

# The Probability of Joint Monophyly of Samples of Gene Lineages for All Species in an Arbitrary Species Tree

ROHAN S. MEHTA,<sup>1</sup> MIKE STEEL,<sup>2</sup> and NOAH A. ROSENBERG<sup>3</sup>

## ABSTRACT

Monophyly is a feature of a set of genetic lineages in which every lineage in the set is more closely related to all other members of the set than it is to any lineage outside the set. Multiple sets of lineages that are separately monophyletic are said to be reciprocally monophyletic, or jointly monophyletic. The prevalence of reciprocal monophyly, or joint monophyly (JM), has been used to evaluate phylogenetic and phylogeographic hypotheses, as well as to delimit species. These applications often make use of a probability of JM under models of gene lineage evolution. Studies in coalescent theory have computed this JM probability for small numbers of separate groups in arbitrary species trees and for arbitrary numbers of separate groups in trivial species trees. In this study, generalizing existing results on monophyly probabilities under the multispecies coalescent, we derive the probability of JM for arbitrary numbers of separate groups in arbitrary species trees. We illustrate how our result collapses to previously examined cases. We also study the effect of tree height, sample size, and number of species on the probability of JM. We obtain relatively simple lower and upper bounds on the JM probability. Our results expand the scope of JM calculations beyond small numbers of species, subsuming past formulas that have been used in simpler cases.

Keywords: coalescent, gene tree, monophyly, probability, species tree.

## 1. INTRODUCTION

Evaluations of the prevalence of reciprocal or joint monophyly (JM) in sampled gene genealogies have been useful in a variety of studies in phylogenetics, phylogeography, and molecular ecology. They have been used for identifying units for conservation (Moritz, 1994), analyzing differing phylogeographic patterns across species (Carstens and Richards, 2007), evaluating the distinctiveness of taxa

---

<sup>1</sup>Department of Physics, Emory University, Atlanta, Georgia, USA.

<sup>2</sup>Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand.

<sup>3</sup>Department of Biology, Stanford University, Stanford, California, USA.

An earlier draft of this article was posted as a preprint at bioRxiv (DOI: 10.1101/2021.12.16.473006).

(Kubatko et al., 2011), and providing context for estimation of species divergence times (Arbogast et al., 2002). JM is fundamental to genealogical perspectives on species delimitation (Hudson and Coyne, 2002; De Queiroz, 2007).

Central to the application of JM is a theoretical prediction of the probability that genealogies show JM as a function of evolutionary parameters. Many studies have used monophyly computations in examinations of the evolutionary relationships among recently diverged species (Birky et al., 2005; Carstens and Knowles, 2007; Carstens and Richards, 2007; Syring et al., 2007; Jansen et al., 2010; Kubatko et al., 2011; Bergsten et al., 2012; Rabeling et al., 2014). These computations have often made use of theoretical results of Rosenberg (2003, 2007), which consider the probability that gene lineages in two populations are jointly monophyletic as a function of population divergence times.

For example, Kubatko et al. (2011) used such computations to assess the taxonomic distinctiveness of two species of *Sistrurus* rattlesnake, each of which was divided into three subspecies. They considered two types of comparisons for each set of three subspecies:  $\hat{\phi}_t$ , that one subspecies was distinct from a hypothetical clade containing the other two, and next, that the two remaining subspecies were distinct from each other. The result of these comparisons was the establishment of the distinctiveness of a seriously threatened subspecies (*Sistrurus catenatus catenatus*), as well as of varying levels of distinctiveness among the remaining subspecies.

Because the probability formulas available were limited to two groups, Kubatko et al. (2011) were restricted to performing a hierarchical set of analyses in which distinctiveness of one subspecies from a taxon that combined the other two subspecies was assessed, followed by distinctiveness of one of the two previously combined taxa from the other. JM computations were likewise restricted to these two hierarchical pairs of subspecies. Although the hierarchical analysis did produce the desired determinations, the analysis of Kubatko et al. (2011) would have been enriched by the ability to simultaneously consider the distinctiveness of one *S. catenatus* subspecies from the two other *S. catenatus* subspecies, rather than being restricted to a hierarchical pairwise comparison that might produce inaccurate probabilities as a result of merging present-day samples from populations that have diverged in the past (Mehta et al., 2016).

Simultaneously studying the relationship between the *S. catenatus* subspecies in relation to the other *Sistrurus* species would have required mathematical results that could accommodate up to six simultaneous monophyly events. Other similar studies involving more than two species or groups have also been restricted to pairwise computations (Carstens and Richards, 2007; Baker et al., 2009; Neilson and Stepien, 2009; Bergsten et al., 2012).

Three theoretical developments now place the possibility of a JM probability computation within reach for taxa related according to an arbitrary species tree. First, Zhu et al. (2011) computed the probability of JM for an arbitrary number of groups for lineages originating within a single population rather than evolving on a species tree. Next, Mehta et al. (2016) found the probability of JM in a species tree of arbitrary size, considering two classes of lineages. Finally, Mehta and Rosenberg (2019) found the full probability of JM for the lineages of species evolving on species trees with three or four species. The  $\hat{\phi}_t$  extension generalized to an arbitrary number of groups whose lineages must be jointly monophyletic. The second produced an algorithm that allows for an arbitrary species tree. The third provided the simplest cases for a synthesis of the other two extensions.

In this study, we obtain the complete generalization: the probability of JM for an arbitrary number of groups in an arbitrary species tree. Figure 1 illustrates the results of Zhu et al. (2011) and Mehta et al.

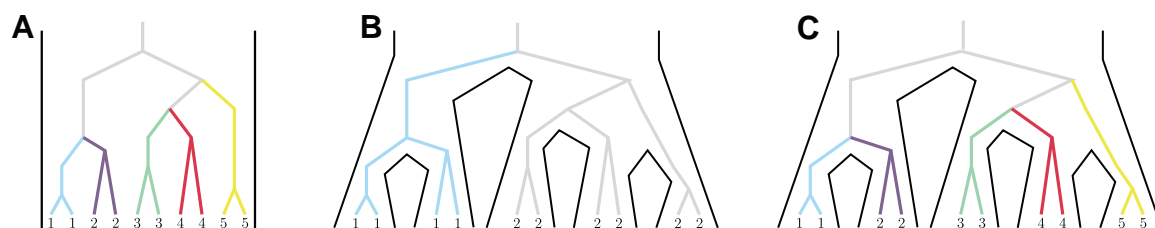


FIG. 1. Schematic of the general joint monophyly calculation. (A) Zhu et al. (2011) computed the probability of joint monophyly of arbitrarily many groups in a single population. (B) Mehta et al. (2016) computed the probability of joint monophyly of two groups in an arbitrary species tree. (C) Here, we compute the probability of joint monophyly of arbitrarily many groups in an arbitrary species tree. In each panel, the numbers and colors indicate groups, and the black lines represent a species tree.

(2016) and how they relate to our recursive computation. We study the effect of species tree parameters, such as tree height and sample size, on this probability. Because the result is computationally intensive, we provide lower and upper bounds on the probability of JM, as well as an alternative, potentially faster, method for numerical computation. Finally, we provide software that encodes the new formulas.

## 2. PRELIMINARIES FOR THE RECURSIVE APPROACH

### 2.1. Model and notation

We consider a binary species tree  $T$  on the species label set  $S$ , consisting of a topology and a set of branch lengths. For each leaf  $S_i$  of  $T$ , we specify a sample size  $s_i \geq 1$ . We use the multispecies coalescent to track the sampled lineages as they travel back in time  $t$  up the species tree. Section 2 describes the terminology and construction of our coalescent model, closely following Mehta et al. (2016) and Mehta and Rosenberg (2019). Figure 2 illustrates some of the notation.

### 2.2. Lineage labels

Genetic lineages are labeled according to the species from which they are sampled. All lineages for a particular species have the same label, and each species has a unique label. We label the species  $1, 2, \dots, k$ , where the number of species is  $|S| = k$ . Lineages that result from a coalescence between lineages of differing labels are called *mixed* lineages and are assigned label  $k + 1$ .

### 2.3. Species tree branches

In our coalescent framework, the bottom of the tree is the present, at time 0, and time increases up the tree, further into the past. Viewed backward in time, an internal node of the species tree represents an event at which two species merge into an ancestral species. Gene lineages enter species tree nodes from the bottom and exit them at the top as time progresses into the past. Because a one-to-one correspondence exists between species tree branches and nodes, we refer to a node and its immediate ancestral branch interchangeably. A particular node  $x$  has lineages enter from both branches directly below it. The length of branch  $x$  is  $T_x$ , the length of time associated with node  $x$ .  $T_x$  is measured in units of  $N$  generations, where  $N$  is the haploid population size on branch  $x$ ; this size is assumed to be constant over all species tree branches. Larger sizes correspond to smaller values of  $T_x$  in coalescent units. The root branch of  $T$  is assumed to contain any coalescence events that have not occurred below the root. Biologically, this assumption is that of a universal common ancestor for all gene lineages, and it is implemented by setting the root branch length to infinity.

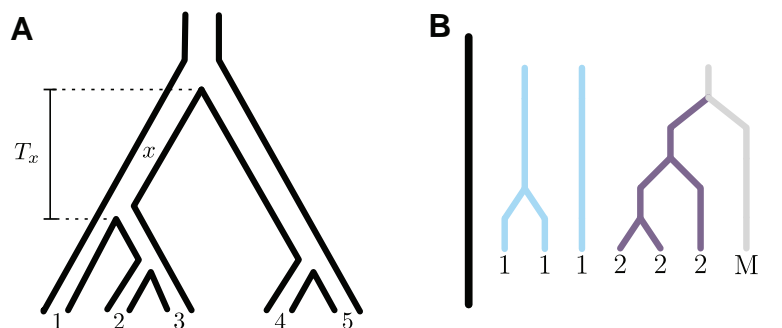


FIG. 2. Notation for input and output lineages. (A) An example of a species tree  $T$ , with 5 species and species label set  $S = \{1, 2, 3, 4, 5\}$ . An example branch  $x$  is highlighted with its branch length  $T_x$ . (B) Coalescences happening within a single branch [branch  $x$  in (A)] of a species tree. In this diagram, three lineages from species 1, three lineages from species 2, and a single mixed lineage enter the branch, and two lineages from species 1 and one mixed lineage exit the branch. Supposing this branch comes from a 6-species tree, the input state is  $n_x^I = (3, 3, 1)$ , and the output state is  $n_x^O = (2, 2, 1)$ . The label 1 is a surviving label, and the label 2 is a lost label.

## 2.4. Input and output states

An output state of a branch  $x$  is a list of nonnegative integers that records the numbers of lineages of each label exiting the branch from the top. In our model, the output state is a random variable. This random variable is a vector  $Z_x$  of length  $k + 1$  whose  $i$ th element is the number of output lineages that possess label  $i$ . A particular instance of this random variable is denoted  $n_x^O$ .

Similarly, an input state for a branch is a list of nonnegative integers that records the numbers of lineages of each label entering the node from the two branches immediately below it. The input state for an internal branch  $x$  is the sum of the two output states for its descendant branches  $x_L$  and  $x_R$ . A particular instance of an input state is  $n_x^I = n_{x_L}^O + n_{x_R}^O$ . Figure 2B provides an example species tree node with its inputs and outputs.

## 2.5. Coalescence sequences

A coalescence sequence is an ordered sequence of coalescence events. For example, consider lineages A, B, C, D, and E. One possible coalescence sequence involving these lineages is  $\{(A, B), (AB, C), (ABC, D), (ABCD, E)\}$ , where A and B coalesce first, then C coalesces with the resulting AB lineage, then D coalesces with the resulting ABC lineage, and finally E coalesces with the resulting ABCD lineage.

Coalescence sequences involving disjoint sets of lineages can be combined into a single coalescence sequence that contains all the coalescences from both sequences, a procedure termed *interweaving* (e.g., Rosenberg, 2003). The same set of coalescence sequences can be interwoven in different ways to form different interwoven coalescence sequences (Fig. 3).

## 2.6. Joint monophyly

Consider a subtree  $T_x$  of  $T$ , defined as the node  $x$ , all of its descendant nodes, and all branches associated with those nodes (including the branch immediately ancestral to  $x$ ). For JM to be achieved, each coalescence in  $T_x$  must be in one of four mutually exclusive classes:

1. The coalescence is between two lineages that have the same label (an intralabel coalescence), and neither label is mixed.

