Algorithms for Molecular Biology

RESEARCH

Open Access

TINNiK: inference of the tree of blobs of a species network under the coalescent model

Elizabeth S. Allman[1*], Hector Baños[2], Jonathan D. Mitchell[3,4] and John A. Rhodes[1]

Abstract

The tree of blobs of a species network shows only the tree-like aspects of relationships of taxa on a network, omitting information on network substructures where hybridization or other types of lateral transfer of genetic information occur. By isolating such regions of a network, inference of the tree of blobs can serve as a starting point for a more detailed investigation, or indicate the limit of what may be inferrable without additional assumptions. Building on our theoretical work on the identifiability of the tree of blobs from gene quartet distributions under the Network Multispecies Coalescent model, we develop an algorithm, TINNiK, for statistically consistent tree of blobs inference. We provide examples of its application to both simulated and empirical datasets, utilizing an implementation in the `MSCquartets 2.0` R package.

Keywords Phylogenetic network, Tree of blobs, Coalescent model

Mathematics Subject Classification 92D15, 92D20

*Correspondence:
Elizabeth S. Allman
e.allman@alaska.edu
[1] Department of Mathematics and Statistics, University of Alaska, Fairbanks, AK, USA
[2] Department of Mathematics, California State University San Bernadino, San Bernadino, CA, USA
[3] School of Natural Sciences (Mathematics), University of Tasmania, Hobart, TAS, Australia
[4] ARC Centre of Excellence for Plant Success in Nature and Agriculture, University of Tasmania, Hobart, TAS, Australia

Background

The availability of genome-scale datasets has led to a shift in focus of methodological work in phylogenetics. The Multispecies Coalescent (MSC) model, which captures how incomplete lineage sorting (ILS) may lead to gene trees discordant with one another and a species tree, now provides the theoretical basis for many approaches to species tree inference [1–7]. However, the analysis of genomic sequence data has also made clear that using trees to model species relationships can be inadequate.

Species networks allow for the description of more complex patterns of sequence evolution produced by hybridization or other forms of lateral gene transfer. Such a network may show tree-like evolution in some parts, with other parts, called *blobs*, displaying reticulations indicating transfers of genetic material between populations. These blobs may range in complexity from simple isolated cycles with a single reticulation to arbitrarily complex structures with numerous reticulations. Since some forms of gene transfer are believed to be more likely among closely related species, and thus occur when ILS is also present, the Network Multispecies Coalescent (NMSC) model is usually adopted to describe the combined effects of both gene transfer and ILS in the formation of gene trees [8–12].

Inference of a species network under the NMSC model, however, poses major challenges. Simultaneous inference of gene trees and species networks from sequences in a Bayesian framework is computationally demanding, with successful attempts limited to very small datasets [13, 14] of few taxa and genes. Inference of gene trees by standard phylogenetic methods, with these inferred gene trees treated as "data" for a second stage of species network inference, allows for the analysis of larger datasets. Since likelihood inference of a network still requires substantial computational effort, optimization of pseudolikelihood on summary statistics may be used instead. Leaving aside whether a Bayesian or pseudolikelihood analysis is preferred, exploring network space completely is impractical even for modest numbers of taxa, and limits on network complexity are often imposed. Data summary network methods hold the most promise for analysis of many-taxon, genome-scale datasets, though more work on computational approaches is still needed.

A pseudolikelihood, data-summary approach is taken by PhyloNet [15], using gene tree rooted triples, and SNaQ [16], using gene quartets, with both requiring pre-specification of the number of reticulations. Additional speed is obtained in SNaQ by limiting networks to a level-1 structure. NANUQ, [17], also based on quartets, attains considerably greater speed by limiting statistical testing to gene tree quartets and then using combinatorial methods for network building. NANUQ also is limited to level-1 networks but can give some indication of when the level-1 hypothesis is violated. Finally, PhyNEST [18] also performs level-1 quartet-based pseudolikelihood inference, but uses genomic site pattern data with the assumption that all sequences on all gene trees were generated under the Jukes-Cantor model of site substitution.

While assuming level-1 structure is helpful computationally, as these methods show, it is unlikely to be justifiable in all biological settings. Nonetheless, some limit on network complexity is necessary for acceptable computational time, and even networks only slightly more complicated than level-1 may lack identifiability from certain data types [19].

