Quartet-based species tree methods enable fast and consistent tree of blobs reconstruction under the network multispecies coalescent

Junyan Dai[1,*[0009−0009−5699−2196]], Yunheng Han[1,*[0000−0003−0200−5924]], and Erin K. Molloy[1,2,⋆[0000−0001−5553−3312]]

[1] Department of Computer Science, University of Maryland, College Park, MD 20742, USA
[2] University of Maryland Institute for Advanced Computer Studies, College Park, MD 20742, USA
ekmolloy@umd.edu

Abstract. Hybridization between species is an important force in evolution, commonly modeled by the network multispecies coalescent. Reconstructing evolutionary histories, called species networks, under this model is notoriously challenging, with the leading methods scaling to just tens of species. Divide-and-conquer is a promising path forward; however, methods with statistical consistency guarantees require the *tree of blobs* (TOB), which displays only the tree-like parts of the network. TOB reconstruction under the NMSC is challenging in its own right, with the only available method TINNiK having time complexity $O(n^5 + n^4k)$, where k is the number of input gene trees and n is the number of species. Here, we present a framework for TOB reconstruction that operates by (1) seeking a refinement of the TOB and then (2) contracting edges in it. For step (1), we show that an optimal solution to Weighted Quartet Consensus is a TOB refinement almost surely, as the number of gene trees increases, motivating the use of fast quartet-based methods for species tree estimation such as ASTRAL or TREE-QMC. For step (2), we contract edges in the refinement tree based on the same hypothesis tests as TINNiK, which are applicable to subsets of four taxa. We show that sampling just $O(n)$ four-taxon subsets around each edge enables statistically consistent TOB estimation, so asymptotic runtime dominated by tree reconstruction. Leveraging TREE-QMC for this step gives our method a time complexity of $O(n^3k)$ and its name TOB-QMC. On simulated data sets, TOB-QMC is typically at least as accurate and often more accurate than TINNiK. Moreover, TOB-QMC scales to larger data sets and enables efficient and interpretable exploration of hyperparameters (significance thresholds); we show the utility of this feature in re-analyzing several phylogenomic data sets. Lastly, our results show how to interpret the ASTRAL tree in the context as gene flow, which is critical given the recent result that network inference based on the ASTRAL can be positively misleading.

Keywords: Phylogenetic Networks · Multispecies Coalescent · Quartets · Tree of Blobs

⋆ Corresponding author
*J. Dai and Y. Han—These authors contributed equally to this work.

1 Introduction

Hybridization between species (or admixture between populations) are an important forces in evolution; however, they complicate the reconstruction of evolutionary histories, requiring them to be represented as phylogenetic networks rather than trees. In this context, phylogenetic networks (also called species or population networks) govern the ancestry of unlinked genomic regions (i.e., gene genealogies), as modeled by the network multi-species coalescent (NMSC) [35]. Despite significant progress over the last decade, phylogenetic network reconstruction under the NMSC continues to be a major scientific challenge. Perhaps the most statistically rigorous methods developed to date (e.g., [36]) are those that employ likelihood functions accounting for gene genealogies (also called gene trees) evolving within a species network under the NMSC [35, 38] as well as molecular sequences evolving down each gene tree [12], integrating over all possible gene trees. The latter step can be avoided by estimating gene trees from molecular sequences prior to network reconstruction, but even this simplified likelihood can be prohibitively expensive, motivating the development of pseudo-likelihood methods [37, 32]. One of the most popular pseudo-likelihood methods, SNaQ [32], is limited to around 30 species [21].

Given scalability limitations, researchers sometimes add edges representing gene flow to a fixed species tree, typically estimated with the popular method ASTRAL [23] (e.g., Fig. 2 in [10]; also see [9]). ASTRAL is a heuristic for the NP-hard Weighted Quartet Consensus (WQC) problem [22], which can be framed as weighting quartets (i.e., unrooted trees on four species) according to their frequencies in the input gene trees and then seeking a species tree to maximize the total weight of the quartets it displays. However, Dinh and Baños (2025) recently showed that this approach can be *positively misleading* (because an optimal solution to WQC is not necessarily a display tree, as the number of (error-free) gene trees tends to infinity [11] (also see [24]). This negative result is impacts numerous prior studies using ASTRAL in the presence of gene flow and suggests scalable phylogenetic network methods are urgently needed.

A promising path forward is divide-and-conquer [17, 40, 21]. As an example, the recently developed InPhyNet method [21] partitions species into subsets, estimates a network on each subset with SNaQ, and then combines the subset networks together, similar to disjoint tree mergers [25, 26]. Of the existing divide-and-conquer methods only InPhyNet [21] claims statistical consistency; however, the proof assumes that species are partitioned into subsets according to the tree of blobs (TOB) [14], which displays only the tree-like parts of the network (note that other network methods, including NANUQ+ [3], also require a tree of blobs). To our knowledge, the first and only available method for TOB reconstruction under the NMSC is TINNiK [2]. TINNiK is statistically consistent and has time complexity $O(n^4k + n^5)$, where k is the number of input gene trees and n is the number of species. The first term comes from computing quartet weights, which are needed to perform hypothesis tests on all subsets of four species, and the second term comes from updating the test results to ensure any remaining signals of network-like evolution are removed. The final results are encoded as a distance matrix so that applying Neighbor Joining [31] and contracting zero-length edges yields the tree of blobs. However, as TINNiK's asymptotic runtime would suggest, it cannot scale to large numbers of species, so InPhyNet implements an alternative approach to subset decomposition on large data sets, without statistical guarantees.

To address this issue, we introduce a new framework for TOB reconstruction that operates by (1) seeking a refinement of the TOB and then (2) contracting edges in it. For step (1), we show that an optimal solution to WQC is a TOB refinement almost surely, as the number of gene trees increases, motivating the use of fast quartet-based methods for species tree estimation such as ASTRAL or TREE-QMC [15]. For step (2), we show that applying the same hypothesis test as TINNiK [1] to just $O(n)$ four-taxon subsets around each edge is sufficient for statistically consistent TOB reconstruction. We implement our framework on top of the quartet-based method TREE-QMC, giving our method a time complexity of $O(n^3k)$ as well as its name TOB-QMC. On simulated data sets, TOB-QMC achieves competitive accuracy with TINNiK. Moreoever, TOB-QMC scales to large data sets (with 200 taxa) and enables fast and interpretable exploration of hyperparameters (significance levels) used in hypothesis testing. We demonstrate the importance of this feature re-analyzing phylogenomic data sets for seed plants, the bee subfamily *Nomiinae*, and the butterfly subfamily *Heliconiinae*.

Overall, our study advances theory and practice for scalable evolutionary analyses in the presence of gene flow. On the theoretical side, our results explain why ASTRAL, and related quartet-based methods, have great utility in the presence of gene flow, provided the results are interpreted as TOB refinements; this context is critical given ASTRAL's widespread use. On the empirical side, our framework enables researchers

to evaluate the impact of network-like evolution on large data sets. Long term, we expect TOB-QMC to have applications in phylogenetic network reconstruction.

2 Results

2.1 Overview of TOB-QMC

To present our method TOB-QMC, consider the phylogenetic network shown in Figure 1A. Internal vertices with in-degree greater than 1 represent hybrid species or admixed populations, whereas internal vertices with in-degree 1 represent speciation events or population splits. Each edge in the network is either a tree edge (meaning its deletion disconnects the network) or a blob edge. The contraction of all blob edges in the network yields a tree of blobs (TOB) (Fig. 1B). Adding edges arbitrarily to the TOB until it is binary (i.e., all non-leaf vertices have degree 3, after suppressing the root vertex) yields a TOB refinement (Fig. 1C). The TOB can be obtained from a refinement of it simply by identifying and contracting all false positive (FP) edges (i.e., edges in the TOB refinement that are not in the TOB). Although our example shows a directed network and rooted trees, our results are for semi-directed networks and unrooted trees; see Methods.

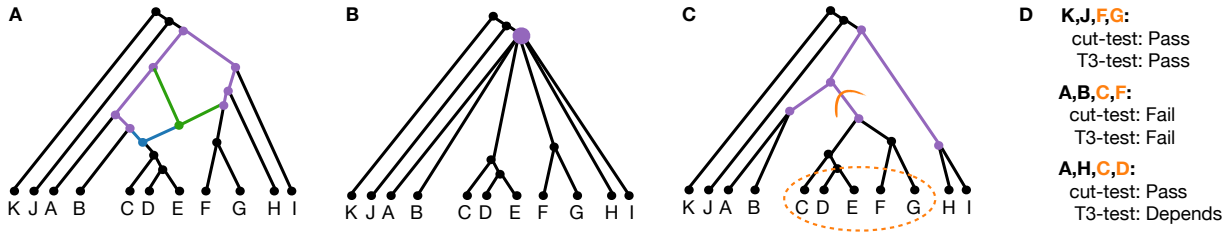

Fig. 1: **A)** A level-2 directed phylogenetic network N^+ with one blob. Reticulation vertices, and their corresponding edges, are colored blue and green. All other edges and vertices in the blob are colored purple. **B)** Tree of blobs (TOB) for N^+. **C)** TOB refinement for that is NOT induced by any display trees of N^+ (note that the display trees are formed by deleting one blue edge and one green edge). False positive edges, highlighted in purple, must be contracted to obtain the TOB. **D)** To decide whether to contract the edge in TOB refinement indicated in orange, TINNiK's tree test is applied to two taxa on one side of edge and two taxa on the other side. Some four-taxon subsets pass the test (i.e., accept the null hypothesis) because the evolutionary relationship among them is "tree-like," but others fail; thus, the selection process matters.

