On the unranked topology of maximally probable ranked gene tree topologies

Filippo Disanto* **Pasquale Miglionico†** **Guido Narduzzi†**

April 5, 2019

Abstract

A ranked tree topology is a tree topology with a temporal ordering of its coalescence events. Under the multispecies coalescent model, we consider ranked gene tree topologies realized along the branches of ranked species trees, where one gene copy is sampled for each species. Previous results have demonstrated that for almost all ranked species tree topologies with at least five species, there exists a set of branch lengths such that the maximally probable ranked gene tree topologies—those generated with the highest probability under the model—do not match the species tree ranked topology. Here, we focus on the agreement of a ranked species tree with its maximally probable ranked gene tree topologies in terms of their unranked topology, that is, disregarding the ordering of the coalescence events. We show that although the set of maximally probable ranked gene tree topologies for a ranked species tree can contain ranked trees with different unranked topologies, at least one of these maximal ranked gene tree topologies must have the same unranked topology as the species tree. Our results contribute to the study of the relationships between gene trees and species trees.

Keywords Gene tree topologies · Phylogenetics · Species trees

Mathematics Subject Classification (2010) 05C05 · 92B10 · 92D15

1 Introduction

The description of evolutionary relationships among organisms using trees or networks is a central theme in Biology. Gene trees represent the evolutionary relationships of a single genetic locus across a set of individuals drawn from different species. Species trees describe the relationships among the populations or species from which individuals are sampled. By making use of probabilistic models of sequence evolution (Felsenstein, 2004), phylogenetic methods enable the inference of gene trees. Species trees can be estimated from the inferred gene trees by different methods. Some of these procedures reconstruct species trees by considering only the topology—i.e. the branching pattern—of the observed gene trees (Degnan et al., 2009; Ewing et al., 2008; Liu et al., 2009; Maddison and Knowles, 2006; Than and Nakhleh, 2009; Wu, 2012). Some others rely on the knowledge of both the topology and branch lengths of gene trees (Heled and Drummond, 2010; Kubatko et al., 2009; Liu and Pearl, 2007; Liu and Yu, 2011; Liu et al., 2010; Mossel and Roch, 2010).

Intermediate between these two approaches is the use of *ranked* topologies for representing gene trees. Ranked gene tree topologies have been introduced by Degnan et al. (2012b) as tree structures that keep track of both the gene tree topology and the relative order (1st, 2nd, and so on) with which the coalescence events occur in the gene tree. More precisely, seen as purely combinatorial object, a ranked tree topology is a leaf labeled tree topology together with a linear ordering—a ranking—of its internal nodes. In particular, the ranked tree topology of a species tree has the ranking of its internal nodes naturally induced by the set of branch lengths of the tree.

The variety of ranked gene tree topologies observed at different loci represents an important source of information that needs to be analyzed in order to provide accurate species tree estimates. A key tool is a proper modeling of the evolution of gene trees. The multispecies coalescent model (Degnan and Rosenberg, 2009;

*Department of Mathematics, University of Pisa, 56126 Italy. Corresponding author. Email: filippo.disanto@unipi.it.
†Scuola Normale Superiore, Pisa, 56126 Italy. Email: pasquale.miglionico@sns.it, guido.narduzzi@sns.it.

Degnan and Salter, 2005; Hudson, 1983; Maddison, 1997; Pamilo and Nei, 1988; Rosenberg, 2002) simulates the descent of genealogical lineages along the branches of a species tree, enabling the calculation of the conditional probability of a ranked gene tree topology (Degnan et al., 2012b; Stadler and Degnan, 2012). Under this model, Degnan et al. (2012b) have demonstrated the existence of anomalous ranked gene tree topologies: ranked gene tree topologies whose probability is larger than that of the ranked gene tree topology that matches the ranked topology of the species tree. In other words, the ranked topology of a species tree that produces anomalous ranked gene tree topologies does not agree with the most likely ranked gene tree topology. Furthermore, since almost all ranked species tree topologies with at least five species produce anomalous ranked gene tree topologies (Degnan et al., 2012a; Disanto and Rosenberg, 2014), the "democratic vote" procedure of estimating the species tree ranked topology as the ranked topology of the most frequently observed ranked gene tree topology may often converge on the wrong species tree ranked topology, as the number of considered loci increases.

Although the most frequent ranked gene tree topology does not in general characterize the species tree ranked topology, it could still provide important information on the *unranked* topology of the species tree, that is, on the species tree topology obtained by disregarding the ordering of the internal nodes. In order to explore this possibility, Degnan et al. (2012a, 2012b) further studied the relationships between anomalous ranked gene tree topologies and species trees in terms of their unranked topology—where, in general, the unranked topology of an anomalous ranked gene tree topology can be either matching or not that of the species tree. In particular, they demonstrated that, under fairly general hypotheses, anomalous ranked gene tree topologies exist that differ from the species tree in the unranked topology (Degnan et al. (2012a), Section 4.2), and asked whether every species tree has its *unranked* topology matching that of the most probable ranked gene tree topologies (Degnan et al. (2012a), Section 5). If true, the latter property could assist in designing inference procedures from ranked gene tree topologies. For instance, knowing that the species tree has the unranked topology of one of the most frequently observed ranked gene tree topologies would allow to restrict the tree space in which we search for a species tree estimate by considering only those unranked tree topologies underlying the most frequent ranked gene tree topologies.

Here, we answer the question raised by Degnan et al. (2012a), considering ranked gene tree topologies realized in species trees under the multispecies coalescent, when exactly one gene copy is sampled for each species. By following the approach introduced by Degnan et al. (2012b) for calculating ranked gene tree probabilities, we prove that among the ranked gene tree topologies that are maximally probable for a species tree S, there is always at least one with the same unranked topology of S. In other words, if anomalous ranked gene tree topologies exist for S, then there is an anomalous ranked gene tree topology of maximal probability that disagrees with S only in the ranking of its unranked topology, which is exactly that of S.

Moreover, we also demonstrate that in general the set of maximally probable ranked gene tree topologies for a given species tree can contain trees with different unranked topologies. Indeed, we exhibit a species tree with 7 taxa for which there exists a ranked gene tree topology of the same size with maximal probability and non-matching unranked topology. Hence, this shows that not all ranked gene tree topologies of maximal probability need to share their unranked topology with the species tree.

Our theoretical results contribute to the study of the inference of species trees from gene tree topologies. In particular, because the unranked topology of a species tree can often disagree with the most likely unranked gene tree topologies (Degnan and Rosenberg, 2006), the existence for every species tree of a maximally probable ranked gene tree topology with a matching unranked topology suggests that, in inferring the unranked topology of a species tree, methods based on ranked gene trees might provide better estimates than methods that use unranked gene trees.

2 Preliminaries

We start by introducing some notation (Table 1) and preliminary results that will be used in the rest of the paper. In Section 2.1, we give some basic tree terminology. Section 2.2 focuses on algorithmic aspects of the computation of ranked gene tree probabilities.

Table 1: Main notation used in the paper

Symbol	Meaning	Symbol	Meaning
t_ℓ / t_r	The left/right root subtree of a planar representation of a tree t	$m_i(h)$	The number of coalescences of G occurring in ith time interval of S
$[t]$	The unranked topology of a ranked tree topology t	$k_{ijz}(G,h)$	The number of lineages that exist on the zth branch-zone of the ith time interval of S right after the jth coalescence of G occurring in that interval
G	The ranked topology of a gene tree	$\mathcal{G}_s$	The set of maximally probable ranked gene tree topologies for S
γ_i	The internal node of rank i in G	G_i	The subtree of G rooted at γ_i
S	The ranked topology of a species tree together with lengths for its time intervals	$S_{\|g}$	The smallest subtree of S containing a subtree g of G
$H(G,S)$	The set of ranked histories h possible for G in S	i^*	The maximum index i such that $[G_i] \neq [S_{\|G_i}]$
h^*	The maximal ranked history in $H(G,S)$ with respect to the lexicographic order	α	The branch of S that coalesces with the root branch of $S_{\|(G_{i*})_\ell}$
t_i	The length of the time interval with exactly i branches of S	β	A point on G contemporary with γ_{i^*} and such that all descending taxa are below α

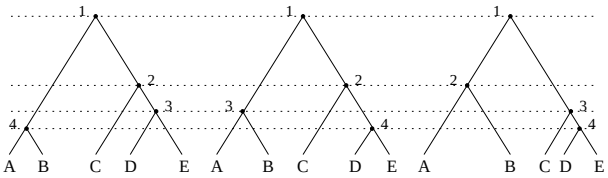

Figure 1: The three ranked tree topologies with unranked topology $((A,B),(C,(D,E)))$.

2.1 Ranked gene tree topologies, ranked species trees, and ranked histories

We recall from Degnan et al. (2012b) the definition of ranked gene tree topology, ranked species tree, and ranked history.

