Phylogenetics

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

Alexey Markin¹,* and Oliver Eulenstein²

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

* To whom correspondence should be addressed.

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., 2011). 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; Youzdihani 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.

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 topology \( P \) as \((T, \omega)\). Here \( T \) is the tree topology and is a binary rooted tree with the designated root vertex, \( \mu(T) \), of degree two, all internal nodes of degree three, and with leaves bijectively labeled by elements of set \( \operatorname{Le}(T) \). 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, \( \omega \) 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, \( \omega \) : \( E(T) \rightarrow Q^+ \). 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 \( \omega_P \) 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 it internal. Further, given a set \( Y \subset \operatorname{Le}(T) \), tree topology \( T|_Y \) is obtained from \( T \) by restricting the leaf-set to \( Y \). A restricted phylogenetic tree \( P|_Y \) 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 \( Y \)).

A rooted tree 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), (middle) and (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., \( T_{P|_{\{a,b,c,d\}}}) \) 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 \( \lambda \) and death (loss) rate \( \mu \) (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.

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

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

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

\[
\begin{align*}
&\text{Fig. 3: An alternative depiction of a locus tree from Figure 2 with red dots representing duplication-points slightly shifted towards a novel locus.}
\end{align*}
\]

\[
\begin{align*}
&\text{Fig. 4: An example of a locus tree that illustrates the b-MSC constraints for Section 2.3.}
\end{align*}
\]

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 coalescent 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):

\[
g_{i,j}(x) = \sum_{k=j}^{i} \binom{i}{j} \frac{(k-1)(-1)^{k-j}}{j!(k-j)!} \prod_{m=0}^{k-1} \frac{(j+m)(i-m)}{i+m}.
\]

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 \( \Delta t \) time. That is, the process is symmetric.

2.3 Bounded MSC (b-MSC) model

The constraints on MSC in the unified DLCcoal model appear due to the duplication points. In particular, all lineages originating below a daughter duplicate must coalesce below the respective duplication node.
For example, in Figure 3, the gene tree lineages corresponding to leaves 2, 3, 4 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 \( a \) and \( b \), which entered the same edge below a duplication \( d \) at time \( d \), coalesce within time \( x \) is as follows (see Figure 3 for a respective locus tree example):

\[
P[a, \ b \ coalesced \ within \ time \ x \ | \ a, \ b \ coalesced \ prior \ to \ duplication] = 1 - e^{-x}
\]

where \( P[a, b \ coalesced \ prior \ to \ duplication] \) 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 these cases.

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

#### 3.1 L is balanced

For convenience, we set \( x := 2 \omega(X) \), \( y := 2 \omega(Y) \) 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,

\[
P[ab|cd \in G] = 1, \quad P[ac|bd \in G] = 1, \quad P[ad|bc \in G] = 1, \quad P[abc|d \in G] = 1
\]

##### 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,

\[
P[ab|cd \in G] = 1; \quad P[ac|bd \in G] = P[ad|bc \in G] = 0.
\]

##### 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.

\[
P[ab|cd \in G] = P[ab|cd \in G | a, b, c, d \ coalesced \ before \ z] \frac{1}{P[a, b, c, d \ coalesced \ before \ z]}
\]

\[
= 1 - P[a, b, c, d \ coalesced \ before \ z] \frac{1}{P[a, b, c, d \ coalesced \ before \ z]}
\]

\[
= 1 - \frac{1}{3} e^{-(x+y)} \frac{\frac{4}{3} e^{-(x+y)} P[4 \ lines \ coalesced \ within \ time \ t]}{P[a, b, c, d \ coalesced \ before \ z]}
\]

\[
= 1 - \frac{1}{3} e^{-(x+y)} \frac{\frac{4}{3} e^{-(x+y)} P[4 \ lines \ coalesced \ within \ time \ t]}{P[a, b, c, d \ coalesced \ before \ z]}
\]

\[
= \frac{1}{3} e^{-(x+y)} \frac{\frac{4}{3} e^{-(x+y)} P[4 \ lines \ coalesced \ within \ time \ t]}{P[a, b, c, d \ coalesced \ before \ z]}
\]
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 \( z := \omega_L(X), \ y := \omega_L(Y) \) 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,

\[
P[ab|cd \in G] = \frac{2}{3} e^{-z};
\]

\[
P[ac|bd \in G] = \frac{1}{3} e^{-z}.
\]

