wQFM-GDL Enables Accurate Quartet-based Genome-scale Species Tree Inference Under Gene Duplication and Loss

Abdur Rafi †[0000−0002−3899−0025], Ahmed Mahir Sultan Rumi†[0009−0006−7273−7417], Sheikh Azizul Hakim[0000−0002−3405−2402], and Md. Shamsuzzoha Bayzid[0000−0002−5640−0615]

Department of Computer Science and Engineering,
Bangladesh University of Engineering and Technology
shams_bayzid@cse.buet.ac.bd
† These authors contributed equally to this work.

Abstract. Species tree estimation from multi-copy gene family trees, including both paralogs and orthologs, is a challenging task due to the gene tree discordance caused by biological processes such as incomplete lineage sorting (ILS) and gene duplication and loss (GDL). Quartet-based species tree estimation methods, such as ASTRAL, Quartet Max-Cut (QMC), and Quartet Fiduccia–Mattheyses (QFM) frameworks have gained substantial popularity for their accuracy and statistical guarantee. However, most of these methods rely on single-copy gene trees and model only ILS, which limits their applicability to large genomic datasets. ASTRAL-Pro incorporates both orthology and paralogy for species tree inference under GDL by employing a refined quartet similarity measure based on the concept of species-driven quartets (SQs). In this study, we show that these SQ-based techniques can be effectively leveraged within the QFM framework. This required substantial algorithmic re-engineering, including the development of efficient techniques for computing the initial bipartition in QFM and novel combinatorial methods for computing refined quartet scores directly from gene family trees. We extensively evaluated our method, wQFM-GDL, on benchmark simulated and real biological datasets and compared it with ASTRAL-Pro3, SpeciesRax, and DupLoss-2. wQFM-GDL outperforms all other methods in 113 out of 124 model conditions considered in this study, with performance differences becoming more pronounced as dataset size increases. In particular, for larger datasets with 200 and 500 taxa, wQFM-GDL significantly outperforms all leading methods in all 72 out of 72 model conditions and achieves, on average, nearly a 25% reduction in reconstruction error compared with ASTRAL-Pro3. wQFM-GDL is freely available in open source form at https://github.com/abdur-rafi/wQFM-GDL.

Keywords: Species tree inference · Gene tree discordance · Gene duplication and loss · Orthologs and Paralogs

1 Introduction

Species tree inference from genes sampled throughout the whole genome is complicated by the fact that gene trees can differ from each other (and from the true species tree) due to several biological processes, including gene duplication and loss (GDL), horizontal gene transfer, incomplete lineage sorting (ILS), and hybridization [22]. In particular, gene duplication and loss is one of the most important processes shaping gene family evolution and often results in multi-copy gene trees [30].

Summary methods, which find a species tree by combining a set of gene trees, are becoming increasingly popular due to their accuracy and statistical guarantee under particular reasons for gene tree discordance [1, 7, 15–20, 23, 26, 29]. A large family of summary methods considers GDL as a source of gene tree discordance. Most of these methods are gene tree parsimony methods that seek a species tree by minimizing the total number of duplications and losses required to explain the observed gene trees. DupTree [38], iGTP [6], and their recent improvement DupLoss-2 [31], DynaDup [2–4], and earlier similar dynamic programming based methods [12] are some of the well-known parsimony-based methods. However, there are other methods, such as PHYLDOG [5] and Guenomu [8], that are more agnostic about the reasons for gene tree discordance and do not necessarily rely on maximum parsimony reconciliation.

Quartet-based summary methods have been among the most widely used approaches for species tree inference over the last decade. These methods address the Maximum Quartet Support Species Tree (MQSST) problem, which seeks to infer a species tree that maximizes the number of quartets induced by the gene trees that are consistent with the species tree. Within this framework, methods can be broadly categorized based on how they solve the MQSST problem. ASTRAL, the most widely used summary method, employs a dynamic programming strategy to optimize quartet support [26]. In contrast, quartet amalgamation approaches, such as QMC (Quartet MaxCut) [1, 36] and QFM (Quartet Fiduccia–Mattheyses) [24, 35], adopt divide-and-conquer strategies to assemble quartets (with or without weights) induced by gene trees. While these approaches offer conceptual simplicity and flexibility–since they can operate directly on quartets, as in [7], without requiring explicit gene tree estimation–their reliance on explicit quartet enumeration poses significant computational challenges for large datasets. Recent methods, notably TREE-QMC [13] and wQFM-TREE [33], address this limitation by eliminating the need for explicit quartet enumeration, thereby substantially improving scalability without sacrificing accuracy. These methods were originally designed to handle gene tree heterogeneity due to ILS and are therefore applicable to single-copy gene trees.

ASTRAL-Pro [43] extends the ASTRAL framework to model GDL and remains the only quartet-based species tree inference method that explicitly accounts for duplication and loss. It introduces a new quartet similarity measure that accounts for both orthology and paralogy [43]. However, despite their effectiveness, no existing quartet amalgamation–based summary methods, such as wQFM and wQMC, explicitly model GDL.

In this study, we introduce wQFM-GDL, an extension of the QFM framework that explicitly accounts for GDL and is the first initiative to model GDL in a quartet amalgamation method. The proposed method includes two variants, wQFM-GDL-Q and wQFM-GDL-T, which operate on quartets and gene trees, respectively. In particular, we make the following key contributions. We extend the wQFM-TREE framework so it can account for speciation-driven quartets (SQs) by developing efficient graph-theoretic and combinatorial techniques to compute them directly from gene family trees without quartet enumeration – leading to wQFM-GDL-T. We also propose wQFM-GDL-Q where we enumerate the SQs from a given set of gene family trees and apply wQFM on this set of SQs. Notably, this set of SQs can be directly given as input to any quartet amalgamation methods (e.g., wQFM, wQMC). We found that, on small to moderate size dataset, where quartet enumeration is feasible, wQFM-GDL-Q is often more accurate than wQFM-GDL-T. Importantly, we also leveraged a locus-aware normalization scheme for improved accuracy. Our extensive experimental analysis on established benchmark datasets shows that wQFM-GDL achieves the highest accuracy among the evaluated methods. Overall, across a wide range of model conditions with varying levels of duplication and loss, ILS, and gene tree estimation error, wQFM-GDL obtained the best performance in 113 out of 124 model conditions considered in this study. wQFM-GDL performs especially well on large datasets such as comprising 200 and 500 taxa, where wQFM-GDL significantly outperformed all other methods across all 72 model conditions with an average 25% reduction in tree error relative to ASTRAL-Pro - making it particularly suitable for large-scale phylogenomics datasets. We also reanalyzed three real biological datasets, Plants83 [39], Vertebrates188 [28], and Arechaea364 [9] and the results provide evidence of wQFM-GDL's robustness and accuracy.