The algorithm presented in this work takes a step toward addressing this problem, by inferring the *tree of blobs* [20] of an arbitrary species network. In this tree only cut edges of the network remain while the blobs are shrunk to nodes. (See, for example, Fig. 1.) Thus multifurcations in the tree of blobs represent potentially quite complicated reticulated structures for which no detailed description is given. This is similar to a "soft polytomy" in an inferred gene tree, which rather than representing a detailed evolutionary relationship indicates merely an inability to obtain the true resolution. The tree of blobs thus serves as a partial answer to how we can efficiently infer species networks, isolating those parts of the network which require additional tools to be applied — and developed — for inferring the detailed reticulate relationships.

In a previous work [21], as a byproduct of proving the theoretical identifiability of the tree of blobs under the NMSC, an algorithm to infer it from gene trees was sketched. Hypothesis tests on counts of quartets displayed across gene trees for sets of 4 taxa allow the putative determination of some *blob quartets*, that is, of sets of 4 taxa that are best related by a single blob, and other quartets that could be related by a tree. Maximum likelihood then allows for assignment of a tree topology to the latter. However since not all blob quartets can be identified directly from gene data on single 4-taxon sets, a new combinatorial inference rule is needed to combine data for multiple 4-taxon sets. As was shown in [21], repeated application of this rule is sufficient to correctly identify all blob quartets. With all blob quartets known, and tree topologies assigned to other sets of 4 taxa, we use a certain intertaxon distance that can be computed from this information and which, assuming no error, exactly fits the tree of blobs. Standard distance tree building methods which are robust to some error can then be used to infer the tree of blobs from data.

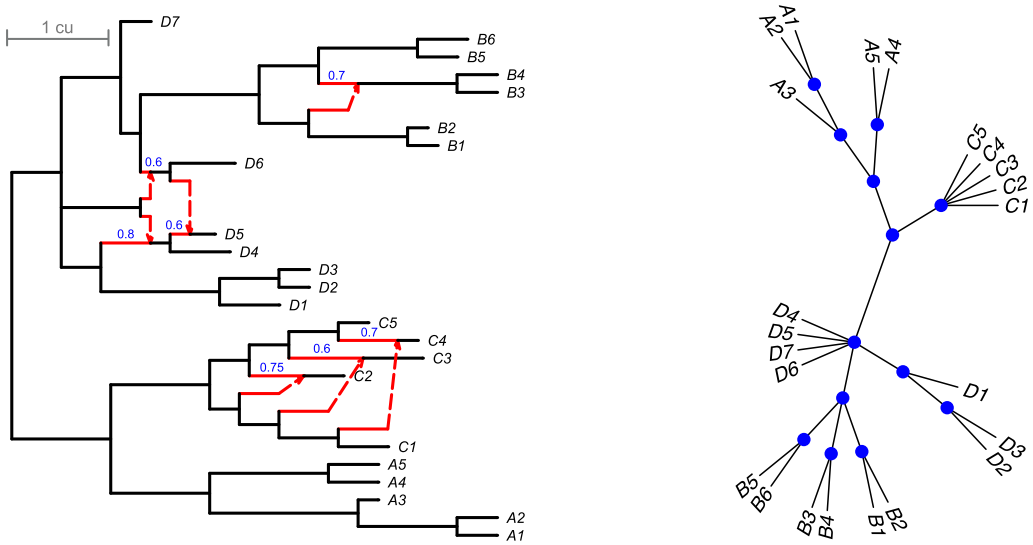

Fig. 1 (L) A species network $\mathcal{N}^+$ with branch lengths in coalescent units, and (R) its tree of blobs. Hybrid edges of $\mathcal{N}^+$ are red, with hybridization parameters in blue above the major hybrid edges. The extended Newick string is given in Appendix A. The network $\mathcal{N}^+$ is non-binary, non-level-1, non-ultrametric and non-tree-child with a 7-blob, a 6-blob, and a 4-cycle. These blobs correspond to nodes of degree 7, 6, and 4 in the true tree of blobs on the right, where blobs are shown as blue dots. The network $\mathcal{N}^+$ is used in simulations to validate the TINNiK algorithm

We provide a detailed algorithm for a fast implementation of this method, called Tree of blobs INference for a NetworK, or TINNiK,[1] as well as an implementation in the `MSCquartets` R package, v. 2.0 [22, 23]. TINNiK is a quartet-based method, motivated in part by our development of quartet-based hypothesis tests and the need for fast scalable algorithms. We show that TINNiK provides a statistically consistent estimate of a network's tree of blobs, provided its input is a sample of gene trees under the NMSC. Since in practice the input will be gene trees inferred from sequence data, some degradation of performance is to be expected. Nonetheless, sample explorations with simulated and empirical data indicate good performance and short computation time.