A *ranked tree topology* t (or ranked tree for short) of size $|t| = n$ is a full binary rooted tree with n labeled terminal nodes, also called the leaves or the taxa of t, and $2n-1$ nodes in total. The $n-1$ internal nodes of t are bijectively associated with a number in $\{1,2,...,n-1\}$ so that each path from the root of t to a leaf contains an increasing sequence of numbers (Fig. 1). The number of ranked trees of size n is given by $[(n-1)!\,n!]/2^{n-1}$ (Rosenberg, 2006). Ranked tree topologies can be represented in Newick format—e.g. $((A,B)_4,(C,(D,E)_3)_2)_1$ for the leftmost tree in Fig. 1—with a subscript for each closed parenthesis that determines the ranking of the associated internal node.

A ranked tree t has its *unranked* topology $[t]$ obtained by ignoring the ordering of its internal nodes, while keeping the labeling of the taxa. Hence, different ranked trees can share the same unranked topology (Fig. 1). The number of possible unranked topologies with $n \geq 2$ taxa is given by $(2n-3)!! = (2n-3)\times(2n-5)\times\cdots\times 1$ (Rosenberg, 2006).

A *ranked gene tree topology* G is a ranked tree in which leaf labels denote different gene copies, whereas the ranking of the internal nodes gives a time ordering of the coalescence events occurring along the branches of the tree looking forward in time. The internal node of G with rank i is denoted by γ_i. The most recent internal node is thus γ_{n-1}, when $|G| = n$. A gene lineage (or lineage for short) of G is a branch of G connecting two nodes, where the term "node" refers to both internal and terminal nodes of G.

A *ranked species tree* S of size n consists of a ranked tree topology together with a vector of positive real numbers $(t_i) \equiv (t_i)_i \equiv (t_1,t_2,...,t_{n-1})$, where $t_i > 0$ measures the length of the interval in which exactly i branches of S coexist (Fig. 2). Interval length is measured in coalescent units of time, where we assume t_1 to be infinite ($t_1 = \infty$). We will often refer to the length t_i of the ith time interval of S as a symbolic variable, without explicitly assigning to it a numerical value.

A branch of a ranked species tree S connects two adjacent nodes of S, which can be both internal or one

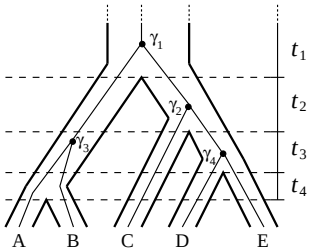

Figure 2: A realization of the ranked gene tree topology $G = ((A, B)_3, (C, (D, E)_4)_2)_1$ in the ranked species tree S (thicker tree) with ranked topology $((A, B)_4, (C, (D, E)_3)_2)_1$ and length t_i for the ith time interval. The ranked history associated with the depicted realization of G in S is the maximal ranked history $h^* = (1, 2, 3, 3)$ of the set $H(G, S)$ given in (1).

terminal and one internal. A *branch-zone* of S is a part of a branch of S that intersects a given time interval of S. For instance, in the ranked species tree S of Fig. 2, the branch that connects the root node of the subtree (A, B) with the root of S is divided into three branch-zones, one for each time interval of S different from the first. The branch that connects the root node of the subtree (D, E) with the root node of the subtree $(C, (D, E))$ consists instead of exactly one branch-zone, as it intersects only the third time interval of S. The ith time interval of S intersects exactly i branch-zones of S that, for a given planar representation of S, can be ordered from left to right: the first branch-zone is the leftmost one, while the ith branch-zone is the rightmost one (Fig. 2).

Treating a ranked species tree S as a fixed parameter, in this paper we study probabilistic properties of random ranked gene tree topologies realized along the branches of S, when exactly one gene copy is sampled for each species (Fig. 2). Thus, each ranked gene tree topology G for S is assumed to have the same size and the same set of taxa of S.

Given a ranked species tree S of size n and a ranked gene tree topology G for S, a *ranked history* h of G in S is a $(n-1)$-tuple $h = (h_1, h_2, ..., h_{n-1})$ of integers representing one of the possible combinatorially different evolutionary scenarios of G realized along the branches of S. More precisely, h encodes the realization of G in S in which the coalescence event of G with rank i occurs in the h_ith time interval of S. For example, $h = (1, 1, ..., 1)$ corresponds to the realization of G in the root branch of S. We denote by $H(G, S)$ the set of ranked histories of G in S. For the ranked gene tree topology G and the ranked species tree S depicted in Fig. 2, the set $H(G, S)$ is given by

$$\{(1,1,1,1), (1,1,1,2), (1,1,1,3), (1,1,2,2), (1,1,2,3), (1,1,3,3), (1,2,2,2), (1,2,2,3), (1,2,3,3) \equiv h^*\}, \quad (1)$$

where the *maximal* ranked history with respect to the lexicographic order is denoted by h^* and it corresponds to the realization depicted in Fig. 2.

As it can be observed from the list given in (1), the ranked histories of a ranked gene tree topology G in a ranked species tree S of size n can be characterized through the maximal ranked history $h^* = (h_1^*, h_2^*, ..., h_{n-1}^*)$ of G in S as the $(n-1)$-tuples of integers $(h_1, h_2, ..., h_{n-1})$ that satisfy the following two conditions: (i) $h_1 = 1$, and (ii) $h_{i-1} \leq h_i \leq h_i^*$ for $2 \leq i \leq n-1$. Indeed, h_i^* corresponds to the index of the most recent time interval of S in which the coalescence event γ_i of G can occur. From the maximal ranked history h^* of G in S the entire set $H(G, S)$ can thus be determined.

2.2 Calculation of the conditional probability of a ranked gene tree topology

For a fixed ranked species tree S of size n, the conditional probability $\text{Prob}(G|S)$ of a ranked gene tree topology G can be calculated under the *multispecies coalescent* model (Degnan and Rosenberg, 2009; Degnan and Salter, 2005; Hudson, 1983; Maddison, 1997; Pamilo and Nei, 1988; Rosenberg, 2002) by following the approach described by Degnan et al. (2012b). In particular, in Eq. (11) of Degnan et al. (2012b) the probability of G is computed by summing the probabilities of the combinatorially different realizations of G in S encoded by the ranked histories of G in S, that is,

$$\text{Prob}(G|S) = \sum_{h \in H(G,S)} \text{Prob}(G \,\&\, h|S). \quad (2)$$

In turn, for each ranked history $h \in H(G,S)$, $\mathrm{Prob}(G\,\&\,h|S)$ is calculated in Eqs. (9) and (10) of Degnan et al. (2012b) as a probability function of the quantities (t_i), i.e. the lengths of the time intervals of S, and $(m_i) \equiv (m_i)_i$ and $(k_{i,j,z}) \equiv (k_{i,j,z})_{i,j,z}$, which are defined as follows:

- $m_i = m_i(h)$ denotes, for $1 \leq i \leq n-1$, the number of coalescence events of G that occur in the ith time interval of S or, in other words, the number of entries i in h. In Fig. 2, $m_1 = 1, m_2 = 1, m_3 = 2$, and $m_4 = 0$.
- $k_{i,j,z} = k_{i,j,z}(G,h)$ denotes, for $1 \leq i \leq n-1$, $0 \leq j \leq m_i$, and $1 \leq z \leq i$, the number of gene lineages present on the zth branch-zone of the ith time interval of S just after the jth coalescence event of G occurring in that time interval. In Fig. 2, we have $k_{1,1,1} = 2, k_{2,1,1} = 1, k_{2,1,2} = 2, k_{3,1,1} = 2, k_{3,1,2} = 1, k_{3,1,3} = 1, k_{3,2,1} = 2, k_{3,2,2} = 1$, and $k_{3,2,3} = 2$. For example, looking at the 3rd time interval of the species tree in the figure, we find $k_{3,1,1} = 2$ as right after the 1st coalescence event γ_3 there are 2 gene lineages in the 1st (leftmost) branch-zone of the interval. Similarly, $k_{3,1,2} = 1$ and $k_{3,1,3} = 1$, because right after γ_3 the 2nd (middle) and the 3rd (rightmost) branch-zone of the interval contain exactly 1 gene lineage. The 2nd coalescence event of G that occurs in the 3rd interval of S is γ_4, where $k_{3,2,1} = 2, k_{3,2,2} = 1$, and $k_{3,2,3} = 2$. Note that $k_{i,0,z}$ has to be interpreted as the number of gene lineages entering the zth branch-zone of S at the boundary of the $(i-1)$th and ith time interval of S. In Fig. 2, $k_{1,0,1} = 1, k_{2,0,1} = 1, k_{2,0,2} = 1, k_{3,0,1} = 1, k_{3,0,2} = 1, k_{3,0,3} = 1, k_{4,0,1} = 2, k_{4,0,2} = 1, k_{4,0,3} = 1$, and $k_{4,0,4} = 1$.

