The Annals of Applied Probability
2022, Vol. 32, No. 6, 4681–4705
https://doi.org/10.1214/22-AAP1799

SPECIES TREE ESTIMATION UNDER JOINT MODELING OF COALESCENCE AND DUPLICATION: SAMPLE COMPLEXITY OF QUARTET METHODS

BY MAX HILL[1,a], BRANDON LEGRIED[2,c] AND SEBASTIEN ROCH[1,b]

[1]*Department of Mathematics, University of Wisconsin–Madison,* [a]*bacharach@wisc.edu,* [b]*roch@wisc.edu*
[2]*Department of Statistics, University of Michigan, Ann Arbor,* [c]*blegried@umich.edu*

We consider species tree estimation under a standard stochastic model of gene tree evolution that incorporates incomplete lineage sorting (as modeled by a coalescent process) and gene duplication and loss (as modeled by a branching process). Through a probabilistic analysis of the model, we derive sample complexity bounds for widely used quartet-based inference methods that highlight the effect of the duplication and loss rates in both subcritical and supercritical regimes.

1. Introduction. Estimating phylogenies from the molecular sequences of existing species is a fundamental problem in computational biology that has been the subject of significant practical and theoretical work [16, 19, 52, 53, 57, 59, 60]. Rigorous statistical guarantees for inference methodologies often involve the probabilistic analysis of Markov models on trees. In particular, these analyses have uncovered deep connections with phase transitions in related statistical physics models [4, 7, 9, 15, 18, 29, 31–34, 37, 38, 44, 46].

In modern datasets, however, phylogeny estimation is confounded by heterogeneity across the genome from processes such as incomplete lineage sorting (ILS), gene duplication and loss (GDL), lateral gene transfer (LGT), and others [25]. Inferred trees depicting the evolution of individual loci in the genome are referred to as *gene trees*, while the tree representing the speciation history is called the *species tree*. Current sequencing technology allows phylogenetic estimates of species relationships for many genes, and a major challenge in reconstructing species trees is that gene trees often disagree for the reasons mentioned above. There is a burgeoning literature on the many ways of extracting speciation histories from collections of gene trees [12, 39, 51, 57, 59].

In this phylogenomic context, the design and analysis of species tree estimation methods require the use of a variety of stochastic processes beyond Markov models on trees, including coalescent processes [42], branching processes [5], random subtree prune-and-regraft operations [17, 24, 47], and tree mixtures [27, 35]. In fact, there is increasing realization in the phylogenetics community that ILS, GDL, LGT, etc. should not be studied in isolation [10, 49, 50, 55] and, as a result, there has been a push to consider more complex models that combine many sources of uncertainty and discordance [1, 3, 8, 23, 26, 28, 36, 43, 45, 48, 56]. We study here a joint coalescent and branching process unifying ILS and GDL, as introduced in [43].

Much is known about estimating species trees in the presence of ILS alone [41], as modeled by the multispecies coalescent (MSC) [42]. The latter model posits that, on a fixed species tree, gene trees evolve backwards in time on each branch according to the Kingman coalescent [20] (see Section 2 for more details). Bayesian approaches are a natural choice under such complex models of evolution [13]. However they do not scale well to large datasets

Received July 2020; revised February 2022.
MSC2020 subject classifications. 92D15.
Key words and phrases. Phylogenetics, gene duplication and loss, incomplete lineage sorting, statistical consistency.

and more computationally efficient procedures have been developed that combine inferred gene trees, sometimes referred to as summary methods [59].

One such method is to deduce the species tree by a plurality vote across gene trees. Unfortunately, that approach is not statistically consistent, that is, it may not converge to the true species tree as the number of gene trees grows to infinity. Indeed, for unrooted species trees with more than four species, the most frequently occurring gene tree topology need not coincide with that of the species tree [11]. However, for every 4-tuple of species—also referred to as quartet—and every locus, the most probable unrooted gene tree topology matches the species tree topology [2]. This implies in particular that the species tree topology is identifiable from the distribution of gene trees. That is, different species tree topologies necessarily produce different gene tree distributions. Further, quartet-based algorithms for combining gene trees [21, 30] are known to be statistically consistent. Tight bounds on their sample complexity, that is, how many gene trees are needed to recover the true species tree with high probability, have also been established [54].

Less is known about estimating species trees in the presence of GDL alone. The model in [5] posits that, on a fixed species tree, the number of copies in a gene family evolves forward in time on each branch according to continuous-time branching process [6] (see Section 2 for more details). Recently, the identifiability of the species tree in the presence of GDL alone was established in [22] by showing that, similar to the ILS alone case, for every quartet the most frequent unrooted gene tree topology matches that of the species tree. As a result, quartet-based inference methods [40] were also shown to be statistically consistent. To date, no sample complexity results have been derived under this GDL model however.

In this paper, we investigate the gene tree evolution model of [43], which unifies the multispecies coalescent and the branching process model of gene duplication and loss discussed above. Given that quartet-based methods have strong guarantees under these models separately, it is natural to consider their performance under the joint model as well (see Section 2 for more details on the methods). Numerical experiments in [14, 22] provide some evidence for the accuracy of certain quartet-based methods. In [26], the authors give a proof of statistical consistency for one such method. In our main result, we give the first known upper bounds on the sample complexity of species tree estimation methods under the joint effect of ILS and GDL. Our proof, which highlights the somewhat counter-intuitive role played by the duplication and loss rates in the supercritical regime (see Section 2), is complicated by the simultaneous forward-in-time/backward-in-time nature of the process.

The rest of the paper is organized as follows. In Section 2, after defining the model and inference methods, we state and discuss our results. In Section 3, we give a proof of identifiability including new quantitative estimates that play a role in our proof of sample complexity. The rest of the proof can be found in Section 4.

2. Background and main results. We first describe the model and then state our results formally.

2.1. *Problem and model.* Our input is a collection $\mathcal{T} = \{t_i\}_{i=1}^k$ of k multilabeled gene trees given without estimation error. We explain the terminology. The process of gene duplication creates multiple copies of a gene within the same individual in a species. We refer to these duplicated genomic segments as *gene copies* and we refer to collections of gene copies from different unrelated genes as *gene families*. A *gene tree* is a depiction of the parental lineages of a gene or multiple gene copies from individuals across several species. Roughly speaking, it shows the joint ancestral history of these related gene copies. In contrast, a *species tree* is a depiction of the evolutionary relationships of a group of species. Roughly speaking, it shows the sequence of speciation events that have produced the current

species. By *multilabeled*, we mean that each leaf of a gene tree is associated with a label from the set S of n species of interest and that the leaf labels need not be unique. That is, multiple leaves of a gene tree may be associated with the same species; this corresponds to observing multiple paralogous copies (i.e., which have arisen from gene duplication) of a gene within the genome. In practice, gene trees are estimated from the molecular sequences of the corresponding genomic segments using a variety of phylogenetic reconstruction methods [16, 19, 53, 57, 59, 60]. For the sake of our results, we assume that gene trees are provided *without estimation error* for a large number of gene families. Our main modeling assumption is that these gene trees have been drawn independently from a distribution for which some n-species tree T is taken as a fixed (i.e., nonrandom)—but unknown—parameter. Our goal is to output this unknown n-species tree T. We define the model more precisely next.

Model. We assume that the gene trees in $\mathcal{T}$ are independent and identically distributed, with each gene tree generated under the DLCoal model [43], a unified model of gene duplication, loss, and coalescence. The process for generating a gene tree under DLCoal involves two steps which are described below; an example realization of these steps is provided in Figure 1. In particular, we introduce yet another tree, a locus tree, which is an unobserved intermediate step in the model generating each gene tree. The process below is repeated independently for each $i = 1, \ldots, k$, thereby producing an independent gene tree t_i.

1. *Locus tree: Birth-death process of gene duplication and loss with daughter edges.* We fix a rooted n-species tree T with edges E directed from root to leaves (i.e., from top to bottom) and edge lengths $\{\eta_e\}_{e \in E}$. In Figure 1, this is the four-species tree on the top left. Note that the species tree is unknown to us.

Starting with a single ancestral copy of a gene at the root of T, a tree is generated by a top-down birth-death process [5] *within* the species tree. That is, on every edge in T, each gene copy independently duplicates at exponential rate $\lambda \geq 0$ and is lost at exponential rate $\mu \geq 0$. In addition, whenever a gene copy reaches a speciation event T, it bifurcates into two child copies, one for each of the descendant edges on T. Both duplications and speciations are indicated in the resulting locus tree by a bifurcation. The locus tree is then pruned of lost copies to give an (unobserved) rooted tree, which we refer to as *locus tree*. In this manner, we obtain a rooted n'-individual locus tree L with edge lengths. Species labels are associated to each leaf of L from the species set S. These steps for generating L are illustrated in the top-middle and top-right trees of Figure 1.

Furthermore, for each vertex in L which corresponds to a gene duplication event—but *not* vertices corresponding to speciation bifurcations—we distinguish between the two child edges by choosing one of them uniformly at random to be the *daughter* edge and the other to be the *mother* edge. This distinction plays an important role in the next step.

2. *Gene tree: Coalescent process on a locus tree.* Gene trees are generated by a backward-in-time (i.e., bottom-up) coalescent process [20, 42] within the locus tree L. The coalescent process begins with exactly one gene copy in each leaf of L. Copies at the bottom of a directed edge in the locus tree undergo the Kingman coalescent process for a time equal to the length of the directed edge. Edge lengths here are assumed to be in so-called coalescent time units, which means that each pair of lineages independently coalesces after an exponential time with mean 1, unless the end of the edge is reached first. Further, we condition on the event that all gene copies at the bottom of any *daughter* edge of the locus tree necessarily coalesces underneath the top of the edge. Continuing upward along ancestral edges of the locus tree, this process eventually yields a gene tree.

This process, which generates a gene tree from a locus tree, is termed the *multilocus coalescent* (MLC) in [43]. A realization of this coalescent process within a locus tree is illustrated in the bottom left of Figure 1, and the tree at the bottom right depicts the resulting gene tree without the overlying locus tree.

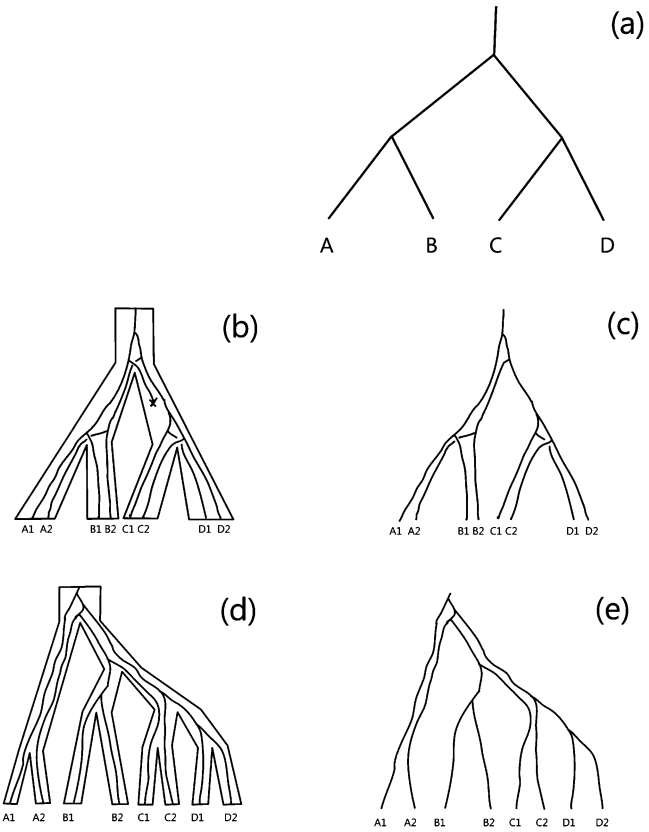

FIG. 1. *A species tree with balanced topology and leaf species A, B, C, D is given in* (*a*). *From* (*a*), *we obtain a realization of the gene duplication and loss* (*GDL*) *process with deaths not yet pruned. The pre-pruned locus tree and its generating species tree are shown in* (*b*). *Nonextant lineages in the tree are then pruned to obtain the locus tree*, *shown in* (*c*). *The gene tree obtained from the multispecies coalescent* (*MSC*) *process and its generating locus tree are shown in* (*d*). *The gene tree we actually observe is shown in* (*e*).

