bioRxiv preprint doi: https://doi.org/10.1101/2021.12.16.473006; this version posted December 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The probability of joint monophyly of all species in an arbitrary species tree

Rohan S. Mehta[1], Mike Steel[2], and Noah A. Rosenberg[3]

[1]Department of Physics, Emory University, Atlanta, GA, USA
[2]Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand
[3]Department of Biology, Stanford University, Stanford, CA, USA

December 16, 2021

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, has been used to evaluate phylogenetic and phylogeographic hypotheses, as well as to delimit species. These applications often make use of a probability of joint monophyly under models of gene lineage evolution. Studies in coalescent theory have computed this joint monophyly probability for small numbers of separate groups in arbitrary species trees, and for arbitrary numbers of separate groups in trivial species trees. Here, generalizing existing results on monophyly probabilities under the multispecies coalescent, we derive the probability of joint monophyly 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 joint monophyly. The result also enables computation of relatively simple lower and upper bounds on the joint monophyly probability. Our results expand the scope of joint monophyly calculations beyond small numbers of species, subsuming past formulas that have been used in simpler cases.

1 Introduction

Evaluations of the prevalence of reciprocal, or joint, monophyly 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 [Kubatko et al., 2011], and providing context for estimation of species divergence times [Arbogast et al., 2002]. Joint monophyly is fundamental to genealogical perspectives on species delimitation [Hudson and Coyne, 2002, De Queiroz, 2007].

Central to the application of joint monophyly is a theoretical prediction of the probability that genealogies show joint monophyly as a function of evolutionary parameters. Many studies have used monophyly probability computations in studies 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 primarily made use of theoretical results of Rosenberg [2003, 2007], which consider the probability that gene lineages in two populations of species 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: first, 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 (*S. catenatus catenatus*), as well as varying levels of distinctiveness among the remaining subspecies.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.16.473006; this version posted December 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

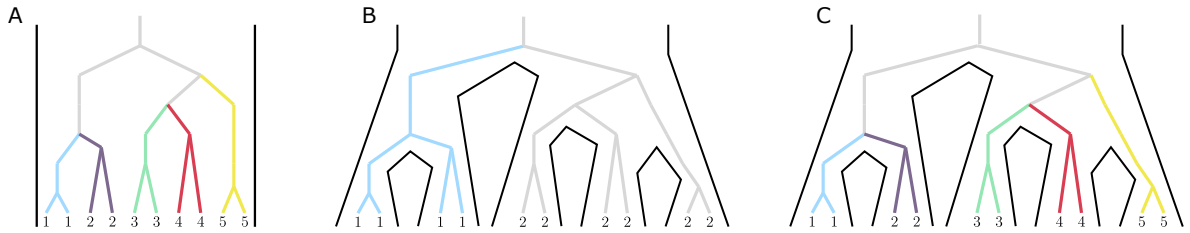

Figure 1: Schematic of the general joint monophyly calculation. (A) Zhu et al. [2011] computed the probability of joint monophyly of arbitrarily many groups in a single population. (B) Mehta et al. [2016] computed the probability of joint monophyly of two groups in an arbitrary species tree. (C) Here, 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.

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. Joint monophyly computations were likewise restricted to these two hierarchical pairs of subspecies. Although this 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]. Furthermore, simultaneously studying the relationship between the *S. catenatus* subspecies in relation to the other *Sistrurus* species would have required results that were able to 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, Kubatko et al., 2011, Bergsten et al., 2012].

Three theoretical developments now place the possibility of a joint monophyly probability computation within reach for taxa related according to an arbitrary species tree. First, Zhu et al. [2011] computed the probability of joint monophyly 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 joint monophyly in a species tree of arbitrary size, considering two classes of lineages. Finally, Mehta and Rosenberg [2019] found the full probability of joint monophyly for the lineages of species evolving on species trees with three or four species. Thus, the first 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 joint monophyly 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. [2016] and how they relate to our 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 joint monophyly, 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 $\mathscr{T}$ on the species label set S, consisting of a topology and a set of branch lengths. For each leaf S_i of $\mathscr{T}$, 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 "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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.16.473006; this version posted December 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

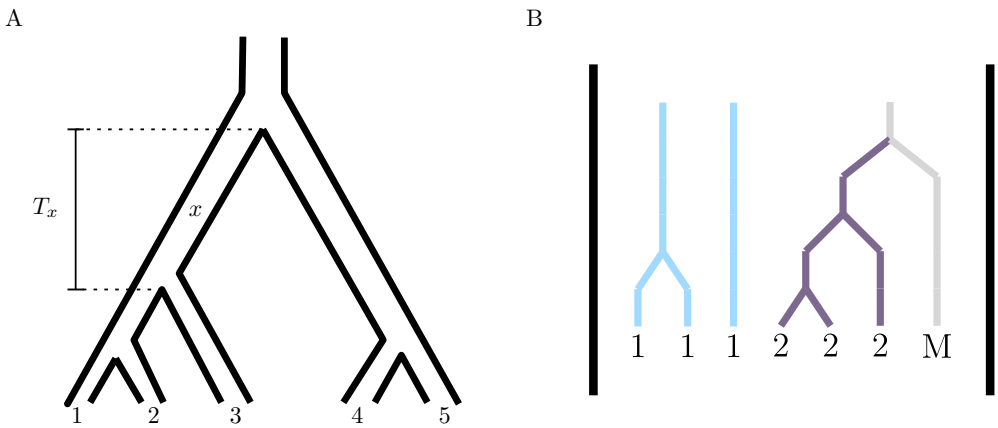

Figure 2: Notation for input and output lineages. (A) An example of a species tree $\mathscr{T}$, with five species and species label set $S = \{1, 2, 3, 4, 5\}$. An example branch x is highlighted with its branch length T_x. (B) Coalescences happening within a single branch (branch x in (A)) of a species 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 five-species tree, the input state is $\mathbf{n_x^I} = (3, 3, 0, 0, 0, 1)$, and the output state is $\mathbf{n_x^O} = (2, 0, 0, 0, 0, 1)$. The label 1 is a surviving label, and the label 2 is a lost label.

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 $1, 2, \ldots, k$, 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 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 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 $\mathscr{T}$ 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.

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 ith element is the number of output lineages that possess label i. A particular instance of this random variable is denoted $\mathbf{n}_x^O$.

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 a 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 $\mathbf{n}_x^I = \mathbf{n}_{x_L}^O + \mathbf{n}_{x_R}^O$. Figure 2B displays an example species tree node with its inputs and outputs.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.16.473006; this version posted December 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

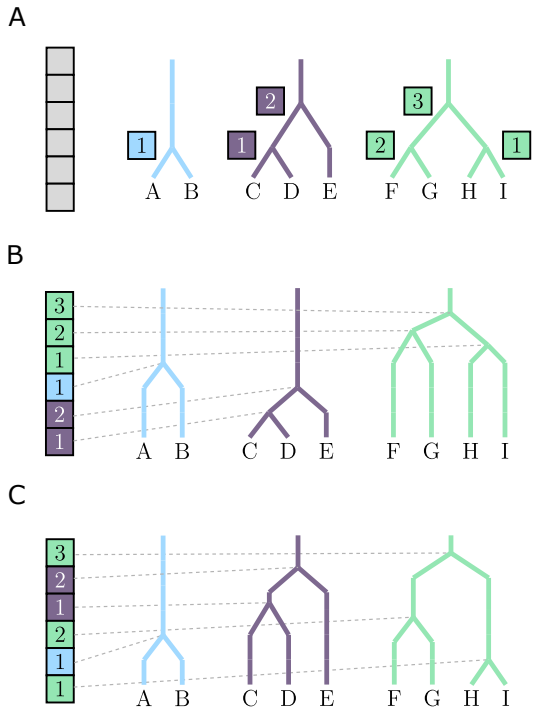

Figure 3: Interweaving of coalescence sequences. (A) Three coalescence sequences. The sequences are represented in three colors. Within a sequence, coalescences occur in a specified order, indicated by numbers within colors. Each of the six coalescences must occur in the interwoven sequence, represented 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).

2.5 Coalescence sequences

A coalescence sequence is an ordered sequence of coalescence events. For example, consider five 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 (Figure 3).

2.6 Joint monophyly

Given a subtree $\mathscr{T}_x$ of $\mathscr{T}$, defined as the node x and all of its descendants, we define the joint monophyly event E_x, representing the event that joint monophyly occurs for gene lineages in the subtree $\mathscr{T}_x$. For joint monophyly to be achieved, each coalescence in $\mathscr{T}_x$ 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.

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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.16.473006; this version posted December 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2.7 Combinatorial functions

We use several combinatorial functions in our calculation. First, $g_{i,j}(T)$ is the probability that i lineages coalesce to j lineages in time T, from Eqn. 6.1 of Tavaré [1984]:

$$g_{i,j}(T) = \sum_{k=j}^{i} e^{-k(k-1)T/2} \frac{(2k-1)(-1)^{k-j} j_{(k-1)} i_{[k]}}{j!\,(k-j)!\,i_{(k)}}, \tag{1}$$

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 $g_{i,1}(\infty)$ for $\lim_{T\to\infty} g_{i,1}(T)$.

