ASTRAL-Pro: quartet-based species tree inference despite paralogy

Chao Zhang,[1] Celine Scornavacca,[2] Erin K. Molloy,[3] and Siavash Mirarab[4*]

[1] *Bioinformatics and Systems Biology, UC San Diego, CA, USA*
[2] *ISEM, Université de Montpellier, CNRS, IRD, EPHE, Montpellier, France*
[3] *Department of Computer Science, University of Illinois at Urbana-Champaign, IL, USA*
[4] *Department of Electrical and Computer Engineering, UC San Diego, CA, USA*
[*] Corresponding author: smirarab@ucsd.edu

Abstract

Species tree inference via summary methods that combine gene trees has become an increasingly common analysis in recent phylogenomic studies. This broad adoption has been partly due to the greater availability of genome-wide data and ample recognition that gene trees and species trees can differ due to biological processes such as gene duplication and gene loss. This increase has also been encouraged by the recent development of accurate and scalable summary methods, such as ASTRAL. However, most of these methods, including ASTRAL, can only handle single-copy gene trees and do not attempt to model gene duplication and gene loss. In this paper, we introduce a measure of quartet similarity between single-copy and multi-copy trees (accounting for orthology and paralogy relationships) that can be optimized via a scalable dynamic programming similar to the one used by ASTRAL. We then present a new quartet-based species tree inference method: ASTRAL-Pro (ASTRAL for PaRalogs and Orthologs). By studying its performance on an extensive collection of simulated datasets and on a real plant dataset, we show that ASTRAL-Pro is more accurate than alternative methods when gene trees differ from the species tree due to the simultaneous presence of gene duplication, gene loss, incomplete lineage sorting, and estimation errors.

Keywords: Species tree inference, gene duplication, gene loss, incomplete lineage sorting

1 Introduction

Evolutionary histories of genes and species can differ for several reasons [1], including incomplete lineage sorting (ILS), duplication and loss (duploss), gene transfer, and hybridization. Species tree inference is a central question in evolutionary biology and dealing with these sources of gene tree discordance is crucial. Many approaches have been proposed for species tree inference, including gene trees-species tree co-estimation [2–6] and species tree inference from sequence data [7–9]. However, the most scalable approach has remained a two-step process: first infer gene trees independently from sequence data and then combine them using a summary method. The goal of a summary method is to find the species tree best explaining the gene trees according to a model of gene tree discordance. While the ultimate goal is to develop summary methods modelling all sources of discordance, the literature mostly focused on separate causes.

A major family of summary methods focuses on duplication and loss processes producing multi-copy gene trees [10–15]. Most of these summary methods rely on maximum parsimony reconciliation [16] and aim at finding the species tree with the minimum reconciliation cost. Example methods include DupTree [10], its later extension iGTP [11, 12], DynaDup [15] and earlier similar dynamic programming algorithms [14]. Other methods take a more agnostic approach and minimize the distance between species trees and the gene trees without necessarily invoking specific reasons for discordance. Example methods of this type include MulRF [17] and *guenomu* [18]. However, a recent result asserts that the optimal solution to the optimization problem solved by MulRF is indeed a statistically consistent estimate of the species tree under a generic duplication-only model of gene evolution [19]. These methods are mostly designed to

handle duplication and loss, and although in simulations some have reasonable accuracy under ILS and gene transfer [20], they have not been widely adopted by biologists.

Several summary methods target ILS as modelled by the multi-species coalescence (MSC) model [21, 22], and many of them are statistically consistent [e.g., 23–30]. However, the most successful summary method for ILS has arguably been ASTRAL [31], which, due to its high accuracy [32–34] and scalability [35, 36], has been used widely in biological analyses. ASTRAL, like several other methods [7, 24, 28], relies on dividing gene trees into unrooted four-taxon trees (called quartets), a feature that allows it to handle ILS and may contribute to its high accuracy. ASTRAL, however, was designed to handle single-copy gene trees reconstructed from sets of orthologous genes. This limitation has restrained the application scope of ASTRAL. For example, two studies on plant transcriptomes had to restrict species tree analyses with ASTRAL to the 400–800 putative single-copy gene trees, discarding thousands of available multi-copy genes [37, 38]. A surprising result asserts that treating gene copies as alleles of a same gene, a feature ASTRAL supports [39], is statistically consistent under a standard parametric model of gene duplication and loss and may be accurate [40]. Others have shown that random sampling of leaves works well empirically [41]. Beyond ASTRAL, several methods have focused on dividing multi-copy gene trees into single-copy genes without apparent duplications [42–46]. However, to our knowledge, no quartet-based methods *designed* to handle duplication and loss currently exist. Extending quartet-based methods to multi-copy gene trees is not trivial if we seek to correctly model orthology and paralogy.

Here, we introduce a quartet-based species tree inference method called ASTRAL-Pro (ASTRAL for PaRalogs and Orthologs). This method requires defining a measure of quartet similarity between single-copy and multi-copy trees accounting for orthology and paralogy. We define such a measure in a principled manner and show how to optimize it using dynamic programming. We test the method on an extensive collection of simulated datasets and on a real plant dataset.

2 Problem Definition

Let $\mathcal{S}$ be a set of n species. Let us suppose that we are given a set of binary gene trees $\mathcal{G}$, and, for each tree $G \in \mathcal{G}$ with leaf set $\mathcal{L}_G = \{1 \dots m_G\}$, we have a mapping $\alpha_G : \mathcal{L}_G \to \mathcal{S}$ specifying in which species each gene is sampled. For a rooted tree G, we denote the set of internal nodes in G by $I(G)$, and, for each $u \in I(G)$, we define $\mathcal{L}_G(u)$ as the set of leaves below u. We define two short-hands: $\alpha_G(A) = \{\alpha_G(i) : i \in A\}$ for $A \subset \mathcal{L}_G$ and $\alpha_G(u) = \alpha_G(\mathcal{L}_G(u))$ for a node u. The notation $G \restriction A$ denotes G restricted to the set A.

We let $\Omega(G)$ be the multi-labelled tree obtained by replacing each leaf $i \in \mathcal{L}_G$ with $\alpha_G(i)$. Multiple copies of the same species in a gene tree G may be created by gene duplication. We assume that each duplication creates a new genomic locus (i.e., a position along the genome) and therefore, each locus, except the original one, has a parent locus (which may or may not have survived to the present day). Thus, each element of $\mathcal{L}_G$ can be theoretically mapped to its parent locus, allowing us to "trace" the locus of each leaf to its ancestors.

In each gene tree G, we refer to a subset Q of four distinct elements of $\mathcal{L}_G$ as a quartet. The subtree of a fully resolved tree G induced on a quartet Q exhibits two degree-three nodes. We refer to these nodes as *anchors of Q on G*. As shown in Fig. 1, for a rooted tree G and for a quartet Q, up to label permutations, $G \restriction Q$ can only have two topologies: an *unbalanced* one (when one anchor descends from the other), denoted as $Q \angle G$, and a *balanced* one (otherwise), denoted as $Q \perp G$. We say a tripartition (P_1, P_2, P_3) of $\mathcal{S}$ "can anchor" a quartet Q of G iff $\forall_i : P_i \cap \alpha_G(Q) \neq \emptyset$.

Definition 1 (Tagged trees)**.** We say a rooted tree G is tagged if every internal node is tagged either as duplication or as speciation. A node u with children u_1 and u_2 can be a speciation if the three sets $\alpha_G(u_1)$, $\alpha_G(u_2)$, and $\alpha_G(\mathcal{L}_G \setminus (\mathcal{L}_G(u_1) \cup \mathcal{L}_G(u_2)))$ are mutually exclusive.

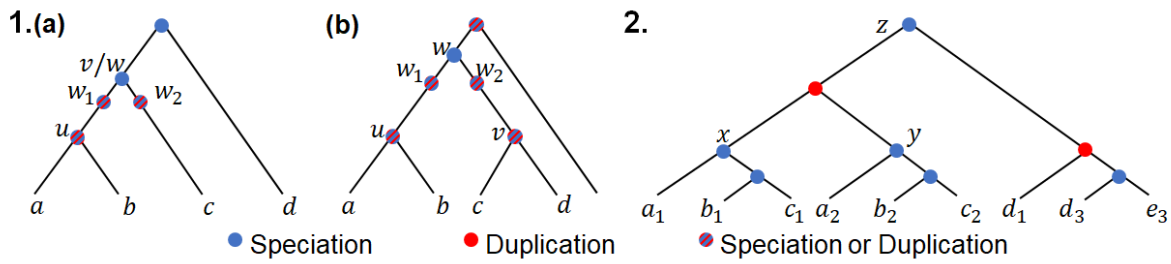


Figure 1: 1. An example of a quartet $Q = \{a, b, c, d\}$ with (a) unbalanced topology $(Q \angle G)$ and (b) balanced topology $(Q \perp G)$. Anchors are u and v, and $w = \psi_G(Q)$ is the anchor LCA. While w has to be a speciation for Q to be considered a SQ, u and v can be either speciation or duplication. 2. An example of equivalent classes. Three equivalent classes are anchored on z: all eight quartets of the form $\{a_i, b_j, d_k, e_3\}$, of the form $\{a_i, c_j, d_k, e_3\}$, and of the form $\{b_i, c_j, d_k, e_3\}$, all with balanced topology. Anchored on x: two equivalent classes with unbalanced topology: $\{a_1, b_1, c_1, d_1\} \sim \{a_1, b_1, c_1, d_3\}$ and $\{a_1, b_1, c_1, e_3\}$. Anchored on y: two equivalent classes: $\{a_2, b_2, c_2, d_1\} \sim \{a_2, b_2, c_2, d_3\}$ and $\{a_2, b_2, c_2, e_3\}$.

We note that these labels may or may not correspond to real speciation and duplication events. In particular, in the presence of deep coalescence across duplication events, a correct tagging corresponding to actual events may not be possible.

Definition 2 (SQ)**.** A quartet Q on a rooted tagged gene tree G is called a speciation-driven quartet (SQ) iff $|\alpha_G(Q)| = 4$ and the LCA of any three out of four leaves of Q is a speciation node. Equivalently, a quartet with topology $ab|cd$ is a SQ if and only if its genes are all contained in different species and the LCA of either a or b with either c or d is tagged as speciation.

Definition 3 (Quartet anchor LCA)**.** Let u and v be anchors of a quartet Q on a *rooted* tree G. We refer to the LCA of u and v as the *anchor LCA* of Q on G and denote it as $\psi_G(Q)$.

The last definition is central to our approach. Note that anchors of a SQ can be speciations or duplications (Fig. 1) and thus SQs are not simply quartets with anchors being speciation nodes. Instead, they are quartets with a topology pre-determined by the speciation event represented by the anchor LCA, regardless of subsequent duplications and losses. Such subsequent duplications and losses may lead to multiple quartets originating from the same speciation event. Since these events include no new information on the speciation event, we count only SQs towards the quartet score of a species tree and weight them in a non-trivial way to avoid double-counting.

Definition 4 (Equivalent SQs)**.** Two SQs on the same 4 species are *equivalent* if they have the same anchor LCA; i.e., for two SQs, $Q_1 \sim Q_2 \iff \alpha_G(Q_1) = \alpha_G(Q_2) \wedge \psi_G(Q_1) = \psi_G(Q_2)$.

Proposition 1. *If Q_1 and Q_2 are equivalent SQs on G, then $\Omega(G \restriction Q_1)$ and $\Omega(G \restriction Q_2)$ are isomorphic.*

Thus, equivalent SQs have the same quartet topology when mapped to species. Proofs of all propositions and lemmas and sketches of proofs of all claims can be found in Appendix A.

Proposition 1 tells us that equivalent SQs do not provide any extra information with respect to each other, and therefore, it is reasonable to count all equivalent SQs as one unit when computing the quartet score of a species tree. This intuition is backed by the following proposition:

Proposition 2. *Assuming a correctly tagged tree G, for all equivalent SQs with a shared anchor LCA w, the three (in the unbalanced case) or four (in the balanced case) quartet leaves below w will all share an ancestral locus at the time of the speciation event corresponding to w.*

We can now provide a natural definition of the quartet score. The equivalence relation (Def. 4) partitions all quartets in equivalence classes and, by Proposition 1, for each equivalent class, we can define a unique quartet tree labelled by $\mathcal{S}$. By Proposition 2, each class corresponds to an ancestral locus.

Definition 5 (Per-locus (PL) Quartet Score)**.** The per-locus quartet score of a species tree S with respect to a tagged gene tree G is the number of equivalent quartet classes with a quartet topology that matches the quartet tree induced by S. More formally,

$$q(S,G) = |\{(\alpha_G(Q), \psi_G(Q)) : Q \subset \mathcal{L}_G, |Q| = |\alpha_G(Q)| = 4, \Omega(G \restriction Q) \simeq S \restriction \alpha_G(Q)\}| \,.$$

The PL quartet score of S with respect to a set of tagged gene trees $\mathcal{G}$ is $q(S,\mathcal{G}) = \sum_{G\in\mathcal{G}} q(S,G)$.

Claim 1. *If all nodes on the path between the root r and a node u are tagged as speciations, changing the root to any branch on the path does not alter the PL quartet score.*

Problem 1 (Maximum per-Locus Quartet score Species Tree (MLQST) problem)**.** *Given a set of rooted tagged gene trees $\mathcal{G}$, find the species tree that maximizes the PL quartet score with respected to input gene trees, i.e.,* $\arg\max_S q(S,\mathcal{G})$.

3 Solving the MLQST problem using dynamic programming

We start by briefly describing the ASTRAL algorithm to solve a related problem (the MQSST problem), and then describe how we extend this approach to the MLQST problem.

3.1 Background: ASTRAL on single-copy gene trees.

A node in a binary unrooted species tree forms a tripartition of $\mathcal{S}$ that implies the topology for all quartets anchored at that node, allowing us to score it against $\mathcal{G}$. Let $P = P_1|P_2|P_3$ and $M = M_1|M_2|M_3$ be two tripartitions, and let $I_{ij} = |M_i \cap P_j|$. Any species tree that displays P will share a certain number of quartets with any gene tree that displays M, and we call this number $QI(P,M)$ (calculations below extends to multifurcations if M is a d-partition). Defining G_3 as the set of all permutations of $\{1,2,3\}$, we have [31, 47]:

$$w(P) = \frac{1}{2}\sum_{G\in\mathcal{G}}\sum_{M\in\mathcal{P}(G)} QI(P,M) \quad \text{where} \quad QI(P,M) = \sum_{(i,j,k)\in G_3} \frac{I_{i1}I_{j2}I_{k3}(I_{i1}+I_{j2}+I_{k3}-3)}{2} \tag{1}$$

and $\mathcal{P}(G)$ is the set of partitions representing internal nodes of G. The quartet score of a species tree is simply the sum of the weights of its tripartitions. The division by half in $w(P)$ is necessary because the sum counts each shared quartet twice (once at each anchor).

ASTRAL finds the tree S that maximizes the quartet score using dynamic programming. It recursively divides $\mathcal{S}$ into subsets, in each step, choosing the division that maximizes the sum of the weights. To avoid exponential running time, instead of considering all ways of partitioning a set $A \subset \mathcal{S}$ into A' and $A \setminus A'$, we constrain the recursion to a given set of bipartitions. Let X be this set and $X' = \{A : A|(\mathcal{S}\setminus A) \in X\}$ and $Y = \{(C,D) : C \in X', D \in X', C \cap D = \emptyset, C \cup D \in X'\}$. Let $V(A)$ be the quartet score of an optimal subtree on the cluster A and set $V(\{a\}) = 0$. Then,

$$V(A) = \max_{(A',A\setminus A')\in Y} V(A') + V(A\setminus A') + w(A'|(A\setminus A')|(\mathcal{S}\setminus A)) \tag{2}$$

3.2 ASTRAL-Pro Algorithm

We extend here ASTRAL to multi-copy gene trees. The input to the new method, called ASTRAL-Pro (A-Pro for short), is a set of rooted tagged gene trees (see in Section 3.3 how unrooted gene trees can be rooted and tagged). This extension involves three changes. (i) To handle multi-copy gene trees, when computing the tripartition associated to each node, we use

α_G to map labels to $\mathcal{S}$. Here, instead of multi-sets, we create sets (counting multiple copies on each side only once). (ii) We change the weight calculation $w(P)$ so that each equivalent class of quartets is counted once instead of twice, only at its LCA anchor. (iii) When computing w, we only sum over internal nodes tagged as speciations.

