A Characterization of the Set of Species Trees that Produce Anomalous Ranked Gene Trees

James H. Degnan, Noah A. Rosenberg, and Tanja Stadler

Abstract—Ranked gene trees, which consider both the gene tree topology and the sequence in which gene lineages separate, can potentially provide a new source of information for use in modeling genealogies and performing inference of species trees. Recently, we have calculated the probability distribution of ranked gene trees under the standard multispecies coalescent model for the evolution of gene lineages along the branches of a fixed species tree, demonstrating the existence of *anomalous ranked gene trees* (ARGTs), in which a ranked gene tree that does not match the ranked species tree can have greater probability under the model than the matching ranked gene tree. Here, we fully characterize the set of unranked species tree topologies that give rise to ARGTs, showing that this set contains all species tree topologies with five or more taxa, with the exceptions of caterpillars and pseudocaterpillars. The results have implications for the use of ranked gene trees in phylogenetic inference.

Index Terms—Anomalous gene trees, coalescent, genealogies, phylogenetics, population genetics

1 Introduction

Under the "multispecies coalescent" [3], [4], [6], [12], [13], [17], a standard model for the evolution of gene trees along the branches of species trees, the topology most likely to be possessed by a random gene tree that evolves on a fixed species tree does not necessarily match the species tree topology [3]. Terming gene tree topologies with probability greater than that of the matching topology *anomalous gene trees* (AGTs), for each species tree topology with five or more taxa, and for asymmetric four-taxon species tree topologies, there exist species tree branch lengths that give rise to AGTs [3].

For a given gene tree topology, in examining the existence of AGTs, all possible sequences of coalescences that can produce the topology—all possible *rankings* or *labeled histories*—are combined into a single topology that is treated as unranked. Evaluations of the properties of AGTs then utilize computations of a formula under the multispecies coalescent for the probabilities of unranked gene tree topologies conditional on species trees [6], [21].

We have recently developed an analogous formula for the probabilities of *ranked* gene tree topologies conditional on species trees [5], [16]. Could there be anomalous *ranked* gene trees, or gene tree labeled histories more likely to be produced than the labeled history that matches the species tree? In a five-taxon example, we demonstrated the existence of 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 [5]. This result was surprising, as the simplest case for anomalous *unranked* gene trees—the four-taxon asymmetric species tree—does not produce ARGTs [5]. Given that ARGTs occur for five taxa, it is then natural to determine the extent to which the existence of ARGTs generalizes beyond the specific five-taxon case.

Here, we perform a complete characterization of the set of species trees that give rise to ARGTs, finding unexpectedly that ARGTs do in fact exist for most species tree topologies. In Section 2, we introduce notation. In Section 3, we state and then prove the main theorem. Section 4 considers a series of consequences of the main result, and the paper concludes with a discussion in Section 5.

2 Notation

In general, the notation and setup follow [5]. We highlight a few key concepts, and refer the reader to [5] for further details. Given a species tree $\mathcal{T}$—a binary rooted tree together with its edge lengths—the *ranked species tree* Ψ associated with $\mathcal{T}$ consists of the labeled topology ψ of $\mathcal{T}$ together with the order in which the speciation events (vertices) of $\mathcal{T}$ occur. We enumerate the vertices starting from the root, so that for a species tree with n species, the root vertex has rank 1, and the most recent interior vertex has rank $n-1$.

Ranked gene trees are defined analogously to ranked species trees. Given a species tree, a ranked gene tree *matches* the ranked species tree if both trees have the same labeled topology and the same ranking. An *anomalous ranked gene tree* is a ranked gene tree that does not match the ranked species tree and that has probability under the multispecies coalescent greater than that of the matching ranked gene tree. We say that a species tree topology (ranked or unranked) *produces* an ARGT if there exist speciation times such that the species tree with the given speciation times has at least one ARGT.

- *J.H. Degnan is with the Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch 8013, New Zealand. E-mail: James.Degnan@canterbury.ac.nz.*
- *N.A. Rosenberg is with the Department of Biology, Stanford University, Stanford, CA 94305-5020. E-mail: noahr@stanford.edu.*
- *T. Stadler is with the Institute of Integrative Biology, ETH Zürich, Universitatsstrasse 16, 8092 Zürich, Switzerland. E-mail: tanja.stadler@env.ethz.ch.*

Manuscript received 31 Mar. 2012; revised 19 July 2012; accepted 24 July 2012; published online 3 Aug. 2012.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2012-03-0080. Digital Object Identifier no. 10.1109/TCBB.2012.110.

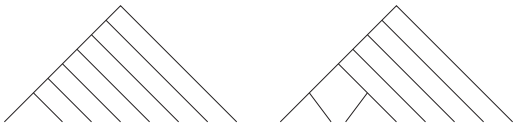

Fig. 1. A caterpillar tree (left) and a pseudocaterpillar tree (right).

We provide names for two special (unranked) topologies (Fig. 1). A *caterpillar* topology contains one interior vertex descended from all other interior vertices. A *pseudocaterpillar* topology has at least five leaves and contains one interior vertex v descended from all except two of the other interior vertices; these two exceptions both descend from v and neither descends from the other.

For a ranked species tree and a ranked gene tree, a *ranked history* is a list of the intervals on the species tree during which the coalescences in the gene tree take place (Fig. 2). Unlike coalescent histories, which identify locations of coalescences according to their species tree edges [6], [14], ranked histories instead locate coalescences by the time intervals during which they occur. The intervals are ordered so that the interval above the root is τ_1, and each subsequent interval is bracketed by two successive interior vertices; for each i from 1 to $n-1$, interval τ_i extends from the time s_{i-1} of vertex $i-1$ to the time s_i of vertex i (s_0 is infinite and s_i is smaller for larger values of i); we further set $t_i = s_{i-1} - s_i$. Thus, with n taxa, a ranked history for a gene tree and species tree is a vector of length $n-1$, whose ith component indicates the interval in which the ith coalescence, ordered forward in time, takes place. A given pair consisting of a gene tree and a species tree potentially has multiple possible ranked histories; given a species tree, a given ranked history can potentially apply to multiple gene trees. Further details regarding ranked histories appear in [5].

Under the multispecies coalescent model, gene trees are generated conditional on a fixed species tree, such that lineages follow separate coalescent processes along each branch of the species tree, and such that coalescences happen independently along separate concurrent branches [4]. The probability that n gene lineages coalesce to i lineages along a species tree branch of length t coalescent time units is a known function $g_{n,i}(t)$ [18]. This function is a linear combination of exponential terms, reflecting the occurrence in coalescent models of exponential distributions that represent waiting times to coalescences. We use here that $g_{n,n}(t) = e^{-\binom{n}{2}t}$ and $\lim_{t\to\infty} g_{n,1}(t) = 1$. Along a single species tree branch, the number of distinct sequences of events in which n lineages coalesce to i lineages is denoted h_n^i [5]. Under the multispecies coalescent, these sequences are equiprobable. We restrict our attention to cases with one gene lineage sampled for each leaf of the species tree.

3 General Existence Results for Anomalous Ranked Gene Trees

3.1 Review of the Five-Taxon Case

We have previously observed that although ARGTs do not exist for three-taxon or four-taxon species trees, they do exist for certain five-taxon species trees [5]. This observation relied on a particular way of ordering vertices in the species tree while lengthening one species tree time interval and shrinking the others.

In brief, for five-taxon species trees that have three taxa on one side of the root (the "left" side) and two on the other side (the "right" side), the species divergences can be ordered so that the single divergence on the right side occurs most recently (Fig. 2). By making this divergence and the two divergences on the left side occur over a short period of time, while setting the root of the species tree far back in the past, a scenario is generated in which all gene tree coalescences except the final one are likely to occur during the long time interval leading up to the root. The three possible sequences of gene tree coalescences—right-left-left (Fig. 2a), left-right-left (Fig. 2b), and left-left-right—are not equiprobable, and in fact, if the long branch leading to the root is sufficiently long and the remaining branches are sufficiently short, then the matching sequence, right-left-left, is the least probable.

