Mathematical Biosciences 235 (2012) 45–55

Contents lists available at SciVerse ScienceDirect

Mathematical Biosciences

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

The probability distribution of ranked gene trees on a species tree

James H. Degnan [a,*], Noah A. Rosenberg [b], Tanja Stadler [c]

[a] Department of Mathematics and Statistics, Private Bag 4800, University of Canterbury, Christchurch 8140, New Zealand
[b] Department of Biology, Stanford University, Stanford, CA 94305, USA
[c] Institute of Integrative Biology, ETH Zürich, Universitätsstrasse 16, 8092 Zürich, Switzerland

ARTICLE INFO

Article history:
Received 11 March 2011
Received in revised form 15 October 2011
Accepted 21 October 2011
Available online 31 October 2011

Keywords:
Anomalous gene trees
Coalescent
Genealogies
Phylogenetics
Population genetics

ABSTRACT

The properties of random gene tree topologies have recently been studied under a coalescent model that treats a species tree as a fixed parameter. Here we develop the analogous theory for random *ranked* gene tree topologies, in which both the topology and the *sequence* of coalescences for a random gene tree are considered. We derive the probability distribution of ranked gene tree topologies conditional on a fixed species tree. We then show that similar to the unranked case, ranked gene trees that do not match either the ranking or the topology of the species tree can have greater probability than the matching ranked gene tree.

1. Introduction

Recent studies have investigated the probability distribution of random gene tree topologies under a particular stochastic evolutionary model, the 'multispecies coalescent' [1–11]. Treating a species tree as a parameter consisting of a fixed labeled topology and fixed branch lengths, Degnan and Salter [3] obtained a probability distribution under the model for the labeled topology of a random gene tree evolving on the species tree.

This probability distribution has generated a wide variety of applications. First, it provides a mathematical basis for studying the properties of gene trees in a standard evolutionary model, enabling predictions about gene tree patterns that the evolutionary process is expected to produce [3,9,12]. Second, it underlies model-based analyses of the consistency properties of species tree inference algorithms, and more generally, of the ways in which different inference approaches behave as increasingly many loci are sampled [1,13–18]. Third, calculations of the probability distribution itself have been incorporated as a component of species tree inference algorithms, namely in likelihood computations applied to gene trees in observed data [19–23]. Finally, the approach used in deriving the distribution has initiated the study of coalescent histories, combinatorial objects that each describe a possible list of branches of the species tree in which the coalescences in a given gene tree can take place [24–26].

The gene tree topologies examined in the probability distribution of Degnan and Salter [3] are unranked, in that they consider only the topological relationship among gene lineages, and not the sequence in which the lineages coalesce. The additional information contained in the coalescence sequence or *labeled history* of a gene tree, however, can potentially lead to a novel method of summarizing gene tree distributions using ranked rather than unranked trees, thereby facilitating new approaches both in problems of evolutionary modeling and in species tree inference problems.

Our interest in ranked gene tree topologies arises partly from the proof of the existence of *anomalous gene trees* in the unranked case. Degnan and Rosenberg [1] showed that under the multispecies coalescent, a species tree can produce anomalous gene trees – unranked gene tree topologies that do not match the species tree topology, and whose probabilities exceed that of the gene tree topology that does match the species tree. The proof relies on the occurrence of variability in the probabilities of different unranked gene tree topologies under the multispecies coalescent. When species tree branch lengths are short, unranked gene tree topologies that can be produced by many possible sequences of coalescences – that is, unranked gene tree topologies with many possible rankings – have greater probabilities than those that have fewer rankings. It might therefore be expected that by considering ranked gene tree topologies, each of which can be produced by exactly one possible sequence of coalescences, *anomalous ranked gene trees* – ranked gene tree topologies that disagree with the ranked species tree topology and whose probabilities exceed the probability of the ranked gene tree topology that matches the ranked species tree topology – could be shown not to exist.

* Corresponding author. Tel.: +64 3 385 2644; fax: +64 3 364 2587.
E-mail addresses: j.degnan@math.canterbury.ac.nz (J.H. Degnan), noahr@stanford.edu (N.A. Rosenberg), tanja.stadler@env.ethz.ch (T. Stadler).

0025-5564/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.mbs.2011.10.006

Here, in a similar manner to previous work on the stochastic theory of unranked gene tree topologies conditional on a species tree, we pursue the theory of ranked gene tree topologies conditional on a species tree. In Section 2 we introduce notation, and in Section 3, analogously to the enumeration of coalescent histories required for computing the probability of a gene tree topology given a species tree in the unranked case, we discuss the number of scenarios that must be enumerated in computing the probability of a ranked gene tree topology given a species tree. We then derive a general expression for the probability of a specific ranked gene tree topology given a species tree (Section 4). Using this expression, in Section 5, we calculate probabilities of ranked gene tree topologies for various small species trees, with three, four, and five taxa. By a close examination of these probabilities in the five-taxon case, we establish that there do in fact exist five-taxon species trees that possess *anomalous ranked gene trees* (ARGTs), in which gene tree labeled histories that disagree with the species tree labeled history are more likely to be produced than the matching labeled history. We end the paper with a discussion in Section 6.

2. Definitions

A *species tree* $\mathcal{T}$ is a binary rooted tree topology together with its edge lengths. In general we consider *labeled* species trees, in which each leaf is associated with a distinct label. We view the lengths of edges as being proportional to time, as described below.

A species tree $\mathcal{T}$ induces a *ranked tree* Ψ as follows. Order the interior vertices of $\mathcal{T}$ by their distance from the root (we assume all distances are distinct). Assign the root rank 1, and assign the ith vertex in the sequence of interior vertices rank i. The ranked tree Ψ associated with species tree $\mathcal{T}$ is obtained by assigning ranks to all interior vertices of $\mathcal{T}$ and disregarding the edge lengths. This ranked tree then induces the *labeled topology* ψ by disregarding the ranks. Note that if two interior vertices in the species tree have the same distance from the root, Ψ is not well-defined, although ψ remains well-defined; we ignore such cases, treating Ψ as unambiguous. This choice to disregard ties is motivated by the fact that under typical stochastic models for species trees, the probability that two speciation events occur simultaneously is zero.

We indicate ranked tree topologies by adding clade ranks to the notation for unranked tree topologies. For each clade, we place the rank of the clade as a subscript after the closing parenthesis associated with that clade. For example, the three possible rankings for the tree (((AB)C)(DE)) can be written $(((AB)_3C)_2(DE)_4)$, $(((AB)_4C)_3(DE)_2)$, and $(((AB)_4C)_2(DE)_3)$, and the species trees in Fig. 1(a), (c), and (d) possess these rankings, respectively. A rank of 4 for a clade on a five-taxon tree, for instance, indicates that the most recent common ancestor (MRCA) for the clade is more recent than the MRCA for any other clade in the tree. We omit the rank 1 in our notation for tree rankings, as the rank 1 applies to the MRCA for the entire tree in any ranking.

For a species tree with n leaves, denote the time of the interior vertex of rank i by s_i. Define $s := (s_1, s_2, \ldots, s_{n-1})$, where time is zero for the leaves of the tree and it increases going back into the past. For $i \in \{2, \ldots, n-1\}$, denote the time interval between the vertices of rank $i-1$ and i by τ_i (Fig. 1). Interval τ_1 extends infinitely far back into the past. We denote the length of interval τ_i by t_i, $i = 1, 2, \ldots, n-1$. The length t_1 is infinite. We focus on the case in which effective population sizes are identical for all populations within each interval τ_i. Each interval length t_i is measured in *coalescent units*, where for a population with N effective gene copies (corresponding to $N/2$ individuals for a diploid species), the length of the interval in units of generations is Nt_i. Because the effective population size is incorporated into the time t_i, we do not necessarily assume that population sizes are equal in different intervals. In Section 4.3, we discuss a relaxation of the assumption of equal population sizes across populations within an interval.

Ranked gene tree topologies are defined analogously to ranked species tree topologies. For a given species tree topology, a *matching gene tree* is a gene tree that has the same topology as the species tree. A *matching ranked gene tree* is a ranked gene tree that has both the same topology and the same ranking as the species tree; we also say that such a ranked gene tree 'is matching.' A gene tree might have a matching unranked topology but a non-matching ranked topology (e.g., Fig. 1(b)). A ranked gene tree topology $\mathcal{G}$ is *anomalous* for a ranked species tree $\mathcal{T} = (\Psi, s)$ if $\mathbb{P}[\mathcal{G}|\mathcal{T}] > \mathbb{P}[\Psi|\mathcal{T}]$ with $\Psi \neq \mathcal{G}$, where $\mathbb{P}[\cdot]$ represents probability under the multispecies coalescent model (as discussed below). A ranked species tree topology Ψ *produces anomalies* if there exists a vector of speciation times s such that the species tree $\mathcal{T} = (\Psi, s)$ has at least one anomalous ranked gene tree (ARGT). ARGTs represent the analogous concept in the ranked case to anomalous gene trees (AGTs) in the unranked case, where an AGT for species tree $\mathcal{T}$ is an unranked gene tree topology G that has greater probability than the matching gene tree topology ψ – that is, an AGT has $\mathbb{P}[G|\mathcal{T}] > \mathbb{P}[\psi|\mathcal{T}]$, as computed using the formula for probabilities of unranked gene tree topologies conditional on species trees [3, Eq. 12].

Given a gene tree that evolves on a species tree $\mathcal{T}$ with n leaves, we define a *ranked history* of the gene tree as $x = (x_1, x_2, \ldots, x_{n-1})$, where for $i = 1, 2, \ldots, n-1$, $x_i = j$ if the ith coalescence occurs in species tree interval τ_j. Here, the coalescence events are ordered forward in time, so that event i is the ith most ancient coalescence event. We focus on gene trees in which each taxon in the species tree is represented by only one gene lineage. For examples of ranked histories, see Fig. 1.

Let X_n be the set of ranked histories of gene trees on n leaves. This set depends only on n, and it contains the same elements for all species tree topologies. The vector $x \in \mathbb{N}^{n-1}$ is a ranked history of some gene tree topology on a given species tree $\mathcal{T}$ if and only if for all $i \in \{2, \ldots, n-1\}$, $x_{i-1} \leqslant x_i$, and for all $i \in \{1, \ldots, n-1\}$, $x_i \leqslant i$. The condition $x_{i-1} \leqslant x_i$ codifies the criterion that events with higher numbers occur more recently than events with lower numbers. The condition $x_i \leqslant i$ specifies that the number of gene tree lineages surviving until the end of time interval τ_i (the older boundary of the interval) is at least i, as i species exist at the end of this interval.

We define the probability under the multispecies coalescent that a ranked gene tree topology $\mathcal{G}$ evolves with ranked history x on species tree $\mathcal{T}$ to be $\mathbb{P}[\mathcal{G}, x|\mathcal{T}]$. Suppose that backward in time, the events in $(0, s_i)$ are compatible with ranked gene tree $\mathcal{G}$ and ranked history x. Then we let $\mathbb{P}_i[\mathcal{G}, x|\mathcal{T}]$ denote the probability that events in $[s_i, s_{i-1})$ are also compatible with $\mathcal{G}$ and x.

