\alpha \approx \sqrt{\log\left(\frac{1}{\alpha}\right)}\cdot\sqrt{n} - \frac{1}{2}\left(1+\log\left(\frac{1}{\alpha}\right)\right) + \frac{1-22\log\left(\frac{1}{\alpha}\right)+5\log^2\left(\frac{1}{\alpha}\right)}{8\sqrt{\log\left(\frac{1}{\alpha}\right)}}\cdot\frac{1}{\sqrt{n}}. \quad (21)$$

In particular, the estimate in (21)—plotted in Fig. 6 (right) for $n = 100$ and different values of α—has been obtained by first substituting $d_\alpha = c_1\sqrt{n}+c_2+c_3/\sqrt{n}$ in the left-hand side of the equation given in (20), and then requiring the first three largest terms of the resulting expression—i.e., the coefficients of $n^{-i/2}$ for $i = 0, 1, 2$—to be identically 0. The values of c_1, c_2, and c_3 found in this way are such that the left-hand side of the polynomial equation in (20) is equal to 0 up to an error term of order $\mathcal{O}(1/n^{3/2})$. Higher precision can be obtained by the substitution $d_\alpha = c_1\sqrt{n}+c_2+c_3/\sqrt{n}+c_4/n$, yielding $c_4 = \frac{1}{4}\left(-4\log^2\left(\frac{1}{\alpha}\right)+22\log\left(\frac{1}{\alpha}\right)-17\right)$ with an error term in the equation of order $\mathcal{O}(1/n^2)$.

The square root behavior of the quantity $d_\alpha(n)$ shows that for increasing tree size a random history will have with high probability a large number of segments in its longest external branch. For instance, setting $\alpha = 0.2$ in (21) we obtain the estimate $d_{0.2} \approx 1.26864\sqrt{n} - 1.30472 - \frac{2.1141}{\sqrt{n}}$, where the longest external branch of a random history of size n has at least $n - d_{0.2}$ segments with probability $1-\alpha = 0.8$.

5.1.2 Longest external branch and root imbalance

In this section, we study how root imbalance affects the number of segments of the longest external branch. Our calculations show that the length of the longest external branch is almost independent of imbalance and affected only by extreme values of the latter parameter. In order to measure root imbalance of a history t of size n, we consider the parameter $\omega(t) \in [1, \lfloor n/2 \rfloor]$ defined as the size of the smallest root subtree of t. For instance, if t is the history of Fig. 1, then $\omega(t) = 1$. Also in this section, we use the fact that the Yule distribution over histories of n leaves is induced by the uniform distribution over ordered histories of n leaves (Section 2).

If t is an ordered history of size n, let t_1 and t_2 be the *rescaled* left and right root subtrees of t of size n_1 and n_2, respectively. If the left root subtree t_ℓ of t has size n_1, then t_1 is obtained from t_ℓ by relabeling its $n_1 - 1$ internal nodes with the integers in $[1, n_1 - 1]$. Each internal node receives the new label $i \in [1, n_1 - 1]$ if the same node has the ith largest label when considered in t_ℓ. Similarly, for the rescaled right root subtree t_2 of t. As an example, consider the ordered history t given by the right root subtree of the history depicted in Fig. 1. In Newick format, $t = ((((\bullet,\bullet)_1,\bullet)_3,(\bullet,\bullet)_4)_5,(\bullet,\bullet)_2)_6$, where the integer next to a closed parenthesis is the label of the associated internal node. The left root subtree of t is given by $t_\ell = (((\bullet,\bullet)_1,\bullet)_3,(\bullet,\bullet)_4)_5$. The rescaled left root subtree is $t_1 = (((\bullet,\bullet)_1,\bullet)_2,(\bullet,\bullet)_3)_4$, which is obtained by replacing the labels $3, 4, 5$ by $2, 3, 4$, respectively, in t_ℓ. The rescaled right root subtree of t is instead $t_2 = (\bullet,\bullet)_1$, which is obtained by taking the right root subtree $(\bullet,\bullet)_2$ of t with the label 2 replaced by 1. Finally, let s_1 (resp. s_2) be the number of segments in the longest external branch of t_1 (resp. t_2), and denote by s'_1 (resp. s'_2) the number of segments in the longest external branch of the non-rescaled left (resp. right) root subtree of t. Thus, $s_1 \le s'_1$ and $s_2 \le s'_2$. For example, if t is the history given by the right root subtree of the history depicted in Fig. 1, then we have $s_1 = 3, s'_1 = 4, s_2 = 1$, and $s'_2 = 2$.