2 Materials and Methods

We first provide a brief overview of wQFM, wQFM-TREE and the duplication-aware quartet similarity measure used in ASTRAL-Pro. Then, we describe the key algorithmic components introduced in wQFM-GDL in detail.

2.1 Overview of wQFM and wQFM-TREE

wQFM [24] is a quartet amalgamation method that takes as input a set of weighted quartets induced by a given set of gene trees and infers a species tree by maximizing quartet consistency. The algorithm follows a divide-and-conquer strategy (Figure 1A): at each divide step, the current taxa set is partitioned into two disjoint subsets, defining two subproblems. For each divide step, wQFM starts from an initial bipartition on the current taxa set and iteratively refines it using a heuristic inspired by the Fiduccia–Mattheyses (FM) algorithm for hypergraph bipartitioning [11], with the objective of maximizing the difference between the number of satisfied and violated quartets. The procedure is applied recursively to the resulting subproblems until each subproblem contains at most

three taxa, for which the solution is trivial. At each divide step, a dummy (artificial) taxon is introduced into each partition to represent the taxa in the complementary partition. During the conquer phase, the solutions to the subproblems (subproblem trees) are merged by connecting them through their corresponding dummy taxa, ultimately producing our final species tree.

Central to wQFM's divide-and-conquer process is the identification of a suitable bipartition of the taxa set at each divide step. Each candidate bipartition is scored with respect to the input quartet set. A quartet $q = ((A, B), (C, D))$ is *satisfied* with respect to a bipartition (P_a, P_b) if taxa A and B lie in one partition and taxa C and D lie in the other, and is *violated* if taxa A and C (or A and D) lie in one partition while taxa B and D (or B and C) lie in the other. Otherwise, the quartet is *deferred*. The *score of a bipartition* is defined as the difference between the number of satisfied and violated quartets.

The main computational bottleneck of wQFM is the explicit enumeration of all quartets induced by the input gene trees. In earlier work, we developed wQFM-TREE [33] to address this limitation by enabling the wQFM framework to operate directly on gene trees without explicit quartet enumeration. This was achieved through novel algorithmic techniques that combine a gene tree consensus–based heuristic for constructing initial bipartitions at each divide step with combinatorial and graph-theoretic methods for computing the scores of candidate bipartitions directly from gene trees. We now revisit the normalization schemes used in TREE-QMC [13] and later adopted by wQFM-TREE, as they play an important role in the accuracy of the methods and are essential for understanding the new locus-aware normalization approach we propose in this study.

Normalization scheme Normalization is a critical component in the QMC and QFM frameworks because the divide-and-conquer process introduces artificial or dummy taxa into subproblems, which enables the subproblem trees to be merged in the conquer phase. Consequently, every non-root subproblem in the recursion contains one or more dummy taxa, whereas the input gene trees and the quartets induced contain only real taxa.

To compute satisfied and violated quartets with respect to a subproblem that contains dummy taxa, the leaves of each gene tree are relabeled appropriately by dummy taxa. A dummy taxon represents all real taxa in its sister subproblem, and the corresponding leaves are relabeled using that dummy taxon. Thus, the resulting gene trees only contain the taxa present in the subproblem, including both real and dummy taxa. An important consideration is that, in this relabeled representation, multiple leaves of a gene tree maybe relabeled by the same dummy taxon, which can inflate the contribution of quartets involving dummy taxa. However, from the perspective of the current subproblem, a dummy taxon should have the same influence as any single real taxon. To address this, quartet weights are normalized so that every subset of four taxa in a subproblem, including those containing dummy taxa, contributes a single vote toward the tally of satisfied or violated quartets.

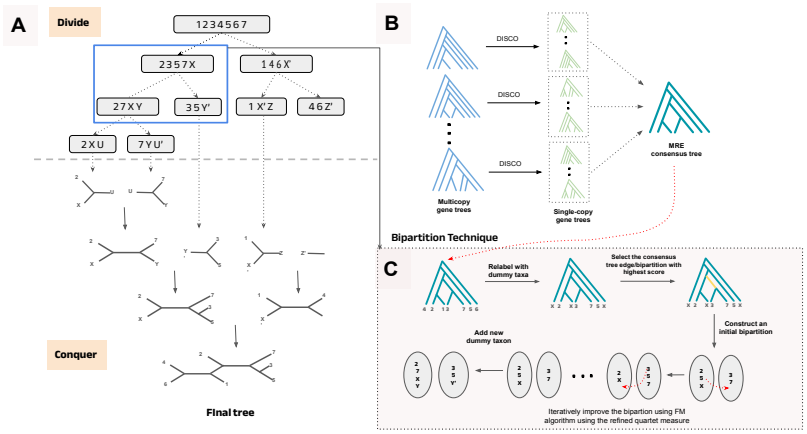

Fig. 1: **Overview of our proposed method.** (A) The divide-and-conquer framework used in all QFM methods. (B) An MRE consensus tree is constructed from the multicopy gene trees after decomposing them into single-copy gene trees using DISCO which will be used to create an initial bipartition. (C) After generating the initial biparition, it is iteratively improved using FM heuristic where the bipartitions are evaluated using the refined quartet measure addressing GDL.

Normalization was used in wQFM prior to TREE-QMC and wQFM-TREE but it was fairly simple because the full quartet set is available as input. Recently, TREE-QMC [13] introduced an effective normalization strategy which directly works on gene trees, and wQFM-TREE adopts a similar approach. Under this strategy, each taxon in a gene tree is assigned a weight and the weight of a quartet is defined as the product of the weights of its four taxa. Consequently, the total contribution of any quartet to the bipartition score reflects the effective representation of its constituent taxa in the current subproblem. The taxa that are absent from a subproblem but are represented through a dummy taxon receive appropriately reduced weights. This ensures that, for every subset of four taxa in a subproblem, including those containing dummy taxa, gets one vote per gene tree. The taxa are weighted non-uniformly depending on the subproblem decomposition in the recursion tree and taxa that are more relevant to a particular subproblem are assigned higher weights. Details of this weighting scheme are provided in Supplementary Section 1. wQFM-GDL extends and improves the normalization scheme of wQFM-TREE to account for GDL (details in Section 2.3).

2.2 Overview of ASTRAL-Pro: Solving Maximum per-Locus Quartet-score Species Tree (MLQST)

To account for orthology and paralogy, ASTRAL-Pro refines the quartet similarity measure of the original ASTRAL in mainly two ways.