In symbols, we write

$$\mathrm{Prob}(G\,\&\,h|S) = P\big((t_i),(m_i),(k_{i,j,z})\big) \tag{3}$$

to indicate that the probability $\mathrm{Prob}(G\,\&\,h|S)$ of G realized in S according to the ranked history h is completely determined by the values of $(t_i), (m_i)$, and $(k_{i,j,z})$.

We have implemented the procedure described by Degnan et al. (2012b) for evaluating ranked gene tree probabilities in a software `RGTProb` available at `https://github.com/PasqM/RGTProb`. This program enables the symbolic calculation of the conditional probability (2) as a function of the lengths (t_i) of the time intervals of S considered as variables. In its current version, it consists of a `Python` script that outputs a `Mathematica` file with an explicit probability formula. For instance, for the ranked gene tree topology G and the ranked species tree S depicted in Fig. 2, our software computes the conditional probability of G given S as

$$\begin{aligned}
\mathrm{Prob}(G|S) \;=\;& \frac{1}{180}\,e^{-t_4}\,e^{-2t_3}\,e^{-4t_2} + \frac{1}{18}\,e^{-t_4}\,e^{-2t_3}\left(\frac{e^{-2t_2}}{2} - \frac{e^{-4t_2}}{2}\right) + \frac{1}{18}\,e^{-t_4}\,(e^{-t_3} - e^{-2t_3})\,e^{-2t_2} \\
&+ \frac{1}{3}\,e^{-t_4}\,e^{-2t_3}\left(\frac{e^{-t_2}}{3} - \frac{e^{-2t_2}}{2} + \frac{e^{-4t_2}}{6}\right) + \frac{1}{3}\,e^{-t_4}\,(e^{-t_3} - e^{-2t_3})\,(e^{-t_2} - e^{-2t_2}) \\
&+ \frac{1}{3}\,e^{-t_4}\left(\frac{1}{2} - e^{-t_3} + \frac{e^{-2t_3}}{2}\right)e^{-t_2} + e^{-t_4}\,e^{-2t_3}\left(\frac{1}{8} - \frac{e^{-t_2}}{3} + \frac{e^{-2t_2}}{4} - \frac{e^{-4t_2}}{24}\right) \\
&+ e^{-t_4}\,(e^{-t_3} - e^{-2t_3})\left(\frac{1}{2} - e^{-t_2} + \frac{e^{-2t_2}}{2}\right) + e^{-t_4}\left(\frac{1}{2} - e^{-t_3} + \frac{e^{-2t_3}}{2}\right)(1 - e^{-t_2}) \\
=\;& \frac{1}{360}\,e^{-4t_2-2t_3-t_4}\left(-40\,e^{2t_2} + 40\,e^{3t_2} + 45\,e^{4t_2} + 80\,e^{2t_2+t_3} - 180\,e^{4t_2+t_3} - 120\,e^{3t_2+2t_3} + 180\,e^{4t_2+2t_3} - 3\right),
\end{aligned} \tag{4}$$

where the ith summand $(1 \leq i \leq 9)$ in Eq. (4) yields the probability of the ith ranked history of G in S listed in (1).

Note that the nine summands present in Eq. (4) match the entries of the second column of Table 4 in Degnan et al. (2012b), where the authors express the probabilities of the ranked histories of G in S in terms of the function $g_{i,j}(t)$ that calculates the probability that i lineages coalesce to $j \leq i$ lineages during time t. Moreover, we have also verified that `RGTProb` gives correct results when used for evaluating unranked gene tree probabilities by summing ranked gene tree probabilities. For example, when S is the ranked species tree with ranked topology $((A,(B,C)_4)_2,((D,E)_5,(F,G)_6)_3)_1$ and interval lengths (t_2,t_3,t_4,t_5,t_6), the probability of the unranked gene tree topology $T = ((A,B),(C,((D,F),(E,G))))$ can be obtained by summing the probabilities of the ten ranked gene tree topologies with unranked topology T:

$$\begin{array}{ll}
((A,B)_2,(C,((D,F)_5,(E,G)_6)_4)_3)_1, & ((A,B)_2,(C,((D,F)_6,(E,G)_5)_4)_3)_1, \\
((A,B)_3,(C,((D,F)_5,(E,G)_6)_4)_2)_1, & ((A,B)_3,(C,((D,F)_6,(E,G)_5)_4)_2)_1, \\
((A,B)_4,(C,((D,F)_5,(E,G)_6)_3)_2)_1, & ((A,B)_4,(C,((D,F)_6,(E,G)_5)_3)_2)_1, \\
((A,B)_5,(C,((D,F)_4,(E,G)_6)_3)_2)_1, & ((A,B)_5,(C,((D,F)_6,(E,G)_4)_3)_2)_1, \\
((A,B)_6,(C,((D,F)_4,(E,G)_5)_3)_2)_1, & ((A,B)_6,(C,((D,F)_5,(E,G)_4)_3)_2)_1.
\end{array}$$

By using `RGTProb` for calculating these ten probabilities, and then substituting $t_2 = t_4 = 1$ and $t_3 = t_5 = t_6 = 0$ in the resulting formula for $\text{Prob}(T|S)$, we recover the value 0.000154... reported by Degnan and Salter (2005) at the bottom of the third column of their Table 1. Indeed, in their setting, the same numerical value yields the conditional probability of the unranked topology T—depicted in their Fig. 1(b) and (d)—when the species tree is characterized by having unranked topology $[S]$—depicted in their Fig. 1(a) and (c)—and all internal branches of length 1.

Besides being of interest from an algorithmic point of view, Eqs. (2) and (3) yield the following lemma, which will be used in the next section to derive our main result.

Lemma 1 *Let S be a ranked species tree, and let G_1 and G_2 be two ranked gene tree topologies for S such that (i) $H(G_1, S) \subseteq H(G_2, S)$, and (ii) for each $h \in H(G_1, S)$, $k_{i,j,z}(G_1, h) = k_{i,j,z}(G_2, h)$ for all possible choices of the indices i, j, z. Then we have $Prob(G_1|S) \leq Prob(G_2|S)$. Furthermore, if $H(G_1, S) = H(G_2, S)$ in (i), then from condition (ii) it follows that $Prob(G_1|S) = Prob(G_2|S)$.*

Proof. If conditons (i) and (ii) are both satisfied, then for every $h \in H(G_1, S)$—where h is a ranked history of G_2 in S as well—we have $\text{Prob}(G_1 \,\&\, h|S) = \text{Prob}(G_2 \,\&\, h|S)$ as it follows from Eq. (3). Hence, from Eq. (2), we obtain $\text{Prob}(G_1|S) < \text{Prob}(G_2|S)$ when $H(G_1, S) \subset H(G_2, S)$, whereas $\text{Prob}(G_1|S) = \text{Prob}(G_2|S)$ when we have the equality $H(G_1, S) = H(G_2, S)$. □

3 Results

For a given ranked species tree S, let $\mathcal{G}_S$ be the set of ranked gene tree topologies that are *maximally* probable for S. In other words, if $G \in \mathcal{G}_S$, then $\text{Prob}(G|S) \geq \text{Prob}(\overline{G}|S)$ for all ranked gene tree topologies $\overline{G}$ that can be realized in S by sampling one gene per species. Note that the set $\mathcal{G}_S$ depends in general on the numerical values of the lengths (t_i) of the time intervals of S.

Here, we show two facts. First, in Section 3.1, we demonstrate that, for every ranked species tree S, there exists $G \in \mathcal{G}_S$ such that $[G] = [S]$. Second, in Section 3.2, we identify a ranked species tree S for which the set $\mathcal{G}_S$ contains a ranked gene tree topology G such that $[G] \neq [S]$. Thus, in general, the set $\mathcal{G}_S$ of maximally probable ranked gene tree topologies for a ranked species tree S can contain ranked trees with different unranked topologies, but at least one of these ranked trees must have the same unranked topology of S.

3.1 Every S has a maximally probable ranked gene tree topology G with $[G] = [S]$

In order to prove that for every ranked species tree S the set $\mathcal{G}_S$ contains a tree with the same unranked topology of S, we describe a procedure that, given a ranked species tree S and a ranked gene tree topology G for S, constructs a ranked gene tree topology G^* such that $[G^*] = [S]$ and $\text{Prob}(G^*|S) \geq \text{Prob}(G|S)$. The idea is to obtain G^* from G by iteratively applying a tree operator $(\cdot)'$ to produce a finite sequence of ranked gene tree topologies of non-decreasing probability,

$$G \equiv G^{(0)}, G^{(1)}, ..., G^{(r)}, G^{(r+1)}, ..., G^{(q-1)}, G^{(q)} \equiv G^*, \tag{5}$$

in which $G^{(r+1)} \equiv \left(G^{(r)}\right)'$ and, for a certain index q, $[G^*] = [G^{(q)}] = [S]$ (Fig. 3).