The particular behavior that leads to ARGTs in our five-taxon example generalizes to larger trees. In particular, when coalescence events occur during the same time interval but in separate populations, different sequences of coalescences need not be equiprobable. As we will show below, scenarios

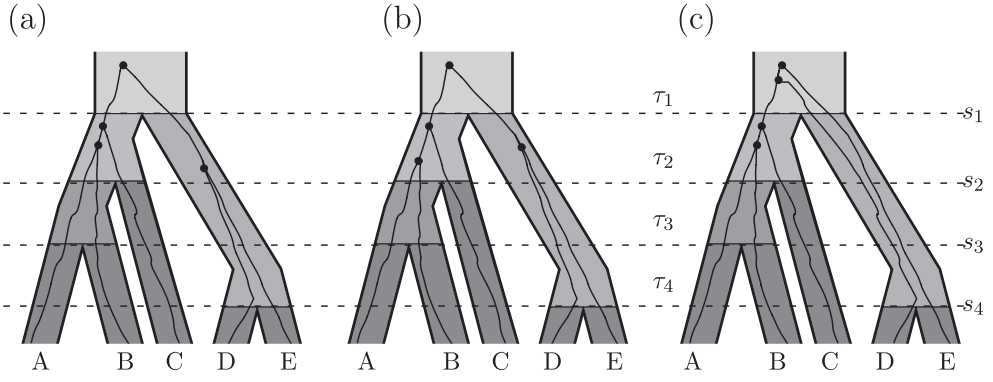

Fig. 2. A ranked species tree with the "right-left-left" sequence of divergences. This ranked labeled species tree is denoted $\mathcal{T}_{\mathrm{RLL}}$. (a) Unranked gene tree topology (((AB)C)(DE)) with ranked gene tree topology $(((\mathrm{AB})_2\mathrm{C})_3(\mathrm{DE})_4)$. (b) Unranked gene tree topology (((AB)C)(DE)) with ranked gene tree topology $(((\mathrm{AB})_2\mathrm{C})_4(\mathrm{DE})_3)$. (c) Unranked gene tree topology ((((AB)C)D)E). The unranked gene tree topologies are identical in (a) and (b); however, the ranked gene tree in (a) matches the ranked species tree, while the ranked gene tree in (b) does not match the ranked species tree. In both (a) and (b), the ranked history is (1,2,2,2), as the three most recent internal vertices of the gene tree occur in interval τ_2 and the root vertex occurs in interval τ_1. In (c), the ranked history is (1,1,2,2).

with different numbers of lineages in simultaneous but separate populations can be used to obtain a general characterization of the set of species tree unranked topologies that have at least one ranking that produces ARGTs.

3.2 Statement of the Main Result

We are now ready to state our main theorem.

Theorem 1. *1) Every noncaterpillar, nonpseudocaterpillar species tree topology with five or more taxa produces anomalous ranked gene trees. 2) Every species tree topology that is a caterpillar or pseudocaterpillar, or that has fewer than five taxa, produces no anomalous ranked gene trees.*

We first outline the proof, and proceed to demonstrate all of the steps. We first show that caterpillar and pseudocaterpillar species tree topologies do not produce ARGTs (Propositions 2 and 3). Three-taxon and four-taxon trees were treated in [5, Sections 5.1 and 5.2].

Otherwise, consider a noncaterpillar, nonpseudocaterpillar species tree topology with five or more taxa. Assuming without loss of generality that the number of "left" descendants of each node is always greater than or equal to the number of "right" descendants, we then identify a subtree of this topology that has $\ell \geq 3$ left-descendant leaves and $r \geq 2$ right-descendant leaves (Lemma 4). This step fails for caterpillars, pseudocaterpillars, and trees with fewer than five taxa.

We assign this subtree a ranking that is not "maximally probable" (in a sense described in Section 3.5). This assignment is always possible for noncaterpillar, nonpseudocaterpillar trees with five or more taxa (Corollary 8). Mimicking our earlier five-taxon work [5], we then choose branch lengths that lead to ARGTs for the subtree, keeping all branch lengths outside this subtree long enough that coalescences of gene lineages outside the subtree are likely to occur in the order suggested by the ranked species tree (Proposition 9 and Lemma 10). This step follows similar reasoning to that used in related proofs with unranked trees [2], [3], [20].

3.3 Caterpillars and Pseudocaterpillars

In this section, we consider the cases of caterpillars and pseudocaterpillars in Theorem 1 (part 2).

Proposition 2. *A caterpillar species tree has no anomalous ranked gene trees.*

Proof. Let Y be the set of ranked histories for nonmatching ranked gene tree $\mathcal{G}$, and let X be the set of ranked histories for the matching ranked gene tree $\mathcal{G}_{\text{cat}}$. We first show that for a given ranked history x and a gene tree $\mathcal{G}$ evolving on an n-taxon caterpillar species tree $\mathcal{T}_{\text{cat}}$, if $x \in Y$, then the probability $\mathbb{P}[\mathcal{G}, x \mid \mathcal{T}_{\text{cat}}]$ does not depend on $\mathcal{G}$. A caterpillar species tree $\mathcal{T}_{\text{cat}}$ has only one internal branch in interval τ_i, $i = 1, \ldots, n-1$, because only one branch in each interval is ancestral to more than one leaf of the species tree. Suppose x is a ranked history both for the matching ranked gene tree $\mathcal{G}_{\text{cat}}$ and for a nonmatching ranked gene tree $\mathcal{G}$. For ranked history x, let $\alpha_i(x)$ and $\beta_i(x)$, respectively, denote the numbers of lineages "entering" and "leaving" the one species tree internal branch in interval τ_i. For each $i \in \{1, \ldots, n-1\}$, using $\mathbb{P}_i$ to denote probability under the multispecies coalescent of the events in species tree interval τ_i,

$$\mathbb{P}_i[\mathcal{G}, x \mid \mathcal{T}_{\text{cat}}] = \mathbb{P}_i[\mathcal{G}_{\text{cat}}, x \mid \mathcal{T}_{\text{cat}}] = g_{\alpha_i(x),\beta_i(x)}(t_i)/h_{\alpha_i(x)}^{\beta_i(x)}. \quad (1)$$

The numerator is the probability that the correct number of coalescences occur in interval i, and the denominator is the number of possible coalescence sequences during the interval, all of which are equiprobable. The probability does not depend on the gene tree because both g and h depend only on the numbers of lineages $\alpha_i(x)$ and $\beta_i(x)$, which in turn depend only on the ranked history

$$\alpha_i(x) = n - i + 1 - \sum_{j=1}^{n-1} I(x_j > i) \quad (2)$$

$$\beta_i(x) = \alpha_i(x) - \sum_{j=1}^{n-1} I(x_j = i), \quad (3)$$

where $I(\cdot)$ is an indicator function that equals 1 if the condition obtains and 0 otherwise. Y is a proper subset of X (because the full set X of ranked histories is the set of ranked histories for the matching ranked gene tree, and $(1, 2, \ldots, n-1) \notin Y$ [5]).

The probability of a ranked gene tree is the sum of the probabilities of its ranked histories. Because each ranked history shared by $\mathcal{G}$ and $\mathcal{G}_{\text{cat}}$ has the same probability for each gene tree, and because $\mathcal{G}_{\text{cat}}$ has more ranked histories, the probability of $\mathcal{G}_{\text{cat}}$ is strictly greater than the probability of any other ranked gene tree $\mathcal{G}$. □

Proposition 3. *A pseudocaterpillar species tree has no anomalous ranked gene trees.*

Proof. Without loss of generality, an n-taxon pseudocaterpillar species tree $\mathcal{T}_{\text{pseudo}}$ has unranked topology $(((((\ldots(\text{AB})(\text{CD}))\text{E}_{n-4})\text{E}_{n-3})\ldots)\text{E}_1)$, where (AB) has rank $n-1$ and (CD) has rank $n-2$, so that (AB) is the most recent divergence on the species tree. If the ranked gene tree does not match the species tree, then exactly five cases exist for the locations of the coalescences in the gene tree:

1. The A and B lineages coalesce in interval τ_{n-1} and the C and D lineages coalesce in interval τ_{n-2} (all lineages coalesce as recently as possible).
2. The A and B lineages coalesce in interval τ_{n-1} and the C and D lineages do not coalesce more recently than s_{n-3}.
3. No coalescences occur in interval τ_{n-1}, and one coalescence occurs in interval τ_{n-2}.
4. No coalescences occur in interval τ_{n-1}, and two coalescences occur in interval τ_{n-2}.
5. No coalescences occur more recently than s_{n-3}.