Based on these observations, we propose to reconstruct the (unrooted) TOB from gene trees in two stages. First, we seek a TOB refinement by solving a weighted quartet consensus (WQC) problem using fast quartet-based species tree methods (e.g., TREE-QMC or ASTRAL). This step is theoretically grounded because we show that optimal solutions to WQC are TOB refinements almost surely, as the number of gene trees generated under the NMSC tends to infinity (see Theorem 1 in Methods). Second, we identify and contract false positive (FP) edges in the TOB refinement estimated in step 1 (Fig. 1C–D). For step (2), we leverage TINNiK's hypothesis tests, treating them as oracles that evaluate whether the evolutionary relationship among four taxa is "tree-like" (this means the induced subnetwork is not a blob with four pendent edges; see Methods for details). Applying these "tree-tests" to all subsets of four taxa around each edge (i.e., taking two taxa from one side of the edge and two taxa from the other side) ensures FP edges will be identified and contracted. Critically, we design targeted sampling strategies that reduce the number of tested four-taxon subsets from $O(n^4)$ to $O(n^2)$ (heuristic search algorithm) or $O(n)$ (3f1a search algorithm) per edge, with the latter maintaining statistical consistency (Theorem 2 in Methods + Theorem 4 in [2]).

To summarize, our approach is statistically consistent (unrooted) TOB estimator, meaning that as the number of gene trees generated under the NMSC given some network $\mathcal{N}$ goes to infinity, it returns the correct (unrooted) TOB for $\mathcal{N}$ almost surely. Our proofs assume five conditions:

1. underlying network is binary (i.e., all vertices have degree 3 after suppressing the root)
2. edges of the network have positive lengths (in coalescent units)

3. edges whose target is a reticulation vertex have inheritance proportions that sum to 1 (and take on values between 0 and 1)
4. network is relaxed quartet-nonanomalous (see Methods)
5. ASTRAL-III is used to compute the TOB refinement

The first three are standard in the phylogenetics literature. The fourth is required for our proof of Theorem 1 (step 1). A stronger version of condition four may be required depending on the tree-test used for hypothesis testing (step 2). In particular, TINNiK introduces two tree-tests (the T3-test and the cut-test) and proves that approaches applying them to all four-taxon subsets are consistent (Theorem 4 in [2]). However, the T3 test requires that the network is quartet-nonanomalous; otherwise, it may incorrectly reject the null. This issue can occur when four taxa form a blob with 3 pendent edges with two of the taxa descending from the hybrid vertex in the blob (Fig. 1D), depending on the numerical parameters (it is well-known that these so-called 3_2-cycles challenge quartet-based network methods; see [32, 5] as well as Section 1.3 in the supplement). Nevertheless, TINNiK's T3 test is the default/recommended option as it performs well in practice [2]. The fifth, and final, condition yields a trivial proof of consistency, thanks to ASTRAL-III's optimality guarantees, but other fast and accurate heuristic for WQC can reasonably be used for step (1). No assumptions about network *level* are made (see Section 1.1 of Supplement). See Methods for more details and theoretical results.

We implemented our proposed approach on top of the quartet-based method TREE-QMC [15, 16], which is used for step (1), thus we call our method TOB-QMC. For step (20), our implementation invokes the code from TINNiK's hypothesis tests on subsets of four taxa. **TOB-QMC-exhaustive** applies a tree-test to all 4-taxons subsets around each edge (bipartition) in the base tree, storing the minimum p-value found and the minimizer four-taxon subsets. **TOB-QMC-3f1a** executes the 3f1a search algorithm (see Methods and Algorithm 1 in Supplement for details). **TOB-QMC-fast** and **TOB-QMC-default** performs heuristic search around the edge (biparition), stopping when the maximum number of iterations is reached, either $2n^2$ and $n^2/4$, respectively, where n is the number of taxa, or a local optima is reached (see Methods and Algorithm 2 in Supplement for details). Importantly, tree-tests have a singularity when the evolutionary relationship among the four species is a star (i.e., an unresolved tree). To rule out this scenario, TOB-QMC applies TINNiK's star-test to the minimizer four-taxon subset and then stores the resulting p-value [1, 2]. Lastly, TOB-QMC contracts edges depending on whether the null hypothesis of the tree-test is rejected (i.e., $p < \alpha$) or the null hypothesis of the star-test is <u>not</u> rejected (i.e., $p \geq \beta$). This requires the user to set α, β vales, same as TINNiK. By storing p-values and minimizers for each edge, the impact of significance thresholds α, β can be evaluated offline, avoiding full re-computation of the TOB. Higher α and lower β result in more edges being contracted by TOB-QMC.

2.2 Method Evaluation

We evaluate TOB-QMC and TINNiK using simulations so that the true network and tree of blobs are known.

Data simulation. Data sets were simulated following the protocol from [21]. First, networks with 50, 100, and 200 taxa were simulated via a birth-death-hybridization process using SiPhyNetwork [19]. For each number of taxa, 50 level-1 and 50 level-2 networks were (randomly) selected to produce 50 replicates, where the true network and its true TOB are known (see Table S1 for network/TOB summary statistics). Second, 1000 gene trees were simulated for each network under NMSC (with Independent Inheritance) using PhyloCoalSimulations [13]. To vary the amount of ILS, we multiplied the branch lengths in the network by scaling factors: 2.0, 1.0, 0.5, and 0.25; this resulted in ILS levels of $\sim$ 45%, 65%, 80%, 90%, respectively (see Fig. S4 in the Supplement). Higher ILS levels correspond to greater gene tree discordance, due to gene lineages failing to coalesce, and thus are more challenging model conditions. Third, sequences were simulated from each gene tree under the GTR+F+G model. Fourth, gene trees were estimated from sequences using IQ-TREE [27, 20]. The mean gene tree estimation error ranged from 40% to 70% (Figs. S5–S8 in Supplement). Lastly, gene trees were given as input to TOB-QMC and TINNiK. A detailed description of the simulation study, including software versions and commands, is provided in the Supplement.

Evaluation metrics. Methods were evaluated based on tree error and runtime. Tree error was computed as the false positive rate (FPR), false negative rate (FNR), and arithmetic mean of FNR and FPR. FNR is the number of branches (bipartitions) in the true TOB not in the estimated TOB, divided by the total number of internal branches in the true TOB. FPR is the number of branches in the estimated TOB not in the true TOB, divided by the total number of internal branches in the estimated TOB.

Hyperparameter tuning. As previously mentioned, both TOB-QMC and TINNiK are based on hypothesis tests applied to subsets of four taxa, specifically the tree-test and the star test. Both methods were run with the T3 version of the tree-test, as this corresponds to the default version of TINNiK. The significance threshold α is used to reject the null hypothesis of the tree-test; low values are typically used to avoid false detection of non-tree-like evolution due to gene tree estimation error or small numbers of gene trees. The significance threshold β is used to not reject the null hypothesis of the star-test; higher values are used to avoid false detection of star-like evolutionary relationships. There is currently no structured approach for selecting α, β [2]. To ensure fair settings for both methods, we evaluated the impact of the α, β hyperparameters on the 50- and 100-taxon data sets using true gene trees for all combinations of $\alpha \in \{10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}, 10^{-9}\}$ and $\beta \in \{0.90, 0.95, 1.00\}$ (note these ranges are similar to the original TINNiK study [2] and that setting $\beta = 1$ ensures the null of the star-test is always rejected so branches are never contracted based on it). We observed that setting $\beta = 0.95$ and $\alpha = 10^{-7}$ gave the good results for both methods, with error increasing for higher settings of α (Figs. S9–S12).

We also explored the impact of TOB-QMC's search algorithms, compared to the base tree computed with TREE-QMC without any edges contracted (Figs. S13–S14 in the Supplement). All algorithms reduced FPR compared to the base tree with little impact on FNR. TOB-QMC-3f1a had the highest FPR error, whereas default and fast modes achieved similar FPR to TOB-QMC-exhaustive. Thus, in the simulation study, we benchmarked TOB-QMC-default and TOB-QMC-fast against TINNiK, with $\beta = 0.95$ and $\alpha = 10^{-7}$, on estimated gene trees.

2.3 Results on Simulated Data Sets

Impact of ILS level and network level. We first evaluated TOB-QMC and TINNiK varying the ILS level and the network level on data sets with estimated gene trees. For data sets with 100 taxa (Fig. 2), TOB-QMC's base tree had the best FNR, but also the worst FPR, of all methods. Its FNR was generally low ($\leq\sim$5% on average) for all model conditions, except for the highest ILS level, in which case FNR increased but was still less than $\sim$10% on average. TOB-QMC-fast and TOB-default had slightly higher FNR compared to the TOB-QMC base tree but much lower FPR, suggesting branch contraction based on the minimum p-value search was effective. Their FPR varied, ranging from $\sim$5% for the easiest model condition (low ILS and level-1) to $\sim$18% for the hardest (high ILS and level-2). Likewise, the performance difference between TOB-QMC and TINNiK increased with model condition difficulty, with TOB-QMC outperforming TINNiK. These overall trends held on data sets with 50 and 200 taxa (Fig. S15–S16), although TINNiK could not be run on the largest data sets with 200 taxa.

Impact of varying the number of taxa. Next we evaluated TOB-QMC and TINNiK for increasing numbers of taxa (Figs. 3). All methods were fast for 50 taxa. At 100 taxa, TINNiK took over 2 hours on average, while both TOB-QMC version took less than 0.5 hours. At 200 taxa, TINNiK could not run within our wall clock limit (48 hours) but TOB-QMC-default took less than 2.5 hours on average; TOB-QMC-fast took less than 1 hour on average. Tree error of TINNiK increased with the number of taxa, going from $\sim$ 5% for 50 taxa to $\sim$ 10% for 100 taxa, on average for level-2 networks (no branch length scaling). In contrast, the tree error of TOB-QMC ($\sim$ 5%) was relatively. Similar trends were observed for the other model conditions (Figs. S17–S20).