#### 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

\[
P[ab|cd \in G] = \frac{1}{4};
\]

\[
P[ac|bd \in G] = \frac{1}{4} e^{-z};
\]

\[
P[ac|bd \in G] = \frac{1}{4} e^{-z}.
\]

#### 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).

\[
P[ac|bd \in G] = \frac{1}{4} e^{-z};
\]

\[
P[ac|bd \in G] = \frac{1}{4} e^{-z}.
\]

### 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 \( P[ab|cd \in G] \) 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 \( d_3 \) on \( L_1 \) always has a
counterpart duplication-point $d_2$ on $L_2$ with the same distance to the root and vice versa (see Figure 7). Then
\[ P_{x}(ab|cd \in G \mid L_1) = P_{x}(ab|cd \in G \mid L_2) = 1, \]
if $L_1$ (and $L_2$) have a duplication-point on edge $X$, and
\[ P_{x}(ab|cd \in G \mid L_1) = P_{x}(ab|cd \in G \mid L_2), \]
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 $P_{x}(ab|cd \in G \mid L) > P_{x}(ac|bd \in G \mid L) = P_{x}(ad|bc \in G \mid L)$.

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 $G$ be a collection of gene trees that independently evolved within $S$ according to the DLCoal process. Then, as the number of trees in $G$ goes to infinity, the probability that $T$, 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 $P_{x}(ab|cd \in G \mid L) > P_{x}(ac|bd \in G \mid L) = P_{x}(ad|bc \in G \mid L)$.

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
\[ P_{x}(ab|cd \in G \mid LR = l) > P_{x}(ac|bd \in G \mid RL = l) = P_{x}(ad|bc \in G \mid RL = l) \]
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 $i_b, i_c, i_d$ 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 $(a, c, d)$ for that scenario, we write $(a, b, d)$ to denote the scenario where all $i_a$ are distinct. Then, by the law of total probability, we have
\[ P_{x}(ab|cd \in G) = \sum_{I} P[I, ab|cd \in G], \]
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 $I \in \{(a, b, c, d), (ab, cd) \cup (ac, bd) \cup (cd, a, b) \cup (ac, bd) \cup (bd, a, c) \cup (bd, c) \cup (acd, b) \cup (bc, ad) \cup (bd, c) \cup (bd, a) \cup (ac, bd); (ad, bc) \cup (ab, d) \cup (bc, a, d) \}$. Observe that we consider all possible scenarios simultaneously here.

Our goal is to prove that $P_{x}(ab|cd \in G) > P_{x}(ac|bd \in G)$. Note that $P_{x}(ad|bc \in G) = P_{x}(ac|bd \in G)$ follows from the fact that swapping $e$
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 
\[ P[A \cup B | I] \geq P[A \cup B | I'] \]
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, \( i_x \) can be dependent on \( i_y \) and \( i_z \) can be dependent on \( i_y \).

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 \( I \)), 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 to (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 \( i_x \) 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 \( i_x \). The same is also true for all other pairs of \( x \in \{a, b\} \) and \( y \in \{c, d\} \).

However, since \( s \) and \( b \) develop (at least partially) in the same species tree branch random variables \( i_x \) and \( i_y \) can be dependent. Similarly for \( i_x \) and \( i_z \).

Observation 4.2. Due to the symmetry of the duplication/loss process, we have 
\[ P[i_x = k] = \frac{1}{l} \]
for any \( x \in \{a, b, c, d\} \) and \( k \in \{1, 2, \ldots, l\} \). Then, by Claim 4.1,
\[ P[i_x = i_y] = \sum_{k=1}^l P[i_x = k] P[i_y = k] = \frac{1}{l^2} = \frac{1}{l} \]
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 \( P[i_a = i_c] \) 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 process. Hence, we have 
\[ P[ab|cd \in G \mid I] = P[ac|bd \in G \mid I] \] 
and at least in one case the strict inequality holds.