Let $\mathcal{G}_{\text{pseudo}}$ be the matching ranked pseudocaterpillar gene tree, and let $\mathcal{G}$ be a nonmatching ranked gene tree. Let X and Y be the sets of ranked histories for $\mathcal{G}_{\text{pseudo}}$ and $\mathcal{G}$, respectively. In all five cases, the probability of a ranked history x for the matching ranked gene tree can be written

$$K_j \prod_{i=1}^{n-3} \mathbb{P}_i[\mathcal{G}_{\text{pseudo}}, x \mid \mathcal{T}_{\text{pseudo}}], \quad (4)$$

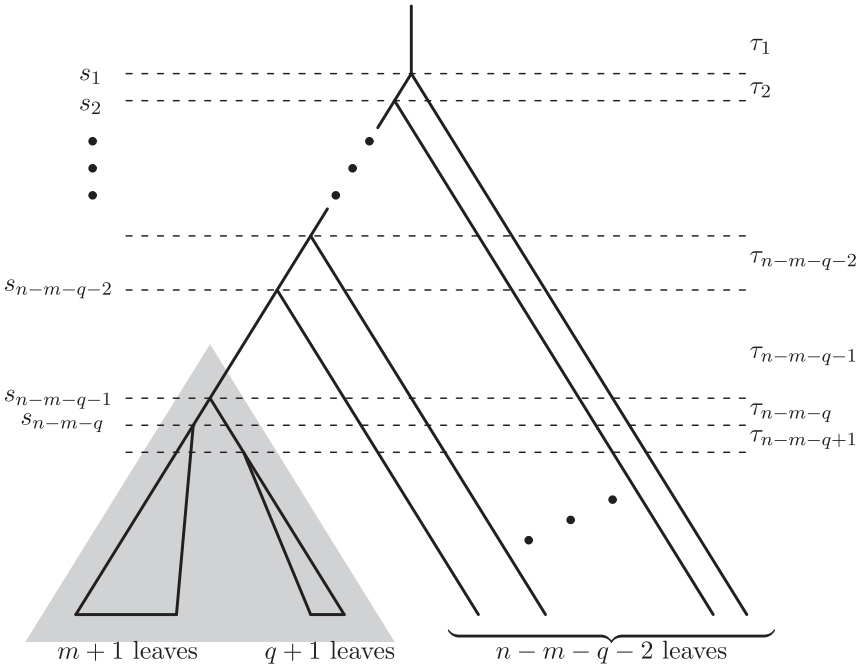

Fig. 3. The subtree $H(\Psi)$. For the tree Ψ, the special subtree $H(\Psi)$ is the subtree rooted at the node occurring at time $s_{n-m-q-1}$. Here, $H(\Psi)$ has $m+1$ left-descendants and $q+1 \leq m+1$ right-descendants. The left and right subtrees of $H(\Psi)$ are depicted as triangles. The most recent common ancestor (MRCA) of the left-descendants of $H(\Psi)$ can be either less recent than the MRCA of the right-descendants of $H(\Psi)$ (as shown), or more recent than the MRCA of the right-descendants. Because $H(\Psi)$ has $m+q+2$ leaves, $n-m-q-2$ leaves of Ψ lie outside of $H(\Psi)$. Note that nodes ancestral to $H(\Psi)$ cannot have more than one right-descendant because the root of $H(\Psi)$ is, by construction, the most ancient node on Ψ with at least two right-descendants.

where for case j, the coefficient K_j is the probability of the events in intervals τ_{n-2} and τ_{n-1}. Similarly, the probability of ranked history x for the nonmatching ranked gene tree is

$$C_j \prod_{i=1}^{n-3} \mathbb{P}_i[\mathcal{G}, x \mid \mathcal{T}_{\text{pseudo}}], \tag{5}$$

where C_j is the probability of the collection of events in intervals τ_{n-2} and τ_{n-1}.

The probabilities of the events more recent than s_{n-3} in the five cases are

$$\begin{aligned} K_1 &= C_1 = g_{2,1}(t_{n-1})g_{2,1}(t_{n-2}) \\ K_2 &= C_2 = g_{2,1}(t_{n-1})g_{2,2}(t_{n-2}) \\ K_3 &= C_3 = g_{2,2}(t_{n-1})g_{2,1}(t_{n-2})g_{2,2}(t_{n-2}) \\ K_4 &= C_4 = g_{2,2}(t_{n-1})\frac{1}{2}g_{2,1}(t_{n-2})g_{2,1}(t_{n-2}) \\ K_5 &= C_5 = g_{2,2}(t_{n-1})g_{2,2}(t_{n-2})g_{2,2}(t_{n-2}). \end{aligned} \tag{6}$$

For each j, $K_j = C_j$. In case 3, the probability does not depend on which pair coalesces in τ_{n-2} (A and B, or C and D). In case 4, the probability does not depend on which coalescence occurs first.

For events more ancient than s_{n-3}, in each time interval, only one species tree branch has more than one gene lineage. Therefore, similarly to the argument in Proposition 2, for each branch more ancient than s_{n-3}, the probability of events in the interval depends only on the number of lineages entering the branch and the number of coalescences. These quantities depend only on the ranked history and not on the gene tree topology. Thus,

$$\begin{aligned} \mathbb{P}_i[\mathcal{G}, x \mid \mathcal{T}_{\text{pseudo}}] &= \mathbb{P}_i[\mathcal{G}_{\text{pseudo}}, x \mid \mathcal{T}_{\text{pseudo}}] \\ &= g_{\alpha_i(x),\beta_i(x)}(t_i)/h_{\alpha_i(x)}^{\beta_i(x)} \end{aligned} \tag{7}$$

for each $i < n-3$, where $\alpha_i(x)$ and $\beta_i(x)$ are defined in (2).

From (6) and (7), each ranked history for $\mathcal{G}$ has the same probability as the same ranked history for $\mathcal{G}_{\text{pseudo}}$. Because Y is a proper subset of X (for example, $(1, 2, \ldots, n-1) \notin Y$), the probability of $\mathcal{G}_{\text{pseudo}}$ exceeds the probability of any other ranked gene tree. □

In the next few sections, we show that the nonexistence of ARGTs for species trees with five or more taxa applies only to caterpillar and pseudocaterpillar topologies. This proof involves first identifying a special subtree of the species tree.

3.4 A Special Subtree

We describe a special subtree that will be used to prove that most ranked species trees produce ARGTs. Consider a ranked binary tree topology Ψ. Without loss of generality, assume that for each node of Ψ, the number of "left" descendants—that is, the number of leaves descended from the left side of the node—is greater than or equal to the number of "right" descendants. By recursively querying nodes of Ψ starting from the root, we identify a specific subtree of Ψ, which we term $H(\Psi)$ (Fig. 3). Note that this subtree depends only on the topology of Ψ and not on the ranking, and it is convenient to apply H both to ranked and to unranked trees.

Starting from the root, we search for a subtree that has at least five taxa and that shares important features of the five-taxon scenario of [5] that generates ARGTs; in particular, this subtree is not a caterpillar or pseudocaterpillar, having

at least three descendants on one side of the root and two on the other. If the root ρ_0 of Ψ has at least three left-descendant leaves and two right-descendant leaves, then we define $H(\Psi)$ to be equal to Ψ. Otherwise, we proceed to the immediate left-descendant node from the root of Ψ, ρ_1. If node ρ_1 has at least three left-descendant leaves and at least two right-descendant leaves, then we define $H(\Psi)$ to be the subtree of Ψ whose root is ρ_1. Otherwise, we proceed to its immediate left-descendant, ρ_2, recursively repeating this process of querying nodes until a suitable subtree is found or until a leaf is queried without a suitable subtree having been found. If no subtree with at least three left-descendants and at least two right-descendants is found, then we set $H(\Psi)$ to be the empty set. In this manner, we define $H(\Psi)$ to be the maximal ranked subtree of Ψ with at least three left-descendant leaves and at least two right-descendant leaves. For most choices of Ψ, $H(\Psi)$ is simply Ψ itself.

Lemma 4. *Consider a tree topology Ψ with three or more leaves. The subtree $H(\Psi)$ is nonempty if and only if Ψ has five or more leaves, is not a caterpillar, and is not a pseudocaterpillar.*

Proof. Consider the sequence of nodes ρ_0, ρ_1, $\rho_2, \ldots$, as constructed above. Let ρ_* be the first node in the sequence that has five or fewer descendant leaves. $H(\Psi)$ is empty if and only if all previous nodes in the sequence (if there are any) each have exactly one right-descendant, and ρ_* itself does not have both three left-descendants and two right-descendants. This condition is obtained if and only if 1) ρ_* has only three or four descendants, or 2) the subtree rooted at ρ_* is a five-taxon caterpillar or pseudocaterpillar and all previous nodes in the sequence each have exactly one right-descendant. Condition 1 obtains if and only if Ψ has three or four leaves. Condition 2 obtains if and only if Ψ is a caterpillar or pseudocaterpillar with five or more leaves. □

3.5 Maximally Probable Coalescence Sequences

Our proof relies on an examination of the relative order of coalescences in the left and right subtrees of $H(\Psi)$ for a ranked species tree topology Ψ. Suppose there are two types of lineages, type ℓ ("left") and type r ("right"). Consider a random sequence of coalescence events W, starting from the present and moving back in time. Each event in the sequence has one of two distinct types, type L and type R. An L event occurs from the coalescence of two lineages of type ℓ; an R event occurs from the coalescence of two lineages of type r. Two ℓ lineages can coalesce or two r lineages can coalesce; an ℓ lineage cannot coalesce with an r lineage. Among the set of pairs of lineages that can coalesce—the set containing all pairs of ℓ lineages and all pairs of r lineages—each pair is equally likely to be the next to coalesce.

