Phylogenetics

Quartet-Based Inference is Statistically Consistent Under the Unified Duplication-Loss-Coalescence Model

Alexey Markin [1,*] and Oliver Eulenstein [2]

[1] Virus and Prion Research Unit, National Animal Disease Center, USDA-ARS, Ames, IA, 50010, USA
[2] Department of Computer Science, Iowa State University, Ames, IA, 50011, USA.

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: The classic multispecies coalescent (MSC) model provides the means for theoretical justification of incomplete lineage sorting-aware species tree inference methods. This has motivated an extensive body of work on phylogenetic methods that are statistically consistent under MSC. One such particularly popular method is ASTRAL, a quartet-based species tree inference method. Novel studies suggest that ASTRAL also performs well when given multi-locus gene trees in simulation studies. Further, Legried et al. recently demonstrated that ASTRAL is statistically consistent under the gene duplication and loss model (GDL). GDL is prevalent in evolutionary histories and is the first core process in the powerful duplication-loss-coalescence evolutionary model (DLCoal) by Rasmussen and Kellis.
Results: In this work we prove that ASTRAL is statistically consistent under the general DLCoal model. Therefore, our result supports the empirical evidence from the simulation-based studies. More broadly, we prove that the quartet-based inference approach is statistically consistent under DLCoal.
Contact: Alexey Markin (alexey.markin@usda.gov)

1 Introduction

The accurate inference of evolutionary histories of species is a grand challenge in evolutionary biology due to the fact that the *true* evolutionary histories are rarely known (Bininda-Emonds, 2004). Consequently, the common strategy in the phylogenetic community is to rely on established statistical models of evolution when evaluating phylogenetic inference methods. One of the most prominent such models is the multispecies coalescent model (Rannala and Yang, 2003) that accounts for incomplete lineage sorting (ILS), also known as deep coalescence. ILS is a prevalent factor that causes discordance between the observed gene tree topologies and the host species tree (Allman *et al.*, 2018). In fact, a large body of work in phylogenetics is dedicated to the design of species tree inference methods that are statistically consistent under MSC. Statistical consistency implies that as the number of observed gene trees grows, the species tree estimate converges to the true species tree that "generated" the observed data. Multiple phylogenetic inference methods have been demonstrated to be statistically consistent, cf. GLASS (Mossel and Roch, 2008), R* (Degnan *et al.*, 2009), STEM (Kubatko *et al.*, 2009), MP-EST (Liu *et al.*, 2010), BUCKy (Larget *et al.*, 2010), STAR/USTAR (Liu *et al.*, 2009; Allman *et al.*, 2016), NJst (Liu and Yu, 2011), ASTRID (Vachaspati and Warnow, 2015), ASTRAL (Zhang *et al.*, 2018), other rooted triplet and unrooted quartet methods (Ewing *et al.*, 2008; Rhodes, 2019; Yourdkhani and Rhodes, 2020), and others.

In recent years ASTRAL became one of the most popular species tree inference methods by practitioners. Note that ASTRAL's objective function is built on the notion of *quartets* (see Figure 1). In particular, the proof that ASTRAL is statistically consistent under MSC stems from two observations. First, Allman et al. (Allman *et al.*, 2011) demonstrated that if a species tree displays a quartet q then q is also the most likely observed (unrooted) gene tree topology. Second, it can be seen that every species tree clade will eventually appear in at least one of the observed gene trees.

More recently, Legried et al. (Legried *et al.*, 2020) studied two natural extensions of ASTRAL that enable processing the multi-locus gene trees. Multi-locus gene trees can have multiple leaves with the same species label (that is, the respective species has multiple copies of the same gene). These extensions allow one to apply ASTRAL to a much broader class of

phylogenetic gene trees and are referred to as *ASTRAL-one* and *ASTRAL-multi*. Given four species (e.g., $\{A, B, C, D\}$) a multi-locus gene tree can have multiple copies of each of the species and therefore can suggest multiple (conflicting) quartets on $\{A, B, C, D\}$. In that case, ASTRAL-one chooses a *single random* copy for each species label and considers the respective quartet type, whereas ASTRAL-multi considers *all* gene copies and all the respective quartets.

Focusing on these two extensions of ASTRAL, Legried et al. proved that both ASTRAL-one and ASTRAL-multi are statistically consistent under the gene duplication and loss model (GDL) (Legried *et al.*, 2020). Note that GDL is a part of the broader and well-recognized unified duplication-loss-coalescence (DLCoal) model of gene tree evolution by Rasmussen and Kellis (Rasmussen and Kellis, 2012). DLCoal simultaneously accounts for three crucial types of evolutionary factors that shape gene family evolution. Namely, duplications, losses, and incomplete lineage sorting. The DLCoal process involves two steps, (i) a birth/death process within the branches of the species tree creates a *locus tree* (i.e., the GDL process), and (ii) a bounded multispecies coalescence process acting on the locus tree generates the observed *gene tree*. See Figure 2 for an example.

In this work, for the first time, we prove that ASTRAL-one is statistically consistent under the general DLCoal model. First, we derive gene tree probabilities (constrained to quartets) under the bounded multispecies coalescent model and draw core observations from that analysis. Second, we build on an idea from Legried et al. to systematically separate different duplication-loss scenarios. Then, for each such scenario, we prove that a random quartet from the gene tree is more likely to agree with the species tree quartet rather than any of the two other quartets. Finally, we extend our result for ASTRAL-one to ASTRAL-multi and demonstrate that ASTRAL-multi is also consistent under DLCoal [1].

Our results provide a theoretical justification to the findings in (Du *et al.*, 2019), which showcased the accuracy of ASTRAL-one in the presence of duplications, losses, and incomplete lineage sorting.

2 Preliminaries

We denote a rooted *(phylogenetic) tree* by $P = (T, \omega)$. Here T is the *tree topology* and is a binary rooted tree with the designated root vertex, $\rho(T)$, of degree two, all internal nodes of degree three, and with leaves bijectively labeled by elements of set $\mathsf{Le}(T)$. For convenience, we identify leaves with their labels. Further, tree topologies are *planted*, implying that an additional *root edge* is attached to the root vertex. Then, ω specifies the lengths of edges in T in *coalescent units* (i.e., the number of generations normalized by the effective population size (Allman *et al.*, 2011)). More formally, $\omega\colon E(T) \to \mathbb{Q}^+$. In particular, we assume that all edge lengths are strictly positive. When the phylogenetic tree P is not clear from the context, we will often use the notation T_P and w_P to refer to its tree topology and its edge-length function, respectively.

An *unrooted (phylogenetic) tree topology* T is similar to the rooted tree topology, but without a designated root and the root edge. That is, in unrooted tree T all non-leaf vertices have degree 3.

We say that an edge e is *external* if it is incident with a leaf vertex, and otherwise we call e *internal*. Further, given a set $Y \subset \mathsf{Le}(T)$, tree topology $T|_Y$ is obtained from T by restricting the leaf-set to Y. A restricted phylogenetic tree $P|_Y = (T|_y, w|_Y)$ is then obtained by choosing the function $w|_Y$ that maintains the same leaf-to-root path lengths as in P (in respect to the leaves in Y).

[1] Our extension of the consistency result to ASTRAL-multi was developed independently from Hill *et al.*, 2020.

A rooted topology T defines a partial order on its nodes: given two nodes x and y we say $x \preceq y$ if x is a descendant of y (and $x \prec y$ if additionally $x \neq y$). We say that two edges in a rooted tree are *parallel* if neither edge is located on the path from the other edge to the root.

Quartets. A quartet is an unrooted tree topology with exactly four leaves. Assuming that the leaves are a, b, c, and d, we denote the quartets in Figure 1(left), 1(middle) and 1(right) as $ab|cd$, $ac|bd$, and $ad|bc$ respectively (based on the two cherries separated by the internal edge).

We say that a quartet q is *displayed* in a phylogenetic tree P, if the unrooted tree topology of P restricted to the leaves in q (i.e., $T_P|_{\mathsf{Le}(q)}$) is equivalent to q. In this case, we write $q \in P$.

2.1 Unified DLCoal model

We now overview the unified duplication-loss-coalescence (DLCoal) model (Rasmussen and Kellis, 2012).

Species tree. A *species tree* $S = (T_S, \omega_S)$ represents an evolutionary history of species. Leaves of T_S are labeled by the extant species names.

Locus tree. A *locus tree* $L = (T_L, \omega_L)$ represents a duplication/loss history of a fixed gene. A locus tree is obtained from a species tree by running the duplication/loss process (Rasmussen and Kellis, 2012; Legried *et al.*, 2020) top-down along the edges of the species tree. More specifically, the duplication/loss process is a birth-death process with a fixed birth (duplication) rate λ and death (loss) rate μ (Arvestad *et al.*, 2003). The birth-death process starts in the root edge of the species tree; whenever it reaches a speciation point, the process splits into two copies and continues independently in the children edges. See Figure 2 for an example. Note that locus tree leaves are labeled by gene names.

A locus tree node is always one of the following two types:

(i) *Speciation*. Such node corresponds to a speciation event/node from the species tree.
(ii) *Duplication*. Such node corresponds to a new locus creation event.

Remark. *A duplication event is asymmetric, as one child (the* mother duplicate*) follows the parent locus, and the other child (the* daughter duplicate*) corresponds to a novel locus (Rasmussen and Kellis, 2012). To account for that, we will often depict duplications as red dots on the locus tree edges immediately below the duplication nodes, shifted towards a daughter duplicate. That is, a red dot on an edge will indicate that this point is a start of a new locus (see Figure 3 for an example). This will ensure a consistent depiction of duplications for Section 3. Further, we will refer to these points as* duplication-points.