Intuitively, in Step 1 above the species tree T is a fat tree that "contains" the birth-death process which generates the skinny tree L. In Step 2, we forget about the species tree and "zoom in" on L so that L becomes the fat tree, with each edge of L containing one or more edges (or parts of edges) of the gene tree.

Species tree estimation methods. Next we describe two quartet-based species tree methods: ASTRAL-one and ASTRAL-multi [14, 30, 40]. Recall that a *quartet* refers to a 4-tuple of species. It is known that the unrooted topology of a species tree is entirely characterized by the collection of its quartet topologies (see, e.g., [53]), and hence a large number of quartet-based approaches to species tree estimation have been developed. Both ASTRAL-one and ASTRAL-multi are practical variants of an intuitive idea (which in the ILS/GDL context is motivated by the results of [2, 22]; see also Propositions 1 and 2 below): (1) for each quartet of species, find the most common topology across gene trees and (2) reconstruct an n-species tree that coincides with as many resulting quartet topologies as possible. The input is a collection of k multilabeled gene trees $\mathcal{T} = \{t_i\}_{i=1}^{k}$. Let S be the set of n species and Σ be the set of m labels (or gene copies). The tree t_i is labeled by the set $\Sigma_i \subset \Sigma$. For any species tree T labeled by S, the extended tree T_{ext} labeled by Σ is built by adding to each leaf of T all gene copies corresponding to that species as a polytomy (i.e., as a vertex with degree possibly higher than 3).

Under ASTRAL-one, we pick one gene copy of each species uniformly at random and restrict the gene tree to these copies, producing a new gene tree $\tilde{t}_i$. For any collection of gene

copies $\mathcal{J} = \{a, b, c, d\}$, let $T^{\mathcal{J}}$ be the restriction of any tree T to those copies. Then the quartet score of every candidate species tree $\widetilde{T}$ is

$$Q_k(\widetilde{T}) = \sum_{i=1}^{k} \sum_{\mathcal{J}=\{a,b,c,d\}\subset\Sigma_i} \mathbf{1}(\widetilde{T}^{\mathcal{J}}_{\text{ext}}, \tilde{t}_i^{\mathcal{J}}),$$

where $\mathbf{1}(T_1, T_2)$ is the indicator for the event that T_1 and T_2 have the same topology. ASTRAL-one selects the candidate tree $\widetilde{T}$ that maximizes that score.

ASTRAL-multi treats copies of a gene in a species as multiple alleles within the species. So, we do not replace t_i with any restricted gene tree $\tilde{t}_i$. The quartet score of T with respect to $\mathcal{T}$ is

$$Q_k(\widetilde{T}) = \sum_{i=1}^{k} \sum_{\mathcal{J}=\{a,b,c,d\}\subset\Sigma_i} \mathbf{1}(\widetilde{T}^{\mathcal{J}}_{\text{ext}}, t_i^{\mathcal{J}}).$$

The candidate tree $\widetilde{T}$ that maximizes this score is chosen by ASTRAL-multi.

Both procedures can be performed in exact mode or in a default constrained mode, which restricts the number of candidate species trees to those displaying the bipartitions in the given gene trees.

2.2. *Statement of main results.* We present a theoretical bound on the number of gene trees needed for ASTRAL-one to reconstruct the model species tree with high probability under the DLCoal model. Similar to the case of the MSC model [54], the sample complexity depends on the length of the shortest species tree branch in coalescent time units. We denote this length by f. However we also highlight the influence of other relevant parameters in the sample complexity: the depth of the species tree, Δ; the duplication and loss rates, λ and μ. For simplicity, we assume throughout that $\mu \neq \lambda$.

THEOREM 1 (Main result: Sample complexity of ASTRAL-one). *Consider a model species tree whose minimum branch length f is finite and assume gene trees are generated under the DLCoal model. Then, for any $\epsilon > 0$, there are universal positive constants C, C' such that the exact version of ASTRAL-one returns the true species tree with probability at least $1 - \epsilon$ if the number of input error-free gene trees satisfies*:

$$k \geq C' \frac{1}{f^2} \frac{e^{C|\mu-\lambda|\Delta}}{(1 - \frac{\lambda}{\mu} \wedge \frac{\mu}{\lambda})^C} \log \frac{n}{\epsilon}. \tag{1}$$

Somewhat surprisingly the subcritical ($\mu > \lambda$) and supercritical ($\mu < \lambda$) regimes exhibit a similar behavior. Indeed one naturally expects a higher sample complexity in the subcritical case as μ/λ becomes large because the absence of any gene copy in a species becomes more likely and leads to the need for more gene trees in order to extract a signal. That prediction is borne out in (1). However the sample complexity similarly increases in the supercritical regime as λ/μ becomes large. As the proof shows, the reasons for this behavior are different in that regime. They have to do with the fact that a large number of copies at the most recent common ancestor of a species quartet tends to produce large numbers of conflicting gene tree quartet topologies, thereby obscuring the signal. It is an open problem whether other inference methods (perhaps not based on quartets) are less sensitive to this last phenomenon.

Our proof of Theorem 1 involves a delicate probabilistic analysis of the DLCoal model. Along the way, we prove other results of interest. First, we show that the unrooted species tree is identifiable from the distribution of multilabeled gene trees $\mathcal{T}$ under the DLCoal model over T. Formally, we show that two distinct unrooted species trees produce different gene tree

distributions. The result is a generalization of [22], Theorem 1, where only GDL is considered. Theorem 2 was first claimed in [26]. Our novel contribution here lies in the proofs of Propositions 1 and 2 below, which give a quantitative version of the identifiability result and play a role in deriving the sample complexity of ASTRAL-one.

In fact, even in the absence of ILS (i.e., in the presence of GDL alone), no sample complexity results were previously available. While identifiability is established through symmetry arguments (see, e.g., Lemma 1 below, whose simple proof closely mirrors the core argument in [22]), sample complexity bounds require a quantitative analysis that is significantly more involved. Our arguments here (see Sections 3 and 4) combine symmetry arguments, explicit computations, and a detailed case analysis of well-chosen events. In addition we incorporate ILS, which further complicates the analysis, in part because of the forward-in-time/backward-in-time nature of the combined process. We expect that the types of arguments developed here will prove useful in analyzing other reconstruction methods under related complex models of genome evolution.

Our main identifiability result is stated next.

THEOREM 2 (Identifiability of species tree). *Let T be a model species tree with at least $n \geq 4$ leaves. Then T, without its root, is identifiable from the distribution of gene trees $\mathcal{T}$ under the DLCoal model over T.*

This identifiability result is established by showing that, for each quartet in the species tree, the most likely gene tree matches the species tree. As in [22, 26], a direct consequence of this proof is the statistical consistency of the ASTRAL-one.

THEOREM 3 (Consistency of ASTRAL-one). *As the number of input gene trees tends toward infinity, the output of ASTRAL-one converges to T almost surely, when run in exact mode or in its default constrained version.*

We use a similar reasoning to prove the consistency of ASTRAL-multi. This result is new.

THEOREM 4 (Consistency of ASTRAL-multi). *As the number of input gene trees tends toward infinity, the output of ASTRAL-multi converges to T almost surely, when run in exact mode or in its default constrained version.*

3. First step: A proof of identifiability of the species tree under the DLCoal model. Let $\mathcal{Q} = \{A, B, C, D\}$ and assume, without loss of generality, that $T^{\mathcal{Q}}$ has unrooted quartet topology $AB|CD$, that is, it has the topology depicted in the top left of Figure 2 (ignoring the root). Let t be a gene tree generated under the DLCoal model on T and let $t^{\mathcal{Q}}$ be its restriction to the gene copies from the species in $\mathcal{Q}$. The high-level idea behind our proof of Theorem 2 is the following:

> *Conditioning on the number of copies in species A, B, C, D in the species in $\mathcal{Q}$, independently pick a uniformly random gene copy a, b, c, d in species A, B, C, D and let q be the corresponding quartet topology under $t^{\mathcal{Q}}$. We show that the most likely outcome is $q = ab|cd$.*

This is the same approach as that used in [22], but the analysis of the model is more involved.

Define $\mathcal{X} = (\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D})$ to be the number of copies in species A, B, C, D, respectively. For example, Figure 1 depicts a realization in which $\mathcal{X} = (2, 2, 2, 2)$ since the locus tree (and hence also the gene tree) has exactly two leaves in each of the species A, B, C, and D. We will let $\mathbb{P}'$ be the probability measure subject to conditioning on the random vector $\mathcal{X}$.

Although we seek to reconstruct an *unrooted* species tree, under the DLCoal model, the locus trees and gene trees are in fact generated from a *rooted* species tree. Therefore when

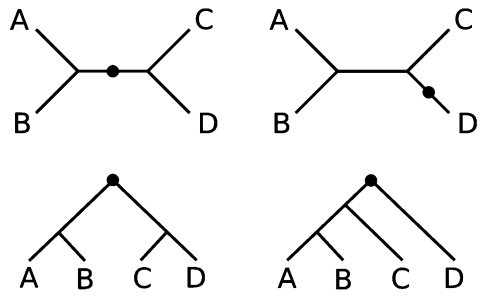

FIG. 2. *The top row shows the two cases of root location on $T^{\mathcal{Q}}$: the root of $T^{\mathcal{Q}}$ may occur either on the internal edge (left) or the pendant edge adjacent to D (right). In both cases the unrooted quartet topology of $T^{\mathcal{Q}}$ is $AB|CD$. However, as shown in the bottom row, the two choices of root location leads to two different rooted topologies: the balanced case (left) and the caterpillar case (right), which we consider separately.*

restricting the species tree to four species, there are two cases of root location to consider: when the root of the species quartet $T^{\mathcal{Q}}$ is located on the internal quartet edge of $T^{\mathcal{Q}}$ (the "balanced case") or on a pendant edge of $T^{\mathcal{Q}}$ (the "caterpillar case"). In the caterpillar case we may take D to be the pendant edge incident with the root. See Figure 2.

For gene copies a, b, c, d from A, B, C, D, define the following events:

$$Q_1 = \{q = ab|cd\}, \qquad Q_2 = \{q = ac|bd\}, \qquad Q_3 = \{q = ad|bc\}.$$

3.1. *Balanced case.* We first consider the balanced case. Without loss of generality, assume the species tree restricted to $\mathcal{Q}$ has rooted topology as in the bottom left of Figure 2. Let R be the most recent common ancestor of $\mathcal{Q}$ in the species quartet $T^{\mathcal{Q}}$ and I be the number of locus copies exiting R forward in time. Let $\mathbb{P}''$ be the probability measure indicating conditioning on I as well as on $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}$. For any selection of copies (a, b, c, d) from each species in the quartet, let $i_x \in \{1, \dots, I\}$ be the ancestral lineage on the locus tree of $x \in \{a, b, c, d\}$ in R. By the law of total probability, we have

$$\mathbb{P}[Q_i] = \mathbb{E}[\mathbb{P}''[Q_i]], \quad i = 1, 2, 3. \tag{2}$$

Hence, in order to show identifiability of the species quartet, it is sufficient to show that

$$\mathbb{P}''[Q_1] > \max\{\mathbb{P}''[Q_2], \mathbb{P}''[Q_3]\}$$

when the copies of (A, B, C, D) are chosen uniformly at random.

We let $\mathcal{X} \geq \vec{1}$ be the event that each of (A, B, C, D) has at least one copy to select from. On the complement of $\mathcal{X} \geq \vec{1}$ (i.e., at least one of (A, B, C, D) fails to have a copy to select), then ASTRAL-one selects Q_1, Q_2, Q_3 each with probability 0. So we consider the case $\mathcal{X} \geq \vec{1}$. We will use the notation $z_1 \wedge z_2 = \min\{z_1, z_2\}$.

