RESEARCH ARTICLE

Statistical inconsistency of the unrooted minimize deep coalescence criterion

Ayed A. R. Alanzi[1], **James H. Degnan**[2]*

1 Mathematics Department, College of Science and Human Studies of Hotat Sudair, Majmaah University, Majmaah, Saudi Arabia, **2** Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM, United States of America

* jamdeg@unm.edu

Abstract

Species trees, which describe the evolutionary relationships between species, are often inferred from gene trees, which describe the ancestral relationships between sequences sampled at different loci from the species of interest. A common approach to inferring species trees from gene trees is motivated by supposing that gene tree variation is due to incomplete lineage sorting, also known as deep coalescence. One of the earliest methods motivated by deep coalescence is to find the species tree that minimizes the number of deep coalescent events needed to explain discrepancies between the species tree and input gene trees. This minimize deep coalescence (MDC) criterion can be applied in both rooted and unrooted settings. where either rooted or unrooted gene trees can be used to infer a rooted species tree. Previous work has shown that MDC is statistically inconsistent in the rooted setting, meaning that under a probabilistic model for deep coalescence, the multispecies coalescent, for some species trees, increasing the number of input gene trees does not make the method more likely to return a correct species tree. Here, we obtain analogous results in the unrooted setting, showing conditions leading to inconsistency of the MDC criterion using the multispecies coalescent model with unrooted gene trees for four taxa and five taxa.

Citation: Alanzi AAR, Degnan JH (2021) Statistical inconsistency of the unrooted minimize deep coalescence criterion. PLoS ONE 16(5): e0251107. https://doi.org/10.1371/journal.pone.0251107

Editor: Michael Scott Brewer, East Carolina University, UNITED STATES

Received: September 20, 2020

Accepted: April 20, 2021

Published: May 10, 2021

Data Availability Statement: The article is theoretical and used data printed in the tables.

Funding: JD was supported by National Institutes of Health R01 GM117590, https://www.nih.gov/ The sponsor played no role in the design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Evolutionary trees estimated at different loci, *gene trees*, vary from one another and from the *species tree*, which represents the history of speciation events. Although there are many causes of such gene tree discordance, one of the most commonly modeled is *deep coalescence*, the failure of two or more gene lineages to coalesce (i.e., be copied from the same gene in the population) in their most recent ancestral population [1, 2]. This phenomenon, also called *incomplete lineage sorting*, is modeled by the *multispecies coalescent*, which makes probabilistic predictions for the probabilities of different gene tree topologies to be observed in a sample of gene trees [3–5]. Although other sources of gene tree heterogeneity are possible, such as gene duplication and loss, hybridization, recombination within genes, and ancient population structure [1, 2, 6, 7] deep coalescence is thought to be quite common for species that underwent rapid radiations [8, 9] and is often used to infer species relationships from gene trees [2, 10, 11].

Species tree inference methods can be based on sequence data or based on analyzing gene trees (e.g., consensus methods). These latter techniques are also called *two-stage methods, meaning that a first stage is estimating the gene trees, and the second stage is combining the information in the gene trees to estimate the species tree. Two-stage techniques are typically computationally faster than sequence-based techniques* and do not have issues with convergence of MCMC algorithms that have arisen for real data sets with many loci for Bayesian methods [12]. Consequently, two-stage methods have remained popular in spite of more sophisticated sequence-based methods. Whether or not a two-stage method is explicitly motivated by the multispecies coalescent, a central concern is whether it is statistically consistent under the multispecies coalescent model. A method is consistent in this setting if the probability that it returns the correct species tree topology tends to 1.0 as the number of loci tends to infinity.

Among two-stage methods for inferring species trees, some have been found to be consistent from known gene trees, while others have been shown to be inconsistent. Two-stage methods that have been shown to be statistically consistent (assuming gene trees are known without error) include rooted triple consensus [13], ASTRAL [14], MP-EST [15], NJ_{st} (also called USTAR) [16, 17], and STAR [18]. Some have been found to be statistically inconsistent, including democratic vote [19], greedy consensus [20], matrix representation with parsimony [21], and the minimize deep coalescence (MDC) method in the rooted setting [22]. The MDC method, the focus of this paper, infers the species tree which minimizes the total number of deep coalescence events needed to explain each gene tree, summed over all the input gene trees.

The idea behind MDC was introduced by Maddison (1997) and was an early method to be implemented for inferring rooted species trees from rooted gene trees motivated by deep coalescence [23–25]. Initially, the method was only applied with rooted gene trees as input, and implementations returned a rooted inferred species tree. To illustrate the idea, consider the species tree and gene trees in Fig 1.

[26], which use approximate Bayesian computation (ABC), is the first method we are aware of to explicitly use the coalescent (i.e., using probabilities from the model) to estimate rooted species trees from unrooted gene trees; however, a version of the minimize deep coalescence

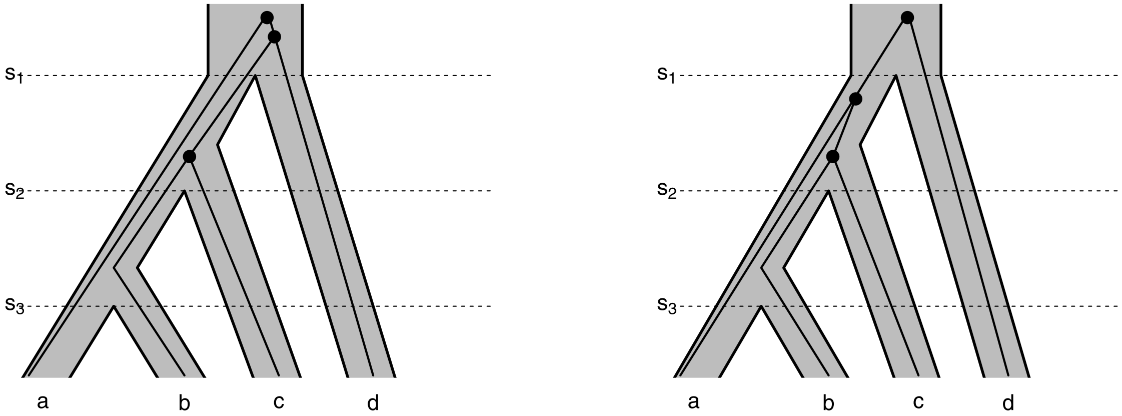

Fig 1. Example species tree—the shaded grey tree with topology (((a,b),c),d)—with two gene trees embedded. In the left example the gene tree is (((B,C),D),A), and on the right the gene tree is (((B,C),A),D). Both gene trees have the same unrooted gene tree, ((B, C),A,D), but different deep coalescence costs. On the left, there are two cases, at time S_1 and S_2 where there are two lineages "exiting" a population (going from the present to the past). In the right-hand example, there is only one population where two lineages "exit". Thus, the gene tree on the left has deep coalescence cost 2, while the gene tree on the right has deep coalescence cost 1. The unrooted version of MDC minimizes coalescence costs over all possible rootings, so the deep coalescence cost for the unrooted gene tree ((B, C),A,D) is 1 for this species tree.