In Section 3.1.1, we define the tree operator $(\cdot)'$ by showing how the output ranked gene tree topology G' is built from a given input ranked gene tree topology G such that $[G] \neq [S]$. In other words, we show how to obtain $G^{(r+1)}$ from $G^{(r)}$ in (5) if $[G^{(r)}] \neq [S]$. In Section 3.1.2, we describe some key features of the introduced tree operator that lead to the proof of our main result.

3.1.1 Construction of G' from G when $[G] \neq [S]$

Fix a ranked species tree S. The tree operator $(\cdot)'$ that we introduce in this section outputs a ranked gene tree topology G' for S starting from a ranked gene tree topolgy G for S such that $[G] \neq [S]$. In order to describe the construction of G', we first need some definitions and notation.

For a given planar representation of a ranked tree t, let t_{left} and t_{right} (or t_ℓ and t_r for short) denote the left and right root subtree of t, respectively. If t is any tree in Fig. 1, then t_ℓ has its set of taxa given by

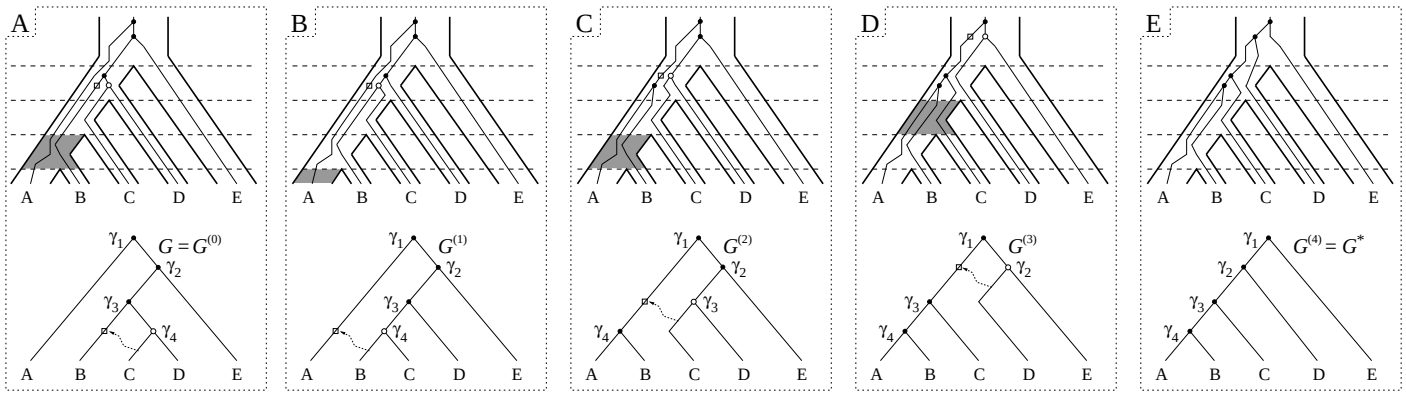

Figure 3: Application of the operator $(\cdot)'$. For a fixed ranked species S tree with ranked topology $((((A,B)_4,C)_3,D)_2,E)_1$, we consider the ranked gene tree topology $G^{(0)} \equiv G = (A,((B,(C,D)_4)_3,E)_2)_1$ realized in S as in **A**. Starting from $G^{(0)}$, we iteratively apply the tree operator $(\cdot)'$ described in the text finding a sequence $G^{(1)} \equiv (G^{(0)})', G^{(2)} \equiv (G^{(1)})', G^{(3)} \equiv (G^{(2)})'$, and $G^* \equiv G^{(4)} \equiv (G^{(3)})'$ of ranked gene tree topologies represented at the bottom of panels **B**, **C**, **D**, and **E**, respectively, and realized in S as depicted at the top of each panel. The sequence ends with $G^* \equiv G^{(4)}$ as $[G^{(4)}] = [S]$.

$\{A,B\}$, while $(t_\ell)_\ell$ consists of the taxon A. Given a ranked gene tree topology G for S, G_i is the subtree of G containing the internal node γ_i of rank i in G and all the nodes descending from it. Given a subtree g of G, $S_{|g}$ is the smallest subtree of S that contains all the taxa present in g. In particular, if g consists of a single labeled taxon τ of G, then $S_{|g} = \tau$, whereas if g has size larger than one, then the root of $S_{|g}$ is the internal node of largest rank in the ranked topology of S whose set of descending taxa contains the taxa present in g. Finally, if $[G] \neq [S]$, then we define

$$i^* = i^*(G) = \max\{i : [G_i] \neq [S_{|G_i}]\}, \tag{6}$$

that is, i^* is the maximum index $1 \leq i \leq |G| - 1$ such that G_i and $S_{|G_i}$ have a different unranked topology. In Fig. 3A, the white node of the ranked gene tree topology G realized inside the ranked species tree S corresponds to the node γ_{i^*}, where $\{C,D\}$ is the set of taxa of G_{i^*} and $\{A,B,C,D\}$ is the set of taxa of $S_{|G_{i^*}}$.

Observe that if S and G are such that $[G] \neq [S]$, then the set of taxa of $S_{|G_{i^*}}$ strictly contains the set of taxa of G_{i^*}. Indeed, from Eq. (6) every proper subtree g of G_{i^*} is such that $[g] = [S_{|g}]$. Hence, if $S_{|G_{i^*}}$ and G_{i^*} were defined over the same set of taxa, then we would have $[G_{i^*}] = [S_{|G_{i^*}}]$ in contrast with Eq. (6). Without loss of generality, we can thus assume that

$$\{\text{taxa of } \left(S_{|G_{i^*}}\right)_\ell\} \nsubseteq \{\text{taxa of } G_{i^*}\}. \tag{7}$$

For instance, in Fig. 3A the taxa of G_{i^*} are C and D, whereas $\{A,B,C\}$ is the set of taxa of $\left(S_{|G_{i^*}}\right)_\ell$.

Starting from a ranked gene tree topology G such that $[G] \neq [S]$, we now describe how G is transformed into the ranked gene tree topology G'. To better follow the steps of the procedure, it is important that we fix a planar representation of the ranked species tree S and consider G realized in S according to one of the ranked histories of G in S. For instance, one can take the realization of G in S associated with the maximal ranked history of $H(G,S)$ (Fig. 3A). The construction of G' from G proceeds by the following three steps (Fig. 3 and 4):

(i) Locate the branch α of S that coalesces with the root branch of $S_{|(G_{i^*})_\ell}$, where the left/right orientation of G_{i^*} is induced by the orientation chosen for S by defining $(G_{i^*})_\ell$ as the root subtree of G_{i^*} that contains those taxa of G_{i^*} that belong to $\left(S_{|G_{i^*}}\right)_\ell$. As an example, consider the ranked species tree S and ranked gene tree topology G depicted at the top of panel A in Fig. 3. G_{i^*} is the subtree of G rooted at the white node γ_{i^*}. $S_{|G_{i^*}}$ has set of taxa $\{A,B,C,D\}$, with $((A,B),C)$ being its left root subtree $(S_{|G_{i^*}})_\ell$. Hence, $(G_{i^*})_\ell$ has its set of taxa given by $\{C\}$, from which we obtain that $S_{|(G_{i^*})_\ell}$ consists only of the taxon C. The branch α that coalesces with the root branch of $S_{|(G_{i^*})_\ell} = C$ is therefore the (grey) branch of S that connects the root node of the clade (A,B) with the root node of the clade $((A,B),C)$. In each panel of Fig. 3 as well as in Fig. 4, α is depicted as the grey branch of the ranked species tree. Note that from Eq. (6) we have $[S_{|(G_{i^*})_\ell}] = [(G_{i^*})_\ell]$. By Eq. (7) it thus follows that $S_{|(G_{i^*})_\ell}$ is *properly* contained in $\left(S_{|G_{i^*}}\right)_\ell$, and therefore α must be a branch of $\left(S_{|G_{i^*}}\right)_\ell$ (Fig. 3 and 4).

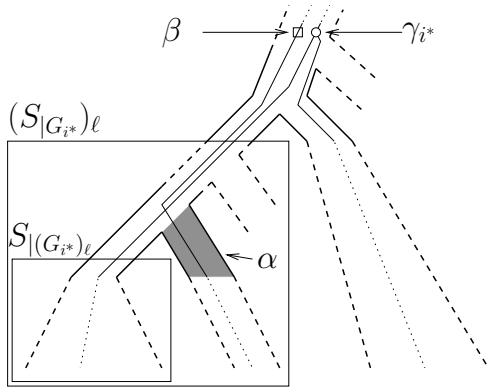

Figure 4: Schematic illustration of the node γ_{i^*} of G (white node), the branch α of S (grey branch), and the point β on G (white square) used for constructing the ranked gene tree topology G' from G. For obtaining G', we let the root lineage of $(G_{i^*})_\ell$ join the point β instead of coalescing with the root lineage of $(G_{i^*})_r$.