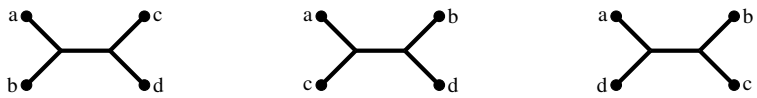

Fig. 1: All three possible quartets on a, b, c, d leaves.

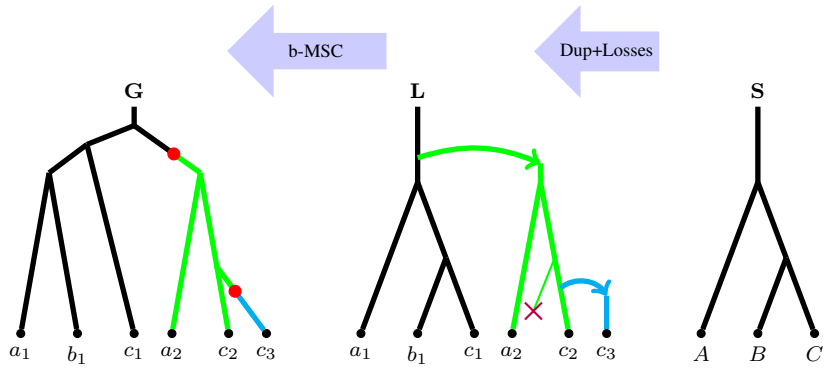

Fig. 2: An example of a gene tree G, locus tree L, and species tree S. Note that the arrows in the locus tree represent the duplication events, and the cross represents a loss event. Further, the red circles on the gene tree represent the duplication-points. As coalescent (b-MSC) runs on the locus tree, the coalescence of the new and the original loci is likely to happen above a duplication event; therefore, the duplication-points can appear in the middle of gene tree edges, as shown in the figure.

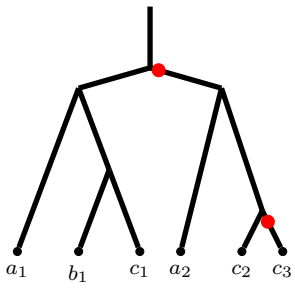

Fig. 3: An alternative depiction of a locus tree from Figure 2 with red dots representing duplication-points *slightly* shifted towards a novel locus.

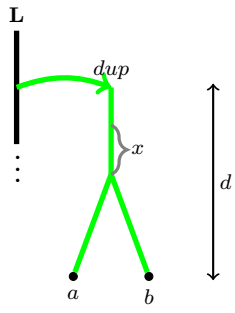

Fig. 4: An example of a locus tree that illustrates the b-MSC constraints for Section 2.3.

Gene tree. A *gene tree* $G = (T_G, \omega_G)$ represents a gene family's evolutionary history. The gene tree is obtained from a locus tree by running the *bounded multispecies coalescent (b-MSC)* process bottom-up along the edges of the locus tree (Rasmussen and Kellis, 2012) (see Section 2.3 for a more detailed description of that process). Figure 2 provides an example of that process.

2.2 Multispecies coalescent (MSC) model

In the standard multispecies coalescent model (Rannala and Yang, 2003) gene lineages are followed backwards in time (from the leaves to the root).

For simplicity, we assume that there is exactly one gene lineage starting in every extant locus tree leaf. If two or more lineages enter the same locus tree edge, then the coalescence history of these lineages is determined by an exponential distribution.

In particular, for any two lineages a, b that entered the same edge the probability that they coalesce within time x (specified in terms of coalescent units) is as follows:

$$P[a, b \text{ coalesced within time } x] = 1 - e^{-x}.$$

More generally, we denote the probability that i lineages coalesce into j lineages within time x ($j \leq i$) by $g_{i,j}(x)$. This value can be computed using the following formula (Tavaré, 1984):

$$g_{i,j}(x) = \sum_{k=j}^{i} \left(\exp\left(-\binom{k}{2} x\right) \frac{(2k-1)(-1)^{k-j}}{j!(k-j)!(j+k-1)} \cdot \prod_{m=0}^{k-1} \frac{(j+m)(i-m)}{i+m} \right).$$

Further, note that if at any given moment in time multiple lineages co-exist in the same edge, then any pair of these lineages have an equal probability of coalescing in the next Δt time. That is, the process is symmetric.

2.3 Bounded MSC (b-MSC) model

The constraints on MSC in the unified DLCoal model appear due to the duplication points. In particular, all lineages originating below a daughter duplicate must coalesce below the respective duplication node.

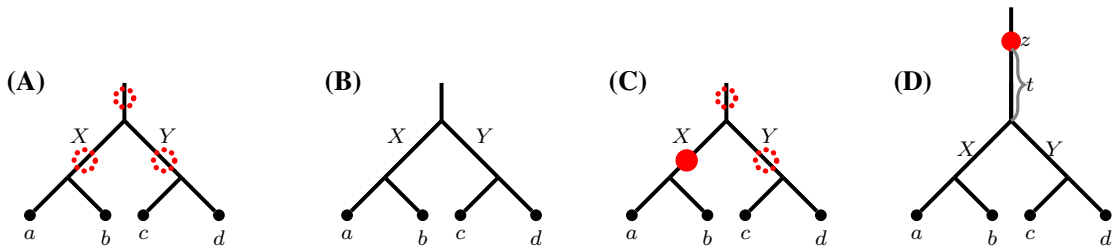

Fig. 5: **(A)**: The balanced quartet representing the locus tree and displaying quartet ab|cd. The dotted circles indicate potential duplication locations that can affect gene tree probabilities. **(B)-(D)**: Specific duplication scenarios corresponding to Sections 3.1.1, 3.1.2, and 3.1.3, respectively.

For example, in Figure 3, the gene tree lineages corresponding to leaves a_2, c_2, and c_3 must coalesce below the root node.

More formally, assume that a duplication occurred at time-point d. Note that, for convenience, we assume that all leaves are aligned in time and are associated with time-point 0; further, we consider time to increase as we go up the trees away from leaves. Now, let a and b be locus tree leaves that are located below the duplication, which is at time-point d (i.e., a and b belong to the new locus created by the duplication). Then we know that lineages a and b must coalesce prior to time-point d. Therefore, generally, the probability that any two lineages a, b, which entered the same edge below a duplication dup at time d, coalesce within time x is as follows (see Figure 4 for a respective locus tree example):

$$P[a,b \text{ coalesced within time } x \mid a,b \text{ coalesced prior to } dup] = \frac{1-e^{-x}}{P[a,b \text{ coalesced prior to } dup]},$$

where $P[a,b$ coalesced prior to $dup]$ is determined by the original, unbounded MSC model.

3 Quartet probabilities under b-MSC

To obtain our main result we need to compute the probabilities of each quartet appearing in the gene tree based on a fixed locus tree topology. Note that Allman *et al.*, 2011 explicitly computed these probabilities for unbounded MSC. In our case we need to incorporate cases, when duplications (locus creation events) appear along the edges of the locus tree.

Remark. *From now on, for convenience, we restrict locus trees to four leaves sampled from different species. That is, choosing (any) four genes $\{a,b,c,d\}$ from distinct species A, B, C, D, we consider the tree $L|_{\{a,b,c,d\}}$. Note that considering only four leaves may suppress other duplication nodes along the locus tree edges. Therefore, we need to allow for additional duplication-points along the locus tree edges. Further, if there are multiple duplication-points along a single edge of $L|_{\{a,b,c,d\}}$, it is sufficient to only consider the lowest duplication-point on that edge since it indicates the lowest point, below which gene lineages must coalesce.*

Without loss of generality assume that the locus tree L displays the quartet ab|cd. Then there are two cases: either (i) L is a balanced rooted tree or (ii) L is a caterpillar tree. We now explore both those cases.

Throughout this section we sometimes use abbreviations 'coal.' for 'coalesce(d)' and 'dup.' for 'duplication'. Further, we abbreviate 'obtained in time t' as simply 'in t'.

3.1 L is balanced

For convenience, we set $x := \omega_L(X), y := \omega_L(Y)$ to be the lengths of edges X and Y, respectively (see Figure 5(A)). We now explore all possibilities of duplication placements on edges of L.

3.1.1 No duplications (unbounded MSC).
In this case quartet probabilities are given by Allman *et al.*, 2011. That is,

$$P[\mathsf{ab}|\mathsf{cd} \in G] = 1 - \frac{2}{3}e^{-(x+y)};$$
$$P[\mathsf{ac}|\mathsf{bd} \in G] = P[\mathsf{ad}|\mathsf{bc} \in G] = \frac{1}{3}e^{-(x+y)}.$$

3.1.2 Duplications along the X or Y edges.
Assume that a duplication has occurred along the X and/or Y edge (see Figure 5(C)). Recall that a duplication point indicates that gene lineages below it in the locus tree must coalesce prior to the duplication (when looking backwards in time). Therefore, if there is a duplication along the X edge, then lineages corresponding to genes a and b must coalesce on that edge. That is, the gene tree must display quartet ab|cd. Similarly, the same is true if a duplication is located on the Y edge. Hence,

$$P[\mathsf{ab}|\mathsf{cd} \in G] = 1;$$
$$P[\mathsf{ac}|\mathsf{bd} \in G] = P[\mathsf{ad}|\mathsf{bc} \in G] = 0.$$

3.1.3 Root edge duplications
Assume that a duplication occurred on the root edge as shown in Figure 5(D), and no duplications appear on X and Y edges. Then the following holds.