A *partial coalescent* H_n^i is the sequence of coalescences of n lineages to i lineages in a single population, retaining the order of coalescences, but disregarding edge lengths. For fixed n and i, each partial coalescent H_n^i is equally likely, and the number of partial coalescents is [27,28]

$$h_n^i := \frac{n!(n-1)!}{2^{n-i} i!(i-1)!}. \tag{1}$$

3. Counting the number of ranked histories

In this section, we count the number of ranked histories associated with a given ranked gene tree topology and a species tree topology. This computation can be viewed as analogous to the enumeration of the number of coalescent histories in the unranked

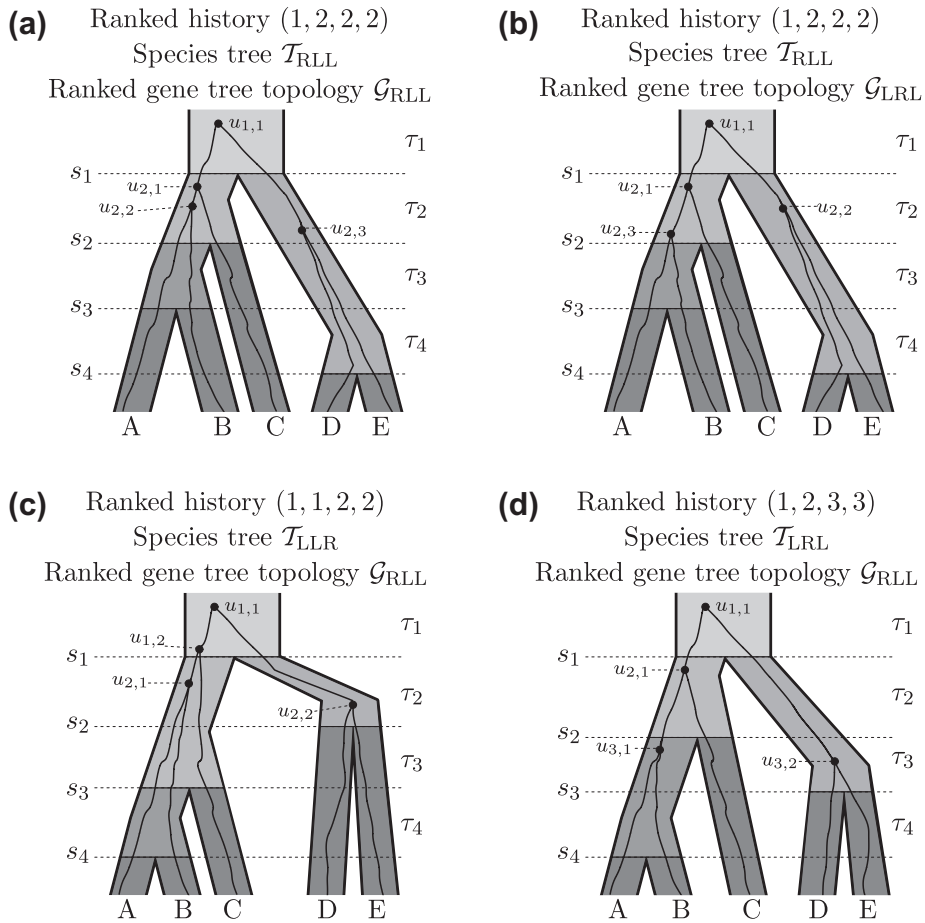

Fig. 1. Gene trees in a species tree, with coalescence times. In each part of the figure, both the species tree and gene tree have unranked topology (((AB)C)(DE)); however, the ranked gene tree topology matches the ranked species tree topology only in (a). The species trees in (b)–(d) each have distinct ranked topologies. We denote these species trees by $\mathcal{T}_{\mathrm{RLL}}$, $\mathcal{T}_{\mathrm{LLR}}$, and $\mathcal{T}_{\mathrm{LRL}}$, respectively (based on the sequence backward in time of 'left' and 'right' species tree coalescences, denoted L and R respectively). In all parts of the figure, the ranked gene tree topology is $\mathcal{G}_{\mathrm{RLL}}$, except in (b), where the ranked gene tree topology is $\mathcal{G}_{\mathrm{LRL}}$. The ranked history in (d) is maximal in the sense that each gene tree coalescence event occurs in the most recent time interval permitted by the combination of the ranked species tree topology and the ranked gene tree topology. For each i, $s_i > 0$ denotes the time of the ith species divergence (using 0 for the present and letting $s_i > s_{i+1}$). The τ_i represent intervals between speciation events, with $\tau_1 = [s_1, \infty)$. For each i and j, $u_{i,j}$ represents the jth coalescences in time interval τ_i (forward in time). The quantities s_i, τ_i, and $u_{i,j}$ are measured in coalescence time units.

case [24–26]. Just as coalescent histories are used in evaluating the probability of an unranked gene tree topology given a species tree, ranked histories are used in evaluating the corresponding probability of a ranked gene tree topology.

Let $\mathcal{G}$ be a ranked gene tree topology on a species tree $\mathcal{T}$ with $n+1$ leaves, and let Y be the set of ranked histories for $\mathcal{G}$. Then $Y \subseteq X_{n+1}$, with equality if and only if $\mathcal{G}$ is the matching ranked gene tree topology. In particular, for any $\mathcal{T}$ with $n+1$ leaves, if $\mathcal{G}$ is a non-matching ranked gene tree topology, then ranked history $(1,2,\ldots,n)$, in which all coalescences occur in the most recent ancestral population allowed by $\mathcal{T}$, is in X_{n+1} but not in Y.

To determine the set Y associated with $\mathcal{G}$ and the ranked species tree topology Ψ of species tree $\mathcal{T}$, consider the ranked history in which each coalescence happens in the time interval τ_j – where j is the maximal possible value for the event, that is, the index for the most recent interval in which the event can occur. Call this ranked history the *maximal ranked history* y^*. Provided that $y \in \mathbb{N}^n$ satisfies $y_{i-1} \leqslant y_i$ for all $i \in \{2,\ldots,n\}$ and $y_i \leqslant i$ for all $i \in \{1,\ldots,n\}$, each y with $y_i \leqslant y_i^*$ for all $i \in \{1,\ldots,n\}$ is also a possible ranked history. Similar to the enumeration of coalescent histories for an unranked gene tree topology and an unranked species tree topology, the number of ranked histories for a *ranked* gene tree topology and a *ranked* species tree topology can be obtained recursively.

Proposition 1. *For a ranked gene tree topology $\mathcal{G}$ and a ranked species tree topology Ψ on $n+1$ leaves, the cardinality of the set of ranked histories, Y, depends only on the maximal ranked history y^* and on n, and it satisfies*

$$|Y| = \sum_{k=1}^{y_n^*} g(n,k),$$

where

$$g(1,1) = 1, \quad g(n,k) = \sum_{j=1}^{\min(k, y_{n-1}^*)} g(n-1,j).$$

Proof. Let $g(n,k)$ be the number of non-decreasing sequences $y = (y_1,\ldots,y_n)$ with $y_i \leqslant y_i^*$ for all i, and $y_n = k$. Then $|Y| = \sum_{k=1}^{y_n^*} g(n,k)$. It remains to establish the recursion for $g(n,k)$. First, because $y_1^* = 1$, we have $g(1,1) = 1$. Second, a sequence of length n with last element k is obtained by appending to a sequence of length $n-1$ that ends with $j \in \{1,\ldots,\min(k, y_{n-1}^*)\}$ an nth element equal to k. □

Corollary 2. *The number of ranked histories on $n+1$ leaves is the nth Catalan number,*

$$|X_{n+1}| = \frac{1}{n+1}\binom{2n}{n}.$$

Proof. The result is obtained by applying Proposition 1 to the maximal ranked history, $y^* = (1,2,\ldots,n)$. Proposition 1 then simplifies to

$$|X_{n+1}| = \sum_{k=1}^{n} g(n,k),$$

where

$$g(1,1) = 1, \quad g(n,k) = \sum_{j=1}^{\min(k,n-1)} g(n-1,j).$$

The recursion for $g(n,k)$ equals the recursion for the Ballot numbers (e.g. [29]). Thus, $|X_{n+1}|$, as the sum of the (n,k)-Ballot numbers over $k = 1,\ldots,n$, is the nth Catalan number. □

Corollary 2 can also be derived by the well-known result that the nth Catalan number is the number of monotonic paths along the edges of an $n \times n$ square grid, where monotonic paths start at the lower left corner, end at the upper right corner, have length $2n$, and do not pass above the diagonal [30]. If we view each step along the horizontal dimension of the grid as an increase in the number of species tree time intervals traversed, and each step along the vertical dimension as a decrease in the number of available gene tree lineages, then the problem of counting monotonic paths is equivalent to the problem in Corollary 2 of counting sequences of length n for which $x_{i-1} \leqslant x_i$ for all $i \in \{2,\ldots,n\}$ and $x_i \leqslant i$ for all $i \in \{1,\ldots,n\}$ [31].

4. Probability of a ranked gene tree on a given species tree

We now derive the probability under the multispecies coalescent model of a ranked gene tree topology on a species tree. This probability, which is analogous to the formula for the probability of an unranked gene tree topology given a species tree [3], can be computed as the sum of the probabilities of the ranked histories for the ranked gene tree topology.

4.1. Special case: at most one gene tree coalescence per species tree time interval

The probability of a ranked history for a given species tree and ranked gene tree topology depends on the interval lengths t_i measured in coalescent units. The ranked history specifies the time intervals in which the various coalescences take place, and therefore, for each time interval, it encodes the number of coalescences occurring in the interval. Because the waiting time to a coalescence for a sample of i lineages from a single population is exponentially distributed with rate $\binom{i}{2}$, the probability that i lineages fail to coalesce in a time interval of length t_i is $e^{-\binom{i}{2}t_i}$. More generally, let $g_{i,j}(t)$ be the probability that i lineages coalesce to $j \leqslant i$ lineages during time t (going backward in time) in a single population, where t is measured in coalescent time units. The probabilities $g_{i,j}(t)$ appear in [32]:

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

The notation $a_{(k)}$ refers to the rising factorial $a_{(k)} = a(a+1)\ldots(a+k-1)$ for $k \geqslant 1$, with $a_{(0)} = 1$, and $a_{[k]}$ refers to the declining factorial $a_{[k]} = a(a-1)\ldots(a-k+1)$ for $k \geqslant 1$, with $a_{[0]} = 1$. We will use the explicit results for small i,

$$\begin{aligned} g_{1,1}(t) &= 1 & g_{2,1}(t) &= 1 - e^{-t} & g_{3,1}(t) &= 1 - \tfrac{3}{2}e^{-t} + \tfrac{1}{2}e^{-3t} \\ & & g_{2,2}(t) &= e^{-t} & g_{3,2}(t) &= \tfrac{3}{2}e^{-t} - \tfrac{3}{2}e^{-3t} \\ & & & & g_{3,3}(t) &= e^{-3t}. \end{aligned} \tag{3}$$

Also, for any $i > 0, g_{i,i}(t) = e^{-\binom{i}{2}t}$, where $\binom{1}{2} = 0$.

