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

Abdur Rafi[1,†], Ahmed Mahir Sultan Rumi[1,†], Sheikh Azizul Hakim[1], and Md. Shamsuzzoha Bayzid[1,*]

[1]Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology
Dhaka-1205, Bangladesh
[†]These authors contributed equally to this work
[*]Corresponding author: shams_bayzid@cse.buet.ac.bd

April 2025

Abstract

Species tree estimation from genome-wide data has transformed evolutionary studies, particularly in the presence of gene tree discordance. Gene trees often differ from species trees due to factors like incomplete lineage sorting (ILS) and gene duplication and loss (GDL). Quartet-based species tree estimation methods have gained substantial popularity for their accuracy and statistical guarantee. However, most of these methods (e.g., ASTRAL, TREE-QMC, wQFM/wQFM-TREE) rely on single-copy gene trees and models ILS and not GDL, limiting their applicability to large genomic datasets. ASTRAL-Pro, a recent advancement, has refined quartet similarity measures to incorporate both orthology and paralogy, improving species tree inference under GDL. Among other quartet-based methods, wQFM-DISCO converts multicopy gene family trees into single-copy gene trees using DISCO and applies the wQFM algorithm on the single-copy trees. However, ASTRAL-Pro has remained the only quartet-based summary method to explicitly model gene duplication and loss. In this study, we propose wQFM-GDL, extending algorithms wQFM and wQFM-TREE which supports gene family trees and models gene duplication and loss, leveraging the concept of speciation-driven quartets introduced in ASTRAL-Pro. Our algorithm consistently outperforms ASTRAL-Pro3 across most model conditions, offering a promising alternative for estimating species trees in the presence of GDL.

1 Introduction

Species tree inference is a fundamental problem in evolutionary biology, complicated by the discordance between gene trees and species trees. This discordance arises due to several biological processes, including incomplete lineage sorting (ILS), gene duplication and loss (DupLoss), horizontal gene transfer, and hybridization (1). Although many approaches have been proposed for species tree inference, the most scalable and widely used approach so far follows a two-step process. At first, gene trees are inferred independently from sequence data. Then they are aggregated to infer the species tree using summary methods based on a model of gene tree discordance.

Summary methods construct species trees by first inferring gene trees from different loci and then integrating these gene trees into a single species tree. Many of these methods are statistically consistent under the multispecies coalescent (MSC) model (2; 3; 4; 5; 6; 7; 8; 9; 10; 11; 12; 13). A key principle in the design of some statistically consistent approaches is the use of fundamental phylogenetic units, such as triplets in rooted analyses and quartets in unrooted analyses (3; 11; 14). ASTRAL, the most popular summary method, tries to solve the Maximum Quartet Support Species Tree (MQSST) problem (11). It takes a set of gene trees as input and seeks to find a species tree so that the number of induced quartets in the gene trees that are consistent with the species tree is maximized. While ASTRAL uses an exact dynamic programming approach, alternative heuristics such as QMC (Quartet MaxCut), QFM (Quartet Fiduccia–Mattheyses), wQMC, and wQFM address the problem by first inferring individual quartets and then constructing a species tree through a divide-and-conquer strategy (14; 15; 16; 17). Although wQFM

has consistently outperformed ASTRAL in multiple studies (17; 18), its limited scalability has hindered widespread adoption. In a previous study, we introduced wQFM-TREE (19) to overcome the computational bottleneck of explicit quartet enumeration in the original wQFM algorithm. By directly traversing gene trees, wQFM-TREE significantly enhanced computational efficiency while preserving the core optimization framework of wQFM and maintaining its superior performance.

Gene Duplication and Loss (GDL) events introduce paralogs — genes that originate from a common ancestor through duplication events—in addition to orthologs, which arise from speciation events (20; 21). As a result, gene families may contain multiple genes with the same species labels, leading to multi-copy gene trees. However, most summary methods, including ASTRAL, TREE-QMC, wQFM, wQFM-TREE, are designed to work with gene trees containing at most one copy of each gene per species, making them unsuitable for use with GDL. Several methods have been developed to estimate species trees from multi-copy gene trees while addressing gene duplications and losses, without the need for orthology determination. These methods include PHYLDOG (22), iGTP (23), DupTree (24), DynaDup (25; 26; 27), MulRF (28), FastMulRFS (29), Guenomu (30), ASTRAL-Pro (31), and SpeciesRax (32). Most of these methods (23; 24; 25; 26) are parsimony-based, estimating a species tree by minimizing a specific distance between the species tree and the given gene trees, such as the total number of duplication and loss events. ASTRAL-Pro is the only quartet-based species tree inference method that explicitly models gene duplication and loss. It takes a set of multi-copy gene family trees and computes a species tree that maximizes similarity to the input gene trees, using a new quartet similarity measure that accounts for both orthology and paralogy (31). Alternative quartet-based methods like TREE-QMC can handle multi-labeled trees but do not incorporate gene duplication and loss modelling. As both wQFM and wQFM-TREE has shown greater accuracy compared to ASTRAL in the past (albeit with higher running time), in this study we propose wQFM-GDL, which builds on the QFM framework by modelling gene duplication and loss. wQFM-GDL has two versions named wQFM-GDL-Q and wQFM-GDL-T, which work on quartets and gene trees, respectively. We make the following contributions in this study.

- We propose wQFM-GDL-T, an extension of wQFM-TREE that incorporates both paralogy and orthology under the gene duplication and loss (GDL) model of (31). We extend the concept of satisfied and violated quartets of the wQFM framework for this task and design graph algorithmic techniques to calculate them efficiently directly from the gene trees.
- We observe that for algorithms that need explicit quartet generation such as wQFM and wQMC, it is sufficient to generate the quartet set with appropriate quartets following the gene duplication and loss model. We devise an algorithm to generate the quartet set for multicopy gene family trees which can be fed to wQFM for analysis of multicopy dataset. We name this version wQFM-GDL-Q. We observe that, for smaller datasets (taxa less than 100) where explicitly generating quartets is feasible, wQFM-GDL-Q outperforms both wQFM-GDL-T and ASTRAL-Pro3.
- We simulate two large datasets of 200 and 500 taxa to test the algorithms extensively with varying duplication and loss rates, ILS rates, number of taxa and genes. wQFM-GDL demonstrate superior performance in most of the model conditions. Interestingly, the difference becomes more prominent with increasing number of taxa.

2 Materials and Methods

We begin with a brief overview of the algorithm. Then, we describe the major steps of the algorithm in detail.

2.1 Overview of wQFM and wQFM-TREE

wQFM (17) and its latest version, wQFM-TREE (19), are both quartet-based summary methods designed to construct a species tree with maximum quartet support from a given set of gene trees over a specified taxa set. wQFM is a divide-and-conquer algorithm where each divide step of the algorithm splits the set of taxa into two disjoint subsets (a bipartition of the taxa set, each representing a subproblem). A unique "dummy taxon" is added to both of the subsets so that the solutions to the subproblems can be combined within a divide-and-conquer framework. The technique to find a bipartition begins with an initial bipartition, followed by a heuristic iterative strategy to improve it. Within this iterative process, each iteration begins with the bipartition from the previous iteration and tries to improve it by moving taxa from one partition to another. The iterative improvement strategy is described in detail in (14).

wQFM needs to generate all the quartets induced by the input gene trees which limits its ability to analyze large datasets. wQFM-TREE solves this major bottleneck by avoiding the quartet generation phase and performing all the steps of the algorithm directly from the input gene trees which enable it to analyze datasets with thousands of taxa

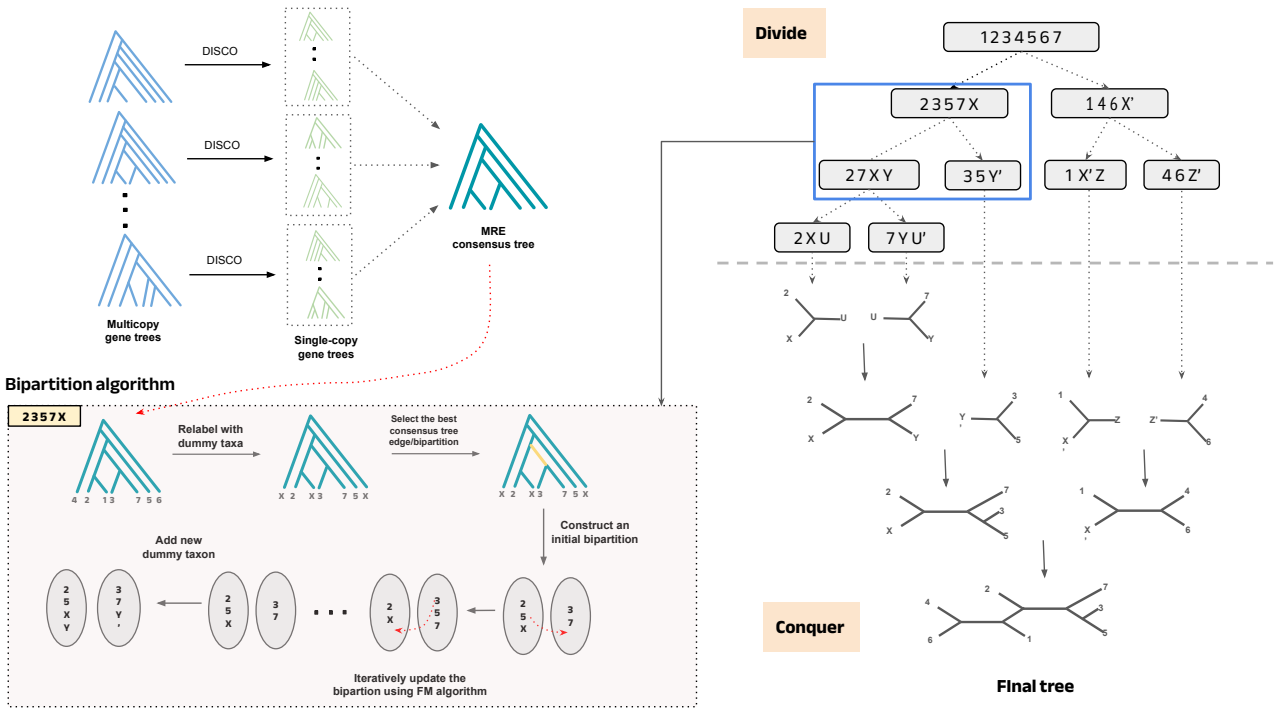


Figure 1: **Overview of our proposed method.** QFM framework originally follows a divide-and-conquer framework. We update the bipartition algorithm in the divide phase to efficiently handle gene duplication and loss. An MRE consensus tree is created from the multicopy gene trees. The initial bipartition is generated using our proposed heuristic from the consensus tree, which is then updated iteratively through the FM algorithm using a duplication-aware quartet score.

and gene trees within hours. In particular, wQFM-TREE employs the following two key strategies: i) using a gene tree consensus-based method to compute an initial bipartition for each divide step, and ii) utilizing combinatorial and graph theoretic techniques to compute the scores of candidate bipartitions directly from gene trees, thereby eliminating the need for generating and computing the weights of the induced quartets required in the original wQFM.