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, \( P[ab|cd \in G \mid (ab, cd)] = P[ac|bd \in G \mid (ac, bd)] = 1 \).

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

We need to show that
\[ P[ab|cd \in G \mid I] = P[ab|cd \in G \mid (ab, cd)] P[ab|cd] + P[ab|cd \in G \mid (ac, bd)] P[ac|bd] \geq P[ac|bd \in G \mid (ab, cd)] P[ab|cd] + P[ac|bd \in G \mid (ac, bd)] P[ac|bd] = P[ac|bd \in G \mid I] \].

Observe the following.

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

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 \( a, b, c, d \).

4.1.4 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 process.
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 at the parent of $C$ and $D$. See Figure 10 for an example of $N_i$ variables. By $N = (N_1, \ldots, N_l)$ and $M = (M_1, \ldots, M_l)$ we denote the vectors of $N_i$ and $M_i$ variables, respectively.

Observe that $P[\{ab, cd\}] = P[\{a, b, c, d\}] = P[\{abd\}]$ and $P[\{ac, bd\}] = P[\{a, c, b, d\}] = P[\{acd\}]$. Further, note that when conditioned on specific values of $N$ and $M$, $i_a = i_b$, $i_c = i_d$, and $i_{ab} = i_{cd}$ 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 $N$ and $M$, 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 $N_i$ lineages independently and uniformly at random. Similarly, $c$ and $d$ choose one of the $M_i$ lineages independently and uniformly at random.

Then, for fixed values of the $N$ and $M$ vectors we have

$$P[i_a = i_b, i_c = i_d \mid N, M] = \frac{1}{N_l} \sum_{j=1}^l P[i_a = j \mid N] P[i_b = j \mid N]$$

$$= \frac{1}{N_l} \sum_{j=1}^l (N_j)^2 \left( \sum_{j=1}^l N_j \right)^2$$

The last equality is due to $P[i_a = j \mid N] = \frac{N_j}{N_l}$. 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 species $A$ and $B$. Similar relations can then be easily derived for $i_a, i_b, i_c, i_d$.

Further, following the same idea, we have

$$P[i_a = i_b, i_c = i_d \mid N, M] = \frac{1}{N_l} \sum_{j=1}^l (N_j M_j) \left( \sum_{j=1}^l N_j \right) \left( \sum_{j=1}^l M_j \right)$$

Then, by Cauchy-Schwarz, $(\sum_{j=1}^l (N_j M_j)^2) \leq \left( \sum_{j=1}^l (N_j)^2 \right) \left( \sum_{j=1}^l (M_j)^2 \right)$ and therefore $P[i_a = i_b, i_c = i_d \mid N, M] \geq P[i_a = i_b, i_c = i_d \mid N, M]$ for any realization of vectors $N$ and $M$. That is, $P[\{ab, cd\}] \geq P[\{ac, bd\}]$.

Using the above results, we have

$$P[\{ab, cd\} \in G \mid I] \geq P[\{ac, bd\} \in G \mid I] \geq P[\{ac, bd\} \in G]$$

4.1.3 Case 1 = (ab, cd) ∨ (cd, ab, bd) ∨ (bd, ac, cd)

For convenience, from now on we denote the event $(ab, cd) \lor (cd, ab, bd)$ by $AB$ and the event $(ac, bd) \lor (bd, ac, cd)$ by $AC$.

We prove that

$$P[\{ab, cd\} \in G \mid I] = P[\{ab, cd\} \in G \mid AB] P[AB] + P[\{ab, cd\} \in G \mid AC] P[AC] \geq P[\{ac, bd\} \in G \mid AB] P[AB] + P[\{ac, bd\} \in G \mid AC] P[AC] = P[\{ac, bd\} \in G \mid I]$$

Consider the following results.

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

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 $L_r$ 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 $(ab, cd, d)$, we have $i_a = i_b = 1, i_c = 2$, and $i_d = 3$. In case $(cd, ab, b)$ we assume $i_a = i_d = 1, i_c = 2$, and $i_b = 3$. Similarly, under $(ac, bd, d)$ we assume $i_a = i_b = 1, i_c = 2$, and $i_d = 3$ and under $(bd, ac, a)$ we assume that $i_a = i_d = 1, i_b = 2, i_c = 3$. Then, a fixed $L_r$ scenario forces the same "top" structure of the locus trees in all four cases.