$$\begin{aligned}
P[\mathsf{ab}|\mathsf{cd} \in G] &= P[\mathsf{ab}|\mathsf{cd} \in G \mid a,b,c,d \text{ coalesced before } z] \\
&= 1 - P\left[\begin{array}{l|l} a,b \text{ did not coal. on } X; & \\ c,d \text{ did not coal. on } Y; & a,b,c,d \text{ coal. before } z \\ \mathsf{ac}|\mathsf{bd} \text{ or } \mathsf{ad}|\mathsf{bc} \text{ in } t & \end{array}\right] \\
&= 1 - \frac{\frac{2}{3}e^{-x}e^{-y}P[4 \text{ lineages coalesced within time } t]}{P[a,b,c,d \text{ coalesced before } z]} \\
&= 1 - \frac{2}{3}e^{-(x+y)}\frac{g_{4,1}(t)}{P[a,b,c,d \text{ coalesced before } z]};
\end{aligned}$$

$$\begin{aligned}
P[\mathsf{ac}|\mathsf{bd} \in G] &= P[\mathsf{ad}|\mathsf{bc} \in G] \\
&= \frac{1}{3}e^{-(x+y)}\frac{g_{4,1}(t)}{P[a,b,c,d \text{ coalesced before } z]}.
\end{aligned}$$

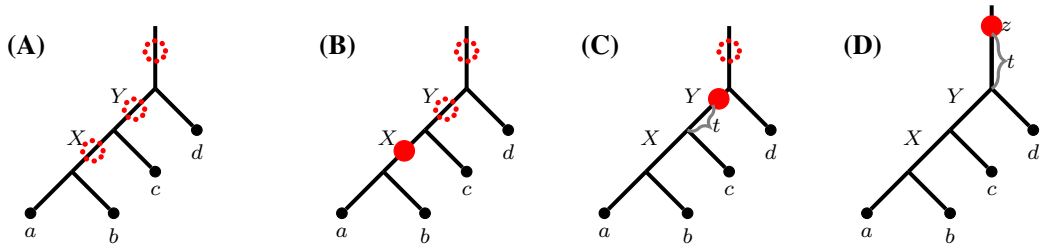

Fig. 6: **(A)**: The caterpillar quartet representing the locus tree and displaying quartet ab|cd. **(B)-(D)**: Specific duplication scenarios corresponding to Sections 3.2.2, 3.2.3, and 3.2.4, respectively.

3.1.4 Duplication at the root vertex

In Sections 4 and 5 we mainly consider cases when the locus tree root corresponds to a locus creation event (i.e., there is a duplication-point on one of the X or Y edges right below the root). In that case the gene tree quartet probabilities are given by Lemma 3.1.

Lemma 3.1. *Let L be a balanced locus tree displaying a quartet q with the root of L corresponding to a duplication. Then $P[q \in G \mid L] = 1$.*

Proof. Since the root of L is a duplication, we place a duplication-point immediately below the root on one of the children edges (i.e., the edge that corresponds to a novel locus). Therefore, quartet probabilities for L are described in Section 3.1.2. That is, $P[q \in G \mid L] = 1$.

Remark. *Note that potential duplications along the external edges do not affect the coalescence process.*

3.2 L is a caterpillar

As above, we set $x := \omega_L(X), y := \omega_L(Y)$ to be the lengths of edges X and Y respectively (see Figure 6(A)). We now similarly explore all possible duplication placements on the edges of L.

3.2.1 No duplications (unbounded MSC).

In this case the quartet probabilities are given by Allman *et al.*, 2011. In particular,

$$P[\mathsf{ab|cd} \in G] = 1 - \frac{2}{3}e^{-x};$$
$$P[\mathsf{ac|bd} \in G] = P[\mathsf{ad|bc} \in G] = \frac{1}{3}e^{-x}.$$

3.2.2 X edge duplication.

Assume that there is a duplication on the X edge (and potentially more duplications on other internal edges) as shown in Figure 6(B). Then, similarly to the balanced case, it is not difficult to see that

$$P[\mathsf{ab|cd} \in G] = 1;$$
$$P[\mathsf{ac|bd} \in G] = P[\mathsf{ad|bc} \in G] = 0.$$

3.2.3 Y edge duplication.

Assume that there is a duplication on Y and there are no duplications on X as shown in Figure 6(C).

$$P[\mathsf{ab|cd} \in G] = P[\mathsf{ab|cd} \in G \mid a, b, c \text{ coal. before duplication}]$$
$$= 1 - P\left[\begin{array}{l} a, b \text{ did not coalesce on } X; \\ \mathsf{ac|bd} \text{ or } \mathsf{ad|bc} \text{ in } t \end{array} \middle| \, a, b, c \text{ coalesced before dup.}\right]$$
$$= 1 - \frac{2}{3}e^{-x}\frac{g_{3,1}(t)}{P[a, b, c \text{ coalesced before duplication}]};$$
$$P[\mathsf{ac|bd} \in G] = P[\mathsf{ad|bc} \in G] = \frac{1}{3}e^{-x}\frac{g_{3,1}(t)}{P[a, b, c \text{ coal. before dup.}]}.$$

3.2.4 Root edge duplication.

Assume that a duplication occurred on the root edge and no duplications occurred along the X and Y edges (see Figure 6(D)).

We start with computing the probability of the ac|bd quartet.

$$P[\mathsf{ac|bd} \in G] = \frac{P\left[\begin{array}{l} a, b \text{ did not coalesce on } X; \\ a, c \text{ coalesced on } Y \text{ first}; \\ \text{remaining lineages coalesced before } z \end{array}\right]}{P[a, b, c, d \text{ coalesced before the duplication}]}$$
$$+ \frac{P\left[\begin{array}{l} a, b \text{ did not coalesce on } X; \\ \text{no coalescence on } Y; \\ \mathsf{ac|bd} \text{ obtained in time } t \end{array}\right]}{P[a, b, c, d \text{ coalesced before the duplication}]}$$
$$= \frac{\frac{1}{3}e^{-x}\left(g_{3,2}(y)g_{3,1}(t) + g_{3,1}(y)g_{2,1}(t) + g_{3,3}(y)g_{4,1}(t)\right)}{P[a, b, c, d \text{ coalesced before the duplication}]}.$$

Further, by symmetry $P[\mathsf{ac|bd} \in G] = P[\mathsf{ad|bc} \in G]$. Therefore, $P[\mathsf{ab|cd} \in G]$ equals

$$1 - \frac{2}{3}e^{-x}\left(\frac{g_{3,2}(y)g_{3,1}(t) + g_{3,1}(y)g_{2,1}(t) + g_{3,3}(y)g_{4,1}(t)}{P[a, b, c, d \text{ coalesced before the duplication}]}\right).$$

3.3 Core observations

It is not difficult to see from the above derivations that for a fixed locus tree topology that displays ab|cd (balanced or caterpillar), if one increases the length of edge X then the probability $P[\mathsf{ab|cd} \in G]$ grows. More formally, see Lemma 3.2.

Lemma 3.2. *Let L_1 and L_2 be two caterpillar trees displaying* ab|cd *with $\omega_{L_1}(X) < \omega_{L_2}(X)$ and $\omega_{L_1}(Y) = \omega_{L_2}(Y)$ as shown in Figure 7. Further, let L_1 and L_2 have identical locations of duplication-points on the internal edges. That is, a duplication-point d_1 on L_1 always has a*

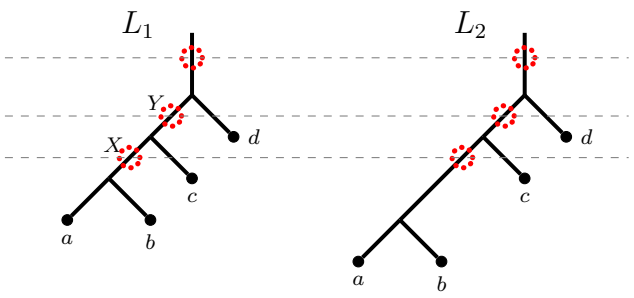

Fig. 7: Locus trees L_1 and L_2 with equal lengths of the Y edges and different lengths of the X edges. The dashed lines highlight that the duplication-points are located identically on the two trees relatively to their roots.

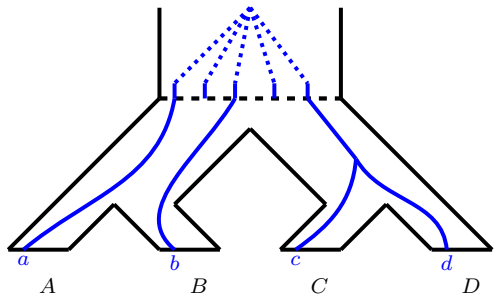

Fig. 8: An example of the partial embedding of a locus tree into balanced S. The blue lineages correspond to the locus tree. Note that the five locus lineages crossing the dashed speciation line are *root lineages*.

counterpart duplication-point d_2 on L_2 with the same distance to the root and vice versa (see Figure 7).Then

$$P[\mathsf{ab}|\mathsf{cd} \in G \mid L_1] = P[\mathsf{ab}|\mathsf{cd} \in G \mid L_2] = 1,$$

if L_1 (and L_2) have a duplication-point on edge X, and

$$P[\mathsf{ab}|\mathsf{cd} \in G \mid L_1] < P[\mathsf{ab}|\mathsf{cd} \in G \mid L_2],$$

otherwise.

Further, from the above derivations we observe the following lemma.

Lemma 3.3. *For a locus tree L that displays* ab|cd *(regardless of duplication locations) we have $P[\mathsf{ab}|\mathsf{cd} \in G \mid L] > P[\mathsf{ac}|\mathsf{bd} \in G \mid L] = P[\mathsf{ad}|\mathsf{bc} \in G \mid L]$.*

The proofs of Lemmas 3.2 and 3.3 are given in Section S1 of the Supplementary Material.

4 Consistency of ASTRAL-one

We now prove ASTRAL-one is statistically consistent under the DLCoal model.