https://doi.org/10.1371/journal.pone.0251107.g001 method (MDC) was also developed to infer rooted species trees from unrooted gene trees [27]. The idea behind the method is to calculate the MDC score contributed by an unrooted gene tree for a candidate rooted species tree by minimizing the cost over all possible rootings of the gene trees.

Although MDC was one of the first methods to be implemented to infer species trees from gene trees [23], this criterion was found to be statistically inconsistent in the rooted setting (i.e., using rooted gene trees as input) in the same year that its unrooted extension was published [22, 27]. For some species trees, the probability that MDC returns an incorrect species tree tends to 1.0 as the number of input rooted gene trees goes to infinity. Although more accurate methods for inferring species trees have been developed, MDC is still sometimes used to quickly estimate a candidate species tree or phylogenetic network [28], which motivates studying its properties. Currently, there are no fast methods for inferring rooted species trees from unrooted gene trees. Consequently, a possible application of MDC in this setting is to generate candidate trees (particularly because MDC can be used to find sub-optimal trees) to reduce the search time needed for other more computationally intensive methods.

We also note that PhyloNet can use unrooted MDC, which we call UMDC, to return a rooted species tree even in the case of four taxa, although four-taxon gene tree topologies do not identify the rooted species tree under the multispecies coalescent [29]. A theoretical result from Allman et al. (2011) is that the true distribution of unrooted genetic tree topologies can be used to infer rooted species tree when there are five or more taxa, but not when there are only four taxa.

Part of the argument for the identifiability of the rooted species tree from unrooted gene trees is that there are certain inequalities that hold in the gene tree probabilities. For instance, consider distinguishing the two rooted species trees $((((a, b), c), d), e)$ versus $((((a, b), c), e), d)$. Both species trees have the same unrooted topology but imply different inequalities in some of the unrooted gene tree topology probabilities. For the first species tree, lineage c is more likely to coalesce with d than e; consequently, for the species tree, the unrooted gene tree $((a, b), e, (c, d))$ is more probable than the unrooted gene tree $((a, b), d, (c, e))$. However, for species tree $((((a, b), c), e), d)$, these inequalities are reversed. Therefore, observing more unrooted gene trees with topology ((a, b), e, (c, d)) than topology $((a, b), d, (c, e))$ gives evidence favoring the first species tree over the second species tree. Interestingly, for distinguishing the two species trees, the frequency of the matching unrooted gene tree $((a, b), c, (d, e))$ is not helpful. Instead, it is frequencies of nonmatching unrooted gene trees that are useful for inferring the rooted species tree [26]. This paper examines features of the distribution of unrooted topological gene trees that occur under the multispecies coalescent model on a species tree, for deriving asymptotic (i.e., large numbers of loci) UMDC behavior from unrooted gene trees for four taxa and five taxa.

Results

Let S be a binary, rooted species tree on a taxon leaf set X, and let λ be a list of branch lengths on S measured in coalescent time units. Here, $\lambda_i = 1$ means that branch i has a length of N_e generations, where N_e is the effective population size. Let $R(X)$ be the set of all rooted, binary trees for taxon set X and let $U(X)$ be the set of all unrooted, binary trees for the same taxon set. Let T denote a rooted gene tree, and S' a candidate species tree. Let $\alpha^*(T, S')$ denote the rooted deep coalescence cost (the minimum number of extra lineages) for a rooted gene tree T and candidate species tree S'. For an unrooted tree U with possible rootings $T^1, T^2, \ldots, T^k$, where $k = 2n - 3$ and n is the number of taxa, the unrooted coalescence cost is

$$\alpha(T_u, S') = \min_{i\in\{1,\ldots,2n-3\}} \alpha^*(T^i, S')$$

The number of extra lineages at a species boundary is the number of lineages greater than 1 passing from a population to its immediate ancestor. The total number of extra lineages for a gene tree-species tree pair is the sum of extra lineages over the entire species tree (Fig 1).

If $\mathcal{G}$ is an observed set of unrooted gene trees, we can think of UMDC as returning the inferred tree $\hat{S}$ that minimizes the average UMDC score:

$$\hat{S} = \arg\min_{S'} \frac{1}{|\mathcal{G}|} \sum_{U \in \mathcal{G}} \alpha(U, S')$$

For a given species tree S with branch lengths λ and candidate species tree topology S', we can define the expected UMDC cost as

$$\begin{aligned} E[\alpha_{S,\lambda}(T, S')] &= \sum_{U_i \in U(X)} \alpha(U_i, S') P(U_i | S, \lambda) \\ &= \sum_{T \in R(X)} \alpha(T, S') P(T | S, \lambda) \end{aligned}$$

Where we interpret $\alpha(T, S')$ for a rooted tree T as $\min_i \alpha^*(T^i, S')$, where T^i are the possible re-rootings of T. In other words, we interpret α applied to a rooted gene tree as minimizing over all possible rootings of the gene tree. The equivalence is due to the fact that we can compute the probability of an unrooted tree by summing over the probabilities of all possible rootings [29].

We note that this expected value depends on the branch lengths of the species tree S, but that branch lengths for S' do not need to be specified since only the topology of S' is used (and estimated). If the expected UMDC score is minimized by some $S' \neq S$, then UMDC is inconsistent since, by the Law of Large Numbers, as the number of loci tends to infinity, the UMDC score will be minimized by a tree other than the species tree with probability 1. To show inconsistency, it is sufficient to find a species trees S and S' and branch lengths λ such that $E[\alpha_{S,\lambda}(T, S')] < E[\alpha_{S,\lambda}(T, S)]$.

