Molecular Phylogenetics and Evolution 161 (2021) 107162

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Heuristics for unrooted, unranked, and ranked anomaly zones under birth-death models

Anastasiia Kim, James H. Degnan

Department of Mathematics and Statistics, University of New Mexico, United States

ARTICLE INFO

Keywords:
Phylogeny
Species trees
Gene trees
Coalescent
Heuristic methods
Anomaly zones

ABSTRACT

Species trees that can generate a nonmatching gene tree topology that is more probable than the topology matching the species tree are said to be in an anomaly zone. We introduce some heuristic approaches to infer whether species trees are in anomaly zones when it is difficult or impossible to compute the entire distribution of gene tree topologies. Here, probabilities of unrooted, unranked, and ranked gene tree topologies under the multispecies coalescent are used. A ranked tree can be viewed as an unranked tree with a temporal ordering of its internal nodes. Overall, considering probabilities of unrooted or unranked gene tree topologies within one nearest neighbor interchange from the species tree topology is a reasonable heuristic to infer the existence of anomalous unrooted or unranked gene trees, respectively. We investigated a test proposed by Linkem et al. (2016) which classifies a species tree as being in an unranked anomaly zone if there is a subset of four taxa in an unranked anomaly zone. We find this test to have high true positive rates, but it can also have high false positive rates. For ranked trees, because at least one of the most probable ranked gene tree topologies must have the same unranked topology as the species tree, we propose to use only those ranked gene trees that have topologies that match the unranked species tree topology. We find that the probability that the species tree is in unrooted and unranked anomaly zones tends to increase with the speciation rate, and the probability of all three types of anomaly zones increases rapidly with the number of taxa. We find that probabilities that species trees are in an anomaly zone can be quite high for moderately high speciation rates.

1. Introduction

The main goal of phylogenetics is to discover the true evolutionary relationships of species. Species trees and gene trees represent how species and individual genes, respectively, evolve from their most recent common ancestors. For a variety of reasons, gene trees may fail to reflect the relationships of the species from which the genes were sampled. Many methods have been developed to infer gene trees from genomic data and then infer the species tree from the set of estimated gene trees. The multispecies coalescent has emerged as a powerful framework that allows modeling sources of gene-species tree incongruence.

Degnan and Rosenberg (2006) proposed the concept of "anomaly zone", a space of the species tree branch lengths that makes a gene tree topology that differs from the species tree topology more probable than the gene tree with the same topology as the species tree. Gene tree topologies more probable than the matching gene tree topology are called *anomalous gene trees*, where a gene tree is said to be matching if it has the same topology as the species tree. This result gave insight for finding when other methods of species tree estimation could be misleading in regions of branch length space resembling, but not identical to, the anomaly zone (Kubatko and Degnan, 2007; Degnan et al., 2009; Wang and Degnan, 2011; Than and Rosenberg, 2011). In particular, many methods that are misleading when the species tree includes short branches often return species tree estimates that correspond to anomalous gene trees (Kubatko and Degnan, 2007; Degnan et al., 2009; Wang and Degnan, 2011; Than and Rosenberg, 2011). Predicting how often such gene trees arise and which topologies can be anomalous can be useful for understanding possible errors for species tree methods such as concatenation. The empiricist should especially be aware of these possibilities for large data sets, which can result in high confidence for an incorrect clade (Kubatko and Degnan, 2007). We also note when there are anomalous gene trees, no particular gene tree can have greater than 1/3 probability (Allman et al., 2011), so that the gene tree distribution is highly heterogeneous. The observation of anomalous gene trees and generally high gene tree heterogeneity also suggests high values for the macroevolutionary parameter of speciation rate.

Although likelihood-based methods, including Bayesian methods (Liu and Pearl, 2007; Heled and Drummond, 2010; Flouri et al., 2018) are not misled by anomalous gene trees, the recognition of the possibility of anomalous gene trees, and especially that they do not exist for three taxa on rooted trees or four taxa on unrooted trees, motivated the development of numerous two-staged methods using rooted triples or

https://doi.org/10.1016/j.ympev.2021.107162
Received 28 December 2019; Received in revised form 21 October 2020; Accepted 23 March 2021
Available online 6 April 2021

quartets [e.g.,] (Ewing et al., 2008; DeGiorgio and Degnan, 2010; Liu et al., 2010; Larget et al., 2010; Mirarab et al., 2014). The concept of the anomaly zone has also been useful for designing simulation studies to test species tree inference methods in challenging regions of parameter space (Kubatko and Degnan, 2007; Liu and Edwards, 2009; Liu et al., 2009; DeGiorgio and Degnan, 2010; Shekhar et al., 2018). Although the theoretical possibility of anomalous gene trees has motivated many methods, the extent that they arise in practice is less clear. Here we find that anomalous gene trees do in fact arise frequently under the widely used birth–death model for speciation, and that they arise more frequently as the number of species and speciation rate increase. Anomalous gene trees have been suggested in empirical papers for skinks (Linkem et al., 2016), gibbons (Shi and Yang, 2017), and flightless birds (Cloutier et al., 2019).

In this paper, we consider three types of anomaly zones, each corresponding to different types of gene trees: unrooted, unranked, and ranked gene trees (Fig. 1), respectively. To study the probability that the species tree is in an anomaly zone, we calculate the probability that the species tree generated from a constant-rate birth–death process lies in unrooted, unranked, or ranked anomaly zones by analytically computing probabilities of gene trees given the simulated species trees.

Because the number of possible tree topologies grows faster than exponentially with the number of species, we propose some heuristic approaches to infer whether larger species trees (i.e., more than eight taxa) are in anomaly zones. The study of various types of anomaly zones can lead to the discovery of the cases when such zones do not overlap with each other for certain species tree topologies and/or branch lengths, meaning that methods based on the one type of gene trees might provide more robust estimates than methods that use other types of gene trees.

2. Materials and methods

The methods in this article involve only topologies (and ranked topologies) of gene trees. A ranked gene tree not only accounts for the topology, as an unranked tree does, but for the order in which lineages join. In practice, branch length information can be used to obtain ranked gene trees. In cases when branch lengths are not estimated very accurately, it might be better to rely on the temporal order of nodes in the gene tree instead (ranked gene tree topology). Rooted unranked or ranked gene tree topologies that are more probable than the unranked or ranked gene tree topology matching the species tree are called anomalous unranked gene trees (AGTs) or anomalous ranked gene trees (ARGTs), respectively. Similarly, unrooted gene trees that are more probable than the matching unrooted gene tree are termed anomalous unrooted gene trees (AUGTs). Species trees that have unrooted, unranked, or ranked anomalous gene trees are said to be in the unrooted, unranked, or ranked anomaly zone (AZ), respectively.

In general, we can uniquely specify an unranked or unrooted gene tree topology by the ranked gene tree topology. The probability of an unranked gene tree topology can be obtained by summing the probabilities of all ranked gene tree topologies that share that unranked topology. Similarly, the probability of an unrooted gene tree topology can be obtained by summing the probabilities of all unranked gene trees with the same unrooted topology. Under the multispecies coalescent model probabilities of unranked gene trees can be computed using coalescent histories (Degnan and Salter, 2005) or ancestral configuration (Wu, 2012). The probabilities of ranked gene trees can be computed as a sum over all ranked histories. The ranked history depicts the sequence of coalescence events, where each coalescence of gene lineages occurs in the species tree interval. The probability for each ranked history can be computed as a product over all speciation intervals.