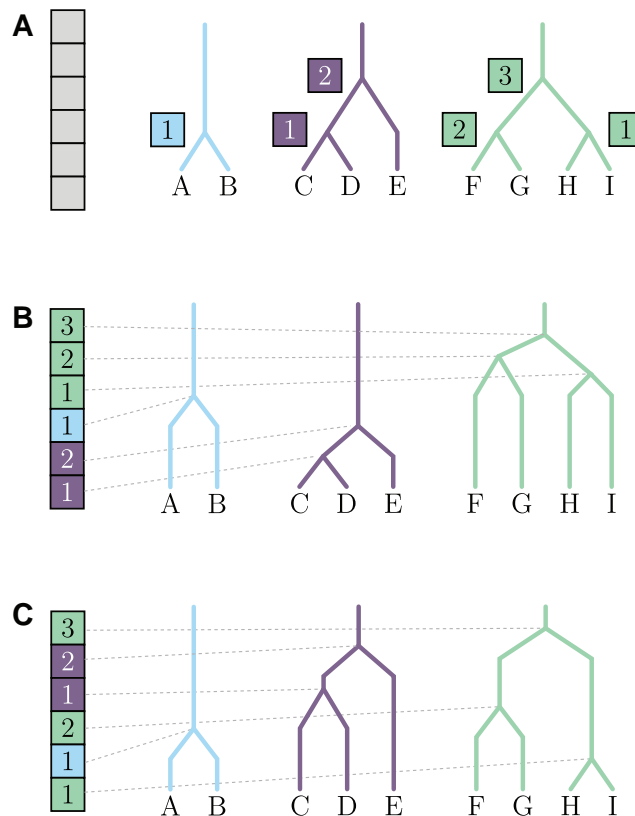


FIG. 3. Interweaving of coalescence sequences. (A) Three coalescence sequences. The sequences are represented in three colors. Within a sequence, coalescences occur in a specific order, indicated by numbers within colors. Each of the six coalescences must occur in the interwoven sequence, represented by the gray blocks. Hence, each coalescence must be mapped to one of the gray blocks, with order increasing from bottom to top for each sequence. (B, C) Two different ways to interweave the sequences from (A).

2. The coalescence is between two lineages with different labels (an interlabel coalescence), neither label is mixed, and both labels have only one existing lineage at the time of the coalescence.
3. The coalescence is between two lineages with different labels, exactly one of which is mixed, and the other label has only one existing lineage at the time of the coalescence.
4. The coalescence is between two mixed lineages.

We define the JM event  $E_x$  for gene lineages in the subtree  $T_x$ ;  $E_x$  is the event that all coalescences in  $T_x$  are in one of the four classes. If at least one coalescence is not in one of these classes, then JM is violated.

## 2.7. Combinatorial functions

We use several combinatorial functions in our calculation. First,  $g_{i,j}(T)$  is the probability that  $i$  lineages coalesce such that at time  $T$ , the number of ancestral lineages is precisely  $j$ . From Equation (6.1) of Tavaré (1984):

$$g_{i,j}(T) = \sum_{k=j}^i e^{-k(k-1)T/2} \frac{(2k-1)(-1)^{k-j} j^{(k-1)} i^{(k)}}{j!(k-j)!i^{(k)}} \quad (1)$$

where  $a_{(k)} = a(a+1)\cdots(a+k-1)$  and  $a_{[k]} = a(a-1)\cdots(a-k+1)$  for  $k \geq 1$ , and  $a_{(0)} = a_{[0]} = 1$ . This function is nonzero when  $i \geq j \geq 1$  and  $T \geq 0$ . We define  $g_{0,0}(T) = 1$ , and we write  $g_{i,j}(1)$  for  $\lim_{T \rightarrow 1} g_{i,j}(T)$ , noting that  $g_{i,j}(1) = 1$  for  $i = j$ .

Second, the number of coalescence sequences that reduce  $n$  lineages to  $k$  lineages is

$$I_{n,k} = \frac{n!(n-1)!}{2^{n-k} k! (k-1)!} \quad (2)$$

This function, from Equation (4) of Rosenberg (2003), is nonzero for  $n \geq k \geq 1$ , and we define  $I_{0,0} = 1$ .

Third, the multinomial coefficient

$$W_k(r_1, r_2, \dots, r_k) = \frac{r_1! + r_2! + \dots + r_k!}{r_1! r_2! \dots r_k!} \quad (3)$$

from Mehta and Rosenberg (2019), is the number of ways that  $k$  coalescence sequences of lengths  $r_1, r_2, \dots, r_k$  coalescent events can be ordered, or interwoven, to create an encompassing coalescence sequence that contains them all as subsequences. This function is defined for  $r_i \geq 0$ ,  $i = 1, 2, \dots, k$ .

Finally,  $Z(s_1, s_2, \dots, s_k)$  is the probability that in a single population in which  $k$  groups are present,  $k$  groups of  $s_1, s_2, \dots, s_k$  gene lineages coalesce to a single lineage while preserving JM of each of the  $k$  groups. This function is taken from Theorem 5.1 of Zhu et al. (2011), as follows.

Suppose that  $A_1, A_2, \dots, A_k$  represent sets of lineages for groups  $1, 2, \dots, k$ , respectively. Under JM of groups  $1, 2, \dots, k$ , each group  $i$  possesses a single lineage  $a_i$  ancestral to all lineages in  $A_i$ . The lineages  $a_i$  possess some labeled topology  $T_k$  from the set of all possible labeled topologies  $rb(k)$  (rooted binary trees). We compute the probability that the  $k$  groups are jointly monophyletic and that their associated single-lineage ancestors possess labeled topology  $T_k$ , and then sum over all possible  $T_k$  to obtain the total probability of JM of the  $k$  groups.

Let  $n = \sum_{i=1}^k s_i$  be the total number of lineages across all groups. Let  $I(T_k)$  be the set of internal nodes of  $T_k$ . For an internal node  $v \in I(T_k)$ , let  $I_v(A_i)$  denote the indicator function that lineage  $a_i$  is a descendant of  $v$  in  $T_k$ . The joint probability of JM of the  $k$  groups and labeled topology  $T_k$  is as follows:

$$Z(s_1, s_2, \dots, s_k; T_k) = \frac{2^{k-1} \prod_{i=1}^k s_i!}{n!} \prod_{v \in I(T_k)} \frac{1}{\prod_{i=1}^k s_i I_v(A_i) - 1} \quad (4)$$

Summing over all  $(2n-2)!/[2^{n-1}(n-1)!]$  possible  $T_k$  in  $rb(k)$ , the total probability of JM is

$$Z(s_1, s_2, \dots, s_k) = \sum_{T_k \in rb(k)} Z(s_1, s_2, \dots, s_k; T_k) \quad (5)$$

Our notation sometimes leads to values of 0 for some of the arguments  $s_i$  of the function in Equation (5); such cases have the interpretation that there is no corresponding label  $i$  among the leaves of  $T_k$ . In those cases, the quantity is properly computed by dropping those arguments.

### 3. MATHEMATICAL RESULTS

For species tree internal node  $x$ , we can compute the probability of the JM event  $E_x$  and a particular output state  $n_x^O$  by recursive decomposition as follows:

$$P(E_x, n_x^O) = \sum_{n_x^I} P(E_x, n_x^O | E_{x_L}, E_{x_R}, n_x^I) P(E_{x_L}, E_{x_R}, n_x^I) \quad (6)$$

$$\sum_{n_{x_L}^O} \sum_{n_{x_R}^O} P(E_x, n_x^O | E_{x_L}, E_{x_R}, n_{x_L}^O, n_{x_R}^O) P(E_{x_L}, n_{x_L}^O) P(E_{x_R}, n_{x_R}^O)$$

where  $x_L$  and  $x_R$  are the daughter nodes of  $x$ , and the second step results from independence of these nodes and the fact that  $n_x^I = n_{x_L}^O + n_{x_R}^O$ . Taking  $x$  to be the species tree root,  $P(E_{\text{root}}, n_{\text{root}}^O) = (0 \dots 0)$  is the JM probability for the entire gene genealogy.

To compute  $P(E_{\text{root}})$ , we use a pruning algorithm, a familiar approach in phylogenetics in general (Felsenstein, 2004, p. 253). First calculating the probability  $P(E_x, n_x^O | E_{x_L}, E_{x_R}, n_{x_L}^O, n_{x_R}^O)$  the probability of obtaining the JM event  $E_x$  and an output state  $n_x^O$  given the events  $E_{x_L}$  and  $E_{x_R}$  and their corresponding output states  $n_{x_L}^O$  and  $n_{x_R}^O$ . We can apply this probability to the root of the tree and then proceed recursively to the leaves, whose inputs are known, ending the recursion. Given  $E_{x_L}$ ,  $E_{x_R}$ , and their output states, the probability of event  $E_x$  and its output state is the probability that all coalescences that occur in the branch  $x$  satisfy JM and result in the specified output state. We can compute this probability by specifying an input state, computing the probability that JM is preserved on branch  $x$  given the input state and output state, and summing over all possible input states for branch  $x$ .

The probability that JM is preserved in a branch with a specified input state and output state requires computation of two quantities: (i) the probability that the correct number of coalescences occurs to convert the input state into the output state, and (ii) among coalescence sequences with the correct number of coalescences, the fraction that satisfy JM.

For (i), the probability that the correct number of coalescences occurs is  $g_{|n_L^I|, |n_R^O|}(T_x)$  (Section 2.7). For (ii), to count the coalescence sequences, the calculation is more involved. It is useful to first classify the  $k$  input labels into two categories: surviving and lost.

#### 3.1. Surviving labels and lost labels

Consider a branch  $x$  with input state  $n_x^I$  and output state  $n_x^O$ . Consider a label  $i \in 1, 2, \dots, k$ . The number of lineages of label  $i$  in the input state is denoted  $n_{x,i}^I$  and its number of lineages in the output state is denoted  $n_{x,i}^O$ . The total number of lineages in an input or output state is denoted  $|n_x^I|$  or  $|n_x^O|$ , respectively.

Suppose  $n_{x,i}^I > 0$ . Two possibilities then exist for label  $i$ :  $n_{x,i}^O > 0$  or  $n_{x,i}^O = 0$ . If  $n_{x,i}^O > 0$ , then label  $i$  is said to be a surviving label on branch  $x$ . To preserve JM on branch  $x$ , lineages of surviving label  $i$  are permitted to undergo intralabel coalescences on the branch, but not interlabel coalescences.

If  $n_{x,i}^O = 0$ , then label  $i$  is said to be a lost label on branch  $x$ . To preserve JM on branch  $x$ , lineages of lost label  $i$  undergo intralabel coalescences until only one lineage of label  $i$  remains. This final lineage undergoes an interlabel coalescence.

In the branch represented in Figure 2B, label 1 survives, whereas label 2 is lost.

#### 3.2. Number of coalescence sequences for each surviving label

Our general approach for counting permissible coalescence sequences within a branch is to split the coalescences within the branch into multiple subsequences that we know how to count, and to then interweave those subsequences together. First, we consider sequences involving surviving labels. Under JM, each lineage in a surviving label must coalesce only with other lineages that possess that same label. Thus, the set of all input lineages of a particular surviving label  $i$ , the coalescences of those lineages, and the output lineages of label  $i$  can be used to define a coalescence subsequence for label  $i$ . The number of distinct

coalescence subsequences for surviving label  $i$  is the number of ways that the  $n_{x\hat{\alpha}}^i$  input lineages of label  $i$  can coalesce to the correct number of output lineages of label  $i$ , or  $n_{x\hat{\alpha}}^0$ . This number of subsequences is  $I_{n_{x\hat{\alpha}}^i, n_{x\hat{\alpha}}^0}$  (Eq. [2]). We compute this quantity for each surviving label.

### 3.3. Enumerating partitions containing lost labels and mixed lineages

We next count coalescence subsequences that involve lost labels and mixed lineages. Unlike for surviving lineages, because a lost label must undergo an interlabel coalescence, coalescence subsequences involving lost labels only produce output mixed lineages. Hence, each output mixed lineage must result from a coalescence subsequence involving (1) at least two mixed lineages and no lost labels, (2) at least two lost labels and no mixed lineages, or (3) at least one lost label and at least one mixed lineage.

To account for every possible coalescence subsequence in one of these three categories, we must assign each output mixed lineage to an element of a partition of the set of lost labels and input mixed lineages. Thus, we partition the input lineages, assigning to each element of the partition a single output mixed lineage. A coalescence subsequence exists for each element of the partition.

We count the number of distinct types of lineages, among the input lineages with lost labels and the input mixed lineages. This quantity equals  $\ell + m_I$ :  $\ell$  input lost labels and  $m_I$  individual input mixed lineages. The number of elements of the partition of output lineages is  $m_O$ : one element for each of the  $m_O$  individual output mixed lineages. Thus, we are partitioning  $\ell + m_I$  labeled elements into  $m_O$  nonzero categories. In particular, these partitions are the ways to place  $\ell + m_I$  labeled balls into  $m_O$  unlabeled boxes, such that each box contains at least one ball (Loehr, 2017). The number of these partitions are Stirling numbers of the second kind,  $S_2(\ell + m_I, m_O)$ . An algorithm for producing these partitions is presented in Knuth (2011).