3.2.1 Weight calculation.

Let w be an internal node of G tagged as speciation and $P = (P_1|P_2|P_3)$ be a tripartition of $\mathcal{S}$.

Definition 6. We say that a SQ equivalent class with LCA anchor w in a gene tree G is mapped from left to a species tree tripartition P iff for each quartet Q in the equivalent class (i) P can anchor Q and (ii) the leaves a and b under the anchor of Q that appear first in a post-order traversal of G (e.g., u in Fig. 1) both map to the same side of P (that is, $\alpha_G(a) \in P_i, \alpha_G(b) \in P_i$ for some $1 \leq i \leq 3$). We denote such quartets by $Q \xrightarrow{w} P$.

We now state a set of lemmas, followed by the main result.

Lemma 1. *If $Q_1 \sim Q_2$ and $Q_1 \xrightarrow{w} P$, then $Q_2 \xrightarrow{w} P$.*

Lemma 2. *For a speciation node w with left child w_1 and right child w_2, let $M_1 = \alpha_G(w_1)$, $M_2 = \alpha_G(w_2)$ and $M_3 = \{\alpha_G(z) : z \in \mathcal{L}_G \setminus \mathcal{L}_G(w),$ and LCA of w and z is tagged as speciation$\}$. Let $M_w = (M_1|M_2|M_3)$. Recall $I_{ij} = |M_i \cap P_j|$. The number of SQ quartet equivalent classes anchored to w and mapped from left to the species partition P can be counted as follows:*

$$\begin{aligned} QI_{pro}(P, M_w) =& |\{\alpha_G(Q) : Q \subset \mathcal{L}_G, Q \xrightarrow{w} P\}| \\ =& \sum_{(i,j,k)\in G_3, j<k} \binom{I_{1i}}{2} I_{2j} I_{2k} + \sum_{(i,j,k)\in G_3} \frac{I_{1i} I_{2j} I_{3k}(I_{1i} + I_{2j} - 2)}{2} \end{aligned} \tag{3}$$

Lemma 3. *If $\Omega(G \restriction Q) \simeq S \restriction \alpha_G(Q)$, there exists a unique $P \in \mathcal{P}(S)$ satisfying $Q \xrightarrow{\psi_G(Q)} P$.*

Lemma 4. *Let $\mathbf{1}_{speciation}(w)$ be 1 for speciation nodes and 0 for duplication nodes and let*

$$w_{pro}(P) = \sum_{G\in\mathcal{G}} \sum_{w\in I(G)} QI_{pro}(P, M_w) \times \mathbf{1}_{speciation}(w) \ .$$

Then: $q(S, \mathcal{G}) = \sum_{P\in\mathcal{P}(S)} w_{pro}(P)$.

Theorem 1. *The ASTRAL-Pro algorithm obtained by replacing $w(P)$ function with $w_{pro}(P)$ in the ASTRAL dynamic programming solves the MLQST problem exactly if $X = 2^{\mathcal{S}}$.*

Proof. By Lemma 4, $\arg\max_S q(S, \mathcal{G}) = \arg\max_S \sum_{P\in\mathcal{P}(S)} w_{pro}(P)$. Thus, ASTRAL dynamic programming can solve the optimization problem exactly given the full search space (the argument is identical to that of ASTRAL and follows from the additive nature of $q(S, \mathcal{G})$). □

We now make two claims, and provide a sketch of proofs in Appendix A. Note that by Claim 3, ASTRAL-Pro has polynomial running time.

Claim 2. *For a set of gene trees $\mathcal{G}$ including only speciations, the tree returned by ASTRAL-Pro is the same as the one returned by ASTRAL.*

Claim 3. *The asymptotic running time of ASTRAL-Pro is $O(D|X|^{1.73}) = O(D(nN)^{1.73})$ where $N = \sum_{G\in\mathcal{G}} |\mathcal{L}_G|$ and D denotes the number of unique gene tree tripartitions tagged as speciations.*

Algorithm 1 Gene tree tagging and rooting.

```
procedure TagAndRoot(G)
    s ← ∞
    for edge e in G do
        root G at e and let r_e be the new root
        s_e ← Tag(r_e)
        if s_e < s then
            r ← r_e
            s ← s_e
    root at r
    Tag(r)

procedure Tag(u)
    if u is a leaf then
        score(u) ← 0
    else
        u_l, u_r ← children of u
        score(u) ← Tag(u_r) + Tag(u_l)
        if α_G(u_l) ∩ α_G(u_r) = ∅ then
            tag u as Speciation
        else
            tag u as Duplication
            if α_G(u_l) = α_G(u) ∨ α_G(u_r) = α_G(u) then
                if α_G(u_l) = α_G(u_r) then
                    score(u) ← score(u) + 1
                else
                    score(u) ← score(u) + 2
            else
                score(u) ← score(u) + 3
    return score(u)
```

3.3 Tagging and rooting gene trees

Gene trees inferred from sequence data are neither rooted nor tagged. We use the heuristic Algorithm 1 to root and tag gene trees, noting that a partially-correct rooting suffices (Claim 1). Given a rooted tree, it tags a node as duplication *only if* the node cannot be tagged as speciation by Definition 1 (i.e., *observable duplication nodes* [43]); other nodes are *assumed* to be speciation.

For rooting, we seek the root position that minimizes the number of duplications and losses while allowing for free ILS. In each gene tree G, for two nodes u and v where $\alpha_G(u) = \alpha_G(v)$, we explain all differences in topologies below u and v by invoking ILS (as opposed to duplication/loss). Then, three scenarios are possible for a node u with children u_l and u_r. (i) When u is duplication and $\alpha_G(u_l) = \alpha_G(u_l)$, we do not need to invoke any loss. One duplication suffices. (ii) If $\alpha_G(u_l) \subset \alpha_G(u_r)$ or vice versa, we need one loss on u_l and an arbitrary amount of deep coalescence. (iii) Else, we need two losses (one in each side) and ILS to describe the differences. Algorithm 1 computes the number of duplication and loss events using this strategy, without penalizing ILS and fixing a cost of one for both duplications and losses. As described, it requires quadratic time per rooting and thus cubic to find optimal rooting. In our implementation, we used memoization to reduce this time to quadratic (details omitted). The LCA-based linear algorithm of Scornavacca *et al.* [43] could also be adapted.

3.3.1 Search Space

As mentioned before, we need to constrain the ASTRAL search space to bipartitions in a set X. To define X, we use Algorithm 2, which relies on several strategies. It uses a sampling algorithm (SampleFull procedure) to create single-copy versions of each gene tree, creating a set $\mathcal{F}$. This sampling algorithm prunes all right (or left) subtrees below the highest duplication nodes in the tree, and recurses on each pruned tree, until no species has multiple copies. In addition, per gene, 2^C (default: $C = 4$) trees are sampled from $\mathcal{F}$, creating a multiset $\mathcal{I}$. This sampling can be probabilistic (taking each side of a duplication with probability $\frac{1}{2}$) for high numbers of duplications. When the number of input trees is small, $\mathcal{I}$ may become too small; in these cases, $\mathcal{I}$ is augmented using another sampling algorithm (SampleExtra procedure). We provide $\mathcal{I}$ as input to the algorithms implemented in ASTRAL-III for building the initial set X (including all of its heuristic methods). Finally, we complete all trees from $\mathcal{F}$ using the tree completion algorithm of ASTRAL-III and add the resulting bipartitions to X. All methods used guarantee that $|X|$ grows polynomially with the number of species, gene trees, and duplication nodes.

Algorithm 2 Building set X. Default constant parameters: $C = 4$, $E_m = 500$, $E_s = 4$.

```
procedure BuildX(𝒢)
    ℱ = ∅ and ℐ = ∅
    for G ← 𝒢 do
        (M, S) ← SampleFull(G, ℒ_G, C)
        ℱ ← ℱ ∪ S
        ℐ ← ℐ + M
    for G ∈ {randomly sample max(0, min(|𝒢|, (E_m − |𝒢|)/E_s)) trees from 𝒢} do
        ℐ ← ℐ+ SampleExtra(G, ℒ_G)
    X ← run ASTRAL-III method (including heuristics) for building X with ℐ as input
    X ← X∪( all bipartitions of {t completed using the ASTRAL-III tree completion method : t ∈ ℱ})
procedure SampleFull(G, A, c)
    if |α_G(A)| = |A| then
        return (multiset: [Ω(G ↾ A) repeated 2^c times], set: {Ω(G ↾ A)})
    else
        A_l ← ∅ and A_r ← ∅
        G_A ← G ↾ A (degree-2 nodes removed)
        for a ∈ A do
            p ← the highest ancestor of a in G_A tagged as a duplication node (or ∅ if it doesn't exist)
            if (p = ∅) ∨ (a is to the left of p) then
                A_l ← A_l ∪ {a}
            if (p = ∅) ∨ (a is to the right of p) then
                A_r ← A_r ∪ {a}
        (L.m, L.s) ← SampleFull(G, A_l, max(c − 1, 0))
        (R.m, R.s) ← SampleFull(G, A_r, max(c − 1, 0))
        if c = 0 then
            return (multiset: randomly select L.m or R.m with equal probabilities, set: L.s ∪ R.s)
        else
            return (multiset: L.m + R.m, set: L.s ∪ R.s)
procedure SampleExtra(G, A)
    if |α_G(A)| = |A| then
        return multiset [Ω(G) repeated once]
    else
        A_l ← ∅ and A_r ← ∅
        G_A ← G ↾ A (degree-2 nodes removed)
        for a ∈ A do
            p ← the highest ancestor of a in G_A tagged as a duplication node (or ∅ if it doesn't exist)
            if (p ≠ ∅) ∧ (a is to the left of p) then
                A_l ← A_l ∪ {a}
            if (p ≠ ∅) ∧ (a is to the right of p) then
                A_r ← A_r ∪ {a}
        B_l ← {x : x ∈ A_l, α_G(x) ∈ α_G(A_l) \ α_G(A_r)}
        B_r ← {x : x ∈ A_r, α_G(x) ∈ α_G(A_r) \ α_G(A_l)}
        G_L ← G ↾ ((ℒ_G \ A_r) ∪ B_r) (degree-2 nodes removed)
        G_R ← G ↾ ((ℒ_G \ A_l) ∪ B_l) (degree-2 nodes removed)
        for internal node u of G_L where ℒ_G(u) ⊂ B_r do
            B_u ← {one leaf node arbitrarily chosen among {x|α_G(x) = s, x ∈ ℒ_G(u)}|s ∈ α_G(u)}
            replace u with a start tree consisting of leaves from the set B_u
        for internal node u of G_R where ℒ_G(u) ⊂ B_l do
            B_u ← {one leaf node arbitrarily chosen among {x|α_G(x) = s, x ∈ ℒ_G(u)}|s ∈ α_G(u)}
            replace u with a start tree consisting of leaves from the set B_u
        R = SampleExtra(G_L, A_l) + SampleExtra(G_R, A_r)
        return R
```

Table 1: Simulation settings for S25 dataset. n=number of ingroup species; k=number of genes; τ = tree height (generations); λ_+= duplication rate; λ_-= loss rate; N_e= Haploid effective population size. Empirically, we estimate: C = mean number of copies per species minus one when $\lambda_- = 0$ and $n = 25$; ILS= mean RF distance between true gene trees and the species tree when $\lambda_+ = 0$. MGTE = mean RF distance between true and estimated gene tree when $\lambda_+ = 0$. See Table S1 for full parameters and Figures S1–S6 for full statistics.

Default model	$n = 25$; $k = 1000$; $\tau \sim LN(21.25; 0.2)$; $\lambda_+ = 4.9 \times 10^{-10}$; $\lambda_- = \lambda_+$; $N_e = 4.7 \times 10^8$ $C \approx 5$; ILS $\approx 70\%$; MGTE = 15% (500bp) or 36% (100bp)
Controlling λ_+, λ_- (duploss rate)	$\lambda_+ \in \{4.9, 2.7, 1.9, 0.52, 0\} \times 10^{-10}$ $\lambda_- \in \{1, 0.5, 0.1, 0\} \times \lambda_+$ $C \approx \{5, 2, 1, 0.2, 0\}$
Controlling λ_+, N_e (dup rate, ILS)	$\{4.9, 1.9, 0\} \times 10^{-10}$; $N_e \in \{4.7 \times 10^8, 1.9 \times 10^8, 4.8 \times 10^7, 1.0 \times 10^4\}$ ILS $\approx\{70, 52, 20, 0\}\%$; $C \approx \{5, 1, 0\}$ MGTE $\approx \{15, 15, 15, 16\}\%$ (500bp) or $\approx \{36, 36, 36, 35\}\%$ (100bp) as N_e changes
Controlling n	$n \in \{10, 25, 100, 250, 500\}$ MGTE $\approx \{15, 15, 17, 18, 18\}$ (500bp) or $\approx \{34, 36, 40, 43, 43\}$ (100bp)
Controlling k	$k \in \{25, 100, 250, 1000, 2500, 10000\}$

3.4 Statistical Consistency

When the input set $\mathcal{G}$ has only speciation nodes, the MLQST problem reduces to the Maximum Quartet Support Species Tree (MQSST) problem solved by ASTRAL [31]. Thus, like the MQSST, the MLQST is NP-hard [48]. Moreover, the solution to MQSST problem is a statistically consistent estimator of the species tree under the MSC model and thus ASTRAL-Pro is also statistically consistent in absence of duplication.

In the presence of gene duplication and losses only, let us consider the birth-death model proposed by Arvestad *et al.* [49] and refer to it as the GDL model.

Proposition 3. *Under the GDL model, every SQ in every correctly tagged rooted gene tree is isomorphic in topology to the species tree.*

Since all quartets in every equivalence class of SQs match the species tree, the per-locus quartet score will be maximized by the species tree. The following theorem follows.

Theorem 2. *Under the GDL model [49], the solution to the MLQST problem is a statistically consistent estimator of the species tree given correctly rooted and tagged gene trees.*

In fact, we conjecture that ASTRAL-Pro is statistically consistent under the GDL model even when gene trees are imperfectly rooted and tagged. We leave the proof to future work. Finally, note that restricting to X does not impact statistical consistency, as each bipartition of the species tree has a non-zero chance of appearing in output of this algorithm.

4 Experiment setup

4.1 Datasets

We use new and existing simulated datasets as well as a biological dataset to test A-Pro.

4.1.1 New simulated dataset (S25)

We perform a set of simulations using SimPhy [50] starting from a default model condition and adjusting five parameters (Table 1). We simulate 50 replicates per condition, and each replicate draws its parameters from prior distributions. Exact commands are given in Appendix B.

Default model: The species tree, simulated under the Yule process with birth rate 5×10^{-9} and the number of generations sampled from a log-normal distribution (mean 2.9×10^9), has 25 ingroup and an outgroup species. Each replicate has 1000 true gene trees simulated under DLCoal with fixed haploid population size $N_e = 4.7 \times 10^8$. Gene trees have mean ILS level in

[60%, 80%] range (mean 70%) across replicates (Fig. S2). The duplication rate $\lambda_+ = 4.9 \times 10^{-10}$; when there is no loss, gene trees on average include 145 leaves ($\approx$ 5 extra copies per species). The loss rate λ_- is set to λ_+; with loss, gene trees have on average 43 leaves. The average number of duplication and loss events are 11 and 9, respectively, but variance is high (Fig. S1). For each gene, we use Indelible [51] to simulate gap-free nucleotide sequences along the gene trees using the GTR+Γ model [52] with 2 different sequence lengths: 500bp and 100bp. We then use FastTree2 [53] to estimate maximum likelihood gene trees under the GTR+Γ model. Gene tree estimation error, measured by the FN rate between the true gene trees and the estimated gene trees, depends on the sequence length and fluctuates significantly (from 0–100%) both within and across replicates (Fig. S3); mean error is 36% and 15% for 100bp and 500bp, respectively.