2.1. Theoretical probability of being in the anomaly zone

Theoretically, the probability that a species tree is an anomaly zone can be obtained by integrating the distribution of species trees under the branching process, using the anomaly zone for limits. In the case of four-taxon unranked trees, the probability under a pure birth model with rate λ is

$$\frac{1}{3}\int_0^\infty \int_0^{a(x)} 6\lambda^2 e^{-\lambda(2x+3y)}\, dy\, dx,$$

where the integrand is the joint density of the branch lengths (Stadler, 2011), the $1/3$ term is the probability that the species tree has a caterpillar topology (i.e., only one two-taxon clade), and

$$a(x) = \log\left[\frac{2}{3} + \frac{3e^{2x}-2}{18(e^{3x}-e^{2x})}\right] \quad (1)$$

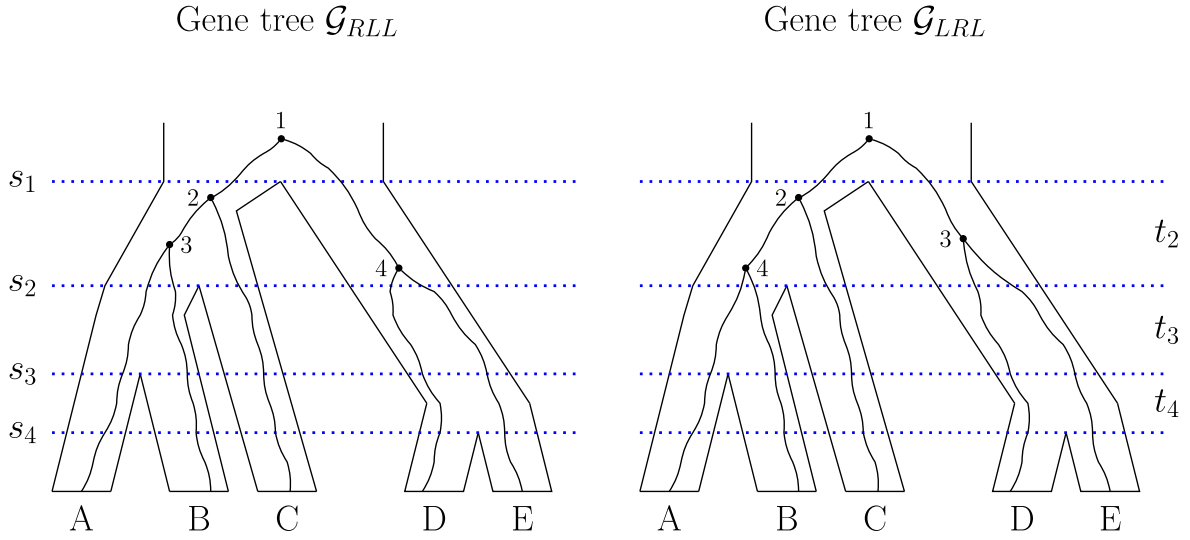

Fig. 1. Ranked gene trees evolving on the five-taxon species trees $\mathcal{T}_{RLL}$. The numbers next to the gene tree nodes indicate rank, with the node 4 indicating the 4th coalescence going from the past to the present. For the left gene tree and the species tree, (D,E) both have rank 4, whereas in the right gene tree, node (D,E) has rank 3. The gene trees have the same unranked topology (((A,B),C), (D,E)) but different ranked topologies. Only the leftmost ranked gene tree topology matches the ranked species tree topology. For each $i = 1, 2, \ldots, n-1, s_i \geq 0$ denotes the time of the ith speciation, t_i represents the interval length between the $(i-1)$th and ith speciation events, and numbers $1, 2, 3, 4$ represent the $1, 2, 3, 4$th coalescence (node with rank $1, 2, 3$, or 4) in the gene tree, respectively.

is the boundary of the anomaly zone. Here x represents the more basal branch in the caterpillar species tree, and for $y < a(x)$, the species tree is in the anomaly zone (Degnan and Rosenberg, 2006). Even in the four-taxon case, however, the integral appears to not be analytically tractable, and requires either numerical or simulation methods to evaluate. For more species, the dimension of the integral would be $n-2$ since the probability requires integrating over all internal branches in the species tree, making the problem more difficult for larger trees. The boundary of the anomaly zone is also more complicated for larger trees (Rosenberg and Tao, 2008). Consequently, we have used the simulation approach for this paper.

2.2. Simulation design

Our simulation approach consisted of the following steps: (1) generation of species trees under a constant rate birth–death model using *TreeSim* (Stadler, 2011), (2) computation of probabilities of gene trees for each species tree, (3) identification of the presence of anomalous gene trees by comparing the probability of the matching gene tree topology to that of the most probable nonmatching gene tree topology, and (4) calculation of the proportion of species trees falling in anomaly zone.

When generating species trees, we let the parameters take values in a biologically plausible range. In particular, the speciation rate λ takes the values of $0.1, 0.5, 1$, and the extinction rate μ depends on λ such that the turnover rate $\frac{\mu}{\lambda}$ is 0 or 0.5. This range of values of (λ, μ) was choosen to observe moderate difference between gene and species trees topologies that allows examining the effect of the species tree parameters on the existence of anomalous gene trees. In this paper, branch lengths in the species tree are in coalescent units $t/(2N)$, where t is the number of generations and $2N$ is the effective population size. The length of a randomly selected interior branch in a Yule (rate λ) tree on n leaves is exponentially distributed with rate 2λ (Stadler and Steel, 2012). Therefore for $\lambda = 0.1$ and $\lambda = 1$, a species tree has a mean branch length of 5 and 0.5 coalescent units, respectively. The range of parameters is also similar to other species tree estimation studies. For example, a study using ASTRAL had differing degrees of incomplete lineage sorting leading to a range of normalized RF distances of gene trees to species trees from 9% to 79% (Mirarab et al., 2016). When $\mu = 0, \lambda = 0.1$ and $\lambda = 1.0$ with $n = 8$ taxa leads to the average normalized RF distances of gene trees to species tree having a range of 13.1% to 59.5%.

Because all three types of anomalous gene trees can exist simultaneously for $n \geqslant 5$-taxon trees, we computed the distributions of gene tree topologies for the 5000 5-, 6-, 7-, and 8-taxon species trees. The results based on the $n \leqslant 8$–taxon trees in Fig. 4 are the same as in Kim et al. (2020) because same species trees were used for these cases.

We proposed some heuristic methods for larger phylogenies and demonstrated their performance on 1000 simulated species trees with 9, 10, 11, and 12 taxa. Using *hybrid-coal* (Zhu and Degnan, 2017), probabilities of unranked gene tree topologies were computed, and from these, probabilities of unrooted topologies for $n \leqslant 8$ were found by summing probabilities of unranked topologies that share the same unrooted topology. We used the *CalGTProb* command from *PhyloNet* (Than et al., 2008) to compute probabilities of unrooted gene trees topologies for $n > 8$. The probabilities of ranked gene tree topologies were computed using *PRANC* (Kim et al., 2020) (https://github.com/anastasiiakim/PRANC). We note that unranked gene tree probabilities can also be computed using *STELLS* (Wu, 2012) and ranked gene tree probabilities can also be computed using *RGTProb* (Disanto et al., 2019).