PROPOSITION 1 (Quartet identifiability: Balanced case). *Let $x = \mathbb{P}''[i_a = i_b]$ and $y = \mathbb{P}''[i_c = i_d]$. On the events $\mathcal{X} \geq \vec{1}$ and $I \geq 1$, we have almost surely*

$$\mathbb{P}''[Q_1] - \mathbb{P}''[Q_2] > \frac{1}{12}\left(x - \frac{1}{I}\right) \wedge \left(y - \frac{1}{I}\right).$$

The proof of Proposition 1 is in Section 3.1.4. We first establish a series of lemmas.

The next lemma shows that $x, y \geq 1/I$, similar to [22], Lemma 1. In that work, it is proved that the probabilities of $\{i_a = i_b\}$ and $\{i_c = i_d\}$ are each at least $1/I$ under a different conditional probability measure. Our proof is otherwise identical.

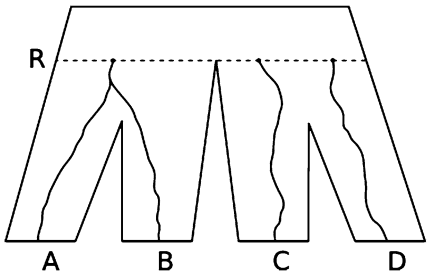

FIG. 3. *An example of a locus tree L satisfying the event F_{ab} for selected gene copies a, b, c, d. In particular, the edges of L are depicted as residing within the fat edges of $T^{\mathcal{Q}}$, however only the portion of L coinciding with the edges of $T^{\mathcal{Q}}$ is shown; the portion of L above R is omitted.*

LEMMA 1. *Let $x = \mathbb{P}''[i_a = i_b]$ and $y = \mathbb{P}''[i_c = i_d]$. On the events $\mathcal{X} \geq \vec{1}$ and $I \geq 1$, we have almost surely*

$$x \wedge y \geq \frac{1}{I}.$$

PROOF. For $j \in \{1, \ldots, I\}$, let N_j be the number of gene copies descending from j in R that survive to the most recent common ancestor of A and B. Conditioning on $(N_j)_j$, the choice of a and b in the locus tree is independent as in [22]. So i_a and i_b are picked proportionally to the N_j's by symmetry. Then

$$\mathbb{P}''[i_a = i_b] = \mathbb{E}''\big[\mathbb{P}''\big[i_a = i_b | (N_j)_j\big]\big] = \mathbb{E}''\bigg[\frac{\sum_{j=1}^I N_j^2}{(\sum_{j=1}^I N_j)^2}\bigg] \geq \frac{1}{I},$$

as in [22], Lemma 1. The same holds for y, completing the proof of the lemma. □

3.1.1. *Ancestral locus configurations.* Conditioned on $\mathcal{X}$ and I, we will characterize the occurrence of Q_1, Q_2, Q_3 based on how $i_x, x \in \{a, b, c, d\}$, are picked at the root R, that is, based on which pairs i_x, i_y are equal. Then, in a worst-case scenario, we will analyze events of the coalescent process above R. For an arbitrary quartet (a, b, c, d), we relate the likelihood of Q_1, Q_2, Q_3 under each of the following events:

$$\begin{aligned}
&E = a - b - c - d,\\
&F_{ab} = ab - c - d, \qquad F_{ac} = ac - b - d, \qquad F_{ad} = ad - b - c,\\
&F_{bc} = bc - a - d, \qquad F_{bd} = bd - a - c, \qquad F_{cd} = cd - a - b,\\
&G_{ab} = ab - cd, \qquad G_{ac} = ac - bd, \qquad G_{ad} = ad - bc,\\
&H_{abc} = abc - d, \qquad H_{abd} = abd - c, \qquad H_{acd} = acd - b, \qquad H_{bcd} = bcd - a,\\
&K = abcd,
\end{aligned} \tag{3}$$

where $-$ indicates separate lineages at R for the chosen copies from A, B, C, D. For example, the event F_{ab} indicates that the i_c and i_d are different and are different from the common ancestral lineage for i_a and i_b. See Figure 3. These events are disjoint and mutually exhaustive. Letting $\mathcal{E}$ run across all the above events, the law of total probability implies

$$\mathbb{P}''[Q_i] = \sum_{\mathcal{E}} \mathbb{P}''[Q_i | \mathcal{E}] \mathbb{P}''[\mathcal{E}]. \tag{4}$$

3.1.2. *Reduction to coalescence above R.* For the rooted locus quartet implied by the four copies a, b, c, d, let $\mathcal{NC}$ be the event that no coalescence event occurs beneath R among the four corresponding lineages. The following lemma shows that conditioning on $\mathcal{NC}$ reduces the probability of Q_1 while increasing that of Q_2 and Q_3.

LEMMA 2. *For any I and any $\mathcal{X} \geq \vec{1}$ and any event*

$$\mathcal{E} \in \{E, F_{ab}, \ldots, G_{ab}, \ldots, H_{abc}, \ldots\},$$

we have

$$\mathbb{P}''[Q_1|\mathcal{E}] \geq \mathbb{P}''[Q_1|\mathcal{E} \cap \mathcal{NC}] \quad \textit{and} \quad \mathbb{P}''[Q_i|\mathcal{E}] \leq \mathbb{P}''[Q_i|\mathcal{E} \cap \mathcal{NC}], \quad i \in \{2, 3\},$$

almost surely.

PROOF. Under $\mathcal{NC}^c$, by definition, at least one of the pairs $\{a, b\}$ or $\{c, d\}$ coalesces below R. Both cases immediately lead to the same quartet topology, so Q_1 is guaranteed. See Figure 5 below for an illustration when $\mathcal{E} = K$. Then the law of total probability implies

$$\mathbb{P}''[Q_1|\mathcal{E}] = \mathbb{P}''\big[\mathcal{NC}^c|\mathcal{E}\big] + \mathbb{P}''[Q_1|\mathcal{E} \cap \mathcal{NC}]\mathbb{P}''[\mathcal{NC}|\mathcal{E}] \geq \mathbb{P}''[Q_1|\mathcal{E} \cap \mathcal{NC}].$$

Similarly

$$\mathbb{P}''[Q_i|\mathcal{E}] = \mathbb{P}''[Q_i|\mathcal{E} \cap \mathcal{NC}]\mathbb{P}''[\mathcal{NC}|\mathcal{E}] \leq \mathbb{P}''[Q_i|\mathcal{E} \cap \mathcal{NC}],$$

for $i \in \{2, 3\}$. □

The event K will play a special role in the proof and we treat it separately. For the other terms, combining (4) and Lemma 2, we have

$$\begin{aligned}
\mathbb{P}''[Q_1] - \mathbb{P}''[Q_2] &= \big(\mathbb{P}''[Q_1|K] - \mathbb{P}''[Q_2|K]\big)\mathbb{P}''[K] \\
&\quad + \sum_{\mathcal{E} \neq K} \big(\mathbb{P}''[Q_1|\mathcal{E}] - \mathbb{P}''[Q_2|\mathcal{E}]\big)\mathbb{P}''[\mathcal{E}] \\
&\geq \big(\mathbb{P}''[Q_1|K] - \mathbb{P}''[Q_2|K]\big)\mathbb{P}''[K] \\
&\quad + \sum_{\mathcal{E} \neq K} \big(\mathbb{P}''[Q_1|\mathcal{E} \cap \mathcal{NC}] - \mathbb{P}''[Q_2|\mathcal{E} \cap \mathcal{NC}]\big)\mathbb{P}''[\mathcal{E}].
\end{aligned}$$

To prove Proposition 1, we derive an explicit bound on this last sum.

Under $\mathbb{P}''$, the events $E, H_{abc}, H_{abd}, H_{bcd}, H_{acd}$ are symmetric in the sense that switching the roles of a and c or the roles of a and d does not change the conditional probability of Q_1 and Q_2. Hence

$$\mathbb{P}''[Q_1|\mathcal{E} \cap \mathcal{NC}] = \mathbb{P}''[Q_2|\mathcal{E} \cap \mathcal{NC}] \quad \forall \mathcal{E} \in \{E, H_{abc}, H_{abd}, H_{bcd}, H_{acd}\}$$

and using this above we get

$$\begin{aligned}
&\mathbb{P}''[Q_1] - \mathbb{P}''[Q_2] \\
&\quad \geq \big(\mathbb{P}''[Q_1|K] - \mathbb{P}''[Q_2|K]\big)\mathbb{P}''[K] \\
&\qquad + \sum_{j \in \{ab, ac, ad\}} \big(\mathbb{P}''[Q_1|G_j \cap \mathcal{NC}] - \mathbb{P}''[Q_2|G_j \cap \mathcal{NC}]\big)\mathbb{P}''[G_j] \\
&\qquad + \sum_{j \in \{ab, \ldots, cd\}} \big(\mathbb{P}''[Q_1|F_j \cap \mathcal{NC}] - \mathbb{P}''[Q_2|F_j \cap \mathcal{NC}]\big)\mathbb{P}''[F_j].
\end{aligned} \tag{5}$$

3.1.3. *The F and G events.* We now consider the events $\{F_{ab}, F_{cd}, F_{ac}, F_{bd}\}$ and $\{G_{ab}, G_{ac}\}$.

Event probabilities. In the next lemma, we compute the probabilities of a given locus tree quartet satisfying the events in $\{F_{ab}, \ldots, F_{cd}, G_{ab}, G_{ac}\}$.

LEMMA 3. *Let $x = \mathbb{P}''[i_a = i_b]$ and $y = \mathbb{P}''[i_c = i_d]$. For $I \geq 2$ and any $\mathcal{X} \geq \vec{1}$, the following hold almost surely*:

$$\begin{aligned}
\mathbb{P}''[F_{ab}] &= \frac{I-2}{I} x(1-y), \\
\mathbb{P}''[F_{cd}] &= \frac{I-2}{I} (1-x)y, \\
\mathbb{P}''[F_{ac}] = \mathbb{P}''[F_{bd}] &= \frac{I-2}{I(I-1)} (1-x)(1-y), \\
\mathbb{P}''[G_{ab}] &= \frac{I-1}{I} xy, \\
\mathbb{P}''[G_{ac}] &= \frac{1}{I(I-1)} (1-x)(1-y).
\end{aligned}$$

PROOF. The calculations for F_{ab} and F_{cd} are similar, except that we condition on different events. Indeed, note that

$$\begin{aligned}
\mathbb{P}''[F_{ab}] &= \mathbb{P}''[F_{ab} | i_a = i_b, i_c \neq i_d] \mathbb{P}''[i_a = i_b] \mathbb{P}''[i_c \neq i_d], \\
\mathbb{P}''[F_{cd}] &= \mathbb{P}''[F_{cd} [i_a \neq i_b, i_c = i_d] \mathbb{P}''[i_a \neq i_b] \mathbb{P}''[i_c = i_d].
\end{aligned}$$

The conditional probability of F_{ab} is then obtained by considering that given the placement of the pair (i_c, i_d) among the I ancestral lineages, the shared lineage $i_a = i_b$ has $I-2$ choices where they do not intersect $\{i_c, i_d\}$. The result in the statement follows. Similarly, for F_ι with $\iota \in \{ac, bd\}$, we have

$$\mathbb{P}''[F_\iota] = \mathbb{P}''[F_\iota | i_a \neq i_b, i_c \neq i_d] \mathbb{P}''[i_a \neq i_b] \mathbb{P}''[i_c \neq i_d].$$

In this case, out of $I(I-1)$ choices for i_a and i_b, the choice of i_c is determined and there are $I-2$ remaining choices for i_d, implying the result.