Controlling λ_+, λ_-: Here, we consider $5 \times 4 = 20$ conditions, changing duplication and loss rates. Our λ_+ settings result in 0 to 5 extra copies per gene, and the $\frac{\lambda_-}{\lambda_+}$ varies between 0 and 1 (Table 1; Fig. S4). All other parameters are identical to the default condition.

Controlling λ_+, N_e: Here, we consider $3 \times 5 = 15$ conditions, fixing λ_- to be equal to λ_+, but changing λ_+ and ILS levels (controlled by N_e). Our λ_+ settings result in 0 to 5 extra copies per gene, and the mean ILS level between true and estimated gene trees varies between 0 and 70% RF. (Table 1; Fig. S5) All other parameters are identical to the default model.

Controlling n: Fixing all parameters, we vary the number of ingroup taxa n from 10 to 500.

Controlling k: Fixing all parameters, we vary the number of gene trees k from 25 to 10,000.

4.1.2 Existing simulations (S100)

We also used an existing dataset that Molloy and Warnow simulated [54] based on a real fungal dataset [55]. The simulation protocol of this dataset is similar to that of S25 dataset, with some notable differences. (*i*) The dataset included 100 species (no outgroup); species tree height, speciation rate, and mutation rates all differed from S25. (*ii*) Shorter gene alignments were also used, resulting in higher MGTE (25bp: 67%, 50bp: 52%, 100bp: 35%, 500bp: 19%). (*iii*) The duplication rate λ_+ was set to 1×10^{-10}, 2×10^{-10}, or 5×10^{-10} (named 1, 2, and 5, respectively), and the duplication rate equaled the loss rate for all model conditions. (*iv*) ILS was much lower than S25; two conditions were simulated with N_e set to 1×10^7 and 5×10^7 (named 1 and 5, respectively), which result in 2% and 12% RF between true gene trees and the species tree. (*v*) Gene trees were estimated using RAxML instead of FastTree2.

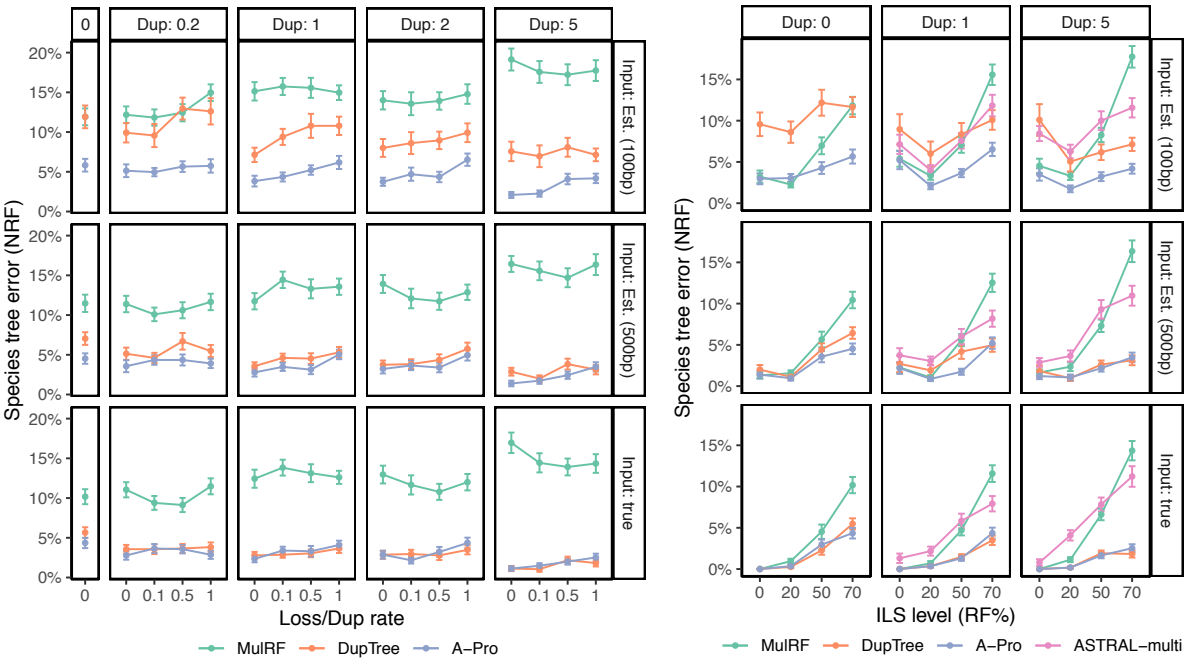
4.1.3 Biological data (1kp)

A transcriptome analysis of 103 plant species has been performed on 424 single-copy gene trees (out of thousands of genes) using both concatenation and ASTRAL [37]. In preliminary analyses, the authors had inferred multi-copy gene trees using RAxML from 9683 genes for 83 of those species, ranging in size between 5 and 2395 leaves. However, not being able to obtain an accurate species tree from the multi-copy gene trees, they abandoned the strategy in later analyses. The gene trees are available on Cyverse [56]. We used gene trees inferred from AA or first two codon positions (C12) as the original study.

4.2 Methods compared

We implemented A-Pro by extending ASTRAL-MP [36] and implementing Algorithms 1 and 2 as part of its native C++ library. We compare A-Pro to the following methods. Another method, STAG [57], is not included because of its poor performance on the S100 dataset [54], including that it fails to run on some model conditions (Fig. S8).

DupTree [10] infers a species tree from rooted or unrooted gene trees minimizing the duplication reconciliation cost [1] under the duplication-only model, but it does not model ILS. We



(a) (b)

Figure 2: Species tree error on the S25 dataset for $n = 25$ ingroup species, $k = 1000$ gene trees, and both true and estimated gene trees from 100bp and 500bp alignments. (a) Controlling duplication rate (box columns; labelled by C) and the loss rate (x-axis; ratio of the loss rate to duplication rate). (b) Controlling the duplication rate (columns; labelled by C) and the ILS level (x-axis; NRF between true gene trees and the species tree for $\lambda_+ = 0$). A-Pro and ASTRAL-multi are identical with $\lambda_+ = 0$. See Table 1 for parameters and Fig. S7 for iGPT-duploss.

provide DupTree with unrooted gene trees. We also tried iGTP, minimizing Dup-Loss score, but we only show results in supplement (Fig. S7) as it was almost universally worse than DupTree.

MulRF [17], based on an extension of the RF distance [58] to multi-labelled trees, is a hill-climbing method that aims at finding the tree with the minimum RF distance to the input. We use MulRF because of its advantage over other methods shown in previous studies [20].

ASTRAL-multi [39] is a feature of ASTRAL designed for handling multiple individuals. A recent paper (concurrently submitted) proposes to use ASTRAL-multi for multi-copy data [40]. Due to its high memory requirements, we were able to include it in only one experiment of S25.

5 Results

5.1 S25 dataset

5.1.1 Controlling duplication and loss rates and the level of ILS

We start by experiments that change the duplication and loss rates (λ_+, λ_-) from the default condition (Fig. 2a). On true gene trees, A-Pro and DupTree are essentially tied in terms of accuracy, except for the case with no duplication and loss where A-Pro is perhaps slightly more accurate. Overall, the accuracy of A-Pro and DupTree is statistically indistinguishable under these conditions (p-value $= 0.79$ according to a multi-variate ANOVA test). Increasing λ_+ *reduces* error ($p < 10^{-5}$), perhaps because additional copies provide more information, akin to increasing the number of loci. Despite statistically significant increases ($p = 0.006$) in error as λ_- increases, both methods are quite robust to loss rates, losing at most 1.5% accuracy on average when $\lambda_- = \lambda_+$ compared to no losses. MulRF has much higher error than other two methods, with errors that range between 10% and 17% across model conditions (we remind the reader that all these conditions have high ILS, a process that MulRF does not model).

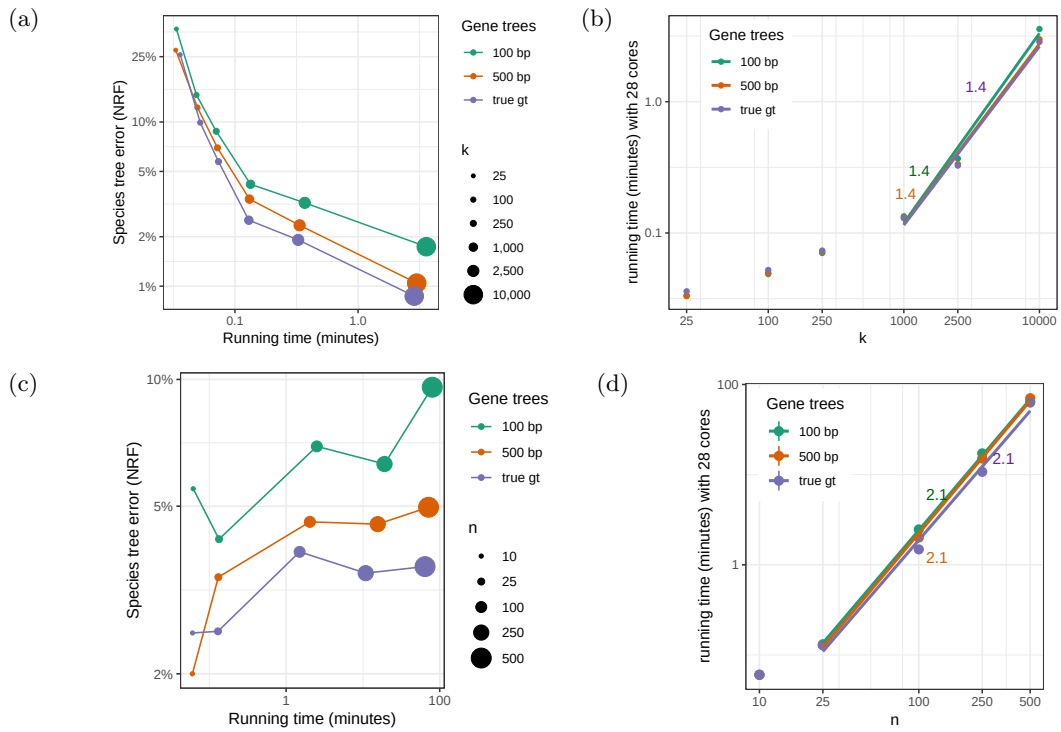


Figure 3: (a,c) Accuracy (y-axis) and running time (x-axis) of A-Pro as the number of genes k (a) or the number of species n (c) changes. Both axis are in log-scale. As k increases, accuracy increases. (b,d) The running time of A-Pro versus k (b) and n (d). We fit a line to the log-log plot of the running time only for $k \geq 1000$ and $n \geq 25$ as smaller runs are too fast to be reliable. We empirically estimate A-Pro to roughly proportionally with $k^{\frac{3}{2}}$ and n^2.

On estimated gene trees, the pattern changes, and the error of DupTree increases dramatically while A-Pro remains relatively accurate. When $\lambda_+ = \lambda_- = 0$, DupTree has on average an 11.5% error whereas A-Pro has only a 4.5% error for 500bp. Adding duplications helps both methods but A-Pro remains more accurate. For example, with 100bp input gene trees, DupTree has an error between 50% to 260% higher than A-Pro. With low-error gene trees, differences are statistically significant ($p < 10^{-5}$) but are more modest in magnitude (the error increase from DupTree to A-Pro across conditions by a median of 28%). The relative accuracy of A-Pro and DupTree is not a function of λ_- (p =0.8) but may depend on λ_+ (p =0.05).

In terms of running time, on the default model condition, we observe that A-Pro is the fastest method, taking less than a minute on this dataset, followed closely by DupTree (Fig. S9). We will revisit running time of A-Pro on larger datasets below.

As we change the ILS level (Table 1), the reason for the poor performance of MulRF becomes clear (Fig. 2b). Without ILS, MulRF has excellent accuracy, often matching A-Pro and beating DupTree on low-error gene trees. As the ILS level increases (especially above 20%), the accuracy of MulRF deteriorates quickly. Overall, ILS has the strongest effect on accuracy ($p \ll 10^{-5}$) but its impact on methods vary ($p \ll 10^{-5}$). DupTree seems as tolerant of ILS as A-Pro, despite the fact that DupTree is not designed specifically for ILS, and both methods are much more tolerant of ILS than MulRF. Nevertheless, once again, DupTree shows extreme sensitivity to gene tree error. To summarize, DupTree is relatively tolerant of ILS but less tolerant of gene tree error; MulRF is tolerant of gene tree error but not of ILS; A-Pro is quite robust to both.

5.1.2 Controlling the number of genes and species

Increasing the number of genes k in the most difficult case of high λ_+, λ_-, and ILS results in continued improvement in accuracy for A-Pro for every value we tested up to $k = 10^4$ (Fig. 3a).

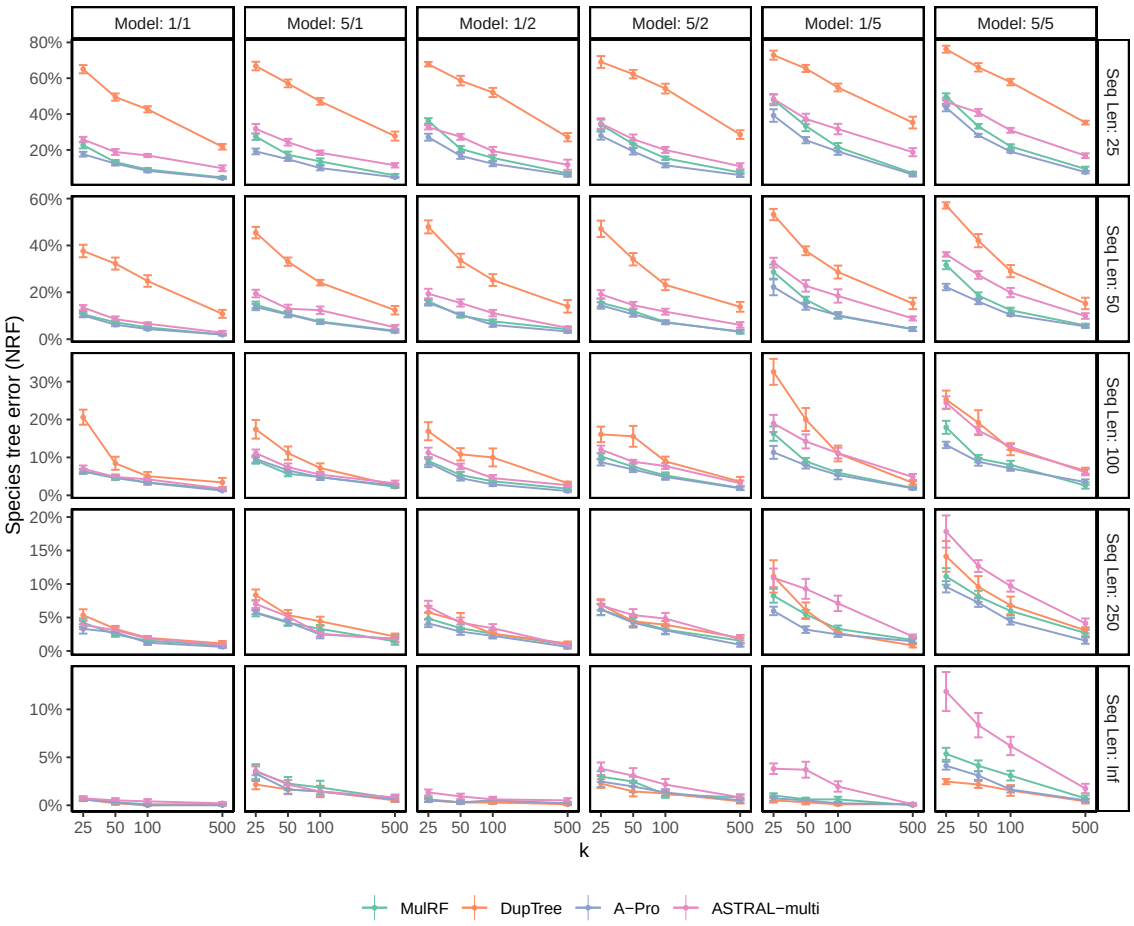


Figure 4: Species tree error on S100 dataset. We compare the species tree error of the four methods, showing mean and standard error over 10 replicates for each model condition, with varying numbers of genes (k) and sequence lengths (with Inf signifying true gene trees). Model conditions are labeled as a/b where a is the level of ILS (1 or 5) and b is the duplication/loss rate (1, 2, or 5).