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

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

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

Third, the multinomial coefficient

$$W_k(r_1, r_2, \ldots, r_k) = \binom{r_1 + \cdots + r_k}{r_1, r_2, \ldots, r_k}, \tag{3}$$

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

Finally, $Z(s_1, s_2, \ldots, s_k)$ is the probability that in a single population in which k groups are present, k groups of $s_1, s_2, \ldots, s_k$ gene lineages coalesce to a single lineage while preserving joint monophyly of each of the k groups. This function is reported by Theorem 5.1 of Zhu et al. [2011], as follows.

Suppose that $A_1, A_2, \ldots, A_k$ represent sets of lineages for groups $1, 2, \ldots, k$, respectively. Under joint monophyly of groups $1, 2, \ldots, k$, 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 $rb(k)$ ("rb" for 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 joint monophyly of the k groups.

Let $n = \sum_{i=1}^{k} s_i$ be the total number of lineages across all groups. Let $\mathscr{I}(T_k)$ be the set of internal nodes of T_k. For an internal node $v \in \mathscr{I}(T_k)$, 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 joint monophyly of the k groups and labeled topology T_k is

$$Z(s_1, s_2, \ldots, s_k; T_k) = \frac{2^{k-1} \prod_{i=1}^{k} s_i!}{n!} \prod_{v \in \mathscr{I}(T_k)} \frac{1}{\left[\sum_{i=1}^{k} s_i I_v(A_i)\right] - 1}. \tag{4}$$

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

$$Z(s_1, s_2, \ldots, s_k) = \sum_{T_k \in rb(k)} Z(s_1, s_2, \ldots, s_k; T_k). \tag{5}$$

Our notation occasionally leads to values of 0 for some of the arguments of the function in Eqn. 5. In those cases, the quantity is properly computed by dropping those arguments.

3 Mathematical results

For species tree node x, we can compute the probability of the joint monophyly event E_x and a particular output state $\mathbf{n}_x^O$ by recursive decomposition as follows:

$$\begin{aligned}\mathbb{P}(E_x, \mathbf{n}_x^O) &= \sum_{\mathbf{n}_x^I} \mathbb{P}(E_x, \mathbf{n}_x^O | E_{x_L}, E_{x_R}, \mathbf{n}_x^I) \cdot \mathbb{P}(E_{x_L}, E_{x_R}, \mathbf{n}_x^I) \\ &= \sum_{\mathbf{n}_{x_L}^O} \sum_{\mathbf{n}_{x_R}^O} \mathbb{P}(E_x, \mathbf{n}_x^O | E_{x_L}, E_{x_R}, \mathbf{n}_x^I) \cdot \mathbb{P}(E_{x_L}, \mathbf{n}_{x_L}^O) \cdot \mathbb{P}(E_{x_R}, \mathbf{n}_{x_R}^O),\end{aligned} \tag{6}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.16.473006; this version posted December 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where x_L and x_R are the daughter nodes of x, and the second step is due to independence of these nodes and the fact that $\mathbf{n}_x^I = \mathbf{n}_{x_L}^O + \mathbf{n}_{x_R}^O$. Taking x to be the species tree root, $\mathbb{P}(E_{\text{root}}, \mathbf{n}_{\text{root}}^O = (0, \ldots, 0, 1))$ is the joint monophyly probability for the entire gene genealogy.

To compute $\mathbb{P}(E_{\text{root}})$, we use a pruning algorithm—a familiar approach in phylogenetics in general [Felsenstein, 2004, p. 253]. First calculating the probability $\mathbb{P}(E_x, \mathbf{n}_x^O | E_{x_L}, E_{x_R}, \mathbf{n}_{x_L}^O, \mathbf{n}_{x_R}^O)$—the probability of obtaining the joint monophyly event E_x and an output state $\mathbf{n}_x^O$ given the events E_{x_L} and E_{x_R} and their corresponding output states $\mathbf{n}_{x_L}^O$ and $\mathbf{n}_{x_R}^O$—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 joint monophyly and result in the specified output state. We can compute this probability by specifying an input state, computing the probability that joint monophyly 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 joint monophyly 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 joint monophyly.

For (i), the probability that the correct number of coalescences occurs is $g_{|\mathbf{n}_x^I|, |\mathbf{n}_x^O|}(T_x)$ (see 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 $\mathbf{n}_x^I$ and output state $\mathbf{n}_x^O$. Consider a label $i \in \{1, 2, \ldots, k\}$. 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 $|\mathbf{n}_x^I|$ or $|\mathbf{n}_x^O|$, 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 joint monophyly 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 joint monophyly 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 joint monophyly, 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 $n_{x,i}^I$ input lineages of label i can coalesce to the correct number of output lineages of label i, or $n_{x,i}^O$. This number of subsequences is $I_{n_{x,i}^I, n_{x,i}^O}$ (Eqn. 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 (i) involving at least two mixed lineages and no lost labels, (ii) involving at least two lost labels and no mixed lineages, or (iii) involving 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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.16.473006; this version posted December 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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 $\ell + m_I$: ℓ 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, $S_2(\ell + m_I, m_0)$. An algorithm for producing these partitions is presented in Knuth [2011].

However, two additional conditions must be met.

1. $m_O \leq m_I + \lfloor \ell \rfloor / 2$.
2. No element of the partition can consist solely of a single one of the ℓ 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 $\lfloor \ell \rfloor / 2$, if the maximal number of pairs of lost labels coalesce with each other.

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 define 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 $\mathscr{P}$ the set of permissible partitions of the set $L \cup M_I$, where L is the set of ℓ lost labels and M_I is the set of m_I input mixed lineages. Let $U = K \setminus L$, where $K = \{1, 2, \ldots, k\}$. Let P be a partition in $\mathscr{P}$.

Consider an element p of P. This element is associated with a set $L_p \subset L$ of lost labels and a set M_p 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 $I_{(\sum_{j \in L_p} r_j) + m_p, 1}$ (Eqn. 2). The fraction of these subsequences that satisfy joint monophyly is $Z(\mathbf{v}_p)$, where Z is the probability of joint monophyly of an arbitrary number of groups in a single population (Eqn. 5). The argument $\mathbf{v}_p$ 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 7-species tree. If partition p contains lost labels 1 and 6 and three input mixed lineages, then $\mathbf{v}_p = (r_1, 0, 0, 0, 0, r_6, 0, 1, 1, 1)$.

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 joint monophyly gives the total number of subsequences that both have the correct number of coalescences and that satisfy joint monophyly:

$$J_p = I_{(\sum_{j \in L_p} r_j) + m_p, 1} \, Z(\mathbf{v}_p). \tag{7}$$

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.16.473006; this version posted December 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

number of coalescences is $s_i - r_i$, noting that U_i is the ith surviving label (enumerated in arbitrary order) and abbreviating $s_i = n^I_{x,U_i}$ and $r_i = n^O_{x,U_i}$ for convenience.

For each i with $|U|+1 \leq i \leq |U|+|P|$, the number of coalescences is $|P_{i-|U|}|-1$ coalescences, where P_j is the jth element of P (again enumerated in arbitrary order). Hence, the number of ways to interweave the $|U|+|P|$ coalescence subsequences is (from Eqn. 3)

$$W_{|U|+|P|}\left(s_1 - r_1, s_2 - r_2, \ldots, s_{|U|} - r_{|U|}, |P_1|-1, |P_2|-1, \ldots, |P_{|P|}|-1\right). \tag{8}$$

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 joint monophyly for a given input state and output state is

$$C_{\mathbf{n}^I_x,\mathbf{n}^O_x} = \sum_{P\in\mathscr{P}} \left(\prod_{i\in U} I_{s_i,r_i}\right)\left(\prod_{p\in P} J_p\right) W_{|U|+|P|}\left(s_1 - r_1, s_2 - r_2, \ldots, s_{|U|} - r_{|U|}, |P_1|-1, |P_2|-1, \ldots, |P_{|P|}|-1\right). \tag{9}$$

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 $\mathscr{P}$.

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 $P \in \mathscr{P}$ 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. Eqn. 9 becomes

$$C_{\mathbf{n}^I_x,\mathbf{n}^O_x} = \left(\prod_{i\in U} I_{s_i,r_i}\right) W_{|U|}\left(s_1 - r_1, s_2 - r_2, \ldots, s_{|U|} - r_{|U|}\right). \tag{10}$$

3.6 Completing the computation

The total number of coalescence sequences in a branch given an input state and an output state is $I_{|\mathbf{n}^I_x|,|\mathbf{n}^O_x|}$ (Eqn. 2). The number that satisfy joint monophyly is $C_{\mathbf{n}^I_x,\mathbf{n}^O_x}$, following Eqn. 9. From Eqn. 1, the probability of obtaining a particular number of coalescences in a branch of length T_x is $g_{|\mathbf{n}^I_x|,|\mathbf{n}^O_x|}(T_x)$.