2.2 Normalization scheme and dummy taxa tree

Normalization is a critical component of both the wQFM and wQFM-TREE algorithms, and we revisit this scheme since it plays an equally essential role in wQFM-GDL. The need for normalization arises from the introduction of dummy taxa within subproblems during the divide phase. When the taxon set is partitioned into two subsets, a dummy taxon is added to each subset so that the solutions to the two resulting subproblems can later be combined in the conquer phase. Each dummy taxon acts as a representative for the species belonging to the sister subproblem (Figure 2). Consequently, every subproblem except the root contains such dummy taxa, leading to quartets that include dummy taxa. These quartets may receive disproportionately higher weights compared to quartets with real taxa because a dummy taxon implicitly represents multiple taxa from the opposite partition. Therefore, normalization is required to ensure that the presence of dummy taxa does not distort the computation of satisfied and violated quartets, which are fundamental to the QFM framework.

In wQFM, normalization is relatively straightforward because the full quartet set is available. As detailed in (17), for each subproblem we construct the subset of quartets restricted to the taxon set of that subproblem. Quartets involving dummy taxa are then down-weighted according to the number of species represented by the corresponding dummy taxon. This adjustment substantially improves the accuracy of wQFM by mitigating the inflation of quartet weights caused by dummy taxa.

For wQFM-TREE, normalization is more challenging because the explicit quartet set is not available. TREE-QMC introduced an effective normalization strategy in (33), and wQFM-TREE adopts a similar approach. In any subproblem containing one or more dummy taxa, there exist two types of quartets: those composed entirely of

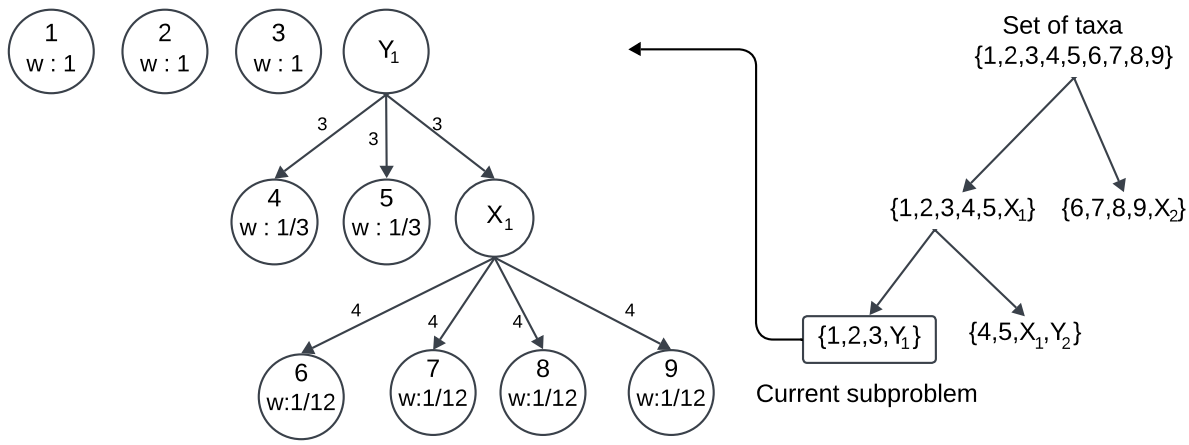


Figure 2: Tree structure of a dummy taxon, and the assignment of weights to real taxa under this dummy taxon. Here Y_1 is a dummy taxon that represents $\{4, 5, X_1\}$, where X_1 is another dummy taxon representing $\{6, 7, 8, 9\}$. The entire tree structure (root labeled by Y_1) is the representation of dummy taxon Y_1. The real taxa under Y_1 appear as leaves in the tree structure. Edge weights for each edge are shown. The weight of a leaf is the multiplicative inverse of the product of the edge weights from the root to the leaf. The figure is taken from (19)

real taxa and those involving one or more dummy taxa. When computing the counts (or weights) of satisfied and violated quartets directly from the tree, the dummy taxa are not present in the underlying gene tree; instead, real taxa belonging to the sister partition are relabeled as the corresponding dummy taxon. Without normalization, this relabeling artificially inflates the weights of quartets that include dummy taxa as there can be a larger number of quartets with taxa relabled by the same dummy taxon. However, for a dummy taxon in a subproblem, it is equivalent and should have the same importance as other real taxa in the subproblem.

To mitigate this effect, we assign asymmetric weights to dummy and real taxa and define the weight of a quartet as the product of the weights of its constituent taxa. Formally, we proceed as follows.

In our divide step, we create a bipartition (A, B) of the taxa set and create two new subproblems, one with taxa set $A \cup \{X_1\}$ and the other with taxa set $B \cup \{X_2\}$, where X_1 and X_2 are dummy taxa. X_1 essentially represents the taxa in B, and X_2 represents the taxa in A. We model X_1 and X_2 as rooted tree structures, where the children of the roots are the taxa in B and A, respectively. If A, B contain any dummy taxon, they will appear as subtrees in the tree structure of X_2, X_1 respectively, resulting in a recursive tree structure (Figure 2). Therefore, the tree representation of a dummy taxon X is a rooted tree with height > 1. The leaves of this tree are real taxa, and any internal node is a dummy taxon which was introduced in an earlier subproblem. We say that the leaves or real taxa in the tree structure of X are *under dummy taxon* X, and denote this set of real taxa under X as X_R.

Now we describe how we assign weight $w(a)$ to a taxon a that is used to compute the score of a candidate bipartition as described in Section 2.4.2. We assign unit weight to the real taxa that are not under any dummy taxa of the current subproblem. Note that the contribution of a dummy taxon to the score of a candidate bipartition should be equal to that of a real taxon as both can be considered equivalent in a subproblem. Therefore, for a dummy taxon X, we assign fractional weights to the real taxa in X_R in such a way that $w(X) = \sum_{a \in X_R} w(a) = 1$. We first assign weights to the edges of the tree structure corresponding to X. We then use these edge weights to calculate $w(a)$ $(a \in X_R)$. Refer to Figure 2 for an example. The weight of an edge (u, v) with u as the parent node, is defined as the number of children of u. Let p_a be the set of edges in the path from root to a. We define $w(a)$ as $(\prod_{(u,v) \in p_a} w_{(u,v)})^{-1}$. Therefore, for any real taxon a, $w(a)$ is defined as follows. As we can see in Figure 2, for taxa set $1, 2, 3, Y_1$ of a subproblem, the real taxa $1, 2, and 3$ are assigned weight 1. All the real taxa under the dummy taxon Y_1 are assigned fractional weights according to the subproblem decomposition, which sum to 1.

$$w(a) = \begin{cases} (\prod_{(u,v) \in p_a} w_{(u,v)})^{-1} & \text{if } a \text{ is under a dummy taxon} \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

2.3 Supporting gene family trees with duplication-aware quartet score

We extend both wQFM and wQFM-TREE for handling multicopy inputs. We utilize the concepts for modelling gene duplication and loss proposed in (31) for ASTRAL-Pro which remain appropriate for wQFM as both of the algorithms are quartet-based summary methods. The key observations of (31) are:

- Internal nodes of a gene tree correspond to either duplication or speciation events. Thus, we can tag nodes of the trees as speciation or duplication.
- Each quartet of leaves in the gene tree defines two anchor nodes, and the Least Common Ancestor (LCA) of the two anchors is defined as the anchor LCA. In a correctly tagged tree, a quartet has information about the speciation events only if it includes four distinct species and if the LCA of any three out of four leaves of the quartet is a speciation node. Thus, to define our measure of quartet similarity, we only include these speciation-driven quartets (SQs) and ignore the rest.
- All the SQs on the same four species that share the same anchor LCA share the same topology. Thus, once the topology of one of them is known, the others do not provide new information. These are SQs equivalent and counted as one equivalence class.

We had to make significant changes to the overall process and heuristics used in both wQFM and wQFM-TREE to incorporate these concepts. For wQFM, we propose an algorithm that constructs a quartet set containing only speciation-driven quartets, with a single representative from each equivalent class. This curated quartet set can then be used as input for algorithms like wQFM or wQMC that require quartet enumeration. No internal modifications to the wQFM algorithm are required.

In case of wQFM-TREE, for the initial bipartition needed for the underlying FM algorithm, we utilize an updated version of the consensus tree-based method used in wQFM-TREE for computing the initial bipartition at each divide step using DISCO and PAUP. Moreover, we design new combinatorial techniques to calculate only the speciation-driven quartets and count all equivalent quartets as one unit efficiently and directly from the gene trees without quartet enumeration.

2.4 Modifications for wQFM-GDL-T

2.4.1 Initial bipartition

In each divide step of our method, we initiate the process by generating an initial bipartition of the taxa set S, which would then be refined iteratively. We employ a consensus tree-based heuristic to construct an initial bipartition. At first, we transform the gene family trees into single-copy gene trees using DISCO (34). Then, we construct a greedy consensus tree using the Majority Rule Extended (MRE) method using the PAUP package (35), which supports incomplete gene trees as well.

Now, we can infer a candidate bipartition (S_A, S_B) of S from each bipartition (A, B) corresponding to each edge of the consensus tree as follows. S may contain both real and dummy taxa (see Figure 1). The real taxa in S are partitioned in a straightforward manner, according to (A, B) meaning that if a taxon $s \in A$, then $s \in S_A$. Otherwise, $s \in S_B$. We now describe how a dummy taxon X in S is assigned to a partition (S_A or S_B). The consensus tree does not contain any dummy taxon. Thus, we use the following heuristic to assign a dummy taxon X in S to a partition (S_A or S_B).

We know that artificial taxa represent the species on the sister subproblem, so that solutions to the two subproblems can be combined during the conquer phase. Let the taxa sets of the two subproblems created because of the current divide step are S and S' and the dummy taxon created at this step is X. X represents the species in S' and we relabel those species in the consensus tree with X. Now, there can be dummy taxa in S' because of the previous divide steps, which represent other real taxa. THus, the real taxa in S' will be directly relablled by X and other real taxa can be relabelled multiplle times in a recursive manner because of the dummy taxa of the upper subproblems. An intuitive method is to directy count the number of relabelled leaves in A and B and assign a partition according to that. However, maintaining the philosophy og non-uniform weighting and normalization discussed before, we assign weights to the relabelled taxa non-uniformly based on the subproblem decomposition, where the taxa that are closer to the current subproblem (such as real taxa in S') are upweighted and the taxa are far (multiple relabelling) are downweighted. The details of this process is discussed in Supplementary Section X.