Theorem 4.1. *Let $S = (T_S, w_S)$ be a fixed species tree and let $\mathcal{G}$ be a collection of gene trees that independently evolved within S according to the DLCoal process. Then, as the number of trees in $\mathcal{G}$ goes to infinity, the probability that $\hat{T}$, the unrooted tree estimate by ASTRAL-one, is equal to the unrooted tree topology T_S goes to 1.*

For this result, it is sufficient (see Legried *et al.*, 2020) to prove the following:

Theorem 4.2. *Let S be a species tree with four leaves that displays quartet $AB|CD$, and let G be a gene tree that evolved in S according to the DLCoal process. If one picks genes a, b, c, d (that correspond to species A, B, C, D respectively) uniformly at random (assuming they exist) from G, then $P[\mathsf{ab}|\mathsf{cd} \in G] > P[\mathsf{ac}|\mathsf{bd} \in G] = P[\mathsf{ad}|\mathsf{bc} \in G]$.*

Theorem 4.2 is sufficient to prove Theorem 4.1, because ASTRAL, as a distance-minimization method, 'prefers' the most dominant quartets among the input trees. Then, by Theorem 4.2, as the number of input trees goes to infinity, the most dominant quartet among input trees for each 4-tuple of species becomes (*almost surely*) the true species tree quartet; hence, it is almost surely picked by ASTRAL-one (see Legried *et al.*, 2020 for a formal proof). Therefore, the remainder of the section is dedicated to the proof of Theorem 4.2. We first prove the theorem for S being balanced and then for S being a caterpillar.

Remark. *To prove Theorem 4.2, we will use some of the results from Legried* et al.*, 2020, who proved that ASTRAL is consistent under the duplication/loss process. To see how their result relates to our problem, observe that a 'gene tree' in Legried* et al.*, 2020 notation is equivalent to the locus tree in the broader DLCoal process. Therefore, below we explicitly use some of Legried et al. results to draw conclusions about the locus tree probabilities.*

4.1 S is balanced

Similarly to Legried *et al.*, 2020, we first implicitly condition our probability space on the event that at least one of each a, b, c and d genes must be present in G. Further, we condition our probability space on a fixed number of locus tree lineages existing at the speciation point at the root of S. That is, consider the duplication/loss (birth/death) process occurring within the root branch of S. Then, let RL be the random variable denoting the number of locus lineages at the speciation point (see Figure 8). We are going to prove that

$$\begin{aligned} P[\mathsf{ab}|\mathsf{cd} \in G \mid RL = l] &> P[\mathsf{ac}|\mathsf{bd} \in G \mid RL = l] \\ &= P[\mathsf{ad}|\mathsf{bc} \in G \mid RL = l] \end{aligned}$$

for any fixed value of $l = \{1, 2, \ldots\}$. Therefore, for convenience, we do not explicitly write the condition $RL = l$ in probability equations throughout the rest of the proof. Further, we refer to the set of these l locus lineages as *root lineages*.

Now let $i_a \in \{1, \ldots, l\}$ be the index of a root lineage, from which gene a has descended. Similarly, we define i_b, i_c, and i_d. For better readability of the remainder of the proof, we introduce the notation to describe scenarios of the type $i_a = i_b = i_c \neq i_d$. In particular, we write (abc, d) for that scenario, we write (ab, cd) to denote the scenario $i_a = i_b \neq i_c = i_d$, and we write (a, b, c, d) to denote the scenario, where all i_x are distinct.

Then, by the law of total probability, we have

$$P[\mathsf{ab}|\mathsf{cd} \in G] = \sum_I P[I, \mathsf{ab}|\mathsf{cd} \in G],$$

where I is one of the above scenarios (i.e., a partition of set $\{a, b, c, d\}$ or a combination of such partitions). In particular $I \in \{(a, b, c, d); (ab, cd) \vee (ac, bd); (ab, c, d) \vee (cd, a, b) \vee (ac, b, d) \vee (bd, a, c); (abc, d) \vee (abd, c) \vee (acd, b) \vee (bcd, a) \vee (abcd); (ad, bc) \vee (ad, b, c) \vee (bc, a, d)\}$. Observe that we cover all possible scenarios/partitions here.

Our goal is to prove that $P[\mathsf{ab}|\mathsf{cd} \in G] > P[\mathsf{ac}|\mathsf{bd} \in G]$. Note that $P[\mathsf{ad}|\mathsf{bc} \in G] = P[\mathsf{ac}|\mathsf{bd} \in G]$ follows from the fact that swapping c

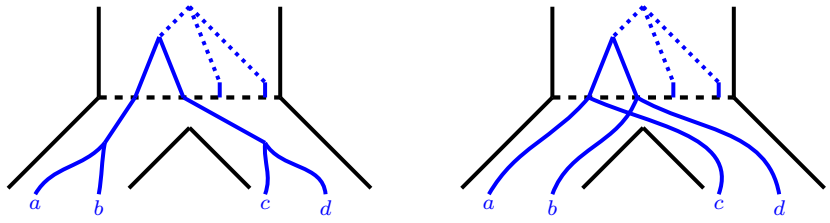

Fig. 9: Left: the embedding of a locus tree $L_{(ab,cd)}$. Right: the embedding of a locus tree $L_{(ac,bd)}$.

and d leaf labels does not affect the probabilities. Let us carry out the proof by considering different values of I. That is, our strategy is to prove that $P[I, \mathsf{ab|cd} \in G] \geq P[I, \mathsf{ac|bd} \in G]$ for all of the above I, and at least in one case the *strict* inequality holds.

To facilitate the proofs in each case, first consider the following observations:

Observation 4.1. *Random variables i_x and i_y are independent for any $x \in \{a, b\}$ and $y \in \{c, d\}$.*
However, i_a can be dependent on i_b and i_c can be dependent on i_d.

Proof. Observe that the duplication/loss process runs independently in the parallel branches of the species tree. Therefore, once we condition the probability space on a fixed number of lineages at the divergence point (i.e., fixed l), the random variables i_x and i_y become independent. In particular, consider any specific realization of the duplication/loss process below the root lineages and let i be a root lineage that a randomly picked locus a belongs two (i.e., $i_a = i$). Then, we can swap the 'left' subtrees below two distinct root lineages i and j (the subtrees that lead to species A and B) so that $i_a = j$ and the probability of that event is not altered due to symmetry. Note that i_c in that case remains the same. Since we can always reshuffle root lineages like that, we can think of a as 'choosing' one of the l root lineages uniformly at random, regardless of a realization of i_c. The same is also true for all other pairs of $x \in \{a, b\}$ and $y \in \{c, d\}$.

However, since a and b develop (at least partially) in the same species tree branch random variables i_a and i_b can be dependent. Similarly for i_c and i_d.

Observation 4.2. *Due to the symmetry of the duplication/loss process, we have*

$$P[i_x = k] = 1/l$$

for any $x \in \{a, b, c, d\}$ and $k \in \{1, 2, \ldots, l\}$. Then, by Claim 4.1,

$$P[i_x = i_y] = \sum_{k=1}^{l} P[i_x = k]P[i_y = k] = l\frac{1}{l^2} = 1/l$$

for any $x \in \{a, b\}$ and $y \in \{c, d\}$.

Lemma 4.1 (Due to Lemma 1 in Legried *et al.*, 2020). *$P[i_a = i_b]$ and $P[i_c = i_d]$ are greater than or equal to $\frac{1}{l}$.*

4.1.1 Case $I = (a, b, c, d)$

By the symmetry of the duplication/loss process, reshuffling the i_a, i_b, i_c, and i_d labels will not change the probability of a fixed duplication/loss history in the root edge. Therefore, we have $P[\mathsf{ab|cd} \in G \mid I] = P[\mathsf{ac|bd} \in G \mid I]$. Hence, $P[\mathsf{ab|cd} \in G, I] = P[\mathsf{ac|bd} \in G, I]$.

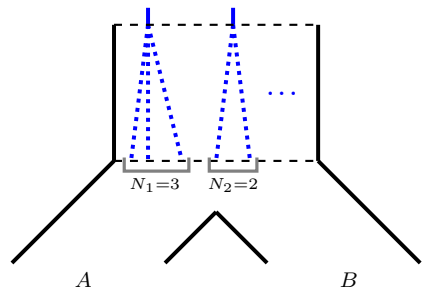

Fig. 10: An example of a partial locus tree embedding in the left part of the species tree below the root speciation. The two shown root lineages expand (through duplication) into $N_1 = 3$ and $N_2 = 2$ lineages at the moment of A/B speciation, respectively

4.1.2 Case $I = (ab, cd) \vee (ac, bd)$

We need to show that

$$\begin{aligned} P[\mathsf{ab|cd} \in G, I] &= P[\mathsf{ab|cd} \in G \mid (ab, cd)]\ P[(ab, cd)] \\ &\quad + P[\mathsf{ab|cd} \in G \mid (ac, bd)]\ P[(ac, bd)] \\ &\geq P[\mathsf{ac|bd} \in G \mid (ab, cd)]\ P[(ab, cd)] \\ &\quad + P[\mathsf{ac|bd} \in G \mid (ac, bd)]\ P[(ac, bd)] \\ &= P[\mathsf{ac|bd} \in G, I]. \end{aligned}$$

Observe the following.

Lemma 4.2. $P[\mathsf{ab|cd} \in G \mid (ab, cd)] = P[\mathsf{ac|bd} \in G \mid (ac, bd)] = 1$.

Proof. Consider the locus trees $L_{(ab,cd)}$ and $L_{(ac,bd)}$ for the (ab, cd) and (ac, bd) cases respectively (see Figure 9). Note that we only consider the part of the locus tree restricted to the four selected genes a, b, c, d. It is not difficult to see that both $L_{(ab,cd)}$ and $L_{(ac,bd)}$ are balanced. Therefore, by Lemma 3.1, $P[\mathsf{ab|cd} \in G \mid (ab, cd)] = P[\mathsf{ac|bd} \in G \mid (ac, bd)] = 1$.

Corollary 4.1. $P[\mathsf{ac|bd} \in G \mid (ab, cd)] = P[\mathsf{ab|cd} \in G \mid (ac, bd)] = 0$.