(ii) Identify a point $\beta \neq \gamma_{i^*}$ on a lineage of G such that β is contemporary with—at the same height of—γ_{i^*} with respect to the considered realization of G in S and all the taxa descending from β are below the branch α of S determined in step (i). In Fig. 3 and 4, the white square indicates possible choices for the point β contemporary with γ_{i^*}, the latter being represented by the white node in the figure. A proof of the existence of such a β is given in the Appendix.

(iii) Define G' by letting the root lineage of $(G_{i^*})_\ell$ join the point β of G found in step (ii), instead of coalescing with the root lineage of $(G_{i^*})_r$. In doing so, we replace the internal node γ_{i^*} of G with an internal node of the same rank in G' placed at the position determined by the point β of G and whose set of descending taxa consists of the taxa of $(G_{i^*})_\ell$ and the taxa descending from β in G. In the sequence of ranked gene tree topologies $(G^{(r)})_r = (G^{(0)}, G^{(1)}, G^{(2)}, G^{(3)})$ depicted at the bottom of panels A, B, C, and D of Fig. 3, the arrow connects the root lineage of $\big((G^{(r)})_{i^*}\big)_\ell$ to the point β (white square) indicating the coalescence event that in $G^{(r+1)} \equiv (G^{(r)})'$ replaces the node γ_{i^*} of $G^{(r)}$ (white node). In general, we observe that as depicted in Fig. 4 the taxa of $(G_{i^*})_\ell$ and those descending from β all belong to $\big(S_{|G_{i^*}}\big)_\ell$. Thus, $S_{|(G')_{i^*}}$—where $i^* = i^*(G)$—must be a (not necessarily proper) subtree of $\big(S_{|G_{i^*}}\big)_\ell$ (Fig. 4).

Remark. It is important to observe that the described construction of the ranked gene tree topology G' depends only on G and S as ranked topologies. In particular, considering different realizations of G in S does not affect the output ranked topology G'. However, note that when the tree operator $(\cdot)'$ is applied to two different realizations of G in S, the resulting realizations of G' in S are different. More precisely, as detailed in the proof of the next lemma, if we start from a realization of G in S associated with a ranked history $h \in H(G,S)$, then the resulting realization of G' in S is associated with a ranked history $h' \in H(G',S)$ such that $h = h'$. For example, in Fig. 3A the ranked gene tree topology G is realized according to the ranked history $(1,1,2,2)$. By iteratively applying the tree operator $(\cdot)'$, the resulting ranked histories of $G^{(1)}, G^{(2)}, G^{(3)}$, and $G^{(4)}$ remain equal to the starting ranked history $(1,1,2,2)$ chosen for G.

3.1.2 The probability of G' and the existence of G^*

We describe some important features (Lemma 2 and 3) of the tree operator introduced in the previous section, and use them for proving that the set $\mathcal{G}_S$ of maximally probable ranked gene tree topologies of a ranked species tree S contains at least one element with the same unranked topology of S (Theorem 1).

Lemma 2 *Let S be a ranked species tree, and let G be a ranked gene tree topology for S such that $[G] \neq [S]$. Then $Prob(G|S) \leq Prob(G'|S)$. Furthermore, if $H(G,S) = H(G',S)$, then $Prob(G|S) = Prob(G'|S)$.*

Proof. Fix an arbitrary ranked history $h \in H(G,S)$ and consider G realized in S according to h (e.g. Fig. 3A). Consider the point β of G defined in step (ii) of the construction of G' from G. Note that β must belong to the same branch b of S that contains the node γ_{i^*} of G. Indeed, since γ_{i^*} is in b, then b has to be ancestral or equal to the root branch of $S_{|G_{i^*}}$, which is in turn ancestral to α as this is a branch of $(S_{|G_{i^*}})_\ell$ (step (i)

of Section 3.1.1). Thus, b is ancestral to α in S. Therefore, any point of G contemporary with γ_{i^*} in S that is not located in the branch b of S cannot have descending lineages passing through α. By definition, β is contemporary with γ_{i^*} in S with descending taxa below α. Hence, β must be in b.

Because γ_{i^*} and β have the same height in S and belong to the same branch b of S, when we apply the operator $(\cdot)'$ to the realization of G in S associated with h, the resulting realization of G' in S is associated with a ranked history h' such that $m_i(h) = m_i(h')$ and $k_{i,j,z}(G,h) = k_{i,j,z}(G',h')$ for all possible indices i, j, z. Indeed, the third step of the construction of G' acts inside that branch-zone of S where the branch b intersects the time interval of S in which γ_{i^*} occurs, replacing the coalescence event γ_{i^*} of G with a contemporary coalescence event that takes place at the position determined by the point β (Fig. 3). This replacement does not alter either the number of coalescence events happening in the different time intervals of S or the number of gene lineages present in the different branch-zones of S after each coalescence event (Fig. 3). In particular, observe that $h' = h$ as $m_i(h) = m_i(h')$ for all possible values of i.

The same argument applied to all the ranked histories h of the ranked gene tree topology G shows that G and the ranked gene tree topology G' satisfy the hypothesis (i) and (ii) of Lemma 1 with $G_1 \equiv G$ and $G_2 \equiv G'$, from which $\text{Prob}(G|S) \leq \text{Prob}(G'|S)$. Moreover, if $H(G,S) = H(G',S)$, then by the same Lemma 1 we obtain $\text{Prob}(G|S) = \text{Prob}(G'|S)$. □

Remark. From the proof of the previous lemma, we see that the property of G' of having at least the same probability as G—or exactly the probability of G when $H(G,S) = H(G',S)$—depends only on the ranked topology of S. In other words, we have $\text{Prob}(G|S) \leq \text{Prob}(G'|S)$—or $\text{Prob}(G|S) = \text{Prob}(G',S)$—for all possible numerical values of the lengths (t_i) of the time intervals of S. For instance, let us consider the ranked species tree S and the ranked gene tree topologies $G^{(1)}$ and $G^{(2)}$ depicted in Fig. 3B and C. We use our software `RGTProb` for calculating the conditional probability $\text{Prob}(G^{(1)}|S)$ and $\text{Prob}(G^{(2)}|S)$ as a symbolic function of the lengths t_2, t_3, and t_4 of the time intervals of S. Subtracting the first probability from the second yields $\text{Prob}(G^{(2)}|S) - \text{Prob}(G^{(1)}|S) = \frac{1}{18}\, e^{-3t_2-t_3-t_4}(3\, e^{2\, t_2} - 2)(e^{t_4} - 1)$, which is larger than or equal to 0 for all values of t_2, t_3, t_4.

We now show that, starting from a ranked gene tree topology G such that $[G] \neq [S]$, the iterative application (5) of the tree operator $(\cdot)'$ yields, at some point, a ranked gene tree topology with the same unranked topology of S. We need two preliminary observations.

First, note that

$$i^*(G') \leq i^*(G). \tag{8}$$

Indeed, in G' all subtrees generated by a node whose rank is strictly larger than $i^*(G)$ are left as in G, that is, $(G')_i = G_i$ for all $i > i^*(G)$. In Fig. 3, we have $\big(i^*(G^{(0)}), i^*(G^{(1)}), i^*(G^{(2)}), i^*(G^{(3)})\big) = (4,4,3,2)$, which also shows that $i^*(G')$ can be equal to $i^*(G)$.

Second, as noticed in step (iii) of Section 3.1.1, $S_{|(G')_{i^*(G)}}$ is contained in $\big(S_{|G_{i^*(G)}}\big)_\ell$, from which

$$|S_{|(G')_{i^*(G)}}| < |S_{|G_{i^*(G)}}|. \tag{9}$$

For example, in Fig. 3 we have $|S_{|(G^{(1)})_{i^*(G^{(0)})}}| = 3 < 4 = |S_{|(G^{(0)})_{i^*(G^{(0)})}}|, |S_{|(G^{(2)})_{i^*(G^{(1)})}}| = 2 < 3 = |S_{|(G^{(1)})_{i^*(G^{(1)})}}|, |S_{|(G^{(3)})_{i^*(G^{(2)})}}| = 3 < 4 = |S_{|(G^{(2)})_{i^*(G^{(2)})}}|$, and $|S_{|(G^{(4)})_{i^*(G^{(3)})}}| = 4 < 5 = |S_{|(G^{(3)})_{i^*(G^{(3)})}}|$.