NOw, let the set of relablled leaves be X_R. X_R can be partitioned to S_A and S_B according to (A, B) similar to how we partition the real taxa in S with respect to (A, B). Next, we assign X to the partition S_A if the sum of weights of the real taxa in X_R that belong to A is greater than the sum of weights of the real taxa in X_R that belong to B. Otherwise, we assign X to the partition S_B.

Thus, we find a bipartition (S_A, S_B) of S with respect to (A, B) in the consensus tree. Next, we score (S_A, S_B) using the scoring scheme described in Section 2.4.2. We find and score all such candidate initia; bipartitions of S corresponding to the edges in the consensus tree. We choose the bipartition of S with the highest score as the initial bipartition.

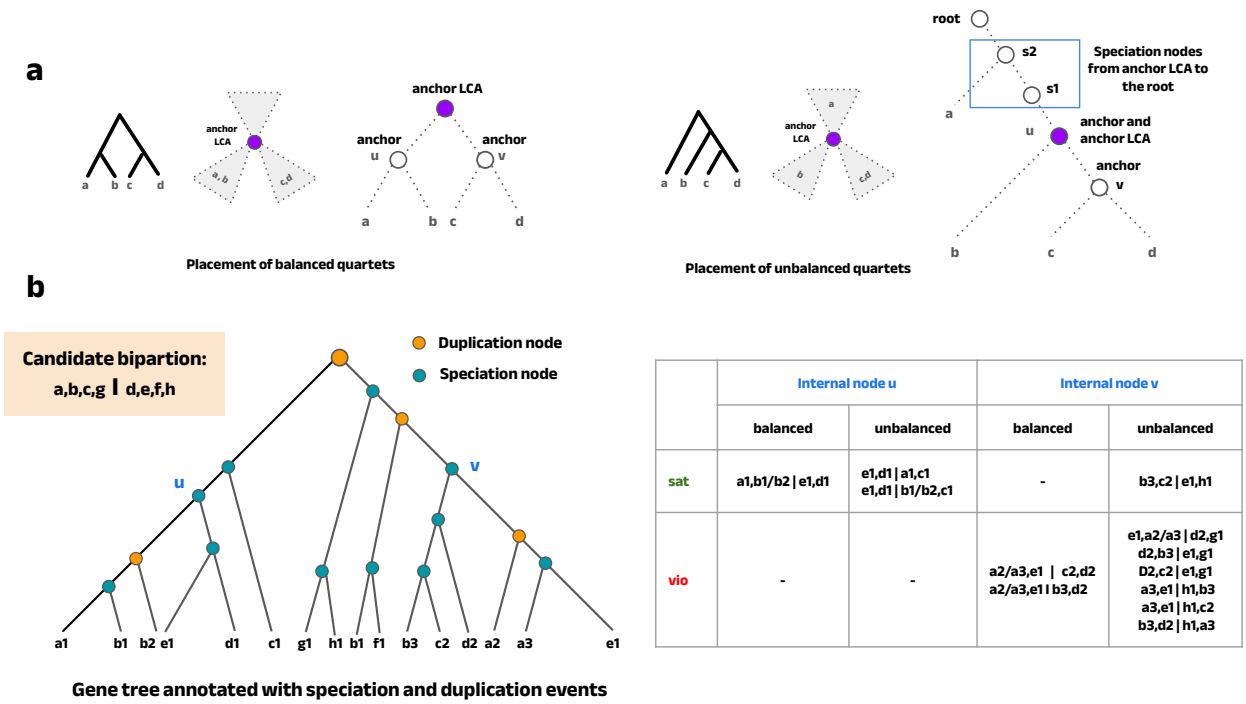
2.4.2 Scoring a candidate bipartition

Let $\mathcal{G}$ be a set of rooted and tagged gene trees on taxa set $\mathcal{X}$. Let $\mathcal{X}^{(g)}$ be the set of real taxa in a gene tree $g \in \mathcal{G}$. Let (A, B) be the given bipartition. Let R_A and D_A be the sets of real and dummy taxa in A, respectively. Let F_A be the set of all real taxa that are either in R_A or under a dummy taxon $X \in D_A$, i.e., $F_A = (\bigcup_{X \in D_A} X_R) \cup R_A$. We similarly define R_B, D_B, and F_B .

Given a bipartition (A, B), each quartet $ab|cd$, where no two taxa in $\{a, b, c, d\}$ are under the same dummy taxon of the current subproblem, can be classified as one of the following three types.

1. **Satisfied:** where either $\{a, b\} \subseteq F_A$ and $\{c, d\} \subseteq F_B$ or $\{a, b\} \subseteq F_B$ and $\{c, d\} \subseteq F_A$
2. **Deferred:** where there exists some $P \subseteq \{a, b, c, d\}$ such that $P \subseteq F_A$ or $P \subseteq F_B$, and $|P| \geq 3$.
3. **Violated:** all other quartets.

Let $S^{(g)}$ and $V^{(g)}$ denote the sets of satisfied and violated speciation-driven quartets, respectively, each containing one representative from every equivalence class in a gene tree $g \in \mathcal{G}$. Note that all quartets within the same equivalence class have equal weights. The score $\text{Score}(A, B, \mathcal{G})$ of a candidate bipartition (A, B) with respect to $\mathcal{G}$ is then defined as $\text{Score}(A, B, \mathcal{G}) = \sum_{g \in \mathcal{G}} \left(w(S^{(g)}) - w(V^{(g)})\right)$.



	Internal node u		Internal node v	
	balanced	unbalanced	balanced	unbalanced
sat	a1,b1/b2 | e1,d1	e1,d1 | a1,c1 e1,d1 | b1/b2,c1	-	b3,c2 | e1,h1
vio	-	-	a2/a3,e1 | c2,d2 a2/a3,e1 | b3,d2	e1,a2/a3 | d2,g1 d2,b3 | e1,g1 D2,c2 | e1,g1 a3,e1 | h1,b3 a3,e1 | h1,c2 b3,d2 | h1,a3

Figure 3: **Calculation of the score of a candidate bipartition.** **(a)** Placement of balanced and unbalanced quartets with respet to an anchor LCA. An important observation is that scores involved with balanced and unbalanced quartets can be computed seperately quite efficiently. **(b)** The weights of quartets belonging to different categories with a fixed speciation node (e.g., internal nodes u and v) as the anchor LCA are calculated seperately, and their sum gives the bipartition score. The quartets of different categories are shown for nodes u and v.

Computing $w(V^{(g)})$, the sum of weights of violated speciation-driven quartets in a gene tree g The main idea is that we can efficiently compute the sum of weights of all speciation-driven quartets with the same anchor

LCA. It is essential to ensure that only speciation-driven quartets are included in the computation and that each equivalence class contributes only a single representative quartet.

The subtree of a fully resolved gene tree, when restricted to a quartet, contains two degree-three nodes. We refer to these nodes as *anchors*. As illustrated in Figure 1a, rooted quartets can have two distinct structural forms. The first, known as the **balanced structure**, consists of two pairs of taxa that are grouped together first, with a root connecting these two pairs. The second, called the **unbalanced structure**, forms a nested grouping where taxa are added sequentially at progressively deeper levels. To compute $w(S^{(g)})$, we divide the calculation into two parts. The weights of the balanced and unbalanced speciation-driven quartets are computed separately and then summed to obtain the final value.

Let u be an internal speciation node in g, with two child branches i and j. We define $PA_i^{(g,u)}$ as the set of unordered pairs $\{a, b\}$, where a and b belong to branch i, are not grouped under the same dummy taxon, and satisfy $a, b \in F_A$. Note that each taxon pair is considered only once, even if multiple copies of a taxon exist in the branch. Similarly, we define $PA_j^{(g,u)}$ for branch j. The sets $PB_i^{(g,u)}$ and $PB_j^{(g,u)}$ are defined analogously.

Now, for the quartets with the balanced structure with anchor LCA u, one of their anchors will be in branch i and the other in branch j, as shown in Figure 1b. This allows us to efficiently compute $w(S_1^{(g)})$, the total weight of quartets with a balanced structure.

$$
\begin{aligned}
w(S_1^{(g,u)}) &= w(PA_i^{(g,u)})w(PB_j^{(g,u)}) + w(PA_j^{(g,u)})w(PB_i^{(g,u)}) \\
w(S_1^{(g)}) &= \sum_u w(S_1^{(g,u)})
\end{aligned}
$$

For the unbalanced ones, one anchor will be present in either branch i or j. The other anchor is u itself. We calculate the weights of all possible pairs that can eventually create an anchor in one of the child branches. Both taxa must belong to partition A or B, forming the set $PA_i^{(g,u)}$ or $PB_i^{(g,u)}$ for branch i, and $PA_j^{(g,u)}$ or $PB_j^{(g,u)}$ for branch j. This is because of the condition that for a speciation node, any three of the four leaves of a quartet must be a speciation node.

Now, let $PA_{i,k}^{(g,u)}$ be the set of unordered pairs $\{a, b\}$ where a belongs to branch i, b belongs to the parent branch k (and is a descendant of a speciation node from u to the root), both are not grouped under the same dummy taxon and satisfy $a, b \in F_A$. $PB_{i,k}^{(g,u)}$ is defined similarly.

From this, we can compute

$$
\begin{aligned}
w(S_2^{(g)}) &= w(PA_i^{(g,u)})w(PB_{j,k}^{(g,u)}) \\
&\quad + w(PA_j^{(g,u)})w(PB_{i,k}^{(g,u)}) \\
&\quad + w(PB_i^{(g,u)})w(PA_{j,k}^{(g,u)}) \\
&\quad + w(PB_j^{(g,u)})w(PA_{i,k}^{(g,u)}) \\
w(S_2^{(g)}) &= \sum_u w(S_2^{(g,u)})
\end{aligned}
$$

The expressions such as $w(PA_i^{(g,u)})$, $w(PA_{i,j}^{(g,u)})$ are calculated efficiently using graph algorithms like BFS and DFS in a single pre-order or post-order traversal.

Finally,

$$
w(S^{(g)}) = w(S_1^{(g)}) + w(S_2^{(g)})
$$

Computing $w(V^{(g)})$, the sum of weights of violated speciation-driven quartets in a gene tree g The calculation of $w(V^{(g)})$ follows a similar approach. We compute the total weight of the violated quartets separately for balanced and unbalanced structures and then sum the results, following the same approach as before.