Trees with four leaves

Here are three unrooted, binary trees on four leaves. For the species tree, we can consider the two cases of symmetric or asymmetric binary species trees (Table 1). We indicate the i^{th} *distinctunrootedgenetreetopology*, i = 1, 2, 3, *as*T_i. The asymmetric species tree is also called a caterpillar, denoted S_C, and we denote the balanced tree as S_B, using one representative labeling for each case. Thus, the species tree is either $(S_C, \lambda) = (((a, b){:}x, c){:}y, d)$ or $(S_B, \lambda) = ((a, b){:}x, (c, d){:}y)$ where x and y are branch lengths in coalescent units. The lengths of the external branches are not used.

To see how this implies that UMDC is inconsistent, let the true species tree be $S_B = (((a, b){:} x, (c, d){:}y)$, then the expected coalescence cost under candidate tree $S_C = (((a, b), c), d)$ is (4/3) $\exp(-(x + y))$. Under candidate tree S_B, the expected cost is $(8/3)\exp(-(x + y))$. Thus, MDC will always give a lower cost to the tree S_C when S_B is the species tree. (Similarly, if S_C is the

Table 1. UMDC for 4-taxon unrooted gene trees.

Gene tree T_i	$P(U_i\|S_C, \lambda)$	$\alpha(U_i, S_C)$	$P(U_i\|S_B, \lambda)$	$\alpha(U_i, S_B)$
$U_1 = ((a, b), c, d)$	$1 - \frac{2}{3}e^{-x}$	0	$1 - \frac{2}{3}e^{-(x+y)}$	0
$U_2 = ((a, c), b, d)$	$\frac{2}{3}e^{-x}$	1	$\frac{2}{3}e^{-(x+y)}$	2
$U_3 = ((a, d), b, c)$	$\frac{2}{3}e^{-x}$	1	$\frac{2}{3}e^{-(x+y)}$	2

https://doi.org/10.1371/journal.pone.0251107.t001 species tree, UMDC will also give a lower cost to S_C than to S_B, regardless of the data.) Thus, MDC is incapable of returning a balanced tree for this scenario. This means that UMDC is inconsistent on four taxa. We also see that if the candidate species tree is $S' = (((a, b), d), c)$ (i.e., swapping taxa *c* and *d* in the species tree), then the deep coalescence costs are also 0, 1, and 1 for gene trees U_1, U_2, and U_3, respectively. Thus, regardless of the unrooted gene trees observed, UMDC will give equal scores to S_C and S', so that there is no way to choose one versus the other except for an arbitrary (or random) tie-break.

These results suggest that UMDC should not be used on unrooted four-taxon gene tree topologies. However, this is not unreasonable because it has been shown that under the MSC, four-taxon rooted species trees are not identifiable from unrooted gene trees. Thus, identifying the rooted species tree from four-taxon unrooted gene tree topologies would also not be possible using maximum likelihood, for example. However, rooted species trees are identifiable from unrooted five-taxon gene trees, which we examine next.

Trees with five leaves

The 15 binary, unrooted trees with five leaves have only one possible shape, whereas rooted trees on five leaves have three shapes, which we call caterpillar, pseudocaterpillar [30], and balanced. We indicate the i^{th} distinct unrooted gene tree topology, $i = 1, \ldots, 15$, as T_i. Although there are 15 possible gene tree topologies, there are 105 rooted species trees possible. An exhaustive approach to UMDC is to compute the UMDC score for all 105 candidate species trees and choose the species tree with the lowest score as the inferred tree.

There are three possible shapes to the rooted species tree when leaf-labels are ignored. The rooted species tree shape is called caterpillar, pseudocaterpillar, or balanced. We use S_C, S_P and S_B, respectively to denote representative trees from each shape:

$$\begin{aligned} S_C &= ((((a,b):x,c):y,d):z,e) \\ S_P &= (((a,b):x,(d,e):y):z,e) \\ S_B &= (((a,b):x,c):y),(d,e):z) \end{aligned}$$

Let $D_{ijk}(\lambda)$ denote the difference in expected UMDC scores for candidate trees S_j and S_k when the true species tree is S_i. Here $i, j, k \in \{C, P, B\}$ to denote caterpillar, pseudocaterpillar, and balanced topologies. Thus,

$$D_{ijk}(\lambda) = \sum_{h=1}^{15} P(U_h|S_i,\lambda)[\alpha(U_h,S_j) - \alpha(U_i,S_k)]$$

Generally, if $D_{ijk}(\lambda) > 0$, then candidate tree S_j has higher expected UMDC score than S_k when the species tree is S_i with branch lengths λ. In particular, if $j = i$, then $D_{iik} > 0$ means that UMDC will tend to rank the incorrect candidate tree S_k as better than the true tree S_i.

The differences in expected values can be obtained from Table 2. To save space, we omit notating the dependence on λ. For example

$$\begin{aligned} D_{CBP} &= 1 \cdot P(U_2|S_C) + 1 \cdot P(U_3|S_C) + 1 \cdot P(U_5|S_C) + 1 \cdot P(U_6|S_C) + 2 \cdot P(U_7|S_C) \\ &+2 \cdot P(U_8|S_C) + \cdots + 1 \cdot P(U_{15}|S_C). \end{aligned}$$

We note that the coefficients of $P(U_i|S_C)$ in D_{CBP} are all positive, indicating that the balanced tree has a higher cost than the pseudocaterpillar when the species tree is S_C. This holds

Table 2. MDC for 5-taxa unrooted gene trees.