The $g_{i,j}(t)$ functions can be used in computing the probability of the events in a particular time interval if coalescences occur in at most one branch in the interval. In many cases, such interval probabilities are sufficient to determine the full probability of a ranked history. The probability of the ranked gene tree topology $\mathcal{G}$ given the species tree $\mathcal{T}$ can then be determined by summing over ranked histories, for each ranked history multiplying probabilities across intervals in $\mathcal{T}$:

$$\mathbb{P}[\mathcal{G}|\mathcal{T}] = \sum_{x \in Y} \mathbb{P}[\mathcal{G}, x|\mathcal{T}] = \sum_{x \in Y} \prod_{i=1}^{n-1} \mathbb{P}_i[\mathcal{G}, x|\mathcal{T}]. \tag{4}$$

4.2. General case: arbitrarily many gene tree coalescences per species tree time interval

When coalescences occur in multiple populations during the same time interval, the probability of the events in the interval generally cannot be written directly as a product of $g_{i,j}(t)$ functions. However, this probability can instead be obtained by integrating over the joint density of coalescence times. We now determine $\mathbb{P}_i[\mathcal{G}, x|\mathcal{T}]$, the probability that the coalescences in time interval τ_i are compatible with the ranked gene tree topology $\mathcal{G}$ and ranked history x given that the events in τ_j for all $j > i$ are also compatible with $\mathcal{G}$ and x. We first describe the joint density of coalescence times in the interval using an approach similar to that of Rannala and Yang [33]. The probability of the events in the interval is then obtained by integrating over this joint density.

Recall that τ_i is the time interval with endpoints s_{i-1} and s_i (Fig. 1). During the interval τ_i, i branches are present in the species tree. Let m_i be the number of coalescences in interval τ_i. Note that m_i depends on the ranked history x. Assume that coalescence event j in interval τ_i occurs at time $u_{i,j}$ on branch $b_{i,j}$, $1 \leqslant j \leqslant m_i$. The events are ordered such that $u_{i,j} > u_{i,j+1}$ (events with smaller j happen farther back in time). Define $u_{i,0} := s_{i-1} > u_{i,1}$ and $u_{i,m_i+1} := s_i < u_{i,m_i}$. These constraints can be written as

$$\begin{aligned} s_{i-1} = u_{i,0} > u_{i,1} &= s_i, & \text{if } m_i &= 0 \\ s_{i-1} = u_{i,0} > u_{i,1} > \cdots > u_{i,m_i} > u_{i,m_i+1} &= s_i, & \text{if } m_i &> 0. \end{aligned} \tag{5}$$

We denote the number of gene lineages on branch z just after the jth coalescence (going forward in time) in τ_i by $k_{i,j,z}$. This quantity is the number of lineages available to coalesce in interval τ_i in population z ($z = 1,2,\ldots,i$) at the jth coalescence in the population. For $j = 0$, we interpret $k_{i,j,z}$ as the number of lineages on branch z at the boundary of intervals τ_i and τ_{i-1} (the number of lineages exiting branch z in interval τ_i).

We write the joint density for an interval τ_i with $m_i \geqslant 0$ coalescence events by subdividing the interval into $m_i + 1$ subintervals and multiplying the densities contributed by the subintervals. If $m_i = 0$, then no coalescences occur in the interval, and $t_i = u_{i,0} - u_{i,1} = s_{i-1} - s_i$. Branch z then contributes a term

$$g_{k_{i,0,z},k_{i,0,z}}(u_{i,0} - u_{i,1}) = e^{-\binom{k_{i,0,z}}{2}(u_{i,0}-u_{i,1})}, \tag{6}$$

the probability from Eq. (2) that no coalescences occur on branch z, where $k_{i,0,z}$ is the number of lineages in interval τ_i on branch z. If

coalescence events do occur in interval τ_i, then exactly one coalescence occurs on one branch of a subinterval, and no coalescences occur in all other branches of the subinterval. Thus, the density for the subinterval includes the density for an exponential distribution whose rate depends on the number of lineages in the branch with the coalescence event. This exponential density is multiplied by probabilities that other branches do not have coalescences in that subinterval. If branch z in population i has no coalescences in the jth subinterval, then it contributes a term

$$g_{k_{i,j,z},k_{i,j,z}}(u_{i,j}-u_{i,j+1}) = e^{-\binom{k_{i,j,z}}{2}(u_{i,j}-u_{i,j+1})} \tag{7}$$

to the density in the jth subinterval. If the number of lineages in a branch $b_{i,j}$ at the jth coalescence event is $k_{i,j,b_{i,j}}$, then the waiting time for this event is exponentially distributed with rate $\binom{k_{i,j,b_{i,j}}}{2}$. The event therefore contributes a term

$$\binom{k_{i,j,b_{i,j}}}{2} e^{-\binom{k_{i,j,b_{i,j}}}{2}(u_{i,j}-u_{i,j+1})}, \tag{8}$$

which must be multiplied by the probability $1\Big/\binom{k_{i,j,b_{i,j}}}{2}$ that two particular lineages are chosen to coalesce, as determined by the ranked gene tree topology. Finally, the full joint density of the coalescence times and coalescence sequence in interval τ_i is

$$\begin{aligned} f_i(u_{i,1},\ldots,u_{i,m_i}) &= \prod_{j=0}^{m_i}\left[e^{-\binom{k_{i,j,b_{i,j}}}{2}(u_{i,j}-u_{i,j+1})}\prod_{z=1,z\neq b_{i,j}}^{i} g_{k_{i,j,z},k_{i,j,z}}(u_{i,j}-u_{i,j+1})\right] \\ &= \prod_{j=0}^{m_i}\left[e^{-\binom{k_{i,j,b_{i,j}}}{2}(u_{i,j}-u_{i,j+1})}\prod_{z=1,z\neq b_{i,j}}^{i} e^{-\binom{k_{i,j,z}}{2}(u_{i,j}-u_{i,j+1})}\right] \\ &= \prod_{j=0}^{m_i}\prod_{z=1}^{i} e^{-\binom{k_{i,j,z}}{2}(u_{i,j}-u_{i,j+1})}, \end{aligned} \tag{9}$$

where $\binom{1}{2}=0$. Because the outer product starts at $j = 0$, this density accounts for the probability that no coalescences occur on any branch more anciently than the most ancient coalescence in τ_i, including on the branch with the most ancient event, $b_{i,1}$.

The probability of the events in interval τ_i is obtained by integrating over the density f_i.

$$\mathbb{P}_i[\mathcal{G},x|\mathcal{T}] = \int_{s_i}^{u_{i,0}}\int_{s_i}^{u_{i,1}}\cdots\int_{s_i}^{u_{i,m_i-1}} f_i(u_{i,1},\ldots,u_{i,m_i})\,du_{i,m_i}\cdots du_{i,2}\,du_{i,1}, \tag{10}$$

where Eq. (5) determines the limits of integration. For the interval above the root, $\tau_1 = [s_1,\infty)$, all lineages coalesce eventually, in random order. Hence, for this interval, the integral in Eq. (10) need not be explicitly evaluated. If the number of lineages above the root is m_1, then by Eq. (1), $\mathbb{P}_1[\mathcal{G},x|\mathcal{T}] = 1/h^1_{m_1}$.

Writing the m_i, $k_{i,j,z}$ and $u_{i,j}$ terms as functions of the ranked history x, the total probability of a ranked gene tree topology $\mathcal{G}$ with ranked history set Y on a species tree $\mathcal{T}$ is

$$\begin{aligned} \mathbb{P}[\mathcal{G}|\mathcal{T}] = \sum_{x\in Y}\frac{1}{h^1_{m_1}(x)}\prod_{i=2}^{n-1}\int_{s_i}^{u_{i,0}(x)}\cdots\int_{s_i}^{u_{i,m_i-1}(x)}\prod_{j=0}^{m_i(x)}\prod_{z=1}^{i} \\ \times\exp\left[-\binom{k_{i,j,z}(x)}{2}(u_{i,j}(x)-u_{i,j+1}(x))\right]du_{i,m_i}(x)\cdots du_{i,1}(x). \end{aligned} \tag{11}$$

4.3. Unequal effective population sizes in the species tree

Thus far, we have assumed that all branches within a time interval use the same effective population size. To allow population sizes to differ within intervals, the density in an interval in Eq. (10) can be written by separating the species divergence times in generations from the population sizes of branches within the interval. In particular, measuring the $u_{i,j}$ in generations rather than coalescent time units, and letting $N_{i,z}$ be the effective size for branch z in interval τ_i, the density in Eq. (9) can be rewritten

$$\prod_{j=0}^{m_i}\prod_{z=1}^{i} e^{-\binom{k_{i,j,z}}{2}\frac{u_{i,j}-u_{i,j+1}}{N_{i,z}}}. \tag{12}$$

Integration over the joint density follows Eq. (10), interpreting the speciation times s_i in generations rather than in coalescent time units. Eq. (12) reduces to Eq. (9) when $N_{i,z}$ has the same value, N_i, for each $z = 1,\ldots,i$.

5. Ranked gene tree probabilities for small numbers of taxa

In this section, by direct computation and using results from Section 4, we calculate probabilities of ranked gene tree topologies for species trees with three, four and five taxa. We then demonstrate the existence of anomalous ranked gene trees (ARGTs) in the five-taxon case.

5.1. Three taxa

The three-taxon case is the smallest case for which comparisons of gene trees and species trees are non-trivial. Each of the three possible topologies has only one ranking, and the ranked and unranked cases are therefore equivalent. Given a species tree with topology ((AB)C) and internal branch $t > 0$ coalescent time units, gene tree topology ((AB)C) has probability $1-(2/3)e^{-t}$, and gene tree topologies ((AC)B) and ((BC)A) each have probability $(1/3)e^{-t}$ [4,7,10]. Because $1-(2/3)e^{-t} > (1/3)e^{-t}$, the most probable gene tree topology is always the topology that matches the species tree topology, and neither anomalous gene trees nor anomalous ranked gene trees exist.