We conclude that in Eqn. 6 for the probability of joint monophyly in branch x together with an output state, the recursive step that computes the conditional probability of joint monophyly and the output state given that joint monophyly is maintained in the daughter branches x_L and x_R and given the input state is

$$\mathbb{P}(E_x, \mathbf{n}^O_x | E_{x_L}, E_{x_R}, \mathbf{n}^I_x) = g_{|\mathbf{n}^I_x|,|\mathbf{n}^O_x|}(T_x)\, \frac{C_{\mathbf{n}^I_x,\mathbf{n}^O_x}}{I_{|\mathbf{n}^I_x|,|\mathbf{n}^O_x|}}, \tag{11}$$

where $C_{\mathbf{n}^I_x,\mathbf{n}^O_x}$ is from Eqn. 9 and $I_{|\mathbf{n}^I_x|,|\mathbf{n}^O_x|}$ is from Eqn. 2. This result, applied recursively starting from $x = \text{root}$ with $\mathbf{n}^O_x = (0,\ldots,0,1)$, yields the probability of joint monophyly over all species $1, 2, \ldots k$.

3.7 Deriving previous results

We can use Eqn. 11 to derive previously-known results on the probability of joint monophyly 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: $\mathbf{n}^I_x = (s_1, s_2, \ldots, s_k, 0)$, where s_i is the sample size of group i. The output state is $\mathbf{n}^O_x = (0,\ldots,0,1)$, with size $|\mathbf{n}^O_x| = 1$. Branch T_x has infinite length. The summation in Eqn. 6 is trivial, and applying Eqn. 11, we obtain

$$\begin{aligned}\mathbb{P}(E_x, \mathbf{n}^O_x) &= \mathbb{P}(E_x, \mathbf{n}^O_x | E_{x_L}, E_{x_R}, \mathbf{n}^I_x)\\ &= g_{|\mathbf{n}^I_x|,1}(\infty)\, \frac{C_{\mathbf{n}^I_x,\mathbf{n}^O_x}}{I_{|\mathbf{n}^I_x|,1}}.\end{aligned}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.16.473006; this version posted December 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Study	Number of Populations	Number 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 paper	Arbitrary	Arbitrary	3.6

Table 1: Analytical results for the probability of joint monophyly for arbitrary sample sizes. Eqn. 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].

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

$$\mathbb{P}(E_x, \mathbf{n}_x^O) = \frac{J_p\, W_1(|\mathbf{n}_x^I| - 1)}{I_{|\mathbf{n}_x^I|,1}} = \frac{J_p}{I_{|\mathbf{n}_x^I|,1}}.$$

Using our notation from Section 3.4, the partition vector is $\mathbf{v}_p = (s_1, s_2, \ldots, s_k)$. We use Eqn. 7 to obtain

$$\begin{aligned}\mathbb{P}(E_x, \mathbf{n}_x^O) &= \frac{I_{|\mathbf{n}_x^I|,1}\, Z(\mathbf{v}_p)}{I_{|\mathbf{n}_x^I|,1}} = Z(\mathbf{v}_p) \\ &= Z(s_1, s_2, \ldots, s_k). \end{aligned} \tag{12}$$

Note that $Z(s_1, s_2, \ldots, s_k)$ is exactly the quantity in Eqn. 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 specific 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 $\mathbf{n}_x^I = (0, \ldots, s_i, \ldots, 0)$, and the output state is $\mathbf{n}_x^O = (0, \ldots, r_i, \ldots, 0)$.

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

$$C_{\mathbf{n}_x^I, \mathbf{n}_x^O} = I_{s_i, r_i} W_1(s_i - r_i) = I_{s_i, r_i}.$$

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

$$\begin{aligned}\mathbb{P}(E_x, \mathbf{n}_x^O) &= g_{|\mathbf{n}_x^I|, |\mathbf{n}_x^O|}(T_x)\, \frac{C_{\mathbf{n}_x^I, \mathbf{n}_x^O}}{I_{|\mathbf{n}_x^I|, |\mathbf{n}_x^O|}} \\ &= g_{s_i, r_i}(T_x)\, \frac{I_{s_i, r_i}}{I_{s_i, r_i}} \\ &= g_{s_i, r_i}(T_x). \end{aligned} \tag{13}$$

Thus, the general computation in Eqn. 11 reduces to Eqn. 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 $r_1 \leq s_1$ and $r_2 \leq s_2$ be the numbers of lineages of species 1 and 2 that enter the root node. There are three species

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.16.473006; this version posted December 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

tree nodes: the root x, leaf x_1 for species 1, and leaf x_2 for species 2. The input and output states are $\mathbf{n}^I_{x_1} = (s_1, 0, 0)$, $\mathbf{n}^I_{x_2} = (0, s_2, 0)$, $\mathbf{n}^O_{x_1} = (r_1, 0, 0)$, $\mathbf{n}^O_{x_2} = (0, r_2, 0)$, $\mathbf{n}^I_x = (r_1, r_2, 0)$, and $\mathbf{n}^O_x = (0, 0, 1)$.

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 $P = \{\{1, 2\} \to m_1\}$.

Because x_1 and x_2 are leaves, from Eqn. 13, $\mathbb{P}(E_{x_1}, \mathbf{n}^O_{x_1}) = g_{s_1,r_1}(T_1)$ and $\mathbb{P}(E_{x_2}, \mathbf{n}^O_{x_2}) = g_{s_2,r_2}(T_2)$. From Eqn. 12, for a particular r_1 and r_2, we have $\mathbb{P}(E_x, \mathbf{n}^O_x | E_{x_1}, E_{x_2}, \mathbf{n}^I_x) = Z(r_1, r_2)$. Substituting into Eqn. 6,

$$\begin{aligned}\mathbb{P}(E_x, \mathbf{n}^O_x) &= \sum_{\mathbf{n}^O_{x_1}} \sum_{\mathbf{n}^O_{x_2}} \mathbb{P}(E_x, \mathbf{n}^O_x | E_{x_1}, E_{x_2}, \mathbf{n}^I_x) \cdot \mathbb{P}(E_{x_1}, \mathbf{n}^O_{x_1}) \cdot \mathbb{P}(E_{x_2}, \mathbf{n}^O_{x_2}) \\ &= \sum_{r_1=1}^{s_1} \sum_{r_2=1}^{s_2} Z(r_1, r_2)\, g_{s_1,r_1}(T_1)\, g_{s_2,r_2}(T_2). \end{aligned} \tag{14}$$

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 Eqns. 4 and 5,

$$\begin{aligned} Z(r_1, r_2) = Z(r_1, r_2; T_2) &= \frac{2 r_1!\, r_2!}{(r_1 + r_2)!} \frac{1}{r_1 + r_2 - 1} \\ &= \frac{2}{r_1 + r_2 - 1} \binom{r_1 + r_2}{r_1}^{-1}, \end{aligned} \tag{15}$$

which matches Lemma 4.3 in Zhu et al. [2011], Eqn. 6 in Brown [1994], and Eqn. 9 in Rosenberg [2003].

Substituting Eqn. 15 into Eqn. 14, we have

$$\mathbb{P}(E_x, \mathbf{n}^O_x) = \sum_{r_1=1}^{s_1} \sum_{r_2=1}^{s_2} g_{s_1,r_1}(T_1)\, g_{s_2,r_2}(T_2) \frac{2}{r_1 + r_2 - 1} \binom{r_1 + r_2}{r_1}^{-1}. \tag{16}$$

We therefore obtain Eqn. 14 from Rosenberg [2003]: the probability of reciprocal monophyly of two species in a two-species tree.

3.7.4 3 species in a 3-species tree

In this section, we recapitulate the probability of joint monophyly for 3 species in a 3-species tree, as provided in Eqn. 5 in Mehta and Rosenberg [2019]. It suffices to describe the reduction of our Eqn. 11 to Eqns. 6 and 9 in Mehta and Rosenberg [2019], giving the conditional probability of joint monophyly within the internal node I of $\mathscr{T}$ given a particular input state $\mathbf{n}^I_I$ and output state $\mathbf{n}^O_I$, and the conditional probability of joint monophyly in the species tree root R given a particular input state $\mathbf{n}^I_R$.