Impact of violating the quartet non-nonanomalous assumption. Lastly, we evaluated the impact of violating the quartet-nonanomalous assumption on TOB-QMC and TINNiK (as this assumption is made by the T3 test used by both methods). This required us to first bin the replicate networks in each model condition based on whether they were quartet-anomalous or not. As this is computationally intensive, we only binned networks with 50 and 100 taxa. We found that the fraction of anomalous networks (out of the 50 replicates per model condition) increased with ILS level and network level (Table S2). For 100 taxa, just 6% of networks were anomalous for the easiest model condition (level: 1; branch scaling factor: 2) but reached 52% for the hardest (level 2: branch scaling factor: 0.25). Sometimes the FPR and FNR of TOB-QMC and TINNiK were greater for anomalous networks compared to non-anomalous ones of the same model condition (e.g., 100 taxa, level: 2, branch scaling factor: 2), as might be expected; however, the opposite trend was observed for other model conditions (e.g., 100 taxa, level: 1, branch scaling factor: 0.25). However, our main findings did not change: TOB-QMC was competitive with TINNiK in terms of tree accuracy, often outperforming it on challenging model conditions, which had the greatest proportion of anomalous networks (Fig. S21–S22).

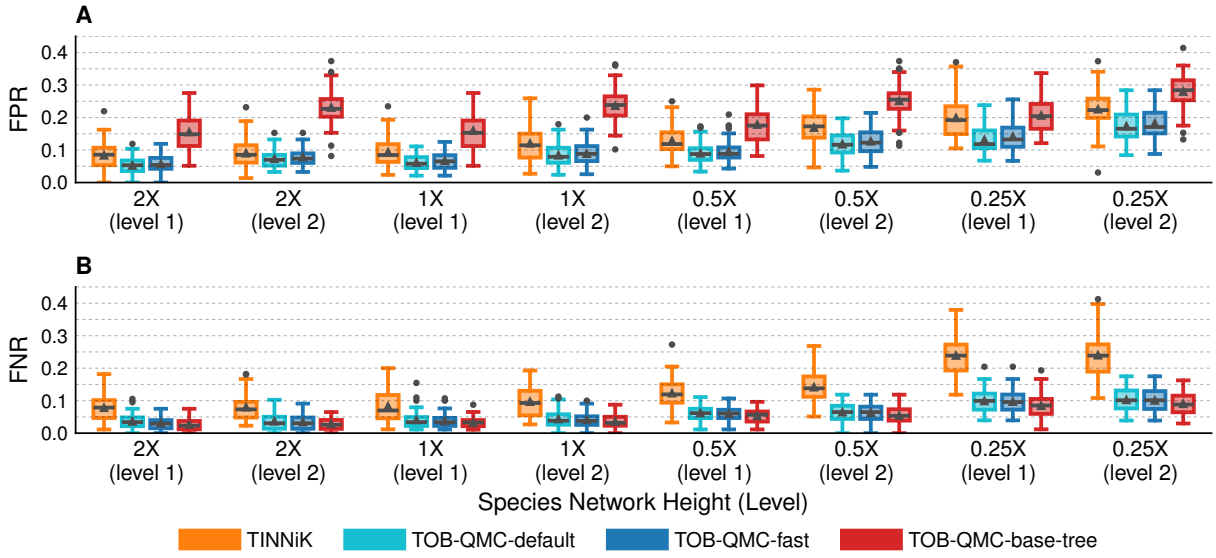

Fig. 2: FPR and FNR for TOB reconstruction methods on 100-taxon data sets with estimated gene trees. Model condition on x-axis indicates network level and branch scaling factor (lower factor is higher ILS).

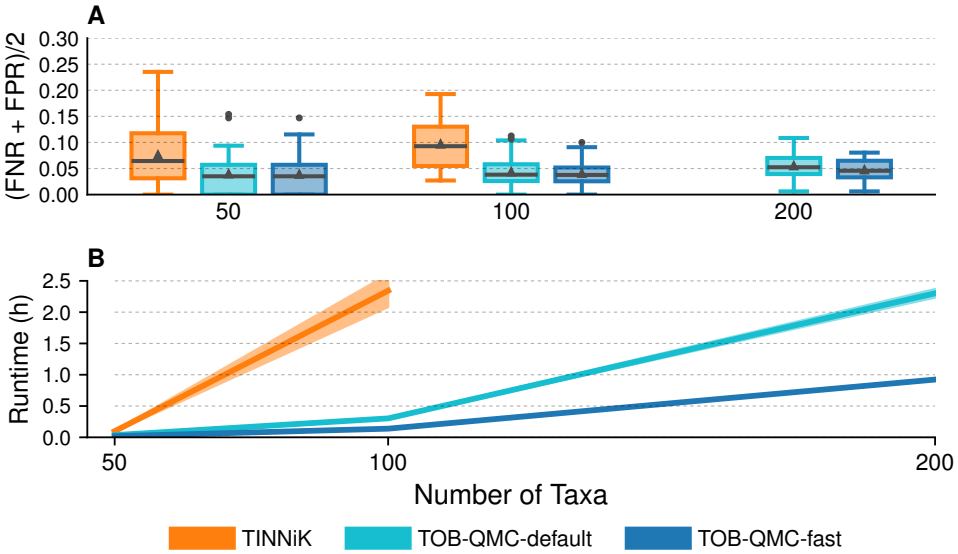

Fig. 3: Tree error and runtime for TOB reconstruction methods for level-2 networks without branch length scaling (ILS level: 65%).

2.4 Application to Phylogenomic Data Sets

TOB-QMC-3f1a had slightly higher FPR in simulations; however, it is fast and simple. Moreover, unlike in simulations, which require significance thresholds to compute FPR and FNR, biologists may wish to use our tool to explore whether there are signals of hybridization in their data. To evaluate the utility of TOB-QMC-3f1a for this purpose, we applied it to phylogenomic analyses data sets for the bee subfamily *Nomiinae* (31 taxa, 852 gene trees) [8, 6], the butterfly subfamily *Heliconiinae* (63 taxa, 3393 gene trees) [10], and seed plants (96 taxa, 406 gene trees) [28, 21]. The bee data set was the smallest so we also ran TINNiK on this data set and compared trees of blobs. Both TINNiK and TOB-QMC-3f1a were fast, taking less than 1

minute, as expected based on our scalability study. The other phylogenomic data sets had larger numbers of taxa, but we were able to analyzed them on a laptop with TREE-QMC-3f1a in just a few minutes.

Nomiinae bees. On the bee data set, both TINNiK and TOB-QMC-3f1a returned the same binary TOB when α was set to 10^{-5} and 10^{-7}, respectively. Increasing α to 10^{-4} and 10^{-6}, respectively, resulted in a blob for the genus *Stictonomia* (Fig. S23; Table S3). A prior study by [8] reported conflicting species tree topologies on this genus, and gene flow was detected by [6] using several different statistical tests for hybridization, including MSCquartets [30], which is similar to TINNiK, as well as HyDe [7] and Patterson's D statistics [29]. As α increased, the TOBs returned by both methods became less refined. Notably, TINNiK needed to be re-run for each setting of α, β, whereas TOB–3f1a did not (this is because edges in its "base tree" are annotated with p values and thus can be contracted for each setting of significance thresholds). This saved time and enabled us to interpret why edges were being contracted. For example, TOB-QMC-3f1a contracted the edge associated with *Stictonomia* because the T3-test applied to the three taxa in this genus plus *Nomiapis bispinosa* was $p = 1.4 \times 10^{-07}$. The associated qCFs were 80/69/21 (0.47/0.41/0.12). These qCFs are suggestive of gene flow (as the T3 test is evaluates whether two qCFs are equal and lesser than the third), there are only 170 gene trees so there is limited data to support this conclusion.

Heliconiinae butterflies. The TOB-QMC-3f1a analysis for butterflies had 14 edges with $\alpha < 10^{-15}$ and $\beta \geq 0.95$; these edges were contracted to get the TOB shown in Fig. 4 (also Fig. S24 and Table S4). The edges were all consistent with the original study, which constructed a species tree with ASTRAL and then annotated it with gene flow events (Fig. 2B in [10]); see the Supplement for details. The original study [10] also detected gene flow among 3 species in *Wallacei* clade (branch 43) and in *Eueides* clade (e.g., branches 48 and 50). Branch 43 had $p = 1.5 \times 10^{-8}$ and qCFs of 2710/205/106 (0.90/0.07/0.04), and branch 48 had $p = 0.0001$ and qCFs of 1533/103/55 (0.91/0.06/0.03). Thus, these events were significant according to the TOB-QMC-3f1 analysis but did not survive the high significance threshold of $\alpha = 10^{-15}$. Both are branches are characterized by the majority of the gene trees ($\sim$90%) supporting the same quartet topology; in the remaining 10% of gene trees, one of the alternative quartet topologies is displayed 2$\times$ more than other, suggestive of gene flow. For branch 43, we exhaustively tested all four-taxon subsets around the induced quadrapartition (i.e., we performed T3 tests on qCFs for *H. burneyi*, *H. wallacei*, *H. egeria*, plus every all other taxa, so 63 tests total). All tests had similar qCFs and p-values (Table S5). Because the vast majority of the genome has a consistent topology for branches 43 and 48, these signals of gene flow may not be relevant to some studies. Overall, this analysis highlights the utility of putting the outcomes of T3 tests (i.e., their p-values but also the underlying qCFs) within the context of the base tree, rather than applying a strict α, β significance threshold prior to TOB reconstruction, as in TINNiK.