- Firstly, it only considers "speciation-driven quartets" (SQs) and excludes duplication quartets (DQs), which contain no information regarding speciation events.
- Secondly, it aggregates the speciation quartets that mandatorily share the same topology and seperately do not provide any new information. These are counted as one unit, forming a quartet equivalence class.

A quartet in a multicopy gene tree is classified as a speciation quartet (SQ) if, for every subset of three leaves, the corresponding most recent common ancestor (MRCA) is an internal node representing a speciation event. The MRCA/LCA of the two anchors of a quartet is called the anchor LCA. All the SQs on the same four species that share the same anchor LCA must also share the same topology [43] (Details in Supplementary Section 1). They are counted as one unit to prevent double counting. Importantly, to classify quartets as SQs or DQs, the input trees must be rooted and tagged. ASTRAL-Pro defines the per-locus quartet score of a species tree with respect to a gene family tree with tagged internal nodes to be the number of quartet equivalence classes of the gene trees agreeing with the species tree. Thus, it tries to solve the Maximum per-Locus Quartet-score Species Tree (MLQST) problem by reconstructing the tree having the maximum total per-locus quartet score with respect to the multicopy gene trees.

2.3 wQFM-GDL-T: Extending wQFM-TREE for GDL

wQFM-GDL-T extends wQFM-TREE to accommodate GDL by solving the MLQST problem as in ASTRAL-Pro through three principal enhancements. First, we extend the consensus-tree-based heuristic to construct an appropriate initial bipartition of wQFM-TREE for multicopy gene trees. Second, we develop combinatorial techniques that enable efficient scoring of candidate bipartitions directly from multicopy gene trees without explicit quartet enumeration. Third, we enhance the normalization strategy of wQFM-TREE by proposing a locus-aware normalization scheme that explicitly accounts for GDL.

Initial bipartition At each divide step, we initiate the FM heuristic by generating an initial bipartition of the taxa set of the subproblem, which would then be refined iteratively, and the quality of this initial bipartition has a substantial impact on the quality of the final bipartition. In wQFM-TREE, an initial bipartition is obtained using a consensus-tree–based heuristic, in which an edge of a consensus tree constructed from gene trees is selected for initial bipartition. wQFM-TREE scores each bipartition according to its scoring scheme and selects the bipartition with the highest score. [33]. However, for multicopy gene trees, a consensus tree is not directly defined. To address this limitation, we employ the method DISCO [41] to decompose multicopy gene trees into single-copy gene trees. We then construct a greedy consensus tree using the Majority Rule Extended (MRE) method implemented in the PAUP package [37], which supports

incomplete gene trees as well. Then, we score each bipartition of the consensus tree using the updated scoring method addressing GDL as described later in this section and select the best one. The consensus tree is constructed once and used for all subproblems.

Scoring a Candidate Bipartition After constructing the initial bipartition, we iteratively improve it. The taxa are transferred from one partition to the other iteratively following the FM heuristic to obtain progressively improved bipartitions. Each bipartition is scored to assess its quality. As described earlier, the score has been defined in the QFM framework as the difference between the number of quartets from the gene trees that the bipartition satisfies and violates.

The scoring mechanism in wQFM-GDL differs substantially from wQFM-TREE, as it must account for the complexities introduced by GDL. In the original wQFM-TREE, all quartets from single-copy gene trees are considered when computing the score. In contrast, for multicopy gene family trees, following the GDL model of ASTRAL-Pro, we restrict the calculation of satisfied and violated quartets only to speciation-driven quartets (SQs) and treat each quartet equivalence class as a single unit to avoid double-counting. To integrate these duplication-aware measures while preserving the scalability of the original wQFM-TREE, we had to introduce substantial modifications to the underlying combinatorial and graph-theoretic techniques of wQFM-TREE to enable the exclusion of all the duplication-driven quartets and the aggregation of equivalent quartets within the scoring framework. The extended terminology and the updated scoring procedure are described below.

Let $\mathcal{G}$ be a set of rooted and tagged (each node labelled as either speciation or duplication) gene trees on the taxa set $\mathcal{X}$. We root and tag using the algorithm of ASTRAL-Pro. Let (A, B) be a candidate bipartition of the taxa set S in a divide step. We define R_A and D_A as the sets of real and dummy taxa in A, respectively. We define F_A as the set of all real taxa that are either in R_A or represented by a dummy taxon $X \in D_A$. We similarly define R_B, D_B, and F_B. Now, a quartet $ab|cd$ in a rooted and tagged gene tree is an SQ only if the MRCAs for all triplets in {a,b,c,d} are speciation nodes in its respective gene tree (Details in Supplementary Section 1). We categorize the SQs using the same traditional approach as in the QFM framework.

1. **Satisfied SQ:** SQ in which either $\{a, b\} \subseteq F_A, \{c, d\} \subseteq F_B$ or vice versa.
2. **Violated SQ:** SQ in which either $\{a, c\} \subseteq F_A, \{b, d\} \subseteq F_B$ or $\{a, d\} \subseteq F_B, \{b, c\} \subseteq F_A$ or vice versa.
3. **Deferred SQ:** Any other SQ.

Now, our algorithm assigns weight to each real taxon in $\mathcal{X}$ according to the normalization process (Supplementary Section 2.4). We consider a quartet as composed of two unordered pairs $\{a, b\}$ and $\{c, d\}$. The weight of a pair, $w(\{a, b\}) = w(a) \cdot w(b)$ and weight of a quartet $ab|cd$ is defined as $w(ab|cd) = w(\{a, b\}) \cdot w(\{c, d\}) = w(a) \cdot w(b) \cdot w(c) \cdot w(d)$. The weight of a set Q of quartets, $w(Q) = \sum_{q \in Q} w(q)$.

Let $S_{SQ}^{(g)}$ and $V_{SQ}^{(g)}$ denote the sets of satisfied and violated SQs, respectively, each containing only one representative from every equivalence class in a gene tree $g \in \mathcal{G}$. The score of a candidate bipartition (A, B) with respect to $\mathcal{G}$ is defined as: $\text{Score}(A, B, \mathcal{G}) = \sum_{g \in \mathcal{G}} \left(w(S_{SQ}^{(g)}) - w(V_{SQ}^{(g)}) \right)$. In this current setting, the total weight of potential satisfied and violated quartets are dependent on the specific speciation and duplication event tagging within each gene family tree which necessitates the use of more complex combinatorial and graph-theoretic techniques to calculate $w(S_{SQ}^{(g)})$ and $w(V_{SQ}^{(g)})$ directly from the gene family trees. We provide a high-level overview of the process below.