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

Then, Figures 11 and 12 depict two possible topologies of the $L_r$ 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 $L_{(ab, cd, d)}$ and $L_{(ac, bd, a)}$ are both cattallars, while in Figure 12, the locus trees are both balanced. This separation is achieved because we condition on a fixed $L_r$ scenario.

We now consider these two cases individually.

(i) $L_{(ab, cd, d)}$ and $L_{(ac, bd, a)}$ are cattallars (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 $(ab, cd, d)$ 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, $P[\{ab, cd\} \in G \mid AB, L_r] \geq P[ac, bd \in G \mid AC, L_r]$. Let $L_r$ be the distance (in coalescent units) from the root speciation event to the divergence of $a$ and $b$ in the locus tree under the $(ab, cd, d)$ case (as shown on the figure). Note that $x_{ab} \geq 0$. There are two cases to consider.

• No duplications along the $x_{ab}$ lineage. Then, as shown in Section 3.2.2, $P[\{ab, cd\} \in G \mid AB, L_r] \geq P[ac, bd \in G \mid AC, L_r]$. Let $L_r$ be the distance (in coalescent units) from the root speciation event to the divergence of $a$ and $b$ in the locus tree under the $(ab, cd, d)$ case (as shown on the figure). Note that $x_{ab} \geq 0$. There are two cases to consider.

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

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

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

$$2P[ab, cd \in G \mid AC] + P[ac, bd \in G \mid AC] = 1$$

$$2P[ac, bd \in G \mid AB] + P[ab, cd \in G \mid AB] = 1$$

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


Downloaded from https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btab414/6287614 by National Animal Disease Centre user on 04 June 2021
Fig. 11: Caterpillar locus trees $L_{(a,b,c,d)}$ (left) and $L_{(a,b,d)}$ (right) embedded into the species tree. The red circles represent the potential duplication locations that could influence the gene tree probabilities. Note that the $L_{(a,b,c,d)}$ 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 $L_{(a,b,c,d)}$ (left) and $L_{(a,b,d)}$ (right) embedded into the species tree.

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

Summarizing the above results we have.

\[ P(ab|cd \in G | AB] P[AB] + P(ab|cd \in G | AC] P[AC] \]
\[ \geq P(ac|bd \in G | AC] P[AB] + P(ac|bd \in G | AB] P[AC] \]
\[ \geq P(ac|bd \in G | AC] P[AC] + P(ac|bd \in G | AB] P[AB]. \]

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.

4.1.4 Case $I = (abc, d) \lor (abd, c) \lor (acd, b) \lor (bc, a, d)$. In all five cases locus tree $L$ displays the quartet $abc|d$. Therefore, by Lemma 3.3 $P[ab|cd \in G | I] = P[ac|bd \in G | I]$. Observe that we obtain the strict inequality in this case.

4.1.5 Case $I = (ad, be) \lor (ad, c) \lor (bc, a, d)$.

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

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, $\mathcal{I} = \{(a, b, c); (a, b); (b, c); (a, c); (abc)\}$ scenarios describe relations between $i_a$, $i_b$, and $i_c$.

We now prove that $P[ab|cd \in G | I] \geq P[ac|bd \in G | I]$ for all $I$ in $\{(a, b, c); (ab, c); (bc, a); (abc)\}$. 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 duplication/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 $P[ab|cd \in G | I] = P[ac|bd \in G | I] = P[ad|bc \in G | I]$.

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

4.2.2 Case $I = (a, b, c) \lor (ac, b)$.

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


Proof. It is sufficient to show that $P[i_a = i_b] \geq P[i_a = i_c]$. By Claim 4.2, $P[i_a = i_b] = 1/l$. 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) \( P[ab|cd \in G \mid \{ab, c\}] \geq P[ac|bd \in G \mid \{ac, b\}] \);

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

(iii) \( P[ac|bd \in G \mid \{ac, b\}] \geq P[ac|bd \in G \mid \{ab, c\}] \).