However, two additional conditions must be met.

1.  $m_O = m_I + \ell$ .
2. No element of the partition can consist solely of a single one of the  $\ell$  lost labels.

In the first condition, the number of output mixed lineages is bounded above by the number of input mixed lineages plus the maximal number of additional mixed lineages that can be produced by coalescences involving the lost labels. Lost labels whose coalescences involve the  $m_I$  input mixed lineages do not generate additional output mixed lineages; however, lost labels whose coalescences involve other lost labels do generate additional output mixed lineages. The maximal number of output mixed lineages that can be generated in this way is  $\ell$ , if the maximal number of pairs of lost labels coalesce.

The second condition codifies the requirement that no output mixed lineage is generated purely by coalescences within a single lost label. Each lost label must coalesce with others or with mixed lineages.

Once all possible partitions of the  $\ell + m_I$  labeled elements into  $m_O$  unlabeled nonempty sets are enumerated, we filter these partitions by the conditions 1 and 2, retaining only those partitions that satisfy both criteria. We denote these partitions to be permissible partitions. For each partition retained, we next describe the enumeration of the coalescence subsequences associated with an element of the partition.

### 3.4. The number of coalescence subsequences for each element of a partition of the set of lost labels and mixed lineages

Denote by  $P$  the set of permissible partitions of the set  $L \cup M_I$ , where  $L$  is the set of  $\ell$  lost labels and  $M_I$  is the set of  $m_I$  input mixed lineages. Let  $P$  be a partition in  $P$ .

Consider an element  $p$  of  $P$ . This element is associated with a set  $L_p \subseteq L$  of  $\ell_p$  lost labels and a set  $M_p \subseteq M_I$  of  $m_p$  input mixed lineages.  $L_p$  or  $M_p$  is possibly empty, but they cannot both be empty. Element  $p$  corresponds to a coalescence subsequence that starts with  $\sum_{j \in L_p} r_j + m_p$  lineages and ends with a single mixed lineage, where  $r_j$  is the number of input lineages of (lost) label  $j$ .

Following Section 2.7, the number of subsequences that coalesce  $\sum_{j \in L_p} r_j + m_p$  lineages to a single lineage is  $I_{\sum_{j \in L_p} r_j, m_p}$  [Eq. (2)]. The fraction of these subsequences that satisfy JM is  $Z(v_p)$ , where  $Z$

is the probability of JM of an arbitrary number of groups in a single population [Eq. (5)]. The argument  $v_p$  is constructed as a vector of length  $k + m_p$ . For elements  $i$  from 1 to  $k$ ,  $v_i = r_i$  if  $i \in L_p$  and  $v_i = 0$  if  $i \notin L_p$ . The last  $m_p$  elements all equal 1. For example, consider a seven-species tree. If partition element  $p$  contains lost labels 1 and 6 and three input mixed lineages, then  $v_p = (r_1, 0, 0, 0, 0, r_6, 1, 1, 1)$ .

Combining the number of subsequences that start from the input lineages in  $p$  and coalesce to a single lineage with the fraction of those subsequences that satisfy JM gives the total number of subsequences that both have the correct number of coalescences and that satisfy JM:

$$J_p = I_{\left(\sum_{j \in L_p} r_j\right) + m_p} Z(v_p) \quad (7)$$

### 3.5. The number of coalescence sequences associated with a set of surviving labels and a partition of the set of lost labels and mixed lineages

We now count, within a single branch of species tree  $T$ , coalescence sequences that contain specific subsequences associated with surviving labels and specific subsequences associated with partitions of lost labels and mixed lineages. Let  $U$  be the set of surviving labels in a branch, and let  $P$  be a partition of the set of lost labels and mixed lineages for the branch. Each of the  $|U|$  surviving labels and each element  $p$  of partition  $P$  creates a coalescence subsequence that must be interwoven with the other such subsequences. There are  $|U| + |P|$  such subsequences. For  $1 \leq i \leq |U|$ , the number of coalescences is  $s_i - r_i$ , noting that  $U_i$  is the  $i$ th surviving label (enumerated in arbitrary order) and abbreviating  $s_i = n_{x \cup U_i}^I$  and  $r_i = n_{x \cup U_i}^O$  for convenience.

For each  $i$  with  $|U| + 1 \leq i \leq |U| + |P|$ , the number of coalescences is  $|P_i| - |U|$  coalescences, where  $P_j$  is the  $j$ th element of  $P$  (again enumerated in arbitrary order). Hence, the number of ways to interweave the  $|U| + |P|$  coalescence subsequences is [from Eq. (3)]

$$W_{|U|+|P|} s_1 - r_1 \hat{\otimes} s_2 - r_2 \hat{\otimes} \dots \hat{\otimes} s_{|U|} - r_{|U|} \hat{\otimes} |P_1| - 1 \hat{\otimes} |P_2| - 1 \hat{\otimes} \dots \hat{\otimes} |P_{|P|}| - 1 : \quad (8)$$

Multiplying the number of ways of interweaving the coalescence subsequences by the product of the numbers of ways of constructing the various subsequences, the total number of sequences that satisfy JM for a given input state and output state is

$$C_{n_x^I \hat{\otimes} n_x^O} = \sum_{P \in \mathcal{P}} \prod_{i \in U} I_{s_i \hat{\otimes} r_i} \prod_{p \in P} J_p W_{|U|+|P|} s_1 - r_1 \hat{\otimes} s_2 - r_2 \hat{\otimes} \dots \hat{\otimes} s_{|U|} - r_{|U|} \hat{\otimes} |P_1| - 1 \hat{\otimes} |P_2| - 1 \hat{\otimes} \dots \hat{\otimes} |P_{|P|}| - 1 : \quad (9)$$

The product over elements of  $U$  is the number of coalescence sequences involving surviving labels. The product over elements of  $P$  is the number of coalescence sequences for a particular partition of lost labels and mixed lineages, and the sum over all  $P$  accounts for all possible partitions in  $\mathcal{P}$ .

If there are no surviving labels, then the product over elements of  $U$  is trivial, equal to 1. If all labels are surviving labels, then trivially, only a single partition in  $\mathcal{P} \subseteq \mathcal{P}$  is possible. We omit the sum over this partition  $P$ , and note that  $J_p = 1$  trivially for the single element  $p$  of this trivial partition  $P$ . Equation (9) becomes

$$C_{n_x^I \hat{\otimes} n_x^O} = \prod_{i \in U} I_{s_i \hat{\otimes} r_i} W_{|U|} s_1 - r_1 \hat{\otimes} s_2 - r_2 \hat{\otimes} \dots \hat{\otimes} s_{|U|} - r_{|U|} : \quad (10)$$

### 3.6. Completing the computation

The total number of coalescence sequences in a branch given an input state and an output state is  $I_{n_x^I \hat{\otimes} n_x^O}$  [Eq. (2)]. The number that satisfy JM is  $C_{n_x^I \hat{\otimes} n_x^O}$ , following Equation (9). From Equation (1), the probability of obtaining a particular number of coalescences in a branch of length  $T_x$  is  $g_{n_x^I \hat{\otimes} n_x^O}(T_x)$ .

We conclude that in Equation (6) for the probability of JM in branch  $x$  together with an output state, the recursive step that computes the conditional probability of JM and the output state given that JM is maintained in the daughter branches  $x_L$  and  $x_R$  and given the input state is

$$P(E_x \hat{\otimes} n_x^O | E_{x_L} \hat{\otimes} E_{x_R} \hat{\otimes} n_x^I) = g_{n_x^I \hat{\otimes} n_x^O}(T_x) \frac{C_{n_x^I \hat{\otimes} n_x^O}}{I_{n_x^I \hat{\otimes} n_x^O}} \hat{\otimes} \quad (11)$$

where  $C_{n_x^I \hat{\otimes} n_x^O}$  is from Equation (9) and  $I_{n_x^I \hat{\otimes} n_x^O}$  is from Equation (2). This result, applied recursively starting from  $x = \text{root}$  with  $n_x^O = (0 \hat{\otimes} \dots \hat{\otimes} 0)$ , yields the probability of JM over all species  $1, 2, \dots, k$ .

### 3.7. Deriving previous results

We can use Equation (11) to derive previously known results on the probability of JM under the multispecies coalescent. In this section, we proceed through several special cases.

**3.7.1.  $k$  groups in one population.** This case has only one branch  $x$ , corresponding to the single population;  $x$  has no daughter nodes. There is only one possible input state into  $x$ :  $n_x^I = (s_1 \hat{a}_2 \hat{a} \dots \hat{a}_k \hat{a})$ , where  $s_i$  is the sample size of group  $i$ . The output state is  $n_x^O = (0 \hat{a} \dots \hat{a} \hat{a})$ , with size  $j n_x^O j = 1$ . Branch  $T_x$  has infinite length. The summation in Equation (6) is trivial, and applying Equation (11), we obtain

$$\begin{aligned} P(E_x \hat{n}_x^O) &= P(E_x \hat{n}_x^O j E_{x_L} \hat{E}_{x_R} \hat{n}_x^I) \\ &= g_{j n_x^I \hat{a}} (1) \frac{C_{n_x^I \hat{n}_x^O}}{I_{j n_x^I \hat{a}}} : \end{aligned}$$

The labels  $1, 2, \dots, k$  are all lost, and there is only one output mixed lineage  $m_1$ . Hence, the set of partitions  $P$  of lost labels and input mixed lineages into output mixed lineages consists of a single partition  $P = \{p\}$ , with single element  $p = \{1 \hat{a} \dots \hat{a} \hat{a}\} \rightarrow m_1$ . Thus, when we use Equation (9), we obtain  $C_{n_x^I \hat{n}_x^O} = J_p W_1 (j n_x^I j - 1)$ . Noting that  $g_{\hat{a}} (1) = 1$ , and from Equation (3),  $W_1 (j n_x^I j - 1) = 1$ , we find

$$P(E_x \hat{n}_x^O) = \frac{J_p W_1 (j n_x^I j - 1)}{I_{j n_x^I \hat{a}}} = \frac{J_p}{I_{j n_x^I \hat{a}}} :$$

Using our notation from Section 3.4, the partition vector is  $v_p = (s_1 \hat{a}_2 \hat{a} \dots \hat{a}_k \hat{a})$ . We use Equation (7) to obtain

$$\begin{aligned} P(E_x \hat{n}_x^O) &= \frac{I_{j n_x^I \hat{a}} Z(v_p)}{I_{j n_x^I \hat{a}}} = Z(v_p) \\ &= Z(s_1 \hat{a}_2 \hat{a} \dots \hat{a}_k \hat{a}) : \end{aligned} \quad (12)$$

Note that  $Z(s_1, s_2, \dots, s_k)$  is exactly the quantity in Equation (5), and we recover the result from Zhu et al. (2011).

**3.7.2. General term for a leaf node.** Next, we consider a series of cases in which monophyletic groups correspond to the lineages of species (Table 1). A leaf node has exactly one input label  $i$  and exactly one surviving label  $i$ , and it has no other types of label. The input state is  $n_x^I = (0 \hat{a} \dots \hat{a}_i \hat{a} \dots \hat{a})$ , and the output state is  $n_x^O = (0 \hat{a} \dots \hat{a}_i \hat{a} \dots \hat{a})$ .

Thus, for a leaf node, using Equation (8), the partition set  $P$  is trivial, producing Equation (10). The set of surviving lineages is  $U = \{i\}$ . Using Equation (10) along with Equation (3), we obtain

$$C_{n_x^I \hat{n}_x^O} = I_{s_i \hat{a}_i} W_1 (s_i - r_i) = I_{s_i \hat{a}_i} :$$

A leaf node has no daughter nodes, and the input state is therefore known; trivially, Equation (6) has a single term. Using Equation (11), we have

$$\begin{aligned} P(E_x \hat{n}_x^O) &= g_{j n_x^I \hat{n}_x^O j} (T_x) \frac{C_{n_x^I \hat{n}_x^O}}{I_{j n_x^I \hat{n}_x^O j}} \\ &= g_{s_i \hat{a}_i} (T_x) \frac{I_{s_i \hat{a}_i}}{I_{s_i \hat{a}_i}} \\ &= g_{s_i \hat{a}_i} (T_x) : \end{aligned} \quad (13)$$

Table 1. Analytical Results for the Probability of Joint Monophyly for Arbitrary Sample Sizes

Study	No. of populations	No. of monophyletic groups	Section
Zhu et al. (2011)	1	Arbitrary	3.7.1
Rosenberg (2003)	2	2	3.7.3
Mehta and Rosenberg (2019)	3	3	3.7.4
Mehta et al. (2016)	Arbitrary	2	3.7.5
This article	Arbitrary	Arbitrary	3.6

Equation (11) in Section 3.6 provides a general calculation from which we recover the other results listed. Other cases with small numbers of populations and monophyletic groups appear in Table 1 in Mehta and Rosenberg (2019).

