Theoretical Population Biology 129 (2019) 133–147

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

The probability of reciprocal monophyly of gene lineages in three and four species

Rohan S. Mehta *, Noah A. Rosenberg

Department of Biology, Stanford University, 371 Serra Mall, Stanford, CA 94305, USA

ARTICLE INFO

Article history:
Received 9 August 2017
Available online 3 May 2018

Keywords:
Coalescent
Gene trees
Monophyly
Species trees

ABSTRACT

Reciprocal monophyly, a feature of a genealogy in which multiple groups of descendant lineages each consist of all of the descendants of their respective most recent common ancestors, has been an important concept in studies of species delimitation, phylogeography, population history reconstruction, systematics, and conservation. Computations involving the probability that reciprocal monophyly is observed in a genealogy have played a key role in criteria for defining taxonomic groups and inferring divergence times. The probability of reciprocal monophyly under a coalescent model of population divergence has been studied in detail for groups of gene lineages for pairs of species. Here, we extend this computation to generate corresponding probabilities for sets of gene lineages from three and four species. We study the effects of model parameters on the probability of reciprocal monophyly, finding that it is driven primarily by species tree height, with lesser but still substantial influences of internal branch lengths and sample sizes. We also provide an example application of our results to data from maize and teosinte.

1. Introduction

A set of gene lineages is monophyletic if all of the lineages are more closely related to each other genealogically than any of them is to any other sampled lineage. Multiple sets of gene lineages in a genealogy are reciprocally monophyletic if each set of lineages is separately monophyletic.

Reciprocal monophyly is likely to occur for the lineages of a pair of populations at some point after they diverge from an ancestral population (Neigel and Avise, 1986; Avise and Ball Jr., 1990). As a result, reciprocal monophyly is often used as a criterion for evaluating the consequences of divergence processes. Criteria for conservation units and species delimitation have frequently been based on levels of reciprocal monophyly (Moritz, 1994; De Queiroz, 2007). Reciprocal monophyly is fundamental to a genealogical concept for describing species (Hudson and Coyne, 2002). It is also useful in understanding the evolutionary processes underlying group divergence, both for groups within species and for groups that represent separate species (e.g. Carstens and Richards 2007; Tavares and Baker 2008; Kubatko et al. 2011; Lohse et al. 2011; Birky 2013; Rabeling et al. 2014; Dearborn et al. 2015).

Under the multispecies coalescent model, in which gene lineages diverge along the branches of a species tree, Rosenberg (2003) computed the probability of reciprocal monophyly for sets of lineages drawn from two species. Rosenberg (2002) provided a recursive formula for the probability of reciprocal monophyly for three species in a model in which the three species descend from a common ancestor via two sequential divergence events. More recently, Zhu et al. (2011) calculated the probability of reciprocal monophyly for an arbitrary partition of sampled lineages for a single species. Eldon and Degnan (2012) obtained the probability of reciprocal monophyly for two species under the Λ-coalescent, which allows asynchronous events in which more than two lineages coalesce. In a generalization of the computation of Rosenberg (2003), Mehta et al. (2016) derived the probability of reciprocal monophyly of a bipartition of sampled lineages for an arbitrary species tree.

Previous work on reciprocal monophyly probabilities has generally been limited to two taxa or two genealogical lineage classes. Although Zhu et al. (2011) permitted arbitrary partitions of a set of lineages into many reciprocally monophyletic groups, their calculation considered lineages in a single population and did not account for species divergence. The three-species computation from Rosenberg (2002) was recursive; although it can be made non-recursive by use of a result in the appendix of Rosenberg (2003), as we will see, the computation has a case that it does not take into account. Because computations for more than two lineage classes have not been available, empirical studies interested in reciprocal monophyly for more than two groups in a genealogy that contains more than two species have often considered reciprocal monophyly of species pairs rather than simultaneous reciprocal monophyly of all groups of interest. Reciprocal monophyly is examined for many pairs of species, or species are combined together

* Corresponding author.
E-mail address: rsmehta@stanford.edu (R.S. Mehta).

https://doi.org/10.1016/j.tpb.2018.04.004
0040-5809/

Table 1
Coalescent-based monophyly probability computations. The current study extends beyond Mehta et al. (2016) in the case of three- and four-taxon species trees by permitting three and four lineage classes rather than two (although unlike in Mehta et al. (2016), we require each lineage class to be identified with a single taxon). It extends beyond Zhu et al. (2011) by considering gene lineages in a species tree model rather than gene lineages within a single population (although unlike in Zhu et al. (2011), we consider only three and four lineage classes). It extends beyond Rosenberg (2002) by correcting an error in the earlier three-lineage-class, three-species computation and by considering four taxa and four lineage classes. Note that Eldon and Degnan (2012) permit asynchronous multiple mergers of gene lineages in their population split model. Jakobsson and Rosenberg (2007) are concerned with counting the number of inter-population coalescences in which one of the coalescing lineages has descendants exclusively in a specific one of the two populations; the monophyly probability is a special case of this computation. Abbreviations: A, analytical; R, recursion; S, simulation; M, monophyly; RM, reciprocal monophyly.

Study	Number of populations	Number of lineage classes	Sample size per population	Approach	Type of computation	Population model
Tajima (1983)	2	2	2	A	M, RM	Population split
Takahata and Nei (1985)	2	2	2	A	M, RM	Population split
Neigel and Avise (1986)	2	2	Arbitrary	S	M, RM	Population split
Takahata and Slatkin (1990)	2	2	2	A	M, RM	Island migration
Wakeley (2000)	2	2	2	A	RM	Population split, island migration
Rosenberg (2002)	2	2	Arbitrary	R	RM	Population split
Rosenberg (2002)	3	3	Arbitrary	R	RM	Species tree
Hudson and Coyne (2002)	2	2	Arbitrary	R	M, RM	Population split
Rosenberg (2003)	2	2	Arbitrary	A	M, RM	Population split
Jakobsson and Rosenberg (2007)	2	2	Arbitrary	A, R	M	Population split
Zhu et al. (2011)	1	Arbitrary	Arbitrary	A	M, RM	Single population
Eldon and Degnan (2012)	2	2	Arbitrary	R	M, RM	Population split
Mehta et al. (2016)	Arbitrary	2	Arbitrary	R	M, RM	Species tree
This study	3	3	Arbitrary	A	RM	Species tree
This study	4	4	Arbitrary	A	RM	Species tree

so that only a pair of groups remains (e.g. Carstens and Richards 2007, Baker et al. 2009, Neilson and Stepien 2009, Kubatko et al. 2011, Bergsten et al. 2012).

Here, we derive the probability of reciprocal monophyly for gene lineages from sets of three or four species under the multispecies coalescent, producing separate results for the unique three-species bifurcating species tree topology and the two distinct four-species topologies. Among monophyly probability computations considering gene lineages evolving on a species tree, the derivation is novel in extending beyond two gene lineage classes to examine three and four classes (Table 1), providing a correction to the one previous three-lineage-class, three-species monophyly probability computation for gene lineages on a species tree. Our approach combines elements of the generalized monophyly computation for pairs of classes of lineages (Mehta et al., 2016) and the earlier efforts to obtain three-species monophyly probabilities (Rosenberg, 2002, 2003). We study the effects of model parameters, such as species tree height, internal branch lengths, and sample sizes, on the probability of reciprocal monophyly. We also examine the distribution of reciprocal monophyly probabilities over grids of choices for the branch lengths. Finally, we provide an example application of our results to data from maize and teosinte.

2. Preliminaries

2.1. Model and notation

We consider a bifurcating species tree $\mathscr{T}$ that consists of a topology and a set of branch lengths. A sample size greater than or equal to 1 is specified for each leaf of $\mathscr{T}$. We use the multispecies coalescent to track the sampled lineages as they travel backward in time up the species tree. In this section, we discuss the terminology and construction of our coalescent model, closely following Mehta et al. (2016).

2.2. Lineage labels

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. Lineages that result from coalescences between lineages with differing labels are called "mixed" lineages and have label M. There are $\ell + 1$ labels for a species tree with ℓ species.

2.3. Species tree branches

In our coalescent framework, time starts at 0 in the present, at the bottom of the species tree, and increases when moving up the species tree and further into the past. From this perspective, an internal node of the species tree is a species-merging event. As time progresses, gene lineages enter species tree nodes from the bottom and exit them from the top. A particular node x has lineages enter from both branches directly below the node. Because each node is associated only with the branch that corresponds to its immediate ancestral species, we refer to branches by the labels of their associated nodes. 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 and is assumed to be constant over all species tree branches. Population size changes can be accommodated by increasing or decreasing the length of a branch, as a constant size change that lasts the duration of a branch amounts to a rescaling of time along the branch. Larger population sizes correspond to smaller values of T_x in coalescent units. The root branch of $\mathscr{T}$ has infinite length.

2.4. Input and output states

An output state of a branch x is a set of integers that represents the numbers of lineages exiting the branch from the top. The random variable for this output state is the vector $\mathbf{Z}_x$ of length $\ell + 1$ whose ith element is the number of output lineages with the ith label. A particular instance of this random variable is denoted $\mathbf{n}_x^O$. An input state for a branch is a set of integers that represents 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 simply the sum of the two output states for its immediate descendant branches x_L and x_R, which are independent random variables. We therefore do not need to define a separate random variable for the input state of a branch. A particular instance of an input state is $\mathbf{n}_x^I = \mathbf{n}_{x_L}^O + \mathbf{n}_{x_R}^O$.

2.5. Coalescence sequences

A coalescence sequence is a sequence of coalescent events. For example, consider three lineages, A, B, and C, coalescing to a single

lineage. One possible coalescence sequence has A and B coalesce, resulting in the lineage AB, then has that lineage coalesce with lineage C, resulting in the lineage ABC. This sequence can be described as (A, B), ((AB), C). Another possible sequence has B coalesce with C first. This sequence can be described by (B, C), (A, (BC)). For our calculations, we must count the number of coalescence sequences that start at a particular input state and end at a particular output state while preserving reciprocal monophyly.

2.6. Reciprocal monophyly events

The reciprocal monophyly event E_x is the event that in $\mathscr{T}_x$ – the subtree of $\mathscr{T}$ that is defined by taking all parts of the species tree descended from the top of branch x – no coalescences between non-M lineages whose labels are different occur until the number of ancestral lineages of each species involved in the coalescence is 1. We aim to compute the probability of this reciprocal monophyly event when x is the root of the species tree $\mathscr{T}$. The probability of reciprocal monophyly for the whole gene genealogy is the probability $\mathbb{P}[E_{\text{root}}]$. We can decompose $\mathbb{P}[E_x]$ recursively, as $\mathbb{P}[E_x] = \mathbb{P}[E_x|E_{x_L}, E_{x_R}]\mathbb{P}[E_{x_L}]\mathbb{P}[E_{x_R}]$.

2.7. Combinatorial functions

The probability $g_{n,j}(T)$ that n lineages coalesce to j lineages in time T is given by Eq. (6.1) of Tavaré (1984). This quantity is nonzero only when $n \geq j \geq 1$ and $T \geq 0$, except that we use the convention $g_{0,0}(T) = 1$. We use this function to compute the probability that a coalescence sequence occurs that contains the correct number of coalescences to reduce a particular input state to a particular output state.

Following Eq. (4) of Rosenberg (2003), $I_{n,k} = [n!(n-1)!]/[2^{n-k}k!(k-1)!]$ is the number of coalescence sequences that reduce n distinct lineages to k lineages. This function is only nonzero when $n \geq k \geq 1$, with $I_{0,0} = 1$ by convention. It is used to count the number of coalescence sequences that lead to particular outcomes.

The multinomial coefficient $W_k(r_1, \ldots, r_k) = \binom{r_1+\cdots+r_k}{r_1,\ldots,r_k}$, which extends a trinomial expression from Rosenberg (2006), is the number of ways that k separate coalescence sequences that do not share any lineages and that consist of $r_1, \ldots, r_k$ coalescences can be ordered in a larger sequence containing all of them as subsequences. $W_k(r_1, \ldots, r_k)$ is defined when $r_i \geq 0$ for each $i = 1, \ldots, k$. This function is used in counting the number of coalescence sequences that produce the same output state from a specific input state.