Seed Plants. Lastly, we applied TOB-QMC to a seed plant data set, previously analyzed by the recently developed divide-and-conquer network method InPhyNet [21]. In the InPhyNet analysis, TINNiK was used to build a TOB for subset decomposition, with $\alpha = 0.001$. The resulting TOB was fairly unresolved, as was the TOB produced in our TOB-QMC-3f1a analysis (25 branches were contracted for $\alpha = 0.001$; see Fig. S25 and Tables S6–S7). Some branches towards the root had weaker evidence; for branches 6 and 7 had p-values less than 10^{-5} but the underlying qCFs were 220/27/3 (0.88/0.11/0.01; $p =$) and 179/26/2 (0.87/0.13/0.01), respectively. This means the T3 test is based on few gene trees (250 and 207) of which the majority ($\sim$87%) support a single topology. In contrast, there branch 25 has p-value less than 10^{-19} with qCFs of 149/134/25 (0.48/0.44/0.08). This result is supported by more gene trees (308) with the majority ($\sim$91%) split between two quartet topologies. This signal is less likely to be due to spurious noise, although problems in identification of single copy orthologs could lead to biases that appear like introgression.

3 Discussion

In this work, we introduced a new framework for tree of blobs (TOBs) reconstruction under the network multispecies coalescent (NMSC) that first seeks a TOB refinement and then contracts edges from it to get a TOB, combining a simple search algorithm with hypothesis testing. Although this approach is simple, we provide theoretical justification. First, we show that an optimal solution to WQC is a TOB refinement almost surely, as the number of gene trees increases. Second, we show that our 3f1a search algorithm, which samples only a linear number of 4-taxon subsets around each branch, detects false positive edges almost surely, as the number of gene trees increases. Together, these results yield a statistically consistent and fast approach for TOB reconstruction, with some standard assumptions.

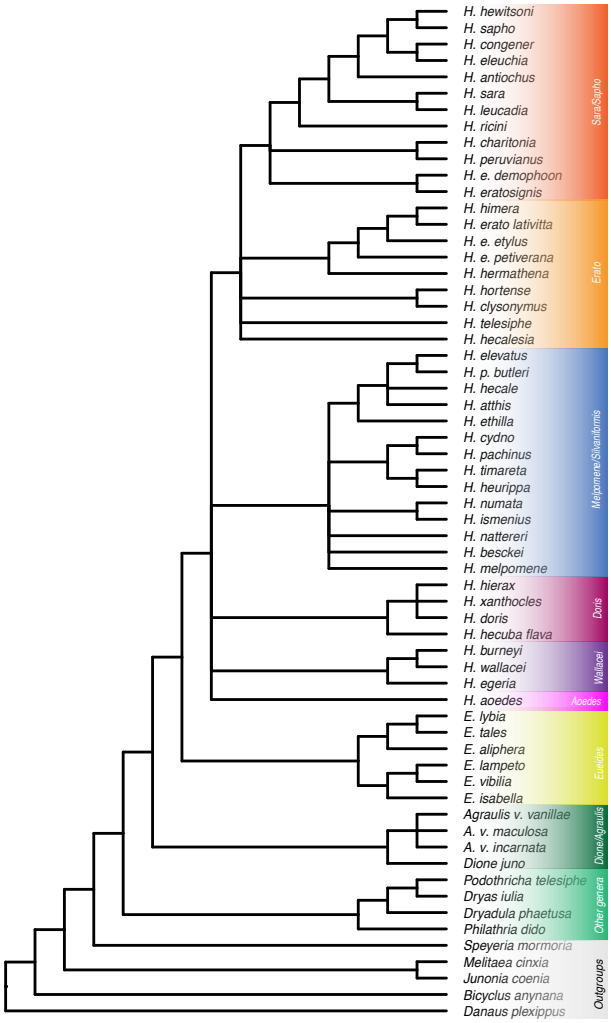

Fig. 4: TOB for butterflies setting $\alpha = 10^{-15}$ and $\beta = 0.95$.

We also present the first large-scale study of TOB reconstruction methods. Benchmarking our method (TOB-QMC) against the only other existing method TINNiK revealed that TOB-QMC was highly competitive in terms of tree accuracy but was also much much faster, as expected from the reduction in time complexity from $O(n^4k+n^5)$ to $O(n^3k)$. Because TOB-QMC and TINNiK were run using the same hypothesis tests (and significance thresholds) and thus were subject to the same model assumptions, this improvement in accuracy likely stems from TOB-QMC's two-stage design, which first aggregates signal across all four-taxon subsets and then applies targeted hypothesis tests to remove false positive edges. In contrast, TINNiK builds the tree based on the outcomes of applying hypothesis tests to all four-taxon subsets individually; by not leveraging global signal across the four-taxon subsets, TINNiK may be more sensitive to noisy or erroneous inputs (or mistakes from applying the T3 test to anomalous four-taxon subsets). On the other hand, TINNiK is deterministic, unlike TOB-QMC, whose output can change depending on the random seeding of the search algorithm (although we expect the overall trends in relative performance to be unaffected by this randomness as methods were run on 50 replicates per model condition).

A limitation of TOB-QMC and TINNiK (with the T3 test) is the assumption that networks are quartet-nonanomalous. Ané *et al.* proved that under the NMSC with Common Inheritance all binary, metric species networks are quartet-nonanomalous [4]. Such strong positive results do not exist for more general models of inheritance, including Independent Inheritance used in our simulations. To evaluate this issue, Ané *et al.* simulated networks under a birth-death-hybridization model and found quartet-anomalous networks to be

rare [4]. In contrast, many of the networks in our simulation study were quartet-anomalous, especially for higher network levels and shorter branch lengths. Binning networks based on whether they were anomalous or not, did not change the conclusions of our study, and perhaps surprisingly, the performance of both TOB-QMC and TINNiK sometimes improved for anomalous networks compared to nonanomalous networks for the same model condition, suggesting correlated properties of those networks made TOB reconstruction easier, although our study design limits the conclusions that can be drawn. The impact of quartet-anomalous networks on TOB reconstruction and the extent to which these model violations are biologically realistic is an important direction for future work.

Another obstacle to TOB reconstruction, with either TOB-QMC or TINNiK, is the burden on users to set significance thresholds: α, β. Because TOB-QMC stores the minimum p-value associated with each edge, users can efficiently evaluate the impact of different α and β settings offline without recomputing the "base tree". We recommend first examining the distribution of these edge-specific p-values to understand the strength of evidence across the tree, and then exploring a grid of α, β values to assess the stability of the inferred TOB across these threshold settings. Selecting appropriate values of $\alpha$and β could be challenging in practice, as the best choice may depends on data characteristics such as ILS level and gene tree estimation error, and the sampling method (e.g. 3f1a, heuristics search, versus exhaustive search), and thus no single setting is universally optimal. Edges with consistently low p-values provide strong evidence of network-like signal and are likely to be contracted across a wide range of α values, whereas edges with p-values near the decision threshold are more sensitive to hyperparameter choice. Accordingly, edges that are contracted only under extreme parameter values (e.g. a very large α, leading to aggressive over-contraction) should be interpreted cautiously, whereas edges that are consistently contracted across a broad range of settings provide stronger evidence of a blob signal. Likewise, stability of branches across parameter choices increases confidence in the inferred structure. This guidance is supported by our tuning experiments, which illustrate how accuracy can vary across threshold settings. Future work should evaluate the effectiveness hypothesis testing (both the tests developed by TINNiK as well as other approaches e.g. [29]) and develop procedures for automatically setting significance thresholds to relieve the burden on users.

It is also worth noting that biologists may use TOB-QMC for exploring whether there are signals of non-tree-like evolution in their data, which does not require fixed settings of α, β. Our re-analysis of phylogenomic data sets suggests that TOB-QMC-3f1a is fast and effective and for this purpose. Relatedly, the TOB-QMC framework is closely connected to how researchers analyze large biological data (i.e., building a species tree with ASTRAL and then annotating it was gene flow or introgression events). This approach to network reconstruction can be positively misleading, as recently shown by [11]. However, our theory supports such analyses, provided the results are interpreted carefully. Specifically, the ASTRAL tree should be interpreted as a refinement of the tree of blobs for network, so the gene flow edges indicate blob-like regions that need to be reconstructed with a network method.

During the time frame of our study, a new network reconstruction method, called Squirrel, was released [18]. Squirrel creates candidate TOBs, through edge contractions, which is similar to TOB-QMC, and then "refines" the blobs to produce a network. However, unlike TINNiK and TOB-QMC, Squirrel is not designed for coalescent-based reconstruction and has no statistical guarantees under the NMSC (it contracts branches based on split support from quarnets estimated from DNA sequences). Existing methods for network reconstruction under the NMSC also depend on TOB reconstruction; for example, NANUQ$^+$ [3] builds a TOB and then reconstructs the cycles, whereas InPhyNet [21] builds a TOB to perform subset decomposition within its divide-and-conquer framework. We therefore anticipate that TOB-QMC will greatly enable scalable network reconstruction under the NMSC; future work should explore this further.

4 Methods

To present TOB-QMC, we briefly review key background and terminology; see the Supplement for details.

4.1 Background and Terminology

Phylogenetic networks. A *directed* phylogenetic network N is a simple directed acyclic graph with the property that there exists a directed path from the root vertex to all other vertices. Leaf vertices are bijectively labeled by a set S of species. Non-leaf (i.e., internal) vertices are either speciation events or reticulation events. Speciation vertices have in-degree 1 and out-degree > 1; *reticulation nodes* vertices have in-degree > 1 and out-degree 1. A tree edge is an arc whose child is a tree vertex; all other arcs are called *reticulation edges.* A

directed network can be converted into a *semi-directed* network or an *undirected* network by suppressing the root and undirecting the tree edges or all edges, respectively. A phylogenetic tree is simply a phylogenetic network with no reticulation vertices. A tree is **displayed** by a directed or semi-directed network if it can be obtained by deleting only incoming edges to reticulation nodes but not their end points (Fig. 1A,D).