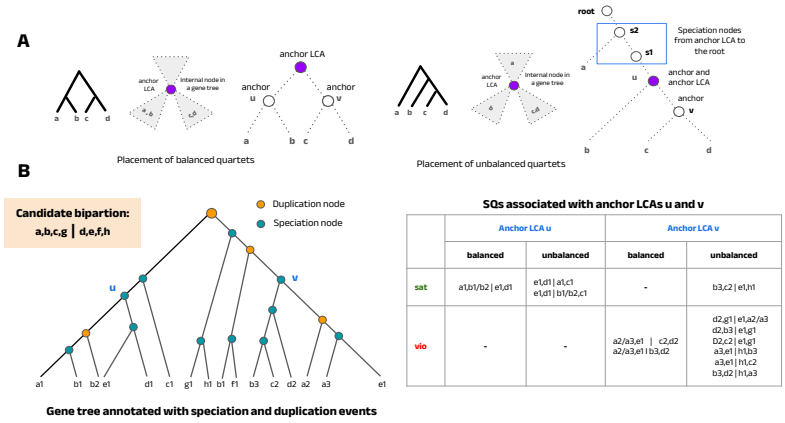

Fig. 2: **Scoring a candidate bipartition.** (A) Placement of a balanced and unbalanced quartet in a gene tree with respect to their anchor LCA. (B) SQs of different categories, illustrated using anchor LCAs u and v for an example bipartition $a, b, c, g|d, e, f, h$. Each speciation node may act as an anchor LCA for some SQs and for scoring a bipartition, we compute the total weight of satisfied and violated SQs associated with each anchor LCA independently, treating balanced and unbalanced quartets separately and then sum the results.

Computing $w(S_{SQ}^{(g)})$ for Multi-copy Trees Our key observation is that the total weight of all satisfied SQs sharing the same anchor LCA (Least Common Ancestor (LCA) of the two anchors of a quartet) can be computed efficiently and independently, and then aggregated for all anchor LCAs. Each internal node in the gene tree can serve as a potential anchor LCA for a set of SQs if and only if it is labeled as a speciation node, and the corresponding SQs associated with it can be computed as follows.

Rooted quartets exhibit two distinct forms: balanced and unbalanced, as shown in Figure 2a. Our observation is that, for each particular anchor LCA, we can efficiently calculate the balanced and unbalanced satisfied SQs separately. Let u be an internal node tagged as speciation with child branches i and j. We

define $PA_i^{(g,u)}$ and $PB_i^{(g,u)}$ as the weights of the unordered pairs of taxa in partitions A and B, respectively, within branch i. $PA_j^{(g,u)}$ and $PB_j^{(g,u)}$ are defined similarly. Importantly, multiple identical pairs may arise within the branches, but we count only one representative instance. This is crucial to ensure that only a single quartet from each equivalence class contributes to the score, as identical quartets formed from these identical pairs share the same anchor LCA and can be considered equivalent (see Supplementary Section 1). Their topology is already determined by the speciation event at the anchor LCA. These quartets arise due to paralogy along branches i and j and do not provide additional speciation information. This represents an important distinction from the calculation procedure used in the original wQFM-TREE.

Balanced Structure. A rooted quartet is balanced if the root splits the taxa into two groups of two. A balanced SQ with anchor LCA u will be anchored by two internal nodes in branch i and j, each subtending exactly two leaves (Figure 2a). The SQ is satisfied if the pairs of leaves two branches belong to opposite partitions (A and B). Let $S_{1,SQ}^{(g,u)}$ denote the weight of balanced satisfied SQs with anchor LCA u. Thus, we compute

$$w(S_{1,SQ}^{(g,u)}) = w(PA_i^{(g,u)}) \cdot w(PB_j^{(g,u)}) + w(PA_j^{(g,u)}) \cdot w(PB_i^{(g,u)})$$

Unbalanced Structure. A rooted quartet is unbalanced if the root separates one taxon from the other three, creating a ladder-like (pectinate) structure. For an unbalanced SQ with anchor LCA u, one anchor lies in a child branch, while the second anchor is the node u itself. In this case, a pair of taxa from one child branch (e.g., branch i) is grouped with a taxon from a sibling branch j and a fourth taxon from the parent lineage k. Importantly, this fourth taxon cannot just be any arbitrary taxon from the parent branch. It must originate from one of the branches associated with speciation nodes along the path from u to the root. Otherwise, the defining condition of a speciation quartet, that the LCA of any selection of three leaves must be a speciation node, would be violated. Now, we define $PA_{i,k}^{(g,u)}$ as the weight of taxon pairs split between branch i and parent lineage k. $PB_{i,k}^{(g,u)}$ is defined similarly. Thus, the weight of unbalanced satisfied SQs with anchor LCA u,

$$\begin{aligned} w(S_{2,SQ}^{(g,u)}) = {} & w(PA_i^{(g,u)}) \cdot w(PB_{j,k}^{(g,u)}) + w(PA_j^{(g,u)}) \cdot w(PB_{i,k}^{(g,u)}) \\ & + w(PB_i^{(g,u)}) \cdot w(PA_{j,k}^{(g,u)}) + w(PB_j^{(g,u)}) \cdot w(PA_{i,k}^{(g,u)}) \end{aligned}$$

Finally, $w(S_{SQ}^{(g,u)}) = w(S_{1,SQ}^{(g,u)}) + w(S_{2,SQ}^{(g,u)})$. Now, we can simply aggregate the results for each anchor LCA. $w(S_{SQ}^{(g)}) = \sum_{u \in \text{SpeciationNodes}} w(S_{SQ}^{(g,u)})$. An example for this process is shown in Figure 2.

Here, we observe that duplication-derived quartets in the gene trees are effectively excluded, as the LCA of at least one triplet among the four taxa of these quartets corresponds to a duplication node and therefore does not satisfy any of the criteria mentioned for getting considered in both balanced and unbalanced quartet calculations.

Computing $w(V_{SQ}^{(g)})$ for Multi-copy Trees The calculation of $w(V_{SQ}^{(g)})$ follows a similar approach. We compute the total weight of violated quartets independently for each anchor LCA, treating balanced and unbalanced structures separately, and then sum the results

Let us consider the same scenario with anchor LCA u. For the balanced violated quartets, each branch will contain a pair, one from partition A and the other from B. Thus, the weight of violated quartets as follows. $w(V_{1,SQ}^{(g,u)}) = w(PA_{i,j}^{(g,u)}) \cdot w(PB_{i,j}^{(g,u)})$. For the unbalanced ones, $w(V_{2,SQ}^{(g)}) = w(PA_{i,k}^{(g,u)}) \cdot w(PB_{i,j}^{(g,u)}) + w(PA_{i,j}^{(g,u)}) \cdot w(PB_{i,k}^{(g,u)}) + w(PA_{j,k}^{(g,u)}) \cdot w(PB_{j,i}^{(g,u)}) + w(PA_{j,i}^{(g,u)}) \cdot w(PB_{j,k}^{(g,u)})$. Then we can add the quantities as before.