Again, we assume an internal speciation node u in g. It contains two child branches i and j. For the balanced violated quartets, each branch will contain a pair (one from partition A, the other from B) and an associated anchor node.

In other words, the pair of taxa from a partition are on separate branches, with one in branch i and the other in branch j. Let $PA_{i,j}^{(g,u)}$ be the set of unordered pairs $\{a, b\}$ where a belongs to branch i, b belongs to branch j, both

are not grouped under the same dummy taxon and satisfy $a, b \in F_A$. $PB_{i,j}^{(g,u)}$ is similarly defined. Thus,

$$w(V_1^{(g,u)}) = w(PA_{i,j}^{(g,u)})w(PB_{i,j}^{(g,u)})$$
$$w(V_1^{(g)}) = \sum_u w(V_1^{(g,u)})$$

For the unbalanced ones, using similar reasoning, we get

$$\begin{aligned} w(V_2^{(g)}) = & \, w(PA_{i,k}^{(g,u)})w(PB_{i,j}^{(g,u)}) \\ & + w(PA_{i,j}^{(g,u)})w(PB_{i,k}^{(g,u)}) \\ & + w(PA_{j,k}^{(g,u)})w(PB_{j,i}^{(g,u)}) \\ & + w(PA_{j,i}^{(g,u)})w(PB_{j,k}^{(g,u)}) \\ w(V_2^{(g)}) = & \sum_u w(V_2^{(g,u)}) \end{aligned}$$

$$w(V^{(g)}) = w(V_1^{(g)}) + w(V_2^{(g)})$$

2.5 Modifications for wQFM-GDL-Q

wQFM takes as input a set of weighted quartets. Here, the weight is the count of the quartet structure in the gene trees. Our main task is to create an appropriate quartet set that only contains speciation-driven quartets and only one representative of an equivalent class. The main wQFM algorithm remain unchanged. Again, we use a similar approach of considering balanced and unbalanced quartet structures separately. However, in this case, we are not calculating the sum of weights of quartets; we have to enumerate them with proper weights. Moreover, there is no notion of satisfied or violated quartets as there is no bipartition to consider in the beginning.

For each speciation node, we generate the speciation-driven quartets with anchor LCA at that node. For the balanced quartets, we enumerate the possible pairs coming form each child branch. Of course, we keep only one copy of a taxa pair as they are equivalent. Then, each of the pairs can create a valid quartet with any pair from the other branch. We generate all possible quartets by joining the pairs.

For the unbalanced quartets, One pair will come from a child branch. We generate all possible pairs for both child branches. For the other pair, one taxa will be present in the other child branch and one taxa in the parent branch. As discussed, not all taxa in the parent branch are considered here (Figure 1b). We list all taxa present in the child branches and the parent branch and merge them to create pairs. The pairs are then merged appropriately to enumerate quartets.

An important point to note is that we just generate the appropriate quartet file for wQFM to work on. The main wQFM algorithm remains the same. The algorithm for generating the quartet file is shown in Algorithm 1.

3 Simulation Studies

We now describe the performance of both versions of wQFM-GDL on simulated datasets. We employ both datasets simulated in prior studies and our own simulated datasets. Numerous methods have been proposed in recent years for analyzing multicopy datasets; among them, we use ASTRAL-Pro3, which is arguably the most widely known approach for this task.

3.1 Simulated Datasets

We primarily use a simulated dataset, called S25 simulated and analyzed in (36). The data set was created by simulating true gene trees under the DLCoal model, simulating a sequence alignment from each true gene family trees, and finally estimating a gene tree from each gene alignment. A large number of model conditions of 25 species were generated, varying parameters suchy as the rate of duplication, the rate of loss, the ILS level, and the number of genes. The number of replicates in each model condition is 50. Each replicate contains 1000 gene trees. Moreover, model conditions with 100, 250, and 500 taxa were generated in the toughest model condition (high rate

Algorithm 1: Algorithm for quartet decomposition for wQFM for multicopy dataset

```
Input  : 𝒢, set of tagged and rooted multicopy gene family trees
Output: The set of speciation-driven quartets in 𝒢 with only one representative for each equivalent class
Q ← ∅ ;                                                        // quartet set initialization
foreach gene tree g ∈G do
    foreach internal speciation node s in g do
        l ← left child of s
        r ← right child of s

        qleft ← getRealTaxainSubtree(l) // list all taxa present in the subtree
        qright ← getRealTaxainSubtree(r)
        pairs_left ← getAllPairs(qleft) // generate all possible taxa pairs from the taxa list
        pairs_right ← getAllPairs(qright)
        Q1 ← generateQuartetsfromPairs(pairs_left,pairs_right)
        /* generate all valid quratets (pair of pairs) from the two list of pairs                    */

        par ← getRealTaxainParents(s)
        /* list all the taxa present in the other child subtree of all the speciation nodes on the path from s
           to root                                                                                    */
        crosspairs_left ← getAllCrossPairs(par,qleft) // generate all possible pairs with one taxa from
           each of the taxa lists
        crosspairs_right ← getAllCrossPairs(par,qright)
        Q2 ← generateQuartetsfromPairs(pairs_left,crosspairs_right)
         ∪generateQuartetsfromPairs(crosspairs_left,pairs_right)

        Q ← Q1 ∪ Q2
    end
end
return Q
```

Function getRealTaxainParents(s)

```
taxa_list ← ∅
current ← s
while current.parent ≠ null do
    if current.parent is speciation node then
        otherChild ← getSibling(current)
        taxa_list ← taxa_list∪ getRealTaxainSubtree(otherChild)
    end
    current ← current.parent
end
return taxa_list
```

of duplication and loss, high level of ILS) to test the scalability of the methods. We also used an existing data set that was simulated by Molloy and Warnow (29) based on a real fungal data set.

Based on our observation that wQFM-TREE outperforms ASTRAL-III under larger model conditions in the ASTRAL-III datasets, we generated additional datasets comprising 200 and 500 taxa to evaluate whether this performance trend extends to this task as well.

3.2 Simulation process

We used Simphy (37) to generate true gene family trees, AliSim (38) to simulate sequence alignment from the true trees. Finally, FastTree (39) was used to estimated maximum likelihood gene trees using GTR+Gamma model. We varied duplication and loss rates, ILS level and the number of gene trees for both of the datasets. The details of the parameters are presented in Table 1. We simulate all model conditions for both 200 and 500 taxa. The duplication rates and haploid effective population sizes are estimated empirically for the the mean number of copies and ILS level (RF rate) for various model conditons. The details of the whole simulation process, additional results and Simphy commands are presented in Supplementary Section 3.

Table 1: Simulation settings for 200 and 500 taxa datasets with varying parameters. For each of the duplication rates, we vary the loss rate, ILS level, and number of gene trees.

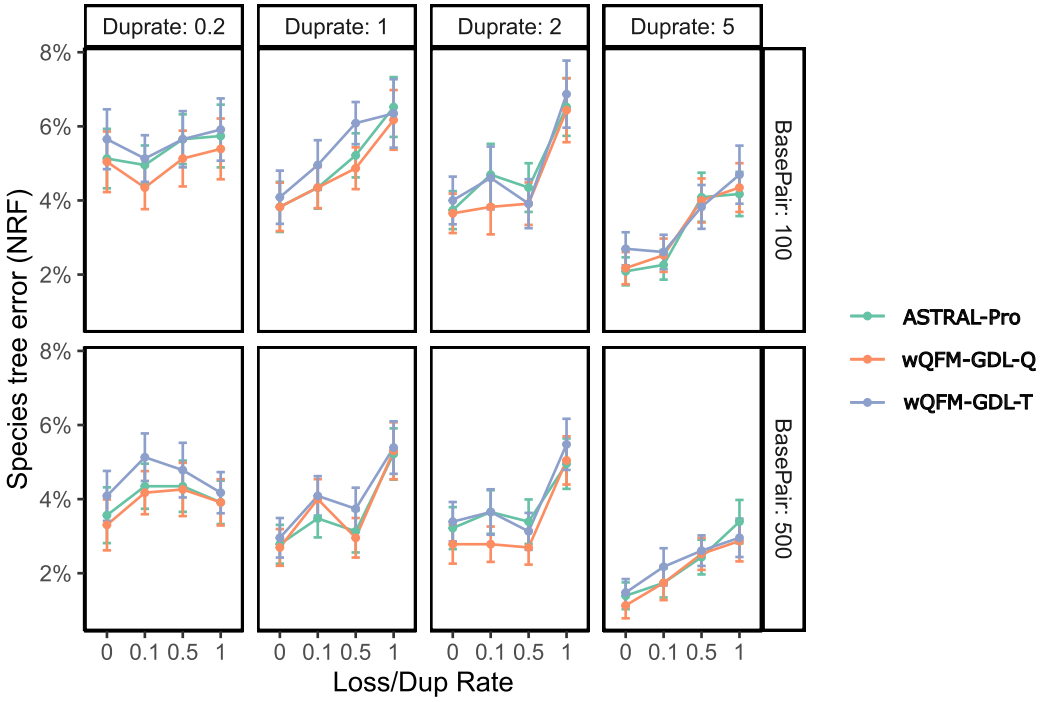
Condition	**Parameter Ranges**
No. of taxa	200,500
Varying no. of gene trees	250, 500 and 1000
Varying duplication rate	0.25, 1 and 3 (Mean number of extra copies of species) Duplication rates for S200: $\{0.8, 2.0, 4.1\} \times 10^{-10}$ Duplication rates for S500: $\{0.8, 2.1, 4.3\} \times 10^{-10}$
Varying loss rate	0.1 and 1 (fraction of duplication rates)
Varying ILS	25% and 70% (mean RF distance between true gene trees and the species tree in case of no duplication or loss) Haploid effective population size for S200: $\{0.5, 1.3, 2.9\} \times 10^8$ Haploid effective population size for S500: $\{0.5, 1.25, 2.7\} \times 10^8$
Sequence length	100bp

