Fast and accurate tree of blobs reconstruction under the network multispecies coalescent

Junyan Dai[⋆,1] Yunheng Han[⋆,1] and Erin K. Molloy[1,2,*]

[1] *Department of Computer Science, University of Maryland, College Park, 20742, USA*

[2] *University of Maryland Institute for Advanced Computer Studies, College Park, MD 20740*

[⋆] *Authors are equal contributors*

[*] *Corresponding author; Email: ekmolloy@umd.edu*

November 5, 2025

Abstract

Gene flow between species or populations is an important force in evolution, modeled by the network multispecies coalescent. Phylogenetic network reconstruction 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, to perform subset decomposition. TOB reconstruction is challenging in its own right, with the leading 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. In prior work, we introduced 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 estimate a tree using TREE-QMC, a fast quartet-based species tree method with time complexity $O(n^3k)$. For step (2), we contract edges based on the same hypothesis tests as TINNiK, which are applicable to subsets of four taxa. On simulated data sets, our method, referred to as TOB-QMC, was faster and more accurate than TINNiK.

Here, we provide the theoretical justification for our framework. For step (1), we show that Weighted Quartet Consensus converges to a TOB refinement with high probability, as the number k of gene trees tends to infinity, motivating the use of fast quartet-based methods for species tree estimation such as TREE-QMC or ASTRAL. ASTRAL, in particular, is statistically consistent in this context. For step (2), we show that sampling just $O(n)$ 4-taxon subsets around each edge is sufficient to ensure that false positive edges in a TOB refinement are contracted with high probability, as k tends to infinity. Thus, our framework enables statistically consistent TOB estimation and is fast, with asymptotic runtime dominated by tree reconstruction in step (1).

1 Introduction

Gene flow between species or populations is an important force in evolution, modeled by the network multispecies coalescent [30, 9]. Phylogenetic network reconstruction under this model is notoriously challenging, with the leading methods scaling to just tens of species [16]. 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, to perform subset decomposition [16]. TOB reconstruction is challenging in its own right, with the leading method TINNiK [2] having time complexity $O(n^5 + n^4k)$, where k is the number of input gene trees and n is the number of species. In prior work [12], we introduced 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 estimate a tree using TREE-QMC [13], a fast quartet-based species tree method with time complexity $O(n^3k)$. For step (2), we contract edges based on the same hypothesis tests as TINNiK, which are applicable to subsets of four taxa [1]. On simulated data sets, our method, referred to as TOB-QMC, was faster and more accurate than TINNiK.

Here, we provide the theoretical justification for this framework. For step (1), we show that Weighted Quartet Consensus [18] converges to a TOB refinement with high probability, as the number k of gene trees tends to infinity, motivating the use of fast quartet-based methods for species tree estimation such as TREE-QMC or ASTRAL [32]. In particular, ASTRAL (version 3 or lower) is statistically consistent in this context. For step (2), we show that sampling just $O(n)$ 4-taxon subsets around each edge is sufficient to ensure that false positive edges in a TOB refinement are contracted with high probability, as k tends to infinity. Thus, our framework enables statistically consistent TOB estimation and is fast, with asymptotic runtime dominated by tree reconstruction in step (1).

The remainder of this paper is organized as follows. We begin by reviewing terminology for phylogenetic networks in Section 2 as well as the network multispecies coalescent (NMSC) model in Section 3. Our model assumptions are detailed in Section 4 and the Weighted Quartet Consensus problem is defined in Section 5), along with related results. Our results are presented in Sections 6 and 7.

2 Preliminaries

We follow standard definitions for phylogenetic networks; see [4] for review. A *phylogenetic network* N is a tuple (n, S, ϕ), where n is a graph, S is a set of taxon (i.e., species) labels, and ϕ is function, mapping leaves (i.e., vertices with degree exactly 1) in n to labels in S. We typically assume that ϕ is a bijection and thus refer to leaves and labels interchangeably for simplicity. We also do not make a distinction between a phylogenetic network N and its graph n, instead saying that N is a phylogenetic network on taxon set S and denoting its edge, vertex, and leaf set as $E(N)$, $V(N)$, and $L(N)$, respectively. A phylogenetic network is **binary** if all non-leaf vertices have degree 3; additionally, the root (i.e., a special vertex with in-degree 0) must have degree 2 if the network is directed, which we now define.

Directed and semi-directed phylogenetic networks. A phylogenetic network N^+ is **directed** if n is a directed acyclic graph with the property that there exists a unique vertex, called the root, with in-degree 0, from which there is a directed path to every other vertex in n. Internal vertices (i.e., non-root, non-leaf) vertices are called *reticulation nodes* if they have in-degree greater than 1; otherwise, they have in-degree 1 and are called *tree nodes*. Vertices with in-degree 1 and out-degree 1 are allowed, which allows modeling changes in population sizes like bottlenecks. A **tree edge** is an arc whose child is a tree node; all other arcs are called *reticulation edges*. Reticulation edges with the same child v are called *partners*. A directed network N^+ can be transformed into a **semi-directed** network N^- by (1) keeping only the edges and vertices that are descendants of the lowest stable ancestor (page 263 in [26]), which is now the root, (2) un-directing tree edges, and (3) suppressing the root vertex (see Fig. 1 in [4]; also see [24]). If there are no reticulation edges, then directed and semi-directed networks correspond to rooted and unrooted trees, respectively.

Undirected networks. A directed network N^+ can be transformed into an **undirected** network N by un-directing edges and suppressing the root (the same transformation can be performed on a semi-directed network N^- except there is no root to suppress). A bridge or **cut edge** is an edge whose deletion increases the number of connected components of the graph, when viewing the network as undirected. All cut edges are tree edges. The maximal bridgeless subgraphs are called **blobs**. A blob is trivial if it consists of a single node. The degree of a blob is the number of cut edges adjacent to it; a blob with degree m is called an m-blob. Edges that are not cut edges are **blob edges**; blob edges may be either tree edges or reticulation edges in the directed or semi-directed version of the network. An undirected network is *level-k* if it can be simplified to a tree by removing at most k edges from each blob; thus, each vertex in a *level-1* network can participate in at most one cycle. An undirected network is level-0 if it has zero reticulation nodes, in which case it is simply an unrooted tree T.

Refinements. A *contraction* operation on edge $e = (s, t)$ in network N deletes e and merges s and t into a new vertex v such that that $E(v) = E(s) \cup E(t) \setminus \{e\}$. A *refinement* operation on v can reverse the contraction, introducing a new edge e. Let N and N' be two networks. If N can be obtained from N' through a sequence of edge contractions, we say that N' is a **refinement** of N or that N' refines N. Every network is a refinement of itself. Typically, we use this terminology in the context of phylogenetic trees [28].

Restricted and induced subnetworks. A directed network N^+ on taxon set S can be restricted to a subset of taxa $Y \subseteq S$ by deleting these species, along with any vertices and edges that are not on any directed path between the lowest stable ancestor for Y and any species in Y (see Fig. S1A–D). The induced subnetwork on Y, denoted $N^+|_Y$, also takes the restriction and then suppresses any internal vertices of degree 2 [25, 4]. Typically, we are interested in the semi-directed or undirected version of the induced subnetwork, denoted $N|_Y$ (Fig. S1E–G).