Lemma 4.3. $P[(ab, cd)] \geq P[(ac, bd)]$.

Proof. Our proof is similar to the proof of Lemma 1 in Legried *et al.*, 2020. In particular, let $N_i \in \{0, 1, \ldots\}$ be the number of locus lineages that descended from a root lineage $i \in \{1, \ldots, l\}$ and that existed *immediately after* the speciation into species A and B. Similarly, we define variables M_i denoting the number of lineages that existed immediately after the speciation at the parent of C and D. See Figure 10 for an example of N_i variables. By $\mathbf{N} = (N_1, \ldots, N_l)$ and $\mathbf{M} = (M_1, \ldots, M_l)$ we denote the vectors of N_i and M_i variables, respectively.

Observe that $P[(ab, cd)] = P[i_a = i_b, i_c = i_d] - P[(abcd)]$ and $P[(ac, bd)] = P[i_a = i_c, i_b = i_d] - P[(abcd)]$. Further, note that when conditioned on specific values of $\mathbf{N}$ and $\mathbf{M}$, $i_a = i_c$ and $i_b = i_d$ events become independent. That is, similarly to Claim 4.1, conditioning on the number of lineages at the divergence point for species A and B eliminates the dependency between i_a and i_b (and similarly for i_c and i_d). After conditioning on $\mathbf{N}$ and $\mathbf{M}$, random variables i_a, i_b, i_c, and i_d are all independent. In particular, we can think of lineages a and b as choosing one of the $\sum N_i$ lineages independently and uniformly at random. Similarly, c and d choose one of the $\sum M_i$ lineages independently and uniformly at random.

Then, for fixed values of the $\mathbf{N}$ and $\mathbf{M}$ vectors we have

$$\begin{aligned} P[i_a = i_b, i_c = i_d \mid \mathbf{N}, \mathbf{M}] &= \sum_{j=1}^{l} \left(P[i_a = j \mid \mathbf{N}] P[i_b = j \mid \mathbf{N}]\right) \\ &\quad \cdot \sum_{j=1}^{l} \left(P[i_c = j \mid \mathbf{M}] P[i_d = j \mid \mathbf{M}]\right) \\ &= \frac{\sum_j (N_j^2)}{(\sum_j N_j)^2} \frac{\sum_j (M_j^2)}{(\sum_j M_j)^2}. \end{aligned}$$

The last equality is due to $P[i_a = j \mid \mathbf{N}] = \frac{N_j}{\sum_{i=1}^{l} N_i}$. That is, as mentioned above, due to the symmetry of the duplication/loss process a has a uniform probability of being 'sampled' from any of the lineages existing at the divergence point of species A and B. Similar relations can then be easily derived for i_b, i_c, and i_d.

Further, following the same idea, we have

$$P[i_a = i_c, i_b = i_d \mid \mathbf{N}, \mathbf{M}] = \frac{\sum_j (N_j M_j)}{(\sum_j N_j)(\sum_j M_j)} \frac{\sum_j (N_j M_j)}{(\sum_j N_j)(\sum_j M_j)}.$$

Then, by Cauchy-Schwartz, $(\sum_j (N_j M_j))^2 \leq \sum_j (N_j^2) \sum_j (M_j^2)$ and therefore $P[i_a = i_b, i_c = i_d \mid \mathbf{N}, \mathbf{M}] \geq P[i_a = i_c, i_b = i_d \mid \mathbf{N}, \mathbf{M}]$ for any realization of vectors $\mathbf{N}$ and $\mathbf{M}$. That is, $P[(ab, cd)] \geq P[(ac, bd)]$.

Using the above results, we have

$$P[\mathsf{ab|cd} \in G, I] = P[(ab, cd)] \geq P[(ac, bd)] = P[\mathsf{ac|bd} \in G, I].$$

4.1.3 Case $I = (ab, c, d) \vee (cd, a, b) \vee (ac, b, d) \vee (bd, a, c)$

For convenience, from now on we denote the event $(ab, c, d) \vee (cd, a, b)$ by $\boldsymbol{AB}$ and the event $(ac, b, d) \vee (bd, a, c)$ by $\boldsymbol{AC}$.

We prove that

$$\begin{aligned} &P[\mathsf{ab|cd} \in G, I] \\ &\quad = P[\mathsf{ab|cd} \in G \mid AB]\ P[AB] + P[\mathsf{ab|cd} \in G \mid AC]\ P[AC] \\ &\quad \geq P[\mathsf{ac|bd} \in G \mid AB]\ P[AB] + P[\mathsf{ac|bd} \in G \mid AC]\ P[AC] \\ &\quad = P[\mathsf{ac|bd} \in G, I]. \end{aligned}$$

Consider the following results.

Lemma 4.4. $P[\mathsf{ab|cd} \in G \mid AB] \geq P[\mathsf{ac|bd} \in G \mid AC]$.

Proof. Note that fixing the number of root lineages allows us to treat the duplication/loss processes independently for the root edge and for the lower edges. Let $\mathcal{L}_r$ be a duplication/loss scenario (i.e., a fixed realization of the duplication/loss process) in the root edge conditioned on $RL = l$. Then, without loss of generality assume that in case (ab, c, d), we have $i_a = i_b = 1$, $i_c = 2$, and $i_d = 3$; in case (cd, a, b) we assume $i_c = i_d = 1$, $i_a = 2$, and $i_b = 3$. Similarly, under (ac, b, d) we assume $i_a = i_c = 1, i_b = 2, i_d = 3$ and under (bd, a, c) we assume that $i_b = i_d = 1, i_a = 2, i_c = 3$. Then, a fixed $\mathcal{L}_r$ scenario forces the same 'top' structure of the locus trees in all four cases.

Given that (ab, c, d) and (cd, a, b) cases are virtually identical for the remainder of the proof (since they are symmetric), for simplicity, we will only consider the (ab, c, d) case. Similarly, under the AC event, we will only consider case (ac, b, d).

Then, Figures 11 and 12 depict two possible topologies of the $\mathcal{L}_r$ scenario when acting on the root lineages $1, 2$, and 3. Observe that the third topology, where root lineages 1 and 3 form a cherry, is identical in terms of analysis to the topology depicted in Figure 11, and therefore is not considered.

Note that in Figure 11, the resulting locus trees $L_{(ab,c,d)}$ and $L_{(ac,b,d)}$ are both caterpillars, while in Figure 12, the locus trees are both balanced. This separation is achieved because we condition on a fixed $\mathcal{L}_r$ scenario. We now consider these two cases individually.

(i) $L_{(ab,c,d)}$ **and** $L_{(ac,b,d)}$ **are caterpillars** (see Figure 11). Let x_{ab} be the distance (in coalescent units) from the root speciation event to the divergence of a and b in the locus tree under the (ab, c, d) case (as shown on the figure). Note that $x_{ab} \geq 0$. There are two cases to consider.

- *There is a duplication along the x_{ab} lineage.* Then, as shown in Section 3.2.2, $P[\mathsf{ab|cd} \in G \mid AB, \mathcal{L}_r] = 1$. That is, $P[\mathsf{ab|cd} \in G \mid AB, \mathcal{L}_r] \geq P[\mathsf{ac|bd} \in G \mid AC, \mathcal{L}_r]$.
- *No duplications along the x_{ab} lineage.* Since $L_{(ab,c,d)}$ and $L_{(ac,b,d)}$ are both caterpillars, we denote their edges by X and Y as shown in Figure 6(A). In particular we denote the X edge in $L_{(ab,c,d)}$ by $X_{(ab,c,d)}$ and the X edge in $L_{(ac,b,d)}$ by $X_{(ac,b,d)}$. Then, $w(X_{(ab,c,d)}) = x' + x_{ab}$, whereas $w(X_{(ac,b,d)}) = x'$ (note that x' is as depicted in Figure 11). Further, the two locus trees are identical in terms of the duplication locations in their internal edges.

  Then, by Lemma 3.2, it is not difficult to see that $P[\mathsf{ab|cd} \in G \mid \mathcal{L}_r, (ab, c, d)] \geq P[\mathsf{ac|bd} \in G \mid \mathcal{L}_r, (ac, b, d)]$ for any fixed $\mathcal{L}_r$. Therefore, the lemma holds.

(ii) $L_{(ab,c,d)}$ **and** $L_{(ac,b,d)}$ **are balanced** (see Figure 12). By Lemma 3.1, $P[\mathsf{ab|cd} \in G \mid AB, \mathcal{L}_r] = 1$ and $P[\mathsf{ac|bd} \in G \mid AC, \mathcal{L}_r] = 1$. Note that we can apply Lemma 3.1, since the roots of the locus trees in these cases must be duplications.

Lemma 4.5. $P[\mathsf{ab|cd} \in G \mid AC] \geq P[\mathsf{ac|bd} \in G \mid AB]$.

Proof. This result follows from Lemma 4.4 (i.e., $P[\mathsf{ab|cd} \in G \mid AB] \geq P[\mathsf{ac|bd} \in G \mid AC]$) and the following relations:

$$\begin{aligned} 2P[\mathsf{ab|cd} \in G \mid AC] + P[\mathsf{ac|bd} \in G \mid AC] = 1; \\ 2P[\mathsf{ac|bd} \in G \mid AB] + P[\mathsf{ab|cd} \in G \mid AB] = 1. \end{aligned}$$

Observation 4.3. *By Lemma 3.3, we have $P[\mathsf{ac|bd} \in G \mid AC] \geq P[\mathsf{ab|cd} \in G \mid AC]$. Then, combining this with Lemma 4.5, we have $P[\mathsf{ac|bd} \in G \mid AC] \geq P[\mathsf{ac|bd} \in G \mid AB]$.*

Lemma 4.6. $P[AB] \geq P[AC]$.