Let $p_1(s'_1, s'_2|s_1, s_2, n_1, n_2) = \mathrm{Prob}(s'_1, s'_2|s_1, s_2, n_1 \ge 2, n_2 \ge 2)$. This is the joint probability of a given number of segments in the longest external branch of the non-rescaled left and right root subtree of a random ordered history of size $n = n_1 + n_2$, given the number of segments s_1 and s_2 in the longest external branch of its rescaled left and right root subtree. The tree decomposition depicted in Fig. 7 yields

$$p_1(s'_1, s'_2|s_1, s_2, n_1, n_2) = \frac{\binom{n_1+n_2-2-s'_2}{n_2-1-s_2}\binom{s'_2-1-s'_1}{s'_2-s_2-s_1}\binom{s'_1-1}{s_1-1} + \binom{n_1+n_2-2-s'_1}{n_2-1-s_1}\binom{s'_1-1-s'_2}{s'_1-s_1-s_2}\binom{s'_2-1}{s_2-1}}{\binom{n_1+n_2-2}{n_1-1}}, \quad (22)$$

where we set $\binom{a}{b} = 0$ if $a < 0$ or $b < 0$. In particular, given two ordered histories t_1 and t_2 of size $n_1 \ge 2$ and $n_2 \ge 2$, respectively, with s_1 and s_2 segments in their longest external branch, we consider the ordered histories

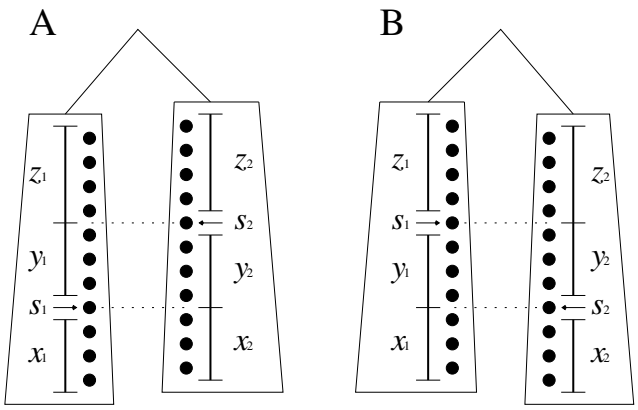

Figure 7: Tree decompositions for calculating Eq. (22).

t of size $n = n_1 + n_2$ having t_1 and t_2 as rescaled root subtrees. The number of such histories t is given by the denominator of the latter formula, which counts the possible orderings of the $n_1 - 1$ internal nodes of t_1 with the $n_2 - 1$ internal nodes of t_2. The enumerator of the same formula counts the orderings of the internal nodes of t_1 and t_2 for which the resulting history t is compatible with the values of s_1' and s_2'. More precisely, when the node s_1 of t_1 has a lower rank in t than the rank assigned to the node s_2 of t_2 (Fig. 7A), the number of orderings is given by the first summand in the enumerator of Eq. (22). When instead s_1 is placed above s_2 in t (Fig. 7B), the number of orderings is given by the second summand of the enumerator.

For the case depicted in Fig. 7A,

$$x_1 = s_1 - 1$$

is the number of internal nodes of t_1 with ranking smaller than s_1. Similarly, $x_2 + y_2 = s_2 - 1$ is the number of internal nodes of t_2 with ranking smaller than s_2. Also, we have $y_1 + z_1 = n_1 - 1 - s_1$ and

$$z_2 = n_2 - 1 - s_2.$$

If x_2 counts the nodes of t_2 that in t are placed below the node s_1 of t_1, then $s_1' = s_1 + x_2$. That is,

$$x_2 = s_1' - s_1.$$

From the values of $x_2 + y_2$ and x_2, we thus find

$$y_2 = s_2 - 1 - s_1' + s_1.$$

If s_1 togehter with the nodes counted by $x_1 + y_1$ are placed below the node s_2 in t, then $s_2' = s_2 + x_1 + 1 + y_1$. From the value of x_1, we find

$$y_1 = s_2' - s_2 - s_1.$$

Finally, from the value of $y_1 + z_1$, we obtain

$$z_1 = n_1 - 1 - s_2' + s_2.$$

The number of orderings compatible with the values of s_1' and s_2' when we consider the decomposition of Fig. 7A is thus given by

$$\binom{x_1 + x_2}{x_1}\binom{y_1 + y_2}{y_1}\binom{z_1 + z_2}{z_2},$$

which is the first summand in the numerator of Eq. (22).