2.8. The general calculation

Consider a branch x. Using the law of total probability, the joint probability of a particular output state of branch x that satisfies the reciprocal monophyly event E_x given the parameters of the species tree $\mathscr{T}$ can be written schematically as

$$\begin{aligned}&\mathbb{P}[\text{outputs of } x, E_x|\mathscr{T}_x]\\ &= \sum_{\substack{\text{possible}\\ \text{inputs of } x}} \mathbb{P}[\text{outputs of } x_L, E_{x_L}|\mathscr{T}_{x_L}]\,\mathbb{P}[\text{outputs of } x_R, E_{x_R}|\mathscr{T}_{x_R}]\\ &\quad\times \mathbb{P}[\text{outputs of } x, E_x|\text{inputs of } x, E_{x_L}, E_{x_R}, \mathscr{T}_x],\end{aligned} \tag{1}$$

and mathematically as

$$\begin{aligned}\mathbb{P}[\mathbf{Z}_x = \mathbf{n}_x^O, E_x|\mathscr{T}_x] &= \sum_{\mathbf{n}_x^I} \mathbb{P}[\mathbf{Z}_{x_L} = \mathbf{n}_{x_L}^O, E_{x_L}|\mathscr{T}_{x_L}]\mathbb{P}[\mathbf{Z}_{x_R} = \mathbf{n}_{x_R}^O, E_{x_R}|\mathscr{T}_{x_R}]\\ &\quad\times\ \mathbb{P}[\mathbf{Z}_x = \mathbf{n}_x^O, E_x|\mathbf{Z}_{x_L} = \mathbf{n}_{x_L}^O, \mathbf{Z}_{x_R} = \mathbf{n}_{x_R}^O, E_{x_L}, E_{x_R}, \mathscr{T}_x].\end{aligned} \tag{2}$$

Eq. (2) is Eq. (2) of Mehta et al. (2016) with a few notational differences.

The joint probability of the output state $\mathbf{n}_x^O$ and the preservation of reciprocal monophyly E_x given the input state $\mathbf{n}_x^I$, or $\mathbb{P}[\mathbf{Z}_x = \mathbf{n}_x^O, E_x|\mathbf{Z}_{x_L} = \mathbf{n}_{x_L}^O, \mathbf{Z}_{x_R} = \mathbf{n}_{x_R}^O, E_{x_L}, E_{x_R}, \mathscr{T}_x]$, can be considered with two events, one conditional on the other. First, the correct number of coalescences, equal to the total number of inputs $|\mathbf{n}_x^I|$ minus the total number of outputs $|\mathbf{n}_x^O|$, must occur. The probability that $|\mathbf{n}_x^O|$ lineages remain from among $|\mathbf{n}_x^I|$ initial lineages is $g_{|\mathbf{n}_x^I|,|\mathbf{n}_x^O|}(T_x)$. Second, given that the correct number of coalescences occurs, the coalescences must yield the correct outputs $\mathbf{n}_x^O$ while also preserving reciprocal monophyly E_x. We represent the probability of this second event by K_x, which is a combinatorial function of the numbers of input and output lineages of the various types. K_x takes the form of a fraction, with denominator $I_{|\mathbf{n}_x^I|,|\mathbf{n}_x^O|}$ – the total number of coalescence sequences – and numerator equal to the total number of possible coalescence sequences that preserve reciprocal monophyly and yield the desired output state. The form of the numerator depends on the specific computation of interest. We derive the forms of K_x in later sections.

Formally, the third multiplicative term in Eq. (2) becomes

$$\begin{aligned}&\mathbb{P}[\mathbf{Z}_x = \mathbf{n}_x^O, E_x|\mathbf{Z}_{x_L} = \mathbf{n}_{x_L}^O, \mathbf{Z}_{x_R} = \mathbf{n}_{x_R}^O, E_{x_L}, E_{x_R}, \mathscr{T}_x]\\ &\quad= g_{|\mathbf{n}_x^I|,|\mathbf{n}_x^O|}(T_x)\,K_x(\mathbf{n}_x^I, \mathbf{n}_x^O).\end{aligned} \tag{3}$$

This quantity corresponds to Eq. (3) from Mehta et al. (2016). K_x is in general a function of input lineage counts and output lineage counts.

The probability of reciprocal monophyly of all ℓ unmixed lineage classes can be obtained by

$$\mathbb{P}[E_{\text{root}}] = \mathbb{P}[\mathbf{Z}_{\text{root}} = (0, \ldots, 0, 1), E_{\text{root}}|\mathscr{T}], \tag{4}$$

as the only possible output of the root is a single M lineage.

The base case of the recursion (Eq. (2)) occurs at the leaves: reciprocal monophyly holds trivially at leaves because no coalescence has occurred. Without loss of generality, we adopt the convention that all leaf inputs come from the left, so that if x is a leaf, then $\mathbb{P}[\mathbf{Z}_{x_L} = \mathbf{n}_x^I, E_{x_L}|\mathscr{T}_{x_L}] = 1$ and $\mathbb{P}[\mathbf{Z}_{x_R} = \mathbf{0}, E_{x_R}|\mathscr{T}_{x_R}] = 1$. Also, every possible coalescence sequence within a leaf branch preserves reciprocal monophyly, as no interspecies coalescence can occur. Consequently, if x is a leaf, then $K_x = 1$. Applying Eq. (2) to a leaf yields $\mathbb{P}[\mathbf{Z}_x = \mathbf{n}_x^O, E_x|\mathscr{T}_x] = g_{|\mathbf{n}_x^I|,|\mathbf{n}_x^O|}(T_x)$.

3. Probability of reciprocal monophyly for lineages in a three-species tree

In this section we derive the formula for the probability of reciprocal monophyly for gene lineages in a three-species tree. Fig. 1A presents a three-species tree with internal nodes labeled and sample sizes, branch lengths, and branch outputs specified. We label the three extant species and their corresponding leaves and lineages A, B, and C (in red, green, and blue, respectively, in Fig. 1A). We indicate the single internal branch by I.

The branch outputs for the three-species model in Fig. 1A are

$$\begin{aligned}\mathbf{Z}_A &= (s, 0, 0, 0)\\ \mathbf{Z}_B &= (0, t, 0, 0)\\ \mathbf{Z}_C &= (0, 0, y, 0)\\ \mathbf{Z}_I &= (w, x, 0, m).\end{aligned}$$

Initially, we have p lineages from species A, which coalesce to s lineages during time T_3. Similarly, q initial lineages from species B coalesce to t lineages during time T_3, and r initial lineages from

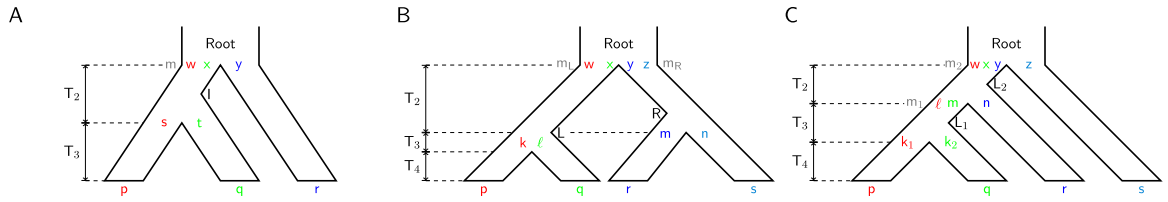

Fig. 1. Schematic diagrams of three- and four-species models. Species A, B, C, and D appear in red, green, blue, and cyan, respectively. (A) Three-species tree. We start with p, q, and r lineages from species A, B, and C, respectively. The m mixed lineages are the result of interspecies coalescences in the internal branch I. (B) Balanced four-species tree. We start with p, q, r, and s lineages from species A, B, C, and D, respectively. The m_L and m_R mixed lineages are the result of interspecies coalescences in the L and R branches, respectively. (C) Caterpillar four-species tree. We start with p, q, r, and s lineages from species A, B, C, and D, respectively. The m_1 and m_2 mixed lineages are the result of interspecies coalescences in the L_1 and L_2 branches, respectively.

species C coalesce to y lineages during time T_3+T_2. Within internal branch I, s lineages from species A and t lineages from species B coalesce to w lineages from species A, x lineages from species B, and m mixed lineages during time T_2, with m possibly equal to 0. Finally, in the root branch of the species tree, w lineages from species A, x lineages from species B, y lineages from species C, and m mixed lineages coalesce to a single mixed lineage.

3.1. Application of the recursion

Applying Eq. (2) to the root of the species tree in Fig. 1A yields

$$\begin{aligned}\mathbb{P}[\mathbf{Z}_{\text{root}} &= (0,0,0,1), E_{\text{root}}|\mathscr{T}] \\ &= \sum_{w=0}^{s}\sum_{x=0}^{t}\sum_{y=1}^{r}\sum_{m=0}^{1}\mathbb{P}[\mathbf{Z}_I=(w,x,0,m),E_I|\mathscr{T}_I] \\ &\quad\times\ \mathbb{P}[\mathbf{Z}_C=(0,0,y,0),E_C|\mathscr{T}_C]g_{w+x+y+m,1}(T_{\text{root}}) \\ &\quad\times\ K_{\text{root}}((w,x,y,m),(0,0,0,1)).\end{aligned}$$

Applying Eq. (2) to the internal node I, we have

$$\begin{aligned}\mathbb{P}[\mathbf{Z}_I =&(w,x,0,m),E_I|\mathscr{T}_I] = \sum_{s=1}^{p}\sum_{t=1}^{q}\mathbb{P}[\mathbf{Z}_A=(s,0,0,0),E_A|\mathscr{T}_A] \\ &\times\ \mathbb{P}[\mathbf{Z}_B=(0,t,0,0),E_B|\mathscr{T}_B]g_{s+t,w+x+m}(T_2) \\ &\times\ K_I((s,t,0,0),(w,x,0,m)).\end{aligned}$$

Finally, applying the recursion to the three leaf nodes yields

$$\begin{aligned}\mathbb{P}[\mathbf{Z}_A=(s,0,0,0),E_A|\mathscr{T}_A] &= g_{p,s}(T_3)\\ \mathbb{P}[\mathbf{Z}_B=(0,t,0,0),E_B|\mathscr{T}_B] &= g_{q,t}(T_3)\\ \mathbb{P}[\mathbf{Z}_C=(0,0,y,0),E_C|\mathscr{T}_C] &= g_{r,y}(T_3+T_2).\end{aligned}$$

Combining all these results and noting that T_{root} is infinite and $\lim_{T\to\infty}g_{w+x+y+m,1}(T) = 1$, the full result for the probability of reciprocal monophyly is

$$\begin{aligned}\mathbb{P}(E) = &\sum_{s=1}^{p}\sum_{t=1}^{q}\sum_{m=0}^{1}\sum_{w=0}^{s}\sum_{x=0}^{t}\sum_{y=1}^{r}g_{p,s}(T_3)g_{q,t}(T_3) \\ &\times\ g_{r,y}(T_2+T_3)g_{s+t,w+x+m}(T_2)K_I(s,t,w,x,m) \\ &\times\ K_{\text{root}}(w,x,y,m),\end{aligned} \tag{5}$$

where we have omitted the full notational form of the K_x terms and replaced them with functions of only the possibly-nonzero variables.

3.2. Combinatorial terms

We now obtain the forms of the K_x terms. First, we find the denominator, the total number of possible coalescence sequences of the correct length. Next, we find the numerator, the number of coalescence sequences that both have the correct length and preserve reciprocal monophyly. For computations of both K_I and K_{root}, we count coalescence sequences using $I_{n,k}$ and $W_k(r_1,\ldots,r_k)$.

3.2.1. K_I

For the internal branch, the value of K_I is

$$K_I(s,t,w,x,m) = \begin{cases}\dfrac{I_{s,w}I_{t,x}W_2(s-w,t-x)}{I_{s+t,w+x}} & \text{Case 1: } s,t,w,x\geq 1, m=0\\[2ex] \dfrac{I_{s,1}I_{t,1}W_2(s-1,t-1)}{I_{s+t,1}} & \text{Case 2: } s,t\geq 1,\\ & w=x=0, m=1\\ 0 & \text{otherwise.}\end{cases} \tag{6}$$

K_I (Eq. (6)) has two possible cases in which reciprocal monophyly can be preserved. Case 1 applies when no interspecies coalescence occurs in the internal branch. The total number of possible coalescence sequences in this branch is $I_{s+t,w+x}$. For a coalescence sequence to preserve reciprocal monophyly and yield the desired output, s lineages from species A coalesce to w lineages (counted by $I_{s,w}$) and t lineages from species B coalesce to x lineages (counted by $I_{t,x}$). The number of ways these two separate sequences can be ordered is $W_2(s-w,t-x)$. The total number of possible coalescence sequences that preserve reciprocal monophyly and yield the desired output is therefore $I_{s,w}I_{t,x}W_2(s-w,t-x)$.

Case 2 applies when an interspecies coalescence occurs in the internal branch. The total number of possible coalescence sequences in this branch is $I_{s+t,1}$. For a coalescence sequence to preserve reciprocal monophyly and yield the desired output, all lineages from each species must separately coalesce to a single lineage and then the resulting two lineages must coalesce, all in the internal branch: s lineages coalesce to 1 lineage (counted by $I_{s,1}$) and t lineages coalesce to 1 lineage (counted by $I_{t,1}$). The number of ways these two separate sequences can be ordered is $W_2(s-1,t-1)$. There is only one way for the interspecies coalescence to occur. The total number of possible coalescence sequences that preserve reciprocal monophyly and yield the desired output is therefore $I_{s,1}I_{t,1}W_2(s-1,t-1)$.