A bridge is an edge whose deletion increases the number of connected components of the graph. The maximal bridgeless subgraphs of an undirected network are called **blobs**. The degree of a blob is given by the number of bridges adjacent to it (an m-blob is a blob with degree m). Edges in blobs are **blob edges**, and all other edges (i.e., bridges) are **cut edges**. The **tree of blobs (TOB)** [14] can be computed from a network by (1) un-directing all edges, (2) contracting all blob edges, and then (3) suppressing all vertices with degree two (Fig. 1A–B). The contraction step results in each blob being collapsed into a single vertex, which may have degree greater than 3 (referred to as a polytomy). We say that a tree T refines the TOB if the TOB can be obtained by contracting edges in T [34]. Any tree displayed by a network is a refinement of the TOB; however, a refinement of the TOB need not be a display tree of the network (Fig. 1C).

The deletion of a vertex in a phylogenetic tree produces an m**-partition** of its species set S, denoted $M = P^1|P^2|...|P^m$; the same holds for an m-blob in a phylogenetic network. Likewise, the deletion of edge e in a tree (or the deletion of a cut edge e in a network) induces a **bipartition**, denoted $Bip(e) = A|B$. If the tree is binary, edge $e = (s, t)$ also induces **quadrapartition** and denoted $Quad(e) = A_1|A_2|B_2|B_2$, formed by deleting e along with its incident vertices s and t. Later we consider sampling 4-taxon subsets **around an edge**, meaning that two leaves are drawn from each block of its bipartition or one leaf is sampled from each block of a quadrapartition. A **quartet** q is an unrooted tree on four species. There are three possible quartet topologies denoted by their non-trivial bipartition: $a, b|c, d$ and $a, c|b, d$ and $a, d|b, c$, where $S = \{a, b, c, d\}$. For complete preliminaries; see Section 1.1 of the Supplement.

Network Multispecies Coalescent (NMSC). The NMSC is parameterized by a species network $\mathcal{N} = (N, \Theta)$, where N is a directed phylogenetic network and Θ is a set of numerical parameters (branch lengths and inheritance proportions). A gene genealogy can be generated under the NMSC by sampling one gene per species and tracing their ancestry backward in time until all gene lineages coalesce (i.e., trace back to a common ancestor); see Section 1.2 of the Supplement for modeling details. The network topology N governs the pool of potential ancestors. Although gene trees are not observed, they can be estimated from molecular sequence data. Our goal is to reconstruct the TOB of a species network from gene trees. A major challenge is that gene trees can differ from the display trees of the network due to **incomplete lineage sorting (ILS)**, which can occur when lineages enter a branch and then fail to coalesce, and due to **reticulations**, which allow lineages to trace back to different ancestral populations. We make standard assumptions about species networks (see Results).

Our first result (Theorem 1) further assumes networks are **relaxed quartet-nonanamolous**, which we now define. Let $\mathbb{P}_{\mathcal{N}}(T)$ denote the probability of an unrooted gene genealogy T under the NMSC given $\mathcal{N}$. This quantity can be computed for T restricted to four species $Y \subseteq S$, denoted $T|_Y$, using the induced subnetwork on Y (see the Section 1.1 of the Supplement).

Definition 1. A 4-taxon binary species network $\mathcal{N}$ on species set S is *relaxed quartet-nonanomalous* if it displays more than one quartet trees <u>or</u> if it displays exactly one quartet tree q_1^S and the following inequalities hold: $\mathbb{P}_{\mathcal{N}}(q_1^S) > \mathbb{P}_{\mathcal{N}}(q_2^S)$ and $\mathbb{P}_{\mathcal{N}}(q_1^S) > \mathbb{P}_{\mathcal{N}}(q_3^S)$ where q_2^S and q_3^S are the other two quartet trees on S. If $|S| > 4$, the above must hold for the induced subnetwork on every four-taxon subset $Y \subseteq S$, denoted $\mathcal{N}|_Y$.

The related Lemma is useful for proving Theorem 1.

Lemma 1 (Lemma 1 in [4])**.** A 4-taxon binary network N displays exactly one quartet if, and only if, it does not contain a 4-blob (also see Fig. S2 in the Supplement).

Although one might think that quartet-based species tree methods would be guaranteed to produce a display tree of the network if it is relaxed quartet-nonanonmalous, Dinh and Baños recently showed this is not the case [11].

Weighted Quartet Consensus (WQC). The popular species tree method ASTRAL-III [39] is a constrained dynamic programming for the NP-hard WQC problem [22]. The input is a set $\mathcal{T}$ of k unrooted gene trees, whose leaves are labeled by a set S of n taxa. The output is an unrooted binary tree T on S that maximizes the quartet score: $QS_{\mathcal{T}}(T) = \sum_{Y \in f_4(S)} w_{\mathcal{T}}(T|_Y)$, where $w_{\mathcal{T}}(q)$ is the number of trees in $\mathcal{T}$ that induce quartet q and $f_4(S)$ denotes all 4-taxon subsets of S, that is $f_4(S) = \{\{a, b, c, d\} : a, b, c, d \in$

$S, a \neq b \neq c \neq d\}$. ASTRAL does not explicitly compute quartets weight, which are referred to as **quartet concordance factors (qCFs)** when grouped by their associated 4-taxon subset.

TINNiK. The TINNiK is the only existing method for consistent TOB reconstruction under the NMSC. It is based on the following Lemma.

Lemma 2. Let N be a network on species set S and T_N be its tree of blobs. For any subset $Y \subseteq S$ of four taxa, $T_N|_Y$ is a *quartet tree* (i.e., a binary tree on four taxa, as opposed to the star tree on four taxa), if and only if, $N|_Y$ is not a 4-blob.

See proof in Supplement. An important consequence is that the TOB is identifiable if given an oracle for determining whether the evolutionary relationship on four taxa is a 4-blob or not. The authors of TINNiK develop a hypothesis testing procedure for this very purpose and show that its application to all four-taxon subsets enables statistically consistent TOB reconstruction (Theorem 4 in [2]).

4.2 WQC is a TOB refinement almost surely as number of gene trees increases

Recall that our method for TOB reconstruction begins by seeking a refinement of the TOB, using a quartet-based method. To motivate this approach, we show that WQC is a TOB refinement almost surely, as the number of gene trees increases. Our proof relies on the following Lemma.

Lemma 3. Let $\mathcal{N} = (N, \Theta)$ be a species network on S that satisfies standard assumptions and is relaxed quartet-nonanomalous. Let T_N denote the tree of blobs for N. Then, for any four-taxon subset $Y \subseteq S$ such that T_N induces a binary quartet tree on Y, denoted q_1^Y, the following inequalities hold: (1) $\mathbb{P}_{\mathcal{N}}(q_1^Y) > \mathbb{P}_{\mathcal{N}}(q_2^Y)$ and (2) $\mathbb{P}_{\mathcal{N}}(q_1^Y) > \mathbb{P}_{\mathcal{N}}(q_3^Y)$, where q_2^Y, q_3^Y denote the two alternative quartet topologies on Y.

Proof. Let $Y = \{a, b, c, d\}$, and let $\mathcal{Q}$ be the set of quartets displayed by $\mathcal{N}|_Y$. If $|\mathcal{Q}| = 1$, the inequalities hold by the definition of relaxed quartet-nonanomalous networks. If $|\mathcal{Q}| > 1$, then by Lemma 1, then the taxa in $\mathcal{N}|_Y$ is a 4-blob, and by Lemma 2, $T_N|_Y$ is a star, not a binary quartet tree. This is a contradiction. □

Now we give our main result.

Theorem 1. Let $\mathcal{N} = (N, \Theta)$ be a binary metric species network on S that satisfies standard assumptions and is relaxed quartet-nonanomalous. Assume the TOB for N is not a star (otherwise all trees on S refine the TOB so the result is trivial). Let T be an unrooted binary tree on S that is <u>not</u> a TOB refinement. Let $\mathcal{T}$ be a set of k gene trees generated under $NMSC$ given $\mathcal{N}$. Then, for any T and any $\epsilon > 0$, there exists a $K > 0$ and a (binary) TOB refinement B_N^* such that, given $k \geq K$ gene trees, $QS_{\mathcal{T}}(B_N^*) > QS_{\mathcal{T}}(T)$ with probability at least $1 - \epsilon$. Put simply, an optimal solution to WQC is a refinement of the TOB for N almost surely, as the number of gene trees increases.

Theorem 1 follows from the two claims. The first claim is that the observed qCFs $w_{\mathcal{T}}(q_i^Y)/k$ converge to their expected values $F_{\mathcal{N}}^*(q_i^Y) = \mathbb{P}_{\mathcal{N}}(q_i^Y)$ for all 4-taxon subsets $Y \in f_4(S)$ and for all quartet topologies $i \in \{1, 2, 3\}$. This claim, which is standard in the phylogenetics literature, means that for any $\epsilon > 0$ and any $\delta > 0$, there exists a $K > 0$ such that, given $k \geq K$ gene trees, $|w_{\mathcal{T}}(q_i^Y)/k - F_{\mathcal{N}}^*(q_i^Y)| < \delta$ for all $i \in \{1, 2, 3\}$ and $Y \in f_4(S)$ with probability at least $1 - \epsilon$. The second claim is that when the absolute error in qCFs is bounded by δ^*, there exists a TOB refinement B_N^* such that $QS_{\mathcal{T}}(B_N^*) > QS_{\mathcal{T}}(T)$ where T is any non-refinement of T_N. The proof of claim 2 (and the setting of δ^*) is given in the Supplement (Lemma 4); it relies heavily on the next result.