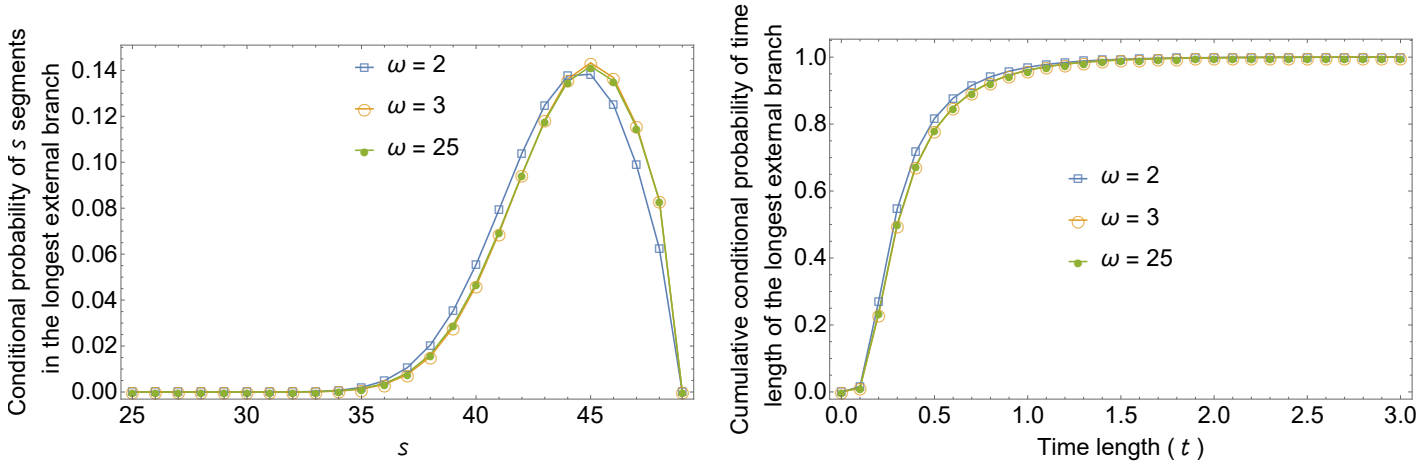

Figure 8: Left: Conditional probability that a Yule-distributed history of size $n = 50$ and $\omega = 2, 3, 25$ has exactly $s \in [\lceil n/2 \rceil, n-1]$ segments in its longest external branch (boxes for $\omega = 2$, circles for $\omega = 3$, and dots for $\omega = 25$).Right: Cumulative conditional probability that the longest external branch of a history of size $n = 50$ and $\omega = 2, 3, 25$ has length at most t (in coalescent time units).

The decomposition of Fig. 7B, in which we consider the case of s_1 placed above s_2 in t, yields by similar calculations the second summand of the enumerator in Eq. (22). Observe that the two summands cannot be both different from zero at the same time. Indeed, the second binomial factor in the first summand is different from zero for $s_2 \geq s_1' - s_1 + 1$, while the second binomial factor of the second summand is not zero for $s_2 \leq s_1' - s_1$.

By using the probability in Eq. (22), we can calculate $p_2(s|s_1, s_2, n_1, n_2) = \text{Prob}(s|s_1, s_2, n_1 \geq 2, n_2 \geq 2)$, that is, the conditional probability of $s' = s$ segments in the longest external branch of a random ordered history of size $n = n_1 + n_2$, given the number of segments s_1 and s_2 in the longest external branch of its rescaled root subtrees. Indeed, if $n_1, n_2 \geq 2$, then the longest external branch of a history of size $n = n_1 + n_2$ is the longest external branch of its non-rescaled root subtrees. Thus, $p_2(s|s_1, s_2, n_1, n_2)$ can be written as

$$p_2(s|s_1, s_2, n_1, n_2) = \sum_{s_2'=s_2}^{s-1} p_1(s, s_2'|s_1, s_2, n_1, n_2) + \sum_{s_1'=s_1}^{s-1} p_1(s_1', s|s_1, s_2, n_1, n_2). \tag{23}$$

Finally, if $\omega \in [1, \lfloor n/2 \rfloor]$ is the size of the smallest root subtree in a random Yule-distributed history (or uniformly distributed ordered history) of size n, then we can calculate the conditional probability $p_n(s|\omega) \equiv \text{Prob}(s|\omega, n)$ of $s' = s$ segments in the longest external branch as

$$p_n(s|\omega) = \delta_{\omega,1}\delta_{s,n-1} + \sum_{s_1=\lceil \omega/2 \rceil}^{\min(s,\omega-1)} \sum_{s_2=\lceil (n-\omega)/2 \rceil}^{\min(s,n-\omega-1)} p_\omega(s_1)\, p_{n-\omega}(s_2)\, p_2(s|s_1, s_2, \omega, n-\omega), \tag{24}$$

where $\delta_{s,n-1}$ is the probability of s segments when $\omega = 1$, and $p_\omega(s_1)$ and $p_{n-\omega}(s_2)$ are respectively the probability of Eq. (17) that an ordered history of size ω and $n - \omega$ has s_1 and s_2 segments in the longest external branch. The probability in Eq. (24) is plotted for $\omega = 2, 3, 25$ in Fig. 8 (left) for random Yule distributed histories of size $n = 50$. Interestingly, we observe that for $\omega = 3$ and $\omega = 25$ the probability of s segments in the longest external branch is basically the same. When $\omega = 2$, the distribution is shifted to the left.