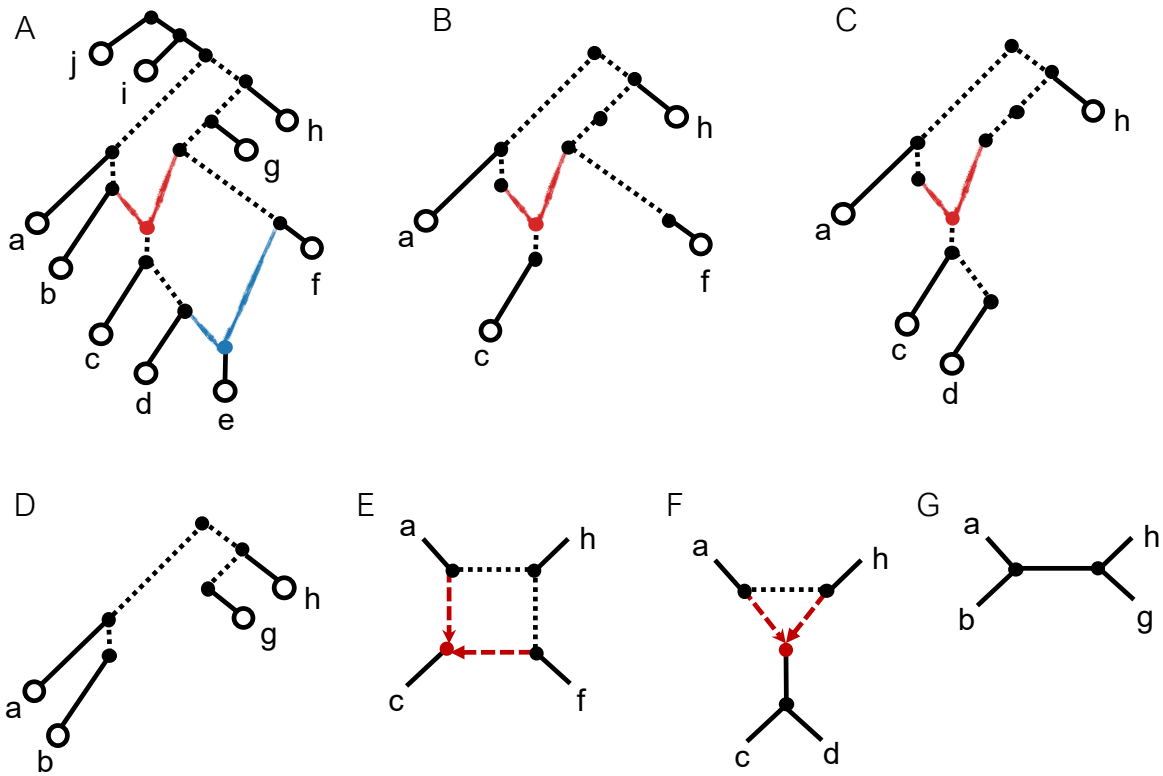

Figure S1: **Restricted and induced subnetworks.** **A)** A level-2 directed phylogenetic network N^+. Tree edges are black, and reticulation edges are red or blue to match the color of their shared child vertex. Reticulation edges are part of the blob, as are the dashed edges. **B)** Network restricted to $\{a, c, f, h\}$. **C)** Network restricted to $\{a, c, d, h\}$. **D)** Network restricted to $\{a, b, g, h\}$. **E)** Induced subnetwork on $\{a, c, f, h\}$; it contains a 4-blob and is class 2, as it has two display trees: $a, c|f, h$ and $a, h|c, f$. **F)** Induced subnetwork on $\{a, c, d, h\}$; it contains a 3-blob and is class 1, as it has one display tree: $a, h|c, d$. **G)** Induced subnetwork on $\{a, b, h, g\}$; it contains no blobs and thus is a tree, not a network.

Display trees and network class. A rooted tree T^+ is a **display tree** of a directed network N^+ if T^+ can be obtained from N^+ by selecting one incoming edge for each of the reticulation nodes, removing their partners but not their end points, and then deleting any vertices not on any undirected path between any two leaves. Vertices with degree 2 are not suppress so that each edge in display tree T^+ maps to a unique edge in N^+ and can inherit any associated model parameters. These definitions extend to semi-directed networks in the natural way. The number of topologically distinct induced display trees is the **class** of the network [4] (Fig. S1E–G).

Induced subtrees and quartets. Similar to directed networks, a rooted or unrooted tree induces an unrooted subtree on taxon set $Y \subseteq S$, denoted $T|_Y$. Typically we are interested in induced *quartets*, which are unrooted binary trees with four uniquely labeled leaves. A quartet q on species set $S = \{a, b, c, d\}$ has

three possible topologies denoted by the separation of species on its only internal edge, that is, $a, b|c, d$ and $a, c|b, d$ and $a, d|b, c$. An unrooted tree T on species set S induces a set $Q(T)$ of quartets, one for every subset of four species in S.

Partitions. Let $M = P^1|P^2|...|P^m$ be an m-partition of species set S such that $P^i \cap P^j = \emptyset$ and $\cup_{i=1}^{m} P^i = S$. Each vertex with degree m in a phylogenetic tree T on S induces a m-partition on S formed by deleting v and all edges incident to it and then assigning labels in the same component to the same block of the partition; the same holds for deleting an m-blob in a phylogenetic network N. Likewise, the deletion of an edge e induces a 2-partition, called a **bipartition** and denoted $Bip(e) = A|B$. Observe that for any $x, y \in A$ and $z, w \in B$ with $x \neq y, z \neq w$, $T|_{\{x,y,z,w\}}$ is the quartet $x, y|z, w$ but after contracting e, $T|_{\{x,y,z,w\}}$ is a star. Thus, we say that a bipartition induces $\binom{|A|}{2}\binom{|B|}{2}$ quartets, from taking two species from each block. Likewise, if T is binary, edge $e = (s, t)$ also induces a 4-partition, called a **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. A quadrapartition induces $\binom{|A_1|}{2}\binom{|A_2|}{2}\binom{|B_2|}{2}\binom{|B_2|}{2}$ quartets, from selecting one species from each block.

Trees of Blobs (TOB). The **tree of blobs (TOB)** for a phylogenetic network N, typically denoted T_N, is formed by (1) un-directing the network if necessary, (2) contracting all edges in the same blob into single vertex, and then (3) suppressing any vertices with degree two [11]; these operations can be performed in linear time [2, 14]. Any unrooted tree induced by a display tree of a directed network or semi-directed network is a TOB refinement; however, the converse is not necessarily true (Fig. S2).

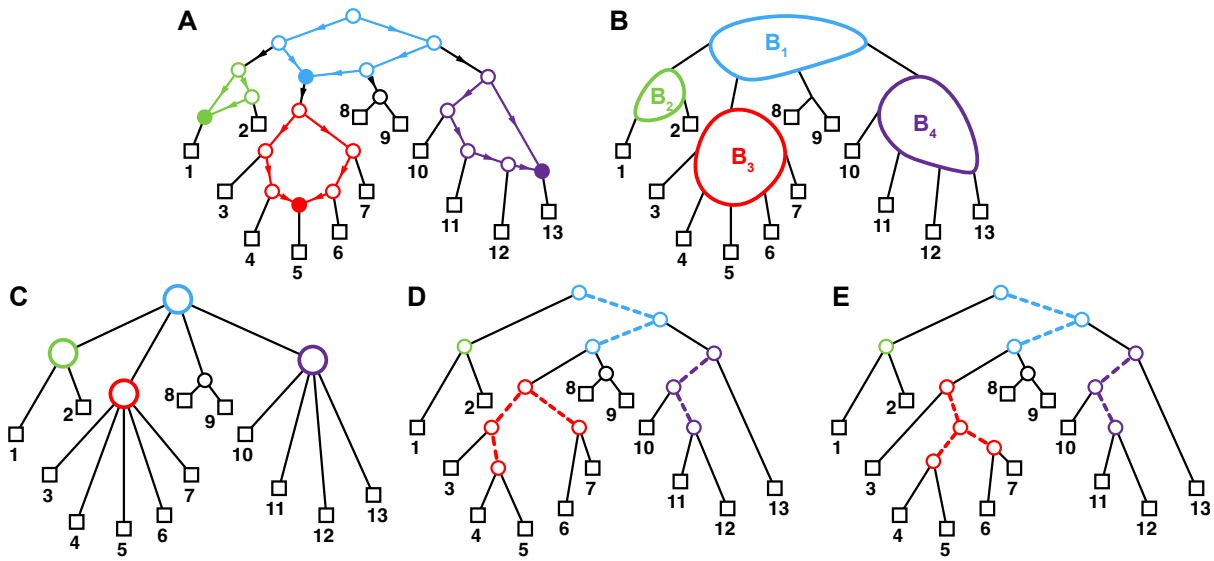

Figure S2: **Phylogenetic Networks, Tree of Blobs, and Refinements. A)** A level-1 directed phylogenetic network N^+ with four reticulation vertices filled in different colors. **B)** The unrooted version of N^+ with blob edges colored based on the single reticulation node in the blob. **C)** Tree of blobs for N^+. **D)** A TOB refinement that is induced by a display tree of N^+. **E)** A TOB refinement that is NOT induced by any display trees of N^+. Dashed edges are contracted to obtain the tree of blobs.

3 Network Multispecies Coalescent (MSC) Model

Our results assume that data (gene trees) are generated under the network multispecies coalescent model, with correlated inheritance. We include a description of the model below for completeness, although it is not necessary for understanding our results.

Network multispecies coalescent (NMSC). The *network multispecies coalescent*, with inheritance correlation, was recently described by Fogg*et al.* [9]. The model is parameterized by the inheritance correlation ρ, which is used to set the concentration parameter $\alpha = \frac{\rho}{1-\rho}$, as well as a directed phylogenetic network N^+ on species/population set S and a set θ of numeric values associated with the edges of N^+:

- Branch lengths $\ell(e) = t(e)/2N_e(e)$ in coalescent units for each arc e, where
  - $N_e(e)$ is the effective population size and
  - $t(e)$ is the number of generations
- Inheritance probability $\gamma(e)$

A species network $\mathcal{N}^+ = (N^+, \Theta)$ is *metric* if each edge e has length $\ell(e) \geq 0$ and is assigned an inheritance probability $\gamma(e) \in [0, 1]$ such that its sum across all partner edges with the same child vertex equals 1. Put simply, each vertex v has an associated inheritance probability distribution $\gamma(v) = \{\gamma(u \mapsto v) : u \in parents(v)\}$. We also typically assume that $0 < \gamma(e) < 1$ if e reticulation vertex (otherwise $\mathcal{N}^+$ can be simplified) and that all tree edges have positive lengths.

Given a species network $\mathcal{N}^+$, a *gene genealogy* T^+ can be generated under the NMSC(ρ) by sampling lineages, typically one lineage per species is sampled for the gene, and then tracing the ancestry of these lineages backward in time with following the procedure.