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:

$$\begin{aligned} \mathbf{n}^I_A &= (p, 0, 0, 0), & \mathbf{n}^O_A &= (s, 0, 0, 0) \\ \mathbf{n}^I_B &= (0, q, 0, 0), & \mathbf{n}^O_B &= (0, t, 0, 0) \\ \mathbf{n}^I_C &= (0, 0, r, 0), & \mathbf{n}^O_C &= (0, 0, y, 0) \\ \mathbf{n}^I_I &= (s, t, 0, 0), & \mathbf{n}^O_I &= (w, x, 0, m) \\ \mathbf{n}^I_R &= (w, x, y, m), & \mathbf{n}^O_R &= (0, 0, 0, 1). \end{aligned}$$

Eqns. 6 and 9 in Mehta and Rosenberg [2019] are special cases of a term in Eqn. 3 from Mehta and Rosenberg [2019], which corresponds to our Eqn. 11. Comparing Eqn. 11 to Eqn. 3 from Mehta and Rosenberg [2019] indicates that to obtain Eqn. 6 of Mehta and Rosenberg [2019], we must show that the quantity K_I from Mehta and Rosenberg [2019] satisfies

$$K_I = \frac{C_{\mathbf{n}^I_I, \mathbf{n}^O_I}}{I_{|\mathbf{n}^I_I|, |\mathbf{n}^O_I|}}.$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.16.473006; this version posted December 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

To obtain Eqn. 9 from Mehta and Rosenberg [2019], we must show that with $\mathbf{v}_p = (w, x, y, m)$, the quantity K_{root} from Mehta and Rosenberg [2019] satisfies

$$K_{\text{root}} = \frac{C_{\mathbf{n}_R^I, \mathbf{n}_R^O}}{I_{|\mathbf{n}_R^I|,1}} = Z(\mathbf{v}_p). \tag{17}$$

First, we consider internal node I. The nontrivial cases of Eqn. 6 from Mehta and Rosenberg [2019] are:

$$K_I = \begin{cases} \frac{I_{s,w}\, I_{t,x}\, W_2(s-w,t-x)}{I_{s+t,w+x}} & \text{Case 1: } s,t,w,x \geq 1; m = 0 \\ \frac{I_{s,1}\, I_{t,1}\, W_2(s-1,t-1)}{I_{s+t,1}} & \text{Case 2: } s,t \geq 1; w = x = 0; m = 1. \end{cases} \tag{18}$$

Eqn. 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 Eqn. 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 $\mathscr{P}$ of the set of lost labels and mixed lineages.

We use Eqn. 10 to obtain

$$\begin{aligned} C_{\mathbf{n}_I^I, \mathbf{n}_I^O} &= \prod_{i=1}^{2} I_{s_i, r_i}\, W_2\left(s_1 - r_1, s_2 - r_2\right) \\ &= I_{s,w}\, I_{t,x}\, W_2\left(s - w, t - x\right), \end{aligned}$$

and so we have:

$$\begin{aligned} \frac{C_{\mathbf{n}_I^I, \mathbf{n}_I^O}}{I_{|\mathbf{n}_I^I|, |\mathbf{n}_I^O|}} &= \frac{I_{s,w}\, I_{t,x}\, W_2\left(s - w, t - x\right)}{I_{s+t,w+x}} \\ &= K_I. \end{aligned} \tag{19}$$

Case 2 in Eqn. 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_I: $P = \{p\}$, with $p = \{1, 2\} \to m_I$. We use Eqn. 9 and Eqn. 7 to obtain:

$$\begin{aligned} C_{\mathbf{n}_I^I, \mathbf{n}_I^O} &= J_p\, W_1\left(s + t - 1\right) = J_p \\ &= I_{\left(\sum_{j \in L_1} r_j\right) + m_1, 1}\, Z(\mathbf{v}_p) \\ &= I_{s+t,1}\, Z(\mathbf{v}_p). \end{aligned}$$

The vector $\mathbf{v}_p$ is $(s, t, 0, 0)$. Note that $Z(s, t, 0, 0) = Z(s, t)$, so we can use Eqns. 15, 2, and 3 to obtain

$$\begin{aligned} \frac{C_{\mathbf{n}_I^I, \mathbf{n}_I^O}}{I_{s+t,1}} &= \frac{2}{s + t - 1} \binom{s+t}{s}^{-1} \\ &= \frac{I_{s,1}\, I_{t,1}\, W_2(s-1, t-1)}{I_{s+t,1}} \\ &= K_I, \end{aligned} \tag{20}$$

where the last step comes from Eqn. 18. Hence, we have $K_I = C_{\mathbf{n}_I^I, \mathbf{n}_I^O} / I_{s+1,1}$, as desired.

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

$$K_{\text{root}} = \begin{cases} f(w,x,y) + f(w,y,x) + f(x,y,w) & \text{Case 1: } w,x,y \geq 1; m = 0, \\ \frac{I_{y,1}}{I_{y+1,1}} & \text{Case 2: } y \geq 1; w = x = 0; m = 1, \end{cases} \tag{21}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.16.473006; this version posted December 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where

$$f(w,x,y) = \frac{\sum_{c=1}^{y} I_{w,1}\, I_{x,1}\, I_{y,c}\, W_3(w-1,x-1,y-c)\, I_{c,1}}{I_{w+x+y,1}}. \tag{22}$$

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

$$Z(w,x,y) = \frac{4w!\,x!\,y!}{(w+x+y)!\,(w+x+y-1)}\left(\frac{1}{x+y-1} + \frac{1}{w+y-1} + \frac{1}{w+x-1}\right). \tag{23}$$

Substituting Eqns. 2 and 3 into Eqn. 22 and simplifying, we have

$$f(w,x,y) = \frac{4w!\,x!\,y!}{(w+x+y)!\,(w+x+y-1)}\,\frac{1}{w+x-1}\,\frac{\sum_{c=1}^{y}\binom{w+x-2+y-c}{w+x-2}}{\binom{w+x+y-2}{w+x-1}} \tag{24}$$

$$= \frac{4w!\,x!\,y!}{(w+x+y)!\,(w+x+y-1)}\,\frac{1}{w+x-1}, \tag{25}$$

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

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

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

Now, from Eqns. 23 and 25, we have

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

as required.

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

$$Z(y,1) = \frac{2}{y}\binom{y+1}{y}^{-1} = \frac{2}{y(y+1)}. \tag{26}$$

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

Having demonstrated that our joint monophyly calculation recovers the combinatorial terms K_I and K_{root}, we have therefore recapitulated the joint monophyly probability for three species, as obtained by Mehta and Rosenberg [2019].

3.7.5 2 groups in a k-species tree

Here we recapitulate the probability of joint monophyly for 2 groups in a k-species tree, as shown in Eqn. 5 from Mehta et al. [2016]. It suffices to describe the reduction of our Eqn. 11 to Eqn. 4 from Mehta et al. [2016], describing the conditional probability of monophyly within the a node x of $\mathscr{T}$ given a particular input state $\mathbf{n}_x^I$ and output state $\mathbf{n}_x^O$. More precisely, we must equate our Eqn. 11 to the scenario of joint monophyly in Eqn. 4 of Mehta et al. [2016], obtained by substituting their Eqn. 5 for Case 2 in their Eqn. 4.

Let the input state for node x be $\mathbf{n}_x^I = (s_1, s_2, m_I)$, and let the output state be $\mathbf{n}_O^I = (r_1, r_2, m_O)$. We assume (as is necessary to achieve joint monophyly) 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

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.16.473006; this version posted December 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and 2. Following the labeling of cases in Mehta et al. [2016], the nontrivial cases of Eqn. 4 from Mehta et al. [2016] in the setting of joint monophyly are:

$$K_{SC} = \begin{cases} 1 & \text{Case 1e: } s_1, r_1 \geq 1; s_2 = r_2 = m_I = m_O = 0 \\ 1 & \text{Case 1b: } s_2, r_2 \geq 1; s_1 = r_1 = m_I = m_O = 0 \\ \frac{I_{s_1,1} I_{s_2,1} W_2(s_1-1, s_2-1)}{I_{s_1+s_2,1}} & \text{Case 2: } s_1, s_2 \geq 1, r_1 = r_2 = m_I = 0, m_O = 1 \\ \frac{I_{s_1,r_1} I_{s_2,r_2} W_2(s_1-r_1, s_2-r_2)}{I_{s_1+s_2,r_1+r_2}} & \text{Case 3: } s_1, s_2, r_1, r_2 \geq 1; m_I = m_O = 0, \end{cases} \tag{27}$$

where Cases 1b and 1e in Eqn. 27 are labeled after their corresponding labels in Mehta et al. [2016].

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

$$K_{SC} = \frac{C_{\mathbf{n}_x^I, \mathbf{n}_x^O}}{I_{|\mathbf{n}_x^I|, |\mathbf{n}_x^O|}}. \tag{28}$$

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

$$\begin{aligned} \frac{C_{\mathbf{n}_x^I, \mathbf{n}_x^O}}{I_{|\mathbf{n}_x^I|, |\mathbf{n}_x^O|}} &= \frac{I_{s_i,r_i} W_1(s_i - r_i)}{I_{s_i,r_i}} = 1 \\ &= K_{SC}, \end{aligned}$$

as required for demonstrating Eqn. 28.