5.2. Four taxa

Although probabilities of ranked four-taxon gene tree topologies can be obtained with the general method of Section 4.2, they can also be computed without explicit integration. If all coalescences in a time interval τ_k occur on a single branch of the species tree (among the k branches extant during the interval), then values of $g_{i,j}(t_k)$ can be used for the probabilities of events on the various branches, where i is the number of lineages 'entering' a branch b and j is the number of lineages on branch b after all $i-j$ coalescences have occurred. A branch with no coalescences in interval τ_k contributes probability $g_{i,i}(t_k) = e^{-\binom{i}{2}t_k}$. For example, if the species tree $\mathcal{T}$ has ranked topology $((AB)_3(CD)_2)$ and the ranked gene tree topology is the asymmetric tree $\mathcal{G}_A = (((CD)_3B)_2A)$ (see Fig. 2 for notation), then the possible ranked histories are (1,1,1) and (1,1,2), with probabilities

$$\mathbb{P}[\mathcal{G}_A,(1,1,1)|\mathcal{T}] = g_{2,2}(t_3)g_{2,2}(t_2)g_{2,2}(t_2)\frac{1}{h^1_4} = e^{-t_3-2t_2}\frac{1}{18},$$

$$\mathbb{P}[\mathcal{G}_A,(1,1,2)|\mathcal{T}] = g_{2,2}(t_3)g_{2,1}(t_2)g_{2,2}(t_2)\frac{1}{h^1_3} = e^{-t_3-t_2}(1-e^{-t_2})\frac{1}{3}.$$

The probability of the ranked gene tree topology is the sum of these two probabilities (see also Table 1 in Appendix A). Because caterpil-

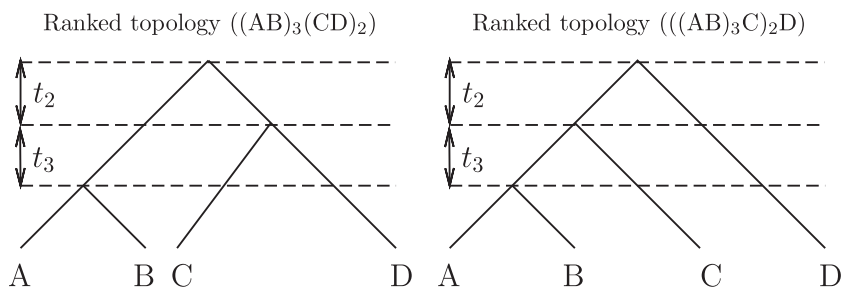

Fig. 2. Notation for four-taxon species tree topologies.

lar tree topologies—which possess one interior vertex descended from all other interior vertices—have only one ranking, the ranked topology $(((CD)_3B)_2A)$ can be written unambiguously as (((CD)B)A), and the probability of a caterpillar ranked gene tree topology is the same as the probability of the caterpillar unranked gene tree topology.

For unranked gene tree topologies, the species tree topology (((AB)C)D) has three AGTs [1]. To search for ARGTs in the ranked case, for each ranked species tree, in Tables 1 and 2 in Appendix A, we exhaustively list the probabilities of four-taxon ranked gene tree topologies given ranked species tree topologies $((AB)_3(CD)_2)$ and $(((AB)_3C)_2)D)$, respectively. In both the symmetric (Table 1) and asymmetric cases (Table 2), the matching ranked gene tree topology is the most probable. Therefore, no ARGTs occur for four-taxon species trees.

We now examine a four-taxon scenario that will be useful for understanding the five-taxon case. Consider a species tree $\mathcal{T}$ with ranked topology $((AB)_3(CD)_2)$ (Fig. 2(a)) when the ranked gene tree is the matching symmetric gene tree $\mathcal{G}_S$. For the ranked history (1,2,2), coalescences occur independently in the same interval, τ_2, along two separate species tree branches. The probability of this ranked history can be computed by integrating over the joint density of the coalescence times. Interval τ_2 has two coalescences and two populations. Hence $k_{2,j,z}$ is defined for $j = 0,1,2$ and $z = 1,2$. Then

$$k_{2,0,1} = 1, \quad k_{2,1,1} = 1, \quad k_{2,2,1} = 2,$$
$$k_{2,0,2} = 1, \quad k_{2,1,2} = 2, \quad k_{2,2,2} = 2,$$

where the branch ancestral to A and B is branch $z = 1$. Note that if $k_{i,j,z} = 1$, then $\binom{k_{i,j,z}}{2} = 0$. Thus, only cases in which $k_{i,j,z} > 1$ need to be considered in writing the density from Eq. (9). Using $s_1 = u_{2,0}$, $s_2 = u_{2,3}$, and $t_2 = s_1 - s_2$, and following Eq. (10) to obtain the probability for events in interval τ_2 yields

$$\begin{aligned}
&\mathbb{P}[\mathcal{G}_S, (1,2,2)|\mathcal{T}] \\
&= g_{2,2}(t_3) \int_{s_2}^{u_{2,0}} \int_{s_2}^{u_{2,1}} e^{-\binom{k_{2,1,2}}{2}(u_{2,1}-u_{2,2})} \\
&\quad \times e^{-\binom{k_{2,2,1}}{2}(u_{2,2}-u_{2,3})} e^{-\binom{k_{2,2,2}}{2}(u_{2,2}-u_{2,3})} \, du_{2,2} \, du_{2,1} \\
&= g_{2,2}(t_3) \int_{s_2}^{s_1} \int_{s_2}^{u_{2,1}} e^{-(u_{2,1}-u_{2,2})} e^{-2(u_{2,2}-s_2)} \, du_{2,2} \, du_{2,1} \\
&= g_{2,2}(t_3) \frac{1}{2}(1 - 2e^{-t_2} + e^{-2t_2}) = g_{2,2}(t_3) \frac{1}{2} [g_{2,1}(t_2)]^2. \qquad (13)
\end{aligned}$$

Eq. (13) can also be obtained by noting that because the two populations in interval τ_2 have equally many lineages, under the model, coalescences occur with equal rates. The two pairs coalesce independently, each pair with probability $g_{2,1}(t_2)$. The events occur in one of two possible sequences (A and B coalesce either more recently or less recently than do C and D). By symmetry, the probability is 1/2 that the sequence of coalescences follows the ranking of the gene tree.

5.3. Five taxa

For species trees with five or more taxa, separate populations can have different non-zero numbers of coalescences in the same time interval. The coalescences in these different populations can occur at different rates, and the symmetry argument in Section 5.2 no longer holds. In this case, the probability of a ranked history can be obtained by the method in Section 4.2. This section illustrates the computation of probabilities of example ranked histories for the case that the species tree is $\mathcal{T}_{\mathrm{RLL}}$ and the ranked gene tree topology is either $\mathcal{G}_{\mathrm{RLL}}$ or $\mathcal{G}_{\mathrm{LRL}}$ (Fig. 1(a) and (b)). If for a ranked history, at most one population in an interval τ_i has coalescences, then the probability of the events in that interval can be determined using the $g_{i,j}(t)$ functions. This approach is useful for computing the probabilities of all ranked histories for species tree $\mathcal{T}_{\mathrm{RLL}}$ and ranked gene tree topologies $\mathcal{G}_{\mathrm{RLL}}$ and $\mathcal{G}_{\mathrm{LRL}}$, except for ranked histories (1,2,2,2) and (1,1,2,2). For these ranked histories, the probability of the events in interval τ_2 is found by integration as in Eq. (9).

As we will see, ranked histories (1,2,2,2) and (1,1,2,2) illustrate the fact that the probability of a ranked history can be greater for a non-matching ranked gene tree topology than for a matching ranked gene tree topology. Probabilities of all ranked histories for the ranked gene tree topologies $\mathcal{G}_{\mathrm{RLL}}, \mathcal{G}_{\mathrm{LRL}}, \mathcal{G}_{\mathrm{LLR}}$, and ((((AB)C)D)E) given ranked species tree topology $(((AB)_3C)_2(DE)_4)$ appear in Tables 3 and 4 in Appendix A. These probabilities are used in Section 5.4 for identifying ARGTs. Computations for example ranked histories are illustrated in Sections 5.3.1 and 5.3.2.

5.3.1. Species tree $\mathcal{T}_{\mathrm{RLL}}$, ranked gene tree topology $\mathcal{G}_{\mathrm{RLL}}$

For an n-taxon species tree, when the ranked gene tree is matching, the maximal ranked history is $(1,2,\ldots,n-1)$. For five taxa, with species tree $\mathcal{T}_{\mathrm{RLL}}$ and ranked gene tree topology $\mathcal{G}_{\mathrm{RLL}}$, in the maximal ranked history $x = (1,2,3,4)$, the D and E lineages coalesce in τ_4, the A and B lineages coalesce in τ_3, and the C lineage coalesces in τ_2 with the ancestor of the A and B lineages. All other ranked histories for this ranked gene tree topology and species tree are obtained by decreasing one or more of the x_i while ensuring that they satisfy $1 \leqslant x_i \leqslant i$ and $x_{i-1} \leqslant x_i$. By Proposition 1, 14 such ranked histories exist. We evaluate the probabilities of three of these ranked histories; the full set of 14 probabilities appears in Table 3 in Appendix A.

The probability of the ranked history (1,1,1,1) is the probability that no coalescences occur more recently than the species tree root. From Eq. (1), the particular sequence of coalescences above the root has probability $1/h_5^1 = 1/180$. Thus,

$$\mathbb{P}[\mathcal{G}_{\mathrm{RLL}}, (1,1,1,1)|\mathcal{T}_{\mathrm{RLL}}] = g_{2,2}(t_4) g_{2,2}(t_3) g_{2,2}(t_3) g_{3,3}(t_2) g_{2,2}(t_2) \frac{1}{180}.$$

Any five-taxon combination of a ranked gene tree topology and ranked species tree topology has ranked history (1,1,1,1), whether

or not the topologies match. The probability of $(1,1,1,1)$ depends on the ranked species tree topology, but is the same for each ranked gene tree topology given a particular ranked species tree topology.

The probability of ranked history (1,2,3,4) is the probability that each pair of lineages coalesces in the first time interval in which both lineages lie in the same ancestral population:

$$\mathbb{P}[\mathcal{G}_{\mathrm{RLL}},(1,2,3,4)|\mathcal{T}_{\mathrm{RLL}}] = g_{2,1}(t_4)g_{2,1}(t_3)g_{2,1}(t_2).$$

For ranked history (1,2,2,2) (Fig. 1(a)), two populations have coalescences in time interval τ_2; however, the coalescence rates differ. Hence, for this interval, we use the approach in Section 4.2 and integrate over the joint density of coalescence times. Because interval τ_2 has two populations and three coalescences, the values of $k_{i,j,z}$ used to obtain the density for events in interval τ_2 are

$$k_{2,0,1}=1,\quad k_{2,1,1}=2,\quad k_{2,2,1}=3,\quad k_{2,3,1}=3,$$
$$k_{2,0,2}=1,\quad k_{2,1,2}=1,\quad k_{2,2,2}=1,\quad k_{2,3,2}=2.$$