- At a **tree node** v, all lineages from descendants entering v (or sampled if v is a leaf) can coalesce on the incoming edge $e = p \mapsto v$ with population size $N_e(e)$ and generation time $t(e)$ following the standard the multi-species coalescent (MSC) [15, 27, 21, 23].
- At a **reticulation node** v with k parent nodes, each lineage entering the node is randomly assigned a parent as follows.
  1. Arbitrarily order the lineages entering v, denoted $l_1, l_2, \ldots$.
  2. Randomly assign a parent to lineage l_1 according to the inheritance probability distribution $\gamma(v)$. Set $i = 1$.
  3. Randomly assign a parent to the next lineage l_{i+1} according to a distribution where the parents assigned to lineages $l_1, \ldots, l_i$ have probability $\frac{1}{\alpha+i}$, and the NULL parent has probability $\frac{\alpha}{\alpha+i}$. If the NULL parent is assigned, randomly assign a parent to l_{i+1} according to the inheritance probability distribution $\gamma(v)$.
  4. Repeat step 3 until all lineages have been assigned a parent.

  This process allows each lineage to have increased preference for having the same ancestry as the previously processed lineages. The correlation between the parent assignment for any pair of lineages l_i and l_j is $\rho = 1/(1 + \alpha)$ [9]. All lineages assigned to the same parent u can coalesce on the incoming edge $u \mapsto v$, if $\ell(u \mapsto v) > 0$, following the standard MSC.

Thus, the resulting gene genealogy T^+ is a rooted, binary phylogenetic tree with each leaf vertex labeled by the species from which the gene lineage was sampled. The probability of T^+ under the NMSC(ρ) given $\mathcal{N}^+$ is denoted $\mathbb{P}_{\mathcal{N}^+,\rho}(T^+)$; for simplicity, we often denote this probability as $\mathbb{P}(T^+)$, omitting the dependencies, as they are implied. The probability of an unrooted gene genealogy T can be computed by simply summing the probabilities of all rooted gene genealogies that induce T.

The NMSC with inheritance correlation was introduced by Fogg *et al.* [9] to generalize and unify the two commonly used traditional network coalescent models: the Common Inheritance model and the Independent Inheritance model, which we now define.

Common Inheritance model. The *Common Inheritance model* is the NMSC(ρ) model with $\rho = 1$, which is equivalent to setting $\alpha = 0$. When $\alpha = 0$, all lineages must be assigned to the same parent as lineage l_1 because $\frac{\alpha}{\alpha+i} = 0$ for any i. Put simply, all lineages sampled for the same gene must evolve within the same display tree.

Independent Inheritance model. The *Independent Inheritance model* is the NMSC(ρ) model with $\rho = 0$, which is equivalent to setting $\alpha = \infty$. When $\alpha = \infty$, the NULL parent is assigned with probability $\frac{\alpha}{\alpha+i} = \frac{1}{1+\frac{i}{\alpha}} = 1$, asymptotically, for any i. Thus, all lineages are independently assigned a parent according to the inheritance probability distribution, so there is no ancestry preference among lineages sampled for the same gene.

Model Comparison. The Independent Inheritance model is more flexible and thus is the most commonly-used model for species network inference [3, 24, 31, 22, 19, 6]. The Common Inheritance model can be viewed as strong selection acting either for or against gene variants associated with donor populations, which results in all of lineages sampled at a particular locus/gene tracing their ancestry back through the same display tree. As a consequence of this strong assumption, the Common Inheritance model has nice statistical properties [10, 29, 17].

4 Model Assumptions and Prior Results

Henceforth, we assume that species networks are binary, metric, and have strictly positive lengths on their tree edges. These same assumptions are made by TINNiK [2]. The T3 tests employed by TINNiK further assume the network is quartet-nonanomalous, which we now define.

Definition 1 (Quartet-nonanomalous network; [8]; also see Definition 5 in [4])**.** *Let $\mathcal{N} = (N, \Theta)$ be a metric binary species network on S. $\mathcal{N}$ is* quartet-nonanomalous *if for every subset Y of four species in S, no quartet on Y that is NOT induced by a display tree of the N has higher probability than a quartet on Y that is induced by a display tree of N. More precisely, for all subsets Y, the following conditions must hold:*

1. $\mathbb{P}(q_1^Y) \geq \mathbb{P}(q_2^Y)$ *and* $\mathbb{P}(q_1^Y) \geq \mathbb{P}(q_3^Y)$ *if* q_1^Y *is the only quartet induced by the display trees of N or*
2. $\mathbb{P}(q_3^Y) \leq min\{\mathbb{P}(q_1^Y), \mathbb{P}(q_2^Y)\}$ *if* q_3^Y *is the only quartet NOT induced by any display trees of N*

where q_1^Y, q_2^Y, and q_3^Y denote the three possible quartets on Y. If there exists some subset Y for which the conditions above do not hold, $\mathcal{N}$ is quartet-anomalous.

This definition allows $\mathbb{P}(q_1^Y) = \mathbb{P}(q_2^Y) = \mathbb{P}(q_3^Y) = \frac{1}{3}$; however, we can safely assume this does not occur due to the requirement that tree edges have strictly positive lengths, as discussed by [2]. Additionally some of our results do not require condition (2) above to hold, so we relax the definition.

Definition 2 (Relaxed quartet-nonanomalous network)**.** *Let $\mathcal{N} = (N, \Theta)$ be a metric binary species network on S with strictly positive lengths on its tree edges. $\mathcal{N}$ is* relaxed quartet-nonanomalous *if, for all subsets Y of four species in S for which there is a exactly one quartet q_1^Y induced by the display trees of N,*

$$\mathbb{P}(q_1^Y) > \mathbb{P}(q_2^Y) \text{ and } \mathbb{P}(q_1^Y) > \mathbb{P}(q_3^Y) \tag{1}$$

where q_2^Y and q_3^Y denote the other two possible quartets on Y.

The related Lemma is useful for proving our results.

Lemma 1 (Lemma 1 in [4])**.** *A 4-taxon binary network N displays exactly quartet (i.e., is of class 1) if, and only if, it does not have a 4-blob.*

An important consideration is whether it is biologically realistic to assume that species networks are quartet-nonanomalous. Ané *et al.* recently studied this question [4]. They proved that all binary metric species networks are quartet-nonanomalous under the Common Inheritance model with $\rho = 1$ (Proposition 12 in [4]). Such strong results do not exist for other settings of ρ.

Take Figure S3A as an example. The inheritance correlation ρ affects the quartet distribution in a 3_2 cycle quartet network (i.e., a 3-blob with the hybrid vertex having 2 descendants). Particularly, when edge lengths $t_1, t_2, t \to 0$ and $t_3 \to \infty$, we have the following [4]

$$\mathbb{P}(AB|CD)|_{\rho=0,t_1,t_2,t\to 0,t_3\to\infty} = \frac{1}{3}(\gamma^2 + (1-\gamma)^2) < \frac{1}{3}$$

Further, Allman *et al.* showed that a 3_2 cycle is not identifiable from quartet probabilities, even when $\rho = 0$; its quartet probabilities may even suggest it is a 4-blob (Proposition 9 in [3]). No doubt, the 3_2 cycle example challenges reconstruction of display trees and the tree of blobs. Moreover, a 3_2 cycle is the necessary condition of a network being quartet anomalous for $\rho < 1$ (Theorem 8 in [4]). On the other hand, the inheritance correlation ρ does not affect the quartet distribution in a 4_1 cycle (Fig. S3B). Moreover, in this case, we can easily verify that this species network is quartet nonanomalous for any arbitrary γ with assume of positive branch lengths for t_1' and t_2'.

To summarize, there are particular configurations where anomalous quartets may be a problem; however, Ané *et al.* found that quartet-anomalous networks are rare in the practice, simulating networks under a birth-death-hybridization model, even with $\rho = 0$ [4] Based on these empirical results, it is reasonable to assume species networks are quartet-nonanomalous, like the T3 test employed by TINNik [2].

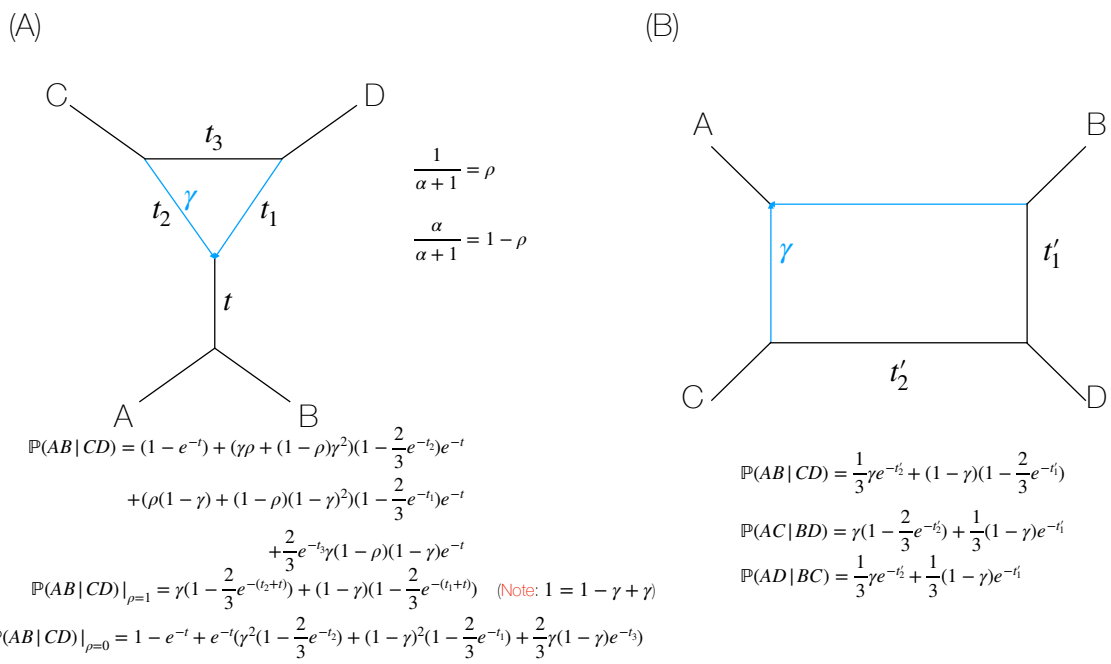

Figure S3: **4-taxon, binary, semi-directed species networks. A)** This network is referred to as a 3_2 cycle (Fig. 18 in [5]). It induces one display tree $A, B|C, D$; likewise, the TOB for this network is $A, B|C, D$. This network has no anomalous quartets when $\rho = 1$ (Common Inheritance model); however, it can have anomalous quartets when $0 \leq \rho < 1$ (Independent Inheritance model). **B)** This network is referred to as a 4_1 cycle (Fig. 17 in [5]). It induces two display trees: $A, B|C, D$ and $A, C|B, D$; the TOB for this network is a star. For this network, the inheritance correlation parameter ρ does not impact the quartet probabilities.