We use the same principle for G_{ab} and G_{ac}. Keeping this in mind, we have

$$\begin{aligned}
\mathbb{P}''[G_{ab}] &= \mathbb{P}''[G_{ab} | i_a = i_b, i_c = i_d] \mathbb{P}''[i_a = i_b] \mathbb{P}''[i_c = i_d], \\
\mathbb{P}''[G_{ac}] &= \mathbb{P}''[G_{ac} | i_a \neq i_b, i_c \neq i_d] \mathbb{P}''[i_a \neq i_b] \mathbb{P}''[i_c \neq i_d],
\end{aligned}$$

and we proceed as before to get the result. □

Using the previous lemma, we collect further bounds on the probabilities of events at the root of the locus tree.

LEMMA 4. *Letting again $x = \mathbb{P}''[i_a = i_b]$ and $y = \mathbb{P}''[i_c = i_d]$, the following statements hold*:

(a) *If $I = 2$ then*

$$\mathbb{P}''[G_{ab}] - \mathbb{P}''[G_{ac}] \geq \left(x - \frac{1}{2}\right) \wedge \left(y - \frac{1}{2}\right).$$

(b) *If* $I \geq 3$ *and* $x \wedge y \geq 1/2$, *then*

$$\mathbb{P}''[G_{ab}] - \mathbb{P}''[G_{ac}] \geq 1/8.$$

(c) *If* $I = 2$,

$$\mathbb{P}''[F_{ab}] - \mathbb{P}''[F_{ac}] - \mathbb{P}''[F_{bd}] + \mathbb{P}''[F_{cd}] = 0.$$

(d) *If* $I \geq 3$ *and* $x \wedge y \leq 1/2$, *then*

$$\mathbb{P}''[F_{ab}] - \mathbb{P}''[F_{ac}] - \mathbb{P}''[F_{bd}] + \mathbb{P}''[F_{cd}] \geq \frac{1}{4}\left(x - \frac{1}{I}\right) \wedge \left(y - \frac{1}{I}\right).$$

PROOF. By Lemma 3, $\mathbb{P}''[G_{ab}] - \mathbb{P}''[G_{ac}] = \frac{1}{2}(x + y - 1)$ which implies part (a). To prove (b), observe that by Lemma 3 again

$$\begin{aligned}
\mathbb{P}''[G_{ab}] - \mathbb{P}''[G_{ac}] &= \frac{I-1}{I}xy - \frac{1}{I(I-1)}(1-x)(1-y) \\
&\geq \frac{1}{2}\left(\frac{I-1}{I}y - \frac{1}{I(I-1)}(1-y)\right) \\
&\geq \frac{1}{4}\left(\frac{I-1}{I} - \frac{1}{I(I-1)}\right) \\
&= \frac{1}{4}\left(\frac{I-2}{I-1}\right),
\end{aligned}$$

where the inequalities are justified by the assumption $x \wedge y \geq 1/2$. Since $I \geq 3$, it follows that $\mathbb{P}''[G_{ab}] - \mathbb{P}''[G_{ac}] \geq 1/8$.

To prove (c) and (d), observe that by Lemma 3,

$$\begin{aligned}
&\mathbb{P}''[F_{ab}] - \mathbb{P}''[F_{ac}] - \mathbb{P}''[F_{bd}] + \mathbb{P}''[F_{cd}] \\
&\quad= \frac{I-2}{I}\big(x(1-y) + y(1-x)\big) - 2\frac{I-2}{I(I-1)}(1-x)(1-y) \\
&\quad= \frac{I-2}{I(I-1)}\big((1-y)\big((I-1)x - (1-x)\big) + (1-x)\big((I-1)y - (1-y)\big)\big) \\
&\quad= \frac{I-2}{I(I-1)}\big((1-y)(Ix-1) + (1-x)(Iy-1)\big).
\end{aligned}$$

Clearly if $I = 2$, the right-hand side is zero, which proves (c). Furthermore, since $x, y \geq 1/I$ by Lemma 1, it follows that both $(1-y)(Ix-1) \geq 0$ and $(1-x)(Iy-1) \geq 0$, and therefore

$$\begin{aligned}
&\mathbb{P}''[F_{ab}] - \mathbb{P}''[F_{ac}] - \mathbb{P}''[F_{bd}] + \mathbb{P}''[F_{cd}] \\
&\qquad\geq \frac{I-2}{I(I-1)}(1-u)(Iv-1)
\end{aligned}$$

for $(u, v) \in \{(x, y), (y, x)\}$. Taking $u = \min(x, y)$ and $v = \max(x, y)$ gives

$$\begin{aligned}
&\mathbb{P}''[F_{ab}] - \mathbb{P}''[F_{ac}] - \mathbb{P}''[F_{bd}] + \mathbb{P}''[F_{cd}] \\
&\quad\geq \frac{I-2}{I(I-1)}\big(1 - \min(x, y)\big)\big(I\max(x, y) - 1\big) \\
&\quad\geq \frac{I-2}{I-1}\big(1 - \min(x, y)\big)\left(\max(x, y) - \frac{1}{I}\right)
\end{aligned}$$

$$\geq \frac{1}{2}(1-\min(x,y))\left(\max(x,y)-\frac{1}{I}\right)$$
$$\geq \frac{1}{4}\left(\max(x,y)-\frac{1}{I}\right),$$

which implies (d). □

Conditional probabilities of quartet topologies. In the following lemma, we give expressions for $\mathbb{P}''[Q_i|\mathcal{E}\cap\mathcal{NC}]$ across the events $\{F_{ab}, F_{cd}, F_{ac}, F_{bd}\}$ and $\{G_{ab}, G_{ac}\}$.

LEMMA 5. (*a*) *For any I and any $\mathcal{X}\geq\vec{1}$, we have*

$$\mathbb{P}''[Q_1|F_{ab}\cap\mathcal{NC}]=\mathbb{P}''[Q_1|F_{cd}\cap\mathcal{NC}]=\mathbb{P}''[Q_2|F_{ac}\cap\mathcal{NC}]=\mathbb{P}''[Q_2|F_{bd}\cap\mathcal{NC}]=:\phi''_+$$

and

$$\mathbb{P}''[Q_2|F_{ab}\cap\mathcal{NC}]=\mathbb{P}''[Q_2|F_{cd}\cap\mathcal{NC}]=\mathbb{P}''[Q_1|F_{ac}\cap\mathcal{NC}]=\mathbb{P}''[Q_1|F_{bd}\cap\mathcal{NC}]=:\phi''_-.$$

(*b*) *For any I and any $\mathcal{X}\geq\vec{1}$, we have*

$$\mathbb{P}''[Q_1|G_{ab}\cap\mathcal{NC}]=1$$

and

$$\mathbb{P}''[Q_2|G_{ac}\cap\mathcal{NC}]=1.$$

PROOF. (a) The quantities ϕ''_+ and ϕ''_- are indeed well-defined as above by symmetry. (b) By switching the roles of b and c, we observe that $\mathbb{P}''[Q_1|G_{ab}\cap\mathcal{NC}]=\mathbb{P}''[Q_2|G_{ac}\cap\mathcal{NC}]$. To see why $\mathbb{P}''[Q_1|G_{ab}\cap\mathcal{NC}]=1$, we again examine the topology above the root with leaves ab and cd. At least one of these leaves descends from a daughter edge, which implies Q_1 is constructed with probability 1. This completes the proof of the lemma. □

The following lemma establishes that, conditioned on F_{ab} and $\mathcal{NC}$, the difference in probability between Q_1 and Q_2 is at least $1/3$.

LEMMA 6. *For $I\geq 1$ and any $\mathcal{X}\geq\vec{1}$, we have*

$$\phi''_+-\phi''_-\geq\frac{1}{3}.$$

PROOF. By definition of ϕ''_+ and ϕ''_-, it suffices to show $\mathbb{P}''[Q_1|F_{ab}\cap\mathcal{NC}]-\mathbb{P}''[Q_2|F_{ab}\cap\mathcal{NC}]\geq 1/3$. Conditioned on $F_{ab}\cap\mathcal{NC}$, no coalescence event between the chosen lineages occurs beneath R. So, we examine the topology of the locus tree above the root with leaf set being the three leaves implied by F_{ab}. Using the law of total probability, we condition further across the three possible rooted locus topologies on the three leaves ab, c, and d. For each $i\in\{ab,c,d\}$, let τ_i be the rooted topology on three leaves in which the outgroup is labeled by i and the other two leaves are labelled from the remaining letters in $\{a,b,c,d\}$. For example, τ_{ab} has c and d as siblings with ab as the outgroup. Then

$$\mathbb{P}''[Q_1|F_{ab}\cap\mathcal{NC}]-\mathbb{P}''[Q_2|F_{ab}\cap\mathcal{NC}]$$
$$=\frac{1}{3}\sum_i(\mathbb{P}''[Q_1|\tau_i,F_{ab}\cap\mathcal{NC}]-\mathbb{P}''[Q_2|\tau_i,F_{ab}\cap\mathcal{NC}]),$$

where we used the fact that $\mathbb{P}''[\tau_i|F_{ab}\cap\mathcal{NC}]=1/3$ for each i. Now we compute the summands. If $i=ab$, then either ab descends from a daughter lineage or the pair (c,d) descends

from a daughter lineage, meaning we observe Q_1 with probability 1 and Q_2 with probability 0. In the other two cases, let $p > 0$ be the probability that a and b coalesce along the pendant edge for ab. If they do not coalesce along the pendant edge, then the lineages from a and b live in the same population as that of, say, c. Then there is probability $1/3$ that the first coalescing pair among a, b, c is a, b. So the probability of observing Q_1 is $p + \frac{1}{3}(1-p)$. There is probability $1/3$ that the first coalescing pair among a, b, c is a, c, so the probability of observing Q_2 is $\frac{1}{3}(1-p)$. Then

$$\phi''_+ - \phi''_- = \frac{1}{3}\left(1 - 0 + 2\left(p + \frac{1}{3}(1-p) - \frac{1}{3}(1-p)\right)\right) = \frac{1}{3}(1+2p) \geq \frac{1}{3}. \qquad \square$$

3.1.4. *Proof of Proposition 1.* With that we can prove Proposition 1.

PROOF OF PROPOSITION 1. In the $I = 1$ case, $\mathbb{P}''[K] = 1$ so

$$\mathbb{P}''[Q_1] - \mathbb{P}''[Q_2] = \mathbb{P}''[Q_1|K] - \mathbb{P}''[Q_2|K] > 0, \tag{6}$$

where we used that, under $K \cap \mathcal{NC}$, the quartets Q_1 and Q_2 occur with equal probability under $\mathbb{P}''$. Since $x - 1/I = y - 1/I = 0$, the claim follows.