By using Eqs. (2) and (24), the conditional probability that the time length τ of the longest external branch in a random history of size n with smaller root subtree of size ω lies in the interval $a < \tau \leq b$ can be calculated

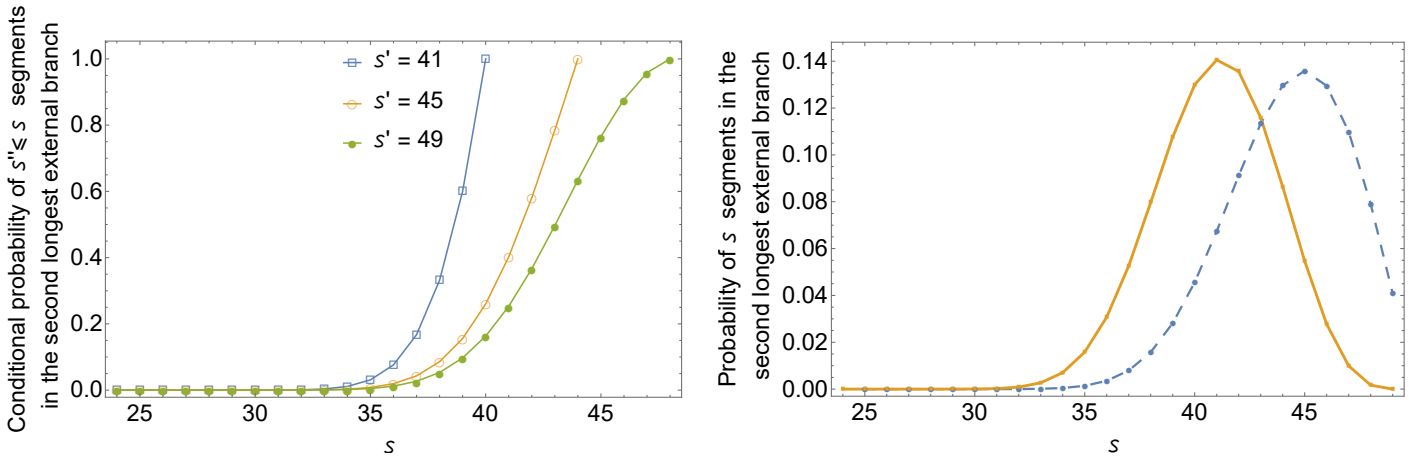

Figure 9: Left: Conditional probability of $s'' \leq s$ segments in the second longest external branch of a random Yule-distributed history of size $n = 50$, when the longest external branch has $s' = 41, 45, 49$ segments (left to right curve). Right: Unconditional probability that a random Yule-distributed history of size $n = 50$ has exactly s segments in its first (dashed line) and second (solid line) longest external branch.

as

$$p_n(a < \tau \leq b|\omega) = \sum_{s=\lceil n/2 \rceil}^{n-1} p_n(s|\omega) \int_a^b f_s(x)dx.$$

For $\omega = 2, 3, 25$, the cumulative conditional probability $p_n(\tau \leq t|\omega)$ is plotted in Fig. 8 (right) setting $n = 50$ and letting t range over the interval $[0, 3]$ in steps of 0.1. As already observed for the number of segments, for $\omega = 3$ and $\omega = 25$ the values of $p_n(\tau \leq t|\omega)$ are basically the same. When $t \leq 1$, the cumulative conditional probability is larger for $\omega = 2$ than for $\omega = 3, 25$.

5.2 The number of segments in the second longest external branch

In Section 5.1, we have studied the random variable s' counting the number of segments in the longest external branch of a random history of given size. Note that for a history t there could be two external branches of size $s'(t)$, when these branches form a cherry in t. Considering the set of external branches of t that are strictly shorter than the longest one(s), we define $s''(t) \in [\lfloor (n-1)/2 \rfloor, s'(t) - 1]$ to be the number of segments in the longest external branch of this set, and we say that $s''(t)$ is the number of segments of the second longest external branch of t. For example, $s''(t) = 4$ when t is the history depicted in Figure 1. In this section, we study the distribution of the random variable $s''(t)$, when t is a random history of size n selected under the Yule probability model. By using the equivalence with the uniform distribution over ordered histories of size n or permutations of size $n-1$, we find that the probability of $s'' = s$ segments in the second longest external branch of a random history of size n can be expressed as a simple function of the probability that a random history of size smaller than n has $s' = s$ segments in its longest external branch.

We first calculate the joint probability of s' and s''. Let $\Pi_n(X)$ denote as in section 5.1 the number of permutations of size n with peak entries matching the elements of the set $X \subseteq [3, n]$, and choose s_1 and s_2 such that $\lfloor n/2 \rfloor \leq s_2 < s_1 \leq n$. For a fixed set $Z \subseteq [3, s_2 - 1]$, by setting $S = Z \cup [s_2 + 1, s_1 - 1]$ and $k = n - s_1$ in Lemma 3.3 of [4], for $n \geq 2$ we find

$$\begin{aligned} \Pi_n(Z \cup [s_2+1, s_1-1] \cup [s_1+1, n]) &= 2(n - s_1 + 1)\, \Pi_{n-1}(Z \cup [s_2+1, s_1-1] \cup [s_1, n-1]) \\ &\quad + (n - s_1)(n - s_1 + 1)\, \Pi_{n-2}(Z \cup [s_2+1, s_1-1] \cup [s_1, n-2]). \end{aligned}$$