5 Weighted Quartet Consensus and Quartet Concordance Factors

A popular trend in species tree and network construction is to utilize information from gene trees restricted subsets of four taxa. In particular, quartet-based species tree methods, including the popular method ASTRAL [20], are often heuristics for the NP-hard **Weighted Quartet Consensus (WQC)** problem [18], which is defined by the following input and output.

- **Input:** A set $\mathcal{T}$ of k unrooted gene trees, each on the same set S of n taxa
- **Output:** An unrooted binary tree T on S that maximizes the quartet score:

$$QS_{\mathcal{T}}(T) = \sum_{Y \in f_4(S)} w_{\mathcal{T}}(T|_Y) \tag{2}$$

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 $f_4(S) = \{\{a,b,c,d\} : a,b,c,d \in S, a \neq b \neq c \neq d\}$.

Relatedly, methods for species network and tree of blob reconstruction, like SNaQ [24] and TINNiK [2] respectively, often leverage quartet concordance factors, which we now define. Let $Y = \{a,b,c,d\}$ be some subset of four species from S. The empirical or **observed quartet concordance factors (qCFs)** for Y are

$$q\hat{C}F_{abcd}(\mathcal{T}) = \left(\frac{w_{\mathcal{T}}(ab|cd)}{k}, \frac{w_{\mathcal{T}}(ac|bd)}{k}, \frac{w_{\mathcal{T}}(ad|bc)}{k}\right) \tag{3}$$

where $w_{\mathcal{T}}(q)$ is the number of gene trees (out of k) that induce quartet q. The **expected values of qCFs** under the NMSC(ρ) given species network $\mathcal{N}$ are

$$qCF_{abcd}(\mathcal{N},\rho) = \left(\mathbb{P}_{\mathcal{N},\rho}(ab|cd), \mathbb{P}_{\mathcal{N},\rho}(ac|bd), \mathbb{P}_{\mathcal{N},\rho}(ad|bc)\right) \tag{4}$$

where the probability of th estar tree on Y is zero, as the NMSC generates binary genealogies. We typically use $F^*_{\mathcal{N},\rho}(T_{ad|bc})$ and the expected quartet concordance factors. Expected qCFs are fully determined by the semi-directed version of $\mathcal{N}^+$ (Proposition 3 in [4]).

Under the assumption that a species network is quartet-nonanomalous, one might (incorrectly) conclude that reconstructing a tree with ASTRAL and then adding reticulation edges would be a consistent approach to species network reconstruction (such an approach was explored in [7]). However, Dinh and Baños recently gave an example of a quartet-nonanomalous network with just 5 taxa that suggests this approach would be misleading, as none of its display trees are solutions to WQC, as the number k of gene trees tends to infinity. This makes our result that WQC converges in probability to a refinement of the tree of blobs for the network more interesting, WQC is a consistent estimator of the tree of blobs, when only considering false negative (FN) branch errors and not false positive (FP) branch errors.

Definition 3 (Statistical Consistency)**.** *Let $\mathcal{T}$ be a set of (unrooted, binary) gene trees generated under $NMSC(p)$ given species network $\mathcal{N}^+ = (N^+, \Theta)$, and let T_N be the (unrooted) tree of blobs for N^+. We say that a TOB estimation method $\mathcal{A}$ is statistically consistent under $NMSC(p)$ if for any $\epsilon > 0$, there exists a constant $K > 0$ such that when $\mathcal{A}$ is given at least K gene trees, $\mathcal{A}$ returns T_N with probability at least $1-\epsilon$.*

6 WQC converges to a TOB refinement with high probability

Our framework to tree of blob reconstruction first estimates a refinement to the TOB and then contracts false positive branches.

We now show that WQC converges to a TOB refinement with high probability, as the number of gene trees tends to infinity, motivating the use of quartet-based methods for step one. The following Lemma is useful for proving our main result.

Lemma 2. *Let $\mathcal{N} = (N, \Theta)$ be a binary metric species network on S with positive lengths on its tree edges, and let $T|_N$ denote the tree of blobs for N. Further assume that $\mathcal{N}$ is relaxed quartet-nonanomalous. Then, for any quartet $a,b|c,d$ induced by $T|_N$, the following inequalities hold*

$$\mathbb{P}(a,b|c,d) > \mathbb{P}(a,c|b,d) \text{ and } \mathbb{P}(a,b|c,d) > \mathbb{P}(a,d|b,c).$$

Proof. Let $\mathcal{Q}$ be the set of quartets on $Y = \{a,b,c,d\}$ that are induced by the display trees of $\mathcal{N}$. If $|\mathcal{Q}| = 1$, the inequalities above hold by the definition of relaxed quartet-nonanomalous networks (Definition 2). If $|\mathcal{Q}| > 1$, then by Lemma 1 in [4], there exists a blob $\mathcal{B}$ in N such that each taxon in $y \in Y$ is separated from the other three taxa in Y by a cut edge, denoted $e_y = (u_y, v_y)$, where vertex v_y is in $\mathcal{B}$, and $v_a \neq v_b \neq v_c \neq v_d$. Thus, e_a induces a bipartition $W|Z$ such that $a \in W$ and $b,c,d \in Z$ or vice versa (and similarly for e_b, e_c, e_d). Observe that all edges on the path between v_a and v_c, denoted $p(v_a, v_c)$, are in $\mathcal{B}$ and thus are contracted when building T_N (and similarly $p(v_a,v_d), p(v_b,v_c), p(v_b,v_d)$). It follows that there exists a vertex in T_N whose deletion induces an m-partition with $P_1|P_2|\ldots|P_m$ such that $a \in P_i, b \in P_j, c \in P_s, d \in P_t$ with $P_i \neq P_j \neq P_s \neq P_t$, so $T|_Y$ is a star. 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 with positive lengths on its tree edges. Further assume that $\mathcal{N}$ is relaxed quartet-nonanomalous. Let $\mathcal{T}$ be a set of k (binary) gene trees generated under $NMSC(\rho)$ given $\mathcal{N}$. Let B be an unrooted binary tree on S that is a refinement of T_N, and let T be an unrooted binary tree on S that is NOT. Then, for any T and any $\epsilon > 0$, there exists a $K > 0$ and B such that, given $k \geq K$ gene trees, $QS_{\mathcal{T}}(B) > QS_{\mathcal{T}}(T)$ with probability at least $1 - \epsilon$. Put simply, Weighted Quartet Consensus converges to a refinement of the tree of blobs for N with high probability, as the number of gene trees tends to infinity.*

Proof. The result follows from the following two claims.

<u>*Claim 1.*</u> Let $f_4(S)$ be the set of all 4-taxon subsets of S, that is $f_4(S) = \{\{a, b, c, d\} : \{a, b, c, d\} \in S, a \neq b \neq c \neq d\}$, and let q_1^Y, q_2^Y, q_3^Y be the three possible quartets on $Y \in f_4(S)$. We claim that $w_{\mathcal{T}}(q_i^Y)/k$ converges in probability to its expected value $F^*_{\mathcal{N},\rho}(q_i^Y) = \mathbb{P}_{\mathcal{N},\rho}(q_i^Y)$ for all $i \in \{1, 2, 3\}$ and $Y \in f_4(S)$, as the number k of gene trees tends to ∞. Claim 1 is a standard assumption in phylogenetic networks literature. Then, by definition, for any pair $\epsilon > 0$ and $\delta > 0$, there exists a $K > 0$ such that, given $k \geq K$ gene trees, $|w_{\mathcal{T}}(q_i^Y)/k - F^*_{\mathcal{N},\rho}(q_i^Y)| < \delta$ for all $i \in \{1, 2, 3\}$ and $Y \in f_4(S)$ with probability at least $1 - \epsilon$.