Suppose random sequence W has m events of type L and q events of type R. Such a sequence describes the coalescence of $m+1$ lineages of type ℓ and $q+1$ lineages of type r to two lineages, one of type ℓ and one of type r. Let W_k denote the subsequence of W that begins at position k (so that $W_1 = W$), and let $W_k^{(j)}$ denote the jth letter in sequence k. The probability that random coalescence of the lineages leads to the particular sequence w can be written recursively as follows:

$$\mathbb{P}[w_1; m, q] = \begin{cases} \dfrac{\binom{m+1}{2}}{\binom{m+1}{2} + \binom{q+1}{2}} \mathbb{P}[w_2; m-1, q] & \text{if} \quad w_1^{(1)} = \mathtt{L} \\ \dfrac{\binom{q+1}{2}}{\binom{m+1}{2} + \binom{q+1}{2}} \mathbb{P}[w_2; m, q-1] & \text{if} \quad w_1^{(1)} = \mathtt{R}. \end{cases} \tag{8}$$

For $m > 0$, we define $\mathbb{P}[w_1; m, 0] = 1$ and for $q > 0$ we define $P[w_1; 0, q] = 1$. Equation (8) can be obtained by noting that because each possible coalescence has the same probability of being the next to occur, the probability that the first event in the sequence has type L is the ratio of the number of ways that an event of type L can occur, or the number of pairs of ℓ lineages, $\binom{m+1}{2}$, to the total number of ways of choosing the next event, or the total number of pairs of lineages of the same type, $\binom{m+1}{2} + \binom{q+1}{2}$. Similar reasoning holds if the first event has type R.

Definition 5. *In the set $F_{m,q}$ of sequences consisting of m events of type L and q events of type R, a sequence is m,q-maximally probable if its probability under the assumption of equal rates of coalescence for all lineage pairs permitted to coalesce (all $\binom{m+1}{2}$ pairs of lineages of type ℓ and all $\binom{q+1}{2}$ pairs of lineages of type r) is greater than or equal to that of every other sequence consisting of m events of type L and q events of type R.*

Proposition 6. *For $m, q \geq 0$ and $m \geq q$, among sequences in $F_{m,q}$, the set of m,q-maximally probable sequences consists of those 2^q sequences that contain $m-q$ events of type L succeeded by q pairs of events, each of which includes one event of type L and one event of type R.*

Proof. By (8), the probability of a sequence $w \in F_{m,q}$ is a product of $m+q$ terms, where the kth term is the nonrecursive quotient obtained in (8) from the kth event in w. Labeling the numerator of this quotient by $N_k(w)$ and the denominator by $D_k(w)$,

$$\mathbb{P}[w; m, q] = \frac{\prod_{k=1}^{m+q} N_k(w)}{\prod_{k=1}^{m+q} D_k(w)}. \tag{9}$$

The product $\prod_{k=1}^{m+q} N_k(w)$ does not depend on w; for each w, this product contains terms $\binom{m+1}{2}, \binom{m}{2}, \ldots, \binom{2}{2}$ and $\binom{q+1}{2}, \binom{q}{2}, \ldots, \binom{2}{2}$, in some order (a different order for each w). As a result, maximizing $\mathbb{P}[w; m, q]$ amounts to finding the sequences that minimize $\prod_{k=1}^{m+q} D_k(w)$.

Denote the minimal $D_k(w)$ across sequences in $F_{m,q}$ by $d_k = \min_{w \in F_{m,q}} D_k(w)$. We have

$$\prod_{k=1}^{m+q} D_k(w) \geq \prod_{k=1}^{m+q} d_k. \tag{10}$$

Thus, if sequences exist for which $\prod_{k=1}^{m+q} D_k(w)$ achieves the minimal value of $\prod_{k=1}^{m+q} d_k$, then such sequences have the maximal value for $\mathbb{P}[w; m, q]$.

We now identify the sequences for which equality holds in (10). We can characterize a sequence $w \in F_{m,q}$ by a vector $(x_1, x_2, \ldots, x_{m+q})$, where x_k denotes the number of events of type L that have occurred prior to the kth event in w. Alternatively, we can characterize w by $(y_1, y_2, \ldots, y_{m+q})$, where y_k denotes the number of events of type R that have occurred prior to the kth event in w. The x_k and y_k satisfy several constraints. First, as

$k-1$ events total have occurred prior to the kth event, $x_k + y_k = k-1$. Also, $0 \le x_k \le \min(k-1, m)$ and $0 \le y_k \le \min(k-1, q)$, because the total number of events of type L in w is m, the total number of events of type R in w is q, and at most $k-1$ events of a single type (L or R) can occur before the kth event.

For each sequence $w \in F_{m,q}$, using (8), $D_k(w)$ can be written as

$$D_k(w) = \binom{m+1-x_k}{2} + \binom{q+1-y_k}{2}. \tag{11}$$

Substituting $k-1-x_k$ for y_k, we can write (11) as

$$D_k(w) = x_k^2 + (q+1-k-m)x_k + \frac{1}{2}[(m+1)m + (q+2-k)(q+1-k)]. \tag{12}$$

Treated as a quadratic function of x_k, the minimum of $D_k(w)$ occurs at $x_k = [(m-q)+(k-1)]/2$. However, because of the constraints on x_k ($0 \le x_k \le \min(k-1, m)$), this minimum need not fall among allowed values of x_k.

If $k-1 < [(m-q)+(k-1)]/2$, or equivalently, if $k-1 < m-q$, then the minimum of $D_k(w)$ occurs at a value that exceeds the largest possible value of x_k, and the minimum among allowed values occurs at $x_k = k-1$. Thus, for each k from 1 to $m-q$, the minimum of $D_k(w)$ within the allowed range occurs when $x_k = k-1$.

If $k-1 \ge [(m-q)+(k-1)]/2$, or equivalently, if $k-1 \ge m-q$, then the minimum of $D_k(w)$ occurs at a value less than or equal to the largest possible value for x_k. If $[(m-q)+(k-1)]/2$ is an integer, then the minimum occurs at $x_k = [(m-q)+(k-1)]/2$. However, if $[(m-q)+(k-1)]/2$ is instead halfway between two integers, then the minimum does not occur at an allowed value of x_k. By symmetry of a quadratic function around its vertex, two minima occur among allowed values of x_k, at $x_k = \lfloor [(m-q)+(k-1)]/2 \rfloor$ and at $x_k = \lceil [(m-q)+(k-1)]/2 \rceil$. Because $[(m-q)+(k-1)]/2$ alternates between integer and noninteger values, as k increases, starting from $m-q+1$, $D_k(w)$ alternates between having one value of x_k that achieves its minimum and having two such values. There are q of these alternating pairs as k ranges from $m-q+1$ to $m+q$.

Thus, by separately identifying the location of the minimum or minima for each of the terms $D_k(w)$, we have found a set of vectors $(x_1, x_2, \ldots, x_{m+q})$, each of which corresponds to a sequence with probability $\prod_{k=1}^{m+q} d_k$. Each such vector begins with $x_k = k-1$ for $k = 1$ to $k = m-q$. The next component, x_{m-q+1}, must equal $m-q$, but x_{m-q+2} can then be either $\lfloor m-q+1/2 \rfloor$ or $\lceil m-q+1/2 \rceil$. Next, x_{m-q+3} must equal $m-q+1$, but x_{m-q+4} can be either $\lfloor m-q+3/2 \rfloor$ or $\lceil m-q+3/2 \rceil$. A total of q of these alternating pairs occur.

Linking a vector $(x_1, x_2, \ldots, x_{m+q})$ to its associated sequence in $F_{m,q}$, the values $x_k = k-1$ for $k = 2$ to $k = m-q+1$ indicate that the first $m-q$ events in a sequence that achieves the minimal product $\prod_{k=1}^{m+q} d_k$—and that therefore has maximal probability—all have type L. The subsequent q alternating pairs of terms, together with the fact that the total number of events of type L is m, indicate that for m,q-maximally probable sequences, starting with event $m-q+1$, either an L or R event is possible in the first event in a pair. However, the second event must then have type opposite to that of the first. □

Note that at each point in each m,q-maximally probable sequence, the type for the next event (either L or R) always corresponds to the type for which more events remain in the sequence; in case of a tie in the numbers of remaining events of type L and of type R, either type of event can occur next. Thus, m,q-maximally probable sequences are "greedy" in the sense that the next type of event to occur is always the type that is more probable.