3.2.2. K_{root}

The root can have lineages from all three species initially present. In order to compute the value of K_{root}, we need to compute the probability that when lineages from all three species enter the root and coalesce to a single mixed lineage, they do so without violating reciprocal monophyly. We call this probability F.

To compute F, we first consider the coalescences of the single lineages of each species ancestral to all lineages of that species, and we then consider the coalescence of the lineages of each species to single lineages. This line of reasoning is similar to a technique used by Zhu et al. (2011).

Three lineages, one from each species, result from complete within-species coalescence. These three lineages can coalesce in $I_{3,1} = 3$ ways. Under reciprocal monophyly, these are the only possible interspecies coalescences. We consider each of these three

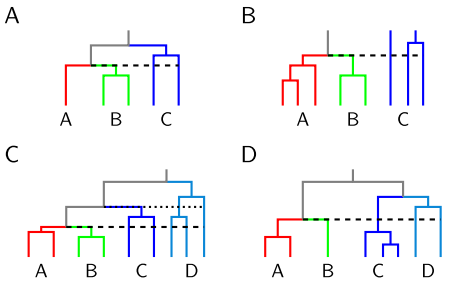

Fig. 2. Example coalescence sequences counted by combinatorial functions. (A) Lineages from three species coalesce to a single mixed lineage; $F(1, 2, 2)$ (Eq. (8), specifically $f(1, 2, 2)$, Eq. (7)). (B) Lineages from three species coalesce to lineages from one species and a mixed lineage; $G(3, 2, 3, 2)$ (Eq. (19)). (C) Lineages from four species coalesce to a single mixed lineage with a caterpillar topology; $H(2, 2, 2, 3)$ (Eq. (15), specifically $h_1(2, 2, 2, 3)$, Eq. (13)) (D) Lineages from four species coalesce to a single mixed lineage with a balanced topology; $H(2, 1, 3, 2)$ (Eq. (15), specifically $h_2(2, 1, 3, 2)$, Eq. (14)). Species A appears in red, species B in green, species C in blue, and species D in cyan. For (A), all lineages from three species coalesce to a single lineage while preserving reciprocal monophyly of all species, with the first interspecies coalescence happening between species A and species B. The dashed line indicates the time of the (A, B) coalescence, at which we count $c = 2$ lineages from species C. In (B), only one interspecies coalescence occurs, between species A and B. The dashed line indicates the time of the (A, B) coalescence, at which $c = 3$ lineages are present from species C. For (C), all lineages from four species coalesce to a single lineage while preserving reciprocal monophyly of all species, with the first interspecies coalescence happening between species A and species B, and the second between species C and the (A, B) lineage resulting from the first interspecies coalescence. The dashed line indicates the time of the (A, B) coalescence, at which we count $c_1 = 2$ lineages from species C and $c_2 = 3$ lineages from species D. The dotted line indicates the time of the ((AB), C) coalescence, at which we count $c_3 = 2$ lineages from species D. For (D), all lineages from four species coalesce to a single lineage while preserving reciprocal monophyly of all species, with the most recent interspecies coalescence happening between species A and species B and the second coalescence between species C and species D. The dashed line indicates the time of the (A, B) coalescence, at which we count $c_1 = 1$ lineage from species C and $c_2 = 2$ lineages from species D.

ways separately. Suppose we begin with w lineages from species A, x lineages from species B, and y lineages from species C.

First, we assume that the two interspecies coalescences follow the sequence (A, B), ((AB), C) (Fig. 2A). There is only one way for the (A, B) coalescence to occur. Until this coalescence happens, lineages from species C cannot coalesce with lineages from either of the other two species. We let c be the number of lineages from species C present when the (A, B) coalescence occurs. The dashed line in Fig. 2A indicates this time point. In Fig. 2A, $c = 2$. The number of ways to reach the (A, B) coalescence is $I_{w,1}I_{x,1}I_{y,c}W_3(w-1, x-1, y-c)$. The (A, B) coalescence occurs (in only one possible way), then the lineages from species C coalesce completely (in $I_{c,1}$ possible ways), and, finally, the ((AB), C) coalescence occurs (in only one possible way). The number of coalescence sequences reducing the lineages from three species to a single lineage while preserving reciprocal monophyly is $I_{w,1}I_{x,1}I_{y,c}W_3(w-1, x-1, y-c)I_{c,1}$. The total number of coalescence sequences reducing $w+x+y$ lineages to a single lineage is $I_{w+x+y,1}$.

Using these results, we construct $f(w, x, y)$, the probability that w lineages from one species, x lineages from a second species, and y lineages from a third species coalesce to a single lineage while preserving reciprocal monophyly, with the first interspecies coalescence taking place between the first two species:

$$f(w, x, y) = \sum_{c=1}^{y} \frac{I_{w,1}I_{x,1}I_{y,c}W_3(w-1, x-1, y-c)I_{c,1}}{I_{w+x+y,1}}. \tag{7}$$

The first two arguments in this function are exchangeable: $f(w, x, y) = f(x, w, y)$. The three nonequivalent assignments of species to arguments of the function f – $f(w, x, y)$, $f(w, y, x)$, and $f(x, y, w)$ – yield the three different orders in which the two interspecies coalescences can occur. Thus, to obtain $F(w, x, y)$, the probability that w lineages from species A, x lineages from species B, and y lineages from species C coalesce to a single lineage while preserving reciprocal monophyly, we sum over these distinct assignments:

$$F(w, x, y) = f(w, x, y) + f(w, y, x) + f(x, y, w). \tag{8}$$

Fig. 2A illustrates a specific coalescence sequence that would be counted by Eq. (8), and in particular by Eq. (7), because the first interspecific coalescence joins the first two species. We note that Eqs. (7) and (8) enable corrections to two results of Rosenberg (2002) that omitted some scenarios (see Appendix).

For the root branch, the value of K_{root} is

$$K_{\text{root}}(w, x, y, m) = \begin{cases} F(w, x, y) & \text{Case 1: } w, x, y \geq 1, m = 0 \\ \dfrac{I_{y,1}}{I_{y+1,1}} & \text{Case 2: } w = x = 0, y \geq 1, m = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Like K_I (Eq. (6)), K_{root} also has two possible cases that preserve reciprocal monophyly (Eq. (9)). Case 1 applies when no interspecies coalescence has occurred below the root branch. Lineages from three species enter the root branch, so we use F (Eq. (8)).

Case 2 applies when an interspecies coalescence has occurred in the internal branch of the species tree. In this case, a single mixed lineage and y lineages from species C are initially present at the root branch. The total number of possible coalescence sequences that reduce $y+1$ lineages to a single lineage is $I_{y+1,1}$. For reciprocal monophyly to occur, the lineages from species C must coalesce to a single lineage (counted by $I_{y,1}$) before the interspecies coalescence. There is only one way for the interspecies coalescence to occur. The total number of coalescence sequences that preserve reciprocal monophyly is therefore $I_{y,1}$.

4. Probability of reciprocal monophyly of lineages in a four-species tree with a balanced topology

In this section, we derive the formula for the probability of reciprocal monophyly for gene lineages in a four-species tree with a balanced topology. Fig. 1B presents a four-species balanced tree topology with internal nodes labeled, and sample sizes, branch lengths, and branch outputs specified. We label the four extant species and their corresponding leaves and lineages A, B, C, and D (in red, green, blue, and cyan, respectively, in Fig. 1B). We indicate the two internal branches by L and R, corresponding to their positions in Fig. 1B. The branch outputs in Fig. 1B are

$$\mathbf{Z}_A = (k, 0, 0, 0, 0)$$
$$\mathbf{Z}_B = (0, \ell, 0, 0, 0)$$
$$\mathbf{Z}_C = (0, 0, m, 0, 0)$$
$$\mathbf{Z}_D = (0, 0, 0, n, 0)$$
$$\mathbf{Z}_L = (w, x, 0, 0, m_L)$$
$$\mathbf{Z}_R = (0, 0, y, z, m_R).$$

We have p lineages from species A, which coalesce to k lineages during time T_4. Similarly, q lineages from species B coalesce to ℓ lineages during time T_4, r lineages from species C coalesce to m lineages during time T_4+T_3, and s lineages from species D coalesce to n lineages during time $T_4 + T_3$. Within branch L, k lineages from species A and ℓ lineages from species B coalesce to w lineages from species A, x lineages from species B, and m_L mixed lineages during time $T_3 + T_2$. Within branch R, m lineages from species C and n lineages from species D coalesce to y lineages from species C, z lineages from species D, and m_R mixed lineages during time T_2. Finally, in the root branch, w lineages from species A, x lineages from species B, y lineages from species C, z lineages from species D, and $m_L + m_R$ mixed lineages coalesce to a single mixed lineage.

$$h_1(w,x,y,z) = \sum_{c_1=1}^{y}\sum_{c_2=1}^{z}\sum_{c_3=1}^{c_2} \frac{I_{w,1}I_{x,1}I_{y,c_1}I_{z,c_2}W_4(w-1,x-1,y-c_1,z-c_2)I_{c_1,1}I_{c_2,c_3}W_2(c_1-1,c_2-c_3)I_{c_3,1}}{I_{w+x+y+z,1}}. \quad (13)$$

Box I.

4.1. Application of the recursion

Compiling these reductions in the numbers of lineages along the various species tree branches and applying Eq. (2) to the root of the species tree in Fig. 1B in three steps, one for each internal node, yields

$$\begin{aligned}\mathbb{P}(E) = &\sum_{k=1}^{p}\sum_{\ell=1}^{q}\sum_{m=1}^{r}\sum_{n=1}^{s}\sum_{m_L=0}^{1}\sum_{m_R=0}^{1}\sum_{w=0}^{k}\sum_{x=0}^{\ell}\sum_{y=0}^{m}\sum_{z=0}^{n} g_{p,k}(T_4)\\ &\times g_{q,\ell}(T_4)g_{r,m}(T_4+T_3)\\ &\times g_{s,n}(T_4+T_3)g_{k+\ell,w+x+m_L}(T_3+T_2)g_{m+n,y+z+m_R}(T_2)\\ &\times K_L(k,\ell,w,x,m_L)K_R(m,n,y,z,m_R)K_{\text{root}}(w,x,y,z,m_L,m_R).\end{aligned} \quad (10)$$

4.2. Combinatorial terms

We now derive the values of K_x for nodes L and R and for the root.

4.2.1. K_L and K_R

The internal nodes L and R both have lineages from two species and no mixed lineages as inputs. Thus, K_L and K_R are exactly analogous to K_I in the three-species case, with (k,ℓ) or (m,n) in the role of (s,t), (w,x) or (y,z) in the role of (w,x), and m_L or m_R in the role of m:

$$K_L(k,\ell,w,x,m_L) = \begin{cases} \dfrac{I_{k,w}I_{\ell,x}W_2(k-w,\ell-x)}{I_{k+\ell,w+x}} & \text{Case 1: } k,\ell,w,x\geq 1,\ m_L=0 \\ \dfrac{I_{k,1}I_{\ell,1}W_2(k-1,\ell-1)}{I_{k+\ell,1}} & \text{Case 2: } k,\ell\geq 1,\ w=x=0, m_L=1 \\ 0 & \text{otherwise,}\end{cases} \quad (11)$$

$$K_R(m,n,y,z,m_R) = \begin{cases} \dfrac{I_{m,y}I_{n,z}W_2(m-y,n-z)}{I_{m+n,y+z}} & \text{Case 1: } m,n,y,z\geq 1,\ m_R=0 \\ \dfrac{I_{m,1}I_{n,1}W_2(m-1,n-1)}{I_{m+n,1}} & \text{Case 2: } m,n\geq 1,\ y=z=0, m_R=1 \\ 0 & \text{otherwise.}\end{cases} \quad (12)$$

4.2.2. K_{root}

The root can have lineages from all four species initially present. In order to compute the value of K_{root}, we need to compute the probability that when lineages from all four species enter the root and coalesce to a single mixed lineage, they do so without violating reciprocal monophyly. We call this probability H. This function is the four-species analogue of F (Eq. (8)).

To derive H, we apply the same strategy used to derive F (Eq. (8)): we consider in the numerator all the ways the single last remaining lineages of each species coalesce with each other. For four taxa, these lineages can coalesce in $I_{4,1} = 18$ different ways, and hence there are 18 terms to sum. These terms involve permutations of the arguments of two distinct functions, in the same way that the three terms in Eq. (8) involve permutations of the arguments of Eq. (7). The two functions, h_1 and h_2, correspond to balanced and caterpillar topologies, respectively. In the three-species case, there is only the caterpillar topology, and there is only one function, f (Eq. (7)). For four species, the caterpillar topology yields 12 of the 18 terms, and the balanced topology yields the remaining 6.