For $I \geq 2$, (5) and Lemma 5 implies that

$$\begin{aligned}
\mathbb{P}''[Q_1] - \mathbb{P}''[Q_2] &\geq (\mathbb{P}''[Q_1|K] - \mathbb{P}''[Q_2|K])\mathbb{P}''[K] \\
&\quad + \mathbb{P}''[G_{ab}] - \mathbb{P}''[G_{ac}] + (\phi''_+ - \phi''_-)(\mathbb{P}''[F_{ab}] + \mathbb{P}''[F_{cd}]) \\
&\quad - (\phi''_+ - \phi''_-)(\mathbb{P}''[F_{ac}] + \mathbb{P}''[F_{bd}]) \\
&> \mathbb{P}''[G_{ab}] - \mathbb{P}''[G_{ac}] \\
&\quad + (\phi''_+ - \phi''_-)(\mathbb{P}''[F_{ab}] - \mathbb{P}''[F_{ac}] - \mathbb{P}''[F_{bd}] + \mathbb{P}''[F_{cd}]),
\end{aligned}$$

where again we used that, under $K \cap \mathcal{NC}$, the quartets Q_1 and Q_2 occur with equal probability. If $I = 2$, then by Lemma 6 and Lemma 4 parts (a) and (c), this leads to

$$\mathbb{P}''[Q_1] - \mathbb{P}''[Q_2] > \left(x - \frac{1}{I}\right) \wedge \left(y - \frac{1}{I}\right). \tag{7}$$

If $I \geq 3$, then by Lemma 4 parts (b) and (d),

$$\mathbb{P}''[Q_1] - \mathbb{P}''[Q_2] > \begin{cases} 1/8 & \text{if } x \wedge y \geq 1/2, \\ \dfrac{1}{12}\left(x - \dfrac{1}{I}\right) \wedge \left(y - \dfrac{1}{I}\right) & \text{if } x \wedge y \leq 1/2. \end{cases} \tag{8}$$

It follows that $\mathbb{P}''[Q_1] - \mathbb{P}''[Q_2] > \frac{1}{12}(x - \frac{1}{I}) \wedge (y - \frac{1}{I})$, finishing the proof of the main claim in the balanced case. $\square$

3.2. *Caterpillar case.* We now consider the caterpillar case. Without loss of generality, assume the species tree restricted to $\mathcal{Q}$ has rooted topology as in the bottom right of Figure 2. Let R be the most recent common ancestor of A, B, C and let I be the number of locus copies exiting R (forward in time). Let $\mathbb{P}''$ be the probability measure indicating conditioning on I and $\mathcal{X}$. Let $i_x \in \{1, \ldots, I\}$ be the ancestral lineage of $x \in \{a, b, c\}$ in R. As with the balanced case, on the complement of $\mathcal{X} \geq \vec{1}$, ASTRAL-one selects Q_1, Q_2, Q_3 each with probability 0. To prove

$$\mathbb{P}[Q_1] > \max\{\mathbb{P}[Q_2], \mathbb{P}[Q_3]\},$$

it is sufficient to prove Proposition 2 below for $\mathcal{X} \geq \vec{1}$.

PROPOSITION 2 (Quartet identifiability: Caterpillar case). *Let $x = \mathbb{P}''[i_a = i_b]$. On the events $I \geq 1$ and $\mathcal{X} \geq \vec{1}$, we have almost surely*

$$\mathbb{P}''[Q_1] - \mathbb{P}''[Q_2] > \frac{1}{3}\left(x - \frac{1}{I}\right).$$

Similar to the balanced case, in order to prove this proposition we consider the following events:

$$\begin{aligned} E &= a - b - c, \\ G_{ab} = ab - c, \qquad G_{ac} &= ac - b, \qquad G_{bc} = bc - a, \\ K &= abc, \end{aligned}$$

where $-$ indicates separation of lineages in R of the selected copies of A, B, C. Letting $\mathcal{E}$ run across all events, the law of total probability implies

$$\mathbb{P}''[Q_i] = \sum_{\mathcal{E}} \mathbb{P}''[Q_i|\mathcal{E}]\mathbb{P}''[\mathcal{E}]. \tag{9}$$

Let $\mathcal{NC}$ be the event that no coalescent event occurs beneath R between the three lineage corresponding to a, b, c.

Analogues to Lemmas 1, 2, 3, 4, and 5 hold with similar proofs.

LEMMA 7. *Let $x = \mathbb{P}''[i_a = i_b]$. On the events $\mathcal{X} \geq \vec{1}$ and $I \geq 1$, we have almost surely*

$$x \geq \frac{1}{I}.$$

LEMMA 8. *Let the species tree be a rooted caterpillar on four leaves A, B, C, D. For all $I \geq 1$ and $\mathcal{X} \geq \vec{1}$, and any event $\mathcal{E} \in \{E, G_{ab}, G_{ac}, G_{bc}\}$,*

$$\mathbb{P}''[Q_1|\mathcal{E}] \geq \mathbb{P}''[Q_1|\mathcal{E} \cap \mathcal{NC}] \quad \text{and} \quad \mathbb{P}''[Q_i|\mathcal{E}] \leq \mathbb{P}''[Q_i|\mathcal{E} \cap \mathcal{NC}], \quad i \in \{2, 3\}.$$

LEMMA 9. *For $I \geq 2$ and any $\mathcal{X} \geq \vec{1}$, let $x = \mathbb{P}''[i_a = i_b]$. Then the following hold:*

$$\begin{aligned} \mathbb{P}''[G_{ab}] &= \frac{I-1}{I}x, \\ \mathbb{P}''[G_{ac}] &= \mathbb{P}''[G_{bc}] = \frac{1}{I}(1-x). \end{aligned}$$

LEMMA 10. *On $I \geq 1$, almost surely*

$$\mathbb{P}''[G_{ab}] - \mathbb{P}''[G_{ac}] = x - \frac{1}{I}.$$

LEMMA 11. *For any I and any $\mathcal{X} \geq \vec{1}$, we have*

$$\mathbb{P}''[Q_1|G_{ab} \cap \mathcal{NC}] = \mathbb{P}''[Q_2|G_{ac} \cap \mathcal{NC}] =: \psi''_+$$

and

$$\mathbb{P}''[Q_2|G_{ab} \cap \mathcal{NC}] = \mathbb{P}''[Q_1|G_{ac} \cap \mathcal{NC}] =: \psi''_-.$$

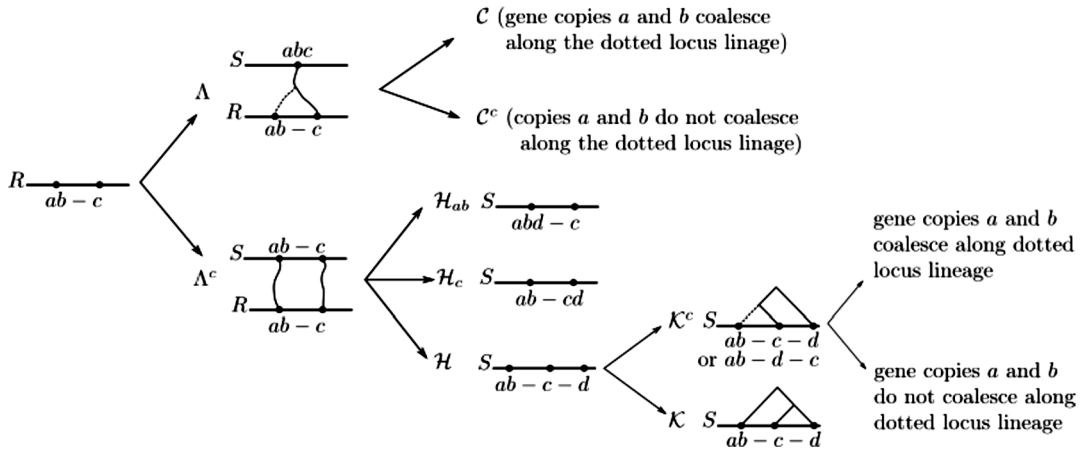

FIG. 4. *Flowchart for case analysis in Lemma 12.*

3.2.1. *The G events.* The following lemma bounds the conditional probability difference for the G events.

LEMMA 12. *On the events $I \geq 1$ and $\mathcal{X} \geq \vec{1}$, we have almost surely*

$$\psi_+'' - \psi_-'' \geq \frac{1}{3}.$$

PROOF. By definition of ψ_+'' and ψ_-'', it suffices to show that $\mathbb{P}''[Q_1|G_{ab} \cap \mathcal{NC}] - \mathbb{P}''[Q_2|G_{ab} \cap \mathcal{NC}] \geq 1/3$. The proof of this inequality involves decomposing $G_{ab} \cap \mathcal{NC}$ into a number of subcases, depicted in Figure 4, and computing the probabilities of Q_1 and Q_2 in each subcase. Let S be the most recent common ancestor of $\mathcal{Q}$ in the species tree. Let Λ be the event that the ab and c individuals in R descend from a common ancestor in S and let $q = \mathbb{P}''[\Lambda|G_{ab} \cap \mathcal{NC}]$. There are two cases:

1 *(Condition on Λ).* Let $\mathcal{C}$ be the event that gene copies a and b coalesce above R and below the MRCA of loci $i_{ab} = i_a = i_b$ and i_c. Let $q' = \mathbb{P}''[\mathcal{C}|\Lambda, G_{ab}, \mathcal{NC}]$. We claim that $q' \geq 1/2$. To see this, observe that conditional on $G_{ab} \cap \mathcal{NC}$ the loci i_{ab} and i_c share the same ancestral locus at S only if there occurred a duplication event between S and R which is ancestral to both of them. Therefore with probability at least $1/2$, the gene copies a, b coalesce along their shared pendant edge in the rooted topology between R and S, proving the claim. Furthermore, it is obvious that

$$\mathbb{P}''[Q_1|\mathcal{C}, \Lambda, G_{ab}, \mathcal{NC}] - \mathbb{P}''[Q_2|\mathcal{C}, \Lambda, G_{ab}, \mathcal{NC}] = 1. \tag{10}$$

On the other hand, conditional on $\mathcal{C}^c$, the copies of a, b, and c enter the same population and are then symmetric, and hence

$$\mathbb{P}''\big[Q_1|\mathcal{C}^c, \Lambda, G_{ab}, \mathcal{NC}\big] - \mathbb{P}''\big[Q_2|\mathcal{C}^c, \Lambda, G_{ab}, \mathcal{NC}\big] = 0. \tag{11}$$

2 *(Condition on Λ^c).* Let $\mathcal{H}_j$ be the event that copies d and j share the same ancestor in the locus tree at S, and define $\mathcal{H} = (\mathcal{H}_{ab} \cup \mathcal{H}_c)^c$ and $r = \mathbb{P}''[\mathcal{H}|\Lambda^c, G_{ab}, \mathcal{NC}]$. Then by symmetry, $\mathbb{P}''[\mathcal{H}_{ab}|\Lambda^c, G_{ab}, \mathcal{NC}] = \mathbb{P}''[\mathcal{H}_c|\Lambda^c, G_{ab}, \mathcal{NC}] = \frac{1-r}{2}$. By a further symmetry argument similar to that made in Case 1 we have

$$\mathbb{P}''\big[Q_1|\mathcal{H}_{ab}, \Lambda^c, G_{ab}, \mathcal{NC}\big] - \mathbb{P}''\big[Q_2|\mathcal{H}_{ab}, \Lambda^c, G_{ab}, \mathcal{NC}\big] \geq 0, \tag{12}$$

where the inequality accounts for the possibility that the lineages from a and b coalesce between R and S. Let τ be the topology of the locus tree restricted to the copies ab, c, d and

restricted to the portion above S (and suppressing nodes of degree 2). Conditioned on $\mathcal{H}_c$, we have $\tau = (ab, cd)$, so it must be the case that either ab descends from a daughter lineage or cd descends from a daughter lineage, meaning we observe Q_1 with probability 1 and Q_2 with probability 0. Therefore

$$\mathbb{P}''[Q_1|\mathcal{H}_c, \Lambda^c, G_{ab}, \mathcal{NC}] - \mathbb{P}''[Q_2|\mathcal{H}_c, \Lambda^c, G_{ab}, \mathcal{NC}] = 1. \tag{13}$$

It remains to consider the case $\mathcal{H} = (\mathcal{H}_{ab} \cup \mathcal{H}_c)^c$. Let $\mathcal{K}$ be the event that $\tau = (ab, (c, d))$. By symmetry, the three possible topologies are equally likely, so $\mathbb{P}''[\mathcal{K}|\mathcal{H}, \Lambda^c, G_{ab}, \mathcal{NC}] = 1/3$. Conditioned on $\mathcal{K}$, either ab descends from a daughter lineage or the pair (c, d) descends from a daughter lineage, and therefore

$$\mathbb{P}''[Q_1|\mathcal{K}, \mathcal{H}, \Lambda^c, G_{ab}, \mathcal{NC}] - \mathbb{P}''[Q_2|\mathcal{K}, \mathcal{H}, \Lambda^c, G_{ab}, \mathcal{NC}] = 1. \tag{14}$$

Conditioned on $\mathcal{K}^c$, let p denote the probability that gene copies a and b coalesce along the pendant edge for ab in the rooted triple above S. Then

$$\mathbb{P}''[Q_1|\mathcal{K}^c, \mathcal{H}, \Lambda^c, G_{ab}, \mathcal{NC}] = p + \frac{1}{3}(1-p)$$

and

$$\mathbb{P}''[Q_2|\mathcal{K}^c, \mathcal{H}, \Lambda^c, G_{ab}, \mathcal{NC}] = \frac{1}{3}(1-p),$$

and hence

$$\mathbb{P}''[Q_1|\mathcal{K}^c, \mathcal{H}, \Lambda^c, G_{ab}, \mathcal{NC}] - \mathbb{P}''[Q_2|\mathcal{K}^c, \mathcal{H}, \Lambda^c, G_{ab}, \mathcal{NC}] \geq p, \tag{15}$$

where again the inequality accounts for the possibility that the lineages from a and b coalesce between R and S.