3.3 Results for S25 dataset

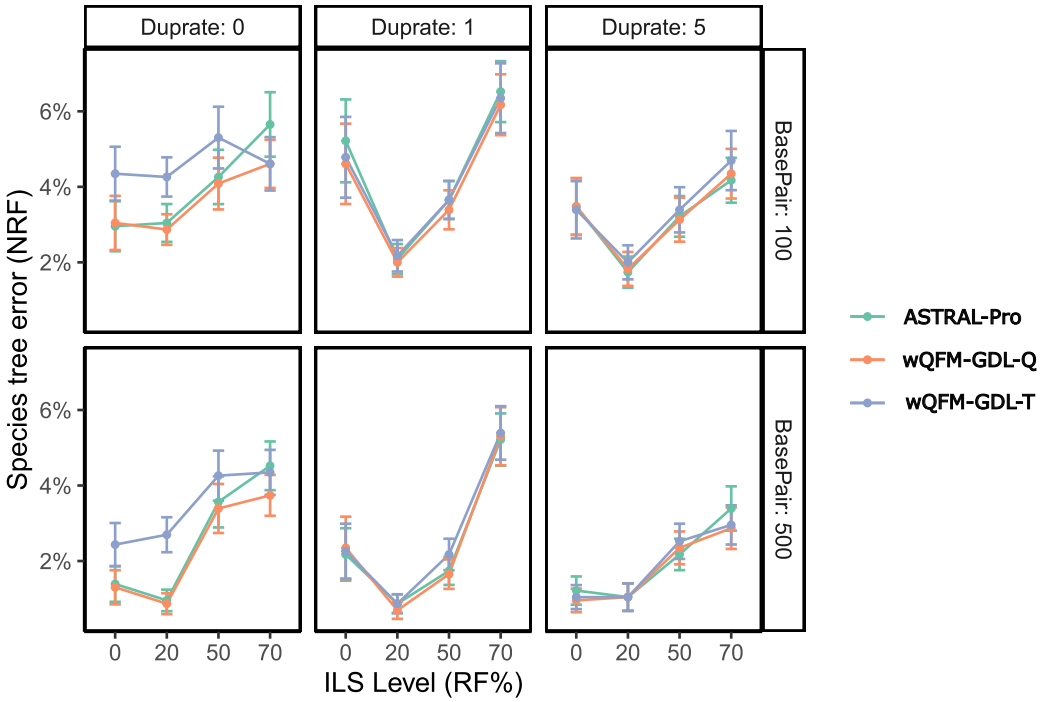
3.3.1 Controlling Duplication and Loss Rates

We present results for model conditions varying duplication and loss rates for estimated gene trees with 100 bp and 500 bp alignments. As shown in Figure 8, wQFM consistently outperforms both ASTRAL and wQFM-TREE in most model conditions. Notably, as the duplication rate increases, all methods improve in performance which is consistent with the findings of (31). This is likely because additional gene copies provide more information, similar to increasing the number of genes. Both wQFM and ASTRAL exhibit similar trends. Although error rates increase with higher loss rates, the increase remains under 2% for wQFM. As expected, longer alignments (500 bp) lead to lower error rates.

Overall, across 32 model conditions, wQFM outperforms ASTRAL in 21 cases (with most differences being statistically significant) and ties in 4 cases. For the 25-taxon dataset, wQFM-TREE shows slightly lower accuracy compared to wQFM and ASTRAL. While wQFM and wQFM-TREE follow the same core strategy, wQFM-TREE introduces heuristic approaches to make wQFM applicable for large datasets—such as constructing consensus tree from single-copy gene trees decomposed from multicopy gene families using DISCO. This may contribute to a slight decrease in accuracy.We recommend using wQFM for small datasets. However, as we will demonstrate, for large datasets, wQFM-TREE significantly outperforms ASTRAL-Pro.



(a) Controlling duplication and loss rate



(b) Controlling duplication and ILS rate

Figure 4: Species tree error on the S25 dataset using 1,000 estimated gene trees from 100 and 500bp alignments, under varying duplication, loss rates and ILS levels.

3.3.2 Controlling ILS level

Naturally, as the ILS level increases, the performance of all the methods decrase significantly as observed in Figure **??**. Again, wQFM outperforms other methods. However, the performance improvements remain relatively modest. Out of 24 model conditions, wQFM outperforms ASTRAL-Pro in 17 model conditions (with most differences being statistically significant) and ties in 1 condition. wQFM-TREE demonstrates slightly lower accuracy.

3.3.3 Controlling the number of species

Increasing taxa count from up to 500 in the most difficult model condition (high duplication and loss rate, high ILS) with estimated gene trees (for 100 bp) shows that both wQFM-TREE and ASTRAL are robust to a large number of species. An important point to note is that as wQFM requires quartet decomposition and it cannot analyze model conditions with more than 100 species. Thus, we only use wQFM-TREE for this experiment. wQFM-TREE shows significantly better performance in case of larger taxa. Specially for 500 taxon dataset, wQFM-TREE constructs species trees with an average RF rate of 6.8% compared to 9.5% of ASTRAL-Pro.

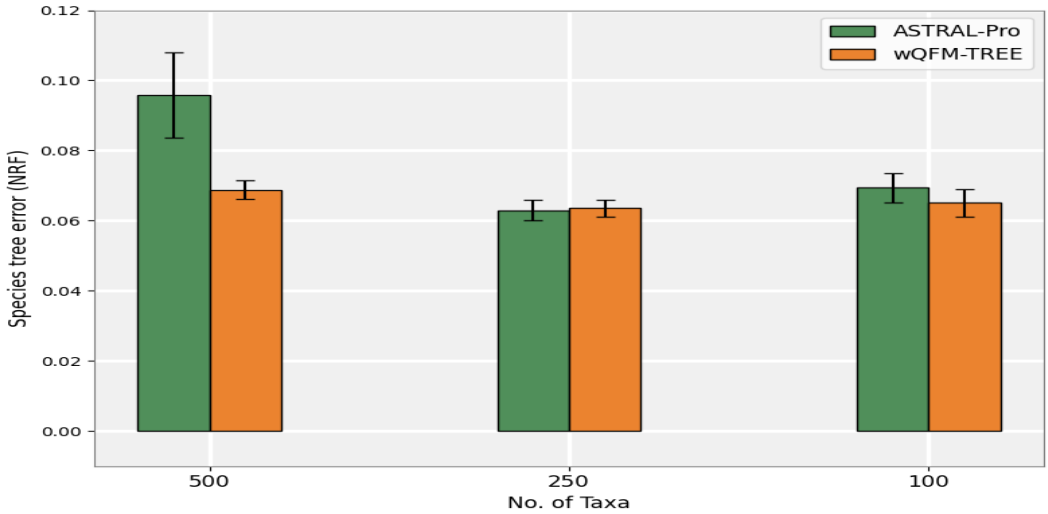


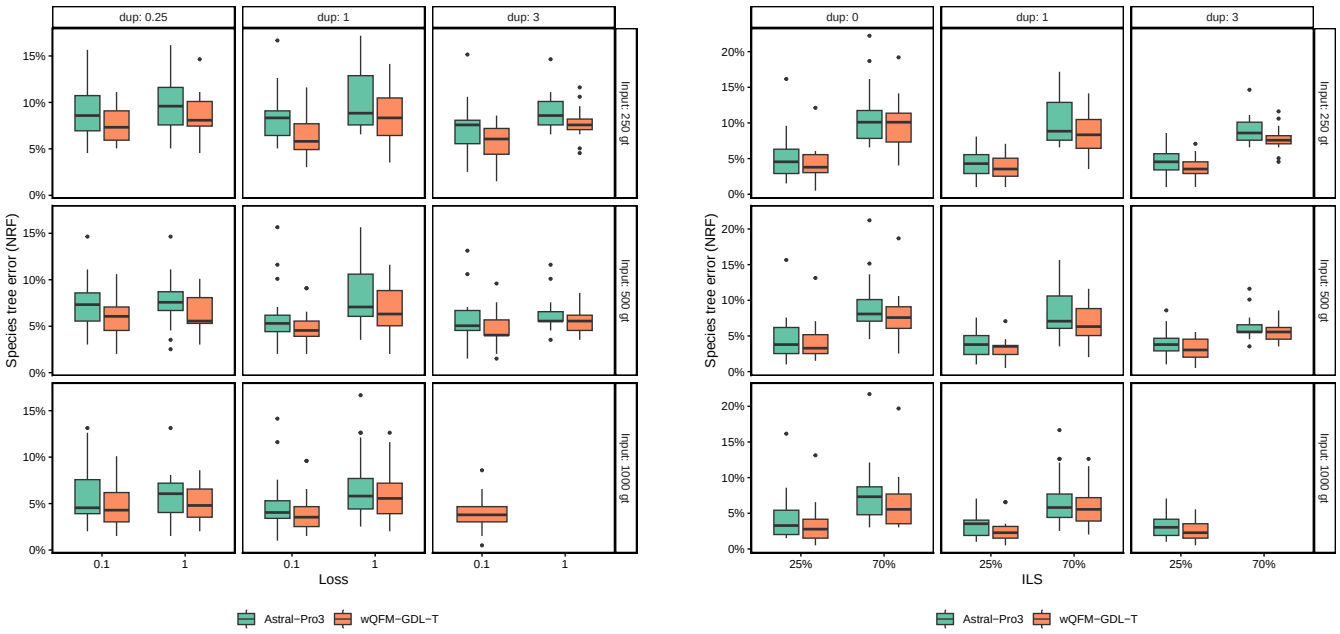
Figure 5: Comparison of methods on varying taxa with fixed ILS level and Loss/Dup rate.

3.4 Results for SIM200 and SIM500 datasets

For the large model conditions involving 200 and 500 taxa, we observe a pronounced difference in the performance between wQFM-GDL and ASTRAL-Pro. Notably, wQFM-GDL-Q could not be executed on these model conditions due to its computationally intensive quartet generation phase, which, based on our experiments, does not practically scale beyond 100 species. In both the SIM200 and SIM500 datasets, we systematically varied the duplication rates, ILS levels, and loss rates. Across all model conditions, wQFM-GDL consistently outperformed ASTRAL-Pro3, with the observed differences being statistically significant. As expected, the accuracy of both methods improved with an increased number of gene trees and reduced ILS levels. However, unlike the smaller S25 model conditions, where the impact of higher loss rates was less evident, we observe that higher loss rates in these larger models have a noticeably negative effect on performance. Finally, the performance gap between wQFM-GDL and ASTRAL-Pro3 in SIM500 is greater than in SIM200, suggesting that as the problem space becomes larger and more complex, the QFM framework is better able to search for the correct species tree and achieves superior performance more decisively.

3.5 Results for S100 dataset

Finally, we also test wQFM-GDL and ASTRAL-Pro on S100 dataset. Here, we can observe the accuracy of the methods with varying sequence length and Psize (effective haploid population size). Again, both versions of wQFM-GDL outperform ASTRAL-Pro in most of the model conditions, with wQFM-GDL-Q being the slightly more accurate version, although at 100 species, the quartet generation phase may take hours. We observe that, as the model



(a) Controlling duplication rate and number of genes

(b) Controlling ILS rate and number of genes.