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

\[
P[ab\mid cd \in G \mid I] = P[ab\mid cd \in G \mid \{ab, c\}] P[(ab, c)]
\]

\[
+ P[ab\mid cd \in G \mid \{ac, b\}] P[(ac, b)]
\]

\[
\geq P[ac\mid bd \in G \mid \{ac, b\}] P[(ab, c)]
\]

\[
+ P[ac\mid bd \in G \mid \{ab, c\}] P[(ac, b)]
\]

\[
\geq P[ac\mid bd \in G \mid \{ac, b\}] P[(ab, c)]
\]

\[
+ P[ac\mid bd \in G \mid \{ab, c\}] P[(ac, b)]
\]

\[
= P[ac\mid bd \in G \mid I].
\]

4.2.3 Case I = \{(bc), a\}.

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

4.2.4 Case I = \{(abc)\}.

The locus tree displays quartet abc|d; therefore, by Lemma 3.3 and the law of total probability, we have \( P[ab\mid cd \in G \mid I] \geq P[ac\mid bd \in G \mid I] \).

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 \( S = (T_G, w_G) \) be a fixed species tree and let \( G \) be a collection of gene trees that independently evolved within \( S \) according to the DRCaD process. Then, as the number of trees in \( G \) goes to infinity, the unrooted tree estimate by ASTRAL-multi converges almost surely to \( T_G \).

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 DRCaD process. Let \( G_{ab|cd} \) (respectively \( G_{ac|bd} \) and \( G_{ad|bc} \)) be the number of \( ab|cd \) (respectively \( ac|bd \) and \( ad|bc \)) 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. \( E[G_{ab|cd}] \geq \max\{E[G_{ac|bd}], E[G_{ad|bc}]\} \).

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 \( E[G_{ab|cd}] \geq E[G_{ac|bd}] \). The general structure of the proof is the same due to the 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 \( G_{ab|cd} \) notation, we let \( L_{ab|cd} \) (respectively \( L_{ac|bd} \) and \( L_{ad|bc} \)) denote the number of \( ab|cd \) (respectively \( ac|bd \) and \( ad|bc \)) quartets in the locus tree \( L \). Further, for a fixed scenario \( I \) (e.g., scenario \( i_1 = i_2 = 1, i_3 = \ldots, i_l = 3 \)), let \( L^I_{ab|cd} \) be the number of \( ab|cd \) quartets in the locus tree that follow the scenario \( I \). Further, \( G^I_{ab|cd} \) be the number of \( ab|cd \) quartets in \( G \) that ‘appeared’ from one of the \( L^I_{ab|cd} \) quartets. Similarly we define \( L^I_{ac|bd}, L^I_{ad|bc}, G^I_{ac|bd}, G^I_{ad|bc} \).

Consider any \( I \neq (ab|cd) \). Note that the root of locus tree \( L^I_{(ab|cd)} \) must be a duplication for such \( I \) (because \( I \) involves at least two root lineages). Then, if \( I \) always defines balanced quartets, we have \( G^I_{ab|cd} = L^I_{ab|cd} \) for any \( I \neq (ab|cd) \) by Lemma 3.1. In particular, we note the following:

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

\[
G^I_q = L^I_q.
\]

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 \( L_q^I \) notation. For example, we write \( L_{ab|cd}^I \) instead of \( L_{ab|cd}^{I|ab|cd} \) since \( ab|cd \) is the only type of quartets that can appear under scenario \( (ab|cd) \).

Given a fixed root lineage \( i \) and let \( A_i \) be the random variable denoting the number of \( a \) leaves generated by that lineage. Similarly, we define random variables \( B_i, C_i, D_i \). By symmetry, \( E[A_1] = E[A_2] = \ldots = E[A_l] \) (with similar relations holding for \( B_i, C_i, D_i \)). Then, observe the following:

Observation 5.2. Since the duplication/loss process runs independently in the parallel branches of the species tree, \( X_i \) is independent from \( Y_j \) for any \( X \subseteq \{A, B, C, D\} \) and \( i, j \in \{1, \ldots, l\} \).