Corollary 7. *The probability of an m,q-maximally probable sequence is*

$$\frac{q+1}{4^q} \prod_{k=q+1}^{m} \frac{(k+1)k}{(k+1)k + (q+1)q}. \tag{13}$$

Proof. The formula is obtained by explicit computation with (8), using the characterization of m,q-maximally probable sequences in Proposition 6. □

We note that maximally probable sequences and (8) and (13) can be employed to confirm the limiting five-taxon results of [5, Section 5.3]. Suppose we label species divergences in increasing order from the root, which is always first. Consider five-taxon species trees with three left-descendants of the root (A, B, and C) and two right-descendants (D and E), and for which the labeled topology is ((AB)C)(DE)). Indicate the ranked species tree by $\mathcal{T}_{\mathrm{RLL}}$ if the two right-descendants have the most recent species divergence (Fig. 2), by $\mathcal{T}_{\mathrm{LRL}}$ if the right-descendants have the second most recent divergence, and by $\mathcal{T}_{\mathrm{LLR}}$ if their divergence is the most ancient divergence in the species tree other than the root. Denote corresponding gene trees by $\mathcal{G}_{\mathrm{RLL}}$, $\mathcal{G}_{\mathrm{LRL}}$, and $\mathcal{G}_{\mathrm{LLR}}$. Such trees have ranked topology $(((\mathrm{AB})_2\mathrm{C})_3(\mathrm{DE})_4)$, $(((\mathrm{AB})_2\mathrm{C})_4(\mathrm{DE})_3)$, and $(((\mathrm{AB})_3\mathrm{C})_4(\mathrm{DE})_2)$, respectively.

Consider the species tree $\mathcal{T}_{\mathrm{RLL}}$. Let $\tau_2 \to \infty$ and let $\tau_3, \tau_4 \to 0$. The species tree approaches a scenario in which two sequences of coalescences proceed concurrently in interval τ_2, one with $m+1 = 3$ lineages and the other with $q+1 = 2$ lineages. One maximally probable sequence generates ranked gene tree $\mathcal{G}_{\mathrm{LRL}}$, and another generates $\mathcal{G}_{\mathrm{LLR}}$. Each of these ranked gene trees has probability $1/8$, whereas the matching ranked gene tree has probability $1/12$.

Corollary 8. *If $m \ge q$, $m \ge 2$, and $q \ge 1$, then at least one sequence in $F_{m,q}$ is not m,q-maximally probable.*

Proof. The number of m,q-maximally probable sequences in $F_{m,q}$ is 2^q. The total number of sequences in $F_{m,q}$ is the number of ways of ordering m events of type L and q events of type R, or $\binom{m+q}{m}$. Therefore, we must show $\binom{m+q}{m} > 2^q$.

If $q = 1$, since $m \ge 2$, the inequality holds. For $q \ge 2$, because $m \ge q$, $\binom{m+q}{m} \ge \binom{2q}{q} = \prod_{k=1}^{q}(k+q)/k$. Each term in the product exceeds 2 except the term for $k = q$, which equals 2. Because $q \ge 2$, at least one term exceeds 2. Thus, $\binom{m+q}{m}$ exceeds a product of q numbers, each of which is greater than or equal to 2, and at least one of which strictly exceeds 2. □

3.6 Proof of the Main Result (Theorem 1)

Consider a ranked species tree topology Ψ. We assume that Ψ has five or more leaves and is not a caterpillar or pseudocaterpillar, so that $H(\Psi)$ is nontrivial. Denote by $m+1$ and $q+1$ the numbers of left- and right-descendants of the root of $H(\Psi)$, respectively, letting $m \geq q$ without loss of generality. Construct the sequence of divergence times of $H(\Psi)$, beginning from the most recent. Label a divergence by L if it occurs on the left side of the root of $H(\Psi)$ and by R if it occurs on the right side. The sequence of labels constructed in this manner, denoted $W(\Psi)$, is in $F_{m,q}$.

Proposition 9. *Consider a ranked species tree topology Ψ with nontrivial $H(\Psi)$ and with $m+1$ left-descendants and $q+1$ right-descendants of $H(\Psi)$. If $W(\Psi)$ is not m,q-maximally probable in $F_{m,q}$, then Ψ produces anomalous ranked gene trees.*

To prove the proposition, we first establish that by choosing intervals to be either short or long, we can make the probability arbitrarily close to 1 that either no coalescences or all possible coalescences occur in the interval.

Lemma 10. *Let $\mathcal{T}$ be a ranked species tree with n leaves. For interval τ_j with length t_j, $j = 2, \ldots, n-1$, label the j species tree branches by b_{jk}, $k = 1, \ldots, j$. For branch b_{jk}, let a_{jk} be the number of lineages available to coalesce and let c_{jk} be the number of coalescence events. Then,*

1. *For any $\varepsilon > 0$, there exists $t_j > 0$ such that $\mathbb{P}[c_{j1} = 0, \ldots, c_{jj} = 0 \mid \mathcal{T}] > 1 - \varepsilon$.*
2. *For any $\varepsilon > 0$, there exists $t_j > 0$ such that $\mathbb{P}[c_{j1} = a_{j1} - 1, \ldots, c_{jj} = a_{jj} - 1 \mid \mathcal{T}] > 1 - \varepsilon$.*

Part 1 states that for a given interval on a ranked species tree, we can choose the interval length short enough that, with probability close to 1, no coalescences occur in that interval. Part 2 states that we can choose the interval length long enough that, with probability close to 1, all lineages that can coalesce in the interval do coalesce in the interval.

Proof. For part 1, the probability that no coalescences occur on the j branches in interval τ_j is

$$\prod_{k=1}^{j} g_{a_{jk},a_{jk}}(t_j) = \prod_{k=1}^{j} e^{-\binom{a_{jk}}{2} t_j}. \tag{14}$$

The values of a_{jk} in (14) depend on the coalescence events that have occurred in intervals τ_i, $i > j$. Because (14) is decreasing in each of the a_{jk}, the equation is minimized when no coalescences occur in any of the intervals τ_i, $i > j$, so that a_{jk} is maximal. Regardless of where coalescences occur, $a_{jk} \leq n$ for each j and k. Hence, choosing t_j such that $t_j < [-\log(1-\varepsilon)]/[\binom{n}{2} j]$,

$$\prod_{k=1}^{j} e^{-\binom{a_{jk}}{2} t_j} \geq \prod_{k=1}^{j} e^{-\binom{n}{2} t_j} > 1 - \varepsilon.$$

For part 2, note that for any branch k in interval j with $a_{jk} = 1$, $\mathbb{P}[c_{jk} = 0] = 1$ because $g_{1,1}(t) = 1$ for any $t > 0$. Using the fact that for all $i \geq 1$, $\lim_{t\to\infty} g_{i,1}(t) = 1$ [18], we can choose t_j large enough such that for each $k = 1, \ldots, j$, $g_{a_{jk},1}(t_j) > 1 - \varepsilon/j$. The probability that $c_{jk} = a_{jk} - 1$ for each k is the product $\prod_{k=1}^{j} g_{a_{jk},1}(t_j)$, which exceeds $1 - \varepsilon$. □

Proposition 9 can now be proven by considering the special subtree $H(\Psi)$. Suppose the root of $H(\Psi)$ occurs at s_i. By making all intervals τ_j short except for τ_{i+1} immediately below the root of $H(\psi)$ (and τ_1 above the root of the full tree), which is made long, the only intervals that are likely to have a coalescence event are τ_{i+1} and τ_1. Because τ_{i+1} has exactly two branches in which coalescence events can occur, if the topology of $H(\Psi)$ has a nonmaximally probable sequence of species divergences, then the most likely ranked gene tree does not match the ranked species tree topology.

Proof of Proposition 9. We let $\mathcal{G}_1 = \Psi$ denote the matching ranked gene tree topology. Without loss of generality, the topology of Ψ can be written as

$$(((\Psi_1, \Psi_2), A_1), \ldots, A_{n-m-q-2}),$$

where $H(\Psi) = (\Psi_1, \Psi_2)$ and each of $A_1, \ldots, A_{n-m-q-2}$ represents a single taxon. If $n - m - q - 2 = 0$, then $H(\Psi) = \Psi$. Here, Ψ_1 has $m+1 \geq 3$ leaves and Ψ_2 has $q+1 \geq 2$ leaves, with $m \geq q$ and $m + q + 2 \leq n$. Labeling the internal nodes of Ψ_1 and Ψ_2 by L and R, respectively, we assume that the LR sequence for the nodes of Ψ, w_1, is not maximally probable. Consider a nonmatching ranked gene tree $\mathcal{G}_2$ with the same topology as Ψ, except that $H(\Psi)$ is replaced by $H(\mathcal{G}_2) = (\Psi_1^*, \Psi_2^*)$, where Ψ_k^* is defined as follows on the same leaves as Ψ_k, $k = 1, 2$. Suppose that labeling the internal nodes of Ψ_1^* and Ψ_2^* by L and R, respectively, results in a maximally probable sequence, w_2, of coalescences. We let $x = (1, 2, \ldots, i, i+1, \ldots, i+1)$, where $i = n - m - q - 1$ by construction of $H(\Psi)$, denote the ranked history in which all left- and right-descendants of $H(\Psi)$ (and $H(\mathcal{G}_2)$) coalesce in interval τ_{i+1}, and in which there is exactly one coalescence event in each interval τ_j, $j \leq i$.