With true gene trees, the error reduces from 26% with $k = 25$ to below 1% with $k = 10^4$. Even with less accurate gene trees, the error reduces to below 2% with increased numbers of genes. Increasing k increases running time, which empirically grows with $k^{1.4}$ (Fig. 3b). Nevertheless, using 28 cores, the running time was never more than 3.5 minutes even with $k = 10^4$.

Increasing n from 25 to 500 shows that A-Pro is relatively robust to a large number of species (Fig. 3c). With true gene trees, the error ranges between 2.5% with 10 species to 3.5% with 500 species. With estimated gene trees, error ranges between 4.1% to 9.5% (for 100bp) and between 2% and 5% (for 500bp). Note that as n increases, the gene tree error also increases (Table 1; Fig S6). The running time of A-Pro increases roughly quadratically with n (Fig. 3d) but is below 2 hours (given 28 cores) even for $n = 500$ (recall that $k = 1000$).

5.2 S100 dataset

Patterns of performance on the S100 dataset are consistent with the S25 dataset (Fig. 4). DupTree is highly accurate with true gene trees and gene trees with low estimation error but quickly degrades in accuracy as gene tree error increases. MulRF is less sensitive to gene tree error but is more sensitive to the ILS level (which is always moderate or low on this dataset). As in S25, here, we see that using ASTRAL-multi to handle duplication and loss is not beneficial. A-Pro works the best overall, ranking first in terms of mean error (rounded to two significant digits) in 105 out of 120 test conditions (Table S2). The second best method is MulRF on this dataset, which is not surprising given the low ILS levels in this dataset.

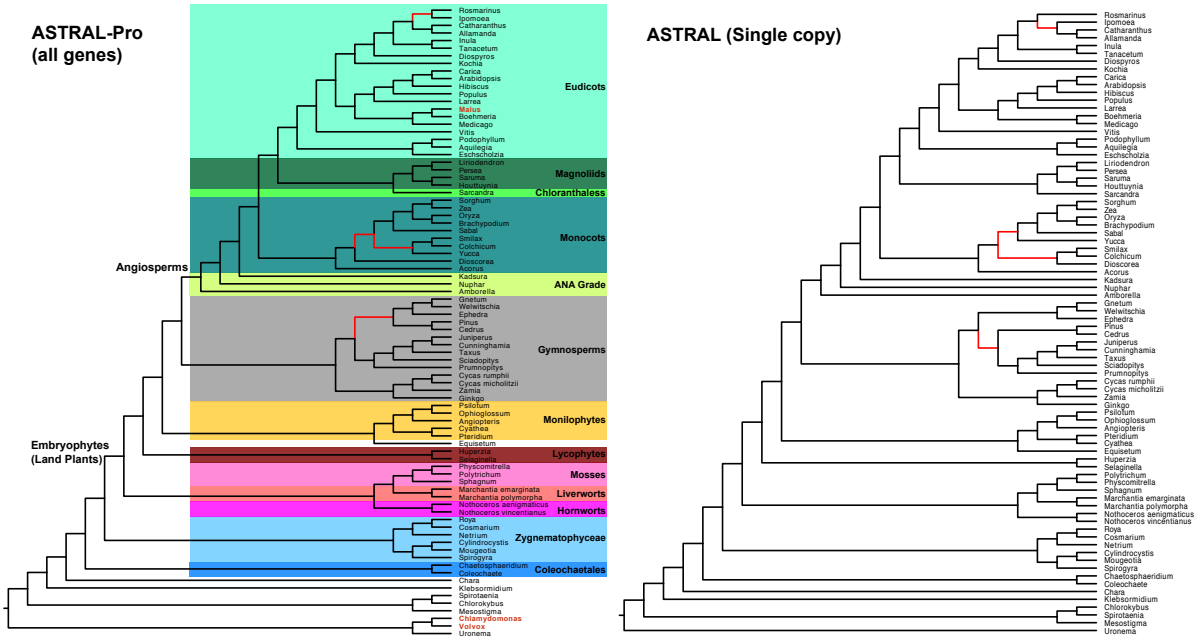


Figure 5: **ASTRAL and ASTRAL-Pro on biological plant paper**. ASTRAL-Pro is run on 9683 multi-copy AA gene trees available online [56]. ASTRAL is run on 424 single-copy gene trees and was reported previously [37]. Three genomes, shown in red, were present in multi-copy gene trees but not in the single-copy analyses. The single-copy tree includes 23 extra species that were not in the multi-copy data and are removed here. The two trees are very similar and differ in only four branches shown in red.

5.3 Biological plant dataset

On the plant dataset [37], A-Pro on AA gene trees returned a species tree (Fig. 5) that was largely similar to the main single-copy ASTRAL tree reported by the original study (only 4 differences). Using C12 gene trees resulted in only one change, swapping Chara and Coleochaetales. In contrast, DupTree differed from the ASTRAL tree in 33 out of 77 branches (21/77 for iGTP-DupLoss) and violated many known biological relationships (Fig. S10). The A-Pro trees are consistent with ASTRAL for major groups, including placing Zygnematales (not Chara) as sister to all land plants, the placement of Amboerlla as sister to the rest of angiosperms, and monophyly of Bryophytes (liverworts, mosses, and hornworts).

The four changes between the ASTRAL and A-Pro trees are interesting. In A-Pro, unlike ASTRAL, Rosmarinus and Ipomoea are grouped together, which is likely the correct result as these species are in the same order (Lamiales). The position of genus Yucca in the A-Pro tree has changed; interestingly, a recent update to this paper using $> 1,000$ species [38] (which samples close genera Asparagales and Liliales) finds Yucca in a position identical to A-Pro. Thus, the A-Pro placement is more likely to be correct. Most consequentially, A-Pro, unlike ASTRAL, recovers the GnePine hypothesis combining Gnetales and Pinaceae, a hypothesis suggested by several studies [59–62] and all concatenation analyses of 1kp [37]. Interestingly, the new 1kp paper [38] uses DiscoVista [63] to examine quartet frequencies around this branch and detects that the second and third most frequent quartets do not match (0.4 vs. 0.1) and are heavily skewed towards GnePine, making the resolution obtained in ASTRAL less reliable.

6 Discussions

We developed a "per-locus" quartet-based measure of similarity between multi-copy gene trees and a species tree. The measure relies on internal nodes of gene trees being tagged as speciation or duplication. Somewhat counter-intuitively, despite being a quartet measure, it needs *partially* rooted trees (Claim 1). The measure defines an equivalence relationship on quartets and counts

each equivalent class only once, avoiding double-counting quartets that are bound to have identical topologies. Avoiding double-counting is at the heart of the approach and likely is a main reason behind its high accuracy on simulated and empirical data we tested.

Astral-Pro, which maximizes the per-locus quartet score, is statistically consistent under MSC and GDL models. This makes one hope that it may also be consistent under both causes of discordance combined. The DLCoal model [55] accounts for ILS, duplication, and loss. Under this model, each duplication immediately creates a daughter locus, which is unlinked from the parent locus; the duplication event gets fixed in all species. Gene trees are seen as generated by first producing a locus tree via a birth-death process that runs on the species tree and then running a MSC process on the locus tree. Because the loci are considered as unlinked, the coalescence processes occur independently between the parent and daughter loci (but the daughter MSC process is "bounded" at the time of duplication). Due to the independence of loci, dividing a multi-copy gene family into its constituent loci can give us distributions on gene tree topologies that behave similarly (though not identically) to the MSC model. The per-locus metric *seeks* to count quartet topologies across loci as they existed at the time of speciation events relevant to a quartet (i.e., at the time of the anchor LCA). When successful, it counts only topologies that are drawn from independent coalescent processes. However, complicated scenarios involving a combination of duplications, losses and ILS can lead to incorrectly tagged gene trees. These scenarios create complications that need to be addressed. We leave it to the future work to study whether ASTRAL-Pro is statistically consistent under the DLCoal model.

To get rooted and tagged gene trees, we used the maximum parsimony principle, with duplication and loss each penalized equally and deep coalescence not penalized at all. There is a large literature on various ways of tagging and rooting gene trees [e.g., 64–66], including other penalties for the duplication and loss events (e.g., there is a suggestion of losses having half the penalty of duplications [67]). It may also be possible to improve tagging of gene trees using probabilistic orthology inference [68, 69] or using synteny information [70, 71]. However, these methods often require a species tree. It may possible to use A-pro in an iterative fashion, where the species tree is inferred, gene trees are re-tagged and re-rooted, and a new species tree is inferred Future work should explore these approaches.

Quartet-based methods for handling multi-copy gene trees are not abundant. Besides our method, one can attempt to sample single-copy gene trees [41], an approach that we plan to test in the future. Most recently, there has been theoretical and empirical evidence that simply treating gene copies as alleles may be sufficient [40]. We showed that this alternative, although attractive in theory, is less accurate and less scalable than A-Pro. We are unaware of other quartet-based species tree inference methods for multi-copy input.

A-Pro has some limitations. Most importantly, in its current form, it can only handle binary trees, which reduces its ability to handle gene tree error [47]. While A-Pro is more robust to gene tree error than alternatives, combining it with co-estimation [3] and tree fixing [72–77] may further improve its accuracy. Future work should also explore ways to extend A-Pro so that it can handle polytomies in input gene trees. Finally, with more algorithmic development, it should be possible to provide all the features ASTRAL provides, including branch length and Local-PP [78], polytomy test [79], and visualization of discordance [63]. All these features should be adopted to A-Pro in the future.

Data availability. The code is available at `https://github.com/chaoszhang/A-pro` and data are made available at `https://github.com/chaoszhang/duploss-pipeline.git`.

Acknowledgments. S.M and C.Z were supported by the National Science Foundation (NSF) grant III-1845967. E.K.M. was supported by the Ira and Debra Cohen Graduate Fellowship in Computer Science.

References

[1] Wayne P. Maddison. Gene Trees in Species Trees. *Systematic Biology*, 46(3):523–536, 9 1997.

[2] Joseph Heled and Alexei J Drummond. Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, 27(3):570–580, 3 2010.

[3] Bastien Boussau, GJ J Szöllősi, and Laurent Duret. Genome-scale coestimation of species and gene trees. *Genome Research*, 23(2):323–330, 12 2013.

[4] G J Szöllõsi, E Tannier, Vincent Daubin, and Bastien Boussau. The inference of gene trees with species trees. *Systematic Biology*, 64(1):e42–e62, 7 2014.

[5] Liang Liu. BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics*, 24(21):2542–2543, 11 2008.

[6] Junjian An, Lihua Zhu, Yingying Zhang, and Heqing Tang. Efficient visible light photo-fenton-like degradation of organic pollutants using in situ surface-modified BiFeO3 as a catalyst. *Journal of environmental sciences (China)*, 25(6):1213–25, 6 2013.

[7] Julia Chifman and Laura S Kubatko. Quartet Inference from SNP Data Under the Coalescent Model. *Bioinformatics*, 30(23):3317–3324, 8 2014.

[8] David Bryant, Remco Bouckaert, Joseph Felsenstein, Noah A. Rosenberg, and Arindam Roychoudhury. Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution*, 29(8):1917–1932, 2012.

[9] Nicola De Maio, Christian Schlötterer, and Carolin Kosiol. Linking Great Apes Genome Evolution across Time Scales Using Polymorphism-Aware Phylogenetic Models. *Molecular Biology and Evolution*, 30(10):2249–2262, 10 2013.

[10] André Wehe, Mukul S Bansal, J Gordon Burleigh, and Oliver Eulenstein. DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics*, 24(13):1540–1541, 2008.

[11] Ruchi Chaudhary, Mukul S Bansal, André Wehe, David Fernández-Baca, and Oliver Eulenstein. iGTP: a software package for large-scale gene tree parsimony analysis. *BMC bioinformatics*, 11(1):574, 1 2010.

[12] Mukul S. Bansal, J. Gordon Burleigh, and Oliver Eulenstein. Efficient genome-scale phylogenetic analysis under the duplication-loss and deep coalescence cost models. *BMC Bioinformatics*, 11(Suppl 1):S42, 2010.

[13] Bin Ma, Ming Li, and Louxin Zhang. From gene trees to species trees. *SIAM Journal on Computing*, 30(3):729–752, 2000.

[14] M. T. Hallett and Jens Lagergren. New algorithms for the duplication-loss model. In *Proceedings of the fourth annual international conference on Computational molecular biology - RECOMB '00*, pages 138–146, New York, New York, USA, 2000. ACM Press.

[15] Md. Shamsuzzoha M.S. Bayzid, Siavash Mirarab, and Tandy Warnow. Inferring optimal species trees under gene duplication and loss. *Pacific Symposium on Biocomputing*, 18:250–261, 2013.

[16] Morris Goodman, John Czelusniak, G. William Moore, A. E. Romero-Herrera, and Genji Matsuda. Fitting the Gene Lineage into its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences. *Systematic Biology*, 28(2):132–163, 1979.

[17] Ruchi Chaudhary, J Gordon Burleigh, and David Fernández-Baca. Inferring species trees from incongruent multi-copy gene trees using the Robinson-Foulds distance. *Algorithms for Molecular Biology*, 8:28, 2013.

[18] Leonardo De Oliveira Martins, Diego Mallo, and David Posada. A Bayesian Supertree Model for Genome-Wide Species Tree Reconstruction. *Systematic Biology*, 65(3):397–416, 5 2016.

[19] Erin Molloy and Tandy Warnow. Large-scale Species Tree Estimation. *ArXiv preprint: 1904.02600*, 4 2019.

[20] Ruchi Chaudhary, Bastien Boussau, J. Gordon Burleigh, and David Fernández-Baca. Assessing approaches for inferring species trees from multi-copy genes. *Systematic Biology*, 64(2):325–339, 2015.

[21] P Pamilo and M Nei. Relationships between gene trees and species trees. *Molecular biology and evolution*, 5(5):568–583, 1988.

[22] Bruce Rannala and Ziheng Yang. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164(4):1645–1656, 2003.

[23] Liang Liu, Lili Yu, Dennis K. Pearl, and Scott V. Edwards. Estimating species phylogenies using coalescence times among sequences. *Systematic Biology*, 58(5):468–477, 10 2009.

[24] Bret R. Larget, Satish K. Kotha, Colin N. Dewey, and Cécile Ané. BUCKy: Gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics*, 26(22):2910–2911, 11 2010.

[25] Elchanan Mossel and Sebastien Roch. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 7(1):166–171, 1 2010.

[26] Liang Liu, Lili Yu, and Scott V Edwards. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, 10(1):302, 2010.

[27] Yufeng Wu. Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution*, 66(3):763–775, 2012.

[28] Erfan Sayyari and Siavash Mirarab. Anchoring quartet-based phylogenetic distances and applications to species tree reconstruction. *BMC Genomics*, 17(S10):101–113, 11 2016.

[29] Liang Liu and Lili Yu. Estimating Species Trees from Unrooted Gene Trees. *Systematic Biology*, 60(5):661–667, 10 2011.

[30] Pranjal Vachaspati and Tandy Warnow. ASTRID: Accurate Species TRees from Internode Distances. *BMC Genomics*, 16(Suppl 10):S3, 2015.

[31] Siavash Mirarab, Rezwana Reaz, Md. Shamsuzzoha Bayzid, Théo Zimmermann, M. S. Swenson, and Tandy Warnow. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548, 9 2014.

[32] Thomas C. Giarla and Jacob A. Esselstyn. The Challenges of Resolving a Rapid, Recent Radiation: Empirical and Simulated Phylogenomics of Philippine Shrews. *Systematic Biology*, 64(5):727–740, 9 2015.