<u>*Claim 2.*</u> Let π be the minimum absolute distance between any expected quartet concordance factor on the same set of species, that is

$$\pi = \min_{>0, Y \in f_4(S)} \{|F^*_{\mathcal{N},\rho}(q_1^Y) - F^*_{\mathcal{N},\rho}(q_2^Y)|, |F^*_{\mathcal{N},\rho}(q_1^Y) - F^*_{\mathcal{N},\rho}(q_3^Y)|, |F^*_{\mathcal{N},\rho}(q_2^Y) - F^*_{\mathcal{N},\rho}(q_3^Y)|\}, \tag{5}$$

and let $\delta_* = \pi/2|f_4(S)|$. We claim that if $|w_{\mathcal{T}}(q_i^Y)/k - F^*_{\mathcal{N},\rho}(q_i^Y)| < \delta_*$ for all $i \in \{1, 2, 3\}$ and $Y \in f_4(S)$, then there exists a B such that $QS_{\mathcal{T}}(B) > QS_{\mathcal{T}}(T)$ for any T. The proof of claim 2 is given in Lemma 3. □

Lemma 3. *Assume the same conditions and notation as Theorem 1. If*

$$|w_{\mathcal{T}}(q_i^Y)/k - F^*_{\mathcal{N},\rho}(q_i^Y)| < \delta_* = \frac{\pi}{2|f_4(S)|}$$

for all $i \in \{1, 2, 3\}$ and $Y \in f_4(S)$, then there exists a refinement of the TOB, denoted B, such that $QS_{\mathcal{T}}(B) > QS_{\mathcal{T}}(T)$ for any arbitrary (non-refinement) tree T.

Proof. In the proof of Theorem 2, we will show how to construct a refinement of T_N that maximizes the quartet score when there is no error in the quartet concordance factors, that is $|w_{\mathcal{T}}(q_i^Y)/k - F^*(q_i^Y)| = 0$. Let B^* denote such a tree. Our proof we partitions the calculation of the quartet score for a binary tree t based on the tree of blobs $T|_N$, specifically

$$QS_{\mathcal{T}}(t) = QS(t)^{tree}_{\mathcal{T}} + QS^{star}_{\mathcal{T}}(t) = \sum_{Y \in \mathcal{S}_{tree}} F^*_{\mathcal{N},\rho}(t|_Y) + \sum_{S \in \mathcal{S}_{star}} F^*_{\mathcal{N},\rho}(t|_Y) \tag{6}$$

where $\mathcal{S}_{tree}$ and $\mathcal{S}_{star}$ denote the 4-taxon subsets on which T_N induces a quartet or a star, respectively. Because we showed that B^* maximizes the quartet score on $\mathcal{S}_{tree}$ and $\mathcal{S}_{star}$ separately, we can let $QS(t)^{tree}_{\mathcal{T}}(*)$ and $QS(t)^{star}_{\mathcal{T}}(*)$ denote the best scores obtainable on each part. Now we want to show that when $|w_{\mathcal{T}}(q_i^Y)/k - F^*_{\mathcal{N},\rho}(q_i^Y)| < \delta$,

$$\begin{aligned} QS_{\mathcal{T}}(B^*) &> QS_{\mathcal{T}}(T) \\ QS^{tree}_{\mathcal{T}}(B^*) - QS^{tree}_{\mathcal{T}}(T) &> QS^{star}_{\mathcal{T}}(T) - QS^{star}_{\mathcal{T}}(B^*) \end{aligned} \tag{7}$$

Right side. To begin, we give a lower bound for the right side of Equation 7. We need only consider the 4-taxon subsets of $\mathcal{S}_{tree}(T_N)$ for which $T|_Y \neq B^*|_Y$, denoted $\mathcal{Y}_{tree}$. Then, by Lemma 2,

$$\begin{aligned} \frac{QS^{tree}_{\mathcal{T}}(B^*)}{k} - \frac{QS^{tree}_{\mathcal{T}}(T)}{k} &= \sum_{Y \in \mathcal{Y}_{tree}} \left(\frac{w_{\mathcal{T}}(B^*|_Y)}{k} - \frac{w_{\mathcal{T}}(T|_Y)}{k} \right) \\ &\geq \sum_{Y \in \mathcal{Y}_{tree}} \left((F^*_{\mathcal{N},\rho}(B^*|_Y) - \delta) - (F^*_{\mathcal{N},\rho}(T|_Y) + \delta) \right) \\ &\geq |\mathcal{Y}_{tree}|\pi - |\mathcal{Y}_{tree}|2\delta. \end{aligned} \tag{8}$$

Left side. Now we give an upper bound for the left side of the Equation 7. Again, we need only consider the 4-taxon subsets of $\mathcal{S}_{star}$ for which $T|_Y \neq B^*|_Y$, denoted $\mathcal{Y}_{star}$. First, observe that

$$Q^{star}_{\mathcal{T}}(B^*)/k = \sum_{Y \in \mathcal{Y}_{star}} \frac{w_{\mathcal{T}}(B^*|_Y)}{k} \geq \sum_{Y \in \mathcal{Y}_{star}} (F^*_{\mathcal{N},\rho}(B^*|_Y) - \delta) = \frac{QS^{star}_{\mathcal{T}}(*)}{k} - |\mathcal{Y}_{star}|\delta.$$

Second, observe that

$$QS^{star}_{\mathcal{T}}(T) = \sum_{Y \in \mathcal{Y}} \frac{w_{\mathcal{T}}(T|_Y)}{k} \leq \frac{QS^{star}_{\mathcal{T}}(*)}{k} + |\mathcal{Y}_{star}|\delta,$$

as this is the maximum score that can be obtained when error is zero plus the maximum amount the score can be *increased* due to error. Putting this together gives us an upper bound

$$\frac{QS^{star}_{\mathcal{T}}(T)}{k} - \frac{QS^{star}_{\mathcal{T}}(B^*)}{k} \geq \left(\frac{QS^{star}_{\mathcal{T}}(*)}{k} + |\mathcal{Y}_{star}|\delta\right) - \left(\frac{QS^{star}_{\mathcal{T}}(*)}{k} - |\mathcal{Y}_{star}|\delta\right) = 2|\mathcal{Y}_{star}|\delta. \tag{9}$$

Combining bounds together. Putting Equations 9 and 8 together gives

$$\frac{QS^{tree}_{\mathcal{T}}(B^*)}{k} - \frac{QS^{tree}_{\mathcal{T}}(T)}{k} \geq |\mathcal{Y}_{tree}|\pi - |\mathcal{Y}_{tree}|2\delta > 2|\mathcal{Y}_{star}|\delta \geq \frac{QS^{star}_{\mathcal{T}}(T)}{k} - \frac{QS^{star}_{\mathcal{T}}(B^*)}{k}$$

Looking at the inner inequality, we have

$$|\mathcal{Y}_{tree}|\pi - |\mathcal{Y}_{tree}|2\delta > 2|\mathcal{Y}_{tree}|\delta$$
$$\frac{|\mathcal{Y}_{tree}|}{|\mathcal{Y}_{tree}| + |\mathcal{Y}_{star}|}\frac{\pi}{2} > \delta$$

where $1 \leq |\mathcal{Y}_{tree}| \leq |\mathcal{S}_{tree}|$ and $|\mathcal{Y}_{tree}| + |\mathcal{Y}_{star}| \leq |\mathcal{S}_{tree}| + |\mathcal{S}_{star}| = |f_4(S)|$ by construction. Therefore, when $\delta_* = \pi/(2|f_4(S)|)$, there exists a TOB refinement, specifically B^*, that has a higher quartet score than any arbitrary (non-refinement) tree T. □

It remains to be shown how to construct B^*.

Theorem 2. *Assume the same conditions and notation as Theorem 1. Given a set $\mathcal{T}$ of k (binary) gene trees generated under $NMSC(\rho)$ given $\mathcal{N}$, then there exists a TOB refinement that maximizes the quartet score when the error in all quartet concordance factors is zero, that is $|w_{\mathcal{T}}(q_i^Y)/k - F^*_{\mathcal{N},\rho}(q_i^Y)| = 0$ for all $i \in \{1,2,3\}$ and $Y \in f_4(S)$.*

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

$$\sum_{Y \in \mathcal{S}_{tree}(T_N)} F^*_{\mathcal{N},\rho}(t|_Y) + \sum_{Y \in \mathcal{S}_{star}(T_N)} F^*_{\mathcal{N},\rho}(t|_Y) \tag{10}$$

provided the input quartet concordances are error-free. We can further partition $Y \in \mathcal{S}_{star}$ as follows. Observe that $Y = \{a,b,c,d\} \in \mathcal{S}_{star}(T_B)$, if and only if, there exists a unique vertex v in T_B 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 $P_i \neq P_j \neq P_s \neq P_t$ (as an induced quartet has exactly two internal vertices of degree three and an induced star has exactly one internal vertex degree). It follows that

$$\sum_{Y \in \mathcal{S}_{star}(T_N)} F^*_{\mathcal{N},\rho}(t|_Y) = \sum_{v \in V_{>3}(T_N)} \sum_{P \in f_4(Part(v))} \sum_{Y \in f_{cover}(P)} F^*_{\mathcal{N},\rho}(t|_Y) \tag{11}$$

where $V_{>3}(T_N)$ is the set of all vertices in T_N with degree greater than 3, $f_4(Part(v))$ is the set of 4-block subsets of $Part(v)$, that is $f_4(Part(v)) = \{P_i, P_j, P_s, P_t : P_i, P_j, P_s, P_t \in Part(v), P_i, \neq P_j \neq$

$P_s \neq P_t\}$, and $f_{cover}(\{P_i, P_j, P_s, P_t\})$ is the set of all 4-taxon subsets of S covering the input, that is, $f_{cover}(\{P_i, P_j, P_s, P_t\}) = \{\{a, b, c, d\} : a \in P_i, b \in P_j, c \in P_s, d \in P_t\}$.