Now, for calculating the terms $w(PA_{i}^{(g,u)}), w(PB_{i}^{(g,u)}), w(PA_{i,j}^{(g,u)})$ and so on efficiently and without actually enumerating the quartet sets, we use further combinatorial techniques and efficient graph traversal algorithms. Details are presented in Supplementary Section 2.

Locus-aware normalization scheme The prior normalization scheme used in wQFM-TREE could also be applied directly to wQFM-GDL. However, we identified an opportunity to improve normalization by introducing different normalization factors for distinct locus-specific regions in multicopy gene trees. Here, the the concept of locus tree plays a crucial role. Rasmussen and Kellis proposed the DLCoal model [34] for jointly modelling GDL and ILS, which introduces a third tree called the locus tree, alongside the gene tree and the species tree.The locus tree is derived from the species tree via a top-down duplication–loss process, with each internal node labeled as either a speciation or a duplication event. A duplication event creates a new daughter locus from the parent locus. Both the parent and daughter loci evolve independently thereafter and can undergo further duplications or losses in subsequent evolutionary steps. Finally, a gene family tree is constructed from the locus tree by applying a multi-locus coalescent process.

Unlike the case with single-copy gene trees, a particular set of four species (e.g., a, b, c, d) can appear multiple times within different locus-specific subtrees of a particular gene tree. Now, note that the set of taxa in the various locus-specific subtrees may not be identical because each locus undergoes a distinct evolutionary history, marked by different duplication and loss events. Importantly, the normalization factors used in our method depends on the taxa set. The normalization scheme used in QMC and QFM families of methods does not consider this and thus any quartet on a, b, c, d in a gene tree is assigned the same weight. However, our observation is that, the quartets on a particular set of four species across different locus-specific subtrees should be assigned different weights as the taxa set of different locus-specific subtrees may vary. Recently, it has been shown that normalization of TREE-QMC can be affected in case of missing taxa in incomplete gene trees and thus needs correction [14]. However, in case of gene family trees, identifying the taxa set associated with a particular loci within a multicopy gene tree is not straightforward. Our key step

is that all quartets within an equivalence class descend from the same ancestral locus at the time of the speciation event corresponding to the anchor LCA, and the normalization factors for quartets associated with each anchor LCA are computed separately. Details are provided in Supplementary Section 2.4. This locus-aware normalization scheme notably improved upon the traditional normalization (used in TREE-QMC and wQFM-TREE), especially under scenarios with substantial duplications and losses (see Supplementary Section 2.4).

2.4 wQFM-GDL-Q: Extending wQFM for GDL

wQFM takes as input a set of weighted quartets where the weight is the frequency of the quartet in the input gene trees. We devise an algorithm to create an appropriate weighted quartet set that only contains SQs and only one representative from an equivalent class. Then, this quartet set is given input to the wQFM algorithm which remains unchanged. We call this method wQFM-GDL-Q. For generating this quartet set, we adopt a similar algorithmic approach to the score calculation process in wQFM-GDL-T; however, in this case, we explicitly enumerate quartets with appropriate weights, rather than computing the sum of quartet weights. As expected, wQFM-GDL-Q is considerably slower than wQFM-GDL-T. However, its main advantage is that under smaller model conditions, where explicit quartet enumeration is feasible, we observe that it outperforms competing methods, including wQFM-GDL-T, in several settings. This may be because the heuristics used in wQFM-GDL-T, which allow it to operate directly on gene trees, can sometimes influence its accuracy. The generated quartet set can be given input to other quartet amalgamation methods like wQMC. Details, pseudocode, and additional results regarding wQFM-GDL-Q and wQMC are presented in Supplementary Section 3.

2.5 Complexity Analysis

The total time complexity for wQFM-GDL-T is $O(n^2k + \alpha n^2 Dd)$, where n is the number of taxa, k is the number of gene trees, D is the number of unique speciation-driven tripartitions, d is the number of dummy taxa, and α is the number of refinement iterations (consistently observed to be ≤ 5 in our studies). It has successfully analyzed 500 taxa datasets with high duplication rate (more than 2,000 leaves) within 20 hours and 16GB of memory. Details of time and memory complexity of wQFM-GDL-T and wQFM-GDL-Q are described in Supplementary Section 5.

3 Experimental results and discussions

We test the performance of wQFM-GDL-T and wQFM-GDL-Q on existing and our new simulated datasets and biological datasets against the leading methods: ASTRAL-Pro3 [42], SpeciesRax [28] and DupLoss-2 [31].

3.1 Datasets

Simulated Datasets We used two benchmark datasets simulted and used in prior studies: S25 from the ASTRAL-Pro study [43], which contains 25 species under diverse duplication, loss, and ILS conditions, and S100, simulated by Molloy and Warnow [27] based on a real fungal dataset. We note that the number of model conditions for large datasets was limited in ASTRAL-Pro study. Therefore, to extensively evaluate the performance under GDL on large datasets, we generated two new large-scale simulated datasets comprising 200 and 500 taxa, which we call SIM200 and SIM500, respectively.

For both SIM200 and SIM500, species trees and gene family trees were simulated using SimPhy [25] under varying duplication rates, loss rates, and ILS levels. Sequence alignments were generated using AliSim [21], and maximum likelihood gene trees were estimated with FastTree [32] under the GTR+Gamma model. Duplication rates and haploid effective population sizes were determined empirically. Detailed parameter settings and the full simulation pipeline is described in Supplementary Section 4. The datasets are publicly available at `https://doi.org/10.5281/zenodo.18605522`.

Biological datasets We analyzed three important empirical datasets with multicopy gene trees, Plants83 [39], Vertebrates188 [28], and Archaea364 [9], which cover a wide range of organisms. The vertebrates dataset is the largest dataset we analyzed, comprising 188 species and more than 30,000 gene families. The Archaea364 dataset is particularly challenging due to the presence of host–symbiont gene transfers in widely conserved marker genes.

3.2 Results on simulated datasets

Results for S25 dataset In S25, both versions of wQFM-GDL and ASTRAL-Pro3 substantially outperform SpeciesRax and DupLoss-2 across nearly all model conditions (Figure 3). We exclude DupLoss-2 for improved visual representation of the results because of their very high error rate but results including DupLoss-2, are provided in Supplementary Section 6.3. While wQFM-GDL and ASTRAL-Pro3 perform similarly, wQFM-GDL-Q demonstrates better performance. It outperforms ASTRAL-Pro3 in 35 out of 52 model conditions, with statistically significant differences in 7 of these cases, and ties occurring in 8 conditions. The S100 results are presented in Supplementary Section 6.2.