[33] Erin K. Molloy and Tandy Warnow. To Include or Not to Include: The Impact of Gene Filtering on Species Tree Estimation Methods. *Systematic Biology*, 67(2):285–303, 3 2018.

[34] Jesús A Ballesteros and Prashant P Sharma. A Critical Appraisal of the Placement of Xiphosura (Chelicerata) with Account of Known Sources of Phylogenetic Error. *Systematic Biology*, pages 1–62, 2 2019.

[35] Siavash Mirarab and Tandy Warnow. ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, 31(12):i44–i52, 6 2015.

[36] John Yin, Chao Zhang, and Siavash Mirarab. ASTRAL-MP: scaling ASTRAL to very large datasets using randomization and parallelization. *Bioinformatics*, 35(20):3961–3969, 10 2019.

[37] Norman J. Wickett, Siavash Mirarab, Nam Nguyen, Tandy Warnow, Eric J Carpenter, Naim Matasci, Saravanaraj Ayyampalayam, Michael S. Barker, J. Gordon Burleigh, Matthew A. Gitzendanner, Brad R. Ruhfel, Eric Wafula, Joshua P. Der, Sean W. Graham, Sarah Mathews, Michael Melkonian, Douglas E. Soltis, Pamela S. Soltis, Nicholas W. Miles, Carl J. Rothfels, Lisa Pokorny, A. Jonathan Shaw, Lisa DeGironimo, Dennis W. Stevenson, Barbara Surek, Juan Carlos Villarreal, Béatrice Roure, Hervé Philippe, Claude W. DePam-

philis, Tao Chen, Michael K. Deyholos, Regina S. Baucom, Toni M. Kutchan, Megan M. Augustin, Jian Jun Wang, Yong Zhang, Zhijian Tian, Zhixiang Yan, Xiaolei Wu, Xiao Sun, Gane Ka-Shu Wong, and James Jim Leebens-Mack. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences*, 111(45):4859–4868, 10 2014.

[38] James H Leebens-Mack, Michael S Barker, Eric J Carpenter, Michael K Deyholos, Matthew A Gitzendanner, Sean W Graham, Ivo Grosse, Zheng Li, Michael Melkonian, Siavash Mirarab, Martin Porsch, Marcel Quint, Stefan A Rensing, Douglas E Soltis, Pamela S Soltis, Dennis W Stevenson, Kristian K Ullrich, Norman J Wickett, Lisa DeGironimo, Patrick P Edger, Ingrid E Jordon-Thaden, Steve Joya, Tao Liu, Barbara Melkonian, Nicholas W Miles, Lisa Pokorny, Charlotte Quigley, Philip Thomas, Juan Carlos Villarreal, Megan M Augustin, Matthew D Barrett, Regina S Baucom, David J Beerling, Ruben Maximilian Benstein, Ed Biffin, Samuel F Brockington, Dylan O Burge, Jason N Burris, Kellie P Burris, Valérie Burtet-Sarramegna, Ana L Caicedo, Steven B Cannon, Zehra Çebi, Ying Chang, Caspar Chater, John M Cheeseman, Tao Chen, Neil D Clarke, Harmony Clayton, Sarah Covshoff, Barbara J Crandall-Stotler, Hugh Cross, Claude W DePamphilis, Joshua P Der, Ron Determann, Rowan C Dickson, Verónica S Di Stilio, Shona Ellis, Eva Fast, Nicole Feja, Katie J Field, Dmitry A Filatov, Patrick M Finnegan, Sandra K Floyd, Bruno Fogliani, Nicolás García, Gildas Gâteblé, Grant T Godden, Falicia (Qi Yun) Goh, Stephan Greiner, Alex Harkess, James Mike Heaney, Katherine E Helliwell, Karolina Heyduk, Julian M Hibberd, Richard G J Hodel, Peter M Hollingsworth, Marc T J Johnson, Ricarda Jost, Blake Joyce, Maxim V Kapralov, Elena Kazamia, Elizabeth A Kellogg, Marcus A Koch, Matt Von Konrat, Kálmán Könyves, Toni M Kutchan, Vivienne Lam, Anders Larsson, Andrew R Leitch, Roswitha Lentz, Fay-Wei Li, Andrew J Lowe, Martha Ludwig, Paul S Manos, Evgeny Mavrodiev, Melissa K McCormick, Michael McKain, Tracy McLellan, Joel R McNeal, Richard E Miller, Matthew N Nelson, Yanhui Peng, Paula Ralph, Daniel Real, Chance W Riggins, Markus Ruhsam, Rowan F Sage, Ann K Sakai, Moira Scascitella, Edward E Schilling, Eva-Marie Schlösser, Heike Sederoff, Stein Servick, Emily B Sessa, A Jonathan Shaw, Shane W Shaw, Erin M Sigel, Cynthia Skema, Alison G Smith, Ann Smithson, C Neal Stewart, John R Stinchcombe, Peter Szövényi, Jennifer A Tate, Helga Tiebel, Dorset Trapnell, Matthieu Villegente, Chun-Neng Wang, Stephen G Weller, Michael Wenzel, Stina Weststrand, James H Westwood, Dennis F Whigham, Shuangxiu Wu, Adrien S Wulff, Yu Yang, Dan Zhu, Cuili Zhuang, Jennifer Zuidof, Mark W Chase, J Chris Pires, Carl J Rothfels, Jun Yu, Cui Chen, Li Chen, Shifeng Cheng, Juanjuan Li, Ran Li, Xia Li, Haorong Lu, Yanxiang Ou, Xiao Sun, Xuemei Tan, Jingbo Tang, Zhijian Tian, Feng Wang, Jun Wang, Xiaofeng Wei, Xun Xu, Zhixiang Yan, Fan Yang, Xiaoni Zhong, Feiyu Zhou, Ying Zhu, Yong Zhang, Saravanaraj Ayyampalayam, Todd J Barkman, Nam-phuong Nguyen, Naim Matasci, David R Nelson, Erfan Sayyari, Eric K Wafula, Ramona L Walls, Tandy Warnow, Hong An, Nils Arrigo, Anthony E Baniaga, Sally Galuska, Stacy A Jorgensen, Thomas I Kidder, Hanghui Kong, Patricia Lu-Irving, Hannah E Marx, Xinshuai Qi, Chris R Reardon, Brittany L Sutherland, George P Tiley, Shana R Welles, Rongpei Yu, Shing Zhan, Lydia Gramzow, Günter Theißen, Gane Ka-Shu Wong, and One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*, 574(7780):679–685, 10 2019.

[39] Maryam Rabiee, Erfan Sayyari, and Siavash Mirarab. Multi-allele species reconstruction using ASTRAL. *Molecular Phylogenetics and Evolution*, 130:286–296, 1 2019.

[40] Brandon Legried, Erin K Molloy, Tandy Warnow, and Sebastin Roch. Polynomial-Time Statistical Estimation of Species Trees under Gene Duplication and Loss. *Submitted to RECOMB 202, available on bioRxiv*, page 821439, 2019.

[41] Peng Du, Matthew W Hahn, and Luay Nakhleh. Species Tree Inference under the Multispecies Coalescent on Data with Paralogs is Accurate. *bioRxiv*, page 498378, 2019.

[42] Marina Marcet-Houben and Toni Gabaldón. TreeKO: a duplication-aware algorithm for the comparison of phylogenetic trees. *Nucleic Acids Research*, 39(10):e66–e66, 5 2011.

[43] C. Scornavacca, V. Berry, and V. Ranwez. Building species trees from larger parts of phylogenomic databases. *Information and Computation*, 209(3):590–605, 3 2011.

[44] Ya Yang and Stephen A. Smith. Orthology Inference in Nonmodel Organisms Using Transcriptomes and Low-Coverage Genomes: Improving Accuracy and Matrix Occupancy for Phylogenomics. *Molecular Biology and Evolution*, 31(11):3081–3092, 11 2014.

[45] Casey W Dunn, Mark Howison, and Felipe Zapata. Agalma: an automated phylogenomics workflow. *BMC bioinformatics*, 14(1):330, 2013.

[46] Jesús A. Ballesteros and Gustavo Hormiga. A New Orthology Assessment Method for Phylogenomic Data: Unrooted Phylogenetic Orthology. *Molecular Biology and Evolution*, 33(8):2117–2134, 8 2016.

[47] Chao Zhang, Maryam Rabiee, Erfan Sayyari, and Siavash Mirarab. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, 19(S6):153, 5 2018.

[48] Manuel Lafond and Celine Scornavacca. On the Weighted Quartet Consensus problem. *Theoretical Computer Science*, 769:1–17, 5 2019.

[49] Lars Arvestad, Jens Lagergren, and Bengt Sennblad. The gene evolution model and computing its associated probabilities. *Journal of the ACM*, 56(2):1–44, 4 2009.

[50] Diego Mallo, Leonardo De Oliveira Martins, and David Posada. SimPhy: Phylogenomic Simulation of Gene, Locus, and Species Trees. *Systematic biology*, 65(2):334–44, 3 2016.

[51] William Fletcher and Ziheng Yang. INDELible: A flexible simulator of biological sequence evolution. *Molecular Biology and Evolution*, 26(8):1879–1888, 2009.

[52] Simon Tavaré. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *Lectures on Mathematics in the Life Sciences*, 17:57–86, 1986.

[53] Morgan N. Price, Paramvir S. Dehal, and Adam P. Arkin. FastTree-2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, 5(3):e9490, 3 2010.

[54] Erin K Molloy and Tandy Warnow. FastMulRFS : Statistically consistent polynomial time species tree estimation under gene duplication. *bioRxiv*, page 835553, 2019.

[55] MD Rasmussen and Manolis Kellis. Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome research*, 22(4):755–765, 2012.

[56] Naim Matasci, Ling-Hong L.-H. Hung, Zhixiang Yan, E.J. Eric J Carpenter, N.J. Norman J Wickett, Siavash Mirarab, Nam Nguyen, Tandy Warnow, Saravanaraj Ayyampalayam, Michael S Barker, J.G. Burleigh, M.A. Gitzendanner, E. Wafula, J.P. Der, C.W. dePamphilis, B. Roure, H. Philippe, B.R. Ruhfel, N.W. Miles, S.W. Graham, S. Mathews, B. Surek, M. Melkonian, D.E. Soltis, P.S. Soltis, C. Rothfels, L. Pokorny, J.A. Shaw, L. DeGironimo, D.W. Stevenson, J.C. Villarreal, T. Chen, T.M. Kutchan, M. Rolf, R.S. Baucom, M.K. Deyholos, R. Samudrala, Z. Tian, X. Wu, X. Sun, Y. Zhang, J. Wang, J. Leebens-Mack, and G.K.S. Wong. Data access for the 1,000 Plants (1KP) project. *GigaScience*, 3(1):17, 2014.

[57] D M Emms, S Kelly, and South Parks Road. STAG: Species Tree Inference from All Genes. *bioRxiv*, page 267914, 1 2018.

[58] DF Robinson and LR Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2):131–147, 1981.

[59] J. Gordon Burleigh and Sarah Mathews. Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life. *American Journal of Botany*, 91(10):1599–1613, 10 2004.

[60] B. Zhong, Takahiro Yonezawa, Yang Zhong, and Masami Hasegawa. The Position of Gnetales among Seed Plants: Overcoming Pitfalls of Chloroplast Phylogenomics. *Molecular Biology and Evolution*, 27(12):2855–2863, 12 2010.

[61] Bojian Zhong, Oliver Deusch, Vadim V. Goremykin, David Penny, Patrick J. Biggs, Robin A. Atherton, Svetlana V. Nikiforova, and Peter James Lockhart. Systematic Error in Seed Plant Phylogenomics. *Genome Biology and Evolution*, 3:1340–1348, 1 2011.

[62] Simon Laurin-Lemay, Henner Brinkmann, and Hervé Philippe. Origin of land plants revisited in the light of sequence contamination and missing data. *Current Biology*, 2012.

[63] Erfan Sayyari, James B. Whitfield, and Siavash Mirarab. DiscoVista: Interpretable visualizations of gene tree discordance. *Molecular Phylogenetics and Evolution*, 122:110–115, 5 2018.

[64] Mukul S. Bansal, Eric J. Alm, and Manolis Kellis. Reconciliation revisited: Handling multiple optima when reconciling with duplication, transfer, and loss. *Journal of Computational Biology*, 2013.

[65] Dannie Durand, Bjarni V Halldórsson, and Benjamin Vernot. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *Journal of Computational Biology*, 13(2):320–335, 2006.

[66] Edwin Jacox, Cedric Chauve, Gergely J Szöllősi, Yann Ponty, and Celine Scornavacca. eccetera: comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics*, 32(13):2056–2058, 2016.

[67] Lawrence A David and Eric J Alm. Rapid evolutionary innovation during an archaean genetic expansion. *Nature*, 469(7328):93, 2011.

[68] Lars Arvestad, Ann-Charlotte Berglund, Jens Lagergren, and Bengt Sennblad. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. In *Proceedings of the eighth annual international conference on Computational molecular biology - RECOMB '04*, pages 326–335, New York, New York, USA, 2004. ACM Press.

[69] Bengt Sennblad and Jens Lagergren. Probabilistic Orthology Analysis. *Systematic Biology*, 58(4):411–424, 8 2009.

[70] Guillaume Bourque, Yasmine Yacef, and Nadia El-Mabrouk. Maximizing Synteny Blocks to Identify Ancestral Homologs. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 21–34. 2005.

[71] Cedric Chauve, Nadia El-Mabrouk, Laurent Guéguen, Magali Semeria, and Eric Tannier. Duplication, Rearrangement and Reconciliation: A Follow-Up 13 Years Later. In Cedric Chauve, Nadia El-Mabrouk, and Eric Tannier, editors, *Models and Algorithms for Genome Evolution*, volume 19 of *Computational Biology*, pages 47–62. Springer London, London, 2013.

[72] Yi-Chiew Wu, Matthew D. Rasmussen, Mukul S Bansal, and Manolis Kellis. TreeFix: Statistically Informed Gene Tree Error Correction Using Species Trees. *Systematic Biology*, 62(1):110–120, 2013.

[73] Manuel Lafond, Magali Semeria, Krister M. Swenson, Eric Tannier, and Nadia El-Mabrouk. Gene tree correction guided by orthology. *BMC Bioinformatics*, 14(S15):S5, 10 2013.

[74] Manuel Lafond, Cedric Chauve, Dondi, and Nadia El-Mabrouk. Polytomy refinement for the correction of dubious duplications in gene trees. *Bioinformatics*, 30(17):i519–i526, 2014.

[75] Celine Scornavacca, Edwin Jacox, and Gergely J. Szöllosi. Joint amalgamation of most parsimonious reconciled gene trees. *Bioinformatics*, 31(6):841–848, 2015.

[76] Emmanuel Noutahi, Magali Semeria, Manuel Lafond, Jonathan Seguin, Bastien Boussau, Laurent Guéguen, Nadia El-Mabrouk, and Eric Tannier. Efficient gene tree correction guided by genome evolution. *PLoS ONE*, 11(8), 2016.

[77] Nadia El-Mabrouk and Emmanuel Noutahi. Gene Family Evolution—An Algorithmic Framework. In *Bioinformatics and Phylogenetics*, pages 87–119. Springer, 2019.

[78] Erfan Sayyari and Siavash Mirarab. Fast Coalescent-Based Computation of Local Branch

Support from Quartet Frequencies. *Molecular Biology and Evolution*, 33(7):1654–1668, 7 2016.

[79] Erfan Sayyari and Siavash Mirarab. Testing for Polytomies in Phylogenetic Species Trees Using Quartet Frequencies. *Genes*, 9(3):132, 2 2018.

[80] Daniel Kane and Terence Tao. A bound on partitioning clusters. *Electr. J. Comb.*, 24:P2.31, 2017.