If we sum both sides of the latter equation over the possible subsets Z of $[3, s_2 - 1]$, then we obtain

$$\begin{aligned}\sum_Z \Pi_n(Z \cup [s_2+1, s_1-1] \cup [s_1+1, n]) &= 2(n-s_1+1) \sum_Z \Pi_{n-1}(Z \cup [s_2+1, s_1-1] \cup [s_1, n-1]) \\ &\quad +(n-s_1)(n-s_1+1) \sum_Z \Pi_{n-2}(Z \cup [s_2+1, s_1-1] \cup [s_1, n-2]).\end{aligned}$$

where the sum $\sum_Z \Pi_n(Z \cup [s_2+1, s_1-1] \cup [s_1+1, n])$ counts the permutations of size n in which the first largest and the second largest non-peak entry are respectively s_1 and s_2, and the sums $\sum_Z \Pi_{n-1}(Z \cup [s_2+1, s_1-1] \cup [s_1, n-1])$ and $\sum_Z \Pi_{n-2}(Z \cup [s_2+1, s_1-1] \cup [s_1, n-2])$ count respectively the permutations of size $n-1$ and $n-2$ in which the largest non-peak entry is s_2. For instance, let us set $n = 6$. For $s_1 = 5$ and $s_2 = 4$, Table 2 of [4] gives $\sum_Z \Pi_n(Z \cup [s_2+1, s_1-1] \cup [s_1+1, n]) = 264$, $\sum_Z \Pi_{n-1}(Z \cup [s_2+1, s_1-1] \cup [s_1, n-1]) = 60$ and $\sum_Z \Pi_{n-2}(Z \cup [s_2+1, s_1-1] \cup [s_1, n-2]) = 12$, where $264 = 2 \cdot 2 \cdot 60 + 1 \cdot 2 \cdot 12$.

By using the correspondence between non-peak entries of permutations of size n and external branches of ordered histories of size $n+1$ (Section 2), the number a_{n+1,s_1,s_2} of ordered histories t of size $n+1$ in which $s'(t) = s_1$ and $s''(t) = s_2$ can be calculated as

$$a_{n+1,s_1,s_2} = 2(n-s_1+1)a_{n,s_2} + (n-s_1)(n-s_1+1)a_{n-1,s_2},$$

where a_{n,s_2} and a_{n-1,s_2} count the number of ordered histories of size n and $n-1$, respectively, in which the longest external branch has s_2 segments. Replacing $n+1$ by n in the latter equation and dividing by the number $(n-1)!$ of ordered histories of size n, the probabilites $p_{n-1}(s_2)$ and $p_{n-2}(s_2)$ of Eq. (17) yield the joint probability

$$p_n(s_1, s_2) = \frac{2(n-s_1)}{n-1} p_{n-1}(s_2) + \frac{(n-s_1)(n-s_1-1)}{(n-1)(n-2)} p_{n-2}(s_2) \tag{25}$$

of a random history of size n in which the longest external branch has $s' = s_1$ segments and the second longest external branch has $s'' = s_2$ segments. The conditional probability of $s'' = s_2$ segments in the second longest external branch given $s' = s_1$ segments in the longest external branch of a random history of size n is thus $p_n(s_2|s_1) = \frac{p_n(s_1,s_2)}{p_n(s_1)}$ (Fig. 9, left). The sum $\sum_{s_1=s_2+1}^{n-1} p_n(s_2|s_1)p_n(s_1)$ gives the unconditional probability of $s'' = s_2$ segments in the second longest external branch of a random history with n leaves (Fig. 9, right).

6 Estimating the longest external branches from experimental data

To compare the theoretical results obtained above with experimental data we estimate the longest external branches of a hypothetical coalescent tree from the maximal counts of singleton mutations in sets of SNP data. Since recombination can disturb tree topology, we concentrate on (i) the non-recombining mitochondrial genomes from *Homo sapiens* (size about 17kb) from three different populations (CEU, CHB and YRI) and (ii) on a short 7.3kb genomic sequence (pre-mRNA) of a single nuclear gene (CTCF) from a wild population of zebrafish. Human data ($1k$ genomes initiative, phase III) were downloaded from `www.1000genomes.org`. The zebrafish sample comes from a larger collection of data, which we sequenced and analyzed as part of a different project [20]. The sample analyzed here has a size of $n = 34$, from 17 individuals collected from the wild (GPS coordinates N022.262 E087.279; sub-population termed 'KG').

We calculated two simple estimates for the relative length of the longest and second longest branches of a coalescent tree: E_1 is based on the total number of SNPs observed (S_{total}), E_2 is based on the observed singletons only (S_{singl}). Since the combined length of all external branches compares to the total tree length as 1 to h_{n-1} [21], we estimate

$$E_1 = \left(\frac{h_{n-1} \cdot S_1}{S_{\text{total}}}, \frac{h_{n-1} \cdot S_2}{S_{\text{total}}} \right)$$

Table 1: Experimental data observed in human (mitochondria) and zebrafish (genome-wide survey).

species	pop.	bp	n°	h^*_{n-1}	S_{total}	$S^{\#}_{\text{singl}}$	$S^{\dagger}_1$	$S^{\ddagger}_2$	E_1	E_2
H. sapiens	CEU	16,569	99	5.167	491	280	15	10	(0.158, 0.105)	(0.054, 0.036)
	CHB		103	5.207	670	407	20	15	(0.155, 0.117)	(0.049, 0.037)
	YRI		107	5.245	655	308	13	13	(0.104, 0.104)	(0.042, 0.042)
D. rerio	KG	7,335	34	4.088	170	62	8	7	(0.192, 0.168)	(0.129, 0.113)