2.3. Heuristic methods

It is necessary to propose some heuristic approaches for nine and more taxa since computing probabilities of gene trees given a species tree for the entire distribution is computationally intensive. The number of tree topologies grows faster than exponentially with the number of species. For instance, there are 56,700, 1,587,600, and 57,153,600 ranked gene trees for 7, 8, and 9 species, respectively.

To see how topologically different AUGTs, AGTs, and ARGTs can be from species trees we consider the Robinson-Foulds (RF) (Robinson and Foulds, 1981) topological distance between the most probable gene tree topology and the species tree topology for different values of the speciation rate λ, extinction rate μ, and the number of taxa n (Table 1). We observed that in the cases where the species tree produces AGTs or AUGTs, the majority of most probable gene tree topologies were not too distant from the species tree topology (RF-distance $\leqslant 2$). However, it should be noted that RF distances can occasionally be nearly maximal for an AGT. In both unranked and unrooted cases for larger turnover μ/λ, the probabilities of anomalous trees are lower but there are more trees within RF $\leqslant 2$ from a species tree topology. To reduce the computational complexity of determining anomaly zones, we consider only gene trees that are exactly one nearest neighbor interchange (NNI) away from the species tree.

The above observations lead to a useful heuristic for unranked gene trees. Given a species tree, we compute probabilities of unranked gene trees that are exactly one NNI of the species trees. If one of these gene trees is anomalous, then the species tree is classified as being in an unranked anomaly zone. If none of the NNI unranked gene trees is in an unranked anomaly zone, then the species tree is classified as not anomalous. Under this heuristic, false positives (judging a species tree to be in an anomaly zone when it is not) cannot occur, but false negatives can occur when a species tree has AGTs but all AGTs are more than one NNI move from the species tree. For n taxa, the heuristic requires only computing $2n-4+1$ unranked gene tree probabilities (the plus 1 is for the matching tree) for each species tree. The heuristic for unrooted trees is the same as that for unranked trees except that $2n-6+1$ unrooted NNI trees (plus 1 for the matching tree) are used. We note that the heuristic will tend to underestimate the probability that the species tree is in an anomaly zone.

Table 2 shows true positive rates of unrooted and unranked species trees that fall in their respective anomaly zones by computing probabilities of all gene tree topologies that are only one NNI step away from a species tree. Since we have found only a few cases in which all anomalous gene trees were more than two NNI steps away from the species tree topology, we propose that considering unranked or unrooted gene tree topologies within one NNI step from the species tree topology is a reasonable heuristic to infer the existence of AGTs or AUGTs. Our simulation shows that even if the most frequent gene tree topology is farther than one NNI step away from the species tree topology, it is likely that there is at least one other gene tree topology within one NNI from the species tree topology that has larger probability than the matching tree topology.

We use a different strategy in the search for ARGTs. Disanto et al. (2019) proved mathematically that at least one of the most probable ranked gene tree topologies must have the same unranked topology as the species tree (it is possible that several conflicting trees are exactly tied for most probable). Based on this, we propose to use only those ranked gene trees that have topologies that match the unranked species tree topology to check for anomalousness. This is an exact test with no false positives and no false negatives. This greatly reduces the number of tree probabilities to be computed when checking whether the species tree has an ARGT. For example, instead of computing 1,587,600 probabilities for the balanced 8-taxon tree, we need to compute only 80 probabilities, since there are 80 possible rankings for the balanced 8-taxon topology.

2.4. Limit of the anomaly zone

Another heuristic test was proposed to identify the unranked anomaly zone in larger trees by Linkem et al. (2016). They consider the limit of the anomaly zone $a(x)$ for the four-taxon caterpillar tree (Fig. 2

Table 1
Frequency distribution of the Robinson-Foulds distances between the most probable gene tree topology and the species tree topology. We simulated 5000 species trees for each combination of n, λ, and μ.

		$\frac{\mu}{\lambda} = 0$						$\frac{\mu}{\lambda} = 0.5$					
		unrooted			unranked			unrooted			unranked		
λ	n	0	2	4–8	0	2	4–10	0	2	4–8	0	2	4–8
0.1	5	99.76	0.24		99.42	0.58		99.78	0.22		99.64	0.36	
	6	99.32	0.68		99.12	0.86	0.02	99.56	0.44		99.30	0.70	
	7	98.80	1.08	0.12	98.54	1.32	0.14	99.40	0.58	0.02	99.24	0.72	0.04
	8	98.54	1.44	0.02	98.30	1.68	0.02	99.26	0.68	0.06	98.96	0.98	0.06
0.5	5	95.68	4.32		90.36	9.08	0.56	96.28	3.72		93.44	6.26	0.30
	6	92.24	7.36	0.40	87.76	11.14	1.10	94.14	5.60	0.26	90.92	8.40	0.68
	7	88.50	10.48	1.02	84.20	13.78	2.02	91.40	7.98	0.62	88.74	10.30	0.96
	8	85.42	13.12	1.46	81.32	16.30	2.38	88.84	10.24	0.92	86.10	12.50	1.40
1	5	90.38	9.62		78.90	18.54	2.56	91.54	8.46		83.90	14.42	1.68
	6	83.66	14.88	1.46	72.68	22.26	5.06	86.60	12.58	0.82	80.16	16.86	2.98
	7	76.60	19.78	3.62	67.20	25.46	7.34	81.02	16.50	2.48	74.94	20.44	4.62
	8	70.84	23.82	5.34	61.34	29.18	9.48	76.20	19.70	4.10	70.90	22.74	6.36

Table 2
True positive rates of species trees that fall in the unrooted and unranked anomaly zones. We simulated 5000 species trees for each combination of n, λ, and μ.

		$\mu = 0$			$\frac{\mu}{\lambda} = 0.5$		
rate (in%)	n	0.1	0.5	1	0.1	0.5	1
unrooted	5	100.0	100.0	100.0	100.0	100.0	100.0
	6	100.0	100.0	99.88	100.0	99.66	100.0
	7	100.0	99.83	100.0	100.0	100.0	99.68
	8	100.0	100.0	99.73	100.0	99.82	99.75
unranked	5	100.0	100.0	99.91	100.0	100.0	100.0
	6	100.0	99.67	99.27	100.0	99.56	100.0
	7	100.0	99.75	99.76	100.0	100.0	99.76
	8	100.0	99.89	99.84	100.0	100.0	99.73

(a)) defined earlier (Eq. 1). Given that y and x are lengths of an internal branch and its immediate ancestor in the species tree, a four-taxon species tree falls in the unranked anomaly zone if it satisfies the condition $y < a(x)$. Linkem et al. (2016) proposed that this condition could be checked for any two consecutive branches in a tree within a larger species tree to conclude whether there is evidence of the unranked anomaly zone. In particular, their examination shows that several pairs of parent–child internodes in the skink phylogeny satisfy this condition, and that this species tree is therefore in the anomaly zone. This may explain a strong conflict between species trees inferred under the coalescent versus using concatenation.

We tested the Linkem et al. (2016) heuristic on small trees to estimate false positive and false negative rates. We then applied this approach on 5–8-taxon species trees since the exact error can be computed by analytically computing probabilities for the entire gene tree distributions. We simulated 5000 species trees under a constant rate birth–death process for each combination of $\lambda = 0.1, 0.5, 1$ and $\mu/\lambda = 0$, 0.5. For each of these species trees, we calculated the probabilities of all possible unranked gene tree topologies and compare a species tree topology with the most probable gene tree topology to see whether a corresponding species tree fell in the unranked anomaly zone.