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

\[
E[X_i] = E[X_j]
\]

for all \( X \subseteq \{A, B, C, D\} \) and \( i, j \in \{1, \ldots, l\} \).

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

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

\[
E[A_i] \geq E[A_1] E[B_1];
\]

\[
E[C_i] \geq E[C_1] E[D_1];
\]

\[
E[I]\sum_{t=1}^{l} t_i = 1, i_2 = 2, i_3 = \ldots, i_l = 3\]

We now outline several key corollary statements.

Corollary 5.1.

\[
E[L^I_{(ab|cd)}] \geq E[L^I_{(ac|bd)}];
\]

\[
E[L^I_{(ab|cd)}] \geq E[L^I_{(ac|bd)}];
\]

\[
E[L^I_{(ab|cd)}] \geq E[L^I_{(ac|bd)}];
\]

\[
E[I]\sum_{t=1}^{l} t_i = 1, i_2 = 2, i_3 = \ldots, i_l = 3\]

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

\[
E[I]\sum_{t=1}^{l} t_i = 1, i_2 = 2, i_3 = \ldots, i_l = 3\]

The other two relationships can be established similarly.
We now consider the following comprehensive set of scenarios: \( I \in \{ (a, b, c, d), (ab, cd) \cup (ac, bd), (ab, d) \cup (ac, b) \cup (bd, a) \cup (cd, b) \cup (bd, c) \cup (cd, a) \} \). For each \( I \) we will prove that \( E[G_{\text{abcd}}^{I}] \geq E[G_{\text{abcd}}^{I}] \) and for at least one \( I \) the strict inequality holds.

### 5.1 Case \( I = (a, b, c, d) \)

By the symmetry of the duplication-loss process in the root edge, we have
\[
E[G_{(a,b,c,d)}^{I}] = E[G_{(a,b,c,d)}^{I}] = E[G_{(a,b,c,d)}^{I}].
\]

### 5.2 Case \( I = (ab, cd) \cup (ac, bd) \)

By Claim 5.1, \( G_{(ab,cd)}^{I} = 1 \) if \( d \) and \( G_{(ac,bd)}^{I} = 1 \) if \( b \) and \( c \). Then, combining this with Corollary 5.1, we have
\[
E[G_{(ab,cd)}^{I}] = E[G_{(ac,bd)}^{I}] \geq E[G_{(ac,bd)}^{I}] = E[G_{(ac,bd)}^{I}].
\]

### 5.3 Case \( I = (ab, c, d) \cup (ac, b, d) \)

Consider a fixed duplication-loss scenario, \( L_{r} \), in the root edge of \( S \). In this section we implicitly condition the probability space on \( L_{r} \). That is, we prove that \( E[G_{(ab,cd)}^{I}] \geq E[G_{(ac,bd)}^{I}] \) for each \( L_{r} \).

Due to symmetry, we consider the following two core scenarios: \( AB: (i_{1} = i_{2} = 1, i_{3} = 2, i_{4} = 3) \) and \( AC: (i_{1} = i_{2} = 1, i_{3} = 2, i_{4} = 3) \). It is then sufficient to show the following:

**Lemma 5.2.**

\[
E[G_{(ab,cd)}^{I}] + E[G_{(ac,bd)}^{I}] \geq E[G_{(ab,cd)}^{I}] + E[G_{(ac,bd)}^{I}].
\]

**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 \( P(ab|bd) \in G \setminus I \) \( \geq P(ac|bd) \in G \setminus I \). Similarly, \( P(ac|bd) \in G \setminus I \) \( \geq P(ac|bd) \in G \setminus I \).