Recalling that $u_{2,0}=s_1$, $u_{2,4}=s_2$, and $s_{i-1}-s_i=t_i$, we use Eq. (10) for interval τ_2 to obtain

$$\begin{aligned}
&\mathbb{P}[\mathcal{G}_{\mathrm{RLL}},(1,2,2,2)|\mathcal{T}_{\mathrm{RLL}}]\\
&\quad= g_{2,2}(t_4)g_{2,2}(t_3)g_{2,2}(t_3)\int_{s_2}^{u_{2,0}}\int_{s_2}^{u_{2,1}}\int_{s_2}^{u_{2,2}}\\
&\qquad\times\left[e^{-\binom{k_{2,1,1}}{2}(u_{2,1}-u_{2,2})}\times e^{-\binom{k_{2,2,1}}{2}(u_{2,2}-u_{2,3})}e^{-\binom{k_{2,3,1}}{2}(u_{2,3}-s_2)}\right.\\
&\qquad\left.\times e^{-\binom{k_{2,3,2}}{2}(u_{2,3}-s_2)}\right]du_{2,3}\,du_{2,2}\,du_{2,1}\\
&\quad= e^{-t_4-2t_3}\left(\frac{1}{12}-\frac{1}{6}e^{-t_2}+\frac{1}{6}e^{-3t_2}-\frac{1}{12}e^{-4t_2}\right).
\end{aligned}\tag{14}$$

5.3.2. Species tree $\mathcal{T}_{\mathrm{RLL}}$, ranked gene tree topology $\mathcal{G}_{\mathrm{LRL}}$ or $\mathcal{G}_{\mathrm{LLR}}$

For species tree $\mathcal{T}_{\mathrm{RLL}}$, the $\mathcal{G}_{\mathrm{LRL}}$ and $\mathcal{G}_{\mathrm{LLR}}$ ranked gene tree topologies have nine and seven ranked histories, respectively (Table 4, Appendix A). The sets of ranked histories are subsets of the 14 ranked histories for the $\mathcal{G}_{\mathrm{RLL}}$ ranked gene tree topology (Table 3, Appendix A). Probabilities of ranked histories that can arise on both non-matching ranked gene tree topologies, such as (1,2,2,2), are the same for $\mathcal{G}_{\mathrm{LRL}}$ and $\mathcal{G}_{\mathrm{LLR}}$ when the species tree is $\mathcal{T}_{\mathrm{RLL}}$, but they differ from the probability of the ranked history (1,2,2,2) for the matching ranked gene tree topology. For $\mathcal{G}_{\mathrm{LRL}}$ and $\mathcal{G}_{\mathrm{LLR}}$, with ranked history (1,2,2,2), the most recent coalescence in interval τ_2 occurs in a population with three rather than two lineages, that is, in a population with a faster coalescence rate. In particular, the probability of ranked history (1,2,2,2) for the $\mathcal{G}_{\mathrm{LRL}}$ and $\mathcal{G}_{\mathrm{LLR}}$ ranked gene tree topologies is

$$\begin{aligned}
&\mathbb{P}[\mathcal{G}_{\mathrm{LRL}},(1,2,2,2)|\mathcal{T}_{\mathrm{RLL}}]\\
&\quad= g_{2,2}(t_4)g_{2,2}(t_3)g_{2,2}(t_3)\int_{s_2}^{u_{2,0}}\int_{s_2}^{u_{2,1}}\int_{s_2}^{u_{2,2}}\\
&\qquad\times\left[e^{-(u_{2,1}-u_{2,2})}e^{-(u_{2,2}-u_{2,3})}\times e^{-(u_{2,2}-u_{2,3})}e^{-3(u_{2,3}-s_2)}e^{-(u_{2,3}-s_2)}\right]\\
&\qquad\times du_{2,3}\,du_{2,2}\,du_{2,1}\\
&\quad= e^{-t_4-2t_3}\left(\frac{1}{8}-\frac{1}{3}e^{-t_2}+\frac{1}{4}e^{-2t_2}-\frac{1}{24}e^{-4t_2}\right).
\end{aligned}\tag{15}$$

This quantity is never less than the probability of (1,2,2,2) for the matching ranked gene tree topology, from Eq. (14):

$$\begin{aligned}
&\mathbb{P}[\mathcal{G}_{\mathrm{LRL}},(1,2,2,2)|\mathcal{T}_{\mathrm{RLL}}]-\mathbb{P}[\mathcal{G}_{\mathrm{RLL}},(1,2,2,2)|\mathcal{T}_{\mathrm{RLL}}]\\
&\quad=\frac{1}{24}e^{-t_4-2t_3}\left(1-e^{-t_2}\right)^4.
\end{aligned}\tag{16}$$

For fixed t_4 and t_3, the difference in Eq. (16) approaches zero as $t_2\to 0$ and is strictly positive for $t_2, t_3, t_4>0$. Hence, for species tree $\mathcal{T}_{\mathrm{RLL}}$, ranked history (1, 2, 2, 2) is more probable for the $\mathcal{G}_{\mathrm{LRL}}$ and $\mathcal{G}_{\mathrm{LLR}}$ ranked gene tree topologies than for the matching ranked gene tree topology. Similarly,

$$\begin{aligned}
&\mathbb{P}[\mathcal{G}_{\mathrm{LRL}},(1,1,2,2)|\mathcal{T}_{\mathrm{RLL}}]-\mathbb{P}[\mathcal{G}_{\mathrm{RLL}},(1,1,2,2)|\mathcal{T}_{\mathrm{RLL}}]\\
&\quad=\frac{e^{-t_4-2t_3-t_2}(1-e^{-t_2})^3}{18},\\
&\mathbb{P}[\mathcal{G}_{\mathrm{LRL}},(1,1,1,2)|\mathcal{T}_{\mathrm{RLL}}]-\mathbb{P}[\mathcal{G}_{\mathrm{RLL}},(1,1,1,2)|\mathcal{T}_{\mathrm{RLL}}]\\
&\quad=\frac{e^{-t_4-2t_3-2t_2}(1-e^{-t_2})^2}{36}.
\end{aligned}\tag{17}$$

These probability differences exceed 0 for $t_2, t_3, t_4>0$. Thus, ranked histories (1,1,1,2), (1,1,2,2), and (1,2,2,2) are each always more probable for non-matching ranked gene tree topologies $\mathcal{G}_{\mathrm{LRL}}$ and $\mathcal{G}_{\mathrm{LLR}}$ than for $\mathcal{G}_{\mathrm{RLL}}$. For these ranked histories, the greater probability for the non-matching ranked gene tree topologies occurs because the faster coalescence rate of $\binom{3}{2}=3$ in the left-hand population compared to $\binom{2}{2}=1$ in the right-hand population increases the probability that the first coalescence involves left-descendants of the root.

5.4. Anomalous ranked gene trees with five taxa

The observation in Section 5.3.2 that for species tree $\mathcal{T}_{\mathrm{RLL}}$, the probability of ranked history (1,2,2,2) is larger for a non-matching ranked gene tree topology than for the matching ranked gene tree topology can be used to construct ARGTs. In particular, branch lengths t_2, t_3, and t_4 can be chosen such that most of the probability of the ranked gene tree topology is concentrated on this ranked history. By setting t_2 large and t_3 and t_4 small, all available lineages in τ_2 coalesce on this interval with high probability, and lineages in τ_3 and τ_4 coalesce with probability close to 0. Because no lineages from species A, B, and C coalesce until their most recent common ancestral population in interval τ_2, each possible pair of coalescences – A with B, A with C, and B with C – is equally likely. Moreover, because t_3 and t_4 are small, it is unlikely that D and E coalesce more recently than the first coalescence among A, B, and C in interval τ_2. For the matching ranked gene tree topology, $\mathcal{G}_{\mathrm{RLL}}$, following Eq. (14), the limiting probability for ranked history (1,2,2,2) when $t_2\to\infty$ and $t_3,t_4\to 0$ is 1/12. However, from Eq. (15), the limiting value for ranked history (1,2,2,2) when the ranked gene tree topology is $\mathcal{G}_{\mathrm{LRL}}$ is 1/8.

Because $g_{i,j}(t)\to 0$ as $t\to 0$ for $j<i$, and $g_{i,i}(t_2)\to 0$ as $t_2\to\infty$ for any $i\geqslant 2$, all ranked histories other than (1,2,2,2) have probabilities near 0 for large t_2 and small t_3 and t_4. Thus, the probabilities $\mathbb{P}[\mathcal{G}_{\mathrm{RLL}}|\mathcal{T}_{\mathrm{RLL}}]$ and $\mathbb{P}[\mathcal{G}_{\mathrm{LRL}}|\mathcal{T}_{\mathrm{RLL}}]$ can be made arbitrarily close to their respective joint probabilities, $\mathbb{P}[\mathcal{G}_{\mathrm{RLL}},(1,2,2,2)|\mathcal{T}_{\mathrm{RLL}}]$ and $\mathbb{P}[\mathcal{G}_{\mathrm{LRL}},(1,2,2,2)|\mathcal{T}_{\mathrm{RLL}}]$. For small t_3 and t_4, as t_2 grows large, the probability eventually grows large for history (1,2,2,2). Because this history is more probable for $\mathcal{G}_{\mathrm{LRL}}$ than for $\mathcal{G}_{\mathrm{RLL}}$ (Eq. (16)), the matching ranked gene tree topology $\mathcal{G}_{\mathrm{RLL}}$ can be less probable than $\mathcal{G}_{\mathrm{LRL}}$ (and similarly, $\mathcal{G}_{\mathrm{LLR}}$) for a large range of values of t_2 (Fig. 3).

The region of branch length space in which species tree $\mathcal{T}_{\mathrm{RLL}}$ has an ARGT – the *ranked anomaly zone* – can be identified by finding the set of values of t_2, t_3, and t_4 for which the probability of some non-matching ranked gene tree topology exceeds the probability of the matching ranked gene tree topology. Fig. 3(a) and (b) depict values of t_2, t_3, and t_4 for which the two non-matching ranked gene tree topologies with the same unranked topology, (((AB)C)(DE)), are more probable than the matching ranked gene tree topology.