We now show how to construct a binary tree that maximizes the quartet score.

Tree part of quartet score. First we consider the "tree part" of the quartet score. For a binary tree T on S that is NOT a refinement of T_N, we have

$$\sum_{Y \in \mathcal{S}_{tree}(T_N)} F^*_{\mathcal{N},\rho}(T_N|_Y) > \sum_{Y \in \mathcal{S}_{tree}(T_N)} F^*_{\mathcal{N},\rho}(T|_Y) \tag{12}$$

from Lemma 2.

Star part of quartet score. Now we just consider the "star part" of the quartet score. It suffices to show that there exists a refinement of T_N, denoted B^*, such that

$$\sum_{P \in f_4(Part(v))} \sum_{Y \in f_{cover}(P)} F^*_{\mathcal{N},\rho}(B^*|_Y) \geq \sum_{P \in f_4(Part(v))} \sum_{Y \in f_{cover}(P)} F^*_{\mathcal{N},\rho}(T|_Y) \tag{13}$$

for all $v \in V_{>3}(T_N)$. Let $T_N(v)$ denote the tree obtained by contracting all edges in the tree of blobs 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 for every $v \in V_{>3}(T_N)$ there exists a refinement of $T_N(v)$, denoted B^*_v, that (1) maximizes Equation 13 and (2) does not break apart blocks of $Part(v)$, meaning that each of its non-trivial bipartitions $A|B$ satisfies $P_i \cap A = P_i$ and $P_i \cap B = \emptyset$ (or vice versa) for all $P_i \in Part(v)$.

It is well-established that any 4-tuple of species covering $P = \{P_i, P_j, P_s, P_t\}$ has the same expected quartet concordance factors, as a lineage sampled from any species in P_i must enter the blob in $\mathcal{N}$, associated with v in T_B, from the same cut edge (and similarly for P_j, P_s, P_t), and only after the four sampled lineages enter this blob can they coalesce; thus, the expected quartet concordance factors are fully dictated by the blob and do not depend on which particular taxon was selected; see [5] for details. This allows us to rewrite the summations in Equation 13 as follows. Define $\alpha_1^P(T)$ as the number of quartets in T that display topology $q_1^P = a, b|c, d$ for all ways of selecting $a \in P_i, b \in P_j, c \in P_s, d \in P_t$ (and similarly for α_2 and $q_2^P = a, c|b, d$ as well as α_3 and $q_3^P = a, c|b, d$). Thus,

$$\sum_{P \in f_4(Part(v))} \sum_{Y \in f_{cover}(P)} F^*_{\mathcal{N},\rho}(T|_Y) = \sum_{P \in f_4(Part(v))} \sum_{i=1}^{3} \alpha_i^P(T) F^*_{\mathcal{N},\rho}(q_i^P). \tag{14}$$

Observe that $\alpha_1^P(T) + \alpha_2^P(T) + \alpha_3^P(T) = |P_i| \cdot |P_j| \cdot |P_s| \cdot |P_t| = \beta^P$ for any binary tree T on S. and that if T is a refinement of $T|_v$, then exactly one of $\alpha_1^P(T), \alpha_2^P(T), \alpha_3^P(T)$ equals β^P and the other two equal 0 because $T|_v$ forces the topology on any four species covering P to be resolved the same way. Moreover, we can maximize the inner sum for any arbitrary $P \in f_4(Part(v))$, if we select q_*^P such that $F^*_{\mathcal{N},\rho}(q_*^P) \geq F^*_{\mathcal{N},\rho}(q_i^P)$ for $i \in \{1, 2, 3\}$. Therefore, it suffices to consider how to achieve the single best quartet resolution for P, which can then be applied all subsets of four species covering P. In other words, a (binary) solution t^* to Weighted Quartet Consensus with the following inputs

- the species set contains just one taxa from each block of $Part(v)$, and
- the weight of a quartet $q = a, b|c, d$ is $\beta^P F^*_{\mathcal{N},\rho}(q)$ where $P = \{g(a), g(b), g(c), g(d)\}$ where g is a function that maps species in S to their partition $Part(v)$ (note that β^P returns 0 if $|P| < 4$)

achieves the maximum score possible for Equation 14. It follows that turning each taxa a in t^* into a vertex incident to all taxa from block $g(a)$ maximizes Equation 13/14. The resulting tree is a refinement of T_v and contains no edges that split taxa in the same block of $Part(v)$, by construction.

Putting the tree and star parts together. It follows from bipartition compatibility that there exists a compatibility supertree for $\{B^*_v : v \in V_{>3}(T_N)\} \cup \{T_N\}$ that maximizes the quartet score when there is no error in the quartet concordance factors (see Chapters 3 and 7 in [28] for a review of compatibility); this gives us our result. □

Corollary 1. *Let $\mathcal{N} = (N, \Theta)$ be a binary metric species network on S with positive lengths on its tree edges. Further assume that $\mathcal{N}$ is relaxed quartet-nonanomalous. Then, given a set of k (binary) gene trees generated under $NMSC(\rho)$, ASTRAL converges to a refinement of the tree of blobs for N with high probability as the number of gene trees goes to infinity.*

This follows from the same logic as ASTRAL's consistency under the standard MSC [20]. At the high level, ASTRAL solves WQC exactly within a constrained search space, containing the bipartitions induced by the input gene trees, and all bipartitions have probability strictly greater than zero under the NMSC.

7 False positive edges can be removed by testing a linear number of quartets around each branch with high probability

Given a TOB refinement, we need to contract false positive edges to obtain the TOB. To determine whether a branch should be contracted, we employ the same hypothesis tests as TINNiK, which are applicable to subsets of four taxa; see [1, 4] for details. In brief, the hypothesis tests takes quartet concordance factors for a 4-taxon subset as input and evaluates whether the evolutionary relationship among the four taxa is tree-like under the NMSC; the rejection of this hypothesis indicates the edge should be contracted [1]; this outcome will occur with high probability as the number of gene trees tends to infinity, because the input quartet concordance factors will converge to their expected values with high probability. Thus, the hypothesis tests enable consistent inference of whether 4 taxa are collectively related through a blob (i.e., the induced subnetwork on the 4 taxa contains a 4-blob; see Fig. S1E and Fig. S3B) while making no assumptions about the blob structure. However, the T3 test, in particular, assumes the network is quartet-nonanomalous. We make this same assumption for reconstructing a TOB refinement and only need to contract an edge in the TOB refinement if the four taxa around the edge are related through a 4-blob, thus, the T3 test is quite useful for our purposes.

However, the specific subset of four taxa sampled around an edge matters. Take Figure S4 as an example. To contract the false positive edge e_{FP} inducing bipartition $Bip(e_{FP}) = 4, 5, 6, 7|1, 2, 3, 8, 9, 19, 11, 12, 13$ and quadrapartition $Quad(e_{FP}) = 4, 5|6, 7|3|1, 2, 8, 9, 10, 11, 12, 13$, we apply the statistical test two 4 taxa sampled around the edge, selecting either two taxa from each block of the bipartition <u>or</u> one taxa from each block of the quadrapartition. We let $Q_2(e)$ and $Q_4(e)$ denote all possible subsets formed in this way, respectively (note that $Q_4(e) \subset Q_2(e)$). Critically, the taxa selected matters for detecting whether the edge is in a blob or not. In our example, the induced subnetwork on $\{3, 5, 7, 8\}$ contains a 4-cycle and displays two quartets: $3, 5|7, 8$ and $3, 8|5, 7$; thus the tests on expected qCFs for this 4-taxon subset will fail, indicating that we should contract the edge e_{FP}. In contrast, the induced subnetwork on $\{3, 4, 7, 8\}$ does not contain a 4-cycle and displays exactly one quartet $3, 4|7, 8$; thus, the tests on the expected qCFs for this 4-taxon subset will not fail, so likewise, we will not contract edge e_{FP}.

We now describe conditions for sampling subsets of four taxa around an edge that when satisfied ensure successful detection and contraction of false positive edges.

Lemma 4. *Let $\mathcal{N}^+ = (N^+, \Theta)$ be a directed binary species network on taxon set S with positive lengths on its tree edges. Let e be a false positive edge in the TOB refinement for N^+, associated with blob $\mathcal{B}$ inducing m-partition $P_1|...|P_m$, with $m \geq 4$. Then, edge e will be successfully contracted based on the results of hypothesis testing on the expected qCFs, if we select 4 species around the e, denoted $(a_1, a_2, a_3, a_*) \in Q_2(e)$, such that the following two conditions holds:*

1. *Each taxon is selected from a different block of the partition, i.e., $a_1 \in P_i, a_2 \in P_j, a_3 \in P_s, a_* \in P_t$ where $i \neq j \neq s \neq t$.*

2. *One of a_1, a_2, a_3, a_* is in a block of the partition that attaches to $\mathcal{B}$ via a cut edge of $\mathcal{B}$ incident to a reticulation vertex in $\mathcal{B}$.*

Proof. As previously mentioned, if the induced subnetwork on a_1, a_2, a_3, a_* contains a 4-blob, the hypothesis test will have a low p-value for rejecting the hypothesis and thus detecting the blob [2, 1]. Therefore, it suffices to show that when these two conditions hold the induced subnetwork on a_1, a_2, a_3, a_* contains a 4-blob.