All pairs of consecutive internode branch lengths were used to check if at least one pair satisfied the anomaly zone limit condition $y < a(x)$ (see Fig. 2(b)). If there was evidence of the unranked anomaly zone based on this condition, the species tree was checked if it was in the unranked anomaly zone based on computing probabilities for the full gene tree distribution.

Table 3 depicts the percentages of species trees that were correctly identified (true positives) to be in the unranked anomaly zone. Table 3 also shows false positive percentages of trees that satisfy the anomaly zone condition $y < a(x)$ but are not in the anomaly zone.

We observed that the false positive rate slowly increases with the number of taxa and speciation rate λ. Despite the relatively high false positive rate, the test is still useful for checking that a tree does not fall in an anomaly zone — if none of two consecutive branches on a path from the root to a tip satisfy $y < a(x)$, then it is very unlikely that the species tree is in an unranked anomaly zone.

We considered a similar test for the ranked anomaly zone. As discussed in Degnan et al. (2012), the $\mathcal{T}_{RLL}$ tree is the only unranked topology that could produce anomalous ranked gene trees (Fig. 1). There are three possible rankings of the unranked $\mathcal{T}_{RLL}$ tree topology. The 5-taxon species tree produces an anomalous ranked gene tree if for the

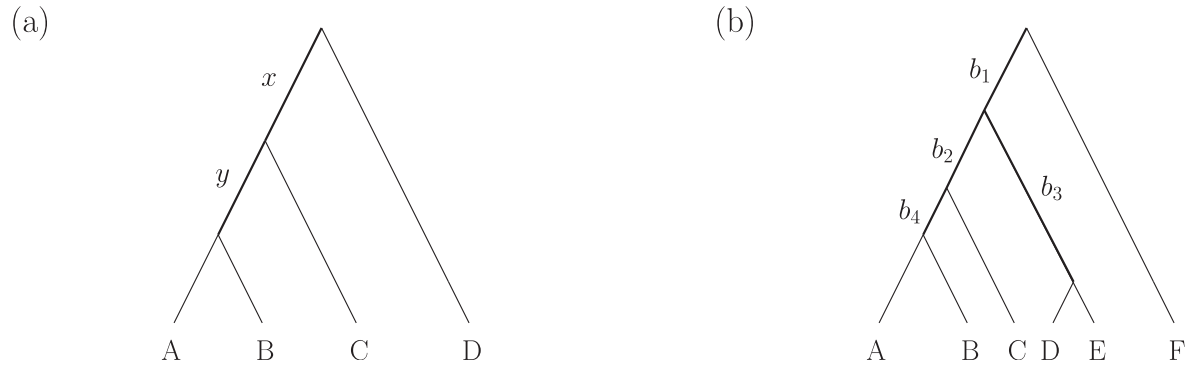

Fig. 2. (a) Using the Linkem et al. (2016) heuristic, the four-taxon species tree is said to be in unranked anomaly zone if two consecutive branches with lengths x and y in coalescent units, satisfy the anomaly zone condition $y < a(x)$. (b) Pairs of two internal consecutive branches (i.e., (b_1, b_2), (b_1, b_3), and (b_2, b_4)) in larger tree that can be checked for anomaly zone condition $y < a(x)$. If at least one pair satisfy condition, then the species tree is likely to be in an unranked anomaly zone.

Table 3
Percentages of species trees that were correctly identified to be in the unranked anomaly zone by satisfying the unranked anomaly zone limit condition $y < a(x)$, where y and x are branch lengths of an internal node and its parental node in the species tree. All consecutive pairs of internode branch lengths were used to check if at least one pair satisfying the anomaly zone limit condition $y < a(x)$. The table also depicts percentages of species trees that were incorrectly identified to be in the unranked anomaly zone. There were 5000 species trees were simulated for each combination of n,λ, and μ. The probabilities that a species tree lies in the unranked anomaly zone were computed based on the full gene tree distribution (true cases).

		$\mu = 0$			$\frac{\mu}{\lambda} = 0.5$		
rate (in%)	n	0.1	0.5	1	0.1	0.5	1
True positive	5	96.55	94.19	92.80	100.00	95.43	94.29
	6	97.73	95.26	94.51	100.00	93.83	94.46
	7	93.15	95.95	95.24	97.37	95.74	96.01
	8	92.94	95.61	95.45	98.08	93.96	97.32
False positive	5	0.00	2.16	7.06	0.06	1.52	4.76
	6	0.18	4.08	9.72	0.06	2.26	7.00
	7	0.12	4.60	11.76	0.02	3.44	9.28
	8	0.18	5.82	13.30	0.12	4.06	9.80

species tree $\mathcal{T}_{RLL}$, the probability of $\mathcal{G}_{LRL}$ is greater than the probability of $\mathcal{G}_{RLL}$ Degnan et al. (2012). Overall, a 5-taxon species tree $\mathcal{T}_{RLL}$ is in the ranked anomaly zone if

$$t_4 < \log\left[\frac{72 - 48e^{-t_2} - 9e^{-t_3}(4 + e^{-t_3}) + e^{-t_2-t_3}(16e^{-t_3} + 16e^{-t_2} - 8e^{-t_2-t_3} + e^{-3t_2-t_3})}{72 - 48e^{-t_2} - 48e^{-t_3} + 4e^{-t_2-t_3}(6 + e^{-2t_2})}\right], \quad (2)$$

where $t_2, t_3,$ and t_4 are lengths of the three consecutive speciation intervals (Fig. 1). We use this expression to determine candidates for being a ranked anomaly zone by checking every three consecutive speciation intervals in larger rooted species trees. Unfortunately, this method has high false positive and low true positive rates (Table 4), and can't be used as a quick test whether a species tree falls into the ranked anomaly zone.

2.5. The anomaly zone of skinks

Linkem et al. (2016) used coalescence based *MP-EST* (Liu et al., 2010) to estimate a species tree from 429 estimated gene trees. They used the inferred species tree topology from *MP-EST* and then estimated median branch lengths by summarizing the posterior distribution from *BP&P* (Yang and Rannala, 2010). We used the estimated 16-taxon skink phylogeny (see Table 3 in Linkem et al. (2016)) and computed unranked probabilities of the 28 gene trees that are one NNI step away from the species tree. We found that there are four anomalous gene trees, and the probability of the unranked matching tree, $1.3987e-06$, is lower than the probability of the most probable unranked nonmatching tree, $2.66115e-06$, which confirms that this species tree is in the unranked anomaly zone. We note that a general difficulty in determining whether a species tree lies in the anomaly zone is that branch lengths in the species tree are not directly estimated in coalescent units. Estimates can be made by using the estimated generation time divided by the estimated population size, although this ratio is not usually directly estimated.

We also found that the probability of the unrooted matching tree $8.972783e-06$ is lower than the probability of the unrooted nonmatching tree $1.719202e-05$, which confirms that this species tree is also in the unrooted anomaly zone. Among 26 one step NNI 16-taxon unrooted trees, four are anomalous.

However, computing probabilities of the $73,920$ 16-taxon ranked gene trees that share the same unranked topology as that of the species tree did not indicate that the species tree is in the ranked anomaly zone. The most probable ranked gene tree has the same ranked topology as the species tree and this probability is $8.166337e-09$.