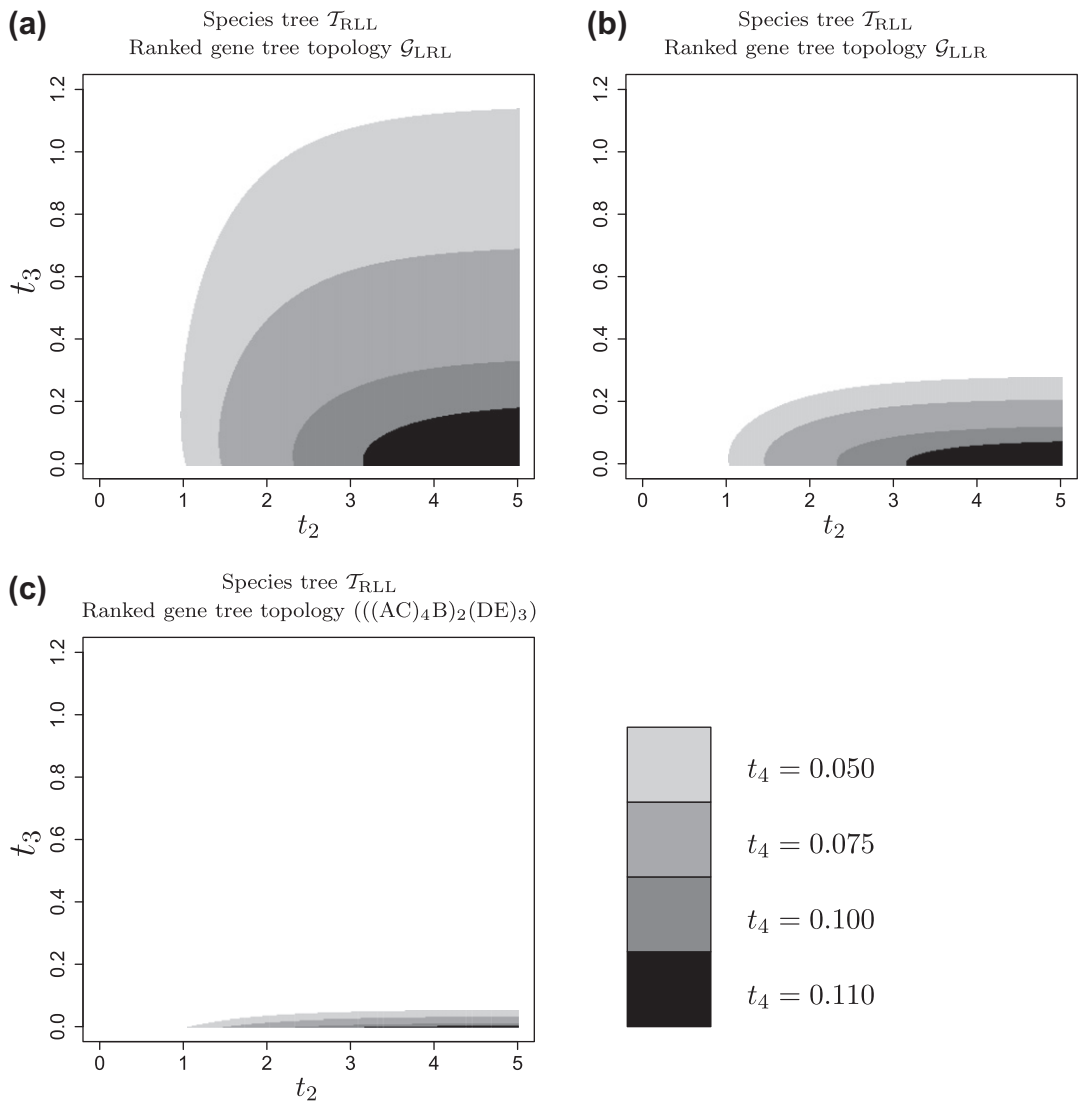

Fig. 3. Slices of the ranked anomaly zone for the species tree $\mathcal{T}_{\mathrm{RLL}}$ and various ranked gene tree topologies. For fixed values of t_4, each shaded region represents the set of points (t_2,t_3) for which the given ranked gene tree topology is more probable than the matching ranked gene tree topology, $\mathcal{G}_{\mathrm{RLL}}$. The color scale applies to all three panels. Each panel was generated by computing probabilities of the matching and alternate ranked gene tree topologies on a grid with $t_2 \in [0.0, 5.0]$ with increments of 5.0/350 and $t_3 \in [0.0, 1.2]$ with increments of 1.2/350. For shaded points, ARGTs exist. In all cases, more darkly shaded regions are subsets of lighter regions. (a) Regions where the ranked gene tree topology $\mathcal{G}_{\mathrm{LRL}}$ is an ARGT ($\mathbb{P}[\mathcal{G}_{\mathrm{LRL}}|\mathcal{T}_{\mathrm{RLL}}] > \mathbb{P}[\mathcal{G}_{\mathrm{RLL}}|\mathcal{T}_{\mathrm{RLL}}]$). (b) Regions where $\mathcal{G}_{\mathrm{LLR}}$ is an ARGT for $\mathcal{T}_{\mathrm{RLL}}$. (c) Regions where $(((\mathrm{AC})_4\mathrm{B})_2(\mathrm{DE})_3)$ is an ARGT for $\mathcal{T}_{\mathrm{RLL}}$.

Fig. 3(c) depicts values for which a ranked gene tree topology with a different unranked topology, $(((\mathrm{AC})_4\mathrm{B})_2(\mathrm{DE})_3)$, has higher probability than the matching ranked gene tree topology. From Table 4 in Appendix A, the probability of ranked gene tree topology $(((\mathrm{AC})_4\mathrm{B})_2(\mathrm{DE})_3)$ is a sum of a subset of the terms in the expression for $\mathbb{P}[\mathcal{G}_{\mathrm{LRL}}|\mathcal{T}_{\mathrm{RLL}}]$:

$$\begin{aligned}\mathbb{P}[(((\mathrm{AC})_4\mathrm{B})_2(\mathrm{DE})_3)] = {} & \mathbb{P}[\mathcal{G}_{\mathrm{LRL}}, (1,1,1,1)|\mathcal{T}_{\mathrm{RLL}}] \\ & + \mathbb{P}[\mathcal{G}_{\mathrm{LRL}}, (1,1,1,2)|\mathcal{T}_{\mathrm{RLL}}] \\ & + \mathbb{P}[\mathcal{G}_{\mathrm{LRL}}, (1,1,2,2)|\mathcal{T}_{\mathrm{RLL}}] \\ & + \mathbb{P}[\mathcal{G}_{\mathrm{LRL}}, (1,2,2,2)|\mathcal{T}_{\mathrm{RLL}}].\end{aligned}$$

By symmetry, we have

$$\mathbb{P}[(((\mathrm{BC})_4\mathrm{A})_2(\mathrm{DE})_3)|\mathcal{T}_{\mathrm{RLL}}] = \mathbb{P}[(((\mathrm{AC})_4\mathrm{B})_2(\mathrm{DE})_3)|\mathcal{T}_{\mathrm{RLL}}].$$

Although it is possible for the matching ranked gene tree topology to have a smaller probability than that of a non-matching ranked gene tree topology with a different unranked topology, in this example, the set of branch lengths with ARGTs that disagree not only in ranked topology but also in unranked topology is considerably smaller than the set of branch lengths for which ARGTs differ only in ranked topology. For example, when $t_2 = 5.0$ and $t_4 = 0.05$, the maximum value of t_3 for which $\mathbb{P}[(((\mathrm{AC})_4\mathrm{B})_2(\mathrm{DE})_3)|\mathcal{T}_{\mathrm{RLL}}] > \mathbb{P}[\mathcal{G}_{\mathrm{RLL}}|\mathcal{T}_{\mathrm{RLL}}]$ is only $t_3 \approx 0.054$, whereas it is considerably greater in Figs. 3a and b. Because the probability of the ranked gene tree topology $(((\mathrm{AC})_4\mathrm{B})_2(\mathrm{DE})_3)$ is always strictly less than the probability of $\mathcal{G}_{\mathrm{LRL}}$, the most probable ranked gene tree topology still has the same *unranked* topology as the species tree for this example.

The ARGT examples above all have the same unlabeled, unranked topology as the species tree (a five-taxon tree with three leaves on one side and two on the other). We note that it is also possible for an ARGT to have an unlabeled topology that is different from that of the species tree. For example, the ranked gene tree topology ((((AC)B)D)E), whose probability $\mathbb{P}[((((\mathrm{AC})\mathrm{B})\mathrm{D})\mathrm{E})|\mathcal{T}_{\mathrm{RLL}}]$ can be obtained from Table 4 in Appendix A by summing joint probabilities of the ranked gene tree topology with ranked histories (1,1,1,1), (1,1,1,2), and (1,1,2,2), can be an ARGT for the ranked species tree topology $\mathcal{T}_{\mathrm{RLL}}$. An example location in the ranked anomaly zone, at which ((((AC)B)D)E) is an ARGT for $\mathcal{T}_{\mathrm{RLL}}$,

is the point at which the intervals between speciation events have values $(t_2, t_3, t_4) = (0.1, 0.001, 0.0005)$.

6. Discussion

This paper has initiated the study of the relationship between ranked gene trees and species trees, deriving a formula for the probability of a ranked gene tree topology given a species tree, and using that formula to investigate cases with three, four, and five taxa, and to uncover the existence of anomalous ranked gene trees. The results have various connections to previous work on unranked trees, as we discuss below.

6.1. Probabilities of ranked and unranked gene tree topologies

The probability of an unranked gene tree topology is the sum of the probabilities of the ranked gene tree topologies that share the unranked topology. Because our method for calculating probabilities of ranked gene tree topologies differs from the existing method for obtaining probabilities of unranked gene tree topologies (coalescent histories used for the unranked calculations [3,24–26] and ranked histories used for ranked gene trees are not the same objects), the relationship between probabilities for the ranked and unranked cases enables a check on computations of either type of probability. It is straightforward to verify, for example, that for ranked species tree topologies $((AB)_3(CD)_2)$ and $((((AB)_3C)_2)D)$, the probability in [9] of each unranked gene tree topology with two ranked gene tree topologies – ((AB)(CD)), ((AC)(BD)), and ((AD)(BC)) – can be obtained by summing the probabilities of the two associated rankings in Tables 1 and 2, respectively.

More generally, our probabilities of ranked gene tree topologies enable probability computations for any partially ranked gene tree topology, in which the ranks of some but not all of the clades are specified. Given a species tree, the probability of a partially ranked gene tree topology can be calculated by summing the probabilities of all completely ranked gene tree topologies that are compatible with the partial ranking. For example, for the unranked gene tree topology (((AB)C)(DE)), if (AB) has occurred more recently than (DE), but the ranks of (DE) and ((AB)C) are unspecified, then we have a partially ranked gene tree topology, $(((AB)_4C)(DE))$, in which (AB) has rank 4. Because this partial ranking is compatible with two complete rankings, $(((AB)_4C)_2(DE)_3)$ and $(((AB)_4C)_3(DE)_2)$, its probability given the species tree is $\mathbb{P}[(((AB)_4C)(DE))] = \mathbb{P}[(((AB)_4C)_2(DE)_3)] + \mathbb{P}[(((AB)_4C)_3(DE)_2)]$.

6.2. ARGTs and AGTs

The connection between ranked and unranked gene trees helps to explain why ARGTs do not exist for four-taxon species tree topologies that have AGTs. For four-taxon caterpillar species tree topologies, as species tree branch lengths approach zero, the probability distribution of ranked gene tree topologies flattens. In the limit, each ranked gene tree topology has probability 1/18 and no ARGTs occur. When unranked topologies are considered, however, three of the 15 unranked gene tree topologies – ((AB)(CD)), ((AC)(BD)), and ((AD)(BC)) – can each be realized by two different ranked gene tree topologies; hence, their probabilities approach 2/18. Consequently, these topologies can be AGTs if the species tree has one of the 12 remaining unranked topologies.

With five taxa, if interval lengths between speciation times all approach zero, then the ranked gene tree distribution still flattens, and probabilities of ranked gene tree topologies approach $1/h_5^1 = 1/180$. Indeed, when all interval lengths between speciation events are equal in coalescent units, the species tree $\mathcal{T}_{\mathrm{RLL}}$ has no ARGTs. In this case, it is only by lengthening interval τ_2 when τ_3 and τ_4 are short that ARGTs can be made to occur.