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

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

We have therefore shown that our Eqn. 11 reduces to Eqn. 28, recapitulating the joint monophyly probability of two groups in an arbitrary species tree from Mehta et al. [2016].

3.8 Lower and upper bounds based on "strong" joint monophyly

The probability in Eqn. 11 involves many steps and is potentially time-consuming to calculate. We can therefore provide a simpler lower bound by introducing the idea of "strong" joint monophyly. We say that a set of lineages sampled from a species satisfies *strong joint monophyly* if the lineages coalesce to a single lineage in the branch associated with that species. The probability of strong joint monophyly can then be computed from the lengths of these branches, so that the rest of the tree structure does not matter.

The probability of strong joint monophyly is

$$\mathbb{P}(SJM) = \prod_{i=1}^{k} g_{s_i,1}(T_i), \tag{29}$$

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

This probability provides a lower bound on Eqn. 11 because it is only one of many ways that joint monophyly can be achieved; if JM denotes the event of joint monophyly, then $\mathbb{P}(JM) \geq \mathbb{P}(SJM)$. 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 Eqn. 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 $\mathbb{P}(SJM)$ enables an *upper* bound on $\mathbb{P}(JM)$, a bound that holds for any species tree and any distribution of gene lineages across species. This bound is:

$$\mathbb{P}(JM) \leq \frac{1}{3} + \frac{2}{3}\mathbb{P}(SJM). \tag{30}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.16.473006; this version posted December 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

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

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

$$\mathbb{P}(JM) = x + \mathbb{P}(JM|\neg SJM)\,(1 - x), \tag{31}$$

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

$$\mathbb{P}(JM|\neg SJM) \leq \frac{1}{3}. \tag{32}$$

The justification of Eqn. 32 is as follows. The coalescent scenarios that comprise the event $\neg SJM$ 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 pendant edge incident with s. However, joint monophyly (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 $\mathscr{T}$ 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, and the next coalescence will involve at least one of the two from s. The probability that the two ancestral lineages of s coalesce with each other (rather than one coalescing with one of the other r lineages present) is $\binom{2}{2}/\binom{r+2}{2} \leq \frac{1}{3}$ for all $r \geq 1$. Thus, $\mathbb{P}(JM|\neg SJM) \leq \frac{1}{3}$ as claimed.

Combining Eqns. 31 and 32 gives Eqn. 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 joint monophyly, 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 joint monophyly amounts to using the same recursive decomposition as in Eqn. 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 joint monophyly using an example. We consider a class of species trees that appears in 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 joint monophyly with Eqn. 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.

Figure 5 displays the effect of number of species, tree height, and sample size on the probability of joint monophyly 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 in order to produce joint monophyly increases. Hence, the joint monophyly 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 joint monophyly probability increases with increasing tree height. As the sample size increases at a fixed tree height, the number of lineages that must

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.16.473006; this version posted December 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

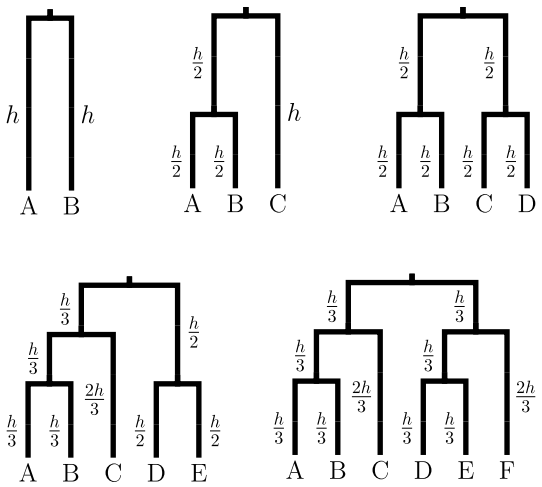

Figure 4: Trees used to explore the effects of tree height and sample size on the probability of joint monophyly.

monophyletically coalesce increases, but no additional time is available for these coalescences; hence, the joint monophyly probability decreases with increasing sample size.

An alternative perspective on the joint monophyly 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 joint monophyly probability that lies below the cutoff. In other words, we calculate the minimum sample size required for an observation of joint monophyly to be improbable at a specified significance level under a specified 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 joint monophyly is observed [Rosenberg, 2007].

Figure 6 plots these minimum sample sizes. They decrease as the cutoff value is increased. In accord with the decrease in joint monophyly probabilities that occurs with an increasing number of species, for fixed tree height, the minimal sample size required for achieving a joint monophyly probability below a specified cutoff decreases with an increasing number of species. The minimal sample size increases with increasing tree height; as tree height grows, joint monophyly is probable even for large samples, so that very large samples might be required for a joint monophyly observation to be surprising. In most scenarios plotted, samples of 6 to 8 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 displays the probability of joint monophyly against the corresponding probability of strong joint monophyly from Eqn. 29. For each combination of a number of species, tree height, and sample size considered in Figure 5, the probability of strong joint monophyly is calculated, and a point is plotted that pairs the probability of strong joint monophyly with the probability of joint monophyly from Figure 5.

As strong joint monophyly is a stricter condition than joint monophyly, the probability of strong joint monophyly is necessarily less than or equal to the probability of joint monophyly (Section 3.8). Traversing the figure from left to right, or from bottom to top, the tree height increases. For large tree heights, joint monophyly is closely approximated by strong joint monophyly, as represented by the proximity of the curves plotted to the $y = x$ line; the event of strong joint monophyly is the primary driver of joint monophyly. For smaller tree heights, the probability of strong joint monophyly is substantially lower than than the probability of joint monophyly, as configurations in which joint monophyly 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 joint monophyly and strong joint monophyly, or of the sample size. Thus, by curve-fitting, it would be possible to empirically transform the easily-computable SJM probability that approximates the joint monophyly probability.

For the case of 2 species and 2 lineages per species, using the 2-species tree in Figure 4, the probability

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.16.473006; this version posted December 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

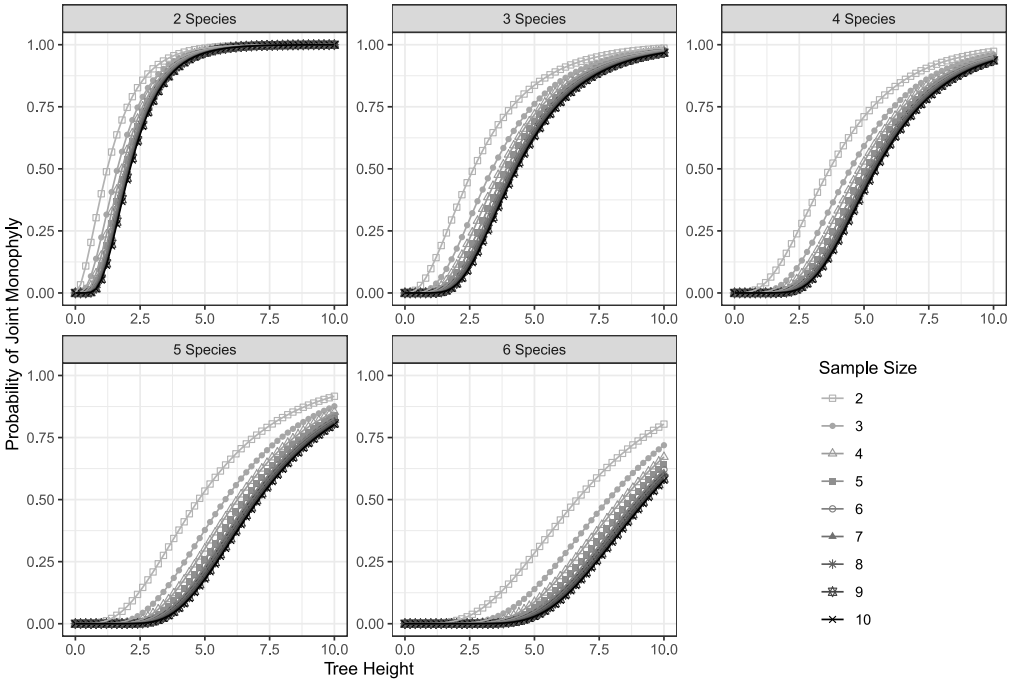

Figure 5: Joint monophyly probabilities for various numbers of species, tree heights, and sample sizes. Probabilities are obtained using Eqn. 11, with the same sample size assigned to each species. Each panel is labeled by the number of species.

of joint monophyly (JM) from Eqn. 16 is

$$\begin{aligned}\mathbb{P}(JM) &= \sum_{r_1=1}^{2}\sum_{r_2=1}^{2} g_{2,r_1}(h)\, g_{2,r_2}(h) \frac{2}{r_1+r_2-1}\binom{r_1+r_2}{r_1}^{-1} \\ &= 1 - \frac{4}{3}e^{-h} + \frac{4}{9}e^{-2h}. \end{aligned} \tag{33}$$