Gene tree U_i	$\mathbf{Pr}(U_i\vert S_C, \lambda)$	$\alpha(U_i, S_C)$	$\mathbf{Pr}(U_i\vert S_P, \lambda)$	$\alpha(U_i, S_P)$	$\mathbf{Pr}(U_i\vert S_B, \lambda)$	$\alpha(U_i, S_B)$
U_1 = (((a, b), c), (d, e))	$1-\frac{2}{3}\mathrm{X}-\frac{2}{3}\mathrm{Y}+\frac{1}{3}\mathrm{XY}+\frac{1}{18}\mathrm{X}Y^3+\frac{1}{90}\mathrm{X}Y^3Z^6$	0	$1-\frac{2}{3}\mathrm{X}-\frac{2}{3}\mathrm{Y}+\frac{4}{9}\mathrm{XY}-\frac{2}{45}\mathrm{XY}Z^6$	0	$1-\frac{2}{3}\mathrm{X}-\frac{2}{3}\mathrm{YZ}+\frac{1}{3}\mathrm{YZ}-\frac{1}{3}\mathrm{XYZ}+\frac{1}{15}\mathrm{XY^3Z}$	0
U_2 = (((a, b), d), (c, e))	$\frac{1}{3}\mathrm{Y}\text{—}\frac{1}{6}\mathrm{XY}\text{—}\frac{1}{9}\mathrm{X}Y^3+\frac{1}{90}\mathrm{X}Y^3Z^6$	1	$\frac{1}{3}\mathrm{Y}\text{—}\frac{5}{18}\mathrm{XY}+\frac{1}{90}\mathrm{XY}Z^6$	1	$\frac{1}{3}\mathrm{YZ}\text{—}\frac{1}{6}\mathrm{XYZ}\text{—}\frac{1}{10}\mathrm{X}Y^3\mathrm{Z}$	2
U_3 = (((a, b), e),(c, d))	$\frac{1}{3}\mathrm{Y}\text{—}\frac{1}{6}\mathrm{XY}\text{—}\frac{1}{18}\mathrm{X}Y^3\text{—}\frac{2}{45}\mathrm{X}Y^3Z^6$	1	$\frac{1}{3}\mathrm{Y}\text{—}\frac{5}{18}\mathrm{XY}+\frac{1}{90}\mathrm{XY}Z^6$	1	$\frac{1}{3}\mathrm{YZ}\text{—}\frac{1}{6}\mathrm{XYZ}\text{—}\frac{1}{10}\mathrm{X}Y^3\mathrm{Z}$	2
U_4 = (((a, c), b), (d, e))	$\frac{1}{3}\mathrm{X}\text{—}\frac{1}{3}\mathrm{XY}+\frac{1}{18}\mathrm{X}Y^3+\frac{1}{90}\mathrm{X}Y^3Z^6$	1	$\frac{1}{3}\mathrm{X}\text{—}\frac{5}{18}\mathrm{XY}+\frac{1}{90}\mathrm{XY}Z^6$	1	$\frac{1}{3}\mathrm{X}\text{—}\frac{1}{3}\mathrm{XYZ}+\frac{1}{15}\mathrm{X}Y^3\mathrm{Z}$	1
U_5 = (((a, c), d), (b, e))	$\frac{1}{6}\mathrm{XY}\text{—}\frac{1}{9}\mathrm{X}Y^3+\frac{1}{90}\mathrm{X}Y^3Z^6$	2	$\frac{1}{18}\mathrm{XY}+\frac{1}{90}\mathrm{XY}Z^6$	2	$\frac{1}{6}\mathrm{XYZ}\text{—}\frac{1}{10}\mathrm{X}Y^3\mathrm{Z}$	3
U_6 = (((a, c), e),(b, d))	$\frac{1}{6}\mathrm{XY}\text{—}\frac{1}{18}\mathrm{X}Y^3\text{—}\frac{2}{45}\mathrm{X}Y^3Z^6$	2	$\frac{1}{18}\mathrm{XY}+\frac{1}{90}\mathrm{XY}Z^6$	2	$\frac{1}{6}\mathrm{XYZ}\text{—}\frac{1}{10}\mathrm{X}Y^3\mathrm{Z}$	3
U_7 = (((a, d), b), (c, e))	$\frac{1}{18}\mathrm{X}Y^3+\frac{1}{90}\mathrm{X}Y^3Z^6$	3	$\frac{1}{18}\mathrm{XY}+\frac{1}{90}\mathrm{XY}Z^6$	2	$\frac{1}{15}\mathrm{X}Y^3\mathrm{Z}$	4
U_8 = (((a, d), c), (b, e))	$\frac{1}{18}\mathrm{X}Y^3+\frac{1}{90}\mathrm{X}Y^3Z^6$	3	$\frac{1}{9}\mathrm{XY}\text{—}\frac{2}{45}\mathrm{XY}Z^6$	2	$\frac{1}{15}\mathrm{X}Y^3\mathrm{Z}$	4
U_9 = (((a, d), e), (b, c))	$\frac{1}{6}\mathrm{XY}\text{—}\frac{1}{18}\mathrm{X}Y^3\text{—}\frac{2}{45}\mathrm{X}Y^3Z^6$	3	$\frac{1}{18}\mathrm{XY}+\frac{1}{90}\mathrm{XY}Z^6$	2	$\frac{1}{6}\mathrm{XYZ}\text{—}\frac{1}{10}\mathrm{X}Y^3\mathrm{Z}$	3
U_{10} = (((a, e), b), (c, d))	$\frac{1}{18}\mathrm{X}Y^3+\frac{1}{90}\mathrm{X}Y^3Z^6$	3	$\frac{1}{18}\mathrm{XY}+\frac{1}{90}\mathrm{XY}Z^6$	2	$\frac{1}{15}\mathrm{X}Y^3\mathrm{Z}$	4
U_{11} = (((a, e), c), (b, d))	$\frac{1}{18}\mathrm{X}Y^3+\frac{1}{90}\mathrm{X}Y^3Z^6$	3	$\frac{1}{9}\mathrm{XY}\text{—}\frac{2}{45}\mathrm{XY}Z^6$	2	$\frac{1}{15}\mathrm{X}Y^3\mathrm{Z}$	4
U_{12} = (((a, e), d), (b, c))	$\frac{1}{6}\mathrm{XY}\text{—}\frac{1}{9}\mathrm{X}Y^3+\frac{1}{90}\mathrm{X}Y^3Z^6$	2	$\frac{1}{18}\mathrm{XY}+\frac{1}{90}\mathrm{XY}Z^6$	2	$\frac{1}{6}\mathrm{XYZ}\text{—}\frac{1}{10}\mathrm{X}Y^3\mathrm{Z}$	3
U_{13} = (((b, c), a), (d, e))	$\frac{1}{3}\mathrm{X}\text{—}\frac{1}{3}\mathrm{XY}+\frac{1}{18}\mathrm{X}Y^3+\frac{1}{90}\mathrm{X}Y^3Z^6$	1	$\frac{1}{3}\mathrm{X}\text{—}\frac{5}{18}\mathrm{XY}+\frac{1}{90}\mathrm{XY}Z^6$	1	$\frac{1}{3}\mathrm{X}\text{—}\frac{1}{3}\mathrm{XYZ}+\frac{1}{15}\mathrm{X}Y^3\mathrm{Z}$	2
U_{14} = (((b, d), a), (c, e))	$\frac{1}{18}\mathrm{X}Y^3+\frac{1}{90}\mathrm{X}Y^3Z^6$	3	$\frac{1}{18}\mathrm{XY}+\frac{1}{90}\mathrm{XY}Z^6$	2	$\frac{1}{15}\mathrm{X}Y^3\mathrm{Z}$	4
U_{15} = (((b, e), a), (c, d))	$\frac{1}{18}\mathrm{X}Y^3+\frac{1}{90}\mathrm{X}Y^3Z^6$	3	$\frac{1}{18}\mathrm{XY}+\frac{1}{90}\mathrm{XY}Z^6$	2	$\frac{1}{15}\mathrm{X}Y^3\mathrm{Z}$	4