Suppose w, x, y, and z lineages from species A, B, C, and D, respectively, enter the root node. We start with the caterpillar sequence for the last three coalescences: (A, B), ((AB), C), and (((AB)C), D) (Fig. 2C). In this situation, the lineage from species A and the lineage from species B coalesce first. Let c_1 and c_2 be the numbers of lineages from species C and species D, respectively, that are present when the (A, B) coalescence occurs. The dashed line in Fig. 2C indicates this point in the coalescence sequence; $c_1 = 2$ and $c_2 = 3$. The number of ways to reach this point in the coalescence sequence is $I_{w,1}I_{x,1}I_{y,c_1}I_{z,c_2}W_4(w-1,x-1,y-c_1,z-c_2)$. Then the (A, B) coalescence occurs. Let c_3 be the number of lineages from species D that are present just before the ((AB), C) coalescence. The dotted line in Fig. 2C indicates this point in the coalescence sequence, with $c_3 = 2$. The number of ways to get from the (A, B) coalescence to this point is $I_{c_1,1}I_{c_2,c_3}W_2(c_1-1,c_2-c_3)$. Then the ((AB), C) coalescence occurs. Finally, the c_3 lineages from species D must coalesce to a single lineage (with $I_{c_3,1}$ possible ways) and then the (((AB), C), D) coalescence must occur. For a specific set of c_i, the number of coalescence sequences that satisfy reciprocal monophyly is $I_{w,1}I_{x,1}I_{y,c_1}I_{z,c_2}W_4(w-1,x-1,y-c_1,z-c_2)I_{c_1,1}I_{c_2,c_3}W_2(c_1-1,c_2-c_3)I_{c_3,1}$.

The number of coalescence sequences that reduce $w+x+y+z$ lineages to a single lineage is $I_{w+x+y+z,1}$. For the probability that the w, x, y, and z lineages from species A, B, C, and D, respectively, are separately monophyletic with the sequence (A, B), ((AB), C), (((AB)C), D) for the coalescences of the final lineages, we have Eq. (13) given in Box I.

As was true for f (Eq. (7)), the first two arguments of h_1 are exchangeable. The twelve nonequivalent permutations of the arguments yield all the sequences by which the final lineages from the four species coalesce in a caterpillar topology. Fig. 2C illustrates a specific coalescence sequence that would be counted by Eq. (13).

The 6 remaining terms come from sequences of the last three coalescences that lead to a balanced gene tree topology, such as (A, B), (C, D), ((AB), (CD)) (Fig. 2D). Let c_1 and c_2 be the numbers of lineages from species C and species D present at the (A, B) coalescence. The dashed line in Fig. 2D indicates this point in the coalescence sequence; $c_1 = 1$ and $c_2 = 2$. The number of ways to reach this point is $I_{w,1}I_{x,1}I_{y,c_1}I_{z,c_2}W_4(w-1,x-1,y-c_1,z-c_2)$. The number of ways to get from the (A, B) coalescence to just before the (C, D) coalescence is $I_{c_1,1}I_{c_2,1}W_2(c_1-1,c_2-1)$. Then the two final coalescences ((C, D) and ((AB), (CD))) happen, each with only one way of occurring. The number of coalescence sequences that satisfy reciprocal monophyly is: $I_{w,1}I_{x,1}I_{y,c_1}I_{z,c_2}W_4(w-1,x-1,y-c_1,z-c_2)I_{c_1,1}I_{c_2,1}W_2(c_1-1,c_2-1)$. The number of coalescence sequences that reduce $w+x+y+z$ lineages to a single lineage is $I_{w+x+y+z,1}$. For the probability that the w, x, y, and z lineages from species A,

$$h_2(w,x,y,z)=\sum_{c_1=1}^{y}\sum_{c_2=1}^{z}\frac{I_{w,1}I_{x,1}I_{y,c_1}I_{z,c_2}W_4(w-1,x-1,y-c_1,z-c_2)I_{c_1,1}I_{c_2,1}W_2(c_1-1,c_2-1)}{I_{w+x+y+z,1}}. \tag{14}$$

Box II.

B, C, and D, respectively, are separately monophyletic and coalesce with sequence (A, B), (C, D), ((AB), (CD)) for the final lineages, we have Eq. (14) given in Box II.

In this case, the first and second arguments of h_2 are exchangeable, as are the third and fourth arguments. The 6 nonequivalent permutations of the arguments yield all the situations in which the final lineages from the four species coalesce in a balanced topology. Fig. 2D illustrates a specific coalescence sequence counted by Eq. (14).

All 18 of the scenarios are mutually exclusive because each corresponds to a unique ordering of coalescences of final lineages. To count the total fraction of coalescence sequences that preserve reciprocal monophyly when four species enter a node and coalesce to a single lineage, we sum the 18 terms:

$$\begin{aligned}H(w,x,y,z)&=h_1(w,x,y,z)+h_1(w,x,z,y)\\&\quad+h_1(w,y,x,z)+h_1(w,y,z,x)+h_1(w,z,x,y)\\&\quad+h_1(w,z,y,x)+h_1(x,y,w,z)+h_1(x,y,z,w)\\&\quad+h_1(x,z,w,y)+h_1(x,z,y,w)\\&\quad+h_1(y,z,w,x)+h_1(y,z,x,w)+h_2(w,x,y,z)\\&\quad+h_2(w,y,x,z)+h_2(w,z,x,y)\\&\quad+h_2(x,y,w,z)+h_2(x,z,w,y)+h_2(y,z,w,x).\end{aligned} \tag{15}$$

For the root node, the value of K_{root} is

$$K_{\text{root}}(w,x,y,z,m_L,m_R)=\begin{cases}H(w,x,y,z) & \text{Case 1: } w,x,y,z\geq 1,\ m_L=m_R=0\\ F(w,x,1) & \text{Case 2: } w,x\geq 1, y=z=m_L=0,\ m_R=1\\ F(y,z,1) & \text{Case 3: } y,z\geq 1, w=x=m_R=0,\ m_L=1\\ 1 & \text{Case 4: } w=x=y=z=0,\ m_L=m_R=1\\ 0 & \text{otherwise.}\end{cases} \tag{16}$$

The root node (Eq. (16)) has four possible cases. Case 1 applies when lineages from all four species are present, so we use H (Eq. (15)).

Cases 2 and 3 apply when two species are present in the root along with a mixed lineage that arises from an interspecies coalescence in an internal branch. The probability of reciprocal monophyly in this scenario is equal to that of a scenario in which lineages from three species are present but one species has already coalesced to a single lineage prior to entering the branch. Thus, in these cases K_{root} takes the value of the function F (Eq. (8)) with two inputs possibly greater than 1 and the third input equal to 1.

Case 4 applies when two mixed lineages and no other lineages are present. Before reaching the root, two interspecies coalescences have already occurred, one in each internal branch, so reciprocal monophyly is already guaranteed at the root as long as it was preserved in the internal branches. Thus, the only possible coalescence will result in reciprocal monophyly, and $K_{\text{root}}=1$.

5. Probability of reciprocal monophyly of lineages in a four-species tree with a caterpillar topology

In this section, we derive the formula for the probability of reciprocal monophyly for gene lineages in a four-species tree with a caterpillar topology. Fig. 1C presents a four-species caterpillar tree topology with internal nodes labeled, and sample sizes, branch lengths, and branch outputs specified. We label the four extant species and their corresponding leaves and lineages A, B, C, and D (in red, green, blue, and cyan, respectively, in Fig. 1C). We indicate the two internal branches by L_1 and L_2, numbered from bottom to top. The branch outputs in Fig. 1C are

$$\begin{aligned}\mathbf{Z}_A&=(k_1,0,0,0,0)\\ \mathbf{Z}_B&=(0,k_2,0,0,0)\\ \mathbf{Z}_C&=(0,0,n,0,0)\\ \mathbf{Z}_D&=(0,0,0,z,0)\\ \mathbf{Z}_{L_1}&=(\ell,m,0,0,m_1)\\ \mathbf{Z}_{L_2}&=(w,x,y,0,m_2).\end{aligned}$$

We have p lineages from species A coalescing to k_1 lineages during time T_4, q lineages from species B coalescing to k_2 lineages during time T_4, r lineages from species C coalescing to n lineages during time T_4+T_3, and s lineages from species D coalescing to z lineages during time $T_4+T_3+T_2$. Within branch L_1, k_1 lineages from species A and k_2 lineages from species B coalesce to ℓ lineages from species A, m lineages from species B, and m_1 mixed lineages during time T_3. Within branch L_2, ℓ lineages from species A, m lineages from species B, n lineages from species C, and m_1 mixed lineages coalesce to w lineages from species A, x lineages from species B, y lineages from species C, and m_2 mixed lineages during time T_2. Finally, in the root branch of the species tree, w lineages from species A, x lineages from species B, y lineages from species C, z lineages from species D, and m_2 mixed lineages coalesce to a single mixed lineage.

5.1. Application of the recursion

Compiling these reductions in the numbers of lineages along the various species tree branches and applying Eq. (2) to the root of the species tree in Fig. 1C in three steps, one for each internal node, yields

$$\begin{aligned}\mathbb{P}(E)&=\sum_{k_1=1}^{p}\sum_{k_2=1}^{q}\sum_{n=1}^{r}\sum_{z=1}^{s}\sum_{m_1=0}^{1}\sum_{m_2=0}^{1}\sum_{\ell=0}^{k_1}\sum_{m=0}^{k_2}\sum_{w=0}^{\ell}\sum_{x=0}^{m}\sum_{y=0}^{n}g_{p,k_1}(T_4)\\&\quad\times g_{q,k_2}(T_4)g_{r,n}(T_4+T_3)\\&\quad\times g_{s,z}(T_4+T_3+T_2)g_{k_1+k_2,\ell+m+m_1}(T_3)g_{\ell+m+n+m_1,w+x+y+m_2}(T_2)\\&\quad\times K_{L_1}(k_1,k_2,\ell,m,m_1)K_{L_2}(\ell,m,n,m_1,w,x,y,m_2)\\&\quad\times K_{\text{root}}(w,x,y,z,m_2).\end{aligned} \tag{17}$$

5.2. Combinatorial terms

We now derive the values of K_x for nodes L_1, L_2, and the root.

5.2.1. K_{L_1}

The internal node L_1 has lineages from two species and no mixed lineages as inputs. Thus, K_{L_1} is exactly analogous to K_I in the three-species case (Eq. (6)), with (k_1, k_2) in the role of (s, t), (ℓ, m) in the role of (w, x), and m_1 in the role of m:

$$K_{L_1}(k_1, k_2, \ell, m, m_1) = \begin{cases} \dfrac{I_{k_1,\ell} I_{k_2,m} W_2(k_1-\ell, k_2-m)}{I_{k_1+k_2, l+m}} & \text{Case 1: } k_1, k_2, \ell, \\ & m \geq 1, m_1 = 0 \\ \dfrac{I_{k_1,1} I_{k_2,1} W_2(k_1-1, k_2-1)}{I_{k_1+k_2,1}} & \text{Case 2: } k_1, k_2 \geq 1, \\ & \ell = m = 0, m_1 = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

5.2.2. K_{L_2}

Node L_2 can have lineages from three species enter but it need not have all lineages coalesce to a single lineage. Non-M lineages ancestral to zero, one, or three species can remain as outputs. If non-M lineages ancestral to three species remain, then the calculation is straightforward. If no non-M lineage remains, then we use F (Eq. (8)). The case in which non-M lineages ancestral to a single species remain requires another new function. In order to compute the value of K_{L_2}, we need to compute the probability that when lineages from three species enter a node and the output consists of a single mixed lineage and lineages from a single species, reciprocal monophyly is preserved. We call this function G.

In computing G, we use the same strategies as we used when computing F (Eq. (8)) and H (Eq. (15)). If non-M lineages from only a single species remain, then a single interspecies coalescence must have occurred. Suppose the interspecies coalescence takes place between a lineage from species A and a lineage from species B. Then the lineages from species A and those from species B must each coalesce to a single lineage, and then the (A, B) coalescence must occur (Fig. 2B). We start with w, x, and y lineages from species A, B, and C, respectively, and c lineages from species C are present at the time of the (A, B) coalescence. The dashed line in Fig. 2B indicates this point in the coalescence sequence; $c = 3$. The number of ways to get to this point is $I_{w,1} I_{x,1} I_{y,c} W_3(w-1, x-1, y-c)$. Next, the (A, B) coalescence occurs. The result of this coalescence is a single mixed lineage (the gray lineage in Fig. 2B). The c lineages from species C coalesce to c_1 lineages; in Fig. 2B, $c_1 = 2$. The number of ways this sequence of coalescences can occur is I_{c,c_1}. The number of coalescence sequences that reduce w, x, and y lineages of three species to c_1 lineages of the third species and a single mixed lineage while preserving reciprocal monophyly is therefore $I_{w,1} I_{x,1} I_{y,c} W_3(w-1, x-1, y-c) I_{c,c_1}$. The total number of coalescence sequences that reduce $w+x+y$ lineages to c_1+1 lineages is I_{w+x+y,c_1+1}. Consequently, the probability that w lineages from one species, x lineages from a second species, and y lineages from a third species coalesce to one mixed lineage and c_1 lineages of the third species is

$$G(w, x, y, c_1) = \sum_{c=c_1}^{y} \frac{I_{w,1} I_{x,1} I_{y,c} W_3(w-1, x-1, y-c) I_{c,c_1}}{I_{w+x+y,c_1+1}}. \tag{19}$$

The first two arguments correspond to the species that coalesce, and they are exchangeable, so that $G(w, x, y, c_1) = G(x, w, y, c_1)$. Fig. 2B illustrates a specific coalescence sequence that would be counted by Eq. (19).