Lemma 5. Assume the same conditions and notation as Theorem 1. Assume the error in all qCFs is zero (i.e., they equal their expected values under the NMSC given $\mathcal{N}$). Then, there exists a (binary) TOB refinement with higher quartet score than any (binary) non-refinement tree T on S.

Proof. As previously mentioned, restricting the TOB, denoted T_N, to any 4-tuple of species $Y \in f_4(S)$, yields either a quartet or a star. Let $\mathcal{S}_{tree}(T_N)$ and $\mathcal{S}_{star}(T_N)$ denote the 4-taxon subsets in $f_4(S)$ partitioned based on those two cases, respectively. Now the quartet score of an unrooted binary tree t can be written as

$$QS(t) = QS^{tree}(t) + QS^{star}(t) = \sum_{Y \in \mathcal{S}_{tree}(T_N)} F_{\mathcal{N}}^*(t|_Y) + \sum_{Y \in \mathcal{S}_{star}(T_N)} F_{\mathcal{N}}^*(t|_Y) \tag{1}$$

as we assumed qCFs equaled their expected values. For an example of tree and star parts; see Figure 5.

Tree part of quartet score. We first consider maximizing the "tree part" of the quartet score. Observe that for a binary tree T on S that is NOT a refinement of T_N, we have

$$QS^{tree}(T|_N) = \sum_{Y \in \mathcal{S}_{tree}(T_N)} F^*_{\mathcal{N}}(T_N|_Y) > \sum_{Y \in \mathcal{S}_{tree}(T_N)} F^*_{\mathcal{N}}(T|_Y) = QS^{tree}(T) \tag{2}$$

by Lemma 3. Thus, the tree part of the quartet score is uniquely maximized by T_N.

Star part of quartet score. We now consider maximizing the "star part" of the quartet score. First, we partition $\mathcal{S}_{star}(T_N)$ based on the observation that $Y = \{a, b, c, d\} \in \mathcal{S}_{star}(T_N)$ implies the existance of a unique vertex v in T_N with degree $m > 3$ whose deletion induces an m-partition $Part(v) = P_1|P_2|\ldots|P_m$ such that $a \in P_i, b \in P_j, c \in P_s, d \in P_t$ with $P_i \neq P_j \neq P_s \neq P_t$ (because an induced quartet has exactly two internal vertices of degree three and an induced star has exactly one internal vertex degree four). This means we can further partition the sum to get

$$QS^{star}(t) = \sum_{v \in V_{>3}(T_N)} QS^v(t) = \sum_{v \in V_{>3}(T_N)} \sum_{Y \in \mathcal{S}_v(T_N)} F^*_{\mathcal{N}}(t|_Y) \tag{3}$$

where $V_{>3}(T_N)$ is the set of vertices in T_N with degree greater than 3 and $\mathcal{S}_v(T_N) \subseteq \mathcal{S}_{star}(T_N)$ is the set of 4-taxon subsets around v. More precisely, let $f_4(Part(v))$ be the set made from all ways of selecting 4 distinct blocks from $Part(v)$, that is $f_4(Part(v)) = \{P_i, P_j, P_k, P_l : P_i, P_j, P_k, P_l \in Part(v), P_i, \neq P_j \neq P_k \neq P_l\}$, and let $f_{cover}(\{P_i, P_j, P_k, P_l\})$ be the set made from all ways of selecting one taxa from each of the input blocks, that is, $f_{cover}(\{P_i, P_j, P_k, P_l\}) = \{\{a, b, c, d\} : a \in P_i, b \in P_j, c \in P_k, d \in P_l\}$. Then, we can write $\mathcal{S}_v(T_N) = \cup_{P \in f_4(Part(v))} \cup_{Y \in f_{cover}(P)} Y$. Now it suffices to show that there exists a refinement of T_N, denoted B^*_N, such that $QS^v(B^*_N) > QS^v(T)$ for all $v \in V_{>3}(T_N)$.

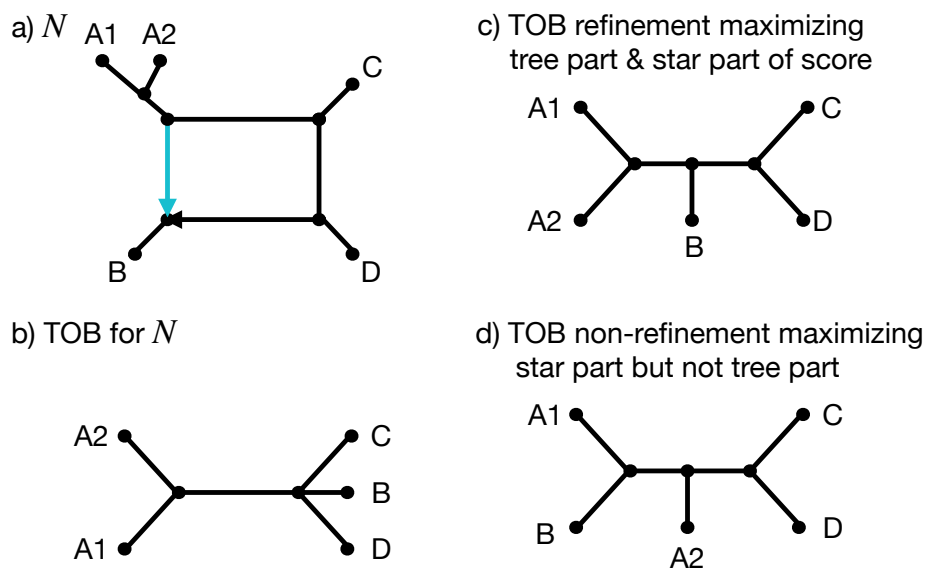

Fig. 5: **a)** A 5-taxon network N, referred to as a good diamond in [32]. **b)** TOB for N has one polytomy v that induces partitions $Part(v) = \{P_1 = \{A_1, A_2\}, P_2 = \{B\}, P_3 = \{C\}, P_4 = \{D\}\}$. Thus, the "star part" is computed from 4-tuples: A_1, B, C, D and A_2, B, C, D (note that A_1 and A_2 are exchangeable when drawing 4-taxon subsets that cover $Part(v)$ and also have the same qCFs). The "tree part" of the quartet score is computed from 4-tuples: A_1, A_2, B, C and A_1, A_2, B, D and A_1, A_2, C, D. **c)** A TOB refinement for N that maximizes the star score assuming $F^*_{\mathcal{N}}(A_1, B|C, D) > \max\{F^*_{\mathcal{N}}(A_1, C|B, D), F^*_{\mathcal{N}}(A_1, C|B, D)\}$. TOB refinements maximize the tree score. **d)** A TOB non-refinement that also maximizes that star score because it induces the same quartets on A_1, B, C, D and A_2, B, C, D as subfigure (c). Howover, TOB non-refinements do not maximize the tree score.

We now focus on the part of the score associated with a single vertex $v \in V_{>3}(T_N)$, which we call a blob vertex as it reflects the contraction of blob edges into a polytomy. Let $T_N(v)$ denote the tree obtained by contracting all internal edges in TOB T_N except those that are incident to v, that is $E(T_N(V)) = \{e : Bip(e) = P_i|S \setminus P_i, P_i \in Part(v)\}$. We ***claim*** that there exists a refinement of $T_N(v)$, denoted $B^*_N(v)$, that (1) maximizes $QS^v(\cdot)$ and (2) does not break apart blocks of $Part(v)$, meaning that each of its non-trivial bipartitions $Z|W$ satisfies $P_i \cap Z = P_i$ and $P_i \cap W = \emptyset$ (or vice versa) for all $P_i \in Part(v)$. Property (2) is important so that $\{B^*_N(v) : v \in V_{>3}(T_N)\}$ are compatible with each other and also T_N. Our proof of this claim is based on the following observation: any 4-tuple of species covering $P = \{P_i, P_j, P_k, P_l\}$ has the same expected qCFs, as a lineage sampled from any species in P_i must enter the blob in $\mathcal{N}$, associated with v in T_N, from the same cut edge (and similarly for P_j, P_k, P_l), and only after the four sampled lineages enter this blob can they coalesce; thus, the expected qCFs are fully dictated by the blob and do not depend on which particular taxon was selected (i.e., they selected taxa are exchangeable); see [5] for details.

We are now ready to prove our claim by contradiction, assuming all trees that maximize $QS^{star}_v(\cdot)$ break apart blocks of $Part(v)$, thus violating property (2). Let T be one such tree. Then, there exists at least one block P_i of $Part(v)$ such that there is no edge in T that splits all of the taxa in P_i from all other taxa. Let $QS^v_x(T)$ denote the part of the quartet score of T from blob vertex v that is associated with taxon x, specifically:

$$QS^v_x(t) = \sum_{Y \in \{Z \in \mathcal{S}_v(T_N) : Z \cap \{x\} \neq \emptyset\}} F^*_{\mathcal{N}}(t|_Y). \tag{4}$$

Next, select taxon $a \in P_i$ such that $QS^v_a(T) \geq QS^v_b(T)$ for all $b \in P_i \setminus \{a\}$. For each selection of b, there are two possible cases: $QS^v_a(T) = QS^v_b(T)$ <u>or</u> $QS^b_a(T) > QS^v_b(T)$. For the first case, consider the following actions: detach b, contract the resulting vertex of degree 2, and then re-attach b so that it is sister to a. In particular b should be attached by sub-diving the edge incident to a, denoted (a, u), if a's current sister(s) is not in P_i; otherwise, b should be attached directly to vertex u. The resulting tree has the same blob v score as T (i.e., $QS^v(T) - QS^v_b(T) + QS^v_a(T) = QS^v(T)$) because for any $Y_b = \{b, x, y, z\} \in \mathcal{S}_v(T_N)$, there exists a $Y_a = \{a, x, y, z\} \in \mathcal{S}_v(T_N)$ with the same expected qCFs (i.e., $F^*(q_i^{Y_b}) = F^*(q_i^{Y_a})$ for all $i \in \{1, 2, 3\}$), by our construction of $\mathcal{S}_v(T_N)$ and exchangeability (also note that a, b cannot appear in the same 4-taxon subset from $\mathcal{S}_v(T_N)$ by construction). For the second case, the same actions can be taken, but now the resulting tree has a blob v score greater than T (i.e., $QS^v(T) - QS^v_b(T) + QS^v_a(T) > QS^v(T)$). Repeating this process for all $b \in P_i \setminus \{a\}$ as well as all blocks $P_i \in Part(v)$ yields a tree that has blob v score greater than or equal to $QS^v(T)$ and satisfies property (2), which is a contradiction. Lastly observe that not only does $B^*_N(v)$ exist but it must be binary after removing all but one taxon from each block of $Part(v)$. This again follows by contradiction as any refinement would increase $QS^v(\cdot)$, after reattaching the removed taxa in the natural way, due to the fact that all qCFs are strictly positive given our assumptions. It is also worth noting that $B^*_N(v)$ is not necessarily a unique maximizer, and moreover, trees that violate property (2) can maximize $QS^v(\cdot)$.