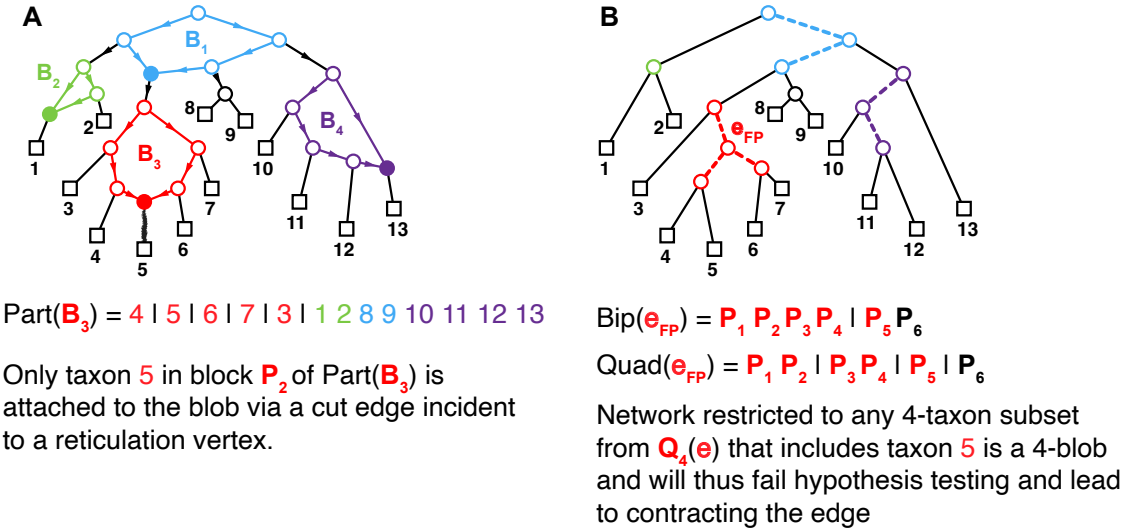

Figure S4: **Testing 4-taxon subsets around edges. A)** A level-1 directed phylogenetic network N^+ with four reticulation vertices filled in different colors. All blob edges and vertices are in the same color. The 6-partition induced by blob B_3 is shown. Only sampling from partition P_2 yields a taxa that attaches to the blob via cut edge incident to a reticulation vertex. **B)** A refinement of the tree of blobs for N^+. A false positive edge e_{FP} that needs to be tested. Drawing one taxa from each block of $Quad(e_{FP})$ will cause the hypothesis test to fail and the edge to be contracted, provided the subset includes taxon 5.

Without loss of generality, let a^* be the descendant of a reticulation node in $\mathcal{B}$ that is incident a cut edge for $\mathcal{B}$ (note that there must exist such a^*, otherwise $\mathcal{B}$ is not a biconnected component). Let $B(a)$ denote that <u>vertex</u> that is incident to the cut edge that connects taxon a to the blob $\mathcal{B}$ in N^+. Since a_1, a_2, a_3, a_* are not in the same P_i for any $i \in [m]$, $B(a_1) \neq B(a_2) \neq B(a_3) \neq B(a_*)$. By the biconnectivity, there must exist two disjoint directed paths from the root of $\mathcal{B}$ to $B(a_*)$. Let $\mathcal{B}|_2$ be the subgraph containing the two disjoint paths. By the definition of blob, we know that for any $a_i, \in \{a_1, a_2, a_3\}$, either $B(a_i)$ is in $\mathcal{B}|_2$ or it is reachable from some vertex in $\mathcal{B}|_2$ in the undirected network N. Let $V_B(a_i)$ be the set of vertices that can reach $B(a_i)$ without visiting any vertex in $\mathcal{B}|_2$ (note that we call $V_B(a_i) = \{B(a_i)\}$ trivial). Then, there must exist $a'_1 \in V_B(a_1), a'_2 \in V_B(a_2), a'_3 \in V_B(a_3)$ such that $a'_1 \neq a'_2 \neq a'_3 \neq B(a_*)$ due to biconnecitivity and the fact that N is binary. Also, by the biconnecitivity, if $V_B(a_i)$ and $V_B(a_j)$ are non-trivial, then there must exist two mutually disjoint paths from $V_B(a_i)$ to a_i and $V_B(a_j)$ to a_j, for any $i, j \in \{1, 2, 3\}$. Only keeping (1) such pairwise disjoint paths or a single vertex $B(a_i)$ if $V_B(a_i)$ is trivial, (2) $\mathcal{B}|_2$, and (3) an arbitrary path from $B(a_i)$ to a_i in N for all $a_i \in \{a_1, a_2, a_3, a_*\}$ yields a 4-blob, after deleting the remaining vertices and suppressing any vertices of degree of 2. □

Importantly, we note that the converse of Lemma 4 is not true. Put simply, there are 4-taxon subsets around e that don't meet these criteria but would still result in rejection of the null hypothesis. As an example, in Figure S1 condition 2 does not hold for $\{a, c, f, g\}$ but the induced subnetwork on these taxa is a 4-blob. However, condition is guaranteed to hold for some taxon as shown in the following Lemma, included for completeness.

Lemma 5. *Let N^+ be a directed binary phylogenetic network with a finite number of vertices. If $\mathcal{B}$ is a blob of N^+, then there exits a reticulation node in $\mathcal{B}$ that is incident to a cut edge of $\mathcal{B}$.*

Proof. For the sake of contradiction, assume all reticulation nodes in v are only incident to blob edges. Let u_0 be a vertex in $\mathcal{B}$ incident to a cut edge. Then we know that u_0 is not a reticulation node, and due to biconnectivity, there must exist two disjoint paths from the root of the blob r' to u_0 in the undirected network N. Then there must exist a reticulation node in one of the two disjoint paths in the undirected network N; otherwise, the orientation of N^+ is invalid, forming a directed cycle. Let the reticulation node

be u_1, and we know that u_0 is an ancestor of u_1. Then u_1 is not incident to a cut edge, so the child of u_1 must be a blob node, and as before, we know that on the two disjoint paths from r' to the child of u_1 in undirected N and there must be a reticulation node u_2 on the paths (noting that u_2 might be the child of u_1 when the N^+ is not a tree child network). We also know that u_1 is an ancestor of u_2, and thus $u_2 \neq u_1, u_0$ Continuing this argument, we can have an arbitrarily large sequence of distinct nodes $\{u_0, u_1, u_2, ...\}$, which contradicts the fact that N^+ has a finite number of vertices. Thus, there must exist a reticulation node that is incident to a cut edge. □

Now we consider how to design algorithms to ensure that we sample at least one 4-taxon subset that meets the conditions of Lemma 4.

Definition 4 (Disjoint cover)**.** *Given a binary directed phylogenetic network N^+ of taxa set S. Let $Q(\mathcal{A}, e)$ denote the 4-taxon subsets sampled around edge e by some algorithm $\mathcal{A}$. Let $Q^*(e)$ 4-taxon subsets around edge e such that the two following condition holds.*

1. *If e is a (false positive) edge e in the TOB refinement for N^+, associated a blob $\mathcal{B}$ inducing m-partition $P_1|...|P_m$, then for any $(a, b, c, d) \in Q^*(e)$, with $m \geq 4$, $a \in P_i, b \in P_j, c \in P_s, d \in P_t$ where $i \neq j \neq s \neq t$.*

2. *Every taxon is included in at least one 4-taxon subset that is tested, that is $\cup_{(a,b,c,d)\in Q^*(\mathcal{A},e)}\{a\} \cup \{b\} \cup \{c\} \cup \{d\} = S$.*

Condition 1 ensures we test subsets with all four taxa drawn from different blocks of the m-partition, as required by condition 1 of Lemma 4; this is guaranteed to hold when sampling around an edge using its induced quadrapartition but not its induced bipartition (as its possible to take two taxa from the same block of the m-partition). Condition 2 ensures that we test at least one subset with a taxon that attaches to the blob via a cut edge that is incident to a reticulation vertex, as required by condition 2 of Lemma 4. It follows that any algorithm $\mathcal{A}$ that disjointly covers a TOB refinement, yields a statistically consistent estimator of the TOB.

Theorem 3. *Let $\mathcal{N}^+ = (N^+, \Theta)$ be a binary metric species network on S with positive lengths on its tree edges, and let T_N denote its tree of blobs. Further assume that $\mathcal{N}^+$ is relaxed quartet-nonanomalous. Then, given a set $\mathcal{T}$ of k (binary) gene trees generated under $NMSC(\rho)$ given $\mathcal{N}^+$ and a refinement of the TOB for N, denoted $T|'_N$, an algorithm A that contracts edges in $T|'_N$ by applying hypothesis tests to disjoint covers of each branch is a consistent estimator of T_N.*

It is easy to see that the following algorithm produces a disjoint cover for branch e with just linear number of 4-taxon subsets.