Therefore, we have
\[
E[G_{(ab,cd)}^{I}] \geq E[G_{(ac,bd)}^{I}] \geq E[G_{(ac,bd)}^{I}] \geq E[G_{(ab,cd)}^{I}].
\]

Further, note that \( P(ac|bd) \in G \setminus I \) \( \geq P(ac|bd) \in G \setminus I \) and \( P(ab|bd) \in G \setminus I \) \( \geq P(ab|bd) \in G \setminus I \) do not depend on the choice of \( \{a, b, c, d\} \) lineages, but only depend on the scenario \( L_{r} \) (see Figures 11 and 12 (right)). Hence,
\[
E[G_{(ab,cd)}^{I}] = E[L_{ac}] \geq E[P(ac|bd) \in G \setminus I],
\]
\[
E[G_{(ac,bd)}^{I}] = E[L_{ab}] \geq E[P(ab|bd) \in G \setminus I],
\]

Combining all of the above relations we have
\[
E[G_{(ab,cd)}^{I}] + E[G_{(ac,bd)}^{I}] \geq E[G_{(ab,cd)}^{I}] + E[G_{(ac,bd)}^{I}] \geq E[L_{ac}] \geq E[P(ac|bd) \in G \setminus I],
\]
\[
E[G_{(ac,bd)}^{I}] + E[G_{(ab,cd)}^{I}] \geq E[G_{(ac,bd)}^{I}] + E[G_{(ab,cd)}^{I}] \geq E[L_{ab}] \geq E[P(ab|bd) \in G \setminus I].
\]

We can now conclude the proof by noting that \( E[L_{ab}] \geq E[L_{ac}] \) (by Corollary 5.1) and \( P(ac|bd) \in G \setminus I \) \( \geq P(ab|bd) \in G \setminus I \) (by Lemma 3.3).

### 5.4 Case \( I = (cd, a, b) \cup (bd, a, c) \)

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

### 5.5 Case \( I = (abc, d) \cup (abd, c) \cup (acd, b) \cup (bcd, a) \cup (abcd) \)

All quartets in the locus tree under each of these scenarios are \( ab|cd \). Then, by Lemma 3.3, \( P(ab|cd) \in G \setminus I \) \( \geq P(ac|bd) \in G \setminus I \) for each of the \( L_{ab|cd} \) quartets. Therefore,
\[
E[G_{(abcd)}^{I}] \geq E[G_{(abcd)}^{I}].
\]

### 5.6 Case \( I = (ad, bc) \cup (ab, d, c) \cup (bc, a, d) \)

All quartets in the locus tree under each of these scenarios are \( ab|bc \). It is then not difficult to see that \( E[G_{(abcd)}^{I}] \geq E[G_{(abcd)}^{I}].

### 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 (I) – see Figure 13. Adapting a similar notation to Section 5.1, let \( A_{c}, B_{c}, C_{c} \) denote the random variables for the number of \( a, b, c \) genes, respectively, below the \( c \)-th ABC-lineage (in the locus tree). Further, let \( D \) denote the total number of \( d \) leaves. It is then not difficult to show that \( D \) is independent from \( X_{i} \) for any \( X_{i} \in \{A_{c}, B_{c}, C_{c}\} \). Further, \( A_{c} \) and \( B_{c} \) are independent from \( C_{c} \) for any \( i, j \in \{1, 2, \ldots, l\} \) (analogously to Case 5.2). Claim 5.3 also upholds when we restrict \( X_{i} \) to \( \{A_{c}, B_{c}, C_{c}\} \). Finally, Lemma 5.1 is applicable in the caterpillar case as well, i.e., \( E[A_{c}B_{c}] \geq E[A_{c}]E[B_{c}] \).

We now need to consider the following scenarios: \( I \in \{ (a, b, c), (ab, d) \cup (ac, b) \cup (bd, a) \cup (cd, b) \cup (bd, c) \cup (cd, a) \} \) and prove that \( G_{\text{abcd}}^{I} \geq G_{\text{abcd}}^{I} \) for all such \( I \). It is then not difficult to do so, since \( I = (a, b, c) \) is analogous to Case 5.1 from Section 5.1. \( I = (ab, c) \cup (ac, b) \cup (bd, a) \cup (cd, b) \cup (bd, c) \cup (cd, a) \) is analogous to Case 5.1.3, \( I = (bc, a) \) is analogous to Case 5.1.6, and \( I = (ab) \) is analogous to Case 5.1.5. Further, under \( I = (ab) \) the inequality \( G_{\text{abcd}}^{I} \geq G_{\text{abcd}}^{I} \) 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