https://doi.org/10.1371/journal.pone.0251107.t002

regardless of the choice of branch lengths λ. Similarly, we see that for any branch lengths λ,

$$D_{CBP} > 0,\ D_{CCP} > 0,\ D_{BBP} > 0,\ D_{BCB} > 0,\ D_{PBC} > 0,$$

$$D_{PPB} < 0,\ D_{PPC} < 0$$

The theoretical expected values show that, for example, if the species tree is a caterpillar, then at least one pseudocaterpillar has lower expected deep coalescence cost than the matching caterpillar species tree, and consequently UMDC is not consistent for recovering the true species tree. Similarly, if the species tree is balanced, then at least one pseudocaterpillar tree always has lower expected coalescence cost the true species tree. In both cases, UMDC will be misleading, tending to return an incorrect species tree as more loci are examined. We note that these relationships hold regardless of the branch lengths of the species tree. Remarkably, these inequalities hold not only asymptotically as the UMDC score approaches its expected value, but even for finite numbers of loci (but using non-strict inequalities). For example, let the number of 5-taxon trees in a sample be $(n_1, n_2, \ldots, n_{15})$ where the subscript indexes the unrooted topologies from Table 2. If the species tree is S_C, then the UMDC score for S'_C minus that for S'_P is

$$S'_C - S'_P = n_7 + n_8 + n_9 + n_{10} + n_{11} + n_{14} + n_{15} \geq 0$$

Since this is always greater than or equal to 0, the matching tree can never have better deep coalescence cost than S'_P (if none of these 7 topologies are observed, the UMDC scores will be tied for these two candidate species trees). If the species tree is S_B, then the situation is even worse, with

$$S'_B - S'_P = n_2 + n_3 + n_5 + n_6 + 2n_7 + 2n_8 + n_9 + 2n_{10} + 2n_{11} + n_{12} + n_{13} + 2n_{14} + 2n_{15} \geq 0.$$

This shows that MDC is misleading for five taxa if the species tree does not have pseudocaterpillar topology.

These examples are sufficient to show that UMDC is inconsistent for trees with five taxa. However, to better understand the behavior of UMDC, it is helpful to also understand the MDC costs for other true species tree and candidate species tree combinations (S1 and S2 Tables in S1 Appendix). An interesting question here is whether UMDC will tend to perform well within a particular unlabeled shape. For example, if it is known (or believed) that the species tree has a particular shape (for example, a caterpillar), will UMDC pick the correct species tree if it restricted to the correct unlabeled shape? This situation can arise in particular if alternative rootings lead to two candidate trees that have the same shape. In this case, we can examine in which cases UMDC might or might not be misleading. A potential use here is that UMDC is used to generate candidate species trees; it could return the best species tree for each tree shape, and more computationally intensive methods could then use these as starting trees.

From S1 and S2 Tables in S1 Appendix, we see that UMDC cannot distinguish two caterpillar trees with the outgroup swapped with the taxon that is an outgroup to all other ingroup taxa, for example species trees $(((a, b), c), d), e)$ and $((((a, b), c), e), d)$ will have exactly the same UMDC score for any data set. To check if, for example, these are expected to be the best scoring caterpillar trees when the species tree is S_C, we can use expected values. For example, to compare the expected score for $S'_1 = ((((a,b),c),d),e)$ with $S'_3 = ((((a,b),d),c),e)$, we note that the MDC cost is the same for these candidate species trees for unrooted gene trees U_3, U_6, U_9, U_{10}, and U_{15}. The difference in expected values therefore depends on the other 10 trees. Collecting terms, the difference is

$$\begin{aligned} f(X,Y,Z) &= E[\alpha(U,S'_3) - \alpha(U,S'_1)] \\ &= -1 - X(2/3 - 2/3 + 2/3) + Y(2/3 + 1/3) + XY(-1/3 - 1/6 + \cdots + 2/3) \\ &\quad + XY^3(1/18 + \cdots + 2/18) + XY^3Z^6(1/90 + \cdots + 2/90) \\ &= -1 - (2/3)X + Y + XY/6 + 4XY^3/18 - 5XY^3Z^6/90 \end{aligned}$$

Because $f(X, Y, Z)$ is decreasing in Z, and therefore $f(X, Y, Z) < f(X, Y, 0)$ for all Z, a sufficient condition to show that that S_C has lower expected score than S'_3 is that

$$-1 - (2/3)X + Y + XY/6 + 4XY^3/18 < 0$$

Note that the expression is equivalent to

$$-1 - X(2/3 + Y/6 + 4Y^3/18) + Y$$

Because $-1 + Y < 0$ and the term in parenthesis is positive, this shows that the difference in expected UMDC costs is negative; hence S_C is expected to be preferred over S'_3 when S_C is the species tree.

Similar arguments can be made, although tediously, to show that S_C has the lowest expected UMDC score (although tied with S_2') among all caterpillar candidate species trees. Because of the ties in the UMDC costs for pairs of candidate caterpillar species trees, this requires evaluating 29 such inequalities.