Thus, the general computation in Equation (11) reduces to Equation (1), the expression describing the probability that  $s_i$  lineages coalesce to  $r_i$  lineages in time  $T_x$ .

**3.7.3. Two species in a two-species tree.** In a two-species tree, let  $s_1$  and  $s_2$  be the initial sample sizes of species 1 and 2, respectively, and let  $r_1$  and  $r_2$  be the numbers of lineages of species 1 and 2 that enter the root node. There are three species tree nodes: the root  $x$ , leaf  $x_1$  for species 1, and leaf  $x_2$  for species 2. The input and output states are  $n_{x_1}^I = (s_1 \text{ 1's})$ ,  $n_{x_2}^I = (s_2 \text{ 2's})$ ,  $n_{x_1}^O = (r_1 \text{ 1's})$ ,  $n_{x_2}^O = (r_2 \text{ 2's})$ ,  $n_x^I = (r_1 \text{ 1's}, r_2 \text{ 2's})$ , and  $n_x^O = (r_1 \text{ 1's}, r_2 \text{ 2's})$ .

For leaf  $x_1$ , label 1 survives and there are no other label types. For leaf  $x_2$ , label 2 survives and there are no other label types. For the root, both species labels are lost, and there is only one output mixed lineage  $m_1$ . Hence, there is only one partition  $P = \{1, 2\}$ .

Because  $x_1$  and  $x_2$  are leaves, from Equation (13),  $P(E_{x_1} | n_{x_1}^O) = g_{s_1 \hat{a}_1}(T_1)$  and  $P(E_{x_2} | n_{x_2}^O) = g_{s_2 \hat{a}_2}(T_2)$ . From Equation (12), for a particular  $r_1$  and  $r_2$ , we have  $P(E_x | n_x^O) = Z(r_1 \hat{a}_1, r_2 \hat{a}_2)$ . Substituting into Equation (6),

$$\begin{aligned} P(E_x | n_x^O) &= \sum_{n_{x_1}^O} \sum_{n_{x_2}^O} P(E_x | n_x^O, n_{x_1}^O, n_{x_2}^O) P(E_{x_1} | n_{x_1}^O) P(E_{x_2} | n_{x_2}^O) \\ &= \sum_{r_1=1}^{s_1} \sum_{r_2=1}^{s_2} Z(r_1 \hat{a}_1, r_2 \hat{a}_2) g_{s_1 \hat{a}_1}(T_1) g_{s_2 \hat{a}_2}(T_2) \end{aligned} \quad (14)$$

It remains to obtain  $Z(r_1, r_2)$ . First, note that there is only one possible labeled topology  $T_2$  for the two ancestral lineages of the two groups  $A_1$  and  $A_2$ , and this topology has a single internal node  $v$  of which both  $A_1$  and  $A_2$  are descendants. So, for  $k=2$ , we have by Equations (4) and (5),

$$\begin{aligned} Z(r_1 \hat{a}_1, r_2 \hat{a}_2) &= Z(r_1 \hat{a}_1, r_2 \hat{a}_2; T_2) = \frac{2r_1!r_2!}{(r_1+r_2)!} \frac{1}{r_1+r_2-1} \\ &= \frac{2}{r_1+r_2-1} \frac{r_1+r_2-1}{r_1} \hat{a}_1 \end{aligned} \quad (15)$$

which matches Lemma 4.3 in Zhu et al. (2011), Equation (6) in Brown (1994), and Equation (9) in Rosenberg (2003).

Substituting Equation (15) into Equation (14), we have

$$P(E_x | n_x^O) = \sum_{r_1=1}^{s_1} \sum_{r_2=1}^{s_2} g_{s_1 \hat{a}_1}(T_1) g_{s_2 \hat{a}_2}(T_2) \frac{2}{r_1+r_2-1} \frac{r_1+r_2-1}{r_1} \hat{a}_1 \quad (16)$$

We therefore obtain Equation (14) from Rosenberg (2003): the probability of reciprocal monophyly of two species in a two-species tree.

**3.7.4. Three species in a three-species tree.** In this section, we recapitulate the probability of JM for three species in a three-species tree, as provided in Equation (5) in Mehta and Rosenberg (2019). It suffices to describe the reduction of our Equation (11) to Equations (6) and (9) in Mehta and Rosenberg (2019), giving the conditional probability of JM within the internal node  $I$  of  $T$  given a particular input state  $n_I^I$  and output state  $n_I^O$ , and the conditional probability of JM in the species tree root  $R$  given a particular input state  $n_R^I$ .

We label the three leaves A, B, and C, and we call the single internal node  $I$  (ancestral to species A and B). The root node is  $R$ . Thus, we can specify branch input and output states:

$$\begin{aligned} n_A^I &= (p \text{ 1's}) & n_A^O &= (s \text{ 1's}) \\ n_B^I &= (q \text{ 1's}) & n_B^O &= (t \text{ 1's}) \\ n_C^I &= (r \text{ 2's}) & n_C^O &= (u \text{ 2's}) \\ n_I^I &= (s \text{ 1's}, t \text{ 1's}) & n_I^O &= (w \text{ 1's}, x \text{ 1's}) \\ n_R^I &= (w \text{ 1's}, x \text{ 1's}, y \text{ 2's}) & n_R^O &= (0 \text{ 1's}, 0 \text{ 1's}, z \text{ 2's}) \end{aligned}$$

Equations (6) and (9) in Mehta and Rosenberg (2019) are special cases of a term in Equation (3) from Mehta and Rosenberg (2019), which corresponds to our Equation (11). Comparing Equation (11) with Equation (3) from Mehta and Rosenberg (2019) indicates that to obtain Equation (6) of Mehta and Rosenberg (2019), we must show that the quantity  $K_I$  from Mehta and Rosenberg (2019) satisfies

$$K_I = \frac{C_{n_I^I \hat{\Theta}_I^0}}{I_{j n_I^I \hat{\Theta}_I^0 j}};$$

To obtain Equation (9) from Mehta and Rosenberg (2019), we must show that with  $v_p = (w, x, 0, m)$ , the quantity  $K_{\text{root}}$  from Mehta and Rosenberg (2019) satisfies

$$K_{\text{root}} = \frac{C_{n_R^I \hat{\Theta}_R^0}}{I_{j n_R^I \hat{\Theta}_R^0 j}} = Z(v_p): \quad (17)$$

First, we consider internal node I. The nontrivial cases of Equation (6) from Mehta and Rosenberg (2019) are:

$$K_I = \begin{cases} \frac{I_{s\hat{\Theta}} I_{t\hat{\Theta}} W_2 i s - w\hat{\Theta} - x}{I_{s+t\hat{\Theta}} + x} & \text{Case 1: } s\hat{\Theta} \hat{\Theta} w\hat{\Theta} \quad 1; m=0 \\ \frac{I_{s\hat{\Theta}} I_{t\hat{\Theta}} W_2 i s - 1\hat{\Theta} - \hat{\Theta}}{I_{s+t\hat{\Theta}}} & \text{Case 2: } s\hat{\Theta} \quad \hat{\Theta}; w=x=0; m=1 \end{cases} \quad (18)$$

Equation (18) concerns the internal node I of a three-species tree, a node that has input lineages from the two species it subtends. Case 1 in Equation (18) occurs when both species labels are surviving labels, as the two quantities that represent the numbers of output lineages from the two input species,  $w$  and  $x$ , are both greater than or equal to 1. In the language of our analysis, the set of surviving labels is  $U = \{1, 2\}$ . There are no output mixed lineages ( $m = 0$ ) and there is no need to consider a set of partitions  $P$  of the set of lost labels and mixed lineages.

We use Equation (10) to obtain

$$\begin{aligned} C_{n_I^I \hat{\Theta}_I^0} &= \left( \sum_{i=1}^{\mathcal{Y}} I_{s_i \hat{\Theta}_i} \right) W_2 i s - r_1 \hat{\Theta}_2 - r_2 \hat{\Theta} \quad \%_0 \quad \ddot{\imath} \quad \hat{\imath} \quad \% \\ &= I_{s\hat{\Theta}} I_{t\hat{\Theta}} W_2 i s - w\hat{\Theta} - x\hat{\Theta} \quad \%_0 \quad \ddot{\imath} \quad \hat{\imath} \quad \%_0 \end{aligned}$$

and so we have:

$$\begin{aligned} \frac{C_{n_I^I \hat{\Theta}_I^0}}{I_{j n_I^I \hat{\Theta}_I^0 j}} &= \frac{I_{s\hat{\Theta}} I_{t\hat{\Theta}} W_2 i s - w\hat{\Theta} - x\hat{\Theta}}{I_{s+t\hat{\Theta}} + x} \quad \%_0 \quad \ddot{\imath} \quad \hat{\imath} \quad \%_0 \\ &= K_I: \end{aligned} \quad (19)$$

Case 2 in Equation (18) occurs when both species labels are lost labels, as the two quantities that represent the number of output lineages from the two input species,  $w$  and  $x$ , are both 0. There is one output lineage, a mixed lineage ( $m = 1$ ). There is only one possible partition of input labels  $\{1, 2\}$  over the single mixed lineage  $m_1$ :  $P = fpg$ , with  $p = f1\hat{\Theta}g$  !  $m_1$ . We use Equations (9) and (7) to obtain:

$$\begin{aligned} C_{n_I^I \hat{\Theta}_I^0} &= J_p W_1 i s + t - \hat{\Theta} = J_p \%_0 \quad \ddot{\imath} \quad \hat{\imath} \quad \%_0 \\ &= I \sum_{j \in L_1} P_{r_j} + m_1 \hat{\Theta} Z(v_p) \\ &= I_{s+t\hat{\Theta}} Z(v_p): \end{aligned}$$

The vector  $v_p$  is  $(s, t, 0, 0)$ . Note that  $Z(s, t, 0, 0) = Z(s, t)$ , so we can use Equations (15), (2), and (3) to obtain

$$\begin{aligned} \frac{C_{n_I^I \hat{\Theta}_I^0}}{I_{s+t\hat{\Theta}}} &= \frac{2}{s+t-1} \frac{s+t-1}{s} \\ &= \frac{I_{s\hat{\Theta}} I_{t\hat{\Theta}} W_2 (s-1\hat{\Theta}-1)}{I_{s+t\hat{\Theta}}} \\ &= K_I \hat{\Theta} \end{aligned} \quad (20)$$

where the last step comes from Equation (18). Hence, we have  $K_I = C_{n_I^I \hat{\Theta}_I^0} / I_{s+t\hat{\Theta}}$ , as desired.

It remains to show that our result accords with the two nontrivial cases of Equation (9) from Mehta and Rosenberg (2019). These cases are:

$$K_{\text{root}} = \frac{f(w\hat{x}\hat{y}) + f(w\hat{y}\hat{x}) + f(x\hat{y}\hat{w})}{I_{y+1}\hat{a}} \quad \begin{array}{l} \text{Case 1 : } w\hat{x}\hat{y} = 1; m = 0\hat{a} \\ \text{Case 2 : } y = 1; w = x = 0; m = 1\hat{a} \end{array} \quad (21)$$

where

$$f(w\hat{x}\hat{y}) = \frac{\prod_{c=1}^y I_{w\hat{a}} I_{x\hat{a}} I_{y\hat{a}} W_3(w - 1\hat{x} - 1\hat{y} - c) I_{c\hat{a}}}{I_{w+x+y}\hat{a}}; \quad (22)$$

Starting from our Equation (17), we must calculate  $Z(v_p)$  for each of these two cases and show that it equals  $K_{\text{root}}$  from Equation (21). Case 1 of Equation (21) occurs when there are input lineages from three species ( $w, x, y \geq 1$ ) and no input mixed lineages ( $m = 0$ ). Thus,  $v_p = (w\hat{x}\hat{y})\hat{a}$ . We note that  $Z(w\hat{x}\hat{y})\hat{a} = Z(w\hat{x}\hat{y})$ . From an unlabeled example in Zhu et al. (2011) immediately following the proof of their Theorem 5.1, we have

$$Z(w\hat{x}\hat{y}) = \frac{4w!x!y!}{(w+x+y)!(w+x+y-1)!} \left( \frac{1}{x+y-1} + \frac{1}{w+y-1} + \frac{1}{w+x-1} \right); \quad (23)$$

Substituting Equation (2) and (3) into Equation (22) and simplifying, we have

$$f(w\hat{x}\hat{y}) = \frac{4w!x!y!}{(w+x+y)!(w+x+y-1)!} \frac{1}{w+x-1} \frac{\prod_{c=1}^{w+x-2+y-c} I_{w+x-2+y-c}\hat{a}}{I_{w+x+y-2}\hat{a}} \quad (24)$$

$$= \frac{4w!x!y!}{(w+x+y)!(w+x+y-1)!} \frac{1}{w+x-1} \hat{a} \quad (25)$$

where the step from Equations (24) to (25) uses the binomial identity [Equation (1) in Section 0.151 from Gradshteyn and Ryzhik (2014)]

$$\sum_{k=0}^{n+k} \binom{n+k}{n} = \binom{n+m+1}{n+1} \hat{a}$$

with  $y - c$  in place of  $k$ ,  $y - 1$  in place of  $m$ , and  $w + x - 2$  in place of  $n$ .