Finally, applying the law of total probability and using equations (10)–(15) gives

$$\begin{aligned}\mathbb{P}''[Q_1|G_{ab}, \mathcal{NC}] - \mathbb{P}''[Q_2|G_{ab}, \mathcal{NC}] &\geq q'q + \left(\frac{1-r}{2} + \frac{1}{3}r + \frac{2}{3}pr\right)(1-q) \\ &\geq \frac{1}{2}q + \frac{1}{3}(1-q) \\ &\geq \frac{1}{3},\end{aligned}$$

where the second inequality follows from $q' \geq 1/2$ and $r \geq 0$. □

3.2.2. *Proofs of Proposition 2 and Theorems 2 and 3.* With that, the proof of Proposition 2 is similar to that of Proposition 1.

In both the balanced and caterpillar cases, observe that $\mathbb{P}''[Q_1] - \mathbb{P}''[Q_3] = \mathbb{P}''[Q_1] - \mathbb{P}''[Q_2]$ by switching the roles of c and d. By Propositions 1 and 2, all species quartet topologies are identifiable and hence we have verified Theorem 2.

Theorem 3 then follows, along similar lines as [22], Theorem 2, from the law of large numbers.

3.3. *Proof of consistency for ASTRAL-multi.* Before finishing the proof of Theorem 1, we give a proof of Theorem 4.

PROOF OF THEOREM 4. Let $\mathcal{N}_{AB|CD}$ (respectively $\mathcal{N}_{AC|BD}, \mathcal{N}_{AD|BC}$) be the number of choices consisting of one gene copy in the gene tree from each species in $\mathcal{Q} = \{A, B, C, D\}$

whose corresponding restriction in t_1 agrees with $AB|CD$ (respectively $AC|BD$, $AD|BC$). Similar to [22], Theorem 3, it suffices to show that

$$\mathbb{E}[\mathcal{N}_{AB|CD}] > \max\{\mathbb{E}[\mathcal{N}_{AC|BD}], \mathbb{E}[\mathcal{N}_{AD|BC}]\}. \tag{16}$$

Letting again $\mathcal{X} = (\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D})$, by taking expectation with respect to I in Propositions 1 and 2, we have on the event $\mathcal{X} \geq \vec{1}$ that

$$\mathbb{P}[q = AB|CD|\mathcal{X}] > \max\{\mathbb{P}[q = AC|BD|\mathcal{X}], \mathbb{P}[q = AD|BC|\mathcal{X}]\}, \tag{17}$$

where q is the topology of a uniformly chosen quartet among A, B, C, D. Let $\mathcal{M} = \mathcal{ABCD}$ be the number of quartet choices and let $q_i, i = 1, \ldots, \mathcal{M}$ be the corresponding topologies ordered arbitrarily. Because q is a uniform choice, we have

$$\mathbb{P}[q = AB|CD|\mathcal{X}] = \frac{1}{\mathcal{M}} \sum_{i=1}^{\mathcal{M}} \mathbb{P}[q_i = AB|CD|\mathcal{X}], \tag{18}$$

and similarly for the other topologies. Since

$$\mathcal{N}_{AB|CD} = \sum_{i=1}^{\mathcal{M}} \mathbf{1}\{q_i = AB|CD\},$$

and similarly for the other topologies, taking expectations and using (17) and (18) gives (16) as claimed. □

4. Proof of sample complexity bound. To prove Theorem 1, our sample complexity result for ASTRAL-one, we use a union bound over all quartets and build on the analysis of Section 3. In particular, the key step of the proof is a more careful analysis of the event K that appeared in the proof of Theorem 2. We first discuss a number of quantities that play an important role in the analysis.

4.1. *Bounds on branching process quantities*. We highlight the role of a number of parameters in the sample complexity: the shortest branch length in the species tree, f; the depth of the species tree, Δ; and the duplication and loss rates, λ and μ. These parameters enter the analysis through three quantities of significance:

- *Coalescence of a pair of lineages on an edge*: In the standard coalescent, the probability that a pair of lineages has coalesced by time f is
$$\gamma = 1 - e^{-f}.$$
- *Survival probability of a quartet*: For a quartet $\mathcal{Q} = \{A, B, C, D\}$, let $\mathcal{X}_\mathcal{Q} = (\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D})$ be the number of gene copies in the corresponding species. The smallest probability over all quartets that a gene family contains a copy in each species will be denoted by
$$\sigma := \min_{\mathcal{Q}} \mathbb{P}[\mathcal{X}_\mathcal{Q} \geq \vec{1}].$$
- *Expected number of lineages at a vertex*: For any vertex R in the species tree, let I_R be the number of copies at R in a single gene family. The largest expectation of I_R over all vertices will be denoted by
$$\alpha = \max_R \mathbb{E}[I_R].$$

These last two quantities can be controlled using branching process theory. See, for example, [57], Section 9.2, for the relevant results in the phylogenetic context. We use the notation $z_1 \vee z_2 = \max\{z_1, z_2\}$.

LEMMA 13. *The following hold*:

- *When* $\mu > \lambda$,

$$\sigma \geq \left[\frac{1}{e^{(\mu-\lambda)\Delta}}\left(1-\frac{\lambda}{\mu}\right)\right]^4.$$

- *When* $\lambda > \mu$,

$$\sigma \geq \left[1-\frac{\mu}{\lambda}\right]^4.$$

PROOF. Let $\mathcal{Q} = \{A, B, C, D\}$ be a quartet and let as before $\mathcal{X}_\mathcal{Q} = (\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D})$ be the number of gene copies in the corresponding species. Assume the species tree topology on $\mathcal{Q}$ is balanced (the argument in the caterpillar case being similar). The probability that $\mathcal{A} \geq 1$ is given by

$$\mathbb{P}[\mathcal{A} \geq 1] = 1 - \frac{\mu}{\lambda} q(\Delta),$$

where

$$q(t) = \lambda \frac{1 - e^{-(\lambda-\mu)t}}{\lambda - \mu e^{-(\lambda-\mu)t}}.$$

It can be checked that, whether $\mu > \lambda$ or $\mu < \lambda$, the function $q(t)$ is increasing in t. Conditioned on $\{\mathcal{A} \geq 1\}$, there is at least one copy in each vertex along the path between the root and A. Hence

$$\mathbb{P}[\mathcal{D} \geq 1 | \mathcal{A} \geq 1] \geq 1 - \frac{\mu}{\lambda} q(\Delta).$$

Repeating this argument for B and C gives

$$\mathbb{P}[\mathcal{X}_\mathcal{Q} \geq \vec{1}] \geq \left(1 - \frac{\mu}{\lambda} q(\Delta)\right)^4.$$

It remains to bound the right-hand side. When $\lambda > \mu$, $q(t) \to 1$ as $t \to +\infty$, which implies $q(t) \leq 1$ by monotonicity. So $1 - \frac{\mu}{\lambda} q(\Delta) \geq 1 - \frac{\mu}{\lambda}$. On the other hand, when $\mu > \lambda$,

$$\begin{aligned} 1 - \frac{\mu}{\lambda} q(\Delta) &= \lambda \frac{1 - e^{-(\lambda-\mu)\Delta}}{\lambda - \mu e^{-(\lambda-\mu)\Delta}} \\ &= \frac{\mu - \lambda}{\mu e^{(\mu-\lambda)\Delta} - \lambda} \\ &\geq \frac{1}{e^{(\mu-\lambda)\Delta}}\left(1 - \frac{\lambda}{\mu}\right). \end{aligned}$$

□

LEMMA 14. *We have*

$$\alpha \leq 1 \vee e^{(\lambda-\mu)\Delta}.$$

PROOF. Recall that we assume there is a single lineage at the top pendant vertex of the species tree. If the time elapsed between this vertex and another vertex U is d, then the expectation number of lineages at U is $e^{(\lambda-\mu)d}$. The result follows from the fact that $d \leq \Delta$ by considering separately the cases $\mu > \lambda$ and $\lambda > \mu$. □

4.2. *Sufficient effective number of samples*. For a quartet $\mathcal{Q}$, let $\mathcal{K}_{\mathcal{Q}}$ be the set of gene trees such that each species in $\mathcal{Q}$ has at least one gene copy. For $k^* \geq 1$, let

$$\mathcal{S}_{k^*} = \{|\mathcal{K}_{\mathcal{Q}}| \geq k^* : \forall \mathcal{Q}\}.$$

LEMMA 15. *For any $k^* \geq 1$ and $\epsilon \in (0,1)$, it holds that $\mathbb{P}[\mathcal{S}_{k^*}] \geq 1-\epsilon$ provided*

$$k \geq \left\{\frac{2k^*}{\sigma}\right\} \vee \left\{\frac{8}{\sigma^2}\log\frac{n}{\epsilon}\right\}.$$

PROOF. For any $\mathcal{Q}$, by the definition of $\mathcal{K}_{\mathcal{Q}}$, if k is the number of loci then

$$\mathbb{E}|\mathcal{K}_{\mathcal{Q}}| \geq k\sigma.$$

Assume k is large enough that $\frac{1}{2}k\sigma \geq k^*$. Then by the union bound and Hoeffding's inequality (see, e.g., [58]), using the fact that the gene trees are independent,

$$\begin{aligned}
\mathbb{P}[\mathcal{S}_{k^*}^c] &\leq \sum_{\mathcal{Q}} \mathbb{P}[|\mathcal{K}_{\mathcal{Q}}| < k^*] \\
&\leq \sum_{\mathcal{Q}} \mathbb{P}\left[|\mathcal{K}_{\mathcal{Q}}| < \frac{1}{2}k\sigma\right] \\
&\leq \sum_{\mathcal{Q}} \mathbb{P}\left[\mathbb{E}|\mathcal{K}_{\mathcal{Q}}| - |\mathcal{K}_{\mathcal{Q}}| > \frac{1}{2}k\sigma\right] \\
&\leq n^4 \exp\left(-2\frac{(k\sigma/2)^2}{k}\right) \\
&\leq \epsilon,
\end{aligned}$$

if

$$k \geq \frac{2}{\sigma^2}\log\frac{n^4}{\epsilon},$$

and since $\epsilon < 1$ this inequality holds whenever

$$k \geq \frac{8}{\sigma^2}\log\frac{n}{\epsilon}.$$

That proves the claim. □

4.3. *The event K*. Using the notation of Section 3, fix a quartet of species $\mathcal{Q} = \{A, B, C, D\}$, let $\mathbb{P}'$ denote the conditional probability given the event $\{\mathcal{X}_{\mathcal{Q}} \geq \vec{1}\}$, and define $\delta' = \mathbb{P}'[Q_1] - 1/3$. Since $\mathbb{P}'[Q_2] = \mathbb{P}'[Q_3]$, we have

$$\delta' = \mathbb{P}'[Q_1] - \frac{\mathbb{P}'[Q_1] + \mathbb{P}'[Q_2] + \mathbb{P}'[Q_3]}{3} = \frac{2}{3}(\mathbb{P}'[Q_1] - \mathbb{P}'[Q_2]). \tag{19}$$