Simulation

To illustrate the theoretical results, we simulated gene trees from the species trees S_C, S_P, and S_B using hybrid-Lambda [31], where all internal branch lengths were 1.0 coalescent units, which allows a moderate amount of incomplete lineage sorting. The species trees were

$$\begin{aligned} S_C &= ((((a:1.0, b:1.0):1.0, c:2.0):1.0, d:3.0):1.0, e:4.0) \\ S_P &= (((a:1.0, b:1.0):1.0, (c:1.0, d:1.0):1.0):1.0, e:4.0) \\ S_B &= (((a:1.0, b:1.0):1.0, c:2.0):1.0, (d:2.0, e:2.0):1.0) \end{aligned}$$

These species trees were also repeated with each branch length multiplied by 0.1 to investigate the effect of shorter branch lengths. For each species tree, independent simulations were run with 50, 100, 200, 400, and 800 loci. For each combination of species tree and number of loci, a set of gene trees was simulated using hybrid-Lambda [31]. Species trees were estimated directly from these known gene trees, and a second set of simulations was done using estimated gene trees. To estimate gene trees, DNA sequences were simulated from the gene trees using seq-gen version 1.3.2x [32] with 500 nucleotides per locus and base frequencies of 0.3, 0.2, 0.2, and 0.3 for nucleotides A, C, G, and T, respectively with a mutation rate of $\theta = 0.01$ under a $GTR + \Gamma + I$ model with four variable rates and 10% invariable sites. Gene trees were then estimated as unrooted using phyml version 20120412 [33] under the correct model. Each of these settings was repeated 100 times, and the proportion of times various tree topologies were inferred was recorded. The UMDC tree was obtained from phylonet using the command `Infer_ST_MDC_UR`. In case of a tie between highest scoring species trees, a tree was picked uniformly at random as the species tree estimate.

As predicted by the theory, regardless of the species tree, the pseudocaterpillar shape tends to be the tree inferred, with probability approaching 1.0 as the sample size (number of loci) increases (Fig 2). In cases where a non-pseudocaterpillar was inferred, this was due to randomly picking one of the trees tied for best UMDC score. Ties for the best MDC score were less likely with larger sample sizes, with 70% of cases (out of 500) with 50 loci having three trees tied for best when the species tree was S_C, and 1.2% of cases being tied with 800 loci for S_C. Results were similar for S_B, but there were much fewer ties for best tree for the pseudocaterpillar species tree, S_P. For S_P, 7% of cases had a tree tied for best with 50 loci, and the best tree was always unique with 100 or more loci.

We investigated the effects of using shorter branch lengths (multiplying each branch length by 0.1) and of estimating gene trees from DNA sequences to investigate the effects of both greater gene tree heterogeneity due to shorter branches and gene tree estimation error. For the caterpillar species tree, there was very little effect of either gene tree estimation error or shortening species tree branches (Fig 2). For all species trees, estimation error had little effect on the inference. When branches were multiplied by 0.1, convergence to the incorrect species tree was more rapid for the balanced tree, while convergence to the correct tree was slightly slower when the species tree was a pseudocaterpillar.

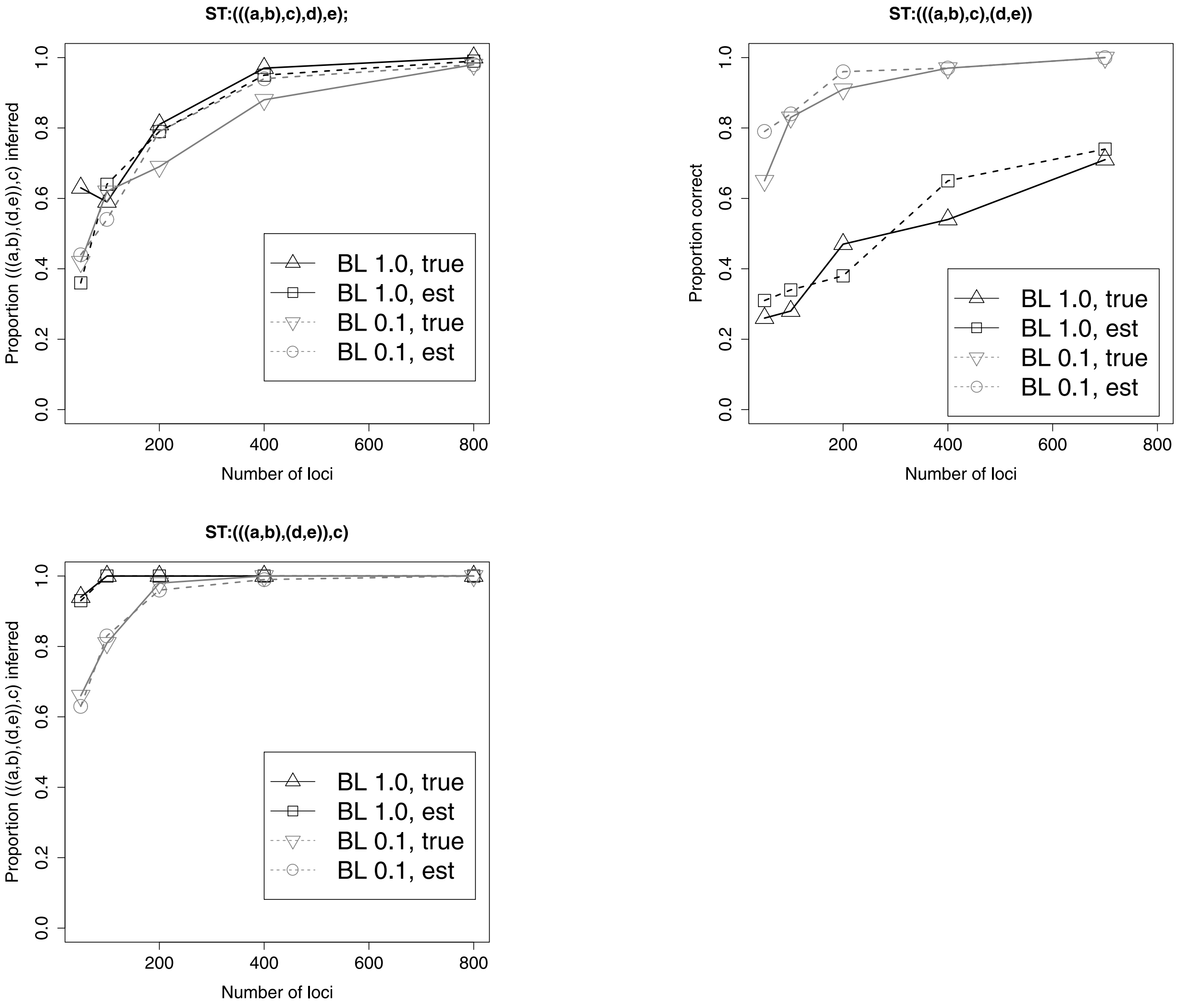

Fig 2. Simulation for species trees S_C, S_B, and S_P. The proportion indicates the number of times out of 100 that the species tree topology (((*a*, *b*), *c*), (*d*, *e*)) was inferred. "BL 1.0" means that internal branches, as well as pendant branches leading to taxa *a* and *b* had length 1.0, while remaining pendant edges had lengths needed to make the trees ultrametric. "BL 0.1" had all branch lengths multiplied by 0.1 in the species tree. The terms "true" and "est" refer to whether gene trees were estimated from simulated DNA sequences or the actual simulated gene trees were used.