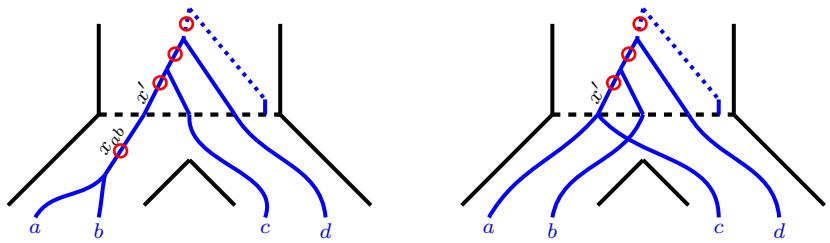

Fig. 11: Caterpillar locus trees $L_{(ab,c,d)}$ (left) and $L_{(ac,b,d)}$ (right) embedded into the species tree. The red circles represent the potential duplication locations that could influence the gene tree probabilities. Note that the $\mathcal{L}_r$ scenarios in the root edges are identical. That is, x' lengths are equal, and the duplication locations above the dashed speciation lines are identical.

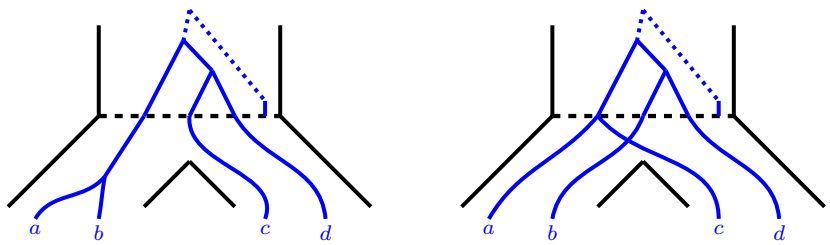

Fig. 12: Balanced locus trees $L_{(ab,c,d)}$ (left) and $L_{(ac,b,d)}$ (right) embedded into the species tree.

Proof. We give the proof in Section S2 of Supplementary Material.

Summarizing the above results we have.

$$P[\mathsf{ab|cd} \in G \mid AB]\ P[AB] + P[\mathsf{ab|cd} \in G \mid AC]\ P[AC]$$
$$\geq P[\mathsf{ac|bd} \in G \mid AC]\ P[AB] + P[\mathsf{ac|bd} \in G \mid AB]\ P[AC]$$
$$\geq P[\mathsf{ac|bd} \in G \mid AC]\ P[AC] + P[\mathsf{ac|bd} \in G \mid AB]\ P[AB].$$

Note that the first inequality is due to Lemmas 4.4 and 4.5. The last inequality is due to Lemma 4.6 and Claim 4.3.

That is, our main statement holds.

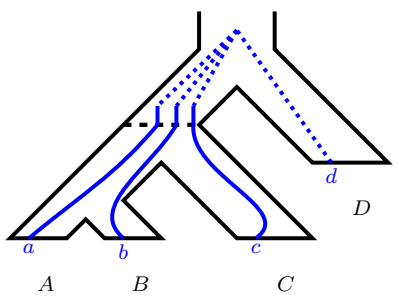

Fig. 13: An example of the locus tree embedding into a caterpillar species tree. The three locus lineages crossing the dashed speciation line are the ABC-lineages.

4.1.4 Case $I = (abc, d) \vee (abd, c) \vee (acd, b) \vee (bcd, a) \vee (abcd)$.

In all five cases locus tree L displays the quartet $\mathsf{ab|cd}$. Therefore, by Lemma 3.3 $P[\mathsf{ab|cd} \in G \mid I] > P[\mathsf{ac|bd} \in G \mid I]$. Observe that we obtain the strict inequality in this case.

4.1.5 Case $I = (ad, bc) \vee (ad, b, c) \vee (bc, a, d)$.

In this case it is not difficult to see that L displays quartet $\mathsf{ad|bc}$. Therefore (as can be seen from the derivations in Section 3), $P[\mathsf{ab|cd} \in G \mid I] = P[\mathsf{ac|bd} \in G \mid I]$.

This concludes the proof for balanced S.

4.2 S is a caterpillar

Without loss of generality assume that S is as it appears in Figure 13. Similarly to the balanced case, we implicitly condition the probability space on a fixed number of loci (lineages) existing at the moment of speciation as shown in the figure. Note that, while in the balanced case we considered root lineages, in the caterpillar scenario we consider lineages at the least common ancestor of A, B, and C. That is, we refer to these lineages/loci as *ABC-lineages*. Finally, as in the balanced case, we denote the number of ABC-lineages by l.

We then use the i_a, i_b, i_c notation in the same way as in the previous section (while referring to indices of ABC-lineages). Further, $\mathcal{I} = \{(a, b, c); (ab, c); (ac, b); (bc, a); (abc)\}$ scenarios describe relations between i_a, i_b, and i_c.

We now prove that $P[\mathsf{ab|cd} \in G, I] \geq P[\mathsf{ac|bd} \in G, I]$ for all I in $\{(a, b, c); (ab, c) \vee (ac, b); (bc, a); (abc)\}$. Moreover, for at least one such I, the strict inequality holds; in particular, see case 4.2.4 below.

4.2.1 Case $I = (a, b, c)$.

By the symmetry of the duplicaion/loss process, reshuffling the i_a, i_b, i_c labels will not affect the probability of a fixed duplication/loss history in the root edge. Therefore, we have $P[\mathsf{ab|cd} \in G \mid I] = P[\mathsf{ac|bd} \in G \mid I] = P[\mathsf{ad|bc} \in G \mid I]$.

Then, $P[\mathsf{ab|cd} \in G, I] = P[\mathsf{ab|cd} \in G \mid I]\ P[I] = P[\mathsf{ac|bd} \in G \mid I]\ P[I] = P[\mathsf{ac|bd} \in G, I]$.

4.2.2 Case $I = (ab, c) \vee (ac, b)$.

The proof in this case is similar to case 4.1.3 for balanced S. In particular, observe the following.

Lemma 4.7. $P[(ab, c)] \geq P[(ac, b)]$.

Proof. It is sufficient to show that $P[i_a = i_b] \geq P[i_a = i_c]$. By Claim 4.2, $P[i_a = i_c] = 1/l$. Further, Legried *et al.*, 2020 showed that $P[i_a = i_b] \geq 1/l$ (see Lemma 1 in Legried *et al.*, 2020).

Lemma 4.8. *The following holds.*

(i) $P[\mathsf{ab|cd} \in G \mid (ab, c)] \geq P[\mathsf{ac|bd} \in G \mid (ac, b)]$;
(ii) $P[\mathsf{ab|cd} \in G \mid (ac, b)] \geq P[\mathsf{ac|bd} \in G \mid (ab, c)]$;
(iii) $P[\mathsf{ac|bd} \in G \mid (ac, b)] \geq P[\mathsf{ac|bd} \in G \mid (ab, c)]$.

Proof. The proofs of these statements are similar to the proofs of the respective statements in Section 4.1.3. In particular, (i) corresponds to Lemma 4.4, (ii) corresponds to Lemma 4.5, and (iii) corresponds to Claim 4.3 from Section 4.1.3.

Then, similarly to Section 4.1.3 we have

$$\begin{aligned} P[\mathsf{ab|cd} \in G, I] &= P[\mathsf{ab|cd} \in G \mid (ab, c)]\ P[(ab, c)] \\ &\quad + P[\mathsf{ab|cd} \in G \mid (ac, b)]\ P[(ac, b)] \\ &\geq P[\mathsf{ac|bd} \in G \mid (ac, b)]\ P[(ab, c)] \\ &\quad + P[\mathsf{ac|bd} \in G \mid (ab, c)]\ P[(ac, b)] \\ &\geq P[\mathsf{ac|bd} \in G \mid (ac, b)]\ P[(ac, b)] \\ &\quad + P[\mathsf{ac|bd} \in G \mid (ab, c)]\ P[(ab, c)] \\ &= P[\mathsf{ac|bd} \in G, I]. \end{aligned}$$

4.2.3 Case $I = (bc, a)$.

In this case $P[\mathsf{ab|cd} \in G \mid I] = P[\mathsf{ac|bd} \in G \mid I]$, since the locus tree displays the third quartet, $\mathsf{ad|bc}$.

4.2.4 Case $I = (abc)$.

The locus tree displays quartet $\mathsf{ab|cd}$; therefore, by Lemma 3.3 and the law of total probability, we have $P[\mathsf{ab|cd} \in G \mid I] > P[\mathsf{ac|bd} \in G \mid I]$.

5 Consistency of ASTRAL-multi

We now extend our consistency result for ASTRAL-one to another variant of ASTRAL adapted to multi-locus input trees, called ASTRAL-multi.

Theorem 5.1. *Let $S = (T_S, w_S)$ be a fixed species tree and let $\mathcal{G}$ be a collection of gene trees that independently evolved within S according to the DLCoal process. Then, as the number of trees in $\mathcal{G}$ goes to infinity, the unrooted tree estimate by ASTRAL-multi converges almost surely to T_S.*

Let S be a species tree with 4 leaves that displays $AB|CD$, and let G be a gene tree that evolved in S according to the DLCoal process. Let $\mathbb{G}_{\mathsf{ab|cd}}$ (respectively $\mathbb{G}_{\mathsf{ac|bd}}$ and $\mathbb{G}_{\mathsf{ad|bc}}$) be the number of $\mathsf{ab|cd}$ (respectively $\mathsf{ac|bd}$ and $\mathsf{ad|bc}$) quartets in G. Then, to prove Theorem 5.1 it is sufficient to show that the following result holds (Legried *et al.*, 2020):

Theorem 5.2. $E[\mathbb{G}_{\mathsf{ab|cd}}] > \max\left(E[\mathbb{G}_{\mathsf{ac|bd}}], E[\mathbb{G}_{\mathsf{ad|bc}}]\right)$.