Table 4
Percentages of species trees that were correctly identified to be in the ranked anomaly zone by satisfying the anomaly zone condition for 5-taxon tree stated in Eq. 2. All consecutive pairs of three interval lengths were used to check if at least one pair satisfying the ranked anomaly zone condition for 5-taxon tree in larger trees. The table also depicts percentages of species trees that were incorrectly identified to be in the ranked anomaly zone. There were 5000 species trees were simulated for each combination of n,λ, and μ. The probabilities that a species tree lies in the ranked anomaly zone were computed based on the full gene tree distribution (true cases).

		$\mu = 0$			$\frac{\mu}{\lambda} = 0.5$		
rate (in%)	n	0.1	0.5	1	0.1	0.5	1
True positive	6	82.61	78.45	72.36	95.65	82.86	73.31
	7	86.21	68.60	65.27	50.00	69.07	65.27
	8	71.83	66.50	61.69	12.99	65.92	62.56
False positive	6	2.02	11.10	14.88	2.00	10.18	15.04
	7	3.58	16.08	19.68	2.96	15.20	20.36
	8	5.24	19.62	22.72	5.32	18.26	23.00

3. Simulation results

Figs. 3 and 4 show probabilities of the species tree being in the unranked, unrooted, and ranked anomaly zones for different combinations of the number of taxa n, speciation rate λ, and extinction rate μ. For all types of trees, the probability of being in an anomaly zone increases with the number of taxa and with λ in this range. Increasing λ makes consecutive short branches more likely to appear in the species tree, which explains the increasing trend in probabilities of the unranked and unrooted anomaly zones.

We also observed the opposite effect of the turnover rate μ/λ on the probability of producing unranked and unrooted versus ranked anomalous gene trees. On average, branches closer to the root are longer than other branches in a tree as the turnover rate increases (Table 5). This leads to a longer speciation interval near the root and explains the decreasing trend in probability in the unranked and unrooted anomaly zones and the increasing trend in the ranked anomaly zone as turnover rate increases since longer branches near the root can produce ARGTs when other branches are short.

To reduce the computational complexity, we use the heuristic described in the *Heuristic methods* section of considering unranked and unrooted gene tree topologies within one NNI step from the species tree topology to infer the existence of AGTs and AUGTs for larger trees ($n > 8$). For ranked gene tree topologies, we consider only those that share the same unranked topology with the species tree.

We used Venn diagrams to visualize results obtained from analyzing larger trees. Figs. 5 and 6 depict the relationships between unrooted, unranked, and ranked anomaly zones (AZ_{UGT}, AZ_{GT} and AZ_{RGT}). Each slice represents the number of species trees in the anomaly zone. We observe that for low speciation rates, AZ_{UGT} is often a subset of AZ_{GT}, and there are not many species trees in any of the three anomaly zones. However, as λ increases, species trees start to produce anomalous gene trees more often in each type of anomaly zone, and the proportion of

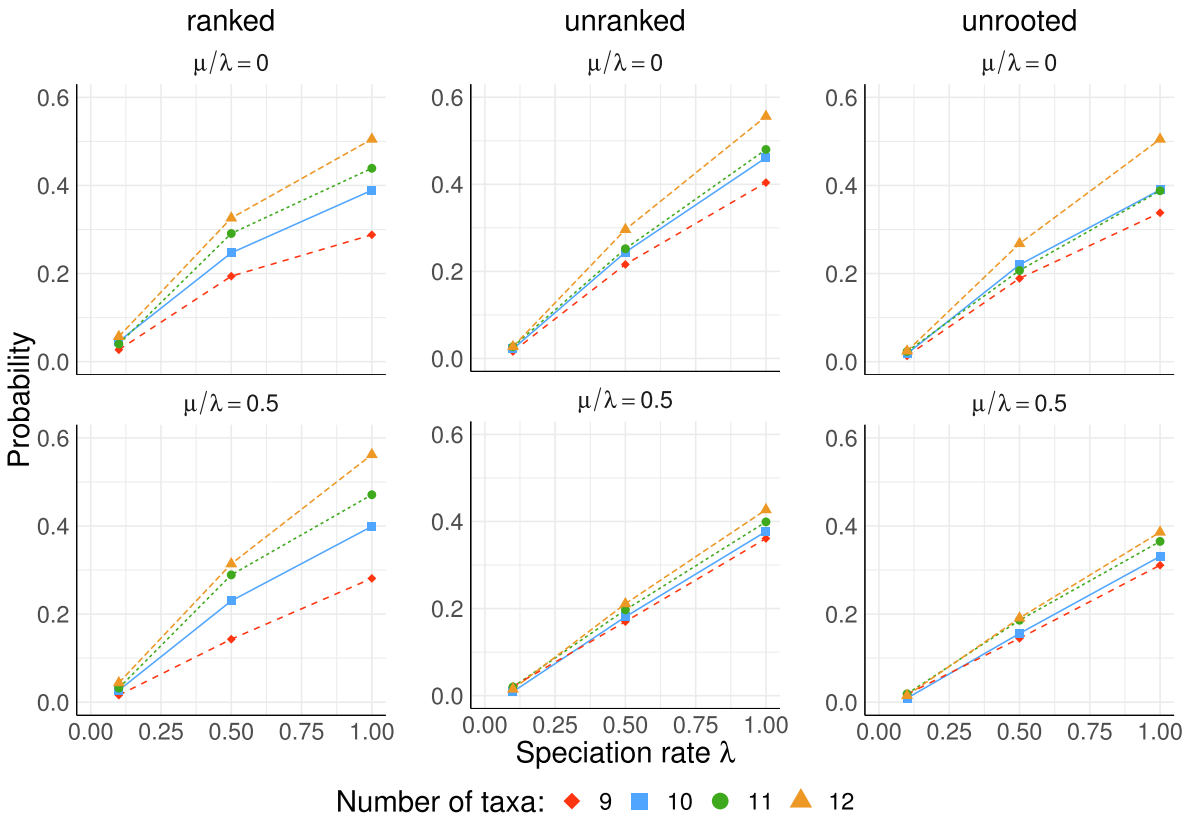

Fig. 3. The impact of the speciation rate parameter λ and turnover rate μ/λ on the existence of ranked, unranked, and unrooted anomaly zones. We simulated 1000 species trees for $n = 9, 10, 11, 12$ taxa using a constant rate birth–death process with rates $\lambda \in \{0.1, 0.5, 1\}$ and $\frac{\mu}{\lambda} \in \{0, 0.5\lambda\}$. For each combination of (n, λ, μ) the probabilities of the species tree being in each type of μ/λ anomaly zone were computed.