A Proofs

Proof of Proposition 1. Denote $Q_1 = \{a, b, c, d\}$ and $Q_2 = \{\tilde{a}, \tilde{b}, \tilde{c}, \tilde{d}\}$ (with obvious correspondence of labels). Let w be the anchor LCA and note that anchor LCA is the LCA of three (if $Q_1 \angle G$) or four (if $Q_1 \perp G$) of the quartet leaves; thus, by Definition 2, w is a speciation node or otherwise Q_1 would not be a SQ. Let denote by w_1 and w_2 the children of w; by Definition 1, $\alpha_G(w_1)$, $\alpha_G(w_2)$ and the remaining leaves ($U = \alpha_G(\mathcal{L}_G - \mathcal{L}_G(w_1) - \mathcal{L}_G(w_2))$) must be mutually exclusive. In the unbalanced case, given that $a, b \in \mathcal{L}_G(w_1)$, $c \in \mathcal{L}_G(w_2)$, $d \in U$, mutual exclusivity is possible only if $\tilde{a}, \tilde{b} \in \mathcal{L}_G(w_1)$, $\tilde{c} \in \mathcal{L}_G(w_2)$, $\tilde{d} \in U$. In the case of balanced topology, mutual exclusivity of $\alpha_G(w_1)$ and $\alpha_G(w_2)$ and the fact that $a, b \in \mathcal{L}_G(w_1)$ and $c, d \in \mathcal{L}_G(w_2)$ implies that $\tilde{a}, \tilde{b} \in \mathcal{L}_G(w_1)$, $\tilde{c}, \tilde{d} \in \mathcal{L}_G(w_2)$. Thus, in either case, $\Omega(G \restriction Q_1) \simeq \Omega(G \restriction Q_2)$. □

Proof of Proposition 2. Each node of a gene tree represents an ancestral or present-day gene and thus belongs to a locus. The children of a speciation node stay in the same locus that their parent, while for a duplication node we have that exactly one of the two children change locus and the other stays in the same locus than its parent. Therefore, all nodes under w, which is a speciation node, belong to the descendants (including itself) of the locus to which w belongs, and when tracing back to the time of speciation event w, they will lead to the same locus. Since all equivalent classes share the same anchor LCA, the result follows. □

Proof of Lemma 1. Note that P can anchor Q_1 only iff any species tree that includes P must match the gene tree topology for Q_1 By Proposition 1, due to equivalence of Q_1 and Q_2, we infer Q_2 must (i) match the same species quartet set as Q_1 and (ii) share the same anchor LCA w. Thus, P can also anchor Q_2. (iii) When $Q_1 \angle G$ as shown in Figure 1, $\tilde{a}, \tilde{b} \in \mathcal{L}_G(w_1)$ are the leaves mapped to the quartet tree and thus mapped to the same partition as a, b; similarly, when $Q_1 \perp G$, the pair of leaves under left subtree of the anchor LCA of both quartets map to the same partition of P. □

Proof of Lemma 2. First note that:

$$|\{\alpha_G(Q) : Q \subset \mathcal{L}_G, Q \xrightarrow{w} P\}| =$$
$$|\{\alpha_G(Q) : Q \subset \mathcal{L}_G, Q \xrightarrow{w} P, Q \perp G\}| + |\{\alpha_G(Q) : Q \subset \mathcal{L}_G, Q \xrightarrow{w} P, Q \angle G\}|$$

We compute each part separately. Recall here that, since $Q \xrightarrow{w} P$, Q is a SQ quartet and thus $|Q| = |\alpha_G(Q)| = 4$.

When $Q \perp G$, let $Q = \{a, b, c, d\}$ with $\alpha_G(a), \alpha_G(b) \in M_1, \alpha_G(c), \alpha_G(d) \in M_2$. Since $Q \xrightarrow{w} P$, leaves $\alpha_G(a)$ and $\alpha_G(b)$ must be in the same partition of P. When $\alpha_G(a), \alpha_G(b) \in P_1$, leaves $\alpha_G(c)$ and $\alpha_G(d)$ must be in partition P_2 and P_3 respectively since P can anchor Q. W.l.o.g., we can assume $\alpha_G(c) \in P_2$. Therefore, $\alpha_G(a), \alpha_G(b) \in M_1 \cap P_1$, $\alpha_G(c) \in M_2 \cap P_2$, $\alpha_G(d) \in M_2 \cap P_3$. The number of such $\alpha_G(Q)$ is $\binom{|M_1 \cap P_1|}{2}|M_2 \cap P_2||M_2 \cap P_3| = \binom{I_{11}}{2} I_{22} I_{23}$. Similarly when $\alpha_G(a), \alpha_G(b) \in M_1 \cap P_2$ and $\alpha_G(a), \alpha_G(b) \in M_1 \cap P_3$, the number of such $\alpha_G(Q)$ is $\binom{I_{12}}{2} I_{21} I_{23}$ and $\binom{I_{13}}{2} I_{21} I_{22}$ respectively. Thus,

$$|\{\alpha_G(Q) : Q \subset \mathcal{L}_G, Q \xrightarrow{w} P, Q \perp G\}| = \binom{I_{11}}{2} I_{22} I_{23} + \binom{I_{12}}{2} I_{21} I_{23} + \binom{I_{13}}{2} I_{21} I_{22}$$

Similarly, when $Q \angle G$, let $Q = \{a, b, c, d\}$ with $\alpha_G(a)$ and $\alpha_G(b)$ in the same partition of P. Notice that, in the unbalanced case, $\alpha_G(a)$ and $\alpha_G(b)$ can be both either in M_1 or either in M_2, and since c and d are not interchangeable as in the balanced case, we can have $\alpha_G(a), \alpha_G(b) \in P_i, \alpha_G(c) \in P_j, \alpha_G(d) \in P_k$ for (i, j, k) with any permutation of $(1, 2, 3)$, from the definition of P anchoring Q. All together we have 12 cases.

In the case that $\alpha_G(a), \alpha_G(b) \in P_1, \alpha_G(c) \in P_2, \alpha_G(d) \in P_3$, and $\alpha_G(a), \alpha_G(b) \in M_1$, we have $\alpha_G(a), \alpha_G(b) \in M_1 \cap P_1$, $\alpha_G(c) \in M_2 \cap P_2$, and $\alpha_G(d) \in M_3 \cap P_3$. The number of such

$\alpha_G(Q)$ is $\binom{|M_1 \cap P_1|}{2}|M_2 \cap P_2||M_3 \cap P_3| = \binom{I_{11}}{2} I_{22} I_{33}$. The other 11 permutations are similar. In total,

$$\begin{aligned} &|\{\alpha_G(Q) : Q \subset \mathcal{L}_G, Q \xrightarrow{w} P, Q \angle G\}| \\ =&\binom{I_{11}}{2}(I_{22}I_{33} + I_{32}I_{23}) + \binom{I_{12}}{2}(I_{21}I_{33} + I_{31}I_{23}) + \binom{I_{13}}{2}(I_{21}I_{32} + I_{31}I_{22}) \\ &+ \binom{I_{21}}{2}(I_{12}I_{33} + I_{32}I_{13}) + \binom{I_{22}}{2}(I_{11}I_{33} + I_{31}I_{13}) + \binom{I_{23}}{2}(I_{11}I_{32} + I_{31}I_{12}) \end{aligned} \tag{4}$$

Thus,

$$\begin{aligned} QI_{pro}(P, M_w) =&|\{\alpha_G(Q) : Q \subset \mathcal{L}_G, Q \xrightarrow{w} P\}| = \\ &\binom{I_{11}}{2} I_{22} I_{23} + \binom{I_{12}}{2} I_{21} I_{23} + \binom{I_{13}}{2} I_{21} I_{22} \\ &+ \binom{I_{11}}{2}(I_{22}I_{33} + I_{32}I_{23}) + \binom{I_{12}}{2}(I_{21}I_{33} + I_{31}I_{23}) + \binom{I_{13}}{2}(I_{21}I_{32} + I_{31}I_{22}) \\ &+ \binom{I_{21}}{2}(I_{12}I_{33} + I_{32}I_{13}) + \binom{I_{22}}{2}(I_{11}I_{33} + I_{31}I_{13}) + \binom{I_{23}}{2}(I_{11}I_{32} + I_{31}I_{12}) \end{aligned} \tag{5}$$

With simple manipulations, it can be shown that the right hand side of this equation can be rewritten as:

$$\sum_{(i,j,k) \in G_3, j<k} \binom{I_{1i}}{2} I_{2j} I_{2k} + \sum_{(i,j,k) \in G_3} \frac{I_{1i} I_{2j} I_{3k}(I_{1i} + I_{2j} - 2)}{2}$$

□

Proof of Lemma 3. Let $\Omega(G \restriction Q)$ be designated by $ab|cd$ and assume w.l.o.g that the anchor corresponding to a and b is the first anchor observed on the post-order traverse of G. It is easy to show (see [31]) that if $\Omega(G \restriction Q) \simeq S \restriction \alpha_G(Q)$ there exist exactly two tripartitions P^1 and P^2 in $\mathcal{P}(S)$ that imply a quartet topology that matches $\Omega(G \restriction Q)$ (condition (ii) of Definition 6). Each of the two tripartitions has two leaves of $\alpha_G(Q)$ in one of its parts and the other two leaves fall on two different parts. Also, the two leaves that are together can only be a and b or c and d and thus, only one of P^1 and P^2 would include both a and b in the same part. Therefore, by condition (iii) of Definition 6, exactly one of $Q \xrightarrow{\psi_G(Q)} P^1$ and $Q \xrightarrow{\psi_G(Q)} P^2$ can be true. □

Proof of Lemma 4.

$$\begin{aligned} q(S, \mathcal{G}) &= \sum_{G \in \mathcal{G}} q(S, G) && (6) \\ &= \sum_{G \in \mathcal{G}} |\{(\alpha_G(Q), \psi_G(Q)) : Q \subset \mathcal{L}_G, |Q| = |\alpha_G(Q)| = 4, \Omega(G \restriction Q) \simeq S \restriction \alpha_G(Q)\}| && (7) \\ &= \sum_{P \in \mathcal{P}(S)} \sum_{G \in \mathcal{G}} |\{(\alpha_G(Q), \psi_G(Q)) : Q \subset \mathcal{L}_G, Q \xrightarrow{\psi_G(Q)} P\}| && (8) \\ &= \sum_{P \in \mathcal{P}(S)} \sum_{G \in \mathcal{G}} \sum_{w \in I(G)} |\{\alpha_G(Q) : Q \subset \mathcal{L}_G, Q \xrightarrow{w} P\}| && (9) \\ &= \sum_{P \in \mathcal{P}(S)} \sum_{G \in \mathcal{G}} \sum_{w \in I(G)} |\{\alpha_G(Q) : Q \subset \mathcal{L}_G, Q \xrightarrow{w} P\}| \times \mathbf{1}_{speciation}(w) && (10) \\ &= \sum_{P \in \mathcal{P}(S)} \sum_{G \in \mathcal{G}} \sum_{w \in I(G)} QI_{pro}(P, M_w) \times \mathbf{1}_{speciation}(w) && (11) \\ &= \sum_{P \in \mathcal{P}(S)} w_{pro}(P) && (12) \end{aligned}$$

The first two lines are implied by Definition 5. Equation (8) follows from Lemma 1 and Lemma 3 that together establish that each equivalence class of quartets maps to exactly one P. Equation (9) follows from Definition 4 combined with a simple rearrangement obtained by counting unique tuples once. Equation (10) follows from the fact that when w is a duplication node, $|\{\alpha_G(Q) : Q \subset \mathcal{L}_G, Q \xrightarrow{w} P\}| = 0$. Equation (11) follows from Lemma 2. □

Proof sketch of Claim 1. The rooting that minimizes the number of duplications and losses (#duploss for short) in Alg. 1 may not be unique. In particular, if a rooted tree G minimizes #duploss, then rooting it at any branch such that the path between the parent node of the branch and the current root (including the two end nodes) does not contain any duplication node will also minimize #duploss. We call a correctly-tagged gene tree partially-correctly-rooted if the path between the parent node of the branch where it is rooted and the root in the correctly-rooted tree does not contain any duplication node. In particular, when gene trees do not have duplications, then any rooting of a gene tree is partially-correctly. We observe that the equivalent classes of quartets in all partially-correctly-rooted trees stay the same (although *all* quartet trees in the same equivalent class may change from balanced to unbalanced or vice versa), and thus any partially-correct-rooting of gene trees will result in the same species tree. □

Sketch of proof of Claim 2. When $\mathcal{G}$ only includes speciation nodes, regardless of rooting, each quartet is a SQ. Since each leaf corresponds to distinct taxa in the species tree, each quartet equivalent class contains only one quartet. Therefore, each quartet is counted exactly once and thus $\sum_{P \in \mathcal{P}(S)} w_{pro}(P) = \sum_{P \in \mathcal{P}(S)} w(P)$ regardless of rooting. □

Sketch of proof of Claim 3 (Running time of ASTRAL-Pro). Let

$$N = \sum_{G \in \mathcal{G}} |\mathcal{L}_G|$$

denote the sum of the number of leaves in the gene trees. Then the number of anchor LCAs in all gene trees is $O(N)$. Let D denote the number of unique gene tree tripartitions tagged as speciations and note $D = O(N)$. By only counting each unique gene tree tripartition once against each species tree tripartition, the running time of ASTRAL-Pro becomes $O(D|X|^{1.73})$ (by an argument that is identical to that provided for ASTRAL-III [31] and follows from results of Kane and Tao [80]). However, while ASTRAL-III guarantees $|X| = O(nk)$ with $k = |\mathcal{G}|$, in ASTRAL-Pro, in the presence of duplications, $|X|$ can be large; in particular with our sampling algorithm (Alg. 2), $|X| = O(nN)$. Thus, the running time of A-Pro is $O(D(nN)^{1.73})$. Note that this analysis is not tight and can be made more precise in the future. Also, in the future, we will explore sub-sampling a constant number of trees from the output of Alg. 2 per gene tree, which will limit the $|X| = nk$ and thus limit the running time of ASTRAL-pro to $O(D(nk)^{1.73})$. □

Proof of Proposition 3. Under GDL, besides leaves, each internal node $u_G \in I(G)$ in a gene tree G corresponds to an internal node $u_S \in I(S)$; if u_G is a duplication node, u_S is the node down the branch in S where the duplication event happened, and if u_G is a speciation node, u_S is the respective speciation node. It is easy to see that $\alpha_G(u_G) \subset \mathcal{L}_S(u_S)$. For each SQ quartet $Q = \{a, b, c, d\}$, assuming w.o.l.g that $G \restriction Q$ has unrooted topology $ab|cd$, let $w_G = \psi_G(Q)$, and u_G and v_G be the children of w_G. Let u_G, v_G, and w_G correspond to u_S, v_S, and w_S in S, respectively. Since w_G is a correctly tagged speciation node, u_S and v_S are descendants from different children of w_S.

When $Q \perp G$, assuming w.o.l.g. $a, b \in \mathcal{L}_G(u_G)$ and $c, d \in \mathcal{L}_G(v_G)$, we get $\alpha_G(a), \alpha_G(b) \in \mathcal{L}_S(u_S)$ and $\alpha_G(c), \alpha_G(d) \in \mathcal{L}_S(v_S)$ and thus $\alpha_G(a)\alpha_G(b)|\alpha_G(c)\alpha_G(d)$ is induced by S.

When $Q \angle G$, assuming w.o.l.g. $a, b \in \mathcal{L}_G(u_G)$, $c \in \mathcal{L}_G(v_G)$, and $d \notin \mathcal{L}_G(w_G)$, we get $\alpha_G(a), \alpha_G(b) \in \mathcal{L}_S(u_S)$ and $\alpha_G(c) \in \mathcal{L}_S(v_S)$. Since d is not under w_G, $\alpha_G(d)$ and w_S are under

different children of the species tree node to which the LCA of d and w_G corresponds. Therefore, $\alpha_G(d) \notin \mathcal{L}_S(w_S)$ and thus $\alpha_G(d) \notin \mathcal{L}_S(u_S)$; since $\alpha_G(a) \in \mathcal{L}_S(u_S)$ and $\alpha_G(b) \in \mathcal{L}_S(u_S)$, it follows that $\alpha_G(a)\alpha_G(b)|\alpha_G(c)\alpha_G(d)$ in S. $\square$

B Simulation details

Simphy command for default parameters:

```
simphy -sl f:25 -rs 50 -rl f:1000 -rg 1 -sb f:0.000000005 -sd f:0
-st ln:21.25,0.2 -so f:1 -si f:1 -sp f:470000000 -su ln:-21.9,0.1
-hh f:1 -hs ln:1.5,1 -hl ln:1.551533,0.6931472 -hg ln:1.5,1 -cs 9644
-v 3 -o default -ot 0 -op 1 -lb f:0.00000000049 -ld f:0.00000000049
-lt f:0
```

Other settings use a similar command with parameters changed according to the table below.