This behavior has a parallel for five-taxon unranked gene trees. For unranked species tree topology (((AB)C)(DE)), the species tree must have one long and two short internal branches for AGTs to occur [12]. Thus, although gene tree discordance is often treated as a consequence of short species tree branch lengths, in many cases, the existence of AGTs and ARGTs depends on a mixture of long and short branches. The choice of which branches must be long and which must be short to produce anomalies differs in the ranked and unranked cases.

In the five-taxon anomaly zone for unranked species tree topology (((AB)C)(DE)) [12], the branch from the root to the (DE) subtree is long and the branch from the root to the ((AB)C) subtree is short. In the ranked anomaly zone, however, both of these branches are

Table 1
Probabilities of the 18 ranked gene tree topologies when the species tree has ranked topology $((AB)_3(CD)_2)$.

Ranked gene tree topology	Probability
$((AB)_3(CD)_2)$	$g_{2,1}(t_3)\left[g_{2,1}(t_2)+g_{2,2}(t_2)\frac{1}{3}\right]+g_{2,2}(t_3)\left[g_{2,1}(t_2)g_{2,1}(t_2)\frac{1}{2}+g_{2,1}(t_2)g_{2,2}(t_2)\frac{1}{3}+g_{2,2}(t_2)g_{2,2}(t_2)\frac{1}{18}\right]$
$((AB)_2(CD)_3)$	$g_{2,2}(t_3)\left[g_{2,1}(t_2)g_{2,1}(t_2)\frac{1}{2}+g_{2,2}(t_2)g_{2,1}(t_2)\frac{1}{3}+g_{2,2}(t_2)g_{2,2}(t_2)\frac{1}{18}\right]$
$((AC)_3(BD)_2)$	$g_{2,2}(t_3)g_{2,2}(t_2)g_{2,2}(t_2)\frac{1}{18}$
$((AC)_2(BD)_3)$	$g_{2,2}(t_3)g_{2,2}(t_2)g_{2,2}(t_2)\frac{1}{18}$
$((AD)_3(BC)_2)$	$g_{2,2}(t_3)g_{2,2}(t_2)g_{2,2}(t_2)\frac{1}{18}$
$((AD)_2(BC)_3)$	$g_{2,2}(t_3)g_{2,2}(t_2)g_{2,2}(t_2)\frac{1}{18}$
$(((AB)_3C)_2D)$	$g_{2,1}(t_3)g_{1,1}(t_2)g_{2,2}(t_2)\frac{1}{3}+g_{2,2}(t_3)\left[g_{2,1}(t_2)g_{2,2}(t_2)\frac{1}{3}+g_{2,2}(t_2)g_{2,2}(t_2)\frac{1}{18}\right]$
$(((AB)_3D)_2C)$	$g_{2,1}(t_3)g_{1,1}(t_2)g_{2,2}(t_2)\frac{1}{3}+g_{2,2}(t_3)\left[g_{2,1}(t_2)g_{2,2}(t_2)\frac{1}{3}+g_{2,2}(t_2)g_{2,2}(t_2)\frac{1}{18}\right]$
$(((AC)_3B)_2D)$	$g_{2,2}(t_3)g_{2,2}(t_2)g_{2,2}(t_2)\frac{1}{18}$
$(((AC)_3D)_2B)$	$g_{2,2}(t_3)g_{2,2}(t_2)g_{2,2}(t_2)\frac{1}{18}$
$(((AD)_3B)_2C)$	$g_{2,2}(t_3)g_{2,2}(t_2)g_{2,2}(t_2)\frac{1}{18}$
$(((AD)_3C)_2B)$	$g_{2,2}(t_3)g_{2,2}(t_2)g_{2,2}(t_2)\frac{1}{18}$
$(((BC)_3A)_2D)$	$g_{2,2}(t_3)g_{2,2}(t_2)g_{2,2}(t_2)\frac{1}{18}$
$(((BC)_3D)_2A)$	$g_{2,2}(t_3)g_{2,2}(t_2)g_{2,2}(t_2)\frac{1}{18}$
$(((BD)_3A)_2C)$	$g_{2,2}(t_3)g_{2,2}(t_2)g_{2,2}(t_2)\frac{1}{18}$
$(((BD)_3C)_2A)$	$g_{2,2}(t_3)g_{2,2}(t_2)g_{2,2}(t_2)\frac{1}{18}$
$(((CD)_3A)_2B)$	$g_{2,2}(t_3)\left[g_{2,2}(t_2)g_{2,1}(t_2)\frac{1}{3}+g_{2,2}(t_2)g_{2,2}(t_2)\frac{1}{18}\right]$
$(((CD)_3B)_2A)$	$g_{2,2}(t_3)\left[g_{2,2}(t_2)g_{2,1}(t_2)\frac{1}{3}+g_{2,2}(t_2)g_{2,2}(t_2)\frac{1}{18}\right]$

The $g_{i,j}(t)$ appear in Eq. (3). For each non-matching ranked gene tree topology, the terms in the probability of the topology are subsumed by the terms for the matching ranked gene tree topology. Therefore, the matching ranked gene tree topology has the highest probability, and ranked species tree topology $((AB)_3(CD)_2)$ has no ARGTs.

Table 2
Probabilities of the 18 ranked gene tree topologies when the species tree has ranked topology $(((AB)_3C)_2D)$.

Ranked gene tree topology	Probability
$((AB)_3(CD)_2)$	$g_{2,1}(t_3)g_{2,2}(t_2)\frac{1}{3}+g_{2,2}(t_3)\left[g_{3,2}(t_2)\frac{1}{9}+g_{3,3}(t_2)\frac{1}{18}\right]$
$((AB)_2(CD)_3)$	$g_{2,2}(t_3)g_{3,3}(t_2)\frac{1}{18}$
$((AC)_3(BD)_2)$	$g_{2,2}(t_3)\left[g_{3,2}(t_2)\frac{1}{9}+g_{3,3}(t_2)\frac{1}{18}\right]$
$((AC)_2(BD)_3)$	$g_{2,2}(t_3)g_{3,3}(t_2)\frac{1}{18}$
$((AD)_3(BC)_2)$	$g_{2,2}(t_3)g_{3,3}(t_2)\frac{1}{18}$
$((AD)_2(BC)_3)$	$g_{2,2}(t_3)\left[g_{3,2}(t_2)\frac{1}{9}+g_{3,3}(t_2)\frac{1}{18}\right]$
$(((AB)_3C)_2D)$	$g_{2,1}(t_3)\left[g_{2,1}(t_2)+g_{2,2}(t_2)\frac{1}{3}\right]+g_{2,2}(t_3)\left[g_{3,1}(t_2)\frac{1}{3}+g_{3,2}(t_2)\frac{1}{9}+g_{3,3}(t_2)\frac{1}{18}\right]$
$(((AB)_3D)_2C)$	$g_{2,1}(t_3)g_{2,2}(t_2)\frac{1}{3}+g_{2,2}(t_3)\left[g_{3,2}(t_2)\frac{1}{9}+g_{3,3}(t_2)\frac{1}{18}\right]$
$(((AC)_3B)_2D)$	$g_{2,2}(t_3)\left[g_{3,1}(t_2)\frac{1}{3}+g_{3,2}(t_2)\frac{1}{9}+g_{3,3}(t_2)\frac{1}{18}\right]$
$(((AC)_3D)_2B)$	$g_{2,2}(t_3)\left[g_{3,2}(t_2)\frac{1}{9}+g_{3,3}(t_2)\frac{1}{18}\right]$
$(((AD)_3B)_2C)$	$g_{2,2}(t_3)g_{3,3}(t_2)\frac{1}{18}$
$(((AD)_3C)_2B)$	$g_{2,2}(t_3)g_{3,3}(t_2)\frac{1}{18}$
$(((BC)_3A)_2D)$	$g_{2,2}(t_3)\left[g_{3,1}(t_2)\frac{1}{3}+g_{3,2}(t_2)\frac{1}{9}+g_{3,3}(t_2)\frac{1}{18}\right]$
$(((BC)_3D)_2A)$	$g_{2,2}(t_3)\left[g_{3,2}(t_2)\frac{1}{9}+g_{3,3}(t_2)\frac{1}{18}\right]$
$(((BD)_3A)_2C)$	$g_{2,2}(t_3)g_{3,3}(t_2)\frac{1}{18}$
$(((BD)_3C)_2A)$	$g_{2,2}(t_3)g_{3,3}(t_2)\frac{1}{18}$
$(((CD)_3A)_2B)$	$g_{2,2}(t_3)g_{3,3}(t_2)\frac{1}{18}$
$(((CD)_3B)_2A)$	$g_{2,2}(t_3)g_{3,3}(t_2)\frac{1}{18}$

The $g_{i,j}(t)$ appear in Eq. (3). For each non-matching ranked gene tree topology, the terms in the probability of the topology are subsumed by the terms for the matching ranked gene tree topology. Therefore, the matching ranked gene tree topology has the highest probability, and ranked species tree topology $(((AB)_3C)_2D)$ has no ARGTs.

Table 3
Probabilities of ranked histories for species tree $\mathcal{T}_{RLL}$ and ranked gene tree topology $\mathcal{G}_{RLL}$.

Ranked history	Probability for ranked gene tree topology $\mathcal{G}_{RLL}$
(1,1,1,1)	$g_{2,2}(t_4)g_{2,2}(t_3)g_{2,2}(t_3)g_{3,3}(t_2)g_{2,2}(t_2)\frac{1}{180}$
(1,1,1,2)	$g_{2,2}(t_4)g_{2,2}(t_3)g_{2,2}(t_3)g_{3,3}(t_2)g_{2,1}(t_2)\frac{1}{18}$
(1,1,1,3)	$g_{2,2}(t_4)g_{2,2}(t_3)g_{2,1}(t_3)g_{3,3}(t_2)\frac{1}{18}$
(1,1,1,4)	$g_{2,1}(t_4)g_{2,2}(t_3)g_{3,3}(t_2)\frac{1}{18}$
(1,1,2,2)	$g_{2,2}(t_4)g_{2,2}(t_3)g_{2,2}(t_3)\left(\frac{1}{6}e^{-t_2}-\frac{1}{2}e^{-3t_2}+\frac{1}{3}e^{-4t_2}\right)\frac{1}{3}$
(1,1,2,3)	$g_{2,2}(t_4)g_{2,2}(t_3)g_{2,1}(t_3)\frac{1}{3}g_{3,2}(t_2)\frac{1}{3}$
(1,1,2,4)	$g_{2,1}(t_4)g_{2,2}(t_3)\frac{1}{3}g_{3,2}(t_2)\frac{1}{3}$
(1,1,3,3)	$g_{2,2}(t_4)\frac{1}{2}g_{2,1}(t_3)g_{2,1}(t_3)g_{2,2}(t_2)\frac{1}{3}$
(1,1,3,4)	$g_{2,1}(t_4)g_{2,1}(t_3)g_{2,2}(t_2)\frac{1}{3}$
(1,2,2,2)	$g_{2,2}(t_4)g_{2,2}(t_3)g_{2,2}(t_3)\left(\frac{1}{12}-\frac{1}{6}e^{-t_2}+\frac{1}{6}e^{-3t_2}-\frac{1}{12}e^{-4t_2}\right)$
(1,2,2,3)	$g_{2,2}(t_4)g_{2,2}(t_3)g_{2,1}(t_3)\frac{1}{3}g_{3,1}(t_2)$
(1,2,2,4)	$g_{2,1}(t_4)g_{2,2}(t_3)\frac{1}{3}g_{3,1}(t_2)$
(1,2,3,3)	$g_{2,2}(t_4)\frac{1}{2}g_{2,1}(t_3)g_{2,1}(t_3)g_{2,1}(t_2)$
(1,2,3,4)	$g_{2,1}(t_4)g_{2,1}(t_3)g_{2,1}(t_2)$

The probability for ranked history (1,2,2,2) is obtained from Eq. (14), and the probability for ranked history (1,1,2,2) is obtained in a similar manner.

long. More precisely, for $\mathcal{T}_{RLL}$ to produce AGTs, τ_2 and τ_3 must be short while τ_4 is long, whereas for $\mathcal{T}_{RLL}$ to have ARGTs, τ_3 and τ_4 must be short while τ_2 is long. It can be shown that for $\mathcal{T}_{RLL}$, the ranked and unranked anomaly zones have no overlap.