Because the sequence w_1 of LR events in $H(\mathcal{G}_1)$ is not maximally probable, and the sequence w_2 in $H(\mathcal{G}_2)$ is maximally probable, $\mathbb{P}[w_2] = \mathbb{P}[w_1] + \delta$ for some $\delta > 0$. We also have that $h := h_{m+1}^1 h_{q+1}^1$ is the number of ways of choosing lineages to coalesce according to ranked history x in interval τ_{i+1}. Thus, given ranked history x and given some sequence of coalescences (either w_1 or w_2), the probability is $1/h$ that the sequence is compatible with a particular ranked gene tree topology. Choose some δ^* such that $0 < \delta^* < \delta$. We select ε such that

$$\varepsilon = \min\left\{1 - \frac{\mathbb{P}[w] + \delta^*}{\mathbb{P}[w] + \delta}, \frac{\delta^*}{h}\right\},$$

and we choose times between divergences on the species tree, $t_2, t_3, \ldots, t_{n-1}$, such that the probability of history x exceeds $1 - \varepsilon$. Thus, as $\mathbb{P}[\mathcal{G}_2|x] = \mathbb{P}[w_2]/h$,

$$\begin{aligned}\mathbb{P}[\mathcal{G}_2] &> \mathbb{P}[\mathcal{G}_2, x] > \frac{1}{h}(1-\varepsilon)\mathbb{P}[w_2] \\ &= \frac{1}{h}(1-\varepsilon)(\mathbb{P}[w_1] + \delta) \\ &\geq \frac{1}{h}(\mathbb{P}[w_1] + \delta^*).\end{aligned}$$

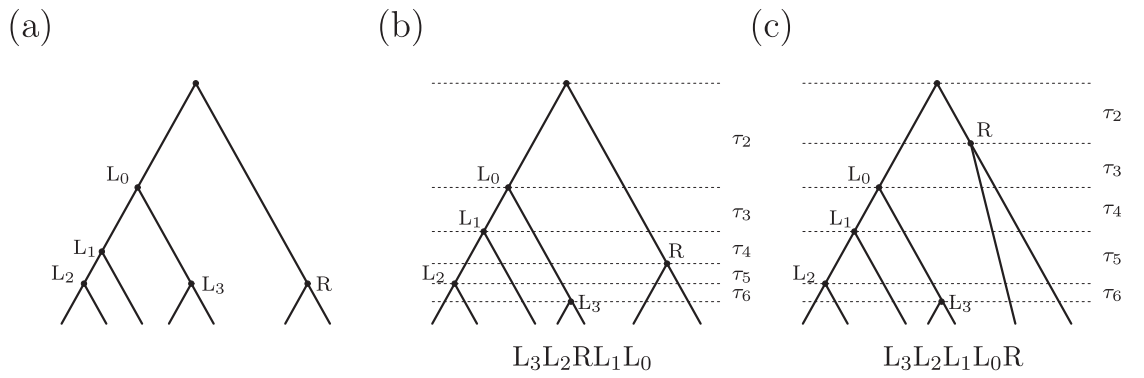

Fig. 4. A rooted species tree topology with two of its rankings that give rise to ARGTs. Each internal node excluding the root is labeled L or R depending on whether it is a left- or right-descendant of the root. The subscripts on the left-descendants provide identifiers, but are otherwise unimportant. (a) The unranked rooted species tree. (b) A ranking of the species tree topology in which ARGTs are detected by the method in the proof of Proposition 9, making τ_2 large and τ_j small for $j > 2$. (c) A ranking of the species tree topology that can be seen to produce ARGTs by a modification of the method in the proof of Proposition 9.

Finally, because $\mathbb{P}[\mathcal{G}_1|x] = \mathbb{P}[w_1]/h$,

$$\begin{aligned}\mathbb{P}[\mathcal{G}_2] &> \mathbb{P}[\mathcal{G}_1 \mid x] + \delta^*/h > \mathbb{P}[\mathcal{G}_1, x] + \delta^*/h \\ &\geq \mathbb{P}[\mathcal{G}_1, x] + \varepsilon > \mathbb{P}[\mathcal{G}_1].\end{aligned}$$

Therefore, $\mathcal{G}_2$ is an ARGT. □

Now, we can finally prove our main result.

Proof of Theorem 1. 1) For each noncaterpillar, nonpseudocaterpillar species tree topology with five or more taxa, by Lemma 4, $H(\Psi)$ is nontrivial. Suppose $H(\Psi)$ has $m+1$ left-descendants and $q+1$ right-descendants. By Corollary 8, there exists a ranking for the sequence of coalescences in $H(\Psi)$ that is not m, q-maximally probable. Choose this ranking for the species tree. By Proposition 9, Ψ produces ARGTs. 2) This result was obtained in the nonexistence results in [5, Sections 5.1 and 5.2] and Propositions 2 and 3. □

4 Properties of ARGTs

Now that we have demonstrated that most ranked species trees produce ARGTs, we examine some further properties of those species trees and the ARGTs that they produce. First, we consider species tree rankings that give rise to ARGTs. Next, we investigate the discordance in topology between species trees and their ARGTs. Finally, we show that unlike in the unranked case, ARGTs can have caterpillar topologies.

4.1 Rankings that Produce ARGTs

The proportion of unranked species tree topologies for which no ranking and choice of branch lengths produces ARGTs is the ratio of the number of caterpillar and pseudocaterpillar trees, $(3/4)n!$, to the number of labeled topologies, or $(2n-3)!!$. The ratio $(3/4)n!/(2n-3)!!$ approaches 0 as $n \to \infty$. Therefore, most unranked species tree topologies can be ranked in a way that produces ARGTs.

In fact, we have proven much more than that each noncaterpillar, nonpseudocaterpillar species tree topology with five or more taxa has a ranking that gives rise to ARGTs. We have proven that *many* rankings give rise to ARGTs, as *all* rankings for $H(\Psi)$ that are not m, q-maximally probable are covered by Proposition 9. Although Theorem 1 completely characterizes which *unranked* species trees have some ranking that gives rise to ARGTs, it stops short of a complete characterization of which *ranked* species trees give rise to ARGTs. Indeed, some species tree rankings not covered by Proposition 9 produce ARGTs. Consider the ranked species tree in Fig. 4c. Because the sequence $\mathrm{L_3L_2L_1L_0R}$ is maximally probable (when the subscripts are ignored), Proposition 9 does not show that this ranked species tree produces ARGTs. However, the subtree rooted at node $\mathrm{L_0}$ has the same ranked topology as $\mathcal{T}_{\mathrm{RLL}}$; hence, by making τ_2, τ_3, and τ_4 long and τ_5 and τ_6 short, the species tree in Fig. 4c produces ARGTs. Thus, while our proof of Theorem 1 utilized the "special subtree" $H(\Psi)$ in the construction of ARGTs, for this particular ranked species tree topology, a different "special subtree" can be used in place of $H(\Psi)$.

For $4 \leq n \leq 8$, Table 1 counts the numbers of ranked species trees on n taxa for which the method of proof of Proposition 9 demonstrates the existence of ARGTs. To illustrate the procedure for counting the number of ranked species trees that produce ARGTs, as determined using Proposition 9, consider the unranked species tree in Fig. 4a. There are $\binom{7}{2}\binom{5}{2}\binom{3}{2} = 630$ ways to label the leaves of this species tree topology, and there are 15 rankings for each labeled topology. For a given leaf-labeling, let the topology in Fig. 4a be Ψ. Then, $H(\Psi) = \Psi$, because the root of Ψ has two right-descendants ($\geq$2) and five left-descendants (≥ 3). Given a ranking of Ψ, such as the one in Fig. 4b, the strategy in Proposition 9 is to let the interval immediately below the root of $H(\Psi)$, τ_2, be long, and to let the intervals $\tau_3, \tau_4, \tau_5, \tau_6$ be short. For each leaf labeling, nine rankings of nodes $\mathrm{L_0}$, $\mathrm{L_1}$, $\mathrm{L_2}$, $\mathrm{L_3}$, and R—one of which appears in Fig. 4b—result in nonmaximally probable LR sequences according to Proposition 6. Therefore, Proposition 9 indicates that $630 \times 9 = 5{,}670$ ranked species tree topologies with the unranked topology in Fig. 4a produce ARGTs.