Table S1: Simphy parameters for all experiments

Parameter name	**Parameter value**
Default Parameters	
Speciation rate	5e-9
Extinction rate	0
Locus trees	1000
Gene trees	1
Number of leaves	25 + an outgroup
Ingroup divergence to the ingroup ratio	1.0
Generations	LogN(21.25,0.2)
Haploid effective population size	4.7e+8
Global substitution rate	LogN(-21.9,0.1)
Lineage specific rate gamma shape	LogN(1.5,1)
Gene family specific rate gamma shape	LogN(1.551533,0.6931472)
Gene tree branch specific rate gamma shape	LogN(1.5,1)
Duplication rate	4.9e-10
Loss rate to duplication rate ratio	1
Seed	9644
Sequence length	500, 100
Sequence base frequencies	Dirichlet(A=36,C=26,G=28,T=32)
Sequence transition rates	Dirichlet(TC=16,TA=3,TG=5,CA=5,CG=6,AG=15)
Controlling Duplication and Loss Rates (5×4 conditions)	
Duplication rate	4.9e-10, 2.7e-10, 1.9e-10, 5.2e-11, 0
Loss rate to duplication rate ratio	1, 0.5, 0.1, 0
Controlling Duplicatoin and ILS Rate (3×4 conditions)	
Duplication rate	4.9e-10, 1.9e-10, 0
Haploid effective population size	4.7e+8, 1.9e+8, 4.8e+7, 1e+4
Controlling n	
Number of leaves	10, 25, 100, 250, 500 + an outgroup
Controlling k	
Locus trees	25, 100, 250, 1000, 2500, 10000

C Tables

	1st	2nd	3rd	4th
MulRF	42	67	10	1
DupTree	28	8	15	69
A-Pro	105	14	1	0
ASTRAL-multi	12	14	71	23

Table S2: Rank of methods on S100 dataset over all 120 test conditions. Ranks are obtained using mean species tree error, rounded to two significant digits to create tie for cases where error values are extremely close.

D Figures

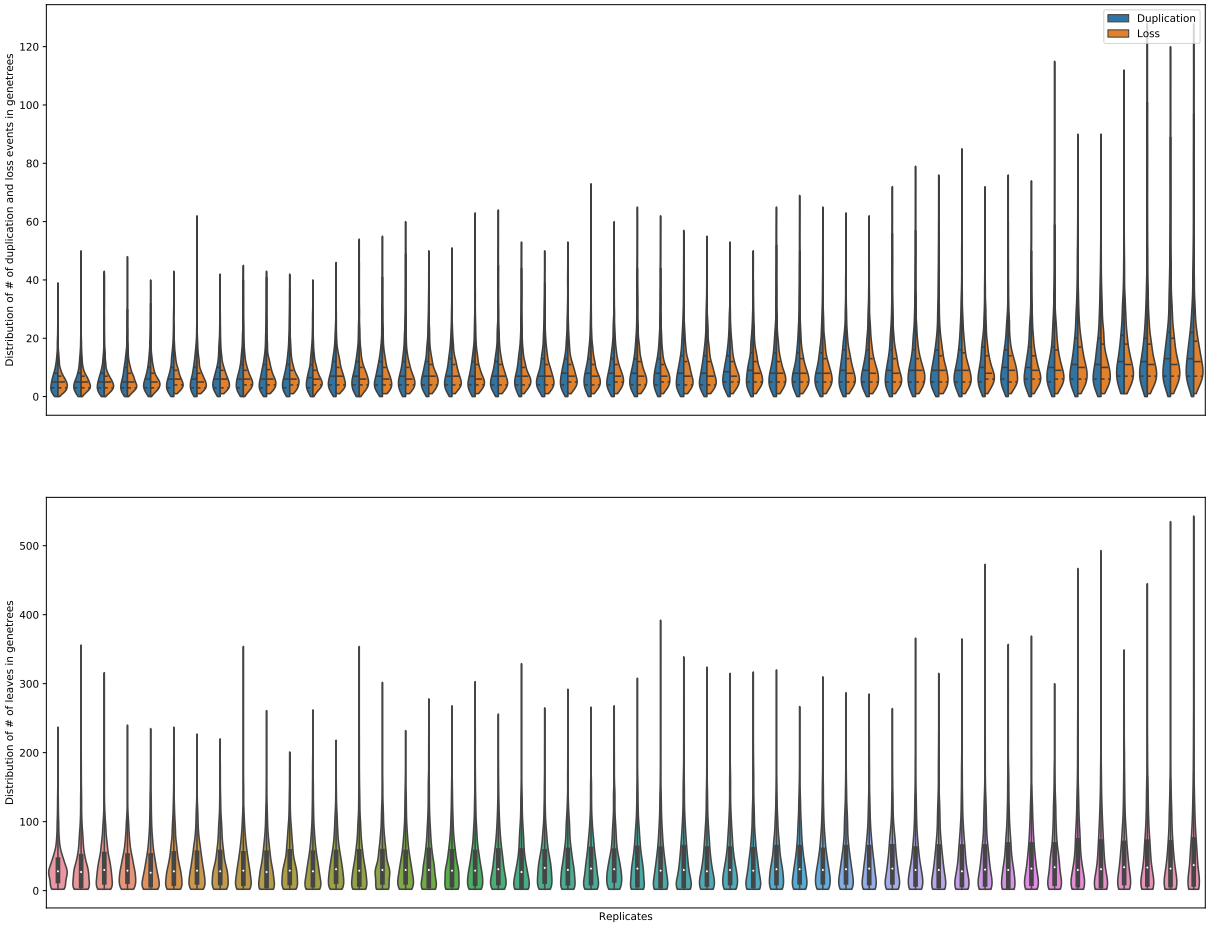


Figure S1: Distribution of the number of duplication events, loss events and sizes of leaf set for gene trees in the default condition by replicates. The figure on the top is sorted by the mean number of duplication events and the figure on the bottom is sorted by mean leaf set size.

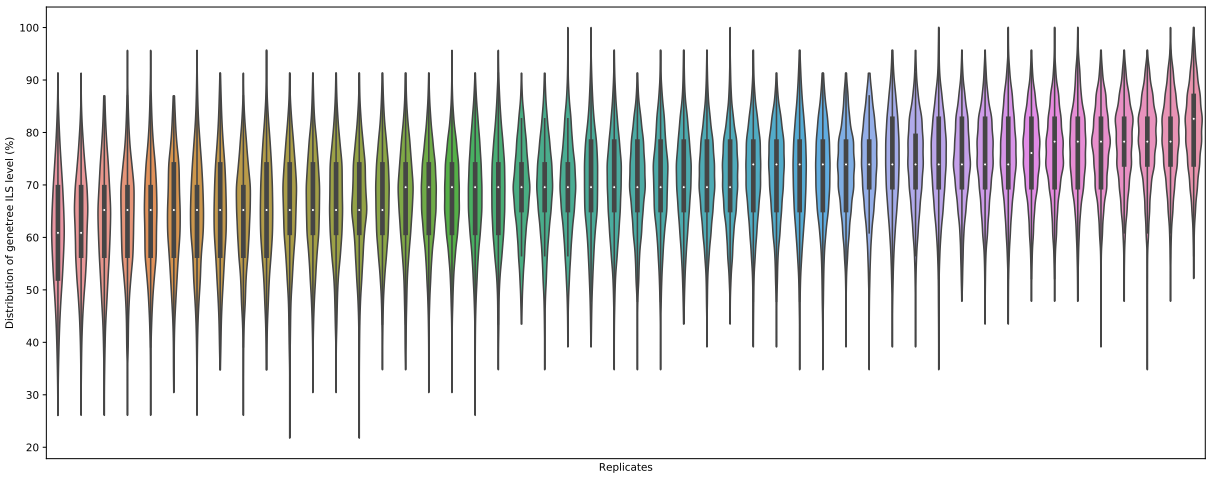


Figure S2: Distribution of genetree ILS, as measured by the normalized RF distance between true gene trees and the true species, in the condition with all default parameters but $\lambda_+ = \lambda_- = 0$. Results are divided by replicates, sorted by mean ILS level.

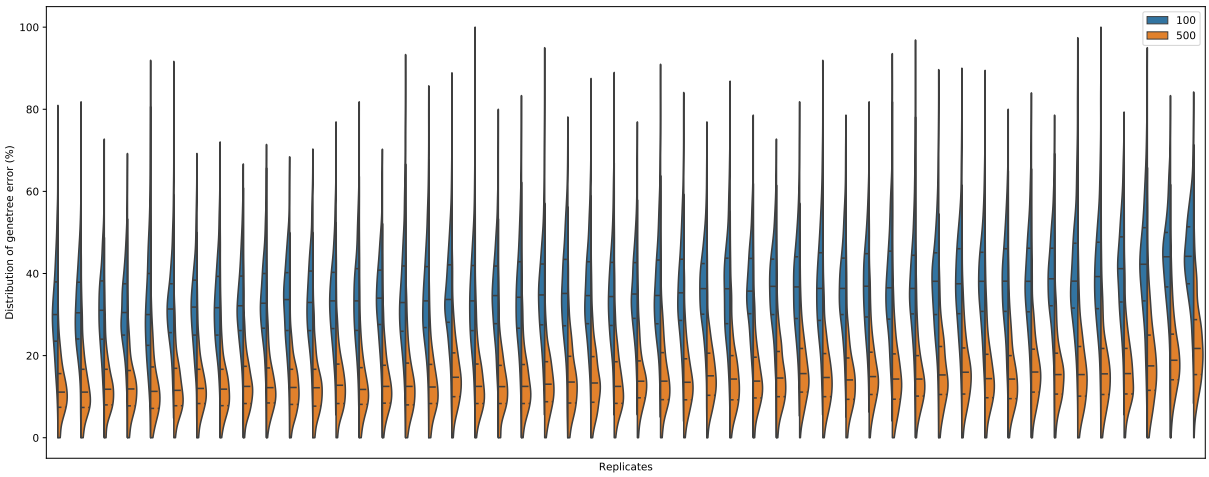


Figure S3: Distribution of the gene tree errors (normalized RF distance between true gene trees and the estimated gene tree) for inferred trees with at least 14 leaves in the default condition. Results are divided by sequence length (100bps or 500bps) and by replicates, sorted by mean gene tree error of the 100bps condition.

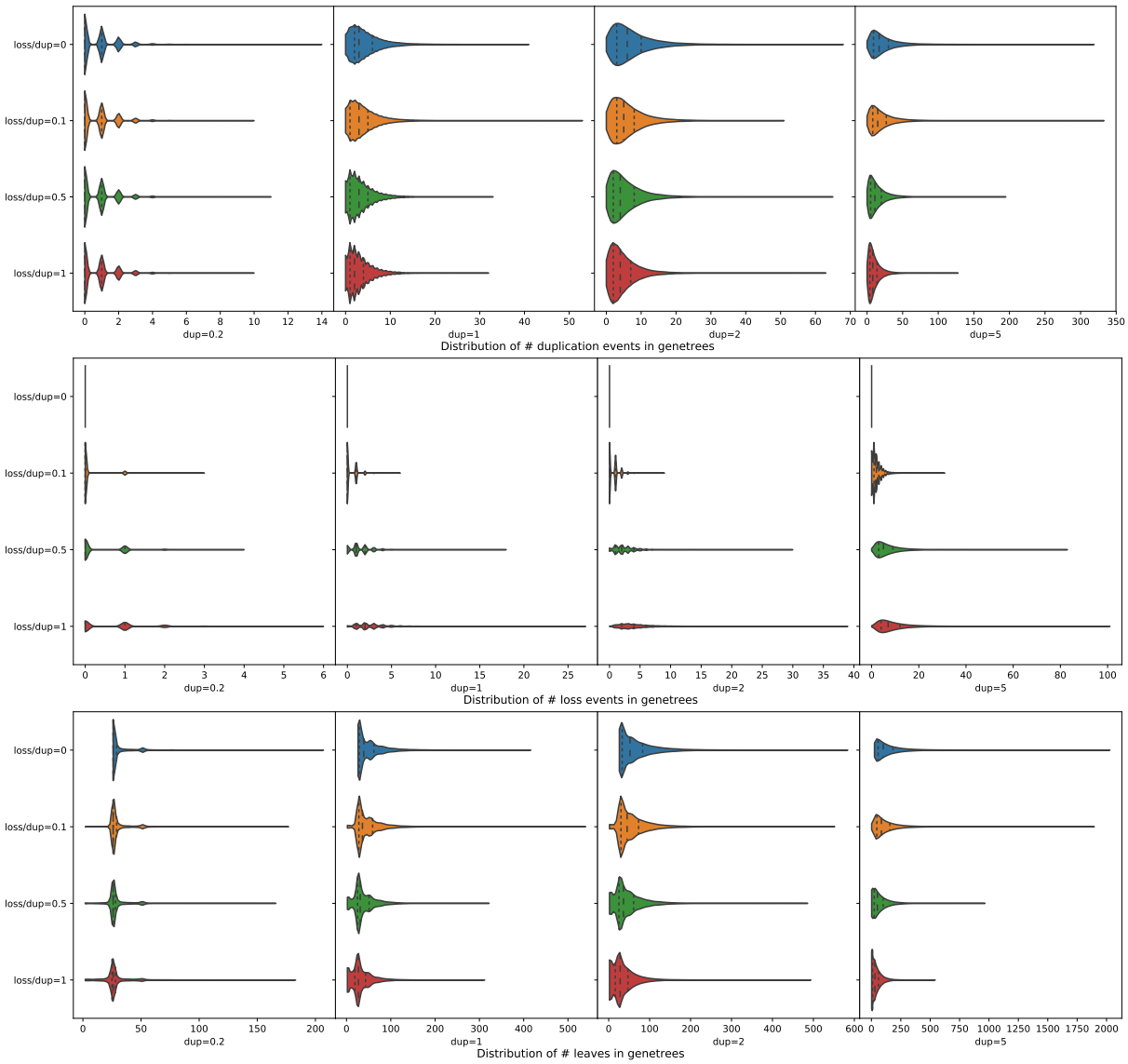


Figure S4: Distribution of the number of duplication events, loss events and sizes of leaf set for gene trees of each replicates sorted by duplication and loss rate.