$^\circ$ sample size; * $(n-1)$st harmonic number; # total number of singletons;

† largest number of singletons observed in one indvidual;

‡ second largest number of singletons observed in one individual;

and

$$E_2 = \left(\frac{S_1}{S_{\text{singl}}}, \frac{S_2}{S_{\text{singl}}} \right) .$$

The observed and estimated data are collected in Table 1. All data fit very well the theoretical prediction. For all populations both coordinates of estimate E_1 are larger than those of E_2. This is compatible with the notion that purifying selection leads to an increase of singleton mutations compared to the neutral expectation. However, purifying selection affects all individuals in the same way, hence is not inducing a bias on the longest or second longest branch. The larger values of Danio compared to human are explained by different sample sizes (also visible in the shift among the solid curves representing the theoretical values). Comparing the derived human populations (CEU and CHB) to the ancestral African population (YRI), one observes a slight increase in both coordinates of E_1 in CEU and CHB compared to YRI. This is explained by the well known stronger increase in the number of singletons in the frequency spectrum of the derived populations due to the bottleneck effect accompanying the migration out of Africa.

7 Conclusions

We have studied probabilistic properties of the number of segments present in the external branches of a random Yule-distributed history of given size. The approach followed in our calculations also provides a combinatorial framework for the analysis of the time length of the external branches of a Kingman coalescent tree of given size.

In Section 3, we have focused on the probability that a random history of fixed size has a given number of external branches of a given number of segments. Eq. (3) together with Eqs. (4-10) enable respectively the calculation of the unconditional and conditional probability of $\mu \in \{0, 1, 2\}$ external branches of $s \in [1, n-1]$ segments in a random history of size n. The unconditional probability of $\mu = 0$ (resp. $\mu = 2$) external branches of s segments increases (resp. decreases) for increasing values of s. In particular, it is interesting to observe that the probability of missing an external branch of s segments symmetrically corresponds to the probability of having two external branches of $n - s + 1$ segments. Also, in a random history, the probability of exactly one branch of s segments is equal to the probability of exactly one branch of $n - s + 1$ segments (Fig. 2). Numerical plots of the conditional probability of μ external branches of s segments given μ_r external branches of r segments are given in Fig. 3.

In Section 4, we have used known combinatorial results [4] on the set of peak entries of a permutation of size $n-1$, for characterizing the possible sets of missing external branches in a history of size n. Furthermore, for a given subset $\overline{\mathcal{E}}$ of $[3, n-1]$, Eq. (12) provides a recursive formula for calculating the probability that the external branches missing in a random history of size n are exactly those whose number of segments is listed in the set $\overline{\mathcal{E}}$ (Fig. 4).

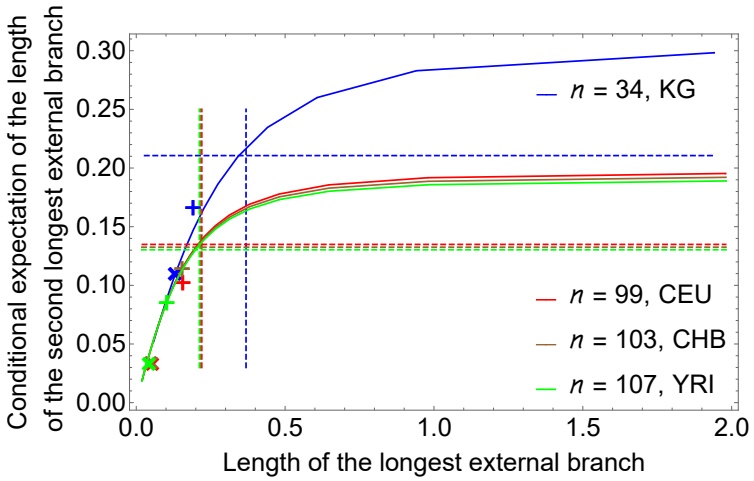

Figure 10: Estimates of the longest and second-longest external branches in coalescent trees. Red curves and symbols: mitochondrial data from *Homo sapiens*, CEU population (Central European origin); green: YRI population (Yoruba in Nigeria); brown: CHB population (Han-Chinese from Beijing). Blue curves and symbols: nuclear data from *Danio rerio*, CB population (Cooch Behar, India). n: sample size (# of chromosomes examined); +-symbol: estimate E_1 (using all available polymorphisms); $\times$-symbol: estimate E_2 (using only singletons). Solid lines: conditional expectation of the length of the second-longest branch, given the length of the longest branch. For a fixed value of n, every integer $s \in [\lceil n/2 \rceil, n-1]$ yields a point (x_s, y_s) of the corresponding theoretical line. The abscissa x_s is the expected time length of an external branch with s segments (1). The ordinate y_s is the expected time length of the second longest external branch, when the first longest has s segments (Section 5.2). Dashed lines: average length of the longest (vertical) and (un-conditioned) average length of the second-longest branch (horizontal). Refer to section 6 for more detailed description.