We know of no other proposed algorithm for tree of blobs inference from biological data. One might consider obtaining such an estimate through a modification of the NANUQ algorithm [17], by collapsing the blobs in the NANUQ splits graph to nodes. However, while it is not hard to see this would give a statistically consistent estimator in the level-1 case, nothing is known about the theoretical behavior of doing so for more general networks.

This paper is structured as follows. Section Networks and models provides basic definitions and restatements of key results from [21] which underlie our algorithm. Section Statistical testing and estimation for cut CFs presents a new statistical test to distinguish 4-taxon networks with a blob from those without one, with the derivation of the test distribution deferred to Appendix C. In Sect. The TINNiK algorithm for inference of the tree of blobs, we present our TINNiK algorithm for the inference of a tree of blobs for a species network, and show its consistency under the NMSC model. Section Simulations and Applications explores performance on both simulated and empirical datasets, with Sect. Conclusions offering concluding comments.

Networks and models

The theoretical underpinnings of the TINNiK algorithm were developed carefully in [21], so we treat the fundamental definitions and background more informally here. Readers should consult the earlier work for a more complete development.

Phylogenetic networks

We denote by $\mathcal{N}^+$ a *rooted phylogenetic network*, that is, a connected, rooted, directed graph with no directed cycles. See Fig. 1 (L) for an example. Taxa in a set X bijectively label the leaves, the degree-1 descendants of the root. Nodes are classified as *tree* or *hybrid* according to whether exactly 1 edge or more enters them. Edges are similarly classified according to their child node. We often focus on *binary* networks, in which the root is degree 2 and all internal vertices are degree 3. For formal definitions of particular classes of networks, including level-k, we recommend [24].

[1] "Tinnik" is the Inupiaq word for bearberry or kinnickinnick, a ground plant found throughout the circumpolar north.

A metric structure on the network specifies numerical parameters for the NMSC model. Edge lengths are measured in *coalescent units* (units of generations/population size), with tree edge lengths positive. Hybrid edges have non-negative lengths (with length 0 modeling instantaneous jumping of a lineage from one population to another). *Hybridization parameters* are positive probabilities that a gene lineage at a hybrid node follows a particular hybrid edge as it moves backward in time toward the root.

The *least stable ancestor* (LSA) of a network is the lowest node through which any path from the root to any taxon must pass. While a network may have a complicated structure above its LSA, our methods do not give us any information about this, nor about the location of the LSA. For this reason, our focus is on the *semidirected phylogenetic network* $\mathcal{N}^-$, obtained from $\mathcal{N}^+$ by deleting nodes above the LSA, undirecting all tree edges, and suppressing the LSA if it became a degree-2 node. Note that $\mathcal{N}^-$ is unrooted, but retains the directions of all hybrid edges. Provided no ambiguity results, the symbol $\mathcal{N}$ may denote either $\mathcal{N}^+$ or $\mathcal{N}^-$ for simplicity.

A rooted phylogenetic network $\mathcal{N}^+$ on a set of taxa X induces a network $\mathcal{N}^+_Y$ on any subset $Y \subset X$, by retaining only edges and nodes ancestral to at least one taxon in Y. Induced networks on 4-taxon sets will play a particularly important role in this work.

Blobs

A *cut edge* in a graph is one whose deletion increases the number of connected components of the graph. The following definition also applies to general graphs.

Definition 1 A *blob* on a graph is a maximal connected subgraph with no cut edges. An edge in a graph is *incident to a blob* if exactly one of its endpoints is in the blob. A blob is an *m-blob* if it has exactly m incident cut edges.

While blobs may have complicated structures, the simplest possible form is a single node, which is a *trivial blob*. For example, on a tree all blobs are trivial. The next simplest form a blob may have is that of an (undirected) cycle.

Definition 2 Gusfield et al. [20] The *strict tree of blobs*, $\mathcal{T}(\mathcal{N})$, for any connected graph, $\mathcal{N}$, is the tree obtained by contracting each of the network's blobs to a vertex, that is, by removing all of the blob's edges and identifying all its vertices.

A blob with m incident cut edges in a network leads to an m-multifurcation in the strict tree of blobs, so 2-blobs give degree-2 nodes. Since our methods cannot detect 2-blobs, we use a variant of the general notion of a tree of blobs.

Definition 3 The *reduced unrooted tree of blobs*, $\mathcal{T} = \mathcal{T}_{rd}(\mathcal{N}^-)$, of a rooted phylogenetic network $\mathcal{N}^+$ is obtained from the strict tree of blobs of the semidirected network $\mathcal{N}^-$ by suppressing all degree 2 nodes.