Now, from Equations (23) and (25), we have

$$Z(w\hat{x}\hat{y}) = f(w\hat{x}\hat{y}) + f(w\hat{y}\hat{x}) + f(x\hat{y}\hat{w}) = K_{\text{root}}\hat{a}$$

as required.

Case 2 of Equation (21) occurs when there are input lineages from one species ( $y \geq 1$ ,  $w = x = 0$ ) and one input mixed lineage ( $m = 1$ ),  $v_p = (0\hat{a}\hat{y})\hat{a}$ . We note that  $Z(0,0,y,1) = Z(y,1)$ , and use Equation (15) to obtain:

$$\begin{aligned} Z(y\hat{a}) &= \frac{2}{y} \frac{y+1}{y} \hat{a}^{-1} \\ &= \frac{2}{y(y+1)}; \end{aligned} \quad (26)$$

Using Equations (2) and (26), we have  $Z(y\hat{a}) = I_{y+1}\hat{a} = K_{\text{root}}$ , as required.

Having demonstrated that our JM calculation recovers the combinatorial terms  $K_I$  and  $K_{\text{root}}$ , we have therefore recovered the JM probability for three species, as obtained by Mehta and Rosenberg (2019).

**3.7.5. Two groups in a  $k$ -species tree.** Here we recapitulate the probability of JM for two groups in a  $k$ -species tree, as shown in Equation (5) from Mehta et al. (2016). It suffices to describe the reduction of our Equation (11) to Equation (4) from Mehta et al. (2016), describing the conditional probability of monophyly within a node  $x$  of  $T$  given a particular input state  $n_x^I$  and output state  $n_x^O$ . More precisely, we must equate our Equation (11) to the scenario of JM in Equation (4) of Mehta et al. (2016), obtained by substituting their Equation (5) for Case 2 in their Equation (4).

Let the input state for node  $x$  be  $n_x^I = (s_1 \hat{c}_2 \hat{m}_1)$ , and let the output state be  $n_O^I = (r_1 \hat{c}_2 \hat{m}_O)$ . We assume (as is necessary to achieve JM) that the input lineages from groups 1 and 2 include all lineages from those groups; that is, species tree node  $x$  is ancestral to all lineages that belong to groups 1 and 2. Following the labeling of cases in Mehta et al. (2016), the nontrivial cases of Equation (4) from Mehta et al. (2016) in the setting of JM are as follows:

$$K_{SC} = \frac{\sum_{s_1, s_2} \frac{I_{s_1 \hat{c}_1} I_{s_2 \hat{c}_2} W_2 (s_1 - 1 \hat{c}_2 - 1)}{I_{s_1 + s_2 \hat{c}_1}}}{\sum_{s_1, s_2} \frac{I_{s_1 \hat{c}_1} I_{s_2 \hat{c}_2} W_2 (s_1 - r_1 \hat{c}_2 - r_2)}{I_{s_1 + s_2 \hat{c}_1 + r_2}}} \quad (27)$$

Case 1e :  $s_1 \hat{c}_1$  1;  $s_2 = r_2 = m_1 = m_O = 0$   
Case 1b :  $s_2 \hat{c}_2$  1;  $s_1 = r_1 = m_1 = m_O = 0$   
Case 2 :  $s_1 \hat{c}_2$  1  $\hat{c}_1 = r_2 = m_1 = 0$   $m_O = 1$   
Case 3 :  $s_1 \hat{c}_2 \hat{c}_1 \hat{c}_2$  1;  $m_1 = m_O = 0$

where cases 1b and 1e in Equation (27) are labeled after their corresponding labels in Mehta et al. (2016).

Equation (4) in Mehta et al. (2016) is a special case of a term in Equation (3) from Mehta et al. (2016). Comparing our Equation (11) to Equation (3) from Mehta et al. (2016), we find that to obtain Equation (4) of Mehta et al. (2016) as a special case of our Equation (11), we must show that the quantity  $K_{SC}$  from Mehta et al. (2016) satisfies

$$K_{SC} = \frac{C_{n_x^I \hat{c}_x^O}}{I_{j n_x^I \hat{c}_x^O}} \quad (28)$$

Cases 1e and 1b from Equation (27) occur when there is one surviving label and no other input lineages. We have  $U = \{i\}$  for  $i = 1, 2$  and  $P$  empty. We use Equations (10), (2), and (3) to obtain:

$$\frac{C_{n_x^I \hat{c}_x^O}}{I_{j n_x^I \hat{c}_x^O}} = \frac{I_{s_i \hat{c}_i} W_1 (s_i - r_i)}{I_{s_i \hat{c}_i}} = 1$$

$$= K_{SC}$$

as required for demonstrating Equation (28).

Case 2 from Equation (27) occurs when there are two lost labels, no surviving labels, and one output mixed lineage  $m_1$ . Thus,  $U$  is empty, and there is one partition  $P = \{1, 2\} \mid m_1$ . We have already shown that Equation (11) produces Equation (20); directly applying the result from Equation (20) yields the result that Equation (28) requires.

Case 3 from Equation (27) occurs when there are two surviving labels, no lost labels, and no input or output mixed lineages. Thus,  $U = \{1, 2\}$  and  $P$  is empty. We have already shown that Equation (11) produces Equation (19); directly applying the result from Equation (19) yields the result required for Equation (28) to be satisfied.

We have therefore shown that our Equation (11) reduces to Equation (28), recapitulating the JM probability of two groups in an arbitrary species tree from Mehta et al. (2016).

### 3.8. Lower and upper bounds based on StrongJM

The probability in Equation (11) involves many steps and is potentially time-consuming to calculate. We can therefore provide a simpler lower bound by introducing the idea of StrongJM (SJM). We say that a set of lineages sampled from a species satisfies SJM if the lineages coalesce to a single lineage in the branch associated with that species. In other words, SJM is the situation in which lineage sorting is complete in the external branches of the species tree and no incomplete lineage sorting occurs in those branches. The probability of SJM can then be computed from the lengths of the external branches of the species tree.

The probability of SJM is

$$P(\text{SJM}) = \prod_{i=1}^k g_{s_i \hat{c}_i}(T_i) \quad (29)$$

where  $T_1, T_2, \dots, T_k$  are the species tree branch lengths associated with species 1, 2, ...,  $k$ .

This probability provides a lower bound on Equation (11) because it is only one of many ways that JM can be achieved;  $P(\text{JM}) \geq P(\text{SJM})$ . This lower bound avoids the pruning step and does not need to track lineage counts at species tree internal nodes, so that its calculation is faster than that of Equation (11). The lower bound is similar in spirit to an upper bound on the probability of gene-tree-species-tree concordance found by Pamilo and Nei (1988).

We can also observe that  $P(\text{SJM})$  enables an upper bound on  $P(\text{JM})$ , a bound that holds for any species tree and any distribution of gene lineages across species. This bound is:

$$P(\text{JM}) \leq \frac{1}{3} + \frac{2}{3}P(\text{SJM}) \quad (30)$$

To prove Equation (30), we observe that if each species has exactly one lineage, then Equation (30) is an equality (i.e.  $\frac{1}{3} + \frac{2}{3}(1) = 1$ ). Thus, we can suppose that at least one species has at least two lineages, so that  $P(\cdot | \text{SJM}) > 0$  and  $P(\text{JM} | \text{SJM})$  is well-defined. In this case, by the law of total probability,

$$P(\text{JM}) = P(\text{JM} | \text{SJM})P(\text{SJM}) + P(\text{JM} | \neg \text{SJM})P(\neg \text{SJM})$$

and because  $P(\text{JM} | \text{SJM}) = 1$ , we obtain:

$$P(\text{JM}) = x + P(\text{JM} | \neg \text{SJM})(1 - x) \quad (31)$$

for  $x = P(\text{SJM})$ . Next, we claim that:

$$P(\text{JM} | \neg \text{SJM}) \leq \frac{1}{3} \quad (32)$$

The justification of Equation (32) is as follows. The coalescent scenarios that comprise the event  $\neg \text{SJM}$  are precisely those for which, for some tip species  $s$ , the (two or more) lineages associated with  $s$  do not coalesce to a single lineage within the external branch incident with  $s$ . However, JM requires that the ancestral lineages of  $s$  coalesce only among themselves (and not with other lineages) until they reach a single lineage along the path in  $T$  back to its root. At some point on this path there will be just two ancestral lineages of  $s$ , along with  $r \geq 1$  other ancestral lineages from other species. The probability that in coalescing to a single lineage, the two ancestral lineages of  $s$  coalesce with each other (rather than one coalescing with one of the other  $r$  lineages present), is given by Equation (11) of Rosenberg (2003), which gives the probability that  $q_A = 2$  lineages are monophyletic when  $q_B = r$  additional lineages are present:

$$\frac{2}{r+2} \cdot \frac{r+1}{r+2} = \frac{2(r+1)}{(r+2)^2} = \frac{2}{3(r+1)} \leq \frac{1}{3} \text{ for all } r \geq 1. \text{ Thus, } P(\text{JM} | \neg \text{SJM}) \leq \frac{1}{3} \text{ as claimed.}$$

Combining Equations (31) and (32) gives Equation (30).

## 4. NUMERICAL RESULTS

### 4.1. Continuous-time Markov chain approach

Although the exact computation in Section 3.6 is instructive and presents new mathematical insight, using this result for computational purposes is inconvenient. To facilitate computation of the probability of JM, we provide a continuous-time Markov chain [CTMC; Grimmett and Stirzaker (2020)] formulation of this probability, described in Appendix A. When constructing this CTMC under the multispecies coalescent, we follow the approach of Hobolth et al. (2011). Computing the probability of JM amounts to using the same recursive decomposition as in Equation (6), but the probabilities are computed by constructing transition matrices for each branch of the tree and using matrix exponentials to obtain the output probabilities given the input probabilities. We use the CTMC formulation in providing numerical results in this section. The approach is implemented in Monophyler (Mehta et al. 2016).

### 4.2. Effects of number of species, tree height, and sample size

We use example species trees to illustrate the effects of tree height and sample size on the probability of JM. We consider a class of species trees that appears in Figure 4. The trees range in size from two to six species, and they are constructed so that the tree height is evenly divided along the branches of the longest topological path length from root to leaf.

Using each tree in Figure 4, we compute the probability of JM with Equation (11). We modulate the tree height  $h$  from 0 to 10 coalescent time units at intervals of 0.2. The number of samples in each leaf ranges from 2 to 10, incremented by 1, with each leaf having the same sample size.

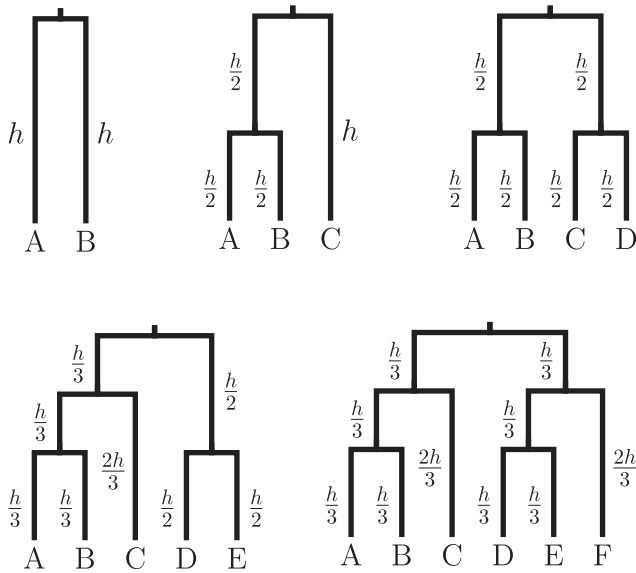


FIG. 4. Trees used to explore the effects of tree height and sample size on the probability of joint monophyly.

Figure 5 provides the effect of number of species, tree height, and sample size on the probability of JM for all trees in Figure 4. As the number of species increases from 2 to 6, the number of separate groups that must be monophyletic to produce JM increases. Hence, the JM probability decreases at fixed values for the tree height and sample size.