From these facts, we have the next result.

Lemma 3 *Let S be a ranked species tree, and let G be a ranked gene tree topology for S such that $[G] \neq [S]$. Then there exists a finite index $q > 0$ such that $[G^{(q)}] = [S]$.*

Proof. Consider the sequence $G \equiv G^{(0)}, G^{(1)}, ..., G^{(r)}, G^{(r+1)}, ...$ of ranked gene tree topologies obtained by iteratively applying the tree operator $(\cdot)'$ starting from G. In this sequence, as long as $[G^{(r)}] \neq [S]$, we construct $G^{(r+1)} \equiv \big(G^{(r)}\big)'$ by following steps (i), (ii), and (iii) of the construction described in the previous section. We show that the assumption

$$[G^{(r)}] \neq [S] \;\; \forall r > 0 \tag{10}$$

yields a contradiction.

For each fixed index $r \geq 0$, define $i_r^* \equiv i^*(G^{(r)})$ as in Eq. (6), where $1 \leq i_r^* \leq |S| - 1$. Notice that $i_{r+1}^* \leq i_r^*$ from Eq. (8). Furthermore, the number of taxa in $S_{|(G^{(r+1)})_{i_r^*}}$ must be strictly smaller than the number of taxa

in $S_{|(G^{(r)})_{i_r^*}}$, as it follows directly from Eq. (9). In particular, by iteratively making use of the latter property, if $i_r^* = i_{r+1}^* = ... = i_{r+w}^*$ for a given integer $w \geq 0$, then $S_{|(G^{(r+s+1)})_{i_r^*}}$ has strictly less taxa than $S_{|(G^{(r+s)})_{i_r^*}}$ for every integer $0 \leq s \leq w$, that is,

$$2 \leq |S_{|(G^{(r+w+1)})_{i_r^*}}| < |S_{|(G^{(r+w)})_{i_r^*}}| < \cdots < |S_{|(G^{(r+1)})_{i_r^*}}| < |S_{|(G^{(r)})_{i_r^*}}|.$$

Therefore, because the number of taxa in $S_{|(G^{(r)})_{i_r^*}}$ is a finite quantity, there must exist an index $r' > r$ such that $i_{r'}^* \neq i_r^*$, that is, $i_{r'}^* < i_r^*$.

Finally, consider the sequence of integers $(i_r^*)_{r \geq 0}$. This sequence contains infinitely many terms, because we are assuming Eq. (10). Moreover, as we have just shown, for every $r \geq 0$ there exists $r' > r$ with $i_{r'}^* < i_r^*$. Hence, $i_r^* \to -\infty$ for increasing values of r, which is in contrast with the fact that $i_r^* \geq 1$ for all r. From Eq. (10) we have thus reached a contradiction, showing that there must exist a finite index $q > 0$ such that $[G^{(q)}] = [S]$. □

Suppose G is a ranked gene tree topology for S such that $[G] \neq [S]$. Let G^* denote the ranked gene tree topology $G^{(q)}$ found by the previous lemma. Hence, $[G^*] = [S]$ and, from Lemma 2, $\text{Prob}(G|S) \leq \text{Prob}(G^*|S)$. We can now prove our main result.

Theorem 1 *For every ranked species tree S, there exists a ranked gene tree topology $G \in \mathcal{G}_S$ with $[G] = [S]$.*

Proof. Suppose $G \in \mathcal{G}_S$ is such that $[G] \neq S$. Consider the ranked gene tree topology G^*. Since $G \in \mathcal{G}_S$, we must have $G^* \in \mathcal{G}_S$ as well. Furthermore, $[G^*] = [S]$. Hence, G^* is a ranked gene tree topology of maximal probability whose unranked topology matches that of S. □

Remark. The set $\mathcal{G}_S$ of the maximally probable ranked gene tree topologies for a ranked species tree S is not independent of the numerical values chosen for the lengths (t_i) of the time intervals of S. By changing these values, the argument used in the proof of the latter proposition will in general lead to identify different ranked gene tree topologies with the property of having maximal probability and same unranked topology as S.

3.2 Existence of a maximally probable ranked gene tree topology G with $[G] \neq [S]$

In Section 3.1, we have shown that for every ranked species tree S, the set $\mathcal{G}_S$ of maximally probable ranked gene tree topologies for S contains a tree whose unranked topology matches that of S. Here, we exhibit a ranked species tree S such that the set $\mathcal{G}_S$ has also one element with an unranked topology that differs from that of S.

Let us consider the ranked species tree S with ranked topology $((((A,B)_4,C)_3,D)_2,((E,F)_6,G)_5)_1$ and interval lengths $(t_2,t_3,t_4,t_5,t_6) = (10, 0.001, 0.001, 0.001, 0.001)$ (Fig. 5). From Proposition 1, the set of maximally probable ranked gene tree topologies for S must contain a ranked gene tree topology with the same unranked topology of S. Since S has a caterpillar with 3 internal nodes and a caterpillar with 2 internal nodes as its root subtrees, there are exactly $\binom{5}{2} = 10$ possible ranked gene tree topologies with the same unranked topology of S. By using our software `RGTProb`, we have calculated the conditional probability of each one of these ten ranked gene tree topologies. Results are reported in Table 2, where the ranked gene tree topology with the ith largest probability is denoted by G_i. Among the ten probability values listed in the table, the first four are equal up to the 7th decimal digit at least—too close to identify which one among G_1, G_2, G_3, and G_4 has the greatest probability without the danger of relying too much upon numerical calculations. The difference between the fourth and the fifth probability value is instead large enough to conclude that at least one among G_1, G_2, G_3, and G_4 has to be maximally probable for the chosen S.

For the selected ranked species tree S, the existence of a maximally probable ranked gene tree topology with a different unranked topology follows by observing that, for each one of G_1, G_2, G_3, and G_4, there exists a ranked gene tree topology with the same conditional probability and with a different unranked topology. In other words, there exist four ranked gene tree topologies for S, $\tilde{G}_1, \tilde{G}_2, \tilde{G}_3$, and $\tilde{G}_4$, such that $\text{Prob}(G_j|S) = \text{Prob}(\tilde{G}_j|S)$ and $[\tilde{G}_j] \neq [G_j] = [S]$, for every $j = 1, 2, 3, 4$. Indeed, let us consider

$$\tilde{G}_1 \equiv ((((A,B)_6,C)_4,D)_2,((E,G)_5,F)_3)_1, \qquad \tilde{G}_2 \equiv ((((A,B)_6,C)_4,D)_3,((E,G)_5,F)_2)_1,$$
$$\tilde{G}_3 \equiv ((((A,B)_6,C)_5,D)_2,((E,G)_4,F)_3)_1, \qquad \tilde{G}_4 \equiv ((((A,B)_6,C)_5,D)_3,((E,G)_4,F)_2)_1.$$