Among the other six rankings of $\mathrm{L_0}$, $\mathrm{L_1}$, $\mathrm{L_2}$, $\mathrm{L_3}$, and R for a given leaf labeling, for two of them, the argument in the proof of Proposition 9 applies with the "special subtree" rooted at $\mathrm{L_0}$ rather than at the species tree root. For example, the ranking in Fig. 4c is not covered by Proposition 9, but by making τ_3 long, and making all other intervals short, ARGTs are produced in the subtree rooted at $\mathrm{L_0}$. Thus, because the ranking in Fig. 4c is one of two

TABLE 1
The Number of Ranked Species Trees that Can be Found to Have ARGTs by Using the Strategy of Proposition 9, and the Number of Ranked Species Trees that Cannot Have ARGTs because They Have Caterpillar or Pseudocaterpillar Topologies

Number of taxa n	Number of labeled topologies	Number of ranked topologies (h_n^1)	Number of ranked species tree topologies found by Props. 2 and 3 to not have ARGTs	Number of ranked species tree topologies found by Prop. 9 to have ARGTs
4	15	18	18	0
5	105	180	90	30
6	945	2700	540	900
7	10395	56700	3780	28980
8	135135	1587600	30240	1003070

The number of ranked species tree topologies for which existence of ARGTs is not determined by our proof of Theorem 1 is the difference between the third column and the sum of the fourth and fifth columns.

rankings with similar behavior, at least $630 \times 2 = 1{,}260$ rankings for the unranked species tree topology in Fig. 4a produce ARGTs undetected by Proposition 9.

4.2 ARGTs with Nonmatching Unranked Topologies

In the five-taxon case in [5], we identified ARGTs with unranked labeled topologies that differ from the unranked labeled topology of the ranked species tree. Such ARGTs exist in general. Under the hypotheses of Theorem 1, our proof is a general demonstration that ARGTs exist that differ in unranked topology from the species tree topology.

Corollary 11. *If an n-taxon species tree topology Ψ has a special subtree $H(\Psi)$ with $m+1$ left-descendants and $q+1$ right-descendants, with $m \geq 2$, $q \geq 1$, and $n \geq m+q+2$, where the sequence of LR divergences on $H(\Psi)$ is not maximally probable, then there exists an ARGT $\mathcal{G}_2$ with a different unranked topology from the species tree.*

Proof. In the proof of Proposition 9, choose $\mathcal{G}_2$ to have a different unranked topology from the species tree as follows: Consider the subtree $H(\Psi) = (\psi_1, \psi_2)$, where ψ_1 is the left subtree of $H(\Psi)$ and has $m+1 \geq 3$ leaves, and ψ_2 is the right subtree of $H(\Psi)$ with $q+1 \geq 2$ leaves. Then, ψ, the unranked topology induced by Ψ, can be written

$$(((\psi_1, \psi_2), A_1), \ldots, A_{n-m-q-2}),$$

where each of $A_1, \ldots, A_{n-m-q-2}$ represents a single taxon.

Suppose ψ_1 is not a caterpillar. Then, let G_2 be an unranked gene tree with topology

$$(((\psi_1^*, \psi_2), A_1), \ldots, A_{n-m-q-2}),$$

where ψ_1^* is a caterpillar topology defined on the same leaves as ψ_1. Choose a maximally probable ranking on the nodes of (ψ_1^*, ψ_2), call it (Ψ_1^*, Ψ_2), and let $\mathcal{G}_2$ be

$$(((\Psi_1^*, \Psi_2), A_1), \ldots, A_{n-m-q-2}).$$

Then, $\mathcal{G}_2$ is a maximally probable ranking of G_2 with nonempty $H(\mathcal{G}_2) = (\Psi_1^*, \Psi_2)$. The proof of Theorem 1 then applies, and Ψ produces $\mathcal{G}_2$ as an ARGT.

Similarly, if the left subtree of $H(\Psi)$ is a caterpillar, say $\psi_1 = (((B_1, B_2), B_3), \ldots, B_{m+1})$, where $B_1, \ldots, B_{m+1}$ are leaves, then the same argument can be repeated using $\psi_1^* = (((B_1, B_3), B_2), \ldots, B_{m+1})$. □

4.3 Caterpillar ARGTs

The behavior of caterpillars provides one of the most noticeable differences in the properties of ranked and unranked gene trees. In the unranked setting, caterpillar species trees are the "most anomalous" case [3], [15], and any noncaterpillar gene tree topology can be an anomalous gene tree for any caterpillar species tree topology. In the ranked setting, caterpillars are the "least anomalous" case, and along with pseudocaterpillars, they are the only species trees that do not produce ARGTs at all. Further, for five-taxon trees, AGTs can never have a caterpillar topology; however, as the following example demonstrates, five-taxon ARGTs can be caterpillars.

Consider species tree $\mathcal{T}_{\mathrm{RLL}}$—with ranked labeled topology $(((AB)_3C)_2(DE)_4)$ as introduced above—and ranked gene tree $\mathcal{G}_{\mathrm{cat}} = ((((\mathrm{AB})\mathrm{C})\mathrm{D})\mathrm{E})$. Using [5, Table 4], which provides the probabilities of ranked histories for the ranked species tree $\mathcal{T}_{\mathrm{RLL}}$, caterpillar gene tree ((((AB)C)D)E) has probability

$$\begin{aligned}\mathbb{P}[\mathcal{G}_{\mathrm{cat}} \mid \mathcal{T}_{\mathrm{RLL}}] = {} & \mathbb{P}[\mathcal{G}_{\mathrm{LLR}}, (1,1,1,1) \mid \mathcal{T}_{\mathrm{RLL}}] \\ & + \mathbb{P}[\mathcal{G}_{\mathrm{LLR}}, (1,1,1,2) \mid \mathcal{T}_{\mathrm{RLL}}] \\ & + \mathbb{P}[\mathcal{G}_{\mathrm{LLR}}, (1,1,1,3) \mid \mathcal{T}_{\mathrm{RLL}}] \\ & + \mathbb{P}[\mathcal{G}_{\mathrm{LLR}}, (1,1,2,2) \mid \mathcal{T}_{\mathrm{RLL}}] \\ & + \mathbb{P}[\mathcal{G}_{\mathrm{LLR}}, (1,1,2,3) \mid \mathcal{T}_{\mathrm{RLL}}].\end{aligned} \tag{15}$$

Because the caterpillar gene tree $\mathcal{G}_{\mathrm{cat}}$ cannot occur if any coalescence events occur in interval τ_4, the probability $\mathbb{P}[\mathcal{G}_{\mathrm{cat}} \mid \mathcal{T}_{\mathrm{RLL}}]$ is decreasing in t_4. The gene tree $\mathcal{G}_{\mathrm{cat}}$ also only has two ranked histories that include a coalescence in interval τ_3; consequently, τ_3 must be small for $\mathcal{G}_{\mathrm{cat}}$ to be more probable than $\mathcal{G}_{\mathrm{RLL}}$. Finally, although t_2 must be sufficiently large for ARGTs to exist, for the caterpillar $\mathcal{G}_{\mathrm{cat}}$, large values of t_2 make the ranked history $(1,2,2,2)$ likely. This history is possible for the matching ranked gene tree but is not a ranked history for the caterpillar. Consequently, the region of the parameter space for which $\mathcal{G}_{\mathrm{cat}}$ is an ARGT has very small values of t_3 and t_4, and small but less extreme values of t_2 (Fig. 5). When $t_3, t_4 = 0$, the maximum of $\mathbb{P}[\mathcal{G}_{\mathrm{cat}} \mid \mathcal{T}_{\mathrm{RLL}}] - \mathbb{P}[\mathcal{G}_{\mathrm{RLL}} \mid \mathcal{T}_{\mathrm{RLL}}]$ is near $t_2 = 0.125$.

5 Discussion

In this paper, we have shown that surprisingly, anomalous ranked gene trees exist for all species tree topologies with five or more taxa, with the exceptions of caterpillar and

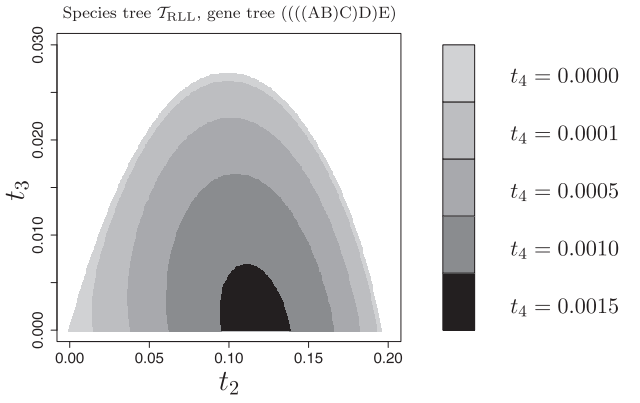

Fig. 5. Anomaly zone for the ranked species tree $\mathcal{T}_{\mathrm{RLL}}$ when the ranked gene tree is ((((AB)C)D)E). For each $t_4 \in \{0, 0.0001, 0.0005, 0.0010, 0.0015\}$, the region was computed by determining the values of $t_2 \in [0, 0.2]$, evaluated in increments of $0.2/350$, and $t_3 \in [0, 0.03]$, evaluated in increments of $0.03/350$, for which $\mathbb{P}[((((\mathrm{AB})\mathrm{C})\mathrm{D})\mathrm{E}) \mid \mathcal{T}_{\mathrm{RLL}}] > \mathbb{P}[\mathcal{G}_{\mathrm{RLL}} \mid \mathcal{T}_{\mathrm{RLL}}]$.

pseudocaterpillar topologies. Further, many different species tree rankings give rise to ARGTs, and ARGTs can disagree with species trees in unranked topology. Our characterization of the set of species trees that give rise to ARGTs parallels the corresponding result in the unranked case that all species trees with five or more taxa produce anomalous unranked gene trees, with the significant exception that caterpillars and pseudocaterpillars produce quite different behavior in the ranked and unranked cases.