https://doi.org/10.1371/journal.pone.0251107.g002

Discussion

In this paper, we used a method of considering expected values of scores to show properties of the unrooted version of MDC. This approach of using expected values has also been used to show inconsistency of the original MDC criterion [22] and matrix representation with parsimony [21]. Although we only showed inconsistency for four- and five-taxon trees, the results apply straightforwardly to larger trees. For example, "caterpillarization" is a technique of making some branches long enough while keeping others short that the distribution of gene trees resembles the the distribution found from a caterpillar species tree, and all the results would apply to these larger trees as well lemmas 3 and 5 in [34]. This means that for larger species trees, there exist branch lengths for which UMDC will be misleading.

To explain this idea in more detail, suppose a six-taxon species tree has topology ((((*a*, *b*), *c*), *d*), (*e*, *f*)). If the branch leading from the most recent common ancestor (MRCA) of *e* and *f* to the root is long, then lineages sampled from *e* and *f* will almost certainly coalesce more recently than the root of the tree. Consequently, almost all gene trees from this species tree will have (*e*, *f*) as a cluster, and the species tree will have very similar properties as a species tree ((((*a*, *b*), *c*), *d*), *x*) with *x* replaced by (*e*, *f*). Consequently, if this branch is sufficiently long, unrooted gene tree distributions on taxa *a*–*f* are concentrated on just the 15 unrooted topologies that occur on five-taxon trees, and we can predict what species tree inference methods will do based on their behavior on 5-taxon trees. Another example is the 6-taxon pseudocaterpillar species tree (((*a*, *b*), (*c*, *d*), *e*), *f*). This tree can be caterpillarized by letting the branch leading from the MRCA of *c* and *d* to the MRCA of *a*, *b*, *c*, and *d* be sufficiently long. Replacing (*c*, *d*) with *x* in this tree, it resembles a five-taxon caterpillar ((((*a*, *b*), *x*), *e*), *f*) for which we can expect UMDC to prefer tree (((*a*, *b*), (*e*, *f*)), *x*) = ((*a*, *b*), ((*c*, *d*), (*e*, *f*)). This approach is sufficient to show that any tree that can be caterpillarized to a five-taxon caterpillar tree can be misleading in the ways shown in this paper given certain branch lengths. A more detailed proof for this particular six-taxon tree would show that of the 105 6-taxon topologies, for any given $\epsilon > 0$, branch lengths in the species tree can be chosen such that the probability is greater than $1 - \epsilon$ that the probability that the gene tree is one of the 15 trees concentrated on taxa *a*, *b*, *x*, *e*, and *f*.

Similarly, many species trees with more than five taxa can mimic the behavior of the 5-taxon balanced tree given certain branch lengths. For example, the species tree (((*a*, *b*), *c*), (*d*, (*e*, *f*))) will mimic the 5-taxon balanced tree in this paper if the branch leading from the MRCA of *e* and *f* to the MRCA of *d*, *e*, and *f* is sufficiently long. A general proof of inconsistency for UMDC for trees with 6–8 taxa would examine some special cases like these and show that there are branch lengths which could make a tree mimic the 5-taxon caterpillar or balanced trees. Trees with 9 or more taxa can always be caterpillarized to a 5-taxon caterpillar [21, 34]

Although the results are negative, MDC and UMDC can still be used to quickly generate starting trees when searching for species trees using other methods. An advantage of MDC and UMDC here is that they can rank trees by score, and therefore return suboptimal trees. Since MDC has a shape bias, tending to make it return more balanced trees [35], it is not surprising that there is a shape bias for UMDC as well. The bias for UMDC is surprisingly extreme however in that UMDC always prefers certain shapes (the pseudocaterpillar for five taxa) for any data. The results of this paper suggest that due to the shape bias of UMDC, a preferable method to generating starting trees might be to return the set of optimal trees within each unlabeled shape. The impact that this could have on species tree inference we leave to future work.

Supporting information

S1 Appendix.
(PDF)

Acknowledgments

We thank the anonymous reviewers for helpful comments.

Author Contributions

Conceptualization: Ayed A. R. Alanzi, James H. Degnan.

Formal analysis: Ayed A. R. Alanzi, James H. Degnan.

Investigation: Ayed A. R. Alanzi.

Methodology: Ayed A. R. Alanzi, James H. Degnan.

Supervision: James H. Degnan.

Writing – original draft: Ayed A. R. Alanzi, James H. Degnan.

Writing – review & editing: Ayed A. R. Alanzi, James H. Degnan.