6.3. Multiple lineages per species

Our work has assumed that one lineage is sampled per species. Cases with more than one lineage per species can be considered, however, by extending the species tree with artificial leaves that have length zero. For example, consider species B, C, D, and E, with two lineages sampled from B and one lineage sampled from the other species, and with ranked species tree topology $((BC)_2(DE)_3)$. The species tree in Fig. 1(d), which includes species A, could be used to represent this case by letting $s_4 = 0$, effectively combining species A and B into a single sampled species. Thus, the four-taxon case in which two lineages are sampled from B can lead to phenomena similar to those observed in the five-taxon case in Fig. 1(d). In particular, from Fig. 1(d), although D and E diverge more recently than do B and C, the ancestral population of B and

Table 4
Probabilities of ranked histories for the species tree $\mathcal{T}_{RLL}$.

Ranked history	Ranked gene tree topology		
	$\mathcal{G}_{LRL}$	$\mathcal{G}_{LLR}$	$(((AC)_4B)_2(DE)_3)$
(1,1,1,1)	$g_{2,2}(t_4)g_{2,2}(t_3)g_{2,2}(t_3)g_{3,3}(t_2)g_{2,2}(t_2)\frac{1}{180}$	Same	Same
(1,1,1,2)	$g_{2,2}(t_4)g_{2,2}(t_3)g_{2,2}(t_3)\frac{1}{3}g_{3,2}(t_2)g_{2,2}(t_2)\frac{1}{18}$	Same	Same
(1,1,1,3)	$g_{2,2}(t_4)g_{2,1}(t_3)g_{2,2}(t_3)g_{2,2}(t_2)g_{2,2}(t_2)\frac{1}{18}$	Same	NA
(1,1,1,4)	NA	NA	NA
(1,1,2,2)	$g_{2,2}(t_4)g_{2,2}(t_3)g_{2,2}(t_3)\left(\frac{1}{3}e^{-t_2}-\frac{1}{2}e^{-2t_2}+\frac{1}{6}e^{-4t_2}\right)\frac{1}{3}$	Same	Same
(1,1,2,3)	$g_{2,2}(t_4)g_{2,1}(t_3)g_{2,2}(t_3)g_{2,1}(t_2)g_{2,2}(t_2)\frac{1}{3}$	Same	NA
(1,1,2,4)	NA	NA	NA
(1,1,3,3)	$g_{2,2}(t_4)\frac{1}{2}g_{2,1}(t_3)g_{2,1}(t_3)g_{2,2}(t_2)\frac{1}{3}$	NA	NA
(1,1,3,4)	NA	NA	NA
(1,2,2,2)	$g_{2,2}(t_4)g_{2,2}(t_3)g_{2,2}(t_3)\left(\frac{1}{8}-\frac{1}{3}e^{-t_2}+\frac{1}{4}e^{-2t_2}-\frac{1}{24}e^{-4t_2}\right)$	Same	Same
(1,2,2,3)	$g_{2,2}(t_4)g_{2,1}(t_3)g_{2,2}(t_3)\frac{1}{2}g_{2,1}(t_2)g_{2,1}(t_2)$	Same	NA
(1,2,2,4)	NA	NA	NA
(1,2,3,3)	$g_{2,2}(t_4)\frac{1}{2}g_{2,1}(t_3)g_{2,1}(t_3)g_{2,1}(t_2)$	NA	NA
(1,2,3,4)	NA	NA	NA

'Same' indicates that the probability of the ranked history is the same for the given ranked gene tree topology as for $\mathcal{G}_{LRL}$; 'NA' indicates that the ranked history does not apply to the combination of ranked gene tree topology and species tree. The probability for ranked history (1,2,2,2) is obtained from Eq. (15), and the probability for ranked history (1,1,2,2) is obtained in a similar manner.

C potentially has three lineages but the ancestral population of D and E has at most two lineages. Given τ_2 sufficiently large and τ_3 and τ_4 sufficiently small, the probability is greater that the most recent interspecific coalescence will occur between lineages from B and C rather than between lineages from D and E.

6.4. Conclusions

A view that ranked trees can produce new tools for evolutionary modeling and species tree inference underlies our work. Our demonstration that ranked gene tree probabilities can be conveniently computed under a standard model provides a basis for exploring the utility of ranked gene trees in modeling and inference problems, and it possesses the same role in the ranked case as the corresponding derivation for unranked trees [3]. However, similarly to the discovery of anomalous unranked gene trees [1], our proof of the existence of ARGTs illustrates that ranked gene trees generate a peculiar phenomenon that is intriguing both mathematically and for what it implies about the evolutionary process. In our subsequent efforts, we will characterize the ARGT phenomenon more fully.

Acknowledgments

This work was supported by NSF grants DEB-0716904 and DBI-1146722 and grants from the New Zealand Marsden Fund and the Burroughs Wellcome Fund. We thank Cuong Than and two anonymous reviewers for comments.

Appendix A

See Tables 1–4.

References

[1] J.H. Degnan, N.A. Rosenberg, Discordance of species trees with their most likely gene trees, PLoS Genet. 2 (2006) 762.
[2] J.H. Degnan, N.A. Rosenberg, Gene tree discordance, phylogenetic inference, and the multispecies coalescent, Trends Ecol. Evol. 24 (2009) 332.
[3] J.H. Degnan, L.A. Salter, Gene tree distributions under the coalescent process, Evolution 59 (2005) 24.
[4] R.R. Hudson, Testing the constant-rate neutral allele model with protein sequence data, Evolution 37 (1983) 203.
[5] L. Liu, L. Yu, L.S. Kubatko, D.K. Pearl, S.V. Edwards, Coalescent methods for estimating phylogenetic trees, Mol. Phylogenet. Evol. 53 (2009) 320.
[6] W.S. Moore, Inferring phylogenies from mtDNA variation: mitochondrial-gene trees versus nuclear-gene trees, Evolution 49 (1995) 718.
[7] M. Nei, Molecular Evolutionary Genetics, Columbia University, New York, 1987.
[8] P. Pamilo, M. Nei, Relationships between gene trees and species trees, Mol. Biol. Evol. 5 (1988) 568.
[9] N.A. Rosenberg, The probability of topological concordance of gene trees and species trees, Theor. Pop. Biol. 61 (2002) 225.
[10] F. Tajima, Evolutionary relationship of DNA sequences in finite populations, Genetics 105 (1983) 437.
[11] N. Takahata, Gene genealogy in three related populations: consistency probability between gene and population trees, Genetics 122 (1989) 957.
[12] N.A. Rosenberg, R. Tao, Discordance of species trees with their most likely gene trees: the case of five taxa, Syst. Biol. 57 (2008) 131.
[13] E.S. Allman, J.H. Degnan, J.A. Rhodes, Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent, J. Math. Biol. 62 (2011) 833.
[14] J.H. Degnan, M. DeGiorgio, D. Bryant, N.A. Rosenberg, Properties of consensus methods for inferring species trees from gene trees, Syst. Biol. 58 (2009) 35.
[15] L. Liu, L. Yu, D.K. Pearl, Maximum tree: a consistent estimator of the species tree, J. Math. Biol. 60 (2010) 95.
[16] L. Liu, L. Yu, D.K. Pearl, S.V. Edwards, Estimating species phylogenies using coalescence times among sequences, Syst. Biol. 58 (2009) 468.
[17] E. Mossel, S. Roch, Incomplete lineage sorting: consistent phylogeny estimation from multiple loci, IEEE/ACM Trans. Comput. Biol. Bioinf. 7 (2010) 166.
[18] C.V. Than, N.A. Rosenberg, Consistency properties of species tree inference by minimizing deep coalescences, J. Comput. Biol. 18 (2010) 1.
[19] B.C. Carstens, L.L. Knowles, Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from *Melanoplus* grasshoppers, Syst. Biol. 56 (2007) 400.
[20] B.R. Larget, S.K. Kotha, C.N. Dewey, C. Ané, BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis, Bioinformatics 26 (2010) 2910.
[21] L. Liu, L. Yu, S.V. Edwards, A maximum pseudo-likelihood approach for estimating species trees under the coalescent model, BMC Evol. Biol. 10 (2010) 303.
[22] C. Meng, L.S. Kubatko, Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model, Theor. Popul. Biol. 75 (2009) 35.
[23] C.-I. Wu, Inferences of species phylogeny in relation to segregation of ancient polymorphisms, Genetics 127 (1991) 429.
[24] N.A. Rosenberg, Counting coalescent histories, J. Comput. Biol. 14 (2007) 360.
[25] N.A. Rosenberg, J.H. Degnan, Coalescent histories for discordant gene trees and species trees, Theor. Pop. Biol. 77 (2010) 145.
[26] C. Than, D. Ruths, H. Innan, L. Nakhleh, Confounding factors in HGT detection: statistical error, coalescent effects, and multiple solutions, J. Comput. Biol. 14 (2007) 517.
[27] J.K.M. Brown, Probabilities of evolutionary trees, Syst. Biol. 43 (1994) 78.
[28] N.A. Rosenberg, The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model, Evolution 57 (2003) 1465.
[29] M. Aigner, Enumeration via ballot numbers, Discrete Math. 308 (2008) 2544.
[30] R. Stanley, Enumerative Combinatorics: Volume 2, Cambridge University Press, Cambridge, 1999.
[31] J.H. Degnan, Gene tree distributions under the coalescent process, Ph.D. Thesis, University of New Mexico, Albuquerque, NM, 2005.
[32] S. Tavaré, Line-of-descent and genealogical processes, and their applications in population genetics models, Theor. Pop. Biol. 26 (1984) 119.
[33] B. Rannala, Z. Yang, Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci, Genetics 164 (2003) 1645.