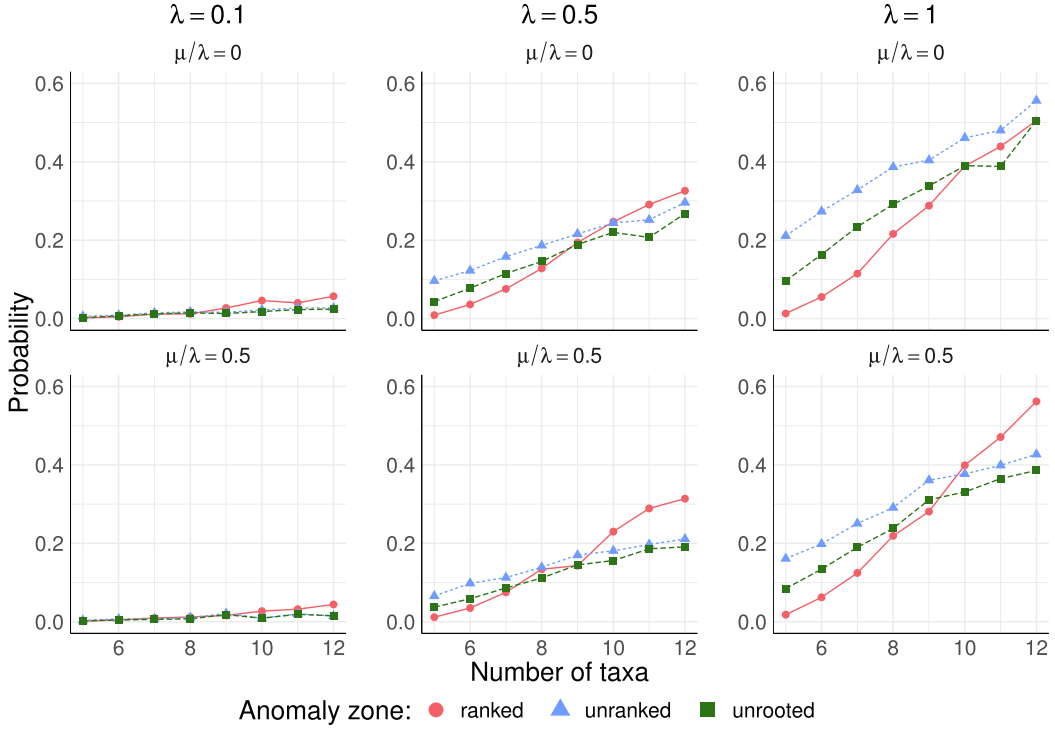

Fig. 4. The impact of the number of taxa n on the existence of ranked, unranked, and unrooted anomaly zones given speciation λ and extinction μ rates. 5000 species trees were simulated for $n = 5, 6, 7, 8$ taxa and 1000 species trees for $n = 9, 10, 11, 12$ taxa using a constant rate birth–death process with rates $\lambda \in \{0.1, 0.5, 1\}$ and $\frac{\mu}{\lambda} \in \{0, 0.5\lambda\}$. For each combination of (n, λ, μ) the probabilities of the species tree being in each type of μ/λ anomaly zone were computed.

Table 5
Average length and average proportion of the intervals in the tree. We generated 10000 8-taxon trees under the constant rate birth–death process with speciation rate $\lambda = 1$ and extinction rates $\mu = 0, 0.5$. The proportion for each tree was calculated by dividing the intreval length by the sum of the interval lengths in the tree. The speciation intervals are represented by t_2 to t_8 from past to present, respectively.

	$\mu = 0$		$\frac{\mu}{\lambda} = 0.5$	
t	length	proportion	length	proportion
t_2	0.50	0.26	0.84	0.30
t_3	0.33	0.19	0.51	0.20
t_4	0.25	0.15	0.35	0.15
t_5	0.20	0.12	0.26	0.12
t_6	0.17	0.10	0.20	0.09
t_7	0.14	0.09	0.16	0.07
t_8	0.12	0.08	0.13	0.06

trees in the intersection of two or more anomaly zones also increases (Figs. 5(b), 5(d)). As shown in Fig. 6, turnover rate does not make a substantial difference to the relationships between different types of anomaly zones.

Overall, using the proposed heuristic approach for $n > 8$, species trees produce more AGTs than AUGTs and ARGTs (Figs. 3–5). Species trees more often fall in AZ_{RGT} than in AZ_{GT} when the speciation rate λ is small, but as λ increases they start to produce more AGTs than ARGTs. The unrooted and unranked anomaly zones have more trees in common, and the ranked anomaly zone is more separated from them.

It is hard to detect what kind of species tree shapes tend to fall more often in the certain types of anomaly zones. We calculate the Colless index (Colless, 1982) for 9-taxon species trees, simulated under the constant rate birth process with $\lambda = 1$, in the ranked and unranked anomaly zones (Fig. 7). More balanced trees tend to fall in the ranked anomaly zone and more imbalanced into the unranked anomaly zone. The average Colless statistic is 9.92 for the ranked anomaly zone and 12.39 for the unranked anomaly zone. For the 12-taxon species trees, the average Colless statistics are 16.96 and 20.01 for the ranked and unranked anomaly zone, respectively.

4. Discussion

Although the theoretical possibility of anomalous gene trees has been known over a decade, it has not been clear how often this phenomenon occurs in practice. This paper addresses this question, albeit indirectly, by estimating how often AGTs, ARGTs, and AUGTs occur under the widely-used birth–death models of speciation.

Ranked gene trees have not been explored as much as unrooted or rooted but unranked gene tree topologies. Ranked gene trees represent a compromise between using only the topologies versus preserving branch length information in the gene trees. From previous work, ARGTs do not exist for four-taxon trees but can for larger trees. We note that ARGTs tend to be much closer topologically to the species tree than AGTs since they usually share the same unranked topology as the species tree but differ in their ranking. Although we recently developed a maximum likelihood method for ranked gene trees (Kim and Degnan, 2020), these results suggest that computationally more efficient consensus tree methods or versions of pseudolikelihood such as MP-EST could also be extended to ranked gene trees on four taxa.

The probability of the species tree having an anomalous gene tree increases with the speciation rate and the number of species sampled. Our simulation was based on unconstrained tree heights for the species tree, so that sampling more species meant that the total expected height of the tree also increased. Even under this design, probabilities of being in anomaly zones increased with more taxa. In practice, the question of

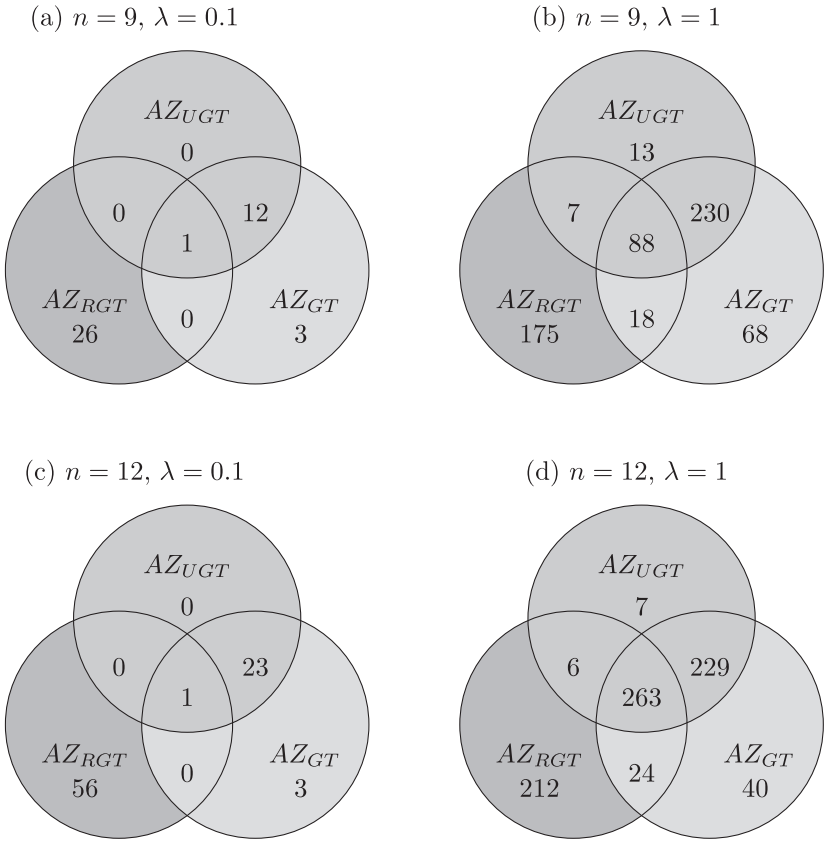

Fig. 5. Relationships between unrooted, unranked, and ranked anomaly zones. We considered 1000 species trees with birth parameters $\lambda = 0.1, 1$ for 9 and 12 taxon species trees. In each Venn diagram, each slice represents the number of anomalous trees found. Note that figures are not drawn to scale.