References

1. Maddison WP. Gene trees in species trees. Syst Biol. 1997; 46(3):523–536. https://doi.org/10.1093/sysbio/46.3.523
2. Degnan JH, Rosenberg NA. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol. 2009; 24(6):332–340. https://doi.org/10.1016/j.tree.2009.01.009 PMID: 19307040
3. Pamilo P, Nei M. Relationship between gene trees and species trees. Mol Biol Evol. 1998; 5:568–583.
4. Rosenberg NA. The probability of topological concordance of gene trees and species trees. Theor Pop Biol. 2002; 61:225–247. https://doi.org/10.1006/tpbi.2001.1568 PMID: 11969392
5. Degnan JH, Salter LA. Gene tree distributions under the coalescent process. Evolution. 2005; 59:24–37. https://doi.org/10.1554/04-385 PMID: 15792224
6. Slatkin M, Pollack JL. Subdivision in an ancestral species creates asymmetry in gene trees. Mol Biol Evol. 2008; 25(10):2241–2246. https://doi.org/10.1093/molbev/msn172 PMID: 18689871
7. DeGiorgio M, Rosenberg NA. Consistency and inconsistency of consensus methods for inferring species trees from gene trees in the presence of ancestral population structure. Theor Popul Biol. 2016; 110:12–24. https://doi.org/10.1016/j.tpb.2016.02.002 PMID: 27086043
8. Suh A, Smeds L, Ellegren H. The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. PLoS Biol. 2015; 13:e1002224. https://doi.org/10.1371/journal.pbio.1002224 PMID: 26284513
9. Pease JB, Haak DC, Hahn MW, Moyle LC. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. PLoS Biol. 2016; 14(2):e1002379. https://doi.org/10.1371/journal.pbio.1002379 PMID: 26871574
10. Edwards SV. Is a new and general theory of molecular systematic biology emerging? Evolution. 2009; 63(1):1–19. https://doi.org/10.1111/j.1558-5646.2008.00549.x PMID: 19146594
11. Liu L, Yu LL, Kubatko L, et al. Coalescent methods for estimating phylogenetic trees. Mol Phylogenet Evol. 2009; 53:320–328. https://doi.org/10.1016/j.ympev.2009.05.033 PMID: 19501178
12. Zimmermann T, Mirarab S, Warnow T. BBCA: Improving the scalability of *BEAST using random binning. BMC Genomics. 2014; 15(6):S11. https://doi.org/10.1186/1471-2164-15-S6-S11 PMID: 25572469
13. Ewing GB, Ebersberger I, Schmidt HA, Von Haeseler A. Rooted triple consensus and anomalous gene trees. BMC Evolut Biol. 2008; 8(1):118. https://doi.org/10.1186/1471-2148-8-118 PMID: 18439266
14. Mirarab S, Reaz R, Bayzid MS, Zimmerman T, Swenson MS, Warnow T. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics. 2014; 30(17):i541–i548. https://doi.org/10.1093/bioinformatics/btu462 PMID: 25161245
15. Liu L, Yu L, Edwards SV. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evolut Biol. 2010; 10(1):302. https://doi.org/10.1186/1471-2148-10-302 PMID: 20937096
16. Liu L, Yu L. Estimating species trees from unrooted gene trees. Syst Biol. 2011; 60(5):661–667. https://doi.org/10.1093/sysbio/syr027 PMID: 21447481
17. Allman ES, Degnan JH, Rhodes JA. Species tree inference from gene splits by unrooted STAR methods. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). 2018; 15(1):337–342. https://doi.org/10.1109/TCBB.2016.2604812 PMID: 28113601
18. Liu L, Yu L, Pearl DK, Edwards SV. Estimating species phylogenies using coalescence times among sequences. Syst Biol. 2009; 58:468–477. https://doi.org/10.1093/sysbio/syp031 PMID: 20525601
19. Degnan JH, Rosenberg NA. Discordance of species trees with their most likely gene trees. PLoS Genet. 2006; 2:762–768. https://doi.org/10.1371/journal.pgen.0020068 PMID: 16733550

20. Degnan JH, DeGiorgio M, Bryant D, Rosenberg NA. Properties of consensus methods for inferring species trees from gene trees. Syst Biol. 2009; 58:35–54. https://doi.org/10.1093/sysbio/syp008 PMID: 20525567

21. Wang Y, Degnan JH. Performance of matrix representation with parsimony for inferring species from gene trees. Stat Appl Genet Mol Biol. 2011; 10:1. https://doi.org/10.2202/1544-6115.1611

22. Than CV, Rosenberg NA. Consistency properties of species tree inference by minimizing deep coalescences. J Comput Biol. 2011; 18:1–15. https://doi.org/10.1089/cmb.2010.0102 PMID: 21210728

23. Maddison WP, Knowles LL. Inferring phylogeny despite incomplete lineage sorting. Syst Biol. 2006; 55 (1):21–30. https://doi.org/10.1080/10635150500354928 PMID: 16507521

24. Bansal MS, Burleigh JG, Eulenstein O. Efficient genome-scale phylogenetic analysis under the duplication-loss and deep coalescence cost models. BMC Bioinformatics. 2010; 11(1):S42. https://doi.org/10.1186/1471-2105-11-S1-S42 PMID: 20122216

25. Than C, Nakhleh L. Species tree inference by minimizing deep coalescences. PLoS Comput Biol. 2009; 5(9):e1000501. https://doi.org/10.1371/journal.pcbi.1000501 PMID: 19749978

26. Alanzi AR, Degnan JH. Inferring rooted species trees from unrooted gene trees using approximate Bayesian computation. Mol Phylogenet Evol. 2017; 116:13–24. https://doi.org/10.1016/j.ympev.2017.07.017 PMID: 28780022

27. Yu Y, Warnow T, Nakhleh L. Algorithms for MDC-based multi-locus phylogeny inference: beyond rooted binary gene trees on single alleles. J Comput Biol. 2011; 18(11):1543–1559. https://doi.org/10.1089/cmb.2011.0174 PMID: 22035329

28. Than C, Ruths D, Nakhleh L. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinformatics. 2008; 9(1):322. https://doi.org/10.1186/1471-2105-9-322 PMID: 18662388

29. Allman ES, Degnan JH, Rhodes JA. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. J Math Biol. 2011; 62:833–862. https://doi.org/10.1007/s00285-010-0355-7 PMID: 20652704

30. Rosenberg NA. Counting coalescent histories. J Comput Biol. 2007; 14:360–377. https://doi.org/10.1089/cmb.2006.0109 PMID: 17563317

31. Zhu S, Degnan JH, Goldstien SJ, Eldon B. Hybrid-Lambda: simulation of multiple merger and Kingman gene genealogies in species networks and species trees. BMC Bioinformatics. 2015; 16(1):292. https://doi.org/10.1186/s12859-015-0721-y PMID: 26373308

32. Rambaut A, Grassly NC. Seq-gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comp Appl Biosci. 1997; 13:235–238. PMID: 9183526

33. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010; 59(3):307–321. https://doi.org/10.1093/sysbio/syq010 PMID: 20525638

34. Degnan JH. Anomalous unrooted gene trees. Syst Biol. 2013; 62:574–590. https://doi.org/10.1093/sysbio/syt023 PMID: 23576318

35. DeGiorgio M, Syring J, Eckert AJ, et al. An empirical evaluation of two-stage species tree inference strategies using a multilocus dataset from North American pines. BMC Evolut Biol. 2014; 14(1):67. https://doi.org/10.1186/1471-2148-14-67 PMID: 24678701