For node L_2, the values of K_{L_2} are

$$K_{L_2}(\ell, m, n, m_1, w, x, y, m_2) = \begin{cases} F(\ell, m, n) & \text{Case 1: } \ell, m, n \geq 1, \\ & w = x = y = m_1 = 0, \\ & m_2 = 1 \\ G(m, n, \ell, w) & \text{Case 2: } \ell, m, n, w \geq 1, \\ & x = y = m_1 = 0, \\ & m_2 = 1 \\ G(\ell, n, m, x) & \text{Case 3: } \ell, m, n, x \geq 1, \\ & w = y = m_1 = 0, \\ & m_2 = 1 \\ G(\ell, m, n, y) & \text{Case 4: } \ell, m, n, y \geq 1, \\ & w = x = m_1 = 0, \\ & m_2 = 1 \\ \dfrac{I_{\ell,w} I_{m,x} I_{n,y} W_3(\ell-w, m-x, n-y)}{I_{\ell+m+n, w+x+y}} & \text{Case 5: } \ell, m, n, w, x, \\ & y \geq 1, \\ & m_1 = m_2 = 0 \\ \dfrac{I_{n,y}}{I_{n+1,y+1}} & \text{Case 6: } n, y \geq 1, \\ & w = x = \ell = m = 0, \\ & m_1 = m_2 = 1 \\ \dfrac{I_{n,1}}{I_{n+1,1}} & \text{Case 7: } n \geq 1, \\ & w = x = \ell = m = y = 0, \\ & m_1 = m_2 = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

Node L_2 (Eq. (20)) has seven possible nontrivial cases. Case 1 applies when lineages of three species enter and a single mixed lineage exits, and the computation is given by F (Eq. (8)) as before.

Cases 2–4 apply when lineages from three species enter and a single interspecies coalescence occurs. These cases correspond to G (Eq. (19)), with different argument assignments that depend on the species identities in the interspecies coalescence.

Case 5 applies when lineages from three species enter and no interspecies coalescences occur. The total number of possible coalescence sequences is $I_{\ell+m+n, w+x+y}$. For reciprocal monophyly to be preserved, ℓ lineages from species A must coalesce to w lineages (counted by $I_{\ell,w}$), m lineages from species B must coalesce to x lineages (counted by $I_{m,x}$), and n lineages from species C must coalesce to y lineages (counted by $I_{n,y}$). The number of ways to order these three separate sequences is $W_3(\ell-w, m-x, n-y)$. The total number of possible coalescence sequences that preserve reciprocal monophyly is therefore $I_{\ell,w} I_{m,x} I_{n,y} W_3(\ell-w, m-x, n-y)$.

Case 6 applies when lineages from species C enter the node along with a mixed lineage that has resulted from an interspecies coalescence between species A and B in the branch L_1, and no interspecies coalescence occurs in branch L_2. The number of possible coalescence sequences is $I_{n+1,y+1}$. For reciprocal monophyly to be preserved, the n lineages from species C must coalesce to y lineages and the mixed lineage must not participate in a coalescence; there exist $I_{n,y}$ such coalescence sequences.

Case 7 applies when lineages from species C enter the node along with a mixed lineage but, unlike in Case 6, an interspecies coalescence occurs, and only a single mixed lineage remains. The number of possible coalescence sequences is $I_{n+1,1}$. For reciprocal monophyly to occur, the n lineages from species C must coalesce to a single lineage while the mixed lineage does not coalesce, and then the remaining C lineage must coalesce with the mixed lineage. The coalescence of the C lineages can occur in $I_{n,1}$ ways and the coalescence with the mixed lineage can occur in only one way. The total number of possible coalescence sequences that preserve reciprocal monophyly is therefore $I_{n,1}$.

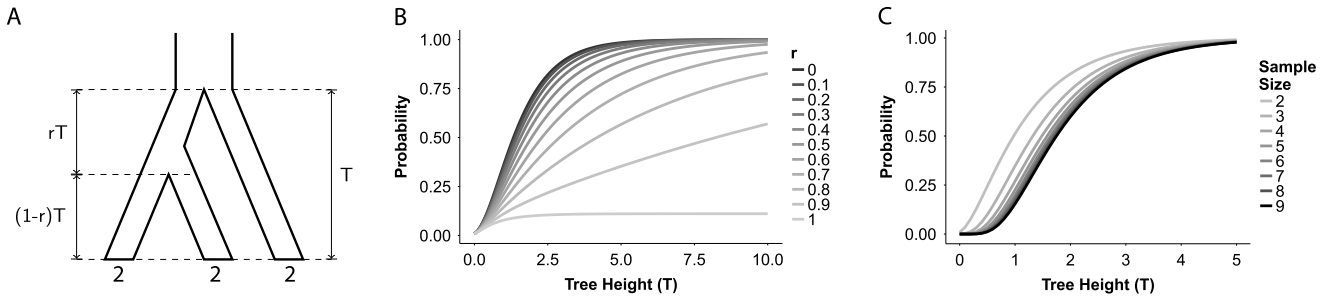

Fig. 3. Probability of reciprocal monophyly for a three-species tree. Species tree height varies from 0 to 5, and the internal branch length parameter r varies from 0 to 1. (A) Species tree with samples of size 2 from each of the three species. (B) Probability of reciprocal monophyly, with height T on the x-axis and shading based on r, with darker shades representing smaller values of r. As $T \to \infty$, the probability for $r = 1$ approaches $\frac{1}{9}$. For all other values of r, the probability approaches 1. (C) Probability of reciprocal monophyly as a function of height for different sample sizes, with the value of r fixed at 0.5. Sample sizes range from 2 to 9 and are assumed to be equal for all three species. Darker shades represent larger sample sizes.

5.2.3. K_{root}

For the root, the values of K_{root} are

$$K_{\text{root}}(w, x, y, z, m_2) = \begin{cases} H(w, x, y, z) & \text{Case 1: } w, x, y, z \geq 1, m_2 = 0 \\ F(w, z, 1) & \text{Case 2: } w, z \geq 1, x = y = 0, m_2 = 1 \\ F(x, z, 1) & \text{Case 3: } x, z \geq 1, w = y = 0, m_2 = 1 \\ F(y, z, 1) & \text{Case 4: } y, z \geq 1, w = x = 0, m_2 = 1 \\ \dfrac{I_{z,1}}{I_{z+1,1}} & \text{Case 5: } z \geq 1, w = x = y = 0, m_2 = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

The root (Eq. (21)) has five possible nontrivial cases. As was true for the case of a balanced four-species topology (Eq. (16)), Case 1 applies when lineages from all four species are present, and we use H (Eq. (15)).

Cases 2–4 apply when exactly one interspecies coalescence occurs below the root branch, and lineages from two species and a mixed lineage enter the root. This scenario corresponds to the calculation of F (Eq. (8)), with one of the values being 1, as was true in the balanced case (Eq. (16)).

Case 5 applies when two interspecies coalescences occur below the root branch, which yields lineages from species D and one mixed lineage entering the root branch. This situation is equivalent to Case 7 for K_{L_2} (Eq. (20)), but with z in place of n.

6. Results

6.1. Species tree height and internal branch lengths: three species

To study the effect of the length of the internal branch on the probability of reciprocal monophyly for three species, we consider a three-species tree with height T, 2 lineages per species, and internal branch length rT, where $r \in [0, 1]$ (Fig. 3A). We compute the probability of reciprocal monophyly according to Eq. (5) for different values of T and r.

Fig. 3B displays probabilities of reciprocal monophyly for this species tree, varying both T and r from 0 to 1. The probability of reciprocal monophyly increases monotonically as T increases with fixed r and as r decreases with fixed T.

When $r = 0$, the species tree has a star shape, globally maximizing the probability of reciprocal monophyly by minimizing the time that lineages from multiple species co-occur in the same population. When $r = 1$, samples from two species are forced into the same branch; as $T \to \infty$, the branch containing these two species becomes infinitely long and resembles a root branch. In this case, the probability of reciprocal monophyly of three species approaches the corresponding probability for a two-species tree, each with sample size 2, and leaf length zero. Using the general closed-form solution for a two-species tree, this probability, originally obtained by Tajima (1983), is equal to $\frac{1}{9}$ (Rosenberg 2003, Zhu et al. 2011, Mehta et al. 2016). If $0 < r < 1$, then as $T \to \infty$, the leaf branch lengths all approach ∞, and the probability of reciprocal monophyly approaches 1. In this limit, each leaf branch only contains lineages from one species, and they all coalesce before they have the opportunity to coalesce with lineages of other species.

6.2. Sample size: three species

To explore the effect of sample size on the probability of reciprocal monophyly for three species, we compute the probability of reciprocal monophyly (Eq. (5)) for the same three-species tree as in Fig. 3A, but we fix $r = 0.5$ and vary the sample size of each species from 2 to 9.

Fig. 3C provides the reciprocal monophyly probabilities over the range of species tree heights T from 0 to 5. The probability of reciprocal monophyly decreases monotonically with increasing sample size. However, as sample size increases for fixed T, the reciprocal monophyly probability approaches a constant. For large T, sample size quickly ceases to become important as the probabilities approach their large-sample limits.

6.3. Species tree height and internal branch lengths: four species

A four-species tree with height fixed at $T = 1$ has two free parameters: the lengths of the two internal branches. We set these lengths to aT and bT, with $a, b \in [0, 1]$. For the balanced topology, we vary a and b across the full range of possible values (Fig. 4A). For the caterpillar topology, we vary a and b with the additional constraint that $a + b \leq 1$, with the remaining height $(1 - a - b)T$ taken up by the shortest leaf branches (Fig. 4D). We compute the probability of reciprocal monophyly for the balanced species tree topology using Eq. (10) and for the caterpillar species tree topology using Eq. (17). Each species has sample size 2.

The probability of reciprocal monophyly for the balanced species tree is symmetric in the two internal branch lengths aT and bT, so the probabilities are symmetric across the $a = b$ line (Fig. 4B). The probability decreases monotonically as a and b each increase, owing to the increased probability of deep coalescence compared to coalescence in the leaves.

The probability of reciprocal monophyly for the caterpillar species tree is not symmetric in the two internal branch lengths aT and bT, so the probabilities are not symmetric across the $a = b$ line (Fig. 4E). As was true for the balanced case, the probability of monophyly decreases monotonically as a and b each increase, owing to the increased probability of deep coalescence compared to coalescence in the leaves.