The remainder of the section is dedicated to the proof of Theorem 5.2. In fact, due to symmetry, it is sufficient to show that $E[\mathbb{G}_{\mathsf{ab|cd}}] > E[\mathbb{G}_{\mathsf{ac|bd}}]$. The general structure of the proof is similar to the proof of consistency for ASTRAL-one in the previous section. We present the proof for balanced S, and then briefly discuss the proof for caterpillar S.

Remark. *Some results in this section hold* almost surely. *Since this is sufficient for the proof of the theorem, we do not specify this explicitly.*

5.1 Proof of Theorem 5.2

As mentioned above, we assume that S is balanced. As before, we implicitly condition the probability space (and the expected values) on a fixed number of root lineages l. That is, we claim that Theorem 5.1 holds for any fixed value of l.

We now introduce our core notation for the proof. Similarly to the $\mathbb{G}_{\mathsf{ab|cd}}$ notation, we let $\mathbb{L}_{\mathsf{ab|cd}}$ (respectively $\mathbb{L}_{\mathsf{ac|bd}}$ and $\mathbb{L}_{\mathsf{ad|bc}}$) denote the number of $\mathsf{ab|cd}$ (respectively $\mathsf{ac|bd}$ and $\mathsf{ad|bc}$) quartets in the locus tree L. Further, for a fixed scenario I (e.g., scenario $i_a = i_b = 1, i_c = 2, i_d = 3$) let $\mathbb{L}^I_{\mathsf{ab|cd}}$ be the number of $\mathsf{ab|cd}$ quartets in the locus tree that follow the scenario I. Further, $\mathbb{G}^I_{\mathsf{ab|cd}}$ be the number of $\mathsf{ab|cd}$ quartets in G that 'appeared' from one of the $\mathbb{L}^I_{\mathsf{ab|cd}}$ quartets. Similarly we define $\mathbb{L}^I_{\mathsf{ac|bd}}, \mathbb{L}^I_{\mathsf{ad|bc}}, \mathbb{G}^I_{\mathsf{ac|bd}}$, and $\mathbb{G}^I_{\mathsf{ad|bc}}$.

Consider any $I \neq (abcd)$. Note that the root of locus tree $L|_{\{a,b,c,d\}}$ must be a duplication for such I (because I involves at least two root lineages). Then, if I always defines balanced quartets, we have $\mathbb{G}^I_q = \mathbb{L}^I_q$ for any $q \in \{\mathsf{ab|cd}, \mathsf{ac|bd}, \mathsf{ad|bc}\}$ by Lemma 3.1. In particular, we note the following:

Observation 5.1. *For any $q \in \{\mathsf{ab|cd}, \mathsf{ac|bd}, \mathsf{ad|bc}\}$ we have*

$$\mathbb{G}_q^{(ab,cd)} = \mathbb{L}_q^{(ab,cd)};$$
$$\mathbb{G}_q^{(ac,bd)} = \mathbb{L}_q^{(ac,bd)}.$$

Further, we will only consider scenarios that uniquely determine the quartet types in the locus tree; therefore, we will typically omit the subscript in the $\mathbb{L}^I_q$ notation. For example, we write $\mathbb{L}^{(ab,cd)}$ instead of $\mathbb{L}^{(ab,cd)}_{\mathsf{ab|cd}}$, since $\mathsf{ab|cd}$ is the only type of quartets that can appear under scenario (ab, cd).

Given a fixed root lineage i, let $\mathcal{A}_i$ be the random variable denoting the number of a leaves generated by that lineage. Similarly, we define random variables $\mathcal{B}_i, \mathcal{C}_i$, and $\mathcal{D}_i$. By symmetry, $E[\mathcal{A}_1] = E[\mathcal{A}_2] = \ldots = E[\mathcal{A}_l]$ (with similar relations holding for $\mathcal{B}_i, \mathcal{C}_i, \mathcal{D}_i$). Then, observe the following:

Observation 5.2. *Since the duplication/loss process runs independently in the parallel branches of the species tree, $\mathcal{X}_i$ is independent from $\mathcal{Y}_j$ for any $\mathcal{X} \in \{\mathcal{A}, \mathcal{B}\}$, $\mathcal{Y} \in \{\mathcal{C}, \mathcal{D}\}$, and $i, j \in \{1, \ldots, l\}$.*

Observation 5.3. *By the symmetry of the duplication/loss process, we have*

$$E[\mathcal{X}_i] = E[\mathcal{X}_j]$$

for all $\mathcal{X} \in \{\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}\}$ and $i, j \in \{1, 2, \ldots, l\}$.

Further, the following lemma is due to Legried et al.

Lemma 5.1 (Lemma 2 in Legried *et al.*, 2020).

$$E[\mathcal{A}_1\, \mathcal{B}_1] \geq E[\mathcal{A}_1]E[\mathcal{B}_1];$$
$$E[\mathcal{C}_1\, \mathcal{D}_1] \geq E[\mathcal{C}_1]E[\mathcal{D}_1].$$

We now outline several key corollary statements.

Corollary 5.1.

$$E[\mathbb{L}^{(ab,cd)}] \geq E[\mathbb{L}^{(ac,bd)}];$$
$$E[\mathbb{L}^{(ab,c,d)}] \geq E[\mathbb{L}^{(ac,b,d)}];$$
$$E[\mathbb{L}^{i_a=i_b=1,\ i_c=2,\ i_d=3}] \geq E[\mathbb{L}^{i_a=i_c=1,\ i_b=2,\ i_d=3}].$$

Proof. To prove the first relationship, note that the duplication/loss process occurs independently below distinct root lineages. Then, we have

$$\begin{aligned} E[\mathbb{L}^{(ab,cd)}] &= l(l-1)E[\mathcal{A}_1\, \mathcal{B}_1]\ E[\mathcal{C}_1\, \mathcal{D}_1] \\ &\geq l(l-1)E[\mathcal{A}_1]\ E[\mathcal{B}_1]\ E[\mathcal{C}_1]\ E[\mathcal{D}_1] \\ &= l(l-1)E[\mathcal{A}_1\, \mathcal{C}_1]\ E[\mathcal{B}_1\, \mathcal{D}_1] = E[\mathbb{L}^{(ac,bd)}]. \end{aligned}$$

The other two relationships can be established similarly.

We now consider the following comprehensive set of scenarios: $I \in \{(a,b,c,d); (ab,cd) \vee (ac,bd); (ab,c,d) \vee (ac,b,d); (cd,a,b) \vee (bd,a,c); (abc,d) \vee (abd,c) \vee (acd,b) \vee (bcd,a) \vee (abcd); (ad,bc) \vee (ad,b,c) \vee (bc,a,d)\}$. For each I we will prove that $E[\mathbb{G}^I_{\mathsf{ab|cd}}] \geq E[\mathbb{G}^I_{\mathsf{ac|bd}}]$ and for at least one I the strict inequality holds.

5.1.1 Case $I = (a,b,c,d)$

By the symmetry of the duplication/loss process in the root edge, we have

$$E[\mathbb{G}^{(a,b,c,d)}_{\mathsf{ab|cd}}] = E[\mathbb{G}^{(a,b,c,d)}_{\mathsf{ac|bd}}] = E[\mathbb{G}^{(a,b,c,d)}_{\mathsf{ad|bc}}].$$

5.1.2 Case $I = (ab,cd) \vee (ac,bd)$

By Claim 5.1, $\mathbb{G}^I_{\mathsf{ab|cd}} = \mathbb{L}^{(ab,cd)}$ and $\mathbb{G}^I_{\mathsf{ac|bd}} = \mathbb{L}^{(ac,bd)}$.

Then, combining this with Corollary 5.1, we have

$$E[\mathbb{G}^I_{\mathsf{ab|cd}}] = E[\mathbb{L}^{(ab,cd)}] \geq E[\mathbb{L}^{(ac,bd)}] = E[\mathbb{G}^I_{\mathsf{ac|bd}}].$$

5.1.3 Case $I = (ab,c,d) \vee (ac,b,d)$

Consider a fixed duplication-loss scenario, $\mathcal{L}_r$, in the root edge of S. In this section we implicitly condition the probability space on $\mathcal{L}_r$. That is, we prove that $E[\mathbb{G}^I_{\mathsf{ab|cd}}] \geq E[\mathbb{G}^I_{\mathsf{ac|bd}}]$ for each $\mathcal{L}_r$.

Due to symmetry, we consider the following two core scenarios: $AB := (i_a = i_b = 1, i_c = 2, i_d = 3)$ and $AC := (i_a = i_c = 1, i_b = 2, i_d = 3)$. It is then sufficient to show the following:

Lemma 5.2.

$$E[\mathbb{G}^{AB}_{\mathsf{ab|cd}}] + E[\mathbb{G}^{AC}_{\mathsf{ab|cd}}] \geq E[\mathbb{G}^{AB}_{\mathsf{ac|bd}}] + E[\mathbb{G}^{AC}_{\mathsf{ac|bd}}].$$

Proof. Due to Lemma 4.4, it is not difficult to see that for any quartet on $\{a,b,c,d\}$ that evolved according to scenario AB or AC, we have $P[\mathsf{ab|cd} \in G \mid AB] \geq P[\mathsf{ac|bd} \in G \mid AC]$. Therefore, we have

$$E[\mathbb{G}^{AB}_{\mathsf{ab|cd}}] \geq E[\mathbb{L}^{AB}]\ P[\mathsf{ac|bd} \in G \mid AC].$$

Similarly, due to Lemma 4.5, we know that $P[\mathsf{ab|cd} \in G \mid AC] \geq P[\mathsf{ac|bd} \in G \mid AB]$. Therefore,

$$E[\mathbb{G}^{AB}_{\mathsf{ac|bd}}] \leq E[\mathbb{L}^{AB}]\ P[\mathsf{ab|cd} \in G \mid AC].$$