Figure 6: Species tree error on the SIM200 dataset for 200 ingroup taxa using estimated gene trees from 100 bp alignments.

conditions become more challenging with small sequences, wQFM-GDL performs noticeably better than ASTRAL-Pro3, whereas with large sequences, the performances are quite similar. For both of the Psize values, we observe similar trends.

3.6 Biological analyses

Now, we describe the results of species tree inference on empirical data sets. We compare the species tree estimated by wQFM-TREE, ASTRAL-Pro and some other methods.

3.6.1 Biological Datasets

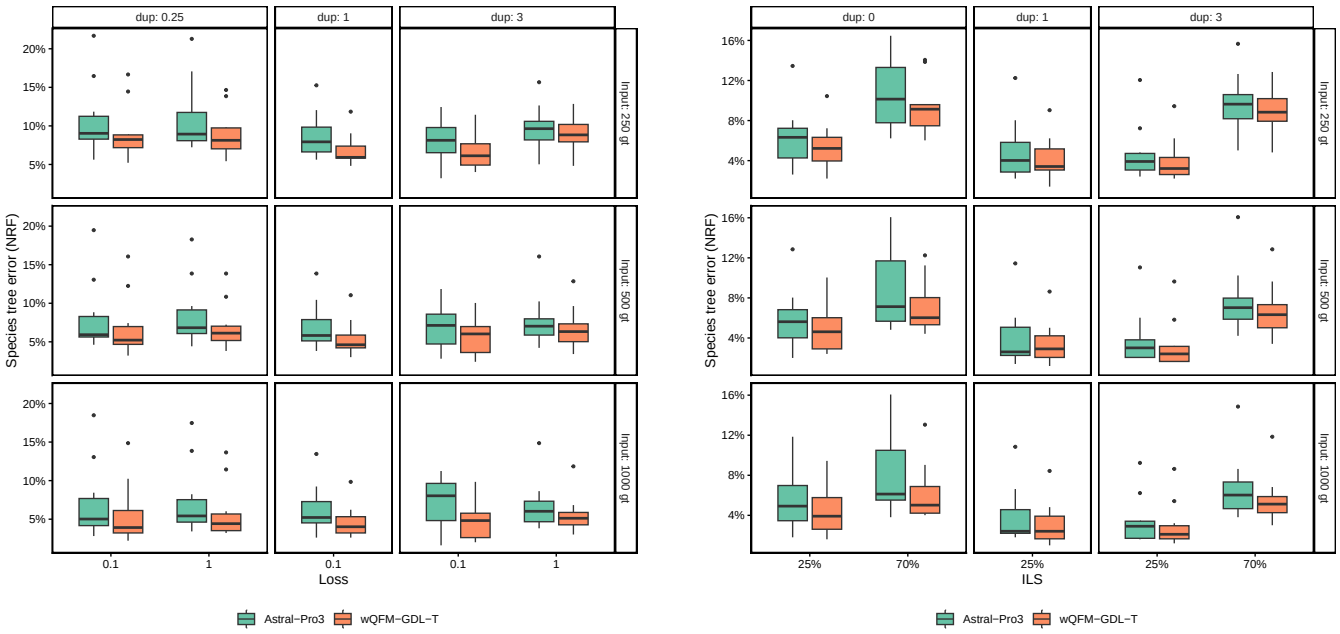
We analyzed some important and popular empirical multicopy datasets. The details of the datasets are shown in Table 2. The datasets cover a wide range of organisms including plants, vertebrates and archaea. The Plants83 dataset correspond to the Plant(1kp) dataset studied in (31). The vertebrates dataset is a largest dataset we analyzed comprising 188 species and more than 30,000 gene families. This dataset was collected from Morel et. al., (32) where they extracted the alignments comprising 199 species from the Ensembl Compara database and inferred the gene trees. The Archaea364 dataset was collected from the original study by DOmbrowski et. al. one important reason for the difficulty in resolving the archaeal tree was the presence of host–symbiont gene transfers in broadly conserved marker genes, in which members of the host-associated DPANN Archaea were sometimes grouped with their hosts.

Table 2: Description of the empirical datasets used in our study

Dataset	Taxa	Families	Source of Discordance	Data Source
Plants	83	9,237	ILS+GDL	(31)
Vertebrates	188	31,612	ILS+GDL	(32)
Archaea	364	150	ILS+GDL+HGT	(40)

3.6.2 Plants83 Dataset

We analyze 9,237 multicopy gene trees which were inferred in the original study by Wicket et. al., (41) but were left unused before the development of methods that can handle multicopy gene trees. The trees estimated by wQFM-



(a) Controlling duplication rate and number of genes

(b) Controlling ILS rate and number of genes.

Figure 7: Species tree error on the SIM500 dataset for 500 ingroup taxa using estimated gene trees from 100 bp alignments..

GDL-T (Figure 9) successfully recover all major relationships and do not violate any known consensus among scientists. The relationships among the Bryophytes (mosses, liverworts, and hornworts) have been contentious and resolved as a grade in some studies. However, following the most prominent recent methods, wQFM-GDL reconfirms the monophyly of Bryophytes (liverworts, mosses, and hornworts). Lycophytes are correctly recovered as the sister group of ferns and seed plants and Gymnosperms are correctly recovered as 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 fairly established now.

However, there are mainly three areas which are quite contentious. Alternative branching orders and quartet supports for those relationships are shown in Figure 9. Firstly, in case of Gymnosperms, wQFM-TREE supports the Gnepine hypothesis as well as ASTRAL-Pro with very high support (1.0 localPP). It is worth mentioning that in the transcriptome analysis using 1,124 species (all single-copy genes), ASTRAL supported the Gnetifer hypothesis but wQFM-TREE still recovered the Gnepine hypothesis. Secondly, for the relative position of Coleochaetales and Chara, wQFM-GDL places Chara as the closest relatives of Zygnematophyceae and Embryophyta (land plants), whereas ASTRAL-Pro places Coleochaetales algae. Both of the cases have almost exactly same quartet support (36%). Finally, In ASTRAL-Pro, Rosmarinus and Ipomoea are grouped together. However, wQFM-GDL puts Ipomoea sister to Catharabthus and Allamanda.

3.6.3 Vertebrates188 Dataset

wQFM-GDL and ASTRAL-Pro recover almost identical trees in this dataset. wQFM-TREE disagrees on five splits with the NCBI taxonomy tree that available on the Ensemble Compara database (42). The full trees and important subtrees recovered by wQFM-GDl and ASTRAL-Pro are shown in Supplementary Section 4.2. All previously tested methods on this dataset violated a well-established phylogenetic relationship. The elephant shark, which is a chondrichtyian, is placed within Osteicthyians, as sister to Sarcopterygii. wQFM-GDL also recovers the same relationship.

The other cases are quite contentious. In Platyrrhini (monkey sub order), when placing Aotidae, Cebus/Saimiri, and Callitrichidae, wQFM-GDL and ASTRAL-Pro group Aotidate with Cebus/Saimiri. The taxonomy places Cebus/Saimiri and Callitrichidae together, while SpeciesRax places Aotidate and Callitrichidae together. There exist studies that agree with SpeciesRax ((43)) and also wQFM-GDL and ASTRAL-Pro ((44)).

In avian subtree of 24 species, the tree estimated by wQFM-GDL perfectly matches the recent bird phylogeny of 363 taxa published by Stiller et. al., (45). However, the taxonomy groups the Estrildidae and Fringillidae

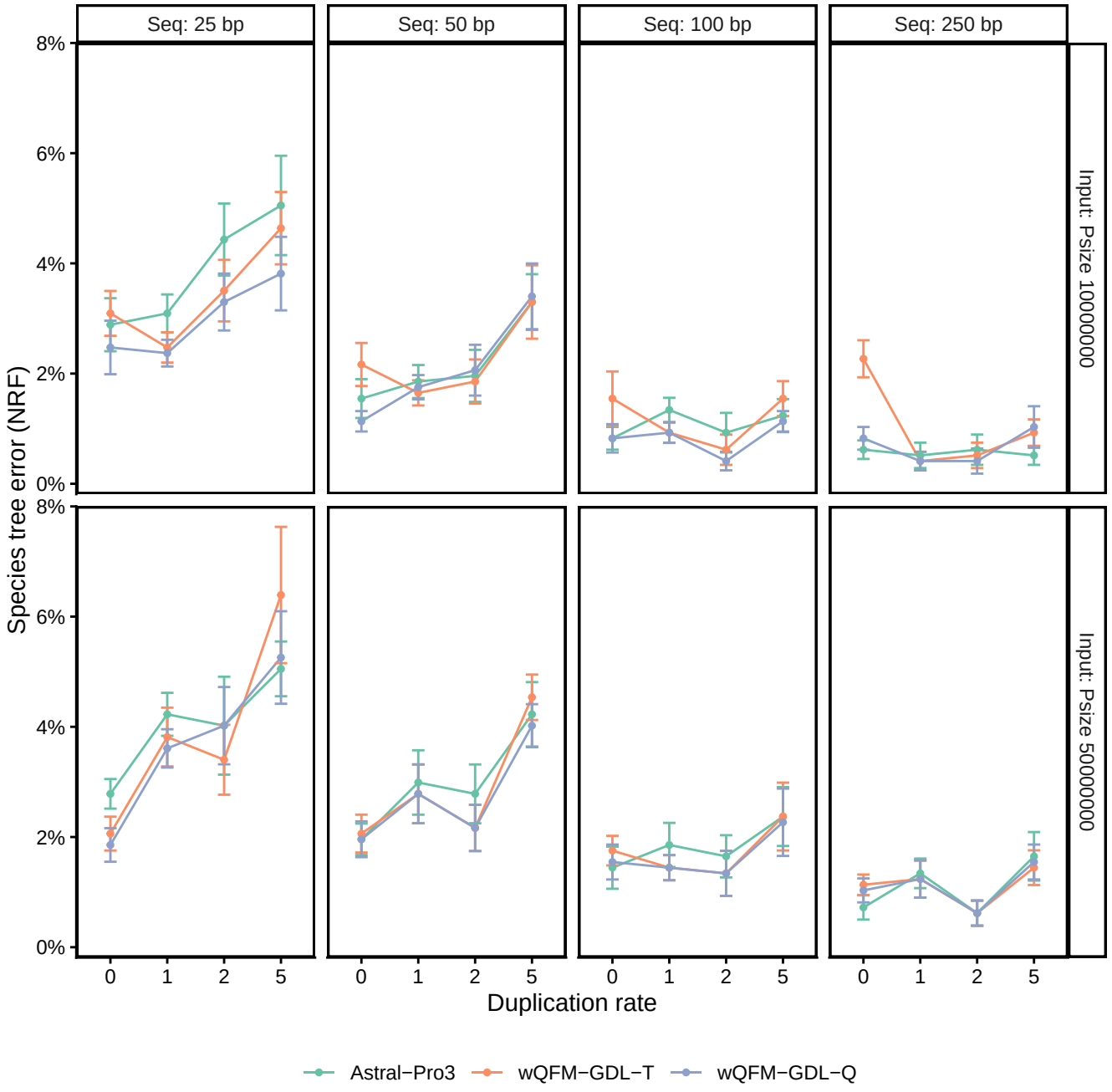


Figure 8: Species tree error on the S100 dataset for 100 ingroup taxa, 1000 gene trees, varying sequence length and the effective haploid population size (labeled as Psize)