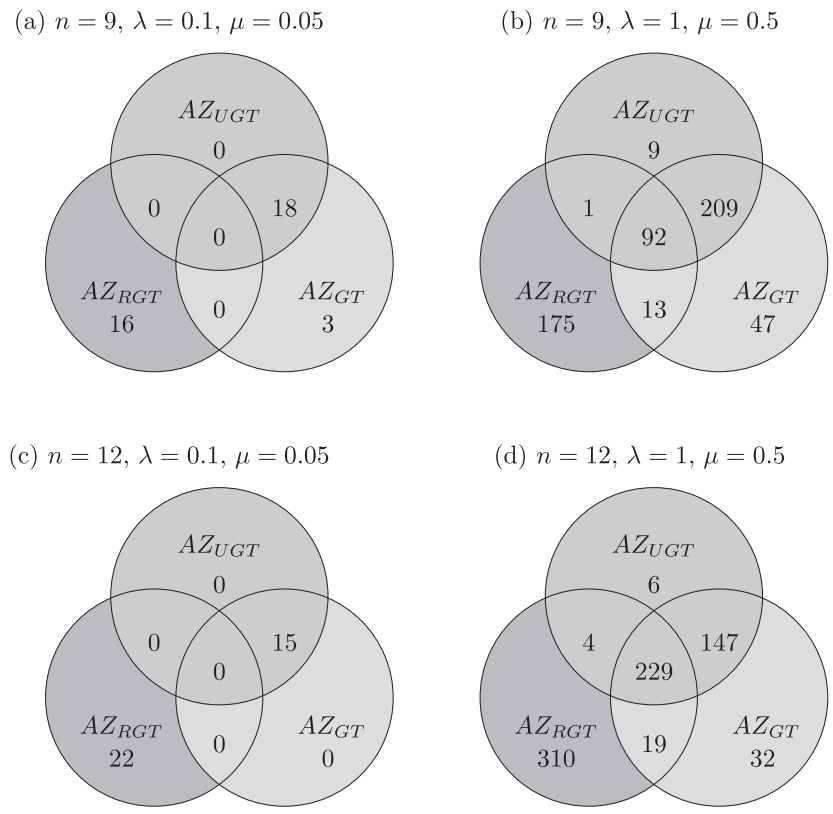

Fig. 6. Relationships between unrooted, unranked, and ranked anomaly zones. We considered 1000 species trees with birth parameters $\lambda = 0.1, 1$ and turnover rates $\frac{\mu}{\lambda} = 0.5$ for 9 and 12 taxon species trees. In each Venn diagram, each slice represents the number of anomalous trees found. Note that figures are not drawn to scale.

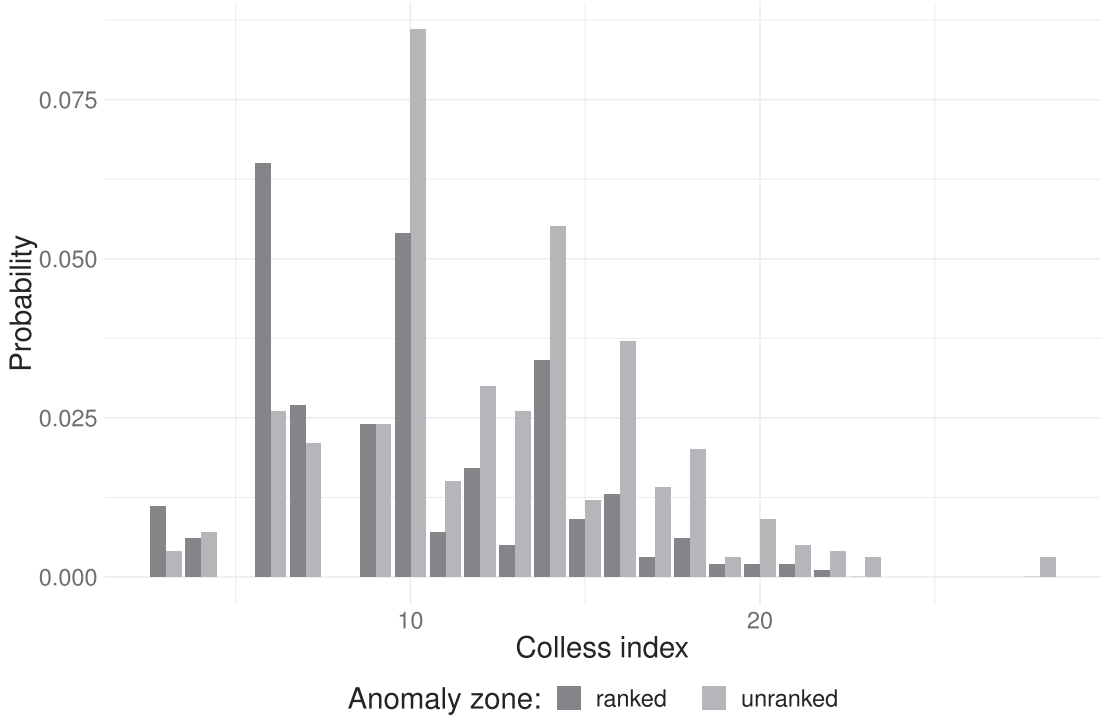

Fig. 7. The probability of species trees being in the unranked and ranked anomaly zones. We simulated 1000 9-taxon species trees under the birth process with $\lambda = 1$. The Colless statistic was computed for all trees in the anomaly zone. The larger the value, the more imbalanced the tree is.

how often anomaly zones arise is still difficult to answer, but if something is known about the speciation rate, particularly in coalescent units, then this study can help to answer that question. It should perhaps not be surprising that with larger trees, there are more possibilities for portions of a tree that can have anomalousness. Empiricists should not necessarily be alarmed that larger trees are very likely to be in anomaly zones,

but the results highlight the value of using statistically sound methods to estimate species trees.

We did not study the effect of sampling more densely within a clade. When more taxa are sampled within a clade, then the total height of the species tree is kept constant but more branches are added, making the branches shorter and more likely to produce at least AGTs and AUGTs. We leave such effects of taxon sampling to future work.