Further, note that $P[\mathsf{ac|bd} \in G \mid AC]$ and $P[\mathsf{ab|cd} \in G \mid AC]$ do not depend on the choice of the $\{a,b,c,d\}$ lineages, but only depend on the scenario $\mathcal{L}_r$ (see Figures 11 and 12 (right)). Hence,

$$E[\mathbb{G}^{AC}_{\mathsf{ac|bd}}] = E[\mathbb{L}^{AC}]\ P[\mathsf{ac|bd} \in G \mid AC];$$
$$E[\mathbb{G}^{AC}_{\mathsf{ab|cd}}] = E[\mathbb{L}^{AC}]\ P[\mathsf{ab|cd} \in G \mid AC]$$

Combining all of the above relations we have

$$\begin{aligned} E[\mathbb{G}^{AB}_{\mathsf{ab|cd}}] + E[\mathbb{G}^{AC}_{\mathsf{ab|cd}}] \geq & E[\mathbb{L}^{AB}]\ P[\mathsf{ac|bd} \in G \mid AC] \\ & + E[\mathbb{L}^{AC}]\ P[\mathsf{ab|cd} \in G \mid AC]; \\ E[\mathbb{G}^{AC}_{\mathsf{ac|bd}}] + E[\mathbb{G}^{AB}_{\mathsf{ac|bd}}] \leq & E[\mathbb{L}^{AC}]\ P[\mathsf{ac|bd} \in G \mid AC] \\ & + E[\mathbb{L}^{AB}]\ P[\mathsf{ab|cd} \in G \mid AC]. \end{aligned}$$

We can now conclude the proof by noting that $E[\mathbb{L}^{AB}] \geq E[\mathbb{L}^{AC}]$ (by Corollary 5.1) and $P[\mathsf{ac|bd} \in G \mid AC] > P[\mathsf{ab|cd} \in G \mid AC]$ (by Lemma 3.3).

5.1.4 Case $I = (cd,a,b) \vee (bd,a,c)$

This case is symmetric to $I = (ab,c,d) \vee (ac,b,d)$. Therefore, the proof is similar.

5.1.5 Case $I = (abc,d) \vee (abd,c) \vee (acd,b) \vee (bcd,a) \vee (abcd)$

All quartets in the locus tree under each of these scenarios are $\mathsf{ab|cd}$. Then, by Lemma 3.3, $P[\mathsf{ab|cd} \in G \mid I] > P[\mathsf{ac|bd} \in G \mid I]$ for each of the $\mathbb{L}^I_{\mathsf{ab|cd}}$ quartets. Therefore,

$$E[\mathbb{G}^I_{\mathsf{ab|cd}}] > E[\mathbb{G}^I_{\mathsf{ac|bd}}].$$

5.1.6 Case $I = (ad,bc) \vee (ad,b,c) \vee (bc,a,d)$

All quartets in the locus tree under each of these scenarios are $\mathsf{ad|bc}$. It is then not difficult to see that $E[\mathbb{G}^I_{\mathsf{ab|cd}}] = E[\mathbb{G}^I_{\mathsf{ac|bd}}]$.

5.2 Caterpillar species tree

We now briefly discuss the proof strategy for Theorem 5.2 when S is a caterpillar. Similarly to Section 4.2, we condition the duplication/loss process on a fixed number of ABC-lineages (l) – see Figure 13. Adapting a similar notation to Section 5.1, let $\mathcal{A}_i, \mathcal{B}_i, \mathcal{C}_i$ denote the random variables for the number of a, b, and c genes, respectively, below the i-th ABC-lineage (in the locus tree). Further, let $\mathcal{D}$ denote the total number of d leaves. It is then not difficult to show that $\mathcal{D}$ is independent from $\mathcal{X}_i$ for any $\mathcal{X} \in \{\mathcal{A}, \mathcal{B}, \mathcal{C}\}$. Further, $\mathcal{A}_i$ and $\mathcal{B}_i$ are independent from $\mathcal{C}_j$ for any $i,j \in \{1,2,\ldots,l\}$ (analogously to Claim 5.2). Claim 5.3 also upholds when we restrict $\mathcal{X}$ to $\{\mathcal{A}, \mathcal{B}, \mathcal{C}\}$. Finally, Lemma 5.1 is applicable in the caterpillar case as well; i.e., $E[\mathcal{A}_1\,\mathcal{B}_1] \geq E[\mathcal{A}_1]E[\mathcal{B}_1]$.

We now need to consider the following scenarios: $I \in \{(a,b,c); (ab,c) \vee (ac,b); (bc,a); (abc)\}$ and prove that $\mathbb{G}^I_{\mathsf{ab|cd}} \geq \mathbb{G}^I_{\mathsf{ac|bd}}$ for all such I. It is then not difficult to do so, since $I = (a,b,c)$ is analogous to Case 5.1.1 from Section 5.1, $I = (ab,c) \vee (ac,b)$ is analogous to Case 5.1.3, $I = (bc,a)$ is analogous to Case 5.1.6, and $I = (abc)$ is analogous to Case 5.1.5. Further, under $I = (abc)$ the inequality $\mathbb{G}^I_{\mathsf{ab|cd}} > \mathbb{G}^I_{\mathsf{ac|bd}}$ is strict, similarly to Case 5.1.5. That is, Theorem 5.2 holds.

6 Conclusion

For the first time, we investigated and established statistical properties of a popular species tree inference method under the powerful duplication-loss-coalescence model. We proved that two natural versions of ASTRAL (adapted for the duplication-loss shaped gene families) are statistically consistent under DLCoal. Our result reinforces the practical value of ASTRAL and other quartet-based methods in the area of evolutionary inference. In addition to our work, Hill et al. (Hill *et al.*, 2020) studied the rate of convergence of ASTRAL under DLCoal. In the future, we anticipate that other statistically consistent methods under DLCoal will be discovered, and the methods will be compared based on their theoretical rate of convergence and simulation studies, advancing the accuracy of evolutionary inference.

7 Acknowledgments

We would like to thank two anonymous reviewers for their excellent work and comments that helped improve the manuscript significantly. This material is based upon work supported by the National Science Foundation under Grant No. 1617626. During the revision and editing, AM was funded by the USDA Agricultural Research Service Research Participation Program of the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and USDA Agricultural Research Service (contract number DE-AC05-06OR23100). Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA, DOE, or ORISE. USDA is an equal opportunity provider and employer.

References

Allman, E. S., Degnan, J. H., and Rhodes, J. A. (2011). Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *Journal of mathematical biology*, **62**(6), 833–862.

Allman, E. S., Degnan, J. H., and Rhodes, J. A. (2016). Species tree inference from gene splits by unrooted star methods. *IEEE/ACM transactions on computational biology and bioinformatics*, **15**(1), 337–342.

Allman, E. S., Degnan, J. H., and Rhodes, J. A. (2018). Split probabilities and species tree inference under the multispecies coalescent model. *Bulletin of mathematical biology*, **80**(1), 64–103.

Arvestad, L., Berglund, A.-C., Lagergren, J., and Sennblad, B. (2003). Bayesian gene/species tree reconciliation and orthology analysis using mcmc. *Bioinformatics*, **19**(suppl1), 7–15.

Bininda-Emonds, O. R., editor (2004). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, volume 4 of *Computational Biology*. Springer Verlag.

Degnan, J. H., DeGiorgio, M., Bryant, D., and Rosenberg, N. A. (2009). Properties of consensus methods for inferring species trees from gene trees. *Systematic Biology*, **58**(1), 35–54.

Du, P., Hahn, M. W., and Nakhleh, L. (2019). Species tree inference under the multispecies coalescent on data with paralogs is accurate. *bioRxiv*.

Ewing, G. B., Ebersberger, I., Schmidt, H. A., and Von Haeseler, A. (2008). Rooted triple consensus and anomalous gene trees. *BMC evolutionary biology*, **8**(1), 118.

Hill, M., Legried, B., and Roch, S. (2020). Species tree estimation under joint modeling of coalescence and duplication: sample complexity of quartet methods. *arXiv preprint arXiv:2007.06697*.

Kubatko, L. S., Carstens, B. C., and Knowles, L. L. (2009). Stem: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*, **25**(7), 971–973.

Larget, B. R., Kotha, S. K., Dewey, C. N., and Ané, C. (2010). BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics*, **26**(22), 2910–2911.

Legried, B., Molloy, E. K., Warnow, T., and Roch, S. (2020). Polynomial-time statistical estimation of species trees under gene duplication and loss. In R. Schwartz, editor, *Research in Computational Molecular Biology*, pages 120–135, Cham. Springer International Publishing.

Liu, L. and Yu, L. (2011). Estimating Species Trees from Unrooted Gene Trees. *Systematic Biology*, **60**(5), 661–667.

Liu, L., Yu, L., Pearl, D. K., and Edwards, S. V. (2009). Estimating species phylogenies using coalescence times among sequences. *Systematic Biology*, **58**(5), 468–477.

Liu, L., Yu, L., and Edwards, S. V. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC evolutionary biology*, **10**(1).

Mossel, E. and Roch, S. (2008). Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **7**(1), 166–171.

Rannala, B. and Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using dna sequences from multiple loci. *Genetics*, **164**(4), 1645–1656.

Rasmussen, M. D. and Kellis, M. (2012). Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome research*, **22**(4), 755–765.

Rhodes, J. A. (2019). Topological metrizations of trees, and new quartet methods of tree inference. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical population biology*, **26**(2), 119–164.

Vachaspati, P. and Warnow, T. (2015). Astrid: accurate species trees from internode distances. *BMC genomics*, **16**(S10).

Yourdkhani, S. and Rhodes, J. A. (2020). Inferring metric trees from weighted quartets via an intertaxon distance. *Bulletin of Mathematical Biology*, **82**(7), 1–22.

Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. (2018). Astral-iii: polynomial time species tree reconstruction from partially resolved gene trees. *BMC bioinformatics*, **19**(6), 153.