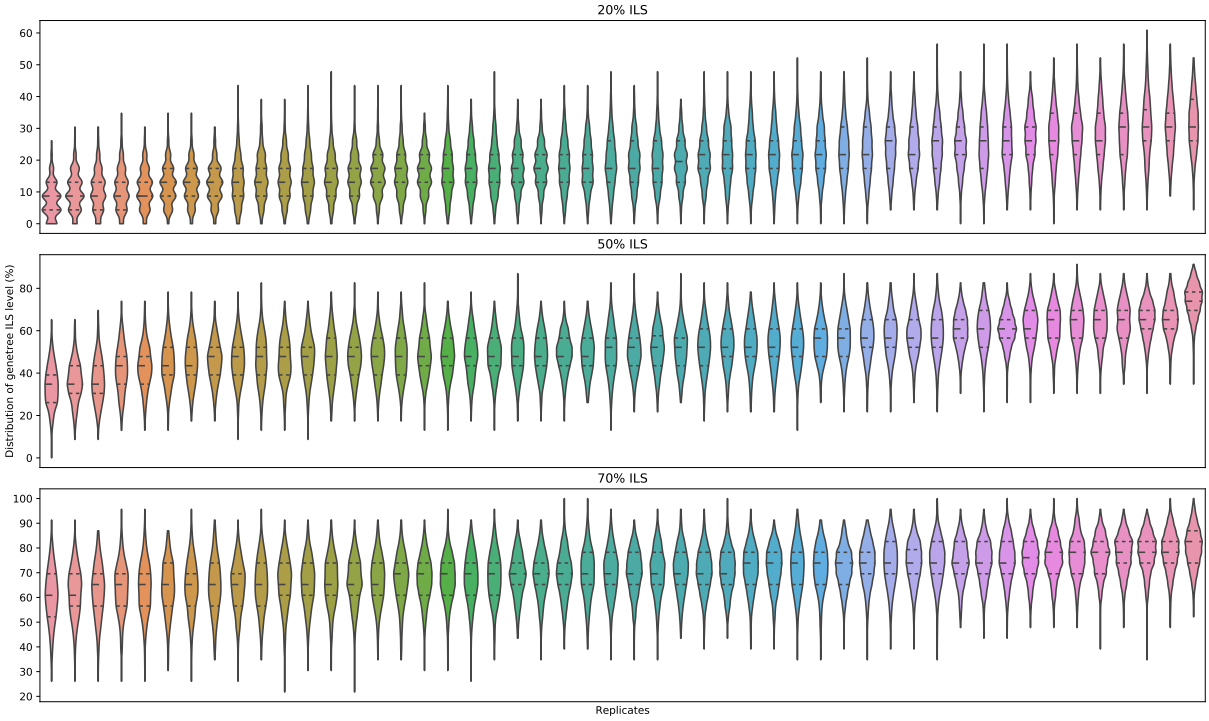


Figure S5: Distribution of gene tree ILS levels by replicates and expected ILS level, sorted by mean ILS level.

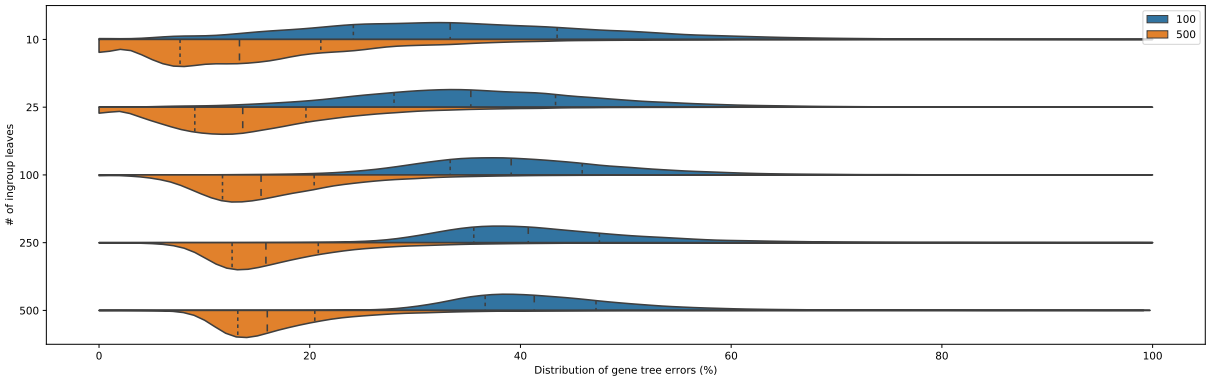


Figure S6: Distribution of gene tree errors by the number of ingroup species n.

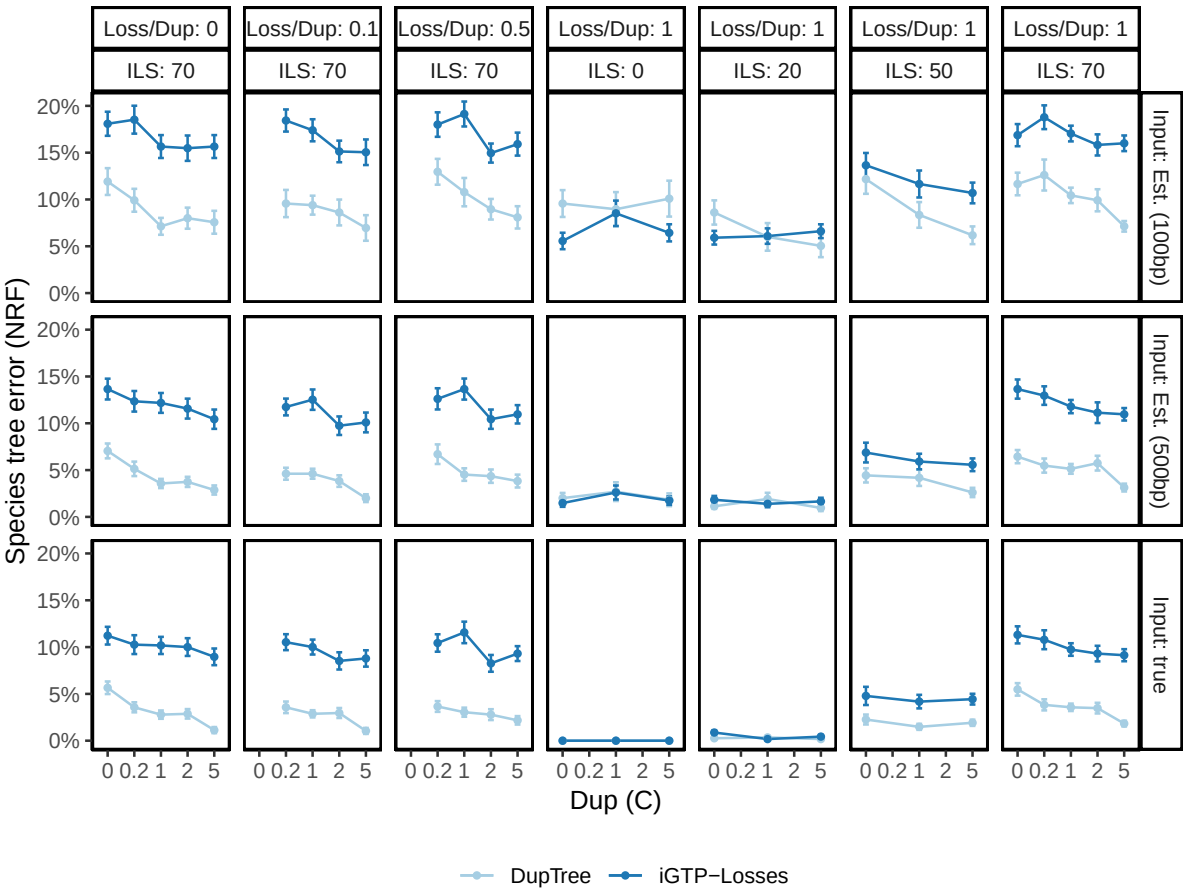


Figure S7: Comparison of DupTree and iGTP-DupLoss methods on all the datasets with $n = 25$ and $k = 1000$. DupTree dominates iGTP-DupLoss in most conditions.

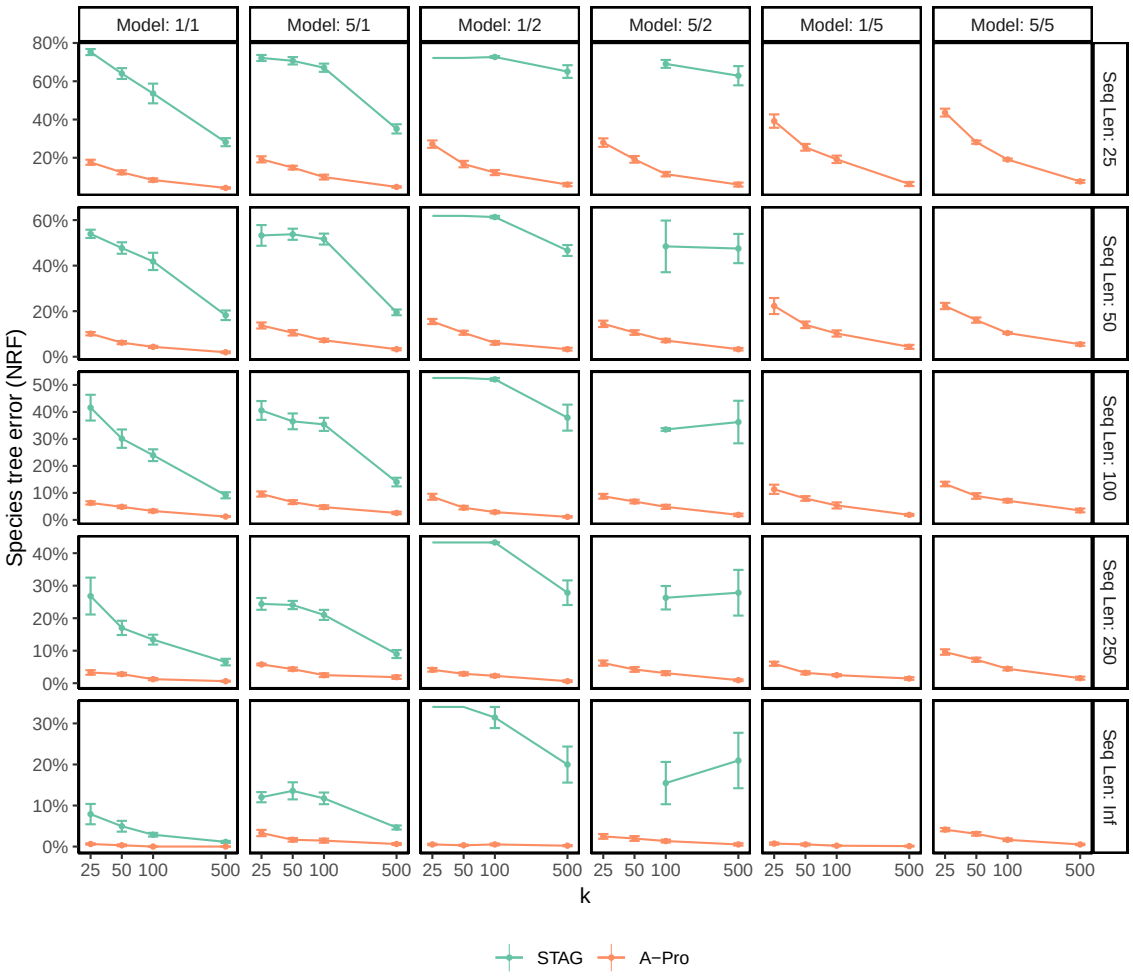


Figure S8: Species tree error on S100 dataset. We compare the species tree error of the STAG method to A-Pro, showing mean and standard error over 10 replicates for each model condition, with varying numbers of genes (k) and sequence lengths (with Inf signifying true gene trees). Model conditions are labeled as a/b where a is the level of ILS (1 or 5) and b is the duplication/loss rate (1, 2, or 5). Cases with missing STAG results are due to STAG failing to run on those model conditions. Note that STAG infers a species tree from the input gene trees that have at least one leaf representing each species of interest; if none of the input gene trees satisfy this requirement, then STAG fails to return a tree.

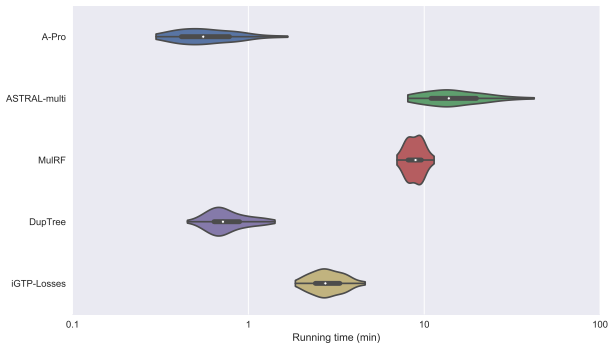


Figure S9: **Comparing running times**, measured on the default model condition, with estimated gene trees (100bp). All methods are run in the single-threaded mode, on the same machine with Intel(R) Xeon(R) CPU E5-2670 0 @ 2.60GHz.

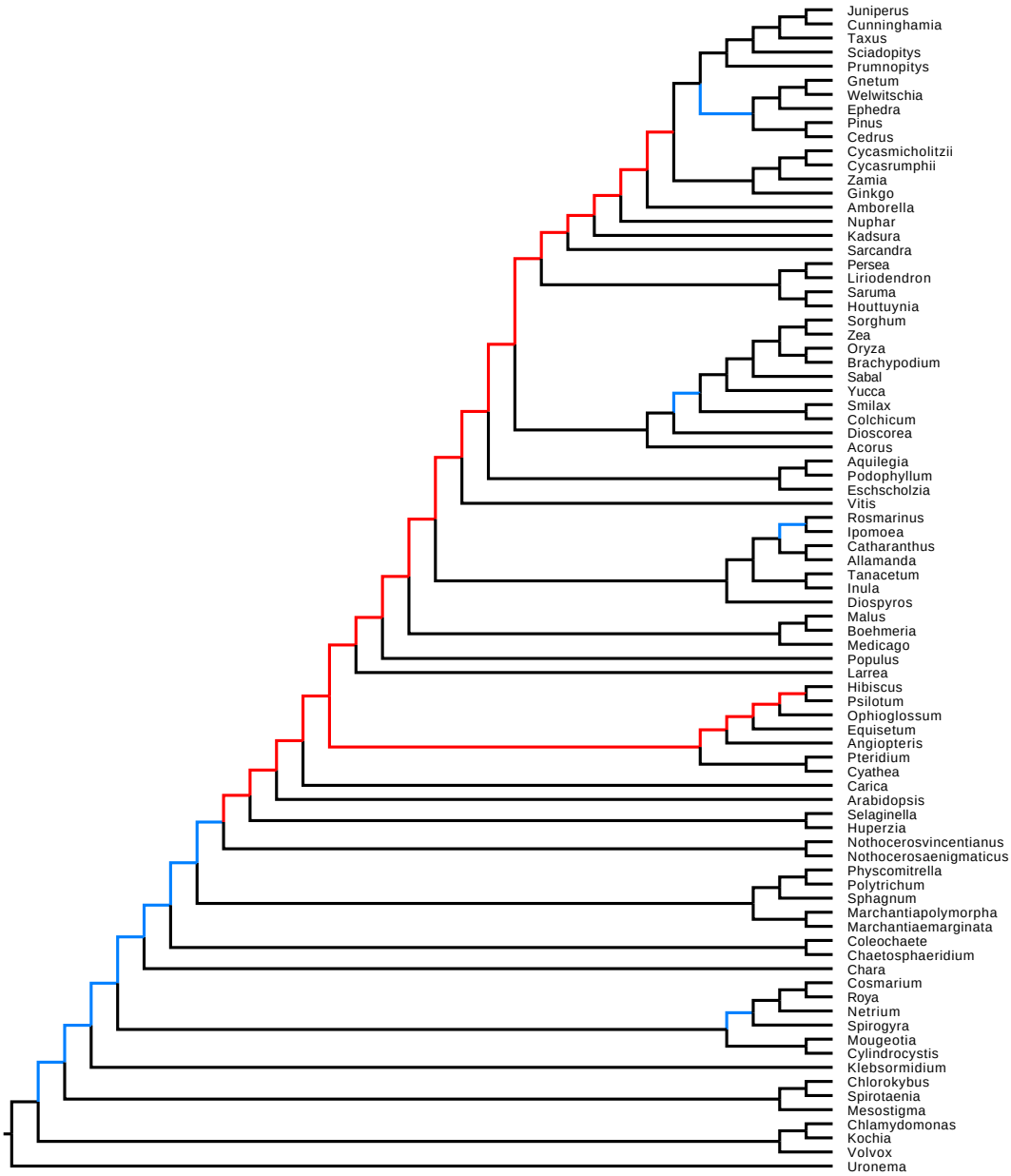


Figure S10: **DupTree on biological plant paper**. DupTree is run on 9683 multi-copy gene trees available online [56] for the plant dataset. Red: Branches that are obviously wrong, because these branches contradict basic biological categorization. Blue: Branches that contradict ASTRAL on single-copy genes that are not so obviously wrong.