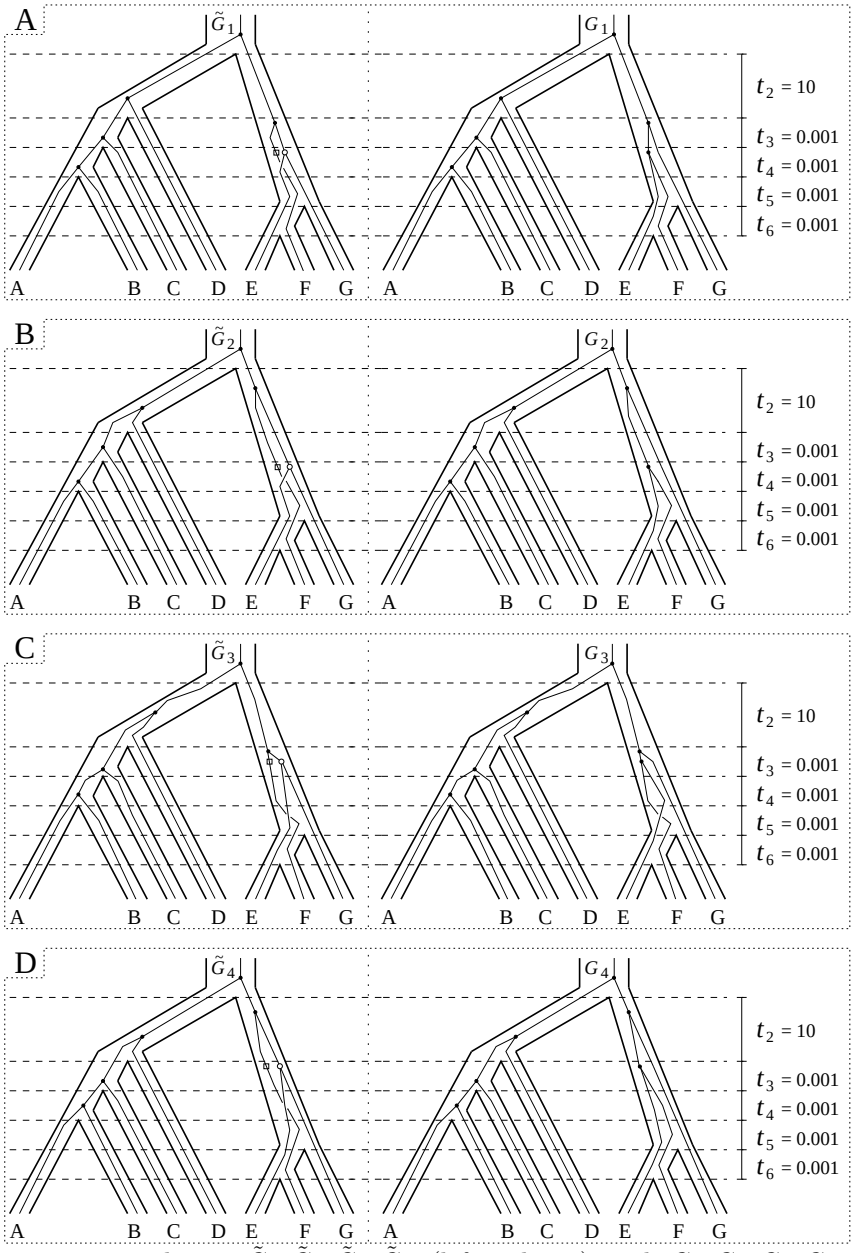

Figure 5: Ranked gene tree topologies $\tilde{G}_1, \tilde{G}_2, \tilde{G}_3, \tilde{G}_4$ (left column) and G_1, G_2, G_3, G_4 (right column) realized according to their maximal ranked history in the ranked species tree S (thicker tree) with ranked topology $((((A, B)_4, C)_3, D)_2, ((E, F)_6, G)_5)_1$. When $(t_2, t_3, t_4, t_5, t_6) = (10, 0.001, 0.001, 0.001, 0.001)$, Table 2 gives the conditional probabilities of G_1, G_2, G_3, and G_4. For $j = 1, 2, 3, 4$, G_j and $\tilde{G}_j$ have the same probability but different unranked topology.

Table 2: Probability of the ranked gene tree topologies whose unranked topology matches that of the ranked species tree of Fig. 5.

Ranked gene tree topology G_i	Prob($G_i\|S$)	Ranked gene tree topology G_i	Prob($G_i\|S$)
$G_1 \equiv ((((A,B)_6,C)_4,D)_2,((E,F)_5,G)_3)_1$	.00232378785813527	$G_6 \equiv ((((A,B)_5,C)_4,D)_3,((E,F)_6,G)_2)_1$	.00202778426225364
$G_2 \equiv ((((A,B)_6,C)_4,D)_3,((E,F)_5,G)_2)_1$	.00232378785768080	$G_7 \equiv ((((A,B)_6,C)_3,D)_2,((E,F)_5,G)_4)_1$	.00154923407574531
$G_3 \equiv ((((A,B)_6,C)_5,D)_2,((E,F)_4,G)_3)_1$	.00232372557171676	$G_8 \equiv ((((A,B)_6,C)_5,D)_4,((E,F)_3,G)_2)_1$	.00154919227511797
$G_4 \equiv ((((A,B)_6,C)_5,D)_3,((E,F)_4,G)_2)_1$	.00232372557151044	$G_9 \equiv ((((A,B)_5,C)_3,D)_2,((E,F)_6,G)_4)_1$	.00135189222396372
$G_5 \equiv ((((A,B)_5,C)_4,D)_2,((E,F)_6,G)_3)_1$	.00202778426403491	$G_{10} \equiv ((((A,B)_4,C)_3,D)_2,((E,F)_6,G)_5)_1$	.000901352849789283

The probabilities have been calculated by using our software `RGTProb` conditioning on the species tree S of Fig. 5 that has ranked topology $((((A,B)_4,C)_3,D)_2,((E,F)_6,G)_5)_1$ and interval lengths $(t_2,t_3,t_4,t_5,t_6) = (10, 0.001, 0.001, 0.001, 0.001)$. In the Supplementary Material, we provide explicit formulas with symbolic variables t_2, t_3, t_4, t_5, and t_6 for the conditional probability Prob($G_i|S$) ($1 \leq i \leq 10$).

Panels A,B,C, and D of Fig. 5 depict the pairs of ranked gene tree topologies $\tilde{G}_1$ and G_1, $\tilde{G}_2$ and G_2, $\tilde{G}_3$ and G_3, and $\tilde{G}_4$ and G_4, respectively, as realized in the ranked species S according to their maximal ranked history. In particular, note that the maximal ranked history $\tilde{h}_j^*$ of $\tilde{G}_j$ in S coincides with the maximal ranked history h_j^* of G_j for every $j = 1,2,3,4$. More precisely, $\tilde{h}_1^* = h_1^* = (1,2,3,3,4,4), \tilde{h}_2^* = h_2^* = (1,2,2,3,4,4), \tilde{h}_3^* = h_3^* = (1,2,3,3,3,4)$, and $\tilde{h}_4^* = h_4^* = (1,2,2,3,3,4)$. Because the set of ranked histories of a ranked gene tree topology G in a ranked species tree S is determined by the maximal ranked history of G in S (Section 2.1), we have that $H(\tilde{G}_j, S) = H(G_j, S)$ for every $j = 1,2,3,4$. Furthermore, observe that the ranked gene tree topology G_j can be obtained from the ranked gene tree topology $\tilde{G}_j$ by a single application of the operator $(\cdot)'$ described in Section 3.1.1, that is, $G_j = (\tilde{G}_j)'$ for every $j = 1,2,3,4$. Indeed, Fig. 5 shows that G_j can be obtained from $\tilde{G}_j$ by replacing the coalescence event identified by the white node—i.e. γ_{i^*} of $\tilde{G}_j$—with the coalescence of the gene lineage ancestral to taxon E—i.e. the root lineage of $\big((\tilde{G}_j)_{i^*}\big)_\ell$—and the gene lineage ancestral to taxon F at the point identified by white square—i.e. the point β of $\tilde{G}_j$. From Lemma 2, we thus have Prob($G_j|S$) = Prob($\tilde{G}_j|S$) for every $j = 1,2,3,4$, as claimed above.

As noticed in the remark following Lemma 2, for every $j = 1,2,3,4$ the ranked gene tree topologies G_j and $\tilde{G}_j$ must have the same conditional probability for all possible numerical values of the lengths $(t_i)_i$ of the time intervals of S—not only when we set $(t_2,t_3,t_4,t_5,t_6) = (10, 0.001, 0.001, 0.001, 0.001)$. Indeed, considering t_2, t_3, t_4, t_5, and t_6 as variables, `RGTProb` returns equivalent formulas for the conditional probabilities Prob($G_j|S$) and Prob($\tilde{G}_j|S$).

4 Discussion

When one gene copy is sampled for each species, we have considered ranked gene tree topologies realized in ranked species trees under the multispecies coalescent model. In particular, in this paper we have addressed a problem left open by Degnan et al. (2012a) on determining whether, for a given ranked species tree S, the most probable ranked gene tree topologies disagree with S in their unranked topology. Theorem 1 shows that, for any ranked species tree S, the set $\mathcal{G}_S$ of the maximally probable ranked gene tree topologies for S contains at least a tree whose unranked topology matches that of S. Equivalently, for a ranked species tree S for which anomalous ranked gene tree topologies exist, there is an anomalous ranked gene tree topology of maximal probability with a matching unranked topology.

The proof of Theorem 1 makes use of a tree operator $(\cdot)'$ that, from a ranked gene tree topology G whose unranked topology differs from that of the fixed ranked species tree S, produces a ranked gene tree topology G' with at least the same conditional probability of G (Lemma 2) and whose unranked topology is "closer" to that of S than it was the unranked topology of G. In particular, the sequence of ranked gene tree topologies $G, G', G'', G''', ...$, which is obtained by iteratively applying the operator $(\cdot)'$ starting from G, terminates at some point with a ranked gene tree topology G^* that is at least as likely as G and whose unranked topology matches that of S (Lemma 3). Notably, the fact that G'—and thus G^*—has at least the same conditional probability of G is true for every set of numerical values that can be assigned to the lengths of the time

intervals of S. In particular, the equality between the conditional probability of G and G' holds exactly when $H(G,S) = H(G',S)$, whereas in general we have $H(G,S) \subseteq H(G',S)$.

In Section 3.2, by using our software `RGTProb`, we have identified a ranked species tree S of relatively small size for which the set $\mathcal{G}_S$ contains also an element whose unranked topology differs from that of S. This shows that in general the set of maximally probable ranked gene tree topologies for a ranked species tree can consist of ranked gene tree topologies with different unranked topologies.