While our characterization identifies some of the key properties of ARGTs, many aspects of ARGTs remain unexplored. For example, for n taxa, how many ranked species trees can have ARGTs? For a given species tree, how many ranked gene tree topologies can be ARGTs? Is there a ranked analogue to wicked forests [3]? Such an analogue would be a set V of ranked species trees $\sigma_1, \sigma_2, \ldots, \sigma_K$, $K \geq 2$, with speciation times, such that for all i, j with $i \neq j$, if $\sigma_i = (\Psi_i, \mathbf{s}_i) \in V$ and $\sigma_j = (\Psi_j, \mathbf{s}_j) \in V$, then Ψ_i is an ARGT for σ_j and Ψ_j is an ARGT for σ_i. By Propositions 2 and 3, if wicked ranked forests exist, then they contain no caterpillars or pseudocaterpillars.

In previous work, we have shown that the democratic vote method of using the most frequent unranked gene tree topology to estimate the unranked species tree topology is statistically inconsistent [3]; the existence of AGTs means that for some parameter values, democratic vote is increasingly likely to estimate an incorrect tree given more gene trees. Here, we have obtained a corresponding result for ranked trees: as the number of gene trees increases, the most frequent ranked gene tree can be increasingly likely to not match the ranked species tree. However, we have not examined if the most probable ranked gene tree always has the same unranked topology as the species tree. If the most probable ranked gene tree matches the species tree in unranked topology, then in inferring species trees, methods that estimate ranked trees might be more robust to gene tree discordance than methods that infer unranked trees. Further, as the space of parameter values that gives rise to ARGTs differs from the space that gives rise to AGTs, combining methods that incorporate rank and methods that ignore it might provide a basis for circumventing anomalous regions that confound one approach but not the other. Thus, as in the case of AGTs, whose investigation has facilitated evaluations of the consistency properties of species tree inference methods that use unranked gene trees [1], [2], [3], [7], [8], [9], [10], [11], [15], [19], [20], the theory of ARGTs could eventually lead to improved inference of species trees on the basis of ranked gene trees.

Acknowledgments

This work was supported by grants from the New Zealand Marsden Fund and the Burroughs Wellcome Fund, and by the US National Science Foundation (NSF) grant DBI-1146722. The work was partly conducted while James H. Degnan was a sabbatical fellow at the National Institute for Mathematical and Biological Synthesis, an Institute sponsored by the National Science Foundation, the US Department of Homeland Security, and the US Department of Agriculture through NSF Award EF-0832858, with additional support from the University of Tennessee, Knoxville. Tanja Stadler was supported by the Swiss National Science Foundation.

References

[1] M. DeGiorgio and J.H. Degnan, "Fast and Consistent Estimation of Species Trees Using Supermatrix Rooted Triples," *Molecular Biology and Evolution,* vol. 27, pp. 552-569, 2010.

[2] J.H. Degnan, M. DeGiorgio, D. Bryant, and N.A. Rosenberg, "Properties of Consensus Methods for Inferring Species Trees from Gene Trees," *Systematic Biology,* vol. 58, pp. 35-54, 2009.

[3] J.H. Degnan and N.A. Rosenberg, "Discordance of Species Trees with Their Most Likely Gene Trees," *PLoS Genetics,* vol. 2, pp. 762-768, 2006.

[4] J.H. Degnan and N.A. Rosenberg, "Gene Tree Discordance, Phylogenetic Inference, and the Multispecies Coalescent," *Trends in Ecology and Evolution,* vol. 24, pp. 332-340, 2009.

[5] J.H. Degnan, N.A. Rosenberg, and T. Stadler, "The Probability Distribution of Ranked Gene Trees on a Species Tree," *Math. Biosciences,* vol. 235, pp. 45-55, 2012.

[6] J.H. Degnan and L.A. Salter, "Gene Tree Distributions under the Coalescent Process," *Evolution,* vol. 59, pp. 24-37, 2005.

[7] G.B. Ewing, I. Ebersberger, H.A. Schmidt, and A. von Haeseler, "Rooted Triple Consensus and Anomalous Gene Trees," *BMC Evolutionary Biology,* vol. 8, article 118, 2008.

[8] E.M. Jewett and N.A. Rosenberg, "iGLASS: An Improvement to the GLASS Method for Estimating Species Trees from Gene Trees," *J. Computational Biology,* vol. 19, pp. 293-315, 2012.

[9] L. Liu and S.V. Edwards, "Phylogenetic Analysis in the Anomaly Zone," *Systematic Biology,* vol. 58, pp. 452-460, 2009.

[10] L. Liu, L. Yu, D.K. Pearl, and S.V. Edwards, "Estimating Species Phylogenies Using Coalescence Times Among Sequences," *Systematic Biology,* vol. 58, pp. 468-477, 2009.

[11] E. Mossel and S. Roch, "Incomplete Lineage Sorting: Consistent Phylogeny Estimation from Multiple Loci," *IEEE/ACM Trans. Computational Biology and Bioinformatics,* vol. 7, no. 1, pp. 166-171, Jan.-Mar. 2010.

[12] P. Pamilo and M. Nei, "Relationships between Gene Trees and Species Trees," *Molecular Biology and Evolution,* vol. 5, pp. 568-583, 1988.

[13] N.A. Rosenberg, "The Probability of Topological Concordance of Gene Trees and Species Trees," *Theoretical Population Biology,* vol. 61, pp. 225-247, 2002.

[14] N.A. Rosenberg, "Counting Coalescent Histories," *J. Computational Biology,* vol. 14, pp. 360-377, 2007.

[15] N.A. Rosenberg and R. Tao, "Discordance of Species Trees with Their Most Likely Gene Trees: The Case of Five Taxa," *Systematic Biology,* vol. 57, pp. 131-140, 2008.

[16] T. Stadler and J.H. Degnan, "A Polynomial Time Algorithm for Calculating the Probability of a Ranked Gene Tree Given a Species Tree," *Algorithms for Molecular Biology,* vol. 7, article 7, 2012.

[17] N. Takahata, "Gene Genealogy in Three Related Populations: Consistency Probability between Gene and Population Trees," *Genetics,* vol. 122, pp. 957-966, 1989.
[18] S. Tavaré, "Line-of-Descent and Genealogical Processes, and Their Applications in Population Genetics Models," *Theoretical Population Biology,* vol. 26, pp. 119-164, 1984.
[19] C.V. Than and N.A. Rosenberg, "Consistency Properties of Species Tree Inference by Minimizing Deep Coalescences," *J. Computational Biology,* vol. 18, pp. 1-15, 2010.
[20] Y. Wang and J.H. Degnan, "Performance of Matrix Representation with Parsimony for Inferring Species From Gene Trees," *Statistical Applications in Genetics and Molecular Biology,* vol. 10, article 21, 2011.
[21] Y. Wu, "Coalescent-Based Species Tree Inference from Gene Tree Topologies under Incomplete Lineage Sorting by Maximum Likelihood," *Evolution,* vol. 66, pp. 763-775, 2012.

James H. Degnan received the PhD degree in statistics in 2005 from the University of New Mexico, where he was advised by Laura Kubatko. He did postdoctoral research at Harvard and the University of Michigan (supervised by Noah Rosenberg). In 2008, he started a faculty position in the Department of Mathematics and Statistics at the University of Canterbury in Christchurch, New Zealand, where is now a senior lecturer.

Noah A. Rosenberg received the PhD degree in biological sciences in 2001 from Stanford University and completed his postdoctoral training at the University of Southern California. He served from 2005 to 2011 on the faculty of the University of Michigan, and he is now an associate professor in the Department of Biology at Stanford University. Research in his laboratory focuses on human evolutionary genetics, population-genetic theory, and mathematical phylogenetics.

Tanja Stadler received the PhD degree in mathematics in 2008 from the Technical University of Munich, and did postdoctoral research in the Institute of Integrative Biology at ETH Zürich. Since 2011, she has been a junior group leader at ETH Zürich, working on developing computational methods at the interface of phylogenetics, macroevolution, and epidemiology.