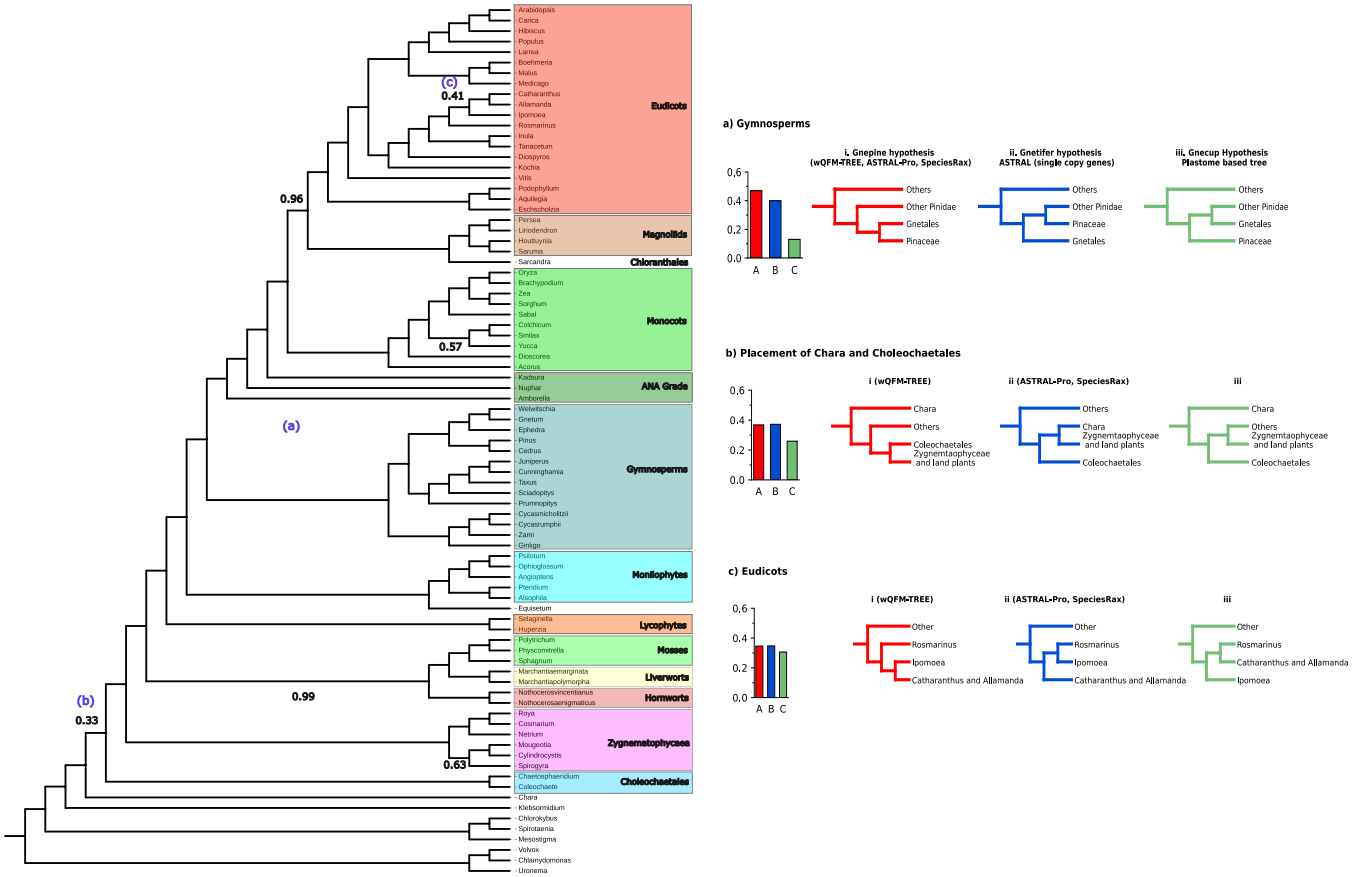


Figure 9: The species tree inferred with wQFM-GDL from the Plants83 dataset and alternative branching orders for contentious relationships.

together, whereas wQFM-TREE, SpeciesRax, and ASTRAL-Pro group Fringillidae and Passerellidae together. Other contentious cases include the placement of Ambassidae. All the tools place Cichliformes as sister to Ambassidae whereas the taxonomy places Pomacentridae as sister to Ambassidae.

3.6.4 Archaea364 Dataset

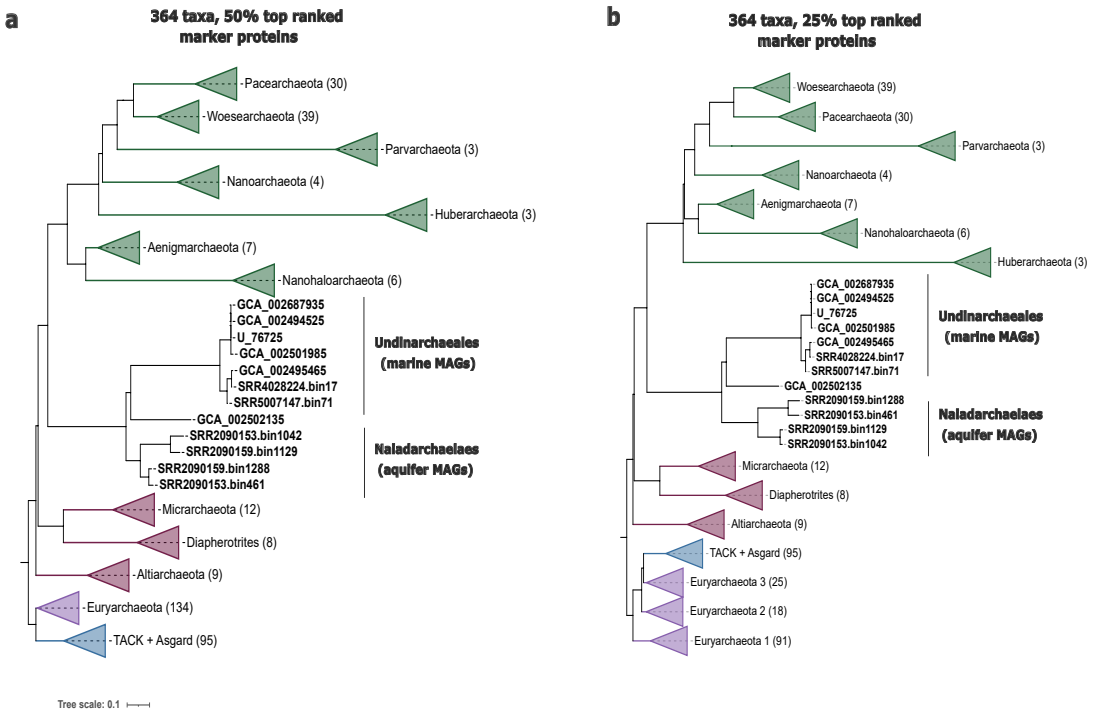


Figure 10: The species trees inferred with wQFM-GDL from the Archaea364 dataset with top 50% and 25% marker proteins

DPANN archaea represent a potentially deep-branching, monophyletic lineage characterized by small cell sizes and genomes. The monophyly, early emergence, and evolutionary significance of the various DPANN clades remain subjects of debate. The main difficulty in resolving the archaeal tree is the presence of host–symbiont gene transfers in the broadly conserved marker genes, in which members of the host-associated DPANN Archaea are sometimes grouped with their hosts in the gene trees. The original study used concatanation with ML for their analysis and did not any summary method analysis. wQFM-GDL successfully reconstructed the species tree from the gene trees provided in (40) as shown in Figure 10. We used not only the full set but also the top 50% and 25% marker proteins for filtering out the incongruent or contaminated markers.

Our analyses recovered DPANN archaea (including Undinarchaeota) as a monophyletic group as a whole. Furthermore, both the species trees from top 50% and 25% marker proteins suggested that Undinarchaeota form a distinct lineage that branches between two other DPANN clans with maximum brnach support. The first clan consists of Altiarchaeota, Micrarchaeota and Diapherotrites, and the second clan consists of Woesearchaeota, Pacearchaeota, Parvarchaeota, Nanoarchaeota, Huberarchaeota, Aenigmarchaeota and Nanohaloarchaeota. Undinarchaeota formed two orders, consisting of aquifer-derived and ocean-derived MAGs, i.e. the Naiadarchaeales and Undinarchaeales, respectively. All these findings match the findings of Dombrowski et. al., (40) done by CA-ML analysis except two slight dissimilarities in the position of Altiarchaeota inside the first clan and the relative position of Parvarchaeota and Nanoarchaeota in the second clan. Astral-Pro also support the monophyly of DPANN and puts Undinarchaeota between two DPANN clans however it puts Micrarchaeota and Diapherotrites in the second clan previously mentioned with only Altiarchaeota in the first clan which deviates from the findings of Dombrowski et. al.

The tree reconstructed from the full set, top 50% and top 25%, all showed similar findings. However, in the species tree of top 25% marker proteins, the Euryarchaeota group is divided into multiple subgroups by all the algorithms.

4 Conclusion

In this study, we proposed algorithms wQMF-GDL-Q and wQFM-GDL-T extending QFM algorithms for gene duplication and loss. Both of the algorithms demonstrate excellent performance and outperform ASTRAL-Pro in most of the model conditions. Although wQFM-GDl-Q requires quartet decomposition which is time-consuming, it performs slightly better than wQFM-GDL-T and we suggest using wQFM-GDL-Q when it is feasible to generate quartet-list from gene trees. ASTRAL-Pro was the only quartet-based summary method to incorporate gene duplication and loss modeling to date. This study positions wQFM-GDL as a strong alternative to ASTRAL-Pro, demonstrating superior accuracy at the expense of increased computational cost.

References

[1] W. P. Maddison, "Gene trees in species trees," *Syst Biol*, vol. 46, pp. 523–536, 1997.