Putting the tree and star parts together. By bipartition compatibility, there exists a compatibility supertree B^*_N for $\{T_N\} \cup \{B^*_N(v) : v \in V_{>3}(T_N)\}$ (see Chapters 3 and 7 in [33]). By construction, B^*_N is a binary TOB refinement that has a greater quartet score than any binary TOB non-refinement because even if the two trees achieve the same score on the star part $QS_{star}(\cdot)$, B^*_N must achieve a greater score on the tree part $QS_{tree}(\cdot)$ (which is why we assume the TOB is not a star). This gives us our result. □

A corollary of Theorem 1 is that ASTRAL (version 3 or earlier) returns a TOB refinement almost surely, as the number of gene trees increases. This is because ASTRAL solves WQC within a constrained search space, defined by the bipartitions induced by the input gene trees, and all bipartitions have probability strictly greater than zero under the NMSC and thus will appear as the number of gene trees increases, as noted by [23]. This motivates the use of ASTRAL, and other WQC heuristics, in seeking a TOB refinement.

4.3 False positives can be identified with a linear number of hypothesis tests

A TOB refinement B_N contains every edge in the TOB T_N plus additional edges that are not in T_N; these edges are false positives (FPs). To obtain a consistent estimator of the TOB, we need to detect (and then contract) all FP branches in a TOB refinement almost surely, as the number of gene trees increases. The following observation is useful for this purpose.

Observation 1. Let N be a network on taxon set S, T_N be its tree of blobs, and B_N be a refinement of T_N. If an edge e is in B_N but not in T_N (i.e., e is a FP edge), there exists at least one subset $Y \subseteq S$ of four taxa around e such that $N|_Y$ is a 4-blob.

This obveration is easy to see because otherwise e is a cut edge and must be in T_N and motivates TINNiK's hypothesis testing procedure for detecting 4-blobs, which we will initially treat as oracle that is always correct. Critically, whether the induced subnetwork is a 4-blob depends on the choice of Y. Taking Fig. 1 as an example, the edge inducing bipartition $C, D, E, F, G|A, B, H, I, J, K$ is an FP. If we select sample $Y = \{K, J, F, G\}$, then $N|_{Y_1}$ is not a 4-blob; $N|_Y$ is a 4-blob if we sample $Y = \{A, B, C, F\}$ (also see Fig. S3 in the Supplement). Unlike TINNiK, our approach just needs to find and test one 4-taxon subset Y around e such that the $N|_Y$ is a 4-blob, if such a Y exists, because as the number of gene trees increases, the edge will be contracted almost surely. A naive approach that tests all 4-taxon subsets around each edge yields a consistent estimator of the TOB; however, it has time complexity $O(n^4k + n^5)$, which is the same as TINNiK. Thus, the challenge is detect FP branches efficiently, without **exhaustive search**. The following lemma gives conditions that when satisfied ensure successful detection of FP edges.

Lemma 6. Let $\mathcal{N} = (N, \Theta)$ be a directed binary species network on taxon set S. Suppose that $\mathcal{B}$ is a blob in N that is contracted into vertex $v_{\mathcal{B}}$ in its TOB T_N. Now suppose that e is an FP edge in a TOB refinement B_N that refines $v_{\mathcal{B}}$. Then, any edge e will be successfully detected as an FP during hypothesis testing if we select and test at least one four-taxon subset $\{a_1, a_2, a_3, a_*\}$ around e that satisfies the following conditions:

1. Each taxon is selected from a different block of the partition induced by $v_{\mathcal{B}}$ in T_N, that is, for $Part(v_{\mathcal{B}}) = P_1|P_2|\ldots|P_m$, we have $a_1 \in P_i, a_2 \in P_j, a_3 \in P_s, a_* \in P_t$ where $i \neq j \neq s \neq t$
2. One of a_1, a_2, a_3, a_* is from a block of the $Part(v_{\mathcal{B}})$ that attaches to $\mathcal{B}$ via a cut edge that is incident to a reticulation vertex in $\mathcal{B}$

as they guarantee $\{a_1, a_2, a_3, a_*\}$ form a 4-blob in N.

See proof in Supplement. There always exists a reticulation node in $\mathcal{B}$ that is incident to a cut edge of $\mathcal{B}$ (Lemma 7 in the Supplement); therefore, we just need to design efficient sampling algorithms that is guaranteed to find one 4-taxon subset that meets these two conditions.

Requiring four-taxon subsets to draw exactly one taxon from each block of the quadrapartition induced by edge e ensures condition (1) is met, and including every taxon in S in at least one four-taxon subset ensures we find a four-taxon subset that satisfies condition (2). This leads us to propose an algorithm that selects one taxon from each block of the quadrapartition, fixes the selection for three of the blocks, and then tests all ways of selecting a different taxa from the fourth block (and repeating so that the "alter" step is performed for every block of the quadrapartition). We refer to this quadraparition search algorithm as **3-fix, 1-alter (3f1a)**; see Section 1.5 in the Supplement.

It is easy to see that 3f1a meets the conditions above and thus returns the true TOB when given an oracle for detecting 4-blobs (also see Theorem 2 in the Supplement). In practice, we do not have such an oracle and instead apply TINNiK's hypothesis tests to qCFs computed from gene trees, with fixed significance thresholds α, β. In this setting, the consistency our approach directly follows from Theorem 4 in [2]. The idea is that as the number of gene trees goes to infinity, the p-values from applying the tree-test to four-taxon subsets that form 4-blobs in the induced subnetwork converge to 0 and otherwise are bounded away from 0, so when the number of gene trees is sufficiently large, there exists an α that enables the tree-tests to return the correct answer for all four-taxon subsets tested with high probability. Unlike TINNiK, which tests all 4-taxon subsets, our 3f1a algorithm tests just $O(n)$ 4-taxon subsets around each edge, so $O(n^2)$ in total. The qCFs for any 4-taxon subset can be computed from gene trees on the fly in $O(k)$ time using the well-known binary lifting algorithm to retrieve the LCA of any pair of leaves in a gene tree in $O(1)$ time. This requires a precomputation step with time and storage complexity of $O(n \log(n)k)$, meaning the total time complexity of applying the 3f1a algorithm to every branch is $O(n^2k + n \log(n)k)$.

As we will show, the 3f1a search algorithm appears to work well for biological data but not as well for simulated data, potentially because it is difficult to set α when qCFs deviate from their expected values, which occurs in practice, as they must be estimated from finite numbers of gene trees, which may have estimation error, missing taxa, etc. In this realistic setting, there can be multiple 4-taxon subsets around an edge that are 4-blobs and some may yield lower p-values than others, so more aggressively searching to find a lower p-value can increase robustness to the choice of α. Sampling around a bipartition (rather than a quadrapartition) allows greater exploration of 4-taxon subsets; however, not every subset sampled will satisfy

condition (1). Indeed, there are cases where $\Omega(n^2)$ 4-taxon subsets must be sampled uniformly at random to ensure that condition (1) is met by one subset with high probability, as discussed in the Supplement. For any deterministic sampling schematics around a bipartition, at least $\Omega(n^2)$ 4-taxons subsets are needed to sample to meet condition (1); otherwise, an adaptive adversary can always ensure that none of our sampled 4-taxons subsets meet condition (1) since there are at least $\Omega(n^2)$ 4-taxon subsets that do not meet condition (1). Ultimately, we settled on a 3f1a-like algorithm that searches around a bipartition until a predetermined number of tests have been performed (see Algorithm 2 in the Supplement). By default, we bound the number of tests by $O(n^2)$ so that the time complexity of FP detection $O(n^3k + n\log(n)k)$ does not exceed the time of seeking a TOB refinement (note that TREE-QMC has time complexity $O(n^3k)$, with some assumptions about the divide-and-conquer, and ASTRAL-III has time complexity $O(n^{1.726}k^{1.726}D)$, where $D = O(nk)$ is the sum of degrees of all unique nodes in the input gene trees). Lastly, we note that combining this bipartition search with 3f1a yields a consistent method and does not increase the time complexity.

Acknowledgments. We are grateful to the anonymous RECOMB reviewers for their feedback that greatly improved the quality of our work. This material is based upon work supported by the U. S. National Science Foundation (NSF) under Grant No. 2441458 (to EKM) as well as the State of Maryland. The authors thank Francesco Cicconardi for sharing gene trees for the analysis of *Heliconiinae* butterflies and helpful discussions. The authors thank Claudia Solís-Lemus and Nathan Kolbow for motivation for studying trees of blobs and discussions regarding phylogenetic network simulations, which were initiated at the NSF-funded Institute for Computational and Experimental Research in Mathematics. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsors.