We seek to bound the right-hand side. Assume first that $T^{\mathcal{Q}}$ is balanced. Let $\mathbb{P}'_i$ indicate $\mathbb{P}'$ conditioned on $\{I = i\}$. By the proof of Proposition 1, specifically the argument leading up to (6), (7), and (8), we have

$$\mathbb{P}''[Q_1] - \mathbb{P}''[Q_2] \geq (\mathbb{P}''[Q_1|K] - \mathbb{P}''[Q_2|K])\mathbb{P}''[K].$$

By further conditioning on the event $\{\mathcal{X}_{\mathcal{Q}} \geq \vec{1}\}$ (which has positive probability when $I \geq 1$), we get for all $i \geq 1$

$$\mathbb{P}'_i[Q_1] - \mathbb{P}'_i[Q_2] \geq (\mathbb{P}'_i[Q_1|K] - \mathbb{P}'_i[Q_2|K])\mathbb{P}'_i[K].$$

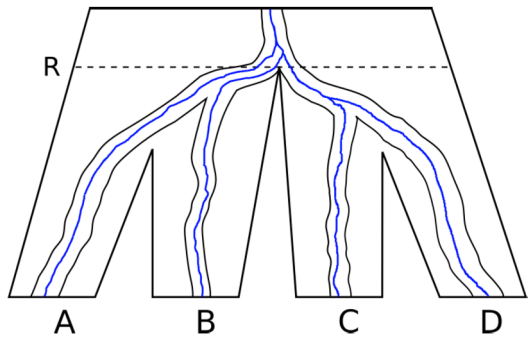

FIG. 5. *An example realization of $K \cap \mathcal{NC}^c$. In particular, a locus tree satisfying event K for copies a, b, c, d is depicted in the case where $T^{\mathcal{Q}}$ is balanced; additionally, a realization of the gene tree is shown (in blue) within the locus tree. This realization satisfies $K \cap \mathcal{NC}^c$ since at least one coalescence event—in this case between the ancestors of gene copies c and d—has occurred on the gene tree below R. Since copies c and d are therefore not separated by an internal edge on the gene tree, the event Q_1 occurs.*

On the event $K \cap \mathcal{NC}$, Q_1, and Q_2 are equally likely by symmetry. On $K \cap \mathcal{NC}^c$, at least one of the pairs $\{a, b\}$ or $\{c, d\}$ coalesces below R, guaranteeing Q_1 (similar to the proof of Lemma 2). See Figure 5 for an illustration. Hence, we have

$$\begin{aligned}\mathbb{P}'_i[Q_1] - \mathbb{P}'_i[Q_2] &\geq \mathbb{P}'_i[Q_1 \cap K] - \mathbb{P}'_i[Q_2 \cap K] \\ &= \mathbb{P}'_i[Q_1 \cap K \cap \mathcal{NC}] - \mathbb{P}'_i[Q_2 \cap K \cap \mathcal{NC}] \\ &\quad + \mathbb{P}'_i\big[Q_1 \cap K \cap \mathcal{NC}^c\big] - \mathbb{P}'_i\big[Q_2 \cap K \cap \mathcal{NC}^c\big] \\ &\geq \mathbb{P}'_i\big[K \cap \mathcal{NC}^c\big].\end{aligned}$$

To bound the right-hand side, we consider the event $\mathcal{C}_{ab}$ that the lineages picked from A and B coalesce below R. Notice in particular that $\mathcal{C}_{ab}$ implies $\{i_a = i_b\}$. The event $K \cap \mathcal{NC}^c$ is implied by $\mathcal{C}_{ab}$ together with $\{i_c = i_d = i_a\}$, which are conditionally independent. By Lemma 1, the latter has probability at least $\mathbb{P}'_i[i_c = i_d]\frac{1}{i} \geq \frac{1}{i^2}$. Hence,

$$\mathbb{P}'_i[Q_1] - \mathbb{P}'_i[Q_2] \geq \mathbb{P}'_i[\mathcal{C}_{ab}]\frac{1}{i^2}. \tag{20}$$

By a similar argument in the caterpillar case, we have

$$\begin{aligned}\mathbb{P}'_i[Q_1] - \mathbb{P}'_i[Q_2] &\geq \mathbb{P}'_i\big[K \cap \mathcal{NC}^c\big] \\ &\geq \mathbb{P}'_i[\mathcal{C}_{ab}]\frac{1}{i} \\ &\geq \mathbb{P}'_i[\mathcal{C}_{ab}]\frac{1}{i^2}.\end{aligned} \tag{21}$$

It remains to bound $\mathbb{P}'_i[\mathcal{C}_{ab}]$.

LEMMA 16. *We have*

$$\mathbb{P}'_i[\mathcal{C}_{ab}] \geq \left\{\gamma \wedge \frac{1}{8}\right\}\frac{1}{i}.$$

PROOF. Similar to the proof of Lemma 1, for copy ℓ at R, let N_ℓ be the number of its descendant copies at R', the most recent common ancestor of A and B, and let $J = \sum_{\ell=1}^{i} N_\ell$. We consider two cases for N_ℓ:

1. In the case $N_\ell = 1$ and $\{i_a = i_b = \ell\}$, let Y_ℓ be the indicator function of the event that the lineages from a and b coalesce before R. So $Y_\ell = 1$ if the lineages from a and b coalesce before R, and 0 otherwise. Under the standard coalescent, the probability of that coalescent event is at least γ. Here, though, we are working under the bounded coalescent. In the case that a daughter edge is ancestral to the lineages from a and b, the additional conditioning on complete coalescence only increases the probability that a and b coalesce before R. So γ remains a lower bound.

2. In the case $N_\ell \geq 2$ and $\{i_a = i_b = \ell\}$, let Z_ℓ be the indicator function of the event that the lineages from a and b coalesce before R. So $Z_\ell = 1$ if the lineages from a and b coalesce before R, and 0 otherwise. Since $N_\ell \geq 2$, there is at least one duplication below ℓ before R'. By symmetry, there is probability at least $1/2$ that the first duplication produces a daughter edge with at least half of the descendants of ℓ below it at R'. Under $\{i_a = i_b = \ell\}$, there is then a probability at least $1/4$ that a and b descend from copies at R' below that daughter edge. So overall there is probability at least $1/8$ that $Z_\ell = 1$ in that case.

Putting these two cases together, we get

$$\begin{aligned}
\mathbb{P}'_i[\mathcal{C}_{ab}] &= \mathbb{E}'_i\Big[\mathbb{E}'_i\Big[\sum_{\ell:N_\ell=1}\Big(\frac{N_\ell}{J}\Big)^2 Y_\ell + \sum_{\ell:N_\ell>1}\Big(\frac{N_\ell}{J}\Big)^2 Z_\ell | (N_\ell)_{\ell=1}^i\Big]\Big] \\
&= \mathbb{E}'_i\Big[\sum_{\ell:N_\ell=1}\Big(\frac{N_\ell}{J}\Big)^2 \mathbb{E}'_i[Y_\ell|N_\ell] + \sum_{\ell:N_\ell>1}\Big(\frac{N_\ell}{J}\Big)^2 \mathbb{E}'_i[Z_\ell|N_\ell]\Big] \\
&\geq \Big\{\gamma \wedge \frac{1}{8}\Big\}\mathbb{E}'_i\Big[\sum_{\ell:N_\ell=1}\Big(\frac{N_\ell}{J}\Big)^2 + \sum_{\ell:N_\ell>1}\Big(\frac{N_\ell}{J}\Big)^2\Big] \\
&\geq \Big\{\gamma \wedge \frac{1}{8}\Big\}\frac{1}{i},
\end{aligned}$$

as in [22], Lemma 1, proving the claim. □

LEMMA 17. *We have*

$$\delta' \geq \frac{2}{3}\Big\{\gamma \wedge \frac{1}{8}\Big\}\frac{\sigma^3}{\alpha^3}.$$

PROOF. By (19), (20), (21), and Lemma 16,

$$\begin{aligned}
\delta' &= \frac{2}{3}(\mathbb{P}'[Q_1] - \mathbb{P}'[Q_2]) \\
&\geq \frac{2}{3}\Big\{\gamma \wedge \frac{1}{8}\Big\}\mathbb{E}'\Big[\frac{1}{I^3}\Big] \\
&\geq \frac{2}{3}\Big\{\gamma \wedge \frac{1}{8}\Big\}\frac{1}{\mathbb{E}'[I]^3},
\end{aligned} \tag{22}$$

where the last line follows from Jensen's inequality. Moreover

$$\begin{aligned}
\alpha &\geq \mathbb{E}[I] \\
&= \mathbb{E}[I|\mathcal{X}_Q \geq \vec{1}]\mathbb{P}[\mathcal{X}_Q \geq \vec{1}] + \mathbb{E}\big[I|\{\mathcal{X}_Q \geq \vec{1}\}^c\big]\mathbb{P}\big[\{\mathcal{X}_Q \geq \vec{1}\}^c\big] \\
&\geq \mathbb{E}'[I]\sigma.
\end{aligned}$$

Plugging back into (22) gives the claim. □

4.4. *Final analysis*.

PROOF OF THEOREM 1. The following, adapted from [54], Lemmas A.1 and A.2, gives a bound on the k^* required to reconstruct the correct species tree with probability $1-\epsilon$ in terms of δ'

$$k^* > 2\log\left(\frac{n^4}{\epsilon}\right)\frac{1}{(\delta')^2}.$$

In particular, for $\epsilon < 1$ this inequality holds whenever

$$k^* > 8\log\left(\frac{n}{\epsilon}\right)\frac{1}{(\delta')^2}.$$

By Lemmas 15 and 17, it suffices to have

$$\begin{aligned} k &\geq \left\{\frac{16}{\sigma}\log\left(\frac{n}{\epsilon}\right)\frac{1}{(\frac{2}{3}\{\gamma \wedge \frac{1}{8}\}\frac{\sigma^3}{\alpha^3})^2}\right\} \vee \left\{\frac{8}{\sigma^2}\log\frac{n}{\epsilon}\right\} \\ &\geq \frac{2304\alpha^6}{\sigma^7\gamma^2}\log\frac{n}{\epsilon}. \end{aligned}$$

The claim follows from Lemmas 13 and 14. □

5. Concluding remarks. Through a probabilistic analysis of the DLCoal model, we established identifiability of the model species tree and statistical consistency of quartet-based species tree estimation methods ASTRAL-one and ASTRAL-multi. In our main new result, we derived an upper bound on the required number of gene trees to reconstruct the species tree with high probability. In particular, we highlighted the roles of the branching process parameters λ and μ as well as the tree depth Δ. These parameters enter naturally through two relevant quantities: the minimum survival probability of a quartet (σ) and the maximum expected number of lineages at a vertex (α).

Our results suggest many open problems. First, can we derive a lower bound (and matching upper bound) on the required number of gene trees for ASTRAL-one and similar methods (including ASTRAL-multi)? In particular, is our dependence on α (in the supercritical regime) and σ (in the subcritical regime) optimal? An improvement on the polynomial dependence on α (see Lemma 17) is likely possible with a more detailed analysis of the events in Section 3. But, perhaps more importantly, are there alternative ways of processing multilabeled gene trees (not necessarily quartet-based) that dampen or even exclude the effect of α?

More generally, it would be interesting to obtain statistical consistency and sample complexity results for models also including LGT, under which at low enough rates quartet-based methods have also been shown to be consistent [47]. One such more general model was recently introduced in [23].

Funding. SR was supported by NSF Grants DMS-1614242, CCF-1740707 (TRIPODS), DMS-1902892, DMS-1916378 and DMS-2023239 (TRIPODS Phase II), as well as a Simons Fellowship and a Vilas Associates Award.

BL was supported by NSF Grants DMS-1614242, CCF-1740707 (TRIPODS), DMS-1902892 and a Vilas Associates Award (to SR).

MH was supported by NSF Grant DMS-1902892 and DMS-2023239 (TRIPODS Phase II) as well as a Vilas Associates Award (to SR).