Further works will investigate the variability of the ranked gene tree topologies belonging to the set $\mathcal{G}_\mathcal{S}$ associated with a given ranked species tree S. For instance, in this direction, preliminary results indicate the existence of ranked species trees with multiple ranked gene tree topologies of maximal probability that have the same unranked topology of the species tree. In other words, we observe that the property of having the same unranked topology as the ranked species tree does not always identify a unique ranked gene tree topology among those of maximal probability.

Results of this paper may find application to the study of the inference of species trees from gene trees. A possible procedure for estimating ranked species trees from ranked gene tree topologies can be derived by following a maximum likelihood approach (Stadler and Degnan, 2012). Roughly speaking, given a collection, or multiset, $\mathcal{G} = \{G_1, \ldots, G_N\}$ of ranked gene tree topologies inferred at N different loci, a maximum likelihood estimate of the "true" unknown ranked species tree S is $\hat{S} = \text{argmax}_{S \in \mathcal{S}} \prod_{i=1}^{N} \text{Prob}(G_i|S)$, where $\mathcal{S}$ is a certain subset of the ranked species tree space (or the entire space). If we assume that the most probable ranked gene tree topologies for S are those appearing with higher frequency in the collection $\mathcal{G}$, a direct application of Theorem 1 gives a possible criterion for defining $\mathcal{S}$. The latter can be taken as the set of ranked species trees whose unranked topology matches the unranked topology of at least one of the most frequent ranked gene tree topologies in $\mathcal{G}$.

Moreover, one can consider to use the tree operator $(\cdot)'$ introduced in this article to further investigate the agreement between the ranked topology of the estimated $\hat{S}$ and the ranked topology of S, with respect to the collection $\mathcal{G}$ of observed ranked gene tree topologies. In theory, if $\hat{S}$ and S share the same ranked topology, then for each ranked gene tree topology $G \in \mathcal{G}$ such that $[G] \neq [\hat{S}]$, the ranked gene tree topology G'—derived from G considering $\hat{S}$ as the ranked species tree—should appear in $\mathcal{G}$ at least as many times as G, if the number N of loci is sufficiently large. Indeed, Lemma 2 shows that $\text{Prob}(G|\hat{S}) \leq \text{Prob}(G'|\hat{S})$ for all possible choices of the lengths of the time intervals of $\hat{S}$, and thus $\text{Prob}(G|S) \leq \text{Prob}(G'|S)$ if $\hat{S}$ and S have the same ranked topology. In practice, if G and G' are found to have respectively a high and low frequency in $\mathcal{G}$, then $\hat{S}$ gives probably a wrong estimate of the ranked topology of the species tree S. In this case, a possibility is to proceed by running again the maximum likelihood procedure sketched above, once all trees with the same ranked topology of $\hat{S}$ have been removed from the examined subset $\mathcal{S}$ of the ranked species tree space.

Acknowledgments We thank Noah A. Rosenberg for discussions. Support was provided by a Rita Levi-Montalcini grant to FD from the Ministero dell'Istruzione, dell'Università e della Ricerca.

Author contributions Wrote the paper: FD. Designed the study: FD, PM, and GN. Derived theorems: PM and GN. Implemented software: PM and GN.

Appendix 1. Proof of the existence of β

We demonstrate the existence of a point β of G as defined in step (ii) of the construction described in Section 3.1.1. In particular, for a fixed realization of G in S, we seek a point $\beta \neq \gamma_{i^*}$ contemporary with γ_{i^*}, i.e. at the same height of γ_{i^*} in S, lying on a lineage of G that has all its descending taxa below the branch α of S determined in step (i) of the same construction.

First, observe that the node γ_{i^*} of G cannot have any of its descending taxa placed below α. Indeed, from step (i), α is a branch of $(S_{|G_{i^*}})_\ell$ external to $S_{|(G_{i^*})_\ell}$—in fact, α coalesces with the root branch of $S_{|(G_{i^*})_\ell}$ (Fig. 4). Hence, none of the lineages of G_{i^*} passes through α.

Second, with respect to the considered realization of G in S, take any lineage (branch) ρ of G containing a point p contemporary with γ_{i^*} and such that *at least one* of the taxa descending from ρ is placed below the branch α. Such a lineage must clearly exists and, as shown above, it cannot contain the node γ_{i^*} of G, i.e. $p \neq \gamma_{i^*}$. We claim that p has *all* its descending taxa below α, and thus it satisfies the definition of β. Suppose a contrario that from p descends a taxon of G that is not placed below α. Since p has also a descending taxon below α, there must be at least one internal node of G descending from p. Let γ_v be the most ancient internal

node of G descending from p, where $v > i^*$. As for the point p, the node γ_v has some descending taxa below α and some others that are not placed below α. Hence, since α is the branch of S that coalesces with the root branch of $S_{|(G_{i^*})_\ell}$ (Fig. 4), $[S_{|G_v}]$ strictly contains $[S_{|(G_{i^*})_\ell}]$. Note that $[S_{|G_v}] = [G_v]$, as $v > i^*$, and $[S_{|(G_{i^*})_\ell}] = [(G_{i^*})_\ell]$. Therefore, the set of taxa of G_v strictly contains the set of taxa of $(G_{i^*})_\ell$, which implies that γ_v must be equal or ancestral to γ_{i^*} in G. This is in contrast with $v > i^*$. □

References

Degnan, J. H., M. DeGiorgio, D. Bryant, and N. A. Rosenberg (2009). Properties of consensus methods for inferring species trees from gene trees. *Syst. Biol. 58*, 35–54.

Degnan, J. H. and N. A. Rosenberg (2006). Discordance of species trees with their most likely gene trees. *PLoS Genet. 2*, 762–768.

Degnan, J. H. and N. A. Rosenberg (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol. 24*, 332–340.

Degnan, J. H., N. A. Rosenberg, and T. Stadler (2012a). A characterization of the set of species trees that produce anomalous ranked gene trees. *IEEE/ACM Trans. Comp. Biol. Bioinf. 9*, 1558–1568.

Degnan, J. H., N. A. Rosenberg, and T. Stadler (2012b). The probability distribution of ranked gene trees on a species tree. *Math. Biosci. 235*, 45–55.

Degnan, J. H. and L. A. Salter (2005). Gene tree distributions under the coalescent process. *Evolution 59*, 24–37.

Disanto, F. and N. A. Rosenberg (2014). On the number of ranked species trees producing anomalous ranked gene trees. *IEEE/ACM Trans. Comp. Biol. Bioinf. 11*, 1229–1238.

Ewing, G. B., I. Ebersberger, H. A. Schmidt, and A. von Haeseler (2008). Rooted triple consensus and anomalous gene trees. *BMC Evol. Biol. 8*, 118.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sunderland, MA: Sinauer.

Heled, J. and A. J. Drummond (2010). Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol. 27*, 570–580.

Hudson, R. R. (1983). Testing the constant-rate neutral allele model with protein sequence data. *Evolution 37*, 203–217.

Kubatko, L. S., B. C. Carstens, and L. L. Knowles (2009). Stem: Species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics 25*, 971–973.

Liu, L. and D. K. Pearl (2007). Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst. Biol. 56*, 504–514.

Liu, L. and L. Yu (2011). Estimating species trees from unrooted gene trees. *Syst. Biol. 60*, 661–667.

Liu, L., L. Yu, and D. K. Pearl (2010). Maximum tree: a consistent estimator of the species tree. *J. Math. Biol. 60*, 95–106.

Liu, L., L. Yu, D. K. Pearl, and S. V. Edwards (2009). Estimating species phylogenies using coalescence times among sequences. *Syst. Biol. 58*, 468–477.

Maddison, W. P. (1997). Gene trees in species trees. *Syst. Biol. 46*, 523–536.

Maddison, W. P. and L. L. Knowles (2006). Inferring phylogeny despite incomplete lineage sorting. *Syst. Biol. 55*, 21–30.

Mossel, E. and S. Roch (2010). Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *IEEE/ACM Trans. Comp. Biol. Bioinf. 7*, 166–171.

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

Rosenberg, N. A. (2002). The probability of topological concordance of gene trees and species trees. *Theor. Pop. Biol. 61*, 225–247.

Rosenberg, N. A. (2006). The mean and variance of the numbers of r-pronged nodes and r-caterpillars in Yule-generated genealogical trees. *Ann. Comb. 10*, 129–146.

Stadler, T. and J. H. Degnan (2012). A polynomial time algorithm for calculating the probability of a ranked gene tree given a species tree. *Alg. Mol. Biol. 7*, 7.

Than, C. and L. Nakhleh (2009). Species tree inference by minimizing deep coalescences. *PLoS Comp. Biol. 5*, e1000501.

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