With increasing tree height and fixed sample size, lineages have more time during which they can coalesce within the species from which they have been sampled, and the JM probability increases with

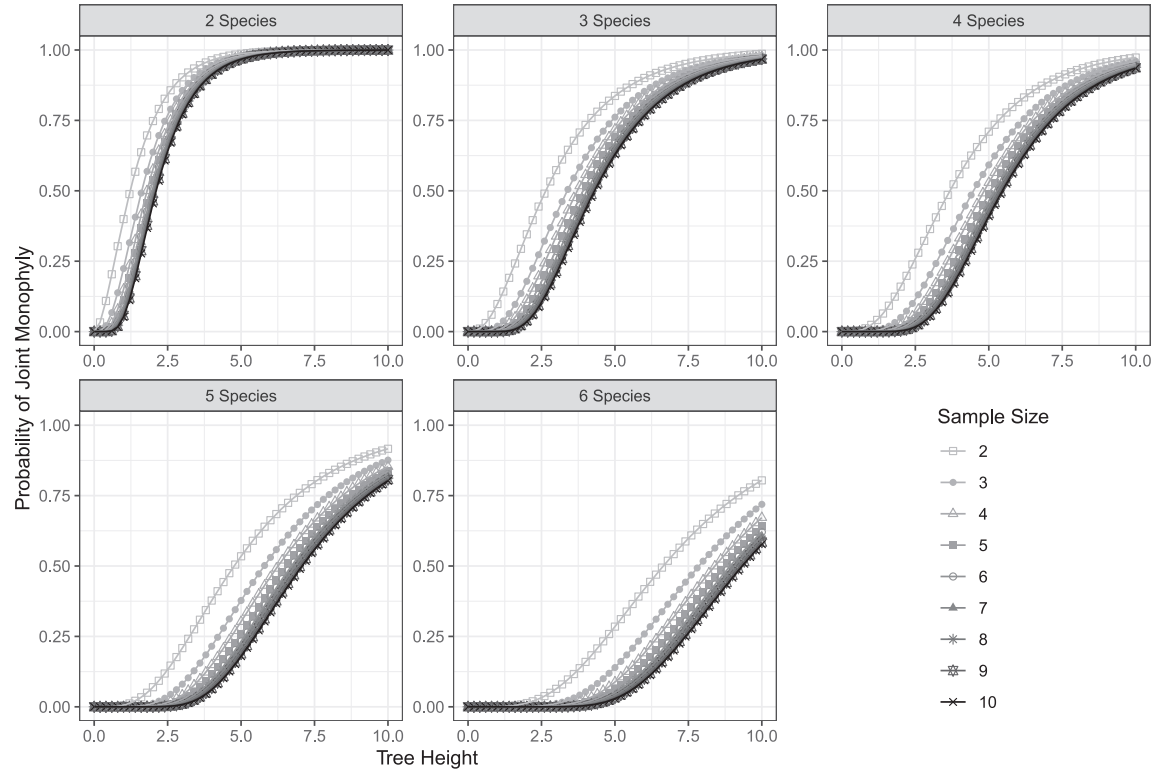


FIG. 5. Joint monophyly probabilities for various numbers of species, tree heights, and sample sizes. Probabilities are obtained using Equation (11), with the same sample size assigned to each species. Each panel is labeled by the number of species.

increasing tree height. As the sample size increases at a fixed tree height, the number of lineages that must monophyletically coalesce increases, but no additional time is available for these coalescences; hence, the JM probability decreases with increasing sample size.

An alternative perspective on the JM probabilities in Figure 5 examines, for a fixed cutoff value representing a level of statistical significance, a fixed number of species, and a fixed tree height, the minimum sample size required for achieving a JM probability that lies below the cutoff. In other words, we calculate the minimum sample size required for an observation of JM to be improbable at a specific significance level under a specific model. Such a computation can assist in understanding the extent to which an observation of monophyly can be regarded as surprising and in designing samples such that a desired level of surprise is achieved if JM is observed (Rosenberg, 2007).

Figure 6 provides these minimum sample sizes. They decrease as the cutoff value is increased. In accordance with the decrease in JM probabilities that occurs with an increasing number of species, for fixed tree height, the minimal sample size required for achieving a JM probability below a specific cutoff decreases with an increasing number of species. The minimal sample size increases with increasing tree height; as tree height grows, JM is probable even for large samples, so that very large samples might be required for a JM observation to be surprising. In most scenarios plotted, samples of six to eight per species suffice to produce probabilities below cutoff 0.001 over most of the domain for tree height.

### 4.3. Strong joint monophyly

Figure 7 provides the probability of JM against the corresponding probability of SJM from Equation (29). For each combination of a number of species, tree height, and sample size considered in Figure 5, the probability of SJM is calculated, and a point is plotted that pairs the probability of SJM with the probability of JM from Figure 5.

As SJM is a stricter condition than JM, the probability of SJM is necessarily less than or equal to the probability of JM (Section 3.8). Traversing the figure from left to right, or from bottom to top, the tree height increases. For large tree heights, JM is closely approximated by SJM, as represented by the proximity of the curves plotted to the  $y = x$  line; the event of SJM is the primary driver of JM. For smaller tree

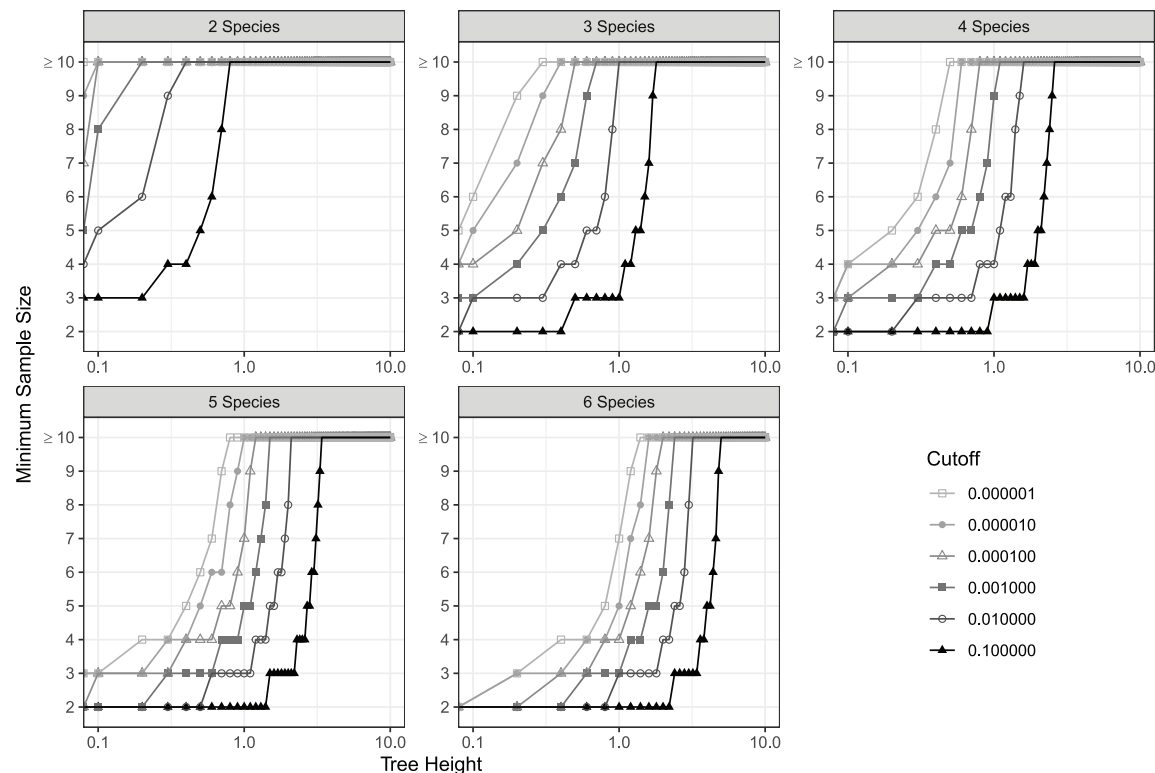


FIG. 6. Minimum sample sizes for the probability of joint monophyly to decrease below a particular cutoff probability, for varying tree height and number of species. Panel title indicates number of species.

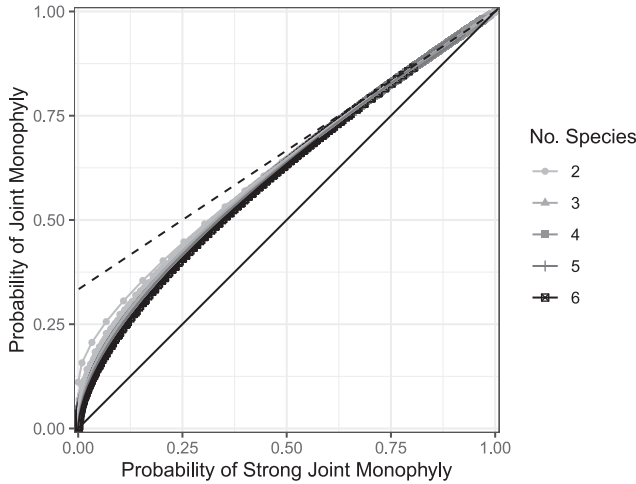


FIG. 7. The probability of joint monophyly [Eq. (11)] in relation to the probability of strong joint monophyly [Eq. (29)]. Strong joint monophyly provides a lower bound for joint monophyly. For each combination consisting of a number of species ( $2 \leq k \leq 6$ ) and a sample size ( $2 \leq n \leq 6$ ), a curve links points with increasing tree height (0 to 10 at intervals of 0.2). Parameter sets (number of species, tree height, sample size) follow Figure 4. The solid line indicates equality of the probabilities of joint monophyly and strong joint monophyly, and the dashed line indicates the upper bound on the probability of joint monophyly provided by Equation (30).

heights, the probability of SJM is substantially lower than the probability of JM, as configurations in which JM is achieved by coalescences that occur deeper in the species tree than the external branches are not improbable.

The plots show relatively little effect of the number of species on the relationship between JM and SJM, or of the sample size. Thus, by curve fitting, it would be possible to empirically transform the easily computable SJM probability to approximate the JM probability.

For the case of two species and two lineages per species, using the two-species tree in Figure 4, the probability of JM from Equation (16) is

$$P(\text{JM}) = \sum_{r_1=1}^{\infty} \sum_{r_2=1}^{\infty} g_{2\hat{w}_1}(h) g_{2\hat{w}_2}(h) \frac{2}{r_1 + r_2 - 1} \frac{r_1 + r_2}{r_1} - 1$$

$$= 1 - \frac{4}{3} e^{-h} + \frac{4}{9} e^{-2h}. \quad (33)$$

The probability of SJM from Equation (29) is  $P(\text{SJM}) = g_{2\hat{w}}(h)^2 = (1 - e^{-h})^2$ . Solving this equation for  $e^{-h}$  and inserting the solution into Equation (33), the probability of JM in terms of the probability of SJM is

$$P(\text{JM}) = \frac{1}{9} [2P(\text{SJM}) + 1]^2. \quad (34)$$

Equation (34) appears in Figure 7 as the curve corresponding to two species and sample size two, visible as the curve with the highest values of the JM probability for low values of the SJM probability.

## 5. DISCUSSION

We have derived the general probability of JM in an arbitrary species tree. The probability that for each species in a  $k$ -species tree, the lineages of that species are monophyletic under the multispecies coalescent. Using this result [Eq. (11)], we have obtained as special cases several previous results for the probability of JM: the cases of arbitrarily many groups of lineages in one species (Section 3.7.1), two lineage groups in two species (Section 3.7.3), three lineage groups in three species (Section 3.7.4), and two lineage groups in arbitrarily many species (Section 3.7.5). Previous results on the probability of JM were restricted to small numbers of groups (four or fewer), small trees (four species or fewer), or both. We were able to fully generalize these results by combining the recursive approach of Mehta et al. (2016) for general species trees and the combinatorial calculations of Zhu et al. (2011) for arbitrary numbers of groups.

Our calculation relies on a pruning algorithm, in which computations are performed recursively at each internal node of a species tree. Pruning algorithms have a long history in phylogenetics, tracing to early efforts to evaluate gene tree probabilities from molecular sequence data in maximum-likelihood phylogenetics (Felsenstein, 1981). Recent algorithms have generalized the pruning approach to gene tree

computations conditional on species trees (Efromovich and Kubatko, 2008; RoyChoudhury et al., 2008; Bryant et al., 2012; RoyChoudhury and Thompson, 2012; Stadler and Degnan, 2012; Wu, 2012; Mehta et al., 2016). The pruning algorithm we have provided accounts for the intricate merging pattern of gene lineages that occurs when two species merge backward in time to their ancestral species.

Although pruning algorithms do lead to exact computations for various quantities of interest, they can suffer from the computational burden of tree traversal as the size of the species tree increases. In addition, although the pruning algorithm renders the tree traversal polynomial-time in the number of species, the computation time is not polynomial-time in the number of species or sample size, owing to the effect on the most computationally complex part of the calculation: enumerating partitions and performing a calculation for each partition (Section 3.7.4). Our analysis includes the instructive formal computations that appear in Section 3 as well as a CTMC approach that is convenient for computation (Appendix A). Using the CTMC approach, we have seen that the JM calculation reproduces sensible patterns in the effects of model parameters on monophyly probabilities.