Results for SIM200 and SIM500 datasets Figure 4 presents the performance of the methods on SIM200 and SIM500 datasets (DupLoss-2 excluded because of high error rate). wQFM-GDL-Q could not be executed on these datasets due to its computationally intensive quartet generation. We could not finish running SpeciesRax on a single replicate with 500 or 1000 genes for some model conditions even after running for more than two days. Therefore, within our resource limits (64 GB of CPU RAM and 48 hours of computation per replicate), we could only analyze SpeciesRax on 250-gene model conditions of SIM200.

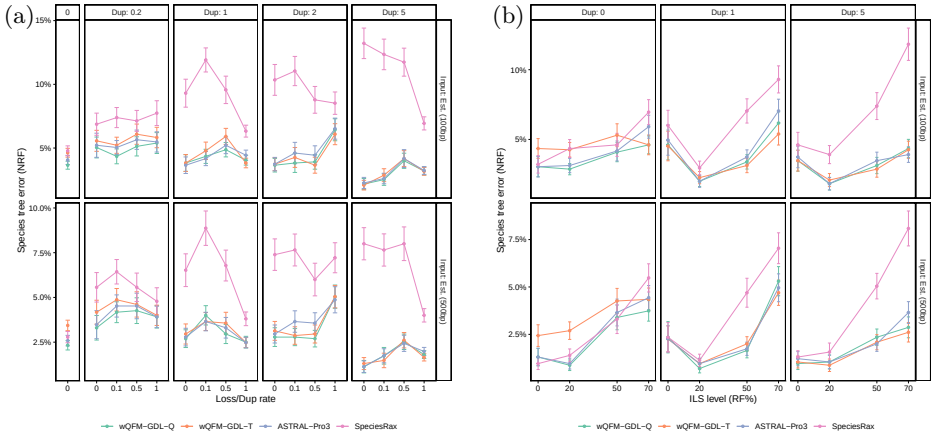

Fig. 3: Species tree error on the S25 dataset using 1,000 estimated gene trees from 100 and 500bp alignments. We show the average RF rates with standard error bars over 50 replicates. (a) Controlling duplication rate (box columns, labelled by mean number of copies per species minus one) and loss rate (labelled by ratio of loss and duplication rate). (b) Controlling duplication rate and ILS level (RF rate between true gene trees and species tree).

Quite remarkably, wQFM-GDL-T convincingly outperformed all the methods in all 72 out of 72 model conditions, differences being statistically significant ($p < 0.05$). wQFM-GDL-T delivers almost 19% and 23% reduction in error compared to ASTRAL-Pro3 in SIM200 and SIM500, respectively. As expected, the performance of all methods decreases with fewer gene trees and higher levels of ILS. Interestingly, both wQFM-GDL and ASTRAL-Pro3 show improved accuracy with higher duplication rates, consistent with observations reported in the ASTRAL-Pro3 study. The bigger performance gap between wQFM-GDL and ASTRAL-Pro3 on larger datasets indicates the superior performance of wQFM-GDL in larger and more complex scenarios.

3.3 Results on biological data

Plants83 Dataset We analyze 9,237 multicopy gene trees which were inferred in the original study by Wicket et. al., [39]. The tree estimated by wQFM-GDL (Figure 5) successfully recovers all major relationships and does not violate any known consensus among scientists. Following the most prominent recent methods, wQFM-GDL reconfirms the monophyly of Bryophytes (mosses, liverworts, and hornworts). Lycophytes are correctly recovered as the sister group of ferns and seed plants, and Gymnosperms are correctly recovered as the sister to flowering plants. It places Zygnematales (not Chara) as sister to all land plants and Amborella as sister to the rest of angiosperms. All of these relationships are well established.

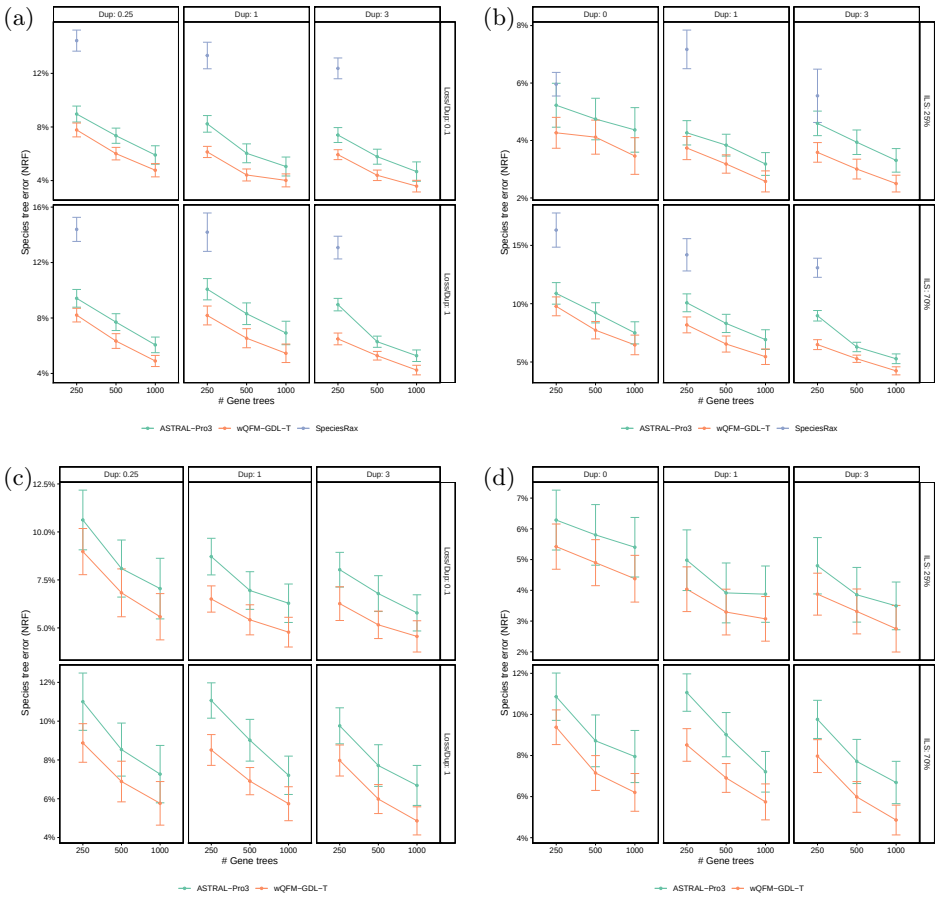

Fig. 4: Species tree error on the SIM200 (a-b) and SIM500 (c-d) dataset using estimated gene trees from 100 bp alignments. (a,c) Controlling duplication and loss rate. (b,d) Controlling duplication rate and ILS level. SpeciesRax only completed 250-gene model conditions of SIM200 within our resource limits.