Knowing how often and what kind of anomalous gene trees can generate species trees can help to design valid simulation studies. In this paper, our proposed heuristic approaches are generally useful for anomaly zone calculation with high true positives and no false positives. Our simulation study revealed that the most probable tree often is not topologically far from the species tree. When the most probable tree is far from the species tree (in terms of RF distance), there are usually other AGTs or AUGTs that are closer to the species tree, even if they are not the highest probability gene trees. Therefore, using the nearest neighbor interchange branch rearrangement technique, we found that considering only unrooted and unranked gene trees within one NNI move from the species tree topology is a good heuristic to infer the existence of anomalous unrooted and unranked gene trees, respectively. This heuristic underestimates the probability that the species tree is in an anomaly zone due to there being few false negatives but no false positives. This tends to reinforce our conclusion that there is a high probability that a species tree can have AGTs or AUGTs for moderately high levels of speciation. Less balanced tree shapes also tend to have higher probabilities of AGTs and AUGTs (Fig. 7). If standard birth–death models underestimate levels of imbalance in real phylogenies (Mooers and Heard, 1997; Bortolussi et al., 2005; Stadler et al., 2016), then we can expect probabilities of AGTs and AUGTs to be even higher than these simulations suggest since less balanced species trees are more easily in anomaly zones (Degnan and Salter, 2005; Degnan and Rosenberg, 2006; Rosenberg and Tao, 2008).

Acknowledgments

This work was supported by National Institutes of Health R01 grant GM117590. We thank the UNM Center for Advanced Research Computing, supported in part by the National Science Foundation, for providing the high performance computing resources used in this work. We are grateful for helpful comments from Noah Rosenberg on the earlier version of this manuscript. We also thank two anonymous reviewers for comments.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.ympev.2021.107162.

References

Allman, E.S., Degnan, J.H., Rhodes, J.A., 2011. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. J. Math. Biol. 62, 833–862.
Bortolussi, N., Durand, E., Blum, M., François, O., 2005. apTreeshape: statistical analysis of phylogenetic tree shape. Bioinformatics 22, 363–364.
Cloutier, A., Sackton, T.B., Grayson, P., Clamp, M., Baker, A.J., Edwards, S.V., 2019. Whole-genome analyses resolve the phylogeny of flightless birds (palaeognathae) in the presence of an empirical anomaly zone. Syst. Biol. 68, 937–955.
Colless, D.H., 1982. Review of phylogenetics: the theory and practice of phylogenetic systematics. Syst. Zool. 31, 100–104.
DeGiorgio, M., Degnan, J.H., 2010. Fast and consistent estimation of species trees using supermatrix rooted triples. Mol. Biol. Evol. 27, 552–569.
Degnan, J.H., Rosenberg, N.A., 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2, 762–768.
Degnan, J.H., Salter, L.A., 2005. Gene tree distributions under the coalescent process. Evolution 59, 24–37.
Degnan, J.H., DeGiorgio, M., Bryant, D., Rosenberg, N.A., 2009. Properties of consensus methods for inferring species trees from gene trees. Syst. Biol. 58, 35–54.
Degnan, J.H., Rosenberg, N.A., Stadler, T., 2012. The probability distribution of ranked gene trees on a species tree. Math. Biosci. 235, 45–55.
Disanto, F., Miglionico, P., Narduzzi, G., 2019. On the unranked topology of maximally probable ranked gene tree topologies. J. Math. Biol. 79, 1205–1225.
Ewing, G.B., Ebersberger, I., Schmidt, H.A., Von Haeseler, A., 2008. Rooted triple consensus and anomalous gene trees. BMC Evol. Biol. 8, 118.
Flouri, T., Jiao, X., Rannala, B., Yang, Z., 2018. Species tree inference with bpp using genomic sequences and the multispecies coalescent. Mol. Biol. Evol. 35, 2585–2593.
Heled, J., Drummond, A.J., 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27, 570–580.
Kim, A., Degnan, J.H., 2020. PRANC: ML species tree estimation from the ranked gene trees under coalescence. Bioinformatics 36 (18), 4819–4821.
Kim, A., Rosenberg, N.A., Degnan, J.H., 2020. Probabilities of unranked and ranked anomaly zones under birth-death models. Mol. Biol. Evol. 37, 1480–1494.
Kubatko, L.S., Degnan, J.H., 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56, 17–24.
Larget, B.R., Kotha, S.K., Dewey, C.N., Ané, C., 2010. Bucky: gene tree/species tree reconciliation with bayesian concordance analysis. Bioinformatics 26, 2910–2911.
Linkem, C.W., Minin, V.N., Leache, A.D., 2016. Detecting the anomaly zone in species trees and evidence for a misleading signal in higher-level skink phylogeny (squamata: Scincidae). Syst. Biol. 65, 465–477.
Liu, L., Edwards, S.V., 2009. Phylogenetic inference in the anomaly zone. Syst. Biol. 58, 452–460.
Liu, L., Pearl, D.K., 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., Yu, L., Pearl, D.K., Edwards, S.V., 2009. Estimating species phylogenies using coalescence times among sequences. Syst. Biol. 58, 468–477.
Liu, L., Yu, L., Edwards, S.V., 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. 10, 302.
Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., Warnow, T., 2014. Astral: genome-scale coalescent-based species tree estimation. Bioinformatics 30, i541–i548.
Mirarab, S., Bayzid, M.S., Warnow, T., 2016. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. Syst. Biol. 65, 366–380.
Mooers, A.O., Heard, S.B., 1997. Inferring evolutionary process from phylogenetic tree shape. Quart. Rev. Biol. 72, 31–54.
Robinson, D.F., Foulds, L.R., 1981. Comparison of phylogenetic trees. Math. Biosci. 53, 131–147.
Rosenberg, N.A., Tao, R., 2008. Discordance of species trees with their most likely gene trees: the case of five taxa. Syst. Biol. 57, 131–140.
Shekhar, S., Roch, S., Mirarab, S., 2018. Species tree estimation using astral: how many genes are enough? IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB) 15, 1738–1747.
Shi, C.-M., Yang, Z., 2017. Coalescent-based analyses of genomic sequence data provide a robust resolution of phylogenetic relationships among major groups of gibbons. Mol. Biol. Evol. 35, 159–179.
Stadler, T., 2011. Simulating trees on a fixed number of extant species. Syst. Biol. 60, 676–684.
Stadler, T., Steel, M., 2012. Distribution of branch lengths and phylogenetic diversity under homogeneous speciation models. J. Theor. Biol. 297, 33–40.
Stadler, T., Degnan, J.H., Rosenberg, N.A., 2016. Does gene tree discordance explain the mismatch between macroevolutionary models and empirical patterns of tree shape and branching times? Syst. Biol. 65, 628–639.
Than, C.V., Rosenberg, N.A., 2011. Consistency properties of species tree inference by minimizing deep coalescences. J. Comput. Biol. 18, 1–15.
Than, Cuong, Ruths, Derek, Nakhleh, Luay, 2008. Phylonet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinform. 9, 322.
Wang, Y., Degnan, J.H., 2011. Performance of matrix representation with parsimony for inferring species from gene trees. Stat. Appl. Genet. Mol. 10, 21.
Wu, Y., 2012. Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. Evol.: Int. J. Organ. Evol. 66, 763–775.
Yang, Z., Rannala, B., 2010. Bayesian species delimitation using multilocus sequence data. Proc. Nat. Acad. Sci. 107, 9264–9269.
Zhu, S., Degnan, J.H., 2017. Displayed trees do not determine distinguishability under the network multispecies coalescent. Syst. Biol. 66, 283–298.