In Section 5, we have studied the probability of a given number of segments in the longest and second longest external branch of a history of size n selected at random. The first probability can be evaluated through Eqs. (16, 17). The joint probability is given by Eq. (25). In Section 5.1.2, we have derived Eqs. (22-24) for computing the conditional probability of the number of segments in the longest external branch of a random history of n leaves given the size $\omega \in [1, \lfloor n/2 \rfloor]$ of its smallest root subtree. Investigations of Section 5 yielded two findings on the structure of Yule histories that were not intuitively clear before. Surprisingly, imbalance does (almost) not influence the length of the longest external branch. For sufficiently large n, the distribution of the length of the longest external branch is roughly the same when we condition on different values of $\omega \geq 3$ (Fig. 8). Furthermore, we observe that conditioning on the discrete length s' of the longest external branch can strongly affect the distribution of the number of segments s'' in the second longest external branch. In Fig. 9 (left), we see that a relative small decrease of the value of s' from $s' = 49$ to $s' = 41$ results in a step function behavior of the conditional probability of s'' being smaller than a given value s.

The study of the number of segments in the external branches of a random Yule-distributed history can assist in the analysis of the time length of the external branches of a Kingman coalescent tree, which can be seen as Yule-distributed history in which the time length of the ith layer is an exponentially distributed variable. Importantly, the probability density function $f_s(x)$ of the time length of an external branch of s segments in a history of size n can be evaluated as in Eq. (2), and our study of the discrete length of external branches be used as a benchmark for biological data scenarios. In particular, when analyzing the mutation frequency spectrum of population samples, one may be interested in the question whether the number of singletons seen in a single haplotype significantly exceeds neutral expectation. For instance, this could be an indication of unaccounted population substructure. In Section 6, we applied our theoretical results on the length of the longest and second longest external branch to sequence samples from human and from a wild population of zebrafish. For the latter, an initial suspicion of sample contamination with non-genuine material was not confirmed with our results.

Several direction of research naturally arise from our work. First, it would be of interest to study the random variables considered in this article under different probability models—e.g., assuming a uniform distribution over the set of unordered histories of given size. Histories with a larger number of cherries have a smaller probability to be generated under the Yule process. Switching to a different distribution could for instance affect the probability of external branches of multiplicity 2 or the correlation between root imbalance and length of the longest external branch. Second, it would be important to extend the approach used in this article for studying the length of the external branches of a Yule-distributed history or coalescent tree to consider also the length of the branches ancestral to a cherry, which are associated with doubletons in the mutation frequency spectrum. Finally, we observe that encoding the ordered histories of size n as permutations of size $n-1$ allows to define a geometric structure over the set of histories of size n, when these are grouped together in equivalence classes according to their set of external branches. In particular, the admissible sets of missing external branches of the histories of size n, that is, the possible peak sets of the permutations of size $n-1$, are shown in [4] to form an abstract simplicial complex over the vertex set $[3, n-1]$. It seems natural to ask for possible biological interpretations of this complex.

Acknowledgments Support was provided by a Rita Levi-Montalcini grant to FD from the Ministero dell'Istruzione, dell'Università e della Ricerca and by a grant from the German Research Foundation (DFG SPP-1590) to TW.

Appendix: Proof of Eq. (15)

The formula in Eq. (15) can be rewritten as $a_{n,n-i} = \sum_{k=0}^{i} (-1)^k \, J(i,k) \, (n-k-1)!$, where $J(i,k) = \frac{(i+1)!}{(i-k+1)!}\binom{i}{k}$. Setting $\tilde{a}_{n,i} = a_{n,n-i}$, Eq. (15) is thus equivalent to

$$\tilde{a}_{n,i} = \sum_{k=0}^{i} (-1)^k \, J(i,k) \, (n-k-1)!. \tag{26}$$

From $a_{n,n-i} = a_{n,n-i-1} + 2(i+1)a_{n-1,n-i-1} + i(i+1)a_{n-2,n-i-1}$—which is Eq. (14) with $s = n-i$—we obtain the recurrence $\tilde{a}_{n,i+1} = \tilde{a}_{n,i} - 2(i+1)\tilde{a}_{n-1,i} - i(i+1)\tilde{a}_{n-2,i-1}$. Replacing $i+1$ by i in the latter expression yields

$$\tilde{a}_{n,i} = \tilde{a}_{n,i-1} - 2\,i\,\tilde{a}_{n-1,i-1} - (i-1)\,i\,\tilde{a}_{n-2,i-2}, \tag{27}$$

which we use to show formula (26) by induction on i. Substituting (26) in the right-hand side of (27), we find

$$\begin{aligned}
\tilde{a}_{n,i} \;=\; & \sum_{k=0}^{i-1} (-1)^k \, J(i-1,k) \, (n-k-1)! - 2i \sum_{k=0}^{i-1} (-1)^k \, J(i-1,k) \, (n-k-2)! \\
& -(i-1)\,i \sum_{k=0}^{i-2} (-1)^k \, J(i-2,k) \, (n-k-3)! \\
=\; & (n-1)!\,J(i-1,0) + (n-2)!\,[-J(i-1,1) - 2i\,J(i-1,0)] \\
& + \sum_{k=2}^{i-1} (-1)^k \, (n-k-1)![J(i-1,k) + 2\,i\,J(i-1,k-1) - (i-1)\,i\,J(i-2,k-2)] \\
& (n-i-1)![-2i\,(-1)^{i-1}\,J(i-1,i-1) - (i-1)\,i\,(-1)^{i-2}\,J(i-2,i-2)].
\end{aligned}$$