[2] D. Bryant, R. Bouckaert, J. Felsenstein, N. A. Rosenberg, and A. RoyChoudhury, "Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis," *Molecular biology and evolution*, vol. 29, no. 8, pp. 1917–1932, 2012.

[3] J. Chifman and L. Kubatko, "Quartet from snp data under the coalescent model," *Bioinformatics*, vol. 30, no. 23, pp. 3317–3324, 2014.

[4] J. Heled and A. J. Drummond, "Bayesian inference of species trees from multilocus data," *Mol Biol Evol*, vol. 27, pp. 570–580, 2010.

[5] M. Islam, K. Sarker, T. Das, R. Reaz, and M. S. Bayzid, "Stelar: A statistically consistent coalescent-based species tree estimation method by maximizing triplet consistency," *BMC genomics*, vol. 21, no. 1, pp. 1–13, 2020.

[6] L. S. Kubatko, B. C. Carstens, and L. L. Knowles, "Stem: Species tree estimation using maximum likelihood for gene trees under coalescence," *Bioinf*, vol. 25, pp. 971–973, 2009.

[7] B. Larget, S. K. Kotha, C. N. Dewey, and C. Ané, "BUCKy: Gene tree/species tree reconciliation with the Bayesian concordance analysis," *Bioinf*, vol. 26, no. 22, pp. 2910–2911, 2010.

[8] L. Liu, L. Yu, D. K. Pearl, and S. V. Edwards, "Estimating species phylogenies using coalescence times among sequences," *Systematic biology*, vol. 58, no. 5, pp. 468–477, 2009.

[9] L. Liu, L. Yu, and S. V. Edwards, "A maximum pseudo-likelihood approach for estimating species trees under the coalescent model," *BMC Evolutinary Biology*, vol. 10:302, 2010.

[10] L. Liu and L. Yu, "Estimating species trees from unrooted gene trees," *Systematic Biology*, vol. 60, no. 5, pp. 661–667, 2011.

[11] S. Mirarab, R. Reaz, M. S. Bayzid, T. Zimmermann, M. S. Swenson, and T. Warnow, "ASTRAL: genome-scale coalescent-based species tree estimation," *Bioinformatics*, vol. 30, no. 17, pp. i541–i548, 2014.

[12] E. Mossel and S. Roch, "Incomplete lineage sorting: consistent phylogeny estimation from multiple loci," *IEEE Comp Biol Bioinform*, vol. 7, no. 1, pp. 166–171, 2011.

[13] P. Vachaspati and T. Warnow, "Astrid: accurate species trees from internode distances," *BMC genomics*, vol. 16, no. 10, p. S3, 2015.

[14] R. Reaz, M. S. Bayzid, and M. S. Rahman, "Accurate phylogenetic tree reconstruction from quartets: A heuristic approach," *PLoS One*, vol. 9, no. 8, p. e104008, 2014.

[15] S. Snir and S. Rao, "Quartet maxcut: a fast algorithm for amalgamating quartet trees," *Molecular phylogenetics and evolution*, vol. 62, no. 1, pp. 1–8, 2012.

[16] E. Avni, R. Cohen, and S. Snir, "Weighted quartets phylogenetics," *Systematic biology*, vol. 64, no. 2, pp. 233–242, 2015.

[17] M. Mahbub, Z. Wahab, R. Reaz, M. S. Rahman, and M. S. Bayzid, "wQFM: highly accurate genome-scale species tree estimation from weighted quartets," *Bioinformatics*, vol. 37, pp. 3734–3743, Nov. 2021.

[18] S. Mahbub, S. Sawmya, A. Saha, R. Reaz, M. S. Rahman, and Md. S. Bayzid, "Quartet Based Gene Tree Imputation Using Deep Learning Improves Phylogenomic Analyses Despite Missing Data," *Journal of Computational Biology*, vol. 29, pp. 1156–1172, Nov. 2022.

[19] A. Rafi, A. M. S. Rumi, S. A. Hakim, Sohaib, M. T. Tahmid, R. J. I. Momin, T. A. Zaman, R. Reaz, and M. S. Bayzid, "wQFM-TREE: highly accurate and scalable quartet-based species tree inference from gene trees," *Bioinformatics Advances*, vol. 5, p. vbaf053, 03 2025.

[20] W. M. Fitch, "Homology: a personal view on some of the problems," *Trends in genetics*, vol. 16, no. 5, pp. 227–231, 2000.

[21] Y. Yang and S. A. Smith, "Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics," *Molecular biology and evolution*, vol. 31, no. 11, pp. 3081–3092, 2014.

[22] B. Boussau, G. J. Szöllősi, L. Duret, M. Gouy, E. Tannier, and V. Daubin, "Genome-scale coestimation of species and gene trees," *Genome research*, vol. 23, no. 2, pp. 323–330, 2013.

[23] R. Chaudhary, M. S. Bansal, A. Wehe, D. Fernández-Baca, and O. Eulenstein, "iGTP: a software package for large-scale gene tree parsimony analysis," *BMC Bioinformatics*, pp. 574–574, 2010.

[24] A. Wehe, M. S. Bansal, J. G. Burleigh, and O. Eulenstein, "Duptree: A program for large-scale phylogenetic analyses using gene tree parsimony," *American Journal of Botany*, vol. 24, no. 13, pp. 1540–1541, 2008.

[25] M. S. Bayzid, S. Mirarab, and T. Warnow, "Inferring optimal species trees under gene duplication and loss," in *Proc. of Pacific Symposium on Biocomputing (PSB)*, vol. 18, pp. 250–261, 2013.

[26] M. S. Bayzid and T. Warnow, "Gene tree parsimony for incomplete gene trees: addressing true biological loss," *Algorithms for Molecular Biology*, vol. 13, p. 1, 2018.

[27] M. S. Bayzid, "Inferring optimal species trees in the presence of gene duplication and loss: Beyond rooted gene trees," *Journal of Computational Biology*, vol. 30, no. 2, pp. 161–175, 2023.

[28] R. Chaudhary, J. G. Burleigh, and D. Fernández-Baca, "Inferring species trees from incongruent multi-copy gene trees using the Robinson-Foulds distance," *Algorithms for Molecular Biology*, vol. 8, pp. 1–12, Dec. 2013.

[29] E. K. Molloy and T. Warnow, "FastMulRFS: fast and accurate species tree estimation under generic gene duplication and loss models," *Bioinformatics*, vol. 36, pp. i57–i65, July 2020.

[30] L. De Oliveira Martins, D. Mallo, and D. Posada, "A bayesian supertree model for genome-wide species tree reconstruction," *Systematic biology*, vol. 65, no. 3, pp. 397–416, 2016.

[31] C. Zhang, C. Scornavacca, E. K. Molloy, and S. Mirarab, "ASTRAL-Pro: Quartet-Based Species-Tree Inference despite Paralogy," *Molecular Biology and Evolution*, vol. 37, pp. 3292–3307, Nov. 2020.

[32] B. Morel, P. Schade, S. Lutteropp, T. A. Williams, G. J. Szöllősi, and A. Stamatakis, "Speciesrax: a tool for maximum likelihood species tree inference from gene family trees under duplication, transfer, and loss," *Molecular biology and evolution*, vol. 39, no. 2, p. msab365, 2022.

[33] Y. Han and E. K. Molloy, "Improving quartet graph construction for scalable and accurate species tree estimation from gene trees," *Genome Research*, vol. gr.277629.122., May 2023.

[34] J. Willson, M. S. Roddur, B. Liu, P. Zaharias, and T. Warnow, "DISCO: Species Tree Inference using Multicopy Gene Family Tree Decomposition," *Systematic Biology*, vol. 71, pp. 610–629, 08 2021.

[35] D. Swofford, *PAUP*: Phylogenetic analysis using parsimony (* and other methods). Ver. 4*. Sinauer Associates, Sunderland, Massachusetts, 2002.

[36] C. Zhang, C. Scornavacca, E. K. Molloy, and S. Mirarab, "Astral-pro: quartet-based species-tree inference despite paralogy," *Molecular biology and evolution*, vol. 37, no. 11, pp. 3292–3307, 2020.

[37] D. Mallo, L. de Oliveira Martins, and D. Posada, "Simphy: phylogenomic simulation of gene, locus, and species trees," *Systematic biology*, vol. 65, no. 2, pp. 334–344, 2016.

[38] N. Ly-Trong, S. Naser-Khdour, R. Lanfear, and B. Q. Minh, "Alisim: a fast and versatile phylogenetic sequence simulator for the genomic era," *Molecular biology and evolution*, vol. 39, no. 5, p. msac092, 2022.

[39] M. N. Price, P. S. Dehal, and A. P. Arkin, "Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix," *Molecular biology and evolution*, vol. 26, no. 7, pp. 1641–1650, 2009.

[40] N. Dombrowski, T. A. Williams, J. Sun, B. J. Woodcroft, J.-H. Lee, B. Q. Minh, C. Rinke, and A. Spang, "Undinarchaeota illuminate dpann phylogeny and the impact of gene transfer on archaeal evolution," *Nature Communications*, vol. 11, no. 1, p. 3939, 2020.

[41] N. J. Wickett, S. Mirarab, N. Nguyen, T. Warnow, E. Carpenter, N. Matasci, S. Ayyampalayam, M. S. Barker, J. G. Burleigh, M. A. Gitzendanner, *et al.*, "Phylotranscriptomic analysis of the origin and early diversification of land plants," *Proceedings of the National Academy of Sciences*, vol. 111, no. 45, pp. E4859–E4868, 2014.

[42] S. Federhen, "The ncbi taxonomy database," *Nucleic acids research*, vol. 40, no. D1, pp. D136–D143, 2012.

[43] P. Perelman, W. E. Johnson, C. Roos, H. N. Seuánez, J. E. Horvath, M. A. Moreira, B. Kessing, J. Pontius, M. Roelke, Y. Rumpler, *et al.*, "A molecular phylogeny of living primates," *PLoS genetics*, vol. 7, no. 3, p. e1001342, 2011.

[44] A. Rodrigues, E. Douzery, *et al.*, "Patterns of macroevolution among primates inferred from a supermatrix of mitochondrial and nuclear dna," *Molecular phylogenetics and evolution*, vol. 53, no. 3, pp. 808–825, 2009.

[45] J. Stiller, S. Feng, A.-A. Chowdhury, I. Rivas-González, D. A. Duchêne, Q. Fang, Y. Deng, A. Kozlov, A. Stamatakis, S. Claramunt, *et al.*, "Complexity of avian evolution revealed by family-level genomes," *Nature*, vol. 629, no. 8013, pp. 851–860, 2024.