However, there are three quite contentious areas. Alternative branching orders and quartet supports for those relationships are shown in Figure 5. Firstly, wQFM-GDL, ASTRAL-Pro3, and SpeciesRax all support the Gnepine hypothesis for Gymnosperms (1.0 localPP), placing Gnetales and Pinaceae as sisters. However, in the 1kp plant analysis [40], ASTRAL with single-copy nuclear gene trees and plastome-based supermatrix analysis strongly suggested otherwise. Secondly, for the relative position of Coleochaetales and Chara, wQFM-GDL places Chara as the closest relative of Zygnematophyceae and Embryophyta (land plants), whereas ASTRAL-Pro3 and SpeciesRax place Coleochaetales algae. Both of the cases have almost exactly the same quartet support (36%).

Finally, in the ASTRAL-Pro3 tree, Rosmarinus and Ipomoea are placed in the same clade. In contrast, wQFM-GDL recovers Ipomoea as sister to Catharanthus and Allamanda.

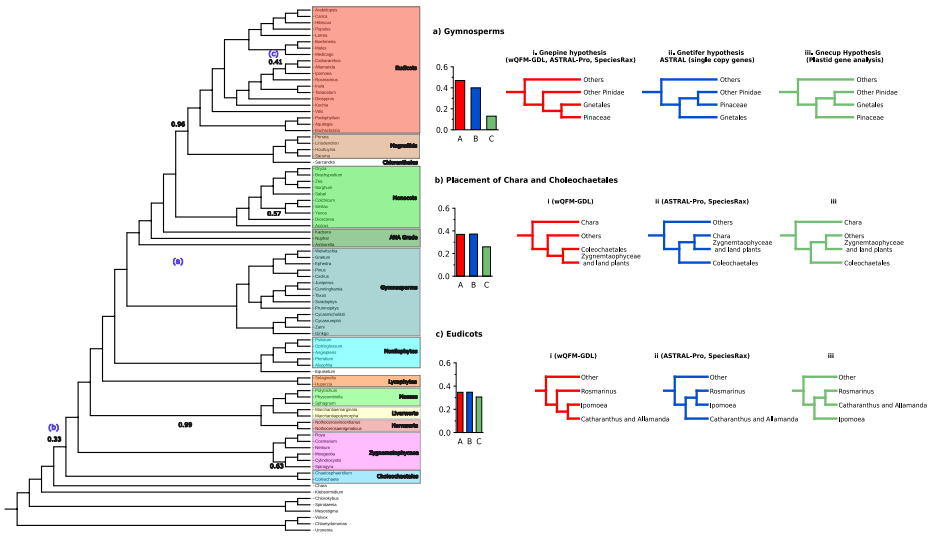

Fig. 5: The species tree inferred using wQFM-GDL for the Plants83 dataset and alternative branching orders for contentious relationships.

Vertebrates188 and Archaea364 Dataset wQFM-GDL and ASTRAL-Pro3 recover nearly identical trees both of these datasets. The Vertebrates species tree inferred by wQFM-GDL is consistent with the NCBI taxonomy tree that is available on the Ensemble Compara database [10]except for only five branches, which remain contentious across the methods. For Archaea364 dataset, our analyses agree with the main findings of Dombrowski et. al., [9], done by CA-ML analysis and recovered DPANN archaea (including Undinarchaeota) as a monophyletic group. Detailed descriptions, along with the full inferred trees and key subtrees recovered by the methods, are provided in Supplementary Section 6.4.

4 Conclusion

In this study, we proposed wQFM-GDL, a scalable and accurate quartet-based summary method, which accounts for both ILS and GDL. We have leveraged the concept of SQs within the QFM framework using appropriate algorithmic re-engineering and used a locus-aware normalization for improved accuracy. It demonstrates excellent performance and outperforms leading methods across most model conditions, particularly on large datasets. wQFM-GDL is capable of analyzing multicopy datasets containing thousands of taxa and gene trees with high duplication and loss rate within hours. Overall, these results position wQFM-GDL as a competitive and practical alternative to existing species tree inference methods for genome-scale data in the presence of GDL.

References

1. Avni, E., Cohen, R., Snir, S.: Weighted quartets phylogenetics. Systematic biology **64**(2), 233–242 (2015)
2. Bayzid, M.S., Mirarab, S., Warnow, T.: Inferring optimal species trees under gene duplication and loss. In: Proc. of Pacific Symposium on Biocomputing (PSB). vol. 18, pp. 250–261 (2013)
3. Bayzid, M.S., Warnow, T.: Gene tree parsimony for incomplete gene trees: addressing true biological loss. Algorithms for Molecular Biology **13**, 1 (2018)
4. Bayzid, M.S.: Inferring optimal species trees in the presence of gene duplication and loss: beyond rooted gene trees. Journal of Computational Biology **30**(2), 161–175 (2023)
5. Boussau, B., Szöllősi, G.J., Duret, L., Gouy, M., Tannier, E., Daubin, V.: Genome-scale coestimation of species and gene trees. Genome Research **23**(2), 323–330 (2013)
6. Chaudhary, R., Bansal, M.S., Wehe, A., Fernández-Baca, D., Eulenstein, O.: iGTP: a software package for large-scale gene tree parsimony analysis. BMC Bioinformatics pp. 574–574 (2010)
7. Chifman, J., Kubatko, L.: Quartet from snp data under the coalescent model. Bioinformatics **30**(23), 3317–3324 (2014)
8. De Oliveira Martins, L., Mallo, D., Posada, D.: A bayesian supertree model for genome-wide species tree reconstruction. Systematic biology **65**(3), 397–416 (2016)
9. Dombrowski, N., Williams, T.A., Sun, J., Woodcroft, B.J., Lee, J.H., Minh, B.Q., Rinke, C., Spang, A.: Undinarchaeota illuminate dpann phylogeny and the impact of gene transfer on archaeal evolution. Nature Communications **11**(1), 3939 (2020)
10. Federhen, S.: The ncbi taxonomy database. Nucleic acids research **40**(D1), D136–D143 (2012)
11. Fiduccia, C., Mattheyses, R.: A linear-time heuristic for improving network partitions. In: 19th Design Automation Conference. pp. 175–181 (June 1982). `https://doi.org/10.1109/DAC.1982.1585498`
12. Hallett, M.T., Lagergren, J.: New algorithms for the duplication-loss model. In: Proc RECOMB. pp. 138–146 (2000)
13. Han, Y., Molloy, E.K.: Improving quartet graph construction for scalable and accurate species tree estimation from gene trees. Genome Research **gr.277629.122.** (May 2023). `https://doi.org/10.1101/gr.277629.122`
14. Han, Y., Molloy, E.K.: Improved robustness to gene tree incompleteness, estimation errors, and systematic homology errors with weighted tree-qmc. Systematic Biology p. syaf009 (2025)
15. Islam, M., Sarker, K., Das, T., Reaz, R., Bayzid, M.S.: Stelar: A statistically consistent coalescent-based species tree estimation method by maximizing triplet consistency. BMC genomics **21**(1), 1–13 (2020)
16. Kubatko, L.S., Carstens, B.C., Knowles, L.L.: Stem: Species tree estimation using maximum likelihood for gene trees under coalescence. Bioinf **25**, 971–973 (2009)
17. Larget, B., Kotha, S.K., Dewey, C.N., Ané, C.: BUCKy: Gene tree/species tree reconciliation with the Bayesian concordance analysis. Bioinf **26**(22), 2910–2911 (2010)
18. Liu, L., Yu, L.: Estimating species trees from unrooted gene trees. Systematic Biology **60**(5), 661–667 (2011). `https://doi.org/10.1093/sysbio/syr027`
19. Liu, L., Yu, L., Edwards, S.V.: A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evolutinary Biology **10:302** (2010)