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

$$\mathbb{P}(JM) = \frac{1}{9}\left[2\sqrt{\mathbb{P}(SJM)}+1\right]^2. \tag{34}$$

Eqn. 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 joint monophyly 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 (Eqn. 11), we have obtained as special cases several previous results for the probability of joint monophyly: 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 joint monophyly were restricted to small numbers of groups (4 or fewer), small trees (4 species or fewer), or both.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.16.473006; this version posted December 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

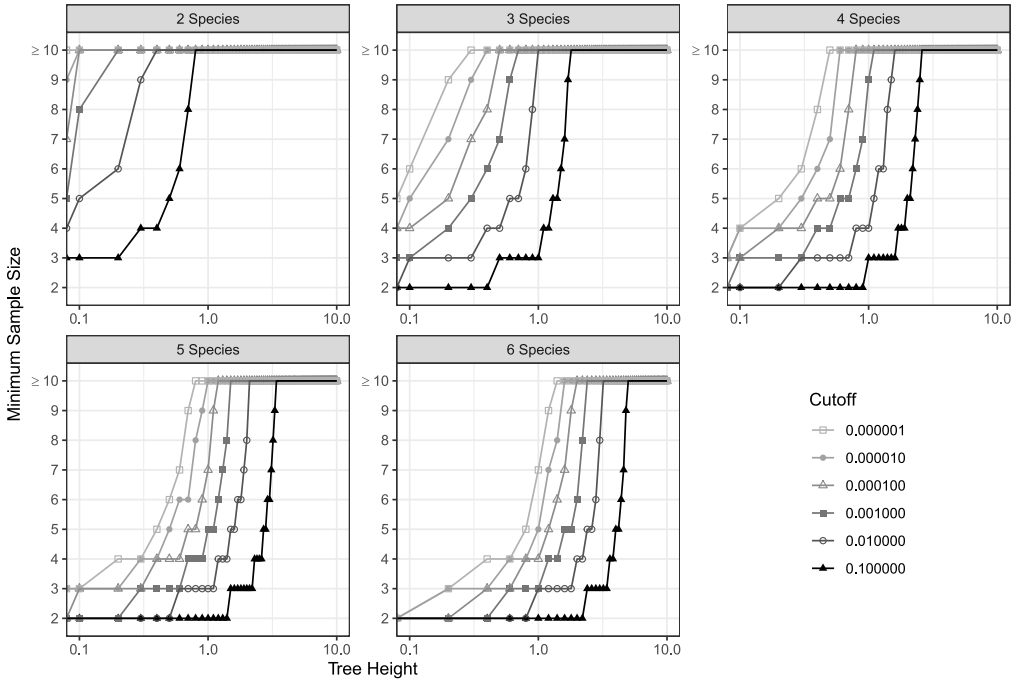

Figure 6: Minimum sample sizes for the probability of joint monophyly to decrease below a particular cutoff probability, for varying tree height and number of species. Panel title indicates number of species.

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, though the pruning algorithm renders the tree traversal to be polynomial-time in the number of species, the computation time is not polynomial-time in the number of species or sample size, due 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 continuous-time-Markov-chain approach that is convenient for computation (Appendix A). Using the CTMC approach, we have seen that the joint monophyly calculation reproduces sensible patterns in the effects of model parameters on monophyly probabilities.

The new algorithmic approach will be useful in problems where joint monophyly of multiple groups is of interest—such as in phylogeographic and species delimination 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 Monophyler [Mehta et al., 2016].

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.16.473006; this version posted December 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

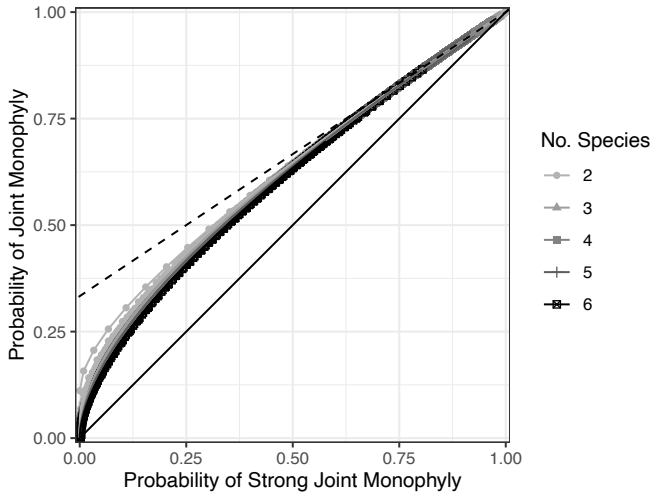

Figure 7: The probability of joint monophyly (Eqn. 11) in relation to the probability of strong joint monophyly (Eqn. 29). Strong joint monophyly provides a lower bound for joint monophyly. For each combination consisting of a number of species (2 to 6) and a sample size (2 to 10), a curve links points 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 joint monophyly provided by Eqn. 30.

A Calculating probabilities with a CTMC approach

A.1 Mathematical approach

We now produce an alternative approach to calculating the probability of joint monophyly: a continuous-time Markov chain (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 first must define a state space. Let $\mathscr{T}_x$ 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 "mixed" labels. Instead, we classify the species labels $\{1, 2, \ldots, k\}$ 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 joint monophyly 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 joint monophyly. Coalescences of free lineages with other free lineages preserve joint monophyly, reducing the number of free lineages. In this formulation, "mixed" lineages are free.

The state space for a branch x therefore consists of a "failure" state F, which represents the situation where joint monophyly has been violated, and a set of vectors $\mathbf{v}_x$ that keep track of the list of lineage counts for the k labels. The ith element of $\mathbf{v}_x$, $v_{x,i}$, is the number of *labels* with i *extant lineages*, with $v_{x,1}$ 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 $\mathscr{T}_x$, as no label can gain lineages through coalescence. If S_x is the set of species in $\mathscr{T}_x$, 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 due to coalescence. Let us define

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

as the total number of lineages for state v_x. For state $\mathbf{v}_x$, we have three possible transitions, corresponding

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.16.473006; this version posted December 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

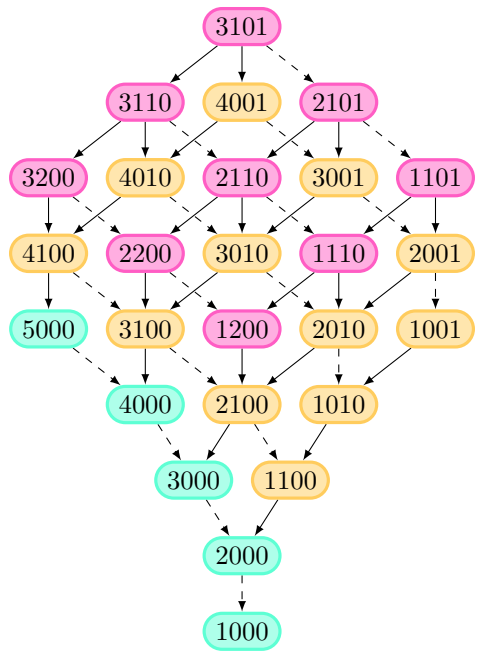

Figure 8: State space for the continuous-time Markov chain for the example branch in Section A.2. States are colored by the number of species for which joint monophyly is not yet determined (pink, two; yellow, one; green, none). Intraspecies transitions use a solid line; interspecies transitions use a dashed line. The Failure state is excluded; all states except those colored green can transition to the failure state.

to intralabel coalescences, interlabel coalescences that preserve joint monophyly, and interlabel coalescences that do not preserve joint monophyly.

1. An intralabel coalescence within a label of size $i > 1$ reduces the number of lineages of that label by 1. $v_{x,i} \to v_{x,i} - 1$, and $v_{x,i-1} \to v_{x,i-1} + 1$. There are $\binom{V_x}{2}$ possible coalescences, and among those, $v_{x,i}\binom{i}{2}$ lead to this state transition. The conditional probability that a coalescence has this transition given that a coalescence occurs is $v_{x,i}\binom{i}{2}/\binom{V_x}{2}$.

2. An interlabel coalescence that preserves joint monophyly can only occur between free lineages. Thus, it reduces $v_{x,1} \to v_{x,1} - 1$. The conditional probability that a coalescence has this transition is $\binom{v_{x,1}}{2}/\binom{V_x}{2}$.

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

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

A.2 Example transitions and transition rate matrices

Consider a branch with input lineages from four species—two with 1 lineage, one with 2 lineages, and one with 4 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 $\mathbf{v}_x = (3, 1, 0, 1)$. The total number of lineages present is $V_x = 9$ (Eqn. 35), so there are $\binom{V_x}{2} = \binom{9}{2} = 36$ possible coalescences. The maximal sample size is $s_{x,m} = 4$.