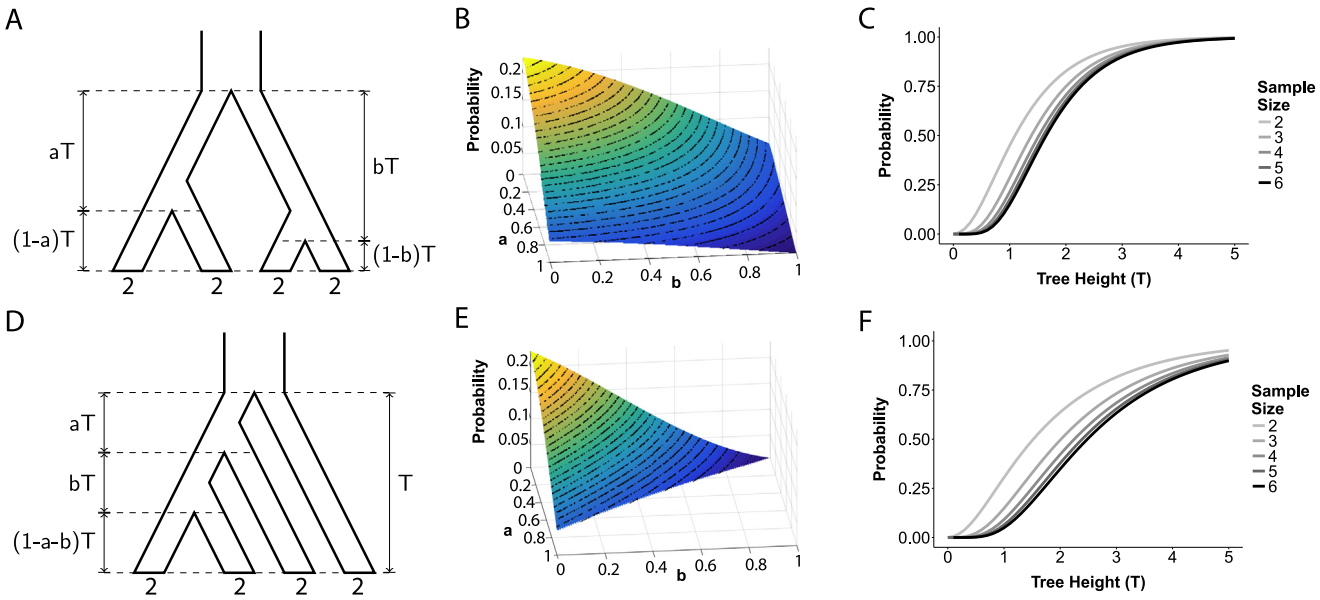

Fig. 4. Probability of reciprocal monophyly for a four-species tree. (A) Balanced species tree with samples of size 2 from each species. (B) Probability of reciprocal monophyly for the balanced species tree with $T = 1$, 2 lineages per species, and $a, b \in [0, 1]$. (C) Probability of reciprocal monophyly for the balanced species tree with $T \in [0, 5]$, $a = b = 0.5$, and sample size per species ranging from 2 to 6. (D) Caterpillar species tree with samples of size 2 from each species. (E) Probability of reciprocal monophyly for the caterpillar species tree with $T = 1$, 2 lineages per species, $a, b \in [0, 1]$, and $a + b \leq 1$. (F) Probability of reciprocal monophyly for the caterpillar species tree with $T \in [0, 5]$, $a = b = \frac{1}{3}$, and sample size per species ranging from 2 to 6. Sample sizes are assumed to be equal for all species.

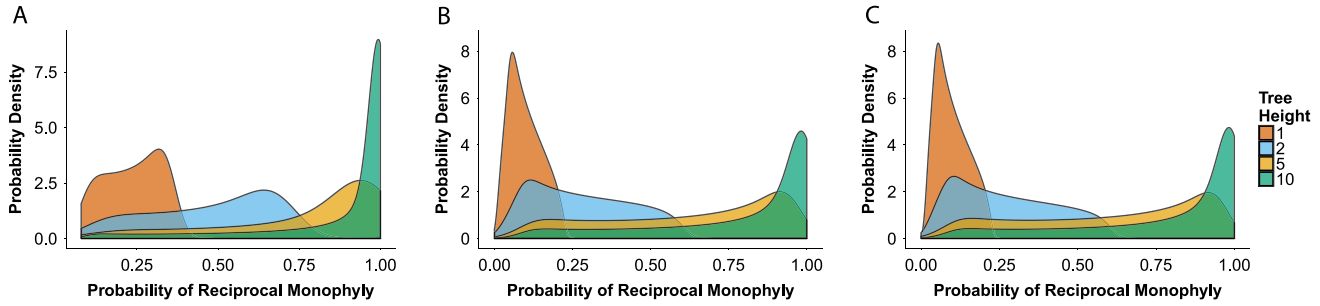

Fig. 5. Distributions of the reciprocal monophyly probability when species tree height T is fixed and internal branch lengths are chosen at each point on a grid of possible values. The kernel density estimates use Gaussian smoothing with bandwidth determined by the *bw.nrd0* function in R. (A) Three species (Fig. 3A) with parameter $r \in \{0, 0.01, 0.02, \ldots, 1\}$. (B) Four species, balanced topology (Fig. 4A) with parameters $(a, b) \in \{0, 0.01, \ldots, 1\} \times \{0, 0.01, \ldots, 1\}$. (C) Four species, caterpillar topology (Fig. 4D) with parameters a and b chosen so that $a + b \leq 1$ and $(a, b) \in \{0, 0.01, \ldots, 1\} \times \{0, 0.01, \ldots, 1\}$. Sample size for each species is fixed at 2.

6.4. Sample size: four species

As we did in the case of three species, to explore the effects of sample size on the reciprocal monophyly probability, we fix the internal branch lengths. We set $a = b = 0.5$ for the balanced species tree (Fig. 4A) and $a = b = \frac{1}{3}$ for the caterpillar species tree (Fig. 4D), varying T from 0 to 5. For both species trees, we vary the sample sizes per species from 2 to 6.

Figs. 4C and 4F display the probabilities of reciprocal monophyly for varying species tree height and sample size for the balanced (Fig. 4A, computed using Eq. (10)) and caterpillar (Fig. 4D, computed using Eq. (17)) species trees, respectively. As we observed for the three-species tree (Fig. 3), for both four-taxon topologies, the probability increases monotonically with increasing species tree height and decreases monotonically with increasing sample size. For a fixed species tree height, the probabilities again converge quickly as the sample sizes increase.

6.5. Distributions of the reciprocal monophyly probability

To provide a summary of the reciprocal monophyly probability for species trees of fixed height with branch lengths unspecified, we consider distributions of reciprocal monophyly probability for a fixed species tree height given a grid of choices for the branch lengths. For the three-species tree (Fig. 3), we examine all values of $r \in \{0, 0.01, 0.02, \ldots, 1\}$. For the four-species tree with balanced topology (Fig. 4A), we consider all $(a, b) \in \{0, 0.01, \ldots, 1\} \times \{0, 0.01, \ldots, 1\}$. For the four-species tree with caterpillar topology (Fig. 4D), we consider all such (a, b) that also satisfy $a + b \leq 1$.

Fig. 5A plots the probability density of reciprocal monophyly probabilities for the three-species case for $T = 1, 2, 5$, and 10. Figs. 5B and 5C plot the densities of reciprocal monophyly probabilities for the four-species balanced and caterpillar cases, respectively, for $T = 1, 2, 5$, and 10.

As species tree height increases, the means and modes of the densities increase. In general, the width of the central part of the distribution increases as height increases from 0. Despite a persistent left tail, the density shifts to the right as the height gets large enough to force most of the distribution close to 1. These patterns are seen for both three and four species (Fig. 5).

A larger height T permits the possibility of larger monophyly probabilities, but if small values are chosen for the internal branch lengths or the branch lengths at the leaves of the species tree, then small monophyly probabilities are still possible. These small probabilities, however, become less likely as height increases. Small

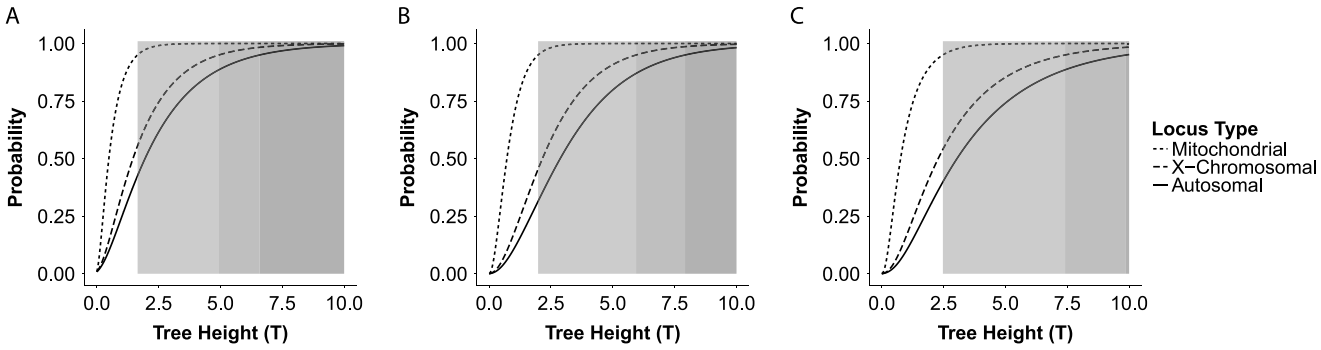

Fig. 6. Probabilities of reciprocal monophyly for autosomal, X-chromosomal, and mitochondrial loci for three and four species, with species tree height $T \in [0, 10]$ for the autosomal locus. All species trees have two lineages per species. (A) Three-species tree with internal branch length $\frac{T}{2}$. (B) Four-species balanced tree topology with internal branch lengths $\frac{T}{2}$. (C) Four-species caterpillar tree topology with internal branch lengths $\frac{T}{3}$. In the lightest shaded region, the probability exceeds 0.95 for mitochondrial loci only. In the next-lightest shaded region, the probability exceeds 0.95 for mitochondrial and X-chromosomal loci. In the darkest shaded region, the probability exceeds 0.95 for all types of loci. This region extends to $T \to \infty$.

probabilities of reciprocal monophyly occur for small values of $(1-r)T$, $(1-a)T$, $(1-b)T$, or $(1-a-b)T$; if T is larger, then r, a, and b must be larger to yield small reciprocal probabilities, and such large values of r, a, or b are less likely to be chosen.

6.6. Multiple types of loci

It is possible to construct gene trees with the same individuals, but from a variety of genetic sources, each with different effective population sizes. In mammals, gene trees can be constructed from autosomal loci, mitochondrial loci, X-chromosomal loci, or Y-chromosomal loci. If we assume equal population size and demography for males and females, we can express the effective population sizes of mitochondria and of the sex chromosomes in terms of the autosomal effective population size N_e^A. For mitochondria and the Y chromosome, $N_e^M = N_e^Y = \frac{1}{4}N_e^A$, and for the X chromosome, $N_e^X = \frac{3}{4}N_e^A$. The branch lengths in our species trees are in coalescent units, which are inversely proportional to population size. Hence, time T for autosomal loci corresponds to time $4T$ for mitochondrial and Y-chromosomal loci and to $\frac{4}{3}T$ for X-chromosomal loci.

Fig. 6 shows reciprocal monophyly probabilities for three and four species for species trees in which height ranges from $T = 0$ to $T = 5$ for autosomal loci. The species tree of height T has height $4T$ for mitochondrial and Y-chromosomal loci and height $\frac{4}{3}T$ for X-chromosomal loci. The three-species tree is from Fig. 3A with $r = 0.5$, the balanced four-species tree is from Fig. 4A with $a = b = 0.5$, and the caterpillar four-species tree is from Fig. 4D with $a = b = \frac{1}{3}$.

For each species tree, a region of heights exists where reciprocal monophyly has probability greater than 0.95 for mitochondrial loci but not for X-chromosomal or autosomal loci. For larger species tree heights, reciprocal monophyly has probability greater than 0.95 for non-autosomal loci but not for autosomal loci. For still larger species tree heights, reciprocal monophyly has probability greater than 0.95 for all types of loci.

6.7. Data example

We now apply our theoretical results to gene trees from maize and teosinte. Mehta et al. (2016) measured monophyly frequencies of individual clades of four groups of maize (two nondomesticated groups: *Zea mays* ssp. *mexicana*, *Zea mays* ssp. *parviglumis*, and two domesticated groups: *Zea mays* ssp. *mays* landraces and improved *Zea mays* ssp. *mays*, identified henceforth as "Mexicana", "Parviglumis", "Landraces", and "Improved", respectively) across many randomly-selected loci for 100 different samples of 8 individuals, 2 from each group, from the dataset reported by Chia et al. (2012). The frequencies measured were compared to theoretical probabilities computed using a species tree adapted from Hufford et al. (2012) and reproduced in Fig. 7A.

Here, instead of single-clade monophyly probabilities, we compute reciprocal monophyly probabilities for all four groups and for all four sets of three of the groups. Table 2 and Fig. 8 present observed frequencies of reciprocal monophyly in the gene trees examined by Mehta et al. (2016) and theoretical probabilities from Eqs. (5) and (17) with two samples per species and the species trees shown in Fig. 7.

The observed reciprocal monophyly frequencies are close to the theoretical probabilities, especially for the four-species tree that includes all groups (Fig. 7A and "All Groups" in Fig. 8) as well as for the three-species trees that contain both nondomesticated groups (Parviglumis and Mexicana; Figs. 7B and 7C, and "No Improved" and "No Landraces" in Fig. 8). These computations suggest that the theoretical probabilities of reciprocal monophyly reasonably capture the behavior of genealogical processes that occur in the phylogeny.

6.8. Software implementation

Eqs. (5), (10) and (17) are implemented in the software MONOPHYLER, which also computes the monophyly probabilities of Mehta et al. (2016) and can be found at http://rosenberglab.stanford.edu/monophyler.html.

7. Discussion

We have derived expressions for the probability of reciprocal monophyly of the lineages of three and four species given a species tree and sample sizes from each species. We have studied the dependence of these probabilities on species tree parameters such as height, lengths of internal branches, and sample sizes. Our results indicate that the probability of reciprocal monophyly increases with increasing height, decreasing sample size, and decreasing lengths of internal branches given a fixed height. Among these parameters, height generally has the strongest effect.