REFERENCES

[1] ALLMAN, E. S., BAÑOS, H. and RHODES, J. A. (2019). NANUQ: A method for inferring species networks from gene trees under the coalescent model. *Algorithms Mol. Biol.* **14** 24. https://doi.org/10.1186/s13015-019-0159-2

[2] 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. *J. Math. Biol.* **62** 833–862. MR2795698 https://doi.org/10.1007/s00285-010-0355-7

[3] ALLMAN, E. S., LONG, C. and RHODES, J. A. (2019). Species tree inference from genomic sequences using the log-det distance. *SIAM J. Appl. Algebra Geom.* **3** 107–127. MR3922905 https://doi.org/10.1137/18M1194134

[4] ANÉ, C., HO, L. S. T. and ROCH, S. (2017). Phase transition on the convergence rate of parameter estimation under an Ornstein–Uhlenbeck diffusion on a tree. *J. Math. Biol.* **74** 355–385. MR3590687 https://doi.org/10.1007/s00285-016-1029-x

[5] ARVESTAD, L., LAGERGREN, J. and SENNBLAD, B. (2009). The gene evolution model and computing its associated probabilities. *J. ACM* **56** Art. 7. MR2535880 https://doi.org/10.1145/1502793.1502796

[6] ATHREYA, K. B. and NEY, P. E. (1972). *Branching Processes*. *Die Grundlehren der Mathematischen Wissenschaften*, *Band* 196. Springer, New York. MR0373040

[7] BORGS, C., CHAYES, J. T., MOSSEL, E. and ROCH, S. (2006). The Kesten–Stigum reconstruction bound is tight for roughly symmetric binary channels. In *FOCS* 518–530.

[8] DASARATHY, G., NOWAK, R. and ROCH, S. (2015). Data requirement for phylogenetic inference from multiple loci: A new distance method. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **12** 422–432.

[9] DASKALAKIS, C., MOSSEL, E. and ROCH, S. (2011). Evolutionary trees and the Ising model on the Bethe lattice: A proof of Steel's conjecture. *Probab. Theory Related Fields* **149** 149–189. MR2773028 https://doi.org/10.1007/s00440-009-0246-2

[10] DEGNAN, J. H. (2018). Modeling hybridization under the network multispecies coalescent. *Syst. Biol.* **67** 786–799.

[11] DEGNAN, J. H. and ROSENBERG, N. A. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genet.* **2** e68.

[12] DEGNAN, J. H. and ROSENBERG, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* **24** 332–340.

[13] DRUMMOND, A. J. and BEAST, A. R. (2007). Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7** 214.

[14] DU, P., HAHN, M. W. and NAKHLEH, L. (2019). Species tree inference under the multispecies coalescent on data with paralogs is accurate. bioRxiv.

[15] FAN, W.-T. and ROCH, S. (2018). Necessary and sufficient conditions for consistent root reconstruction in Markov models on trees. *Electron. J. Probab.* **23** Paper No. 47. MR3814241 https://doi.org/10.1214/18-ejp165

[16] FELSENSTEIN, J. (2003). *Inferring Phylogenies*. Sinauer.

[17] GALTIER, N. (2007). A model of horizontal gene transfer and the bacterial phylogeny problem. *Syst. Biol.* **56** 633–642.

[18] GANESH, A. and ZHANG, Q. (2019). Optimal sequence length requirements for phylogenetic tree reconstruction with indels. In *STOC'*19*—Proceedings of the* 51*st Annual ACM SIGACT Symposium on Theory of Computing* 721–732. ACM, New York. MR4003378 https://doi.org/10.1145/3313276.3316345

[19] GASCUEL, O., ed. (2007). Mathematics of Evolution and Phylogeny. Oxford Univ. Press, Oxford. MR2389362

[20] KINGMAN, J. F. C. (1982). The coalescent. *Stochastic Process. Appl.* **13** 235–248. MR0671034 https://doi.org/10.1016/0304-4149(82)90011-4

[21] LARGET, B. R., KOTHA, S. K., DEWEY, C. N. and ANÉ BUCKY, C. (2010). Gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics* **26** 2910–2911.

[22] LEGRIED, B., MOLLOY, E. K., WARNOW, T. and ROCH, S. (2020). Polynomial-time statistical estimation of species trees under gene duplication and loss. In *Research in Computational Molecular Biology*. *Lecture Notes in Computer Science* **12074** 120–135. Springer, Cham. MR4139323 https://doi.org/10.1007/978-3-030-45257-5_8

[23] LI, Q., GALTIER, N., SCORNAVACCA, C. and CHAN, Y.-B. (2020). The multilocus multispecies coalescent: A flexible new model of gene family evolution. bioRxiv.

[24] LINZ, S., RADTKE, A. and VON HAESELER, A. (2007). A likelihood framework to measure horizontal gene transfer. *Mol. Biol. Evol.* **24** 1312–1319.

[25] MADDISON, W. (1997). Gene trees in species trees. *Syst. Biol.* **46** 523–536.

[26] MARKIN, A. and EULENSTEIN, O. (2020). Quartet-based inference methods are statistically consistent under the unified duplication-loss-coalescence model.
[27] MATSEN, F. A. and STEEL, M. (2007). Phylogenetic mixtures on a single tree can mimic a tree of another topology. *Syst. Biol.* **56** 767–775.
[28] MENG, C. and SALTER KUBATKO, L. (2009). Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. *Theor. Popul. Biol.* **75** 35–45.
[29] MIHAESCU, R., HILL, C. and RAO, S. (2013). Fast phylogeny reconstruction through learning of ancestral sequences. *Algorithmica* **66** 419–449. MR3028647 https://doi.org/10.1007/s00453-012-9644-4
[30] MIRARAB, S., REAZ, R., BAYZID, M. S., ZIMMERMANN, T., SWENSON, M. S. and WARNOW, T. (2014). ASTRAL: Genome-scale coalescent-based species tree estimation. *Bioinformatics* **30** i541–i548.
[31] MOSSEL, E. (2003). On the impossibility of reconstructing ancestral data and phylogenies. *J. Comput. Biol.* **10** 669–678.
[32] MOSSEL, E. (2004). Phase transitions in phylogeny. *Trans. Amer. Math. Soc.* **356** 2379–2404. MR2048522 https://doi.org/10.1090/S0002-9947-03-03382-8
[33] MOSSEL, E. (2004). Survey: Information flow on trees. In *Graphs, Morphisms and Statistical Physics. DIMACS Ser. Discrete Math. Theoret. Comput. Sci.* **63** 155–170. Amer. Math. Soc., Providence, RI. MR2056226
[34] MOSSEL, E. and PERES, Y. (2003). Information flow on trees. *Ann. Appl. Probab.* **13** 817–844. MR1994038 https://doi.org/10.1214/aoap/1060202828
[35] MOSSEL, E. and ROCH, S. (2012). Phylogenetic mixtures: Concentration of measure in the large-tree limit. *Ann. Appl. Probab.* **22** 2429–2459. MR3024973 https://doi.org/10.1214/11-AAP837
[36] MOSSEL, E. and ROCH, S. (2017). Distance-based species tree estimation under the coalescent: Information-theoretic trade-off between number of loci and sequence length. *Ann. Appl. Probab.* **27** 2926–2955. MR3719950 https://doi.org/10.1214/16-AAP1273
[37] MOSSEL, E., ROCH, S. and SLY, A. (2011). On the inference of large phylogenies with long branches: How long is too long? *Bull. Math. Biol.* **73** 1627–1644. MR2802440 https://doi.org/10.1007/s11538-010-9584-6
[38] MOSSEL, E. and STEEL, M. (2004). A phase transition for a random cluster model on phylogenetic trees. *Math. Biosci.* **187** 189–203. MR2043825 https://doi.org/10.1016/j.mbs.2003.10.004
[39] NAKHLEH, L. (2013). Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends Ecol. Evol.* **28** 719–728.
[40] RABIEE, M., SAYYARI, E. and MIRARAB, S. (2019). Multi-allele species reconstruction using ASTRAL. *Mol. Phylogenet. Evol.* **130** 286–296. https://doi.org/10.1016/j.ympev.2018.10.033
[41] RANNALA, B., EDWARDS, S. V., LEACHÉ, A. and YANG, Z. (2020). The multi-species coalescent model and species tree inference. In *Phylogenetics in the Genomic Era* (C. Scornavacca, F. Delsuc and N. Galtier, eds.) 3.3:1–3.3:21. No commercial publisher | Authors open access book.
[42] RANNALA, B. and YANG, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using dna sequences from multiple loci. *Genetics* **164** 1645–1656.
[43] RASMUSSEN, M. D. and KELLIS, M. (2012). Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Res.* **22** 755–765.
[44] ROCH, S. (2010). Toward extracting all phylogenetic information from matrices of evolutionary distances. *Science* **327** 1376–1379. MR2650508 https://doi.org/10.1126/science.1182300
[45] ROCH, S., NUTE, M. and WARNOW, T. (2018). Long-branch attraction in species tree estimation: Inconsistency of partitioned likelihood and topology-based summary methods. *Syst. Biol.* **68** 281–297.
[46] ROCH, S. and SLY, A. (2017). Phase transition in the sample complexity of likelihood-based phylogeny inference. *Probab. Theory Related Fields* **169** 3–62. MR3704765 https://doi.org/10.1007/s00440-017-0793-x
[47] ROCH, S. and SNIR, S. (2013). Recovering the treelike trend of evolution despite extensive lateral genetic transfer: A probabilistic analysis. *J. Comput. Biol.* **20** 93–112. MR3021672 https://doi.org/10.1089/cmb.2012.0234
[48] ROCH, S. and STEEL, M. (2015). Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theor. Popul. Biol.* **100** 56–62.
[49] ROCH, S. and WARNOW, T. (2015). On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. *Syst. Biol.* **64** 663–676.
[50] SCHREMPF, D. and SZÖLLÖSI, G. (2020). The sources of phylogenetic conflicts. In *Phylogenetics in the Genomic Era* (C. Scornavacca, F. Delsuc and N. Galtier, eds.) 3.1:1–3.1:23. No commercial publisher | Authors open access book.
[51] SCORNAVACCA, C., DELSUC, F. and GALTIER, N. (2020). *Phylogenetics in the Genomic Era*. No commercial publisher | Authors open access book.

[52] SÉBASTIEN, R. (2019). Hands-on Introduction to Sequence-Length Requirements in Phylogenetics. In *Bioinformatics and Phylogenetics* 47–86. Springer, Cham.

[53] SEMPLE, C. and STEEL, M. (2003). *Phylogenetics*. *Oxford Lecture Series in Mathematics and Its Applications* **24**. Oxford Univ. Press, Oxford. MR2060009

[54] SHEKHAR, S., ROCH, S. and MIRARAB, S. Species tree estimation using ASTRAL: How many genes are enough? *IEEE/ACM Trans. Comput. Biol. Bioinform.* **15** 1738–1747. https://doi.org/10.1109/TCBB.2017.2757930

[55] SIMION, P., DELSUC, F. and TO, H. P. (2020). What extent current limits of phylogenomics can be overcome? In *Phylogenetics in the Genomic Era* (C. Scornavacca, F. Delsuc and N. Galtier, eds.) 2.1:1–2.1:34. No commercial publisher | Authors open access book.

[56] SOLÍS-LEMUS, C. and ANÉ, C. (2016). Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genet.* **12** 1–21.

[57] STEEL, M. (2016). *Phylogeny—Discrete and Random Processes in Evolution*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **89**. SIAM, Philadelphia, PA. MR3601108 https://doi.org/10.1137/1.9781611974485.ch1

[58] VERSHYNIN, R. (2018). *High-Dimensional Probability*: *An Introduction with Applications in Data Science*. *Cambridge Series in Statistical and Probabilistic Mathematics* **47**. Cambridge Univ. Press, Cambridge. MR3837109 https://doi.org/10.1017/9781108231596

[59] WARNOW, T. (2017). *Computational Phylogenetics*: *An Introduction to Designing Methods for Phylogeny Estimation*. Cambridge Univ. Press, Cambridge.

[60] YANG, Z. (2014). *Molecular Evolution*: *A Statistical Approach*. OUP, Oxford.