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

$$\frac{v_{x,2}\binom{2}{2}}{36} = \frac{1(1)}{36} = \frac{1}{36},$$

This transition converts a species with 2 lineages to one with 1 lineage, or $(3, 1, 0, 1) \to (4, 0, 0, 1)$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.16.473006; this version posted December 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

$$\frac{v_{x,4}\binom{4}{2}}{36} = \frac{1(6)}{36} = \frac{1}{6}.$$

The species with 4 lineages transitions to one with 3 lineages. The state transition is $(3, 1, 0, 1) \to (3, 1, 1, 0)$.

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

$$\frac{\binom{v_{x,1}}{2}}{36} = \frac{\binom{3}{2}}{36} = \frac{3}{36} = \frac{1}{12}.$$

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

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

$$\begin{aligned}\mathbb{P}((3,1,0,1) \to F) &= 1 - \frac{\binom{v_{x,1}}{2}}{36} - \sum_{i=2}^{4} \frac{v_{x,i}\binom{i}{2}}{36} \\ &= 1 - \frac{\binom{3}{2}}{36} - \frac{(1)\binom{2}{2}}{36} - \frac{(0)\binom{3}{2}}{36} - \frac{(1)\binom{4}{2}}{36} \\ &= \frac{13}{18}.\end{aligned}$$

The full transition matrix for this species tree branch includes every state attainable by any number of coalescences beginning with state $(3, 1, 0, 1)$. Figure 8 displays 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 Table 2.

To get the CTMC, the transition matrix for a branch must be converted into a transition rate matrix, or a Q matrix. We first scale the transition rates by noting that for a state $\mathbf{v}_x$, 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 first 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 Table 3.

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 is [Grimmett and Stirzaker, 2020]

$$p_x \cdot \exp(Q_x T_x). \tag{36}$$

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 `getnodeoutput` on root node of tree;

The tree structure is created within Python from a string in Newick format by using the function `Tree` in the package `ete3`. 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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.16.473006; this version posted December 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The recursive function `getnodeoutput` is described in the following psuedocode:

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 $i \neq s_x$;
  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 Eqn. 36;
  return vector of output probabilities;
else
  apply `getnodeoutput` to left daughter node of node x;
  apply `getnodeoutput` 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 Eqn. 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.

We first 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, $v_{x,i}$, is $v_{L,i} + v_{R,i}$ for each $i > 1$. Similarly, free lineages in the output states of nodes L and R remain free as input lineages to x, so $v_{x,1} = v_{L,1} + v_{R,1}$. Thus, summing vectors, an input state $\mathbf{v}_x$ obtained from a pair of output states $\mathbf{v}_L$ and $\mathbf{v}_R$ is $\mathbf{v}_x = \mathbf{v}_L + \mathbf{v}_R$. The probability of the input state $\mathbf{v}_x$ obtained by summing output states $\mathbf{v}_L$ and $\mathbf{v}_R$ is the product of the probabilities of the output states $\mathbf{v}_L$ from node L and $\mathbf{v}_R$ 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 $\mathbb{P}(F)_L$ and $\mathbb{P}(F)_R$ are the output failure probabilities for L and R, respectively, then the input probability of failure is $\mathbb{P}(F)_L\,[1-\mathbb{P}(F)_R]+[1-\mathbb{P}(F)_L]\,\mathbb{P}(F)_R+\mathbb{P}(F)_L\,\mathbb{P}(F)_R$.

The result of Step (*) is a vector of input states $\mathbf{I}_x$ and a vector of their probabilities $\mathbf{p}_{\mathbf{I}_x}$.

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 $\mathbf{p}_x$ from $\mathbf{p}_{\mathbf{I}_x}$;
(IV) compute output state probabilities using $\mathbf{p}_x$, Q_x, and Eqn. 36;

To use Eqn. 36 to obtain output probabilities, the state space of $\mathbf{p}_x$ must include all possible output states. Thus, it is necessary to find all possible output states $\mathbf{O}_x$ for a set of input states $\mathbf{I}_x$.

(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

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.16.473006; this version posted December 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

states $\mathbf{O}_x$ is computed via a recursive algorithm that finds 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 $\mathbf{O}_x$ 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 $\mathbf{I}_x$ and the corresponding probabilities $\mathbf{p}_{\mathbf{I}_x}$ 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 $\mathbf{O}_x$ that are not part of $\mathbf{I}_x$. The rearranged input probability vector is $\mathbf{p}_x$.

(Step IV) Once $\mathbf{p}_x$ is obtained, Eqn. 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`.

Acknowledgments. 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. We acknowledge support from NIH grant R01 GM131404.

References

Brian S Arbogast, Scott V Edwards, John Wakeley, Peter Beerli, and Joseph B Slowinski. Estimating divergence times from molecular data on phylogenetic and population genetic timescales. *Ann Rev Ecol Syst*, 33(1):707–740, 2002.

Allan J Baker, Erika S Tavares, and Rebecca F Elbourne. Countering criticisms of single mitochondrial DNA gene barcoding in birds. *Mol Ecol Resour*, 9:257–268, 2009.

Johannes Bergsten, David T Bilton, Tomochika Fujisawa, Miranda Elliott, Michael T Monaghan, Michael Balke, Lars Hendrich, Joja Geijer, Jan Herrmann, Garth N Foster, et al. The effect of geographical scale of sampling on DNA barcoding. *Syst Biol*, 61(5):851–869, 2012.

C William Birky, Cynthia Wolf, Heather Maughan, Linnea Herbertson, and Elena Henry. Speciation and selection without sex. *Hydrobiologia*, 546(1):29–45, 2005.

James KM Brown. Probabilities of evolutionary trees. *Syst Biol*, 43(1):78–91, 1994.

David Bryant, Remco Bouckaert, Joseph Felsenstein, Noah A Rosenberg, and Arindam RoyChoudhury. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol Biol Evol*, 29(8):1917–1932, 2012.

Bryan C Carstens and L Lacey Knowles. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from *Melanoplus* grasshoppers. *Syst Biol*, 56(3):400–411, 2007.

Bryan C Carstens and Corinne L Richards. Integrating coalescent and ecological niche modeling in comparative phylogeography. *Evolution*, 61(6):1439–1454, 2007.

Alison Cloutier, Timothy B Sackton, Phil Grayson, Michele Clamp, Allan J Baker, and Scott V Edwards. Whole-genome analyses resolve the phylogeny of flightless birds (palaeognathae) in the presence of an empirical anomaly zone. *Syst Biol*, 68(6):937–955, 2019.

Kevin De Queiroz. Species concepts and species delimitation. *Syst Biol*, 56(6):879–886, 2007.

Sam Efromovich and Laura S Kubatko. Coalescent time distributions in trees of arbitrary size. *Stat Appl Genet Mol Biol*, 7(1), 2008.

Joseph Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, 17(6): 368–376, 1981.

Joseph Felsenstein. *Inferring Phylogenies*, volume 2. Sinauer Associates, Sunderland, MA, 2004.

Izrail Solomonovich Gradshteyn and Iosif Moiseevich Ryzhik. *Table of integrals, series, and products*. Academic press, 2014.

Geoffrey R Grimmett and David S Stirzaker. *Probability and Random Processes (4th ed.)*. Oxford University Press, Oxford, UK, 2020.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.16.473006; this version posted December 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Asger Hobolth, Lars Nørvang Andersen, and Thomas Mailund. On computing the coalescence time density in an isolation-with-migration model with few samples. *Genetics*, 187(4):1241–1243, 2011.

Richard R Hudson and Jerry A Coyne. Mathematical consequences of the genealogical species concept. *Evolution*, 56(8):1557–1565, 2002.

Gunther Jansen, Riitta Savolainen, and Kari Vepsäläinen. Phylogeny, divergence-time estimation, biogeography and social parasite–host relationships of the Holarctic ant genus *Myrmica* (Hymenoptera: Formicidae). *Mol Phylogenet Evol*, 56(1):294–304, 2010.

Donald E Knuth. *The Art of Computer Programming, Volume 4A: Combinatorial Algorithms, Part 1*. Addison-Wesley Professional, Boston, MA, 2011.

Laura S Kubatko, H Lisle Gibbs, and Erik W Bloomquist. Inferring species-level phylogenies and taxonomic distinctiveness using multilocus data in *Sistrurus* rattlesnakes. *Syst Biol*, 60(4):393–409, 2011.

Nicholas A Loehr. *Combinatorics*. Chapman and Hall/CRC, London, UK, 2017.

Rohan S Mehta and Noah A Rosenberg. The probability of reciprocal monophyly of gene lineages in three and four species. *Theor Popul Biol*, 129:133–147, 2019.

Rohan S Mehta, David Bryant, and Noah A Rosenberg. The probability of monophyly of a sample of gene lineages on a species tree. *Proc Natl Acad Sci USA*, 113(29):8002–8009, 2016.

Craig Moritz. Defining 'evolutionarily significant units' for conservation. *Trends Ecol Evol*, 9(10):373–375, 1994.