With other factors held constant, any increase in the amount of time that lineages from the same species can coalesce only among themselves increases the probability of reciprocal monophyly. The trends seen in Figs. 3B, 4B, and 4E reflect this observation: increasing height T and decreasing the internal branch length parameter r (Fig. 3) or a and b (Fig. 4) increases the amount of time that lineages from the same species can coalesce only among themselves, thereby increasing the reciprocal monophyly probability.

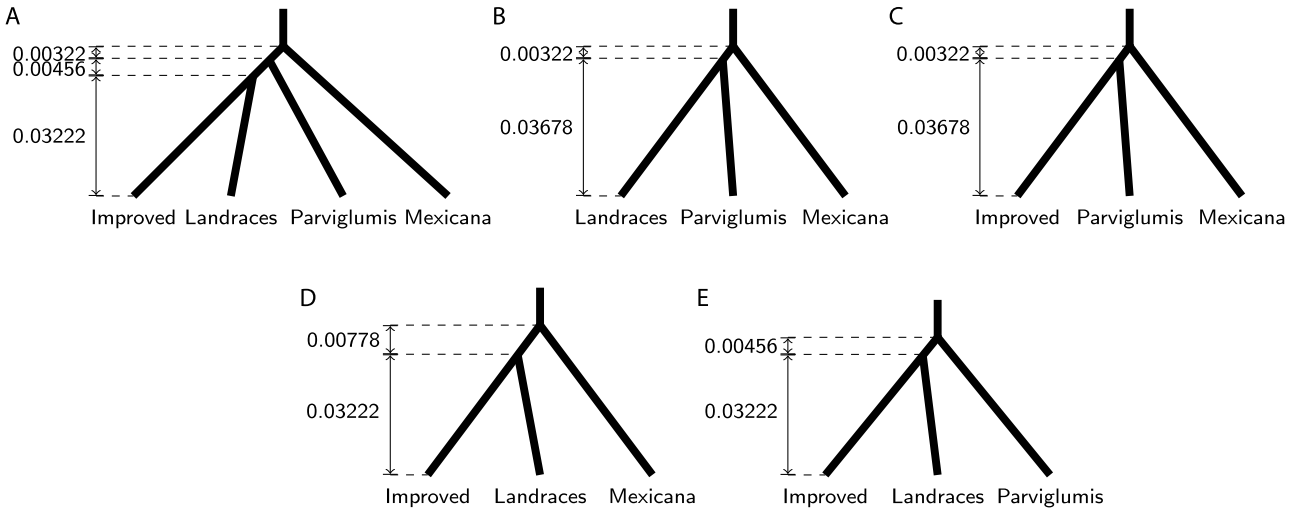

Fig. 7. Species trees from maize and teosinte used for comparing observed reciprocal monophyly frequencies to theoretical probabilities. (A) All groups. Each three-species tree (B–E) is the four-species tree in (A) with a single leaf removed, as follows: (B) Improved. (C) Landraces. (D) Parviglumis. (E) Mexicana. Note that all species trees have height 0.04 except (E), which has height 0.03678.

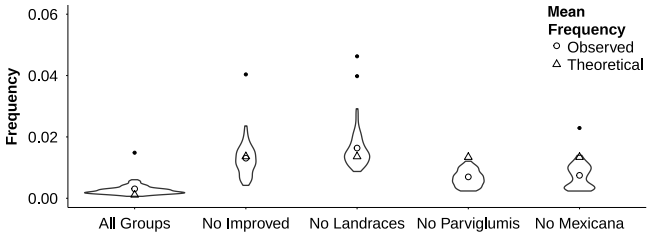

Fig. 8. Reciprocal monophyly frequencies for gene trees from maize and teosinte. Triangles represent theoretical probabilities, and circles represent observed mean frequencies across 100 sets of sampled individuals. Violins indicate distributions across the 100 sets. Outliers are plotted as separate points.

Table 2
Comparison of observed frequencies and theoretical probabilities of reciprocal monophyly for the species trees in Fig. 7. All groups: Improved, Landraces, Parviglumis, Mexicana. For each of the other four rows, one of the four groups is omitted from the species tree. The four-species theoretical probability is taken from Eq. (17), and three-species theoretical probabilities are from Eq. (5). Observed means and standard deviations consider 100 samples of individuals from Chia et al. (2012), evaluating reciprocal monophyly probabilities for a set of loci along the genome as in Mehta et al. (2016).

Tree	Number of groups	Theoretical probability	Observed mean	Observed SD
All groups	4	0.0011	0.0027	0.0017
No Improved	3	0.0136	0.0127	0.0052
No Landraces	3	0.0136	0.0160	0.0058
No Parviglumis	3	0.0134	0.0066	0.0026
No Mexicana	3	0.0131	0.0071	0.0038

The relative lack of importance of sample size in Figs. 3C, 4C, and 4F is consistent with the quick convergence of coalescent processes as sample sizes increase, previously seen in the two-species analysis of monophyly probabilities (Rosenberg, 2003). In all cases explored here, the greatest change in probability occurs when sample sizes for each species are increased from 2 to 3. Every subsequent increase produces smaller changes in probability; as additional lineages are incorporated, they do not necessarily produce a substantial increase in the number of ancestral lineages at the time of the first species divergence, so that sample size has relatively little effect after the branch lengths are sufficiently long. As the per-species sample size increases beyond 5, except at the smallest values of height T, the probability of reciprocal monophyly closely approaches the infinite-sample-size case.

To explore the effect of internal branch lengths on the probability of reciprocal monophyly, we have computed the distribution of reciprocal monophyly probabilities over grids of internal branch lengths for a fixed species tree height. The resulting reciprocal monophyly probability distributions are relatively narrow, suggesting that species tree height is a more important factor than the internal branch length for determining reciprocal monophyly probability. However, configurations of internal branch lengths that result in low probabilities of reciprocal monophyly always exist even if the species tree height is large, so internal branch length cannot be ignored entirely. In particular, if there is the possibility of recent divergence in the species tree, then internal branch lengths become important for determining the reciprocal monophyly probability. The distributions of reciprocal monophyly probabilities are similar for balanced and caterpillar topologies (compare Figs. 5B and 5C).

Extending computations of Moore (1995) and Rosenberg (2003) that examined mitochondrial and autosomal loci in species pairs, we have also studied the situation in which data are collected from multiple different genetic sources, each with a different effective population size. For a given species tree height with respect to autosomal loci, mitochondrial loci and Y-chromosomal loci are expected to have the highest levels of reciprocal monophyly, followed by X-chromosomal loci, and finally by autosomal loci. A period exists during which mitochondrial loci, but not X-chromosomal or autosomal loci, are expected to be reciprocally monophyletic with $>95\%$ probability, followed by a period during which mitochondrial and X-chromosomal loci, but not autosomal loci, are expected to be reciprocally monophyletic with $>95\%$ probability. After a sufficient length of time, all classes of loci are expected to be reciprocally monophyletic with $>95\%$ probability. Thus, studies in which genes are sampled from sources with different effective population sizes allow a richer array of reciprocal monophyly patterns that can be used to assess genealogies more thoroughly than if just a single source were used.

Finally, in a data example using SNP data from maize, we computed frequencies of reciprocal monophyly for three and four groups of maize subspecies and compared those frequencies to theoretical probabilities obtained from our results under a specific species tree model. We found that the observed frequencies generally match the theoretical predictions. The same caveats that were

present in a parallel analysis of monophyly probabilities by Mehta et al. (2016) apply here. First, there is uncertainty in the species tree model that we do not take into account and that might be the cause of the deviations from our theoretical predictions observed in some of the three-species analyses. Second, we do not account for the gene flow between groups that is likely present in this system. Third, we estimate the gene trees from the same data from which the species tree was estimated, so that monophyly frequencies might be implicitly compatible. We do not expect this dual use of the data to have a major effect on the interpretation of our reciprocal monophyly probabilities: reciprocal monophyly largely reflects species tree height, and the height of our species tree was obtained from a study that did not use the same data we used (see Mehta et al., 2016 for details).

Starting with small numbers of taxa and then generalizing to arbitrary numbers has been a successful strategy for probability computations for features of gene trees conditional on species trees. The relationship between gene trees and species trees was studied in detail for three species (Nei, 1987; Takahata, 1989; Rosenberg, 2002) and then extended (Pamilo and Nei, 1988; Degnan and Salter, 2005). For monophyly problems, the probability of monophyly when lineages are classified into two subsets was initially solved for two taxa (Rosenberg, 2003), and the extension to arbitrary numbers of taxa by Mehta et al. (2016) built upon the techniques of the two-taxon case. Our analysis extends this two-class analysis to the case of three and four classes for species trees with three and four taxa.

A next step for this work is to further generalize the theory from three and four classes to n classes and from three and four species to n species. Although our computations have been specific to the species trees and lineage classes considered, it would be possible in principle to proceed by a similar approach for a specified larger tree of interest, with a larger number of lineage classes. A general algorithm that could perform the computation for an arbitrary number of species and lineage classes would be desirable; because the computation could quickly become unwieldy, a possible generalization might involve a method to construct the analytical formula by computer algebra instead of a calculation of the analytical formula itself. It may be possible to aggregate the combinatorial work of Zhu et al. (2011) on multiple classes of lineages in one species and the work of Mehta et al. (2016) on recursive analysis of arbitrary species trees in order to achieve this generalization.

Acknowledgments

We thank Scott Edwards for conversations that facilitated this project. We acknowledge support from NIH grant R01 GM117590 and a Stanford Graduate Fellowship.

Appendix

A.1. Correction of results from Rosenberg (2002)

Rosenberg (2002) presented a formula for the probability that lineages from three species are separately monophyletic and that, in addition, the order of interspecies coalescences mimics the topology of the species tree, terming this situation "monophyletic concordance." However, Rosenberg (2002) mistakenly assumed that no interspecies coalescences occur during the root branch of the species tree until all lineages have coalesced to three total lineages. It is, in fact, possible for an interspecies coalescence to occur before all species have intraspecifically coalesced to single lineages while still permitting reciprocal monophyly or monophyletic concordance. Fig. 2A provides an example of a coalescence sequence that violates the assumption made by Rosenberg (2002) but still gives rise to reciprocal monophyly and monophyletic concordance, assuming the species tree follows Fig. 1A.

Eqs. 11–13 of Rosenberg (2002) therefore omit some cases of coalescence sequences that permit reciprocal monophyly or monophyletic concordance. Our Eq. (5) provides the correct probability corresponding to Eq. (12) of Rosenberg (2002), and substituting $T_2 = 0$ in Eq. (5) provides the correct probability corresponding to Eq. (13) of Rosenberg (2002). The correct monophyletic concordance probability $\mathbb{P}(E_{MC})$ corresponding to Eq. (11) of Rosenberg (2002) is

$$\mathbb{P}(E_{MC}) = \sum_{s=1}^{p}\sum_{t=1}^{q}\sum_{m=0}^{1}\sum_{w=0}^{s}\sum_{x=0}^{t}\sum_{y=1}^{r} g_{p,s}(T_3)g_{q,t}(T_3)g_{r,y}(T_2+T_3) \times g_{s+t,w+x+m}(T_2)K_I(s,t,w,x,m)K^{mc}_{\text{root}}(w,x,y,m), \quad (22)$$

where K_I from Eq. (6) appears because monophyletic concordance is coincident with reciprocal monophyly in the internal branch. K^{mc}_{root}, describing the probability that lineages from all three species enter the root and coalesce to a single mixed lineage with monophyletic concordance, is:

$$K^{mc}_{\text{root}}(w,x,y,m) = \begin{cases} f(w,x,y) & \text{Case 1: } w,x,y \geq 1, m=0 \\ \dfrac{I_{y,1}}{I_{y+1,1}} & \text{Case 2: } y \geq 1, w=x=0, m=1 \\ 0 & \text{otherwise.} \end{cases} \quad (23)$$

Note that in K^{mc}_{root}, f (Eq. (7)) is used instead of F (Eq. (8)) because monophyletic concordance only occurs when the (A, B) coalescence occurs first.

Table 3 shows three-species reciprocal monophyly and monophyletic concordance probability computations from the sixth and seventh columns of Table 2 in Rosenberg (2002) and the corrected values of those probabilities obtained using Eqs. (5) and (22). Table 3 indicates that the values in Rosenberg (2002) and the corrected values are numerically quite close. The new values are slightly larger, as they account for scenarios that were not permitted by Rosenberg (2002). The difference is greater when branch lengths are shorter, the setting in which the case not taken into account by Rosenberg (2002) is most probable. When $r = 1$, the case omitted from Rosenberg (2002) does not occur with monophyletic concordance, and the monophyletic concordance probabilities match exactly. However, the reciprocal monophyly probabilities do not necessarily match when $r = 1$ because interspecies coalescences can still occur before each species reduces to a single lineage.