Increasingly many studies are now considering genealogical discordance, phylogeography, and species delimitation using samples with many individuals per species and many loci. Our computations are well-suited to such scenarios, as we evaluate monophyly probabilities based on multiple individuals within species, and multilocus studies enable comparisons of model-based monophyly probabilities to empirical estimates from loci across the genome (Mehta et al., 2016). The new algorithmic approach will be useful particularly where JM of multiple groups is of interest such as in problems that have been examined in taxon groups including rotifers (Birky et al., 2005), birds (Cloutier et al., 2019), and snakes (Kubatko et al., 2011), among others. We have implemented the new algorithms in the software Monophyl er (Mehta et al., 2016).

## ACKNOWLEDGMENT

We are pleased to contribute to the Mike Waterman special issue this application of a recursive algorithmic approach to a problem in coalescent theory and phylogenetics. The Monophyl er software is available at <http://rosenberglab.stanford.edu>.

## AUTHOR DISCLOSURE STATEMENT

The authors declare they have no competing financial interests.

## FUNDING INFORMATION

The authors acknowledge support from National Institutes of Health Grant No. R01 GM131404 and National Science Foundation Grant No. BCS-2116322.

## REFERENCES

- Arbogast, B.S., Edwards, S.V., Wakeley, J., et al. 2002. Estimating divergence times from molecular data on phylogenetic and population genetic timescales. *Ann. Rev. Ecol. Syst.* 33, 707–740.
- Baker, A.J., Tavares, E.S., and Elbourne, R.F. 2009. Countering criticisms of single mitochondrial DNA gene barcoding in birds. *Mol. Ecol. Resour.* 9, 257–268.
- Bergsten, J., Bilton, D.T., Fujisawa, T., et al. 2012. The effect of geographical scale of sampling on DNA barcoding. *Syst. Biol.* 61, 851–869.
- Birky, C.W., Wolf, C., Maughan, H., et al. 2005. Speciation and selection without sex. *Hydrobiologia.* 546, 29–65.
- Brown, J.K. 1994. Probabilities of evolutionary trees. *Syst. Biol.* 43, 78–91.
- Bryant, D., Bouckaert, R., Felsenstein, J., et al. 2012. Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.* 29, 1917–1932.
- Carstens, B.C., and Knowles, L.L. 2007. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: An example from *Melanoplus* grasshoppers. *Syst. Biol.* 56, 400–411.
- Carstens, B.C., and Richards, C.L. 2007. Integrating coalescent and ecological niche modeling in comparative phylogeography. *Evolution* 61, 1439–1454.

- Cloutier, A., Sackton, T.B., Grayson, P., et al. 2019. Whole-genome analyses resolve the phylogeny of toothless birds (Palaeognathae) in the presence of an empirical anomaly zone. *Syst. Biol.* 68, 937–955.
- De Queiroz, K. 2007. Species concepts and species delimitation. *Syst. Biol.* 56, 879–886.
- Efromovich, S., and Kubatko, L.S. 2008. Coalescent time distributions in trees of arbitrary size. *Stat. Appl. Genet. Mol. Biol.* 7, 2.
- Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* 17, 368–376.
- Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.
- Gradshteyn, I.S., and Ryzhik, I.M. 2014. *Table of Integrals, Series, and Products*. Academic Press, Cambridge, MA.
- Grimmett, G.R., and Stirzaker, D.S. 2020. *Probability and Random Processes*, 4th ed. Oxford University Press, Oxford, UK.
- Hobolth, A., Andersen, L.N., and Mailund, T. 2011. On computing the coalescence time density in an isolation-with-migration model with few samples. *Genetics* 187, 1241–1243.
- Hudson, R.R., and Coyne, J.A. 2002. Mathematical consequences of the genealogical species concept. *Evolution* 56, 1557–1565.
- Jansen, G., Savolainen, R., and Vepsäläinen, K. 2010. Phylogeny, divergence-time estimation, biogeography and social parasite-host relationships of the Holarctic ant genus *Myrmica* (Hymenoptera: Formicidae). *Mol. Phylogenet. Evol.* 56, 294–304.
- Knuth, D.E. 2011. *The Art of Computer Programming, Volume 4A: Combinatorial Algorithms, Part 1*. Addison-Wesley Professional, Boston, MA.
- Kubatko, L.S., Gibbs, H.L., and Bloomquist, E.W. 2011. Inferring species-level phylogenies and taxonomic distinctiveness using multilocus data in *Sistrurus* rattlesnakes. *Syst. Biol.* 60, 393–409.
- Loehr, N.A. 2017. *Combinatorics*. Chapman and Hall/CRC, London, UK.
- Mehta, R.S., Bryant, D., and Rosenberg, N.A. 2016. The probability of monophyly of a sample of gene lineages on a species tree. *Proc. Natl. Acad. Sci. USA*. 113, 8002–8009.
- Mehta, R.S., and Rosenberg, N.A. 2019. The probability of reciprocal monophyly of gene lineages in three and four species. *Theor. Popul. Biol.* 129, 133–147.
- Moritz, C. 1994. Defining evolutionarily significant units for conservation. *Trends Ecol. Evol.* 9, 373–375.
- Neilon, M.E., and Stepien, C.A. 2009. Evolution and phylogeography of the tubenose goby genus *Proterorhinus* (Gobiidae: Teleostei): Evidence for new cryptic species. *Biol. J. Linn. Soc.* 96, 664–684.
- Pamilo, P., and Nei, M. 1988. Relationships between gene trees and species trees. *Mol. Biol. Evol.* 5, 568–583.
- Rabeling, C., Schultz, T.R., Pierce, N.E., et al. 2014. A social parasite evolved reproductive isolation from its fungus-growing ant host in sympatry. *Curr. Biol.* 24, 2047–2052.
- Rosenberg, N.A. 2003. The shapes of neutral gene genealogies in two species: Probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. *Evolution* 57, 1465–1477.
- Rosenberg, N.A. 2007. Statistical tests for taxonomic distinctiveness from observations of monophyly. *Evolution* 61, 317–323.
- RoyChoudhury, A., Felsenstein, J., and Thompson, E.A. 2008. A two-stage pruning algorithm for likelihood computation for a population tree. *Genetics* 180, 1095–1105.
- RoyChoudhury, A., and Thompson, E.A. 2012. Ascertainment correction for a population tree via a pruning algorithm for likelihood computation. *Theor. Popul. Biol.* 82, 59–65.
- Stadler, T., and Degnan, J.H. 2012. A polynomial time algorithm for calculating the probability of a ranked gene tree given a species tree. *Algorithms Mol. Biol.* 7, 7.
- Syring, J., Farrell, K., Businsky, R., et al. 2007. Widespread genealogical nonmonophyly in species of *Pinus* subgenus *Strobus*. *Syst. Biol.* 56, 163–181.
- Tavaré, S. 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol.* 26, 119–164.
- Wu, Y. 2012. Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution* 66, 763–775.
- Zhu, S., Degnan, J.H., and Steel, M. 2011. Clades, clans, and reciprocal monophyly under neutral evolutionary models. *Theor. Popul. Biol.* 79, 220–227.

Address correspondence to:  
 Dr. Rohan S. Mehta  
 Department of Physics  
 Emory University  
 Atlanta, GA 30322  
 USA

E-mail: rsmehrt4@emory.edu

## Appendix

### A. CALCULATING PROBABILITIES WITH A CTMC APPROACH

#### A.1. Mathematical approach

We now produce an alternative approach to calculating the probability of JM: a CTMC (Grimmett and Stirzaker, 2020). We define a transition rate matrix for each species tree branch and traverse the species tree from the leaves to the root. For each branch of the traversal, we use the probability of the input states and the transition rate matrix to obtain the probability of the output states. The output states of two daughter branches combine to form input states of the parent branch.

Each branch of the species tree has its own Markov chain. For a particular branch  $x$ , we must define a state space. Let  $T_x$  be the subtree below and including branch  $x$ . For this appendix, we track lineage labels differently from Section 3. We no longer keep track of lost, surviving, or fixed labels. Instead, we classify the species labels  $\{1, 2, \dots, k\}$  by their numbers of extant lineages. A label  $i$  for one of the  $k$  species starts at a leaf with  $s_i$  lineages, the sample size of the species. If JM is preserved, then the  $s_i$  lineages eventually decrease to a single ancestral lineage. Once the single lineage is reached, the label and its single associated extant lineage become free, in that any coalescence involving this label no longer affects its contribution to JM. Coalescences of free lineages with other free lineages preserve JM, reducing the number of free lineages. In this formulation, fixed lineages are free.

The state space for a branch  $x$  therefore consists of a failure state  $F$ , which represents the situation where JM has been violated, and a set of vectors  $v_x$  that keep track of the list of lineage counts for the  $k$  labels. The  $i$ th element of  $v_x$ ,  $v_{x,i}$ , is the number of labels with  $i$  extant lineages, with  $v_{x,0}$  counting the number of free lineages. For a branch  $x$ , the maximum number of lineages a label can have is the largest sample size of any species in  $T_x$ , as no label can gain lineages through coalescence. If  $S_x$  is the set of species in  $T_x$ , then the vectors in the state space for the chain for branch  $x$  have length  $s_{x,m} = \max_{i \in S_x} s_i$ .

State transitions in this process occur owing to coalescence. Let us define

$$V_x = \sum_{i=1}^{s_{x,m}} i v_{x,i} \quad (35)$$

as the total number of lineages for state  $v_x$ . For state  $v_x$ , we have three possible transitions, corresponding to intralabel coalescences, interlabel coalescences that preserve JM, and interlabel coalescences that do not preserve JM.

1. An intralabel coalescence within a label of size  $i > 1$  reduces the number of lineages of that label by 1.

$v_{x,i} \rightarrow v_{x,i} - 1$ , and  $v_{x,i-1} \rightarrow v_{x,i-1} + 1$ . There are  $\frac{V_x}{2}$  possible coalescences, and among those,

$v_{x,i} \frac{i}{2}$  lead to this state transition. The conditional probability that a coalescence has this transition

given that a coalescence occurs is  $v_{x,i} \frac{i}{2} \frac{V_x}{2}$ .

2. An interlabel coalescence that preserves JM can only occur between free lineages. Thus, it reduces

$v_{x,0} \rightarrow v_{x,0} - 1$ . The conditional probability that a coalescence has this transition is  $\frac{v_{x,0} - 1}{2} \frac{V_x}{2}$ .

3. Finally, any other coalescence is an interlabel coalescence that violates JM. Hence,  $v_x \rightarrow F$ . This transition has conditional probability  $1 - \frac{v_{x,0} - 1}{2} \frac{V_x}{2} - \sum_{i=2}^{s_{x,m}} \frac{v_{x,i} i}{2} \frac{V_x}{2}$ .

These probabilities yield a transition matrix for transitions conditional on occurrence of a coalescence.

(Appendix continues / )

## A.2. Example transitions and transition rate matrices

Consider a branch with input lineages from four species: two with one lineage, one with two lineages, and one with four lineages as well as a single input mixed lineage. There are three free lineages: those of species 1 and 2 and the mixed lineage. The input state is  $v_x = (3,1,0,1)$ . The total number of lineages present is  $V_x = 9$  [Eq (35)], so there are  $\frac{V_x}{2} = \frac{9}{2} = 4.5$  possible coalescences. The maximal sample size is  $s_{x\hat{m}} = 4$ .

Four types of coalescences are possible. An intralabel coalescence can occur in the species with two lineages,  $i = 2$ . This coalescence has probability

$$\frac{v_{x\hat{2}}}{36} = \frac{\frac{2}{2}}{36} = \frac{1}{36} = \frac{1}{36}.$$

This transition converts a species with two lineages to one with one lineage, or  $(3,1,0,1) \rightarrow (4,0,0,1)$ .

An intralabel coalescence can also occur in the species with four lineages. In this case,  $i = 4$ , so that the transition probability is

$$\frac{v_{x\hat{4}}}{36} = \frac{\frac{4}{2}}{36} = \frac{2}{36} = \frac{1}{18}.$$

The species with four lineages transitions to one with three lineages. The state transition is  $(3,1,0,1) \rightarrow (3,1,1,0)$ .

Interlabel coalescences can occur between free lineages ( $i = 1$ ). This transition occurs with probability

$$\frac{v_{x\hat{1}}}{36} = \frac{\frac{3}{2}}{36} = \frac{3}{72} = \frac{1}{24}.$$

It reduces three free lineages to two free lineages, and the state transition is  $(3,1,0,1) \rightarrow (2,1,0,1)$ .