20. Liu, L., Yu, L., Pearl, D.K., Edwards, S.V.: Estimating species phylogenies using coalescence times among sequences. Systematic biology **58**(5), 468–477 (2009)
21. Ly-Trong, N., Naser-Khdour, S., Lanfear, R., Minh, B.Q.: Alisim: a fast and versatile phylogenetic sequence simulator for the genomic era. Molecular biology and evolution **39**(5), msac092 (2022)
22. Maddison, W.P.: Gene trees in species trees. Systematic Biology **46**, 523–536 (1997)
23. Mahbub, M., Wahab, Z., Reaz, R., Rahman, M.S., Bayzid, M.S.: wQFM: highly accurate genome-scale species tree estimation from weighted quartets. Bioinformatics **37**(21), 3734–3743 (Nov 2021). `https://doi.org/10.1093/bioinformatics/btab428`
24. Mahbub, M., Wahab, Z., Reaz, R., Rahman, M.S., Bayzid, M.S.: wqfm: highly accurate genome-scale species tree estimation from weighted quartets. Bioinformatics **37**(21), 3734–3743 (2021)
25. Mallo, D., de Oliveira Martins, L., Posada, D.: Simphy: phylogenomic simulation of gene, locus, and species trees. Systematic biology **65**(2), 334–344 (2016)
26. Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., Warnow, T.: ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics **30**(17), i541–i548 (2014)
27. Molloy, E.K., Warnow, T.: FastMulRFS: fast and accurate species tree estimation under generic gene duplication and loss models. Bioinformatics **36**($Supplement_1$), i57–i65 (Jul 2020). `https://doi.org/10.1093/bioinformatics/btaa444`
28. Morel, B., Schade, P., Lutteropp, S., Williams, T.A., Szöllősi, G.J., Stamatakis, A.: Speciesrax: a tool for maximum likelihood species tree inference from gene family trees under duplication, transfer, and loss. Molecular biology and evolution **39**(2), msab365 (2022)
29. Mossel, E., Roch, S.: Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. IEEE Comp Biol Bioinform **7**(1), 166–171 (2011)
30. Ohno, S.: Evolution by gene duplication. Springer Science & Business Media (2013)
31. Parsons, R.A., Bansal, M.S.: Duploss-2: Improved phylogenomic species tree inference under gene duplication and loss. Systematic Biology p. syaf073 (2025)
32. Price, M.N., Dehal, P.S., Arkin, A.P.: Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. Molecular biology and evolution **26**(7), 1641–1650 (2009)
33. Rafi, A., Rumi, A.M.S., Hakim, S.A., Sohaib, Tahmid, M.T., Momin, R.J.I., Zaman, T.A., Reaz, R., Bayzid, M.S.: wQFM-TREE: highly accurate and scalable quartet-based species tree inference from gene trees. Bioinformatics Advances **5**(1), vbaf053 (03 2025). `https://doi.org/10.1093/bioadv/vbaf053`, `https://doi.org/10.1093/bioadv/vbaf053`
34. Rasmussen, M.D., Kellis, M.: Unified modeling of gene duplication, loss, and coalescence using a locus tree. Genome Research **22**(4), 755 (Apr 2012). `https://doi.org/10.1101/gr.123901.111`
35. Reaz, R., Bayzid, M.S., Rahman, M.S.: Accurate phylogenetic tree reconstruction from quartets: A heuristic approach. PLoS One **9**(8), e104008 (2014)
36. Snir, S., Rao, S.: Quartet maxcut: a fast algorithm for amalgamating quartet trees. Molecular phylogenetics and evolution **62**(1), 1–8 (2012)
37. Swofford, D.: PAUP*: Phylogenetic analysis using parsimony (* and other methods). Ver. 4. Sinauer Associates, Sunderland, Massachusetts (2002)
38. Wehe, A., Bansal, M.S., Burleigh, J.G., Eulenstein, O.: Duptree: A program for large-scale phylogenetic analyses using gene tree parsimony. American Journal of Botany **24**(13), 1540–1541 (2008)

39. Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M.S., Burleigh, J.G., Gitzendanner, M.A., et al.: Phylotranscriptomic analysis of the origin and early diversification of land plants. Proceedings of the National Academy of Sciences **111**(45), E4859–E4868 (2014)
40. Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M.S., Burleigh, J.G., Gitzendanner, M.A., et al.: Phylotranscriptomic analysis of the origin and early diversification of land plants. Proceedings of the National Academy of Sciences **111**(45), E4859–E4868 (2014)
41. Willson, J., Roddur, M.S., Liu, B., Zaharias, P., Warnow, T.: DISCO: Species Tree Inference using Multicopy Gene Family Tree Decomposition. Systematic Biology **71**(3), 610–629 (08 2021). https://doi.org/10.1093/sysbio/syab070, https://doi.org/10.1093/sysbio/syab070
42. Zhang, C., Nielsen, R., Mirarab, S.: Aster: a package for large-scale phylogenomic reconstructions. Molecular Biology and Evolution **42**(8), msaf172 (2025)
43. Zhang, C., Scornavacca, C., Molloy, E.K., Mirarab, S.: ASTRAL-Pro: Quartet-Based Species-Tree Inference despite Paralogy. Molecular Biology and Evolution **37**(11), 3292–3307 (Nov 2020). https://doi.org/10.1093/molbev/msaa139