A.2. Four-taxon monophyletic concordance probability

We note that an analogous four-taxon monophyletic concordance probability is straightforward to obtain by our approach (see Degnan, 2010 for an alternative approach). Computing the probabilities of monophyletic concordance for four-species trees requires a similar modification of our results to that used in the three-species case in Appendix A.1: we substitute $K_L^{mc,b}$, $K_R^{mc,b}$, and $K_{\text{root}}^{mc,b}$ for K_L, K_R, and K_{root} in Eq. (10) for the balanced topology and $K_{L_1}^{mc,c}$, $K_{L_2}^{mc,c}$, and $K_{\text{root}}^{mc,c}$ for K_{L_1}, K_{L_2}, and K_{root} in Eq. (17) (superscripts mc, b, and c denote monophyletic concordance, the balanced topology, and the caterpillar topology, respectively). For the four-species tree with balanced topology, in the internal branches, monophyletic concordance is coincident with reciprocal monophyly, so $K_L^{mc,b} = K_L$ from Eq. (11) and $K_R^{mc,b} = K_R$ from Eq. (12). In the root branch, only two terms of $H(w,x,y,z)$ yield monophyletic concordance: $h_2(w,x,y,z)$ and $h_2(y,z,w,x)$. Only single terms of $F(w,x,1)$ and $F(y,z,1)$ yield monophyletic concordance: $f(w,x,1)$ and $f(y,z,1)$. Note that monophyletic concordance does

Table 3
A comparison of three-species reciprocal monophyly and monophyletic concordance probabilities obtained from Table 2 in Rosenberg (2002) with corrected probabilities obtained by Eqs. (5) and (22). The corrected probabilities are systematically larger, and the effect is more pronounced for shorter branch lengths, when the scenarios not taken into account by Rosenberg (2002) are most probable. The monophyletic concordance probabilities match when $r = 1$, at which the result of Rosenberg (2002) is correct.

(p, q, r)	T_3	T_2	Reciprocal monophyly		Monophyletic concordance	
			Rosenberg (2002)	Eq. (5)	Rosenberg (2002)	Eq. (22)
(1, 1, 1)	≥0	0	1	1	0.333	0.333
		0.05	1	1	0.366	0.366
		0.5	1	1	0.596	0.596
		5	1	1	0.996	0.996
(2, 2, 1)	0.05	0	0.048	0.063	0.016	0.016
		0.05	0.056	0.072	0.019	0.019
		0.5	0.109	0.117	0.049	0.049
		5	0.134	0.134	0.133	0.133
	0.5	0	0.247	0.277	0.082	0.082
		0.05	0.260	0.288	0.092	0.092
		0.5	0.329	0.337	0.175	0.175
		5	0.355	0.355	0.353	0.353
	5	0	0.989	0.990	0.330	0.330
		0.05	0.989	0.990	0.362	0.362
		0.5	0.991	0.991	0.590	0.590
		5	0.991	0.991	0.987	0.987
(2, 2, 2)	0.05	0	0.011	0.015	0.004	0.005
		0.05	0.015	0.020	0.005	0.007
		0.5	0.059	0.065	0.028	0.030
		5	0.133	0.133	0.132	0.132
	0.5	0	0.124	0.145	0.041	0.048
		0.05	0.137	0.158	0.049	0.056
		0.5	0.234	0.244	0.127	0.131
		5	0.354	0.354	0.352	0.352
	5	0	0.983	0.984	0.328	0.328
		0.05	0.984	0.985	0.360	0.360
		0.5	0.987	0.988	0.588	0.588
		5	0.991	0.991	0.987	0.987
(5, 5, 1)	0.05	0	0.0002	0.0003	0.00007	0.00007
		0.05	0.0003	0.0004	0.0001	0.0001
		0.5	0.001	0.002	0.0005	0.0005
		5	0.002	0.002	0.002	0.002
	0.5	0	0.031	0.040	0.010	0.010
		0.05	0.035	0.045	0.012	0.012
		0.5	0.066	0.071	0.030	0.030
		5	0.082	0.082	0.081	0.081
	5	0	0.978	0.979	0.326	0.326
		0.05	0.978	0.980	0.358	0.358
		0.5	0.981	0.981	0.584	0.584
		5	0.982	0.982	0.978	0.978
(5, 5, 5)	0.05	0	0.000002	0.000002	0.0000005	0.0000007
		0.05	0.000005	0.000008	0.000002	0.000003
		0.5	0.0003	0.0004	0.0001	0.0002
		5	0.002	0.002	0.002	0.002
	0.5	0	0.006	0.008	0.002	0.003
		0.05	0.007	0.010	0.003	0.003
		0.5	0.030	0.033	0.014	0.016
		5	0.081	0.081	0.081	0.081
	5	0	0.967	0.969	0.322	0.323
		0.05	0.968	0.970	0.354	0.355
		0.5	0.975	0.976	0.580	0.580
		5	0.982	0.982	0.978	0.978

not require the order of the (A,B) and (C,D) interspecies coalescences to match that of the species tree. Thus, $K_{\text{root}}^{mc,b}$ is obtained by substituting $h_2(w, x, y, z) + h_2(y, z, w, x)$ in Case 1, $f(w, x, 1)$ in Case 2, and $f(y, z, 1)$ in Case 3 of Eq. (16).

For a four-species tree with caterpillar topology, in internal branch L_1, monophyletic concordance is coincident with reciprocal monophyly. Thus, $K_{L_1}^{mc,c} = K_{L_1}$ from Eq. (18). For the second internal branch L_2, only one term of $F(\ell, m, n)$ yields monophyletic concordance: $f(\ell, m, n)$. Also, only $G(\ell, m, n, y)$ and not $G(m, n, \ell, w)$ or $G(\ell, n, m, x)$ yields monophyletic concordance. Thus, $K_{L_2}^{mc,c}$ is obtained by substituting $f(\ell, m, n)$ for $F(\ell, m, n)$ in Case 1 and setting Cases 2 and 3 to zero in Eq. (20). Finally, for the root branch, only one term of $H(w, x, y, z)$ yields monophyletic concordance: $h_1(w, x, y, z)$. Similarly, only one term of $F(y, z, 1)$ yields monophyletic concordance: $f(1, y, z)$. Finally, $F(w, z, 1)$ and $F(x, z, 1)$ cannot yield monophyletic concordance. Thus, $K_{\text{root}}^{mc,b}$ is obtained by substituting $h_1(w, x, y, z)$ in Case 1 and $f(1, y, z)$ in Case 4, and setting Cases 2 and 3 to zero in Eq. (21).

References

Avise, J.C., Ball Jr., R.M., 1990. Principles of genealogical concordance in species concepts and biological taxonomy. In: Futuyma, D., Antonovics, J. (Eds.), Oxford Surveys in Evolutionary Biology, vol. 7. Oxford University Press, Oxford, pp. 45–67.

Baker, A.J., Tavares, E.S., Elbourne, R.F., 2009. Countering criticisms of single mitochondrial DNA gene barcoding in birds. Mol. Ecol. Resour. 9, 257–268.

Bergsten, J., Bilton, D.T., Fujisawa, T., Elliott, M., Monaghan, M.T., Balke, M., Hendrich, L., Geijer, J., Herrmann, J., Foster, G.N., et al., 2012. The effect of geographical scale of sampling on DNA barcoding. Syst. Biol. 61, 851–869.

Birky, C.W., 2013. Species detection and identification in sexual organisms using population genetic theory and DNA sequences. PLoS One 8, e52544.
Carstens, B.C., Richards, C.L., 2007. Integrating coalescent and ecological niche modeling in comparative phylogeography. Evolution 61, 1439–1454.
Chia, J.-M., Song, C., Bradbury, P.J., Costich, D., de Leon, N., Doebley, J., Elshire, R.J., Gaut, B., Geller, L., Glaubitz, J.C., et al., 2012. Maize HapMap2 identifies extant variation from a genome in flux. Nature Genet. 44, 803–807.
De Queiroz, K., 2007. Species concepts and species delimitation. Syst. Biol. 56, 879–886.
Dearborn, D.C., Gager, A.B., Gilmour, M.E., McArthur, A.G., Hinerfeld, D.A., Mauck, R.A., 2015. Non-neutral evolution and reciprocal monophyly of two expressed Mhc class II B genes in Leach's storm-petrel. Immunogenetics 67, 111–123.
Degnan, J.H., 2010. Probabilities of gene trees with intraspecific sampling given a species tree. In: Knowles, L.L., Kubatko, L.S. (Eds.), Estimating Species Trees: Practical and Theoretical Aspects. Wiley-Blackwell, Hoboken, NJ, pp. 53–78.
Degnan, J.H., Salter, L.A., 2005. Gene tree distributions under the coalescent process. Evolution 59, 24–37.
Eldon, B., Degnan, J.H., 2012. Multiple merger gene genealogies in two species: monophyly, paraphyly, and polyphyly for two examples of lambda coalescents. Theor. Popul. Biol. 82, 117–130.
Hudson, R.R., Coyne, J.A., 2002. Mathematical consequences of the genealogical species concept. Evolution 56, 1557–1565.
Hufford, M.B., Xu, X., Van Heerwaarden, J., Pyhäjärvi, T., Chia, J.-M., Cartwright, R.A., Elshire, R.J., Glaubitz, J.C., Guill, K.E., Kaeppler, S.M., et al., 2012. Comparative population genomics of maize domestication and improvement. Nature Genet. 44, 808–811.
Jakobsson, M., Rosenberg, N.A., 2007. The probability distribution under a population divergence model of the number of genetic founding lineages of a population or species. Theor. Popul. Biol. 71, 502–523.
Kubatko, L.S., Gibbs, H.L., Bloomquist, E.W., 2011. Inferring species-level phylogenies and taxonomic distinctiveness using multilocus data in *Sistrurus* rattlesnakes. Syst. Biol. 60, 393–409.
Lohse, K., Nicholls, J.A., Stone, G.N., 2011. Inferring the colonization of a mountain range—refugia vs. nunatak survival in high alpine ground beetles. Mol. Ecol. 20, 394–408.
Mehta, R.S., Bryant, D., Rosenberg, N.A., 2016. The probability of monophyly of a sample of gene lineages on a species tree. Proc. Natl. Acad. Sci. USA 113, 8002–8009.
Moore, W.S., 1995. Inferring phylogenies from mtdna variation: mitochondrial-gene trees versus nuclear-gene trees. Evolution 49, 718–726.
Moritz, C., 1994. Defining 'evolutionarily significant units' for conservation. Trends Ecol. Evol. 9, 373–375.
Nei, M., 1987. Molecular Evolutionary Genetics. Columbia University Press, New York.
Neigel, J.E., Avise, J.C., 1986. Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. In: Nevo, E., Karlin, S. (Eds.), Evolutionary Processes and Theory. Academic Press, New York, pp. 515–534.
Neilson, M.E., Stepien, C.A., 2009. Evolution and phylogeography of the tubenose goby genus *Proterorhinus* (Gobiidae: Teleostei): evidence for new cryptic species. Biol. J. Linn. Soc. 96, 664–684.
Pamilo, P., Nei, M., 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5, 568–583.
Rabeling, C., Schultz, T.R., Pierce, N.E., Bacci Jr., M., 2014. A social parasite evolved reproductive isolation from its fungus-growing ant host in sympatry. Curr. Biol. 24, 2047–2052.
Rosenberg, N.A., 2002. The probability of topological concordance of gene trees and species trees. Theor. Popul. Biol. 61, 225–247.
Rosenberg, N.A., 2003. The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. Evolution 57, 1465–1477.
Rosenberg, N.A., 2006. The mean and variance of the numbers of r-pronged nodes and r-caterpillars in Yule-generated genealogical trees. Ann. Combinator. 10, 129–146.
Tajima, F., 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105, 437–460.
Takahata, N., 1989. Gene genealogy in three related populations: consistency probability between gene and population trees. Genetics 122, 957–966.
Takahata, N., Nei, M., 1985. Gene genealogy and variance of interpopulational nucleotide differences. Genetics 110, 325–344.
Takahata, N., Slatkin, M., 1990. Genealogy of neutral genes in two partially isolated populations. Theor. Popul. Biol. 38, 331–350.
Tavaré, S., 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. 26, 119–164.
Tavares, E.S., Baker, A.J., 2008. Single mitochondrial gene barcodes reliably identify sister-species in diverse clades of birds. BMC Evol. Biol. 8, 81.
Wakeley, J., 2000. The effects of subdivision on the genetic divergence of populations and species. Evolution 54, 1092–1101.
Zhu, S., Degnan, J.H., Steel, M., 2011. Clades, clans, and reciprocal monophyly under neutral evolutionary models. Theor. Popul. Biol. 79, 220–227.