Code and Data Availability. TOB-QMC is implemented within TREE-QMC. The source code is publicly available on Github at `https://github.com/molloy-lab/TREE-QMC`.

References

1. Allman, E.S., Baños, H., Mitchell, J.D., Rhodes, J.A.: The tree of blobs of a species network: identifiability under the coalescent. Journal of Mathematical Biology **86**(1), 10 (2023)
2. Allman, E.S., Baños, H., Mitchell, J.D., Rhodes, J.A.: Tinnik: inference of the tree of blobs of a species network under the coalescent model. Algorithms for Molecular Biology **19**(1), 23 (2024)
3. Allman, E.S., Baños, H., Rhodes, J.A., Wicke, K.: NANUQ+: A divide-and-conquer approach to network estimation
4. Ané, C., Fogg, J., Allman, E.S., Baños, H., Rhodes, J.A.: Anomalous networks under the multispecies coalescent: theory and prevalence. Journal of Mathematical Biology **88**(3), 29 (2024). https://doi.org/10.1007/s00285-024-02050-7
5. Baños, H.: Identifying species network features from gene tree quartets under the coalescent model. Bulletin of Mathematical Biology **81**(2), 494–534 (2019). https://doi.org/10.1007/s11538-018-0485-4
6. Bjorner, M.B., Molloy, E.K., Dewey, C.N., Solís-Lemus, C.: Detectability of varied hybridization scenarios using genome-scale hybrid detection methods. Bulletin of the Society of Systematic Biologists **3**(1) (2024). https://doi.org/10.18061/bssb.v3i1.9284
7. Blischak, P.D., Chifman, J., Wolfe, A.D., Kubatko, L.S.: HyDe: A python package for genome-scale hybridization detection. Systematic Biology **67**(5), 821–829 (03 2018). https://doi.org/10.1093/sysbio/syy023
8. Bossert, S., Murray, E.A., Pauly, A., Chernyshov, K., Brady, S.G., Danforth, B.N.: Gene tree estimation error with ultraconserved elements: An empirical study on Pseudapis bees. Systematic Biology **70**(4), 803–821 (2020). https://doi.org/10.1093/sysbio/syaa097
9. Cao, Z., Zhu, J., Nakhleh, L.: Empirical Performance of Tree-Based Inference of Phylogenetic Networks. In: Huber, K.T., Gusfield, D. (eds.) 19th International Workshop on Algorithms in Bioinformatics (WABI 2019). Leibniz International Proceedings in Informatics (LIPIcs), vol. 143, pp. 21:1–21:13. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany (2019). https://doi.org/10.4230/LIPIcs.WABI.2019.21
10. Cicconardi, F., Milanetti, E., Pinheiro de Castro, E.C., Mazo-Vargas, A., Van Belleghem, S.M., Ruggieri, A.A., Rastas, P., Hanly, J., Evans, E., Jiggins, C.D., Owen McMillan, W., Papa, R., Di Marino, D., Martin, A., Montgomery, S.H.: Evolutionary dynamics of genome size and content during the adaptive radiation of heliconiini butterflies. Nature Communications **14**(1), 5620 (Sep 2023). https://doi.org/10.1038/s41467-023-41412-5
11. Dinh, V., Baños, H.: Misspecification strikes: Astral can mislead in the presence of hybridization, even for nonanomalous scenarios. Molecular Biology and Evolution **42**(3), msaf049 (2025). https://doi.org/10.1093/molbev/msaf049
12. Felsenstein, J.: Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution **17**(6), 368–376 (1981). https://doi.org/10.1007/BF01734359

13. Fogg, J., Allman, E.S., Ané, C.: PhyloCoalSimulations: A simulator for network multispecies coalescent models, including a new extension for the inheritance of gene flow. Systematic Biology **72**(5), 1171–1179 (2023). https://doi.org/10.1093/sysbio/syad030
14. Gusfield, D., Bansal, V., Bafna, V., Song, Y.S.: A decomposition theory for phylogenetic networks and incompatible characters. Journal of Computational Biology **14**(10), 1247–1272 (2007). https://doi.org/10.1089/cmb.2006.0137
15. Han, Y., Molloy, E.K.: Improving quartet graph construction for scalable and accurate species tree estimation from gene trees. Genome Research **33**(7), 1042–1105 (2023). https://doi.org/10.1101/gr.277629.122
16. 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). https://doi.org/10.1093/sysbio/syaf009
17. Hejase, H.A., VandePol, N., Bonito, G.M., Liu, K.J.: Fastnet: Fast and accurate statistical inference of phylogenetic networks using large-scale genomic sequence data. In: Blanchette, M., Ouangraoua, A. (eds.) Comparative Genomics. pp. 242–259. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-00834-5_14
18. Holtgrefe, N., Huber, K.T., van Iersel, L., Jones, M., Martin, S., Moulton, V.: Squirrel: Reconstructing semi-directed phylogenetic level-1 networks from four-leaved networks or sequence alignments. Molecular Biology and Evolution **42**(4), msaf067 (2025). https://doi.org/10.1093/molbev/msaf067
19. Justison, J.A., Solis-Lemus, C., Heath, T.A.: SiPhyNetwork: An R package for simulating phylogenetic networks. Methods in Ecology and Evolution **14**(7), 1687–1698 (2023). https://doi.org/10.1111/2041-210X.14116
20. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K., Von Haeseler, A., Jermiin, L.S.: Modelfinder: fast model selection for accurate phylogenetic estimates. Nature methods **14**(6), 587–589 (2017)
21. Kolbow, N., Kong, S., Solís-Lemus, C.: A method for massively scalable phylogenetic network inference. bioRxiv (2025). https://doi.org/10.1101/2025.05.05.652278
22. Lafond, M., Scornavacca, C.: On the weighted quartet consensus problem. Theoretical Computer Science **769**, 1–17 (2019). https://doi.org/10.1016/j.tcs.2018.10.005
23. 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). https://doi.org/10.1093/bioinformatics/btu462
24. Molloy, E.K., Durvasula, A., Sankararaman, S.: Advancing admixture graph estimation via maximum likelihood network orientation. Bioinformatics **37**(Supplement_1), i142–i150 (2021). https://doi.org/10.1093/bioinformatics/btab267
25. Molloy, E.K., Warnow, T.: Statistically consistent divide-and-conquer pipelines for phylogeny estimation using NJMerge. Algorithms for Molecular Biology **14**(1), 14 (2019). https://doi.org/10.1186/s13015-019-0151-x
26. Molloy, E.K., Warnow, T.: TreeMerge: a new method for improving the scalability of species tree estimation methods. Bioinformatics **35**(14), i417–i426 (2019). https://doi.org/10.1093/bioinformatics/btz344
27. Nguyen, L.T., Schmidt, H.A., Von Haeseler, A., Minh, B.Q.: Iq-tree: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular biology and evolution **32**(1), 268–274 (2015)
28. One Thousand Plant Transcriptomes Initiative: One thousand plant transcriptomes and the phylogenomics of green plants. Nature **574**(7780), 679–685 (2019). https://doi.org/10.1038/s41586-019-1693-2
29. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., Reich, D.: Ancient admixture in human history. Genetics **192**(3), 1065–1093 (2012). https://doi.org/10.1534/genetics.112.145037
30. Rhodes, J.A., Baños, H., Mitchell, J.D., Allman, E.S.: MSCquartets 1.0: quartet methods for species trees and networks under the multispecies coalescent model in R. Bioinformatics **37**(12), 1766–1768 (2020). https://doi.org/10.1093/bioinformatics/btaa868
31. Saitou, N., Nei, M.: The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution **4**(4), 406–425 (1987)
32. Solís-Lemus, C., Ané, C.: Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLOS Genetics **12**(3), 1–21 (2016). https://doi.org/10.1371/journal.pgen.1005896
33. Warnow, T.: Computational Phylogenetics: An Introduction to Designing Methods for Phylogeny Estimation. Cambridge University Press, Cambridge, United Kingdom (2017)
34. Xu, J., Ané, C.: Identifiability of local and global features of phylogenetic networks from average distances. J. Math. Biol. **86**, 12 (2023). https://doi.org/10.1007/s00285-022-01847-8
35. Yu, Y., Degnan, J.H., Nakhleh, L.: The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLOS Genetics **8**(4), 1–10 (04 2012). https://doi.org/10.1371/journal.pgen.1002660
36. Yu, Y., Dong, J., Liu, K.J., Nakhleh, L.: Maximum likelihood inference of reticulate evolutionary histories. Proceedings of the National Academy of Sciences **111**(46), 16448–16453 (2014). https://doi.org/10.1073/pnas.1407950111

37. Yu, Y., Nakhleh, L.: A maximum pseudo-likelihood approach for phylogenetic networks. BMC Genomics **16 Suppl 10**(S10), S10 (2015). https://doi.org/10.1186/1471-2164-16-S10-S10
38. Yu, Y., Ristic, N., Nakhleh, L.: Fast algorithms and heuristics for phylogenomics under ils and hybridization. BMC Bioinformatics **14**(15), S6 (2013). https://doi.org/10.1186/1471-2105-14-S15-S6
39. Zhang, C., Rabiee, M., Sayyari, E., Mirarab, S.: ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics **19**(6), 153 (2018). https://doi.org/10.1186/s12859-018-2129-y
40. Zhu, J., Liu, X., Ogilvie, H.A., Nakhleh, L.K.: A divide-and-conquer method for scalable phylogenetic network inference from multilocus data. Bioinformatics **35**(14), i370–i378 (2019). https://doi.org/10.1093/bioinformatics/btz359