Finally, any other coalescence leads to the failure state. Hence,  $(3,1,0,1) \rightarrow F$  with probability

$$\begin{aligned} P((3,1,0,1) \rightarrow F) &= 1 - \frac{v_{x\hat{2}}}{36} - \sum_{i=2}^4 \frac{v_{x\hat{i}}}{36} \\ &= 1 - \frac{1}{36} - \frac{2}{36} - \frac{3}{36} - \frac{4}{36} \\ &= \frac{13}{36}. \end{aligned}$$

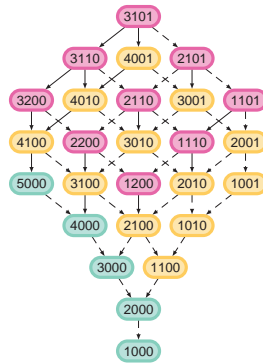
The full transition matrix for this species tree branch includes every state attainable by any number of coalescences beginning with state  $(3,1,0,1)$ . Appendix Figure A1 provides the state space for the branch, along with all possible transitions. The complete transition matrix can be obtained by using similar reasoning for all possible states and appears in Appendix Table A1.

To get the CTMC, the transition matrix for a branch must be converted into a transition rate matrix, or a  $Q$  matrix. We scale the transition rates by noting that for a state  $v_x$ , coalescences occur at rate  $V_x$ . Next, we subtract each row sum from the associated diagonal entry of the matrix. Thus, to obtain our transition rate matrix, we must multiply each row by the total number of possible coalescences for the state corresponding to that row, and then subtract the row sum from the diagonal entry. Therefore, the  $Q$  matrix for the branch follows the matrix in Appendix Table A2.

Given a vector of probabilities  $p_x$  that represents the input probability distribution over all possible states in a species branch  $x$  with length  $T_x$ , the distribution of output states follows (Grimmett and Stirzaker, 2020):

$$p_x \exp(Q_x T_x): \quad (36)$$

(Appendix continues / )



APPENDIX FIG. A1. State space for the continuous-time Markov chain for the example branch in Section A.2. States are colored by the number of species for which JM is not yet determined (pink, two; yellow, one; green, none). Intraspecies transitions use a solid line; interspecies transitions use a dashed line. The Failure state is excluded; all states except those colored green can transition to the failure state. JM, joint monophyly.

### A.3. Algorithm

The CTMC algorithm consists of two components. First, the species tree structure is created. Second, a recursive function is applied to the root node of the tree, and this function returns the probability result. The following pseudocode describes the creation of the tree structure:

```

read tree from Newick string;
read sample size information;
assign sample sizes to leaves of tree;
do recursive function get nodeout put on root node of tree.

```

The tree structure is created from a string in Newick format by using the function `Tree` in the package `ete3` in Python. The user must specify both the Newick tree and the sample size information, which consists of two lists: one specifying the leaf names (the same names as in the Newick tree), and the other specifying the sample sizes of those leaves in the same order.

The recursive function `get nodeout put` is described in the following pseudocode:

```

if node x is a leaf then
    set input state to be the vector  $v$  such that  $v_{s_x} = 1$  for  $s_x$  the sample size of the node  $x$ , and
     $v_i = 0$  for all  $i \notin s_x$ ;
    set input state probability to 1;
    set current failure probability to 0;
    extract branch length of node  $x$  from input tree;
    (**) compute vector of output probabilities of node  $x$  given input state probabilities and branch length
    according to Equation (36);
    return vector of output probabilities;
else
    apply get nodeout put to left daughter node of node  $x$ ;
    apply get nodeout put to right daughter node of node  $x$ ;
    (*) combine the output states and sum the output probabilities of the left and right daughter nodes to
    get input states and input probabilities for node  $x$ ;
    extract branch length of node  $x$  from input tree;
    (**) compute vector of output probabilities of node  $x$  given input state probabilities and branch length
    according to Equation (36);
    return vector of output probabilities;
end

```

Step (\*): combining inputs. Step (\*), combining the output states and summing the output probabilities of the left and right daughter nodes  $L$  and  $R$ , respectively, of a node  $x$ , proceeds as follows.

(Appendix continues / )

Appendix Table A1. Transition Matrix for the Example Continuous-Time Markov Chain in Section A.2

Current state	Next state																			
	3101	3110	4001	2101	3200	4010	2110	3001	1101	4100	2200	3010	1110	2001	5000	3100	1200	2010	1001	4000
(3,1,0,1)	0	$\frac{1}{6}$	$\frac{1}{36}$	$\frac{1}{12}$	0	$\frac{1}{28}$	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(3,1,1,0)	0	0	0	0	$\frac{3}{28}$	$\frac{1}{28}$	$\frac{3}{28}$	0	0	0	0	0	0	0	0	0	0	0	0	0
(4,0,0,1)	0	0	0	0	0	$\frac{3}{14}$	0	$\frac{3}{14}$	0	0	0	0	0	0	0	0	0	0	0	0
(2,1,0,1)	0	0	0	0	0	0	$\frac{3}{14}$	$\frac{1}{28}$	$\frac{1}{28}$	0	$\frac{1}{7}$	0	0	0	0	0	0	0	0	0
(3,2,0,0)	0	0	0	0	0	$\frac{2}{21}$	0	0	0	$\frac{1}{7}$	0	0	$\frac{2}{7}$	0	0	0	0	0	0	0
(4,0,1,0)	0	0	0	0	0	$\frac{1}{7}$	0	0	0	0	0	$\frac{1}{21}$	0	0	0	0	0	0	0	0
(2,1,1,0)	0	0	0	0	0	0	0	0	0	0	0	$\frac{1}{7}$	$\frac{1}{21}$	0	0	0	0	0	0	0
(3,0,0,1)	0	0	0	0	0	0	0	0	0	0	0	$\frac{2}{7}$	0	$\frac{1}{21}$	0	0	0	0	0	0
(1,1,0,1)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(4,1,0,0)	0	0	0	0	0	0	0	0	0	0	0	0	$\frac{2}{7}$	0	$\frac{1}{15}$	$\frac{2}{5}$	0	0	0	0
(2,2,0,0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	$\frac{1}{15}$	$\frac{1}{15}$	0	0	0
(3,0,1,0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	$\frac{1}{5}$	0	0
(1,1,1,0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	$\frac{1}{5}$	$\frac{1}{15}$	0	0
(2,0,0,1)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	$\frac{2}{5}$	0	0
(5,0,0,0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	$\frac{1}{15}$	0	0
(3,1,0,0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(1,2,0,0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(2,0,1,0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(1,0,0,1)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(4,0,0,0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(2,1,0,0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(1,0,1,0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(3,0,0,0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(1,1,0,0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(2,0,0,0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
(1,0,0,0)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

(Appendix continues / )

Appendix Table A2. Transition Rate Matrix ( $Q_x$ ) for the Example Continuous-Time Markov Chain in Section A.2

		Next state																										
		3101	3110	4001	2101	3200	4010	2110	3001	1101	4100	2200	3010	1110	2001	5000	3100	1200	2010	1001	4000	2100	1010	3000	1100	2000	1000	F
Current state	(3,1,0,1)	-36	6	1	3	0																					0	26
	(3,1,1,0)	0	-28	0	0	3	1	3	0																		0	21
	(4,0,0,1)	0	0	-28	0	0	6	0	6	0																0	16	
	(2,1,0,1)	0	0	0	-28	0	0	6	1	1	0															0	20	
	(3,2,0,0)	0	0	0	0	-21	0	0	0	0	2	3	0	0												0	16	
	(4,0,1,0)	0				0	-21	0	0	0	3	0	6	0												0	12	
	(2,1,1,0)	0					0	-21	0	0	0	3	1	1	0											0	16	
	(3,0,0,1)	0					0	0	-21	0	0	0	6	0	3	0										0	12	
	(1,1,0,1)	0						0	0	-21	0	0	0	6	1	0										0	14	
	(4,1,0,0)	0							0	0	-15	0	0	0	0	1	6	0								0	8	
	(2,2,0,0)	0							0	0	0	0	-15	0	0	0	2	1	0							0	12	
	(3,0,1,0)	0							0	0	0	0	0	0	0	3	0	3	0	3	0					0	9	
	(1,1,1,0)	0							0	0	0	0	0	-15	0	0	0	0	3	1	0					0	11	
	(2,0,0,1)	0							0	0	0	0	0	0	-15	0	0	0	6	1	0					0	8	
	(5,0,0,0)	0							0	0	0	0	-10	0	0	-10	0	0	0	0	10	0	0			0	0	0
	(3,1,0,0)	0							0	0	0	0	0	0	0	0	0	-10	0	0	1	3	0	0		0	6	
	(1,2,0,0)	0							0	0	0	0	0	0	0	0	0	0	-10	0	0	2	0	0		0	8	
	(2,0,1,0)	0							0	0	0	0	0	0	0	0	0	0	0	0	0	3	1	0	0	0	6	
	(1,0,0,1)	0							0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	4	
	(4,0,0,0)	0							0	0	0	0	0	0	0	0	0	0	0	0	-6	0	0	6	0	0	0	0
	(2,1,0,0)	0							0	0	0	0	0	0	0	0	0	0	-6	0	0	-6	0	1	1	0	0	4
	(1,0,1,0)	0							0	0	0	0	0	0	0	0	0	0	0	0	0	0	-6	0	3	0	0	3
	(3,0,0,0)	0							0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-3	0	3	0	0
	(1,1,0,0)	0							0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-3	1	0	2
	(2,0,0,0)	0							0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	1	0	1
	(1,0,0,0)	0							0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	F	0																									0	0

(Appendix continues / )

We note that there are no shared species labels between daughter nodes. Hence, all species labels with  $i$  extant lineages in the output of node L or node R also have  $i$  extant lineages in the input of node x. The number of species labels with  $i$  extant lineages as inputs of node x,  $v_{x\hat{a}}$ , is  $v_{L\hat{a}} + v_{R\hat{a}}$  for each  $i > 1$ . Similarly, free lineages in the output states of nodes L and R remain free as input lineages to x, so  $v_{x\hat{a}} = v_{L\hat{a}} + v_{R\hat{a}}$ . Thus, summing vectors, an input state  $v_x$  obtained from a pair of output states  $v_L$  and  $v_R$  is  $v_x = v_L + v_R$ . The probability of the input state  $v_x$  obtained by summing output states  $v_L$  and  $v_R$  is the product of the probabilities of the output states  $v_L$  from node L and  $v_R$  from node R.

The set of possible input states to node x is obtained by considering all possible sums of an output state for daughter node L and daughter node R, using vector summation. The probability of a possible input state to x is the sum over all pairs of output states that result in that input state, where for each pair, the product of the probabilities of the two output states in that pair is summed.

This vector addition procedure omits the failure state F, which occurs as an input state of node x when it is an output state of L, R, or both L and R. If  $P(F)_L$  and  $P(F)_R$  are the output failure probabilities for L and R, respectively, then the input probability of failure is  $P(F)_L [1 - P(F)_R] + [1 - P(F)_L] P(F)_R + P(F)_L P(F)_R$ :

The result of Step (\*) is a vector of input states  $I_x$  and a vector of their probabilities  $p_{I_x}$ .

Step (\*\*): computing outputs. Step (\*\*), the computation of the output states and probabilities given the input states and probabilities, is described by the following pseudocode:

- (I) generate possible output states from input states;
- (II) generate  $Q_x$ , the Q matrix for node x, considering all possible input states and output states;
- (III) rearrange the order of input states to match the order of output states and construct a rearranged probability vector  $p_x$  from  $p_{I_x}$ ;
- (IV) compute output state probabilities using  $p_x$ ,  $Q_x$ , and Equation (36).

To use Equation (36) to obtain output probabilities, the state space of  $p_x$  must include all possible output states. Thus, it is necessary to include all possible output states  $O_x$  for a set of input states  $I_x$ .

(Step I) Possible output states consist of all states that are accessible from any number of transitions starting from the set of input states, and they include the input states themselves. The set of possible output states  $O_x$  is computed using a recursive algorithm that includes all states that are accessible through a one-step transition from the current set of states, and runs until all such transitions are already included in the set.

(Step II) Once the state space  $O_x$  is enumerated, the Q matrix can be constructed using the procedure described in Section A.1.

(Step III) As a minor technical point, to apply matrix operations, the input state vector  $I_x$  and the corresponding probabilities  $p_{I_x}$  must be rearranged to match the order of states enumerated in Step I, and an input probability of 0 must be assigned to the states in  $O_x$  that are not part of  $I_x$ . The rearranged input probability vector is  $p_x$ .

(Step IV) Once  $p_x$  is obtained, Equation (36) is used to compute the output state probabilities. The matrix exponential in our algorithm is computed by the function `linalg.expm` in the package `scipy` in Python.

## APPENDIX REFERENCE

Grimmett, G.R., and Stirzaker, D.S. 2020. Probability and Random Processes, 4th ed. Oxford University Press, Oxford, UK.