Because the algorithm above must be applied to each edge in a TOB refinement and testing each 4-taxon subset takes constant time, the entire testing step is $O(n^2)$, where n is the number of taxa. Thus, our framework is dominated by quartet-based tree reconstruction. Although WQC is NP-hard, there are fast and accurate methods available. TREE-QMC, for example, has time complexity $O(n^3k)$, where n is the number of species and k is the number of genes. 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 (tripartitions) in the input gene trees, and has the advantage of a consistency guarantee due to its constrained dynamic programming approach.

Remarks on sampling. Our linear time algorithm, called 3 fix 1, relies on quadrapartitions to ensures that condition 1 of Lemma 4 is met. Simultaneously, it avoids testing the same 4-taxon subsets on different edges because for any two edge e_1, e_2, $Q_4(e_1) \cap Q_4(e_2) = \emptyset$ (the intuition is that 4-taxon subsets are sampled according to the anchor vertices for the induced quartet in the tree, as discussed by many [20]).

Thus, leveraging the quadrapartition enables efficient sampling. In contrast, bipartition-based sampling allows greater exploration of this space and has that 4-taxon subsets yielding low p-values may be re-used on on as many edges as they are around (the idea is that if one could identify the taxon that attaches to the blob via cut edge is incident to a reticulation vertex, then you leverage it to contract multiple edges in the TOB refinement that are associated with the blob). Additionally, as larger number of 4-taxon subsets can be tested, some of them may yield lower p-values. However, it can be difficult to ensure that the conditions

Algorithm 1: 3 fix 1 sampling

input : Euadrapartition induced by edge e, denoted $Quad(e)$
output: Set of 4-taxon subsets for statistical testing around e

$A, B, C, D \leftarrow Quad(e)$;
$a_* \in A$ `// select taxon from` A
$b_* \in B$ `// select taxon from` B
$c_* \in C$ `// select taxon from` C
$d_* \in D$ `// select taxon from` D
$Q(e) \leftarrow \{\}$;
for $a \in A$ **do**
  Add (a, b_*, c_*, d_*) to $Q(e)$;
for $b \in B$ **do**
  Add (a_*, b, c_*, d_*) to $Q(e)$;
for $c \in C$ **do**
  Add (a_*, b_*, c, d_*) to $Q(e)$;
for $d \in D$ **do**
  Add (a_*, b_*, c_*, d) to $Q(e)$;
return $Q(e)$

needed for disjoint coverability are met when performing bipartition-based sampling, as two taxa may be drawn from the same block of an m-partition associated with blob $\mathcal{B}$.

As an example, consider the quadrapartition $Quad(e) = A|B|C|D$ induced by an edge e associated with topology $A, B|C, D$. Suppose $|A| = |C| = 1$ and that $|B| = |D| = \frac{n-2}{2}$. The probability you pick a taxon from block A, C is $\frac{2}{n-1}$, and the probability you pick a taxon from B, D is $\frac{n-2}{n-1}$. Thus, the probability that you select exactly one taxa from A, B, C, D, is $\left(\frac{2}{n-1}\right)^2\left(\frac{n-2}{n-1}\right)^2 = \frac{4(n-2)^2}{(n-1)^4} \approx \frac{1}{n^2}$. Therefore, you need to sample a quadratic number of quartets uniformly at random to ensure you sample one taxa from A, B, C, D with high probability. However, this does not consider whether any of the selected taxa meet condition 2 of Lemma 4. Given these trade offs, further empirical studies evaluating sampling heuristics are merited.

8 Acknowledgements

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 Claudia Solís-Lemus and Nathan Kolbow for discussions regarding phylogenetic network simulations and applications of tree of blobs, 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.

References

[1] Elizabeth S Allman, Hector Baños, Jonathan D Mitchell, and John A Rhodes. The tree of blobs of a species network: identifiability under the coalescent. *Journal of Mathematical Biology*, 86(1):10, 2023.

[2] Elizabeth S Allman, Hector Baños, Jonathan D Mitchell, and John A Rhodes. Tinnik: inference of the tree of blobs of a species network under the coalescent model. *Algorithms for Molecular Biology*, 19(1):23, 2024.

[3] Elizabeth S Allman, Hector Baños, and John A Rhodes. NANUQ: a method for inferring species networks from gene trees under the coalescent model. *Algorithms for Molecular Biology*, 14(1):24, December 2019.

[4] Cécile Ané, John Fogg, Elizabeth S Allman, Hector Baños, and John A Rhodes. Anomalous networks under the multispecies coalescent: theory and prevalence. *Journal of Mathematical Biology*, 88(3):29, 2024.

[5] Hector Baños. Identifying species network features from gene tree quartets under the coalescent model. *Bulletin of Mathematical Biology*, 81(2):494–534, 2019.

[6] Christopher Blair and Cécile Ané. Phylogenetic trees and networks can serve as powerful and complementary approaches for analysis of genomic data. *Systematic Biology*, 69(3):593–601, 2020.

[7] Zhen Cao, Jiafan Zhu, and Luay Nakhleh. Empirical Performance of Tree-Based Inference of Phylogenetic Networks. In Katharina T. Huber and Dan Gusfield, editors, *19th International Workshop on Algorithms in Bioinformatics (WABI 2019)*, volume 143 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 21:1–21:13, Dagstuhl, Germany, 2019. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[8] Vu Dinh and Hector Baños. Misspecification strikes: Astral can mislead in the presence of hybridization, even for nonanomalous scenarios. *Molecular Biology and Evolution*, 42(3):msaf049, 2025.

[9] John Fogg, Elizabeth S Allman, and Cécile Ané. 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.

[10] David Gerard, H Lisle Gibbs, and Laura Kubatko. Estimating hybridization in the presence of coalescence using phylogenetic intraspecific sampling. *BMC evolutionary biology*, 11:1–12, 2011.

[11] Dan Gusfield, Vikas Bansal, Vineet Bafna, and Yun S. Song. A decomposition theory for phylogenetic networks and incompatible characters. *Journal of Computational Biology*, 14(10):1247–1272, 2007.

[12] Yunheng Han. *Accurate and scalable phylogeny estimation via graph cuts*. PhD thesis, University of Maryland, College Park, MD, 2025. Available at https://www.proquest.com/dissertations-theses/accurate-scalable-phylogeny-estimation-via-graph/docview/3250294073/se-2.

[13] Yunheng Han and Erin K Molloy. Improving quartet graph construction for scalable and accurate species tree estimation from gene trees. *Genome Research*, 33(7):1042–1105, 2023.

[14] John Hopcroft and Robert Tarjan. Algorithm 447: efficient algorithms for graph manipulation. *Communications of the ACM*, 16(6):372—-378, 1973.

[15] John Frank Charles Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248, 1982.

[16] Nathan Kolbow, Sungsik Kong, and Claudia Solís-Lemus. A method for massively scalable phylogenetic network inference. *bioRxiv*, 2025.

[17] Sungsik Kong, David L Swofford, and Laura S Kubatko. Inference of phylogenetic networks from sequence data using composite likelihood. *Systematic Biology*, 74(1):53–69, 2025.

[18] Manuel Lafond and Celine Scornavacca. On the weighted quartet consensus problem. *Theoretical Computer Science*, 769:1–17, 2019.

[19] Sarah Lutteropp, Céline Scornavacca, Alexey M Kozlov, Benoit Morel, and Alexandros Stamatakis. Netrax: accurate and fast maximum likelihood phylogenetic network inference. *Bioinformatics*, 38(15):3725–3733, 2022.

[20] S. Mirarab, R. Reaz, Md. S. Bayzid, T. Zimmermann, M. S. Swenson, and T. Warnow. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548, 2014.

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

[22] Charles-Elie Rabier, Vincent Berry, Marnus Stoltz, Joao D Santos, Wensheng Wang, Jean-Christophe Glaszmann, Fabio Pardi, and Celine Scornavacca. On the inference of complex phylogenetic networks by markov chain monte-carlo. *PLoS Computational Biology*, 17(9):e1008380, 2021.

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

[24] Claudia Solís-Lemus and Cécile Ané. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLOS Genetics*, 12(3):1–21, 2016.

[25] Claudia Solís-Lemus, Paul Bastide, and Cécile Ané. Phylonetworks: A package for phylogenetic networks. *Molecular Biology and Evolution*, 34(12):3292–3298, 2017.

[26] Mike Steel. *Phylogeny: discrete and random processes in evolution*, volume 89 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics, 2016.

[27] Simon Tavaré. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology*, 26(2):119–164, 1984.

[28] T. Warnow. *Computational Phylogenetics: An Introduction to Designing Methods for Phylogeny Estimation*. Cambridge University Press, Cambridge, United Kingdom, 2017.

[29] Yufeng Wu. Inference of population admixture network from local gene genealogies: a coalescent-based maximum likelihood approach. *Bioinformatics*, 36(Supplement_1):i326–i334, 2020.

[30] Yun Yu, James H. Degnan, and Luay Nakhleh. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLOS Genetics*, 8(4):1–10, 04 2012.

[31] Yun Yu, Jianrong Dong, Kevin J. Liu, and Luay Nakhleh. Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences*, 111(46):16448–16453, 2014.

[32] Chao Zhang, Maryam Rabiee, Erfan Sayyari, and Siavash Mirarab. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, 19(6):153, 2018.