For $0 \le k \le i$, the coefficients of $(n-k-1)!$ in the latter expression can be easily seen to satisfy

$$\begin{aligned}
J(i-1,0) &= J(i,0), \\
-J(i-1,1) - 2i\,J(i-1,0) &= -J(i,1), \\
J(i-1,k) + 2\,i\,J(i-1,k-1) - (i-1)\,i\,J(i-2,k-2) &= J(i,k), \quad (2 \le k \le i-1), \\
-2i\,(-1)^{i-1}\,J(i-1,i-1) - (i-1)\,i\,(-1)^{i-2}\,J(i-2,i-2) &= (-1)^i J(i,i).
\end{aligned}$$

Hence, the formula obtained recursively for $\tilde{a}_{n,i}$ matches the sum in Eq. (26).

References

[1] 1000 Genomes Project Consortium, *A global reference for human genetic variation*, Nature 526 (2015): 68–74.

[2] M. Bibinger, *Notes on the sum and maximum of independent exponentially distributed random variables with different scale parameters*, Arxiv (2013): `http://arxiv.org/abs/1307.3945`.

[3] M.G.B. Blum, O. François, *Minimal clade size and external branch length under the neutral coalescent*, Adv. Appl. Probab. 37 (2005): 647–662.

[4] P. Bouchard, H. Chang, J. Ma, J. Yeh, Y.H. Yeh, *Value-Peaks of Permutations*, Electron. J. Comb. 17 (2010): article # R46.

[5] A. Caliebe, R. Neininger, M. Krawczak, U. Rosler, *On the length distribution of external branches in coalescence trees: genetic diversity within species*, Theor. Popul. Biol. 72 (2007): 245–252

[6] C. Diehl, G. Kersting, *External branch lengths of Λ-coalescents without a dust component*, Arxiv (2019): `https://arxiv.org/abs/1811.07653`.

[7] W.J. Ewens, *A note on the sampling theory for infinite alleles and infinite sites models*, Theor. Popul. Biol. 6 (1974): 143–148

[8] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. II, Wiley & Sons, New York (1971).

[9] F. Freund, M. Möhle, *On the time back to the most recent common ancestor and the external branch length of the Bolthausen-Sznitman coalescent*, Markov Process. Related Fields 15 (2009): 387–416.

[10] I.P. Goulden, D.M. Jackson, *Combinatorial enumeration*, Wiley, Chichester (1983).

[11] E.F. Harding, *The probabilities of rooted tree-shapes generated by random bifurcation*, Adv. Appl. Probab. 3 (1971): 44–77.

[12] R.R. Hudson, *Gene genealogies and the coalescent process*, Oxf. Surv. Evol. Biol. 7 (1990): 1–44.

[13] R.R. Hudson, *ms: a program for generating samples under neutral models*, Bioinformatics 18 (2002): 337–338.

[14] S. Karlin, J. McGregor, *The number of mutant forms maintained in a population*, in Proc. of the Fifth Berkeley Symposium on Mathematics, Statistics and Probability 4 (1967), 415–438.

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

[16] J. F. C. Kingman, *Origins of the coalescent: 1974–1982*, Genetics 156 (2000): 1461–1463.

[17] N.A. Rosenberg, *The Mean and Variance of the Numbers of r-Pronged Nodes and r-Caterpillars in Yule-Generated Genealogical Trees*, Ann. Comb. 10 (2006): 129–146.

[18] C. Semple, P. Daniel, W. Hordijk, R.D.M. Page, M. Steel, *Supertree algorithms for ancestral divergence dates and nested taxa*, Bioinformatics 20 (2004): 2355–2360.

[19] M. Steel, A. McKenzie, *Properties of phylogenetic trees generated by Yule-type speciation models*, Math. Biosci. 170 (2001): 91–112.

[20] J. Suurväli, A. Whitley, Y. Zheng, K. Gharbi, M. Leptin, T. Wiehe, *The laboratory domestication of zebrafish: from diverse populations to inbred substrains*, Biorxiv (2019): `https://www.biorxiv.org/content/10.1101/706382v1`.

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

[22] G.U. Yule, *A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis, F.R.S.*, Philos. Trans. Roy. Soc. Lond. Ser. B 213 (1924): 21–87.

[23] S. Zhu, J.H. Degnan, M. Steel, *Clades, clans, and reciprocal monophyly under neutral evolutionary models*, Theor. Popul. Biol. 79 (2011): 220–227.