Matthew E Neilson and Carol A Stepien. Evolution and phylogeography of the tubenose goby genus *Proterorhinus* (Gobiidae: Teleostei): evidence for new cryptic species. *Biol J Linn Soc*, 96(3):664–684, 2009.

Pekka Pamilo and Masatoshi Nei. Relationships between gene trees and species trees. *Mol Biol Evol*, 5(5):568–583, 1988.

Christian Rabeling, Ted R Schultz, Naomi E Pierce, and Maurício Bacci Jr. A social parasite evolved reproductive isolation from its fungus-growing ant host in sympatry. *Curr Biol*, 24(17):2047–2052, 2014.

Noah A Rosenberg. The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. *Evolution*, 57(7):1465–1477, 2003.

Noah A Rosenberg. Statistical tests for taxonomic distinctiveness from observations of monophyly. *Evolution*, 61(2): 317–323, 2007.

Arindam RoyChoudhury and Elizabeth A Thompson. Ascertainment correction for a population tree via a pruning algorithm for likelihood computation. *Theor Popul Biol*, 82(1):59–65, 2012.

Arindam RoyChoudhury, Joseph Felsenstein, and Elizabeth A Thompson. A two-stage pruning algorithm for likelihood computation for a population tree. *Genetics*, 180(2):1095–1105, 2008.

Tanja Stadler and James H Degnan. A polynomial time algorithm for calculating the probability of a ranked gene tree given a species tree. *Algorithm Mol Biol*, 7(1):1–9, 2012.

John Syring, Kathleen Farrell, Roman Businskỳ, Richard Cronn, and Aaron Liston. Widespread genealogical nonmonophyly in species of *Pinus* subgenus *Strobus*. *Syst Biol*, 56(2):163–181, 2007.

Simon Tavaré. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor Popul Biol*, 26(2):119–164, 1984.

Yufeng Wu. Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution*, 66(3):763–775, 2012.

Sha Zhu, James H Degnan, and Mike Steel. Clades, clans, and reciprocal monophyly under neutral evolutionary models. *Theor Popul Biol*, 79(4):220–227, 2011.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.16.473006; this version posted December 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Next State

Current 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
(3,1,0,1)	0	$\frac{1}{6}$	$\frac{1}{36}$	$\frac{1}{12}$	0	$\cdots$																			$\cdots$	0	$\frac{13}{18}$
(3,1,1,0)	0	0	0	0	$\frac{3}{28}$	$\frac{1}{28}$	$\frac{3}{28}$	0	$\cdots$																$\cdots$	0	$\frac{3}{4}$
(4,0,0,1)	0	$\cdots$		$\cdots$	0	$\frac{3}{14}$	0	$\frac{3}{14}$	0	$\cdots$															$\cdots$	0	$\frac{4}{7}$
(2,1,0,1)	0	$\cdots$			$\cdots$	0	$\frac{3}{14}$	$\frac{1}{28}$	$\frac{1}{28}$	0	$\cdots$														$\cdots$	0	$\frac{5}{7}$
(3,2,0,0)	0	$\cdots$						$\cdots$	0	$\frac{2}{21}$	$\frac{1}{7}$	0	$\cdots$												$\cdots$	0	$\frac{16}{21}$
(4,0,1,0)	0	$\cdots$						$\cdots$	0	$\frac{1}{7}$	0	$\frac{2}{7}$	0	$\cdots$											$\cdots$	0	$\frac{4}{7}$
(2,1,1,0)	0	$\cdots$							$\cdots$	0	$\frac{1}{7}$	$\frac{1}{21}$	$\frac{1}{21}$	0	$\cdots$										$\cdots$	0	$\frac{16}{21}$
(3,0,0,1)	0	$\cdots$								$\cdots$	0	$\frac{2}{7}$	0	$\frac{1}{7}$	0	$\cdots$									$\cdots$	0	$\frac{4}{7}$
(1,1,0,1)	0	$\cdots$									$\cdots$	0	$\frac{2}{7}$	$\frac{1}{21}$	0	$\cdots$									$\cdots$	0	$\frac{2}{3}$
(4,1,0,0)	0	$\cdots$											$\cdots$	0	$\frac{1}{15}$	$\frac{2}{5}$	0	$\cdots$							$\cdots$	0	$\frac{8}{15}$
(2,2,0,0)	0	$\cdots$												$\cdots$	0	$\frac{2}{15}$	$\frac{1}{15}$	0	$\cdots$						$\cdots$	0	$\frac{4}{5}$
(3,0,1,0)	0	$\cdots$												$\cdots$	0	$\frac{1}{5}$	0	$\frac{1}{5}$	0	$\cdots$					$\cdots$	0	$\frac{3}{5}$
(1,1,1,0)	0	$\cdots$													$\cdots$	0	$\frac{1}{5}$	$\frac{1}{15}$	0	$\cdots$					$\cdots$	0	$\frac{11}{15}$
(2,0,0,1)	0	$\cdots$														$\cdots$	0	$\frac{2}{5}$	$\frac{1}{15}$	0	$\cdots$				$\cdots$	0	$\frac{8}{15}$
(5,0,0,0)	0	$\cdots$																$\cdots$	0	1	0	$\cdots$			$\cdots$	0	0
(3,1,0,0)	0	$\cdots$																$\cdots$	0	$\frac{1}{10}$	$\frac{3}{10}$	0	$\cdots$		$\cdots$	0	$\frac{3}{5}$
(1,2,0,0)	0	$\cdots$																	$\cdots$	0	$\frac{1}{5}$	0	$\cdots$		$\cdots$	0	$\frac{4}{5}$
(2,0,1,0)	0	$\cdots$																	$\cdots$	0	$\frac{3}{10}$	$\frac{1}{10}$	0	0	0	0	$\frac{3}{5}$
(1,0,0,1)	0	$\cdots$																		$\cdots$	0	$\frac{3}{5}$	0	0	0	0	$\frac{2}{5}$
(4,0,0,0)	0	$\cdots$																			$\cdots$	0	1	0	0	0	0
(2,1,0,0)	0	$\cdots$																			$\cdots$	0	$\frac{1}{6}$	$\frac{1}{6}$	0	0	$\frac{2}{3}$
(1,0,1,0)	0	$\cdots$																				$\cdots$	0	$\frac{1}{2}$	0	0	$\frac{1}{2}$
(3,0,0,0)	0	$\cdots$																					$\cdots$	0	1	0	0
(1,1,0,0)	0	$\cdots$																					$\cdots$	0	$\frac{1}{3}$	0	$\frac{2}{3}$
(2,0,0,0)	0	$\cdots$																						$\cdots$	0	1	0
(1,0,0,0)	0	$\cdots$																						$\cdots$	0	1	0
F	0	$\cdots$																							$\cdots$	0	1

Table 2: Transition matrix for the example continuous-time Markov chain in Section A.2.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.16.473006; this version posted December 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Next State

Current 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
(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	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	−21	0	0	0	6	0	3	0	⋯									⋯	0	12
(1,1,0,1)	0	⋯					⋯	0	−21	0	0	0	6	1	0	⋯									⋯	0	14
(4,1,0,0)	0	⋯						⋯	0	−15	0	0	0	0	1	6	0	⋯							⋯	0	8
(2,2,0,0)	0	⋯							⋯	0	−15	0	0	0	0	2	1	0	⋯						⋯	0	12
(3,0,1,0)	0	⋯								⋯	0	−15	0	0	0	3	0	3	0	⋯					⋯	0	9
(1,1,1,0)	0	⋯									⋯	0	−15	0	0	0	3	1	0	⋯					⋯	0	11
(2,0,0,1)	0	⋯										⋯	0	−15	0	0	0	6	1	0	⋯				⋯	0	8
(5,0,0,0)	0	⋯											⋯	0	−10	0	0	0	0	10	0	⋯			⋯	0	0
(3,1,0,0)	0	⋯												⋯	0	−10	0	0	0	1	3	0	⋯		⋯	0	6
(1,2,0,0)	0	⋯													⋯	0	−10	0	0	0	2	0	⋯		⋯	0	8
(2,0,1,0)	0	⋯														⋯	0	−10	0	0	3	1	0	0	0	0	6
(1,0,0,1)	0	⋯															⋯	0	−10	0	0	6	0	0	0	0	4
(4,0,0,0)	0	⋯																⋯	0	−6	0	0	6	0	0	0	0
(2,1,0,0)	0	⋯																	⋯	0	−6	0	1	1	0	0	4
(1,0,1,0)	0	⋯																		⋯	0	−6	0	3	0	0	3
(3,0,0,0)	0	⋯																			⋯	0	−3	0	3	0	0
(1,1,0,0)	0	⋯																				⋯	0	−3	1	0	2
(2,0,0,0)	0	⋯																					⋯	0	−1	1	0
(1,0,0,0)	0	⋯																						⋯	0	0	0
F	0	⋯																							⋯	0	0

Table 3: Transition rate matrix (Q_x) for the example continuous-time Markov chain in Section A.2.