For the remainder of this work, the strict tree of blobs plays no role. Therefore, we refer to the reduced unrooted tree of blobs simply as the 'tree of blobs $\mathcal{T}$.' See Fig. 1 (R) for an example.

Quartets

We use two distinct classifications of sets of 4 taxa as quartets, expressing different relationships of these sets to the structure of a network. The first is the standard notion of a quartet [25] in which, for instance, $ab|cd$ refers to an unrooted topological tree with a cut edge separating the taxa a, b from c, d, and $abcd$ refers to the star tree.

A different notion of quartet captures the relationship of a set of 4 taxa to the blobs of a network. A set of 4 taxa *defines* a blob $\mathcal{B}$ if there are 4 disjoint undirected paths from $\mathcal{B}$ to these taxa. The taxa define $\mathcal{B}$ precisely when deleting $\mathcal{B}$ and its incident edges leaves the 4 taxa in distinct connected components.

Definition 4 Allman et al. [21] A set $Q = \{a, b, c, d\}$ of 4 taxa on an n-taxon network is a *Blob quartet*, or *B-quartet*, if there is a blob on the network which is defined by Q.

If a set of 4 taxa is not a B-quartet on a network, then it is a *tree-like quartet*, or *T-quartet*.

A B-quartet $Q = \{a, b, c, d\}$ on $\mathcal{N}^+$ induces the unresolved quartet topology $abcd$ on the tree of blobs $\mathcal{T}$ of $\mathcal{N}^+$, while a T-quartet induces a resolved quartet topology on $\mathcal{T}$. Note, however, that a B-quartet on $\mathcal{N}^+$ may become a T-quartet on an induced network $\mathcal{N}^+_Y$. For instance, if $\mathcal{N}^+$ is a 5-taxon network with a single blob which is a 5-cycle (i.e., a 5-*sunlet* network) and Q is the 4 taxa not descended from the hybrid node, then Q is a B-quartet on $\mathcal{N}^+$, but a T-quartet on $\mathcal{N}^+_Q$. See, for example, Figure 11 in Appendix B. In contrast, T-quartets on a large network remain T-quartets on induced subnetworks.

Quartet concordance factors and the B-quartet inference rule

The NMSC model on a metric phylogenetic network determines a distribution of binary metric gene trees, and, through marginalization, distributions of binary topological gene trees on subsets of taxa. For subsets of 4 taxa, these distributions have a special name.

Definition 5 Let $\mathcal{N}^+$ be a metric rooted phylogenetic network on a taxon set X, and $a, b, c, d \in X$ distinct taxa. The *(quartet) concordance factor* $CF_{ab|cd} = CF_{ab|cd}(\mathcal{N}^+)$ is the probability under the NMSC model on $\mathcal{N}^+$ that a gene tree displays the quartet $ab|cd$. The *(vector quartet) concordance factor*, $CF_{abcd} = CF_{abcd}(\mathcal{N}^+)$ is the ordered triple

$$CF_{abcd} = (CF_{ab|cd}, CF_{ac|bd}, CF_{ad|bc})$$

of concordance factors of each resolved quartet on a, b, c, d.

Since under the NMSC on any phylogenetic network all gene trees are binary and all have positive probability, the entries of CF_{abcd} for any a, b, c, d are positive and sum to 1.

Definition 6 CF_{abcd} is said to be *cut* if two of its entries are equal, and *strictly cut* if in addition the third entry is distinct. If CF_{abcd} is strictly cut with $CF_{ab|cd} \neq CF_{ac|bd} = CF_{ad|bc}$, then we say CF_{abcd} is *strictly (ab|cd)-cut*. If CF_{abcd} is not cut, we say it is *non-cut*.

The terminology "cut" is motivated by the following theorem.

Theorem 1 *Allman et al.* [21] (*CF-detectability of 4-blobs on 4-taxon networks*) *Consider a 4-taxon rooted binary phylogenetic network $\mathcal{N}^+$ on taxa $\{a, b, c, d\}$ with quartet concordance factor CF_{abcd} and tree of blobs $\mathcal{T}$. Then under the NMSC model for generic parameters*:

(*a*) *$\mathcal{T}$ has the quartet tree topology ab|cd if, and only if, CF_{abcd} is strictly (ab|cd)-cut.*
(*b*) *$\mathcal{T}$ has the unresolved quartet topology if, and only if, CF_{abcd} is non-cut.*

In contrast to the notions of B- and T-quartets, which refer to the relationship of 4 taxa through the topology of a full network $\mathcal{N}^+$, the notions of cut and non-cut *CF*s refer to properties of the probability distribution under the NMSC, and thus depend only on the induced 4-taxon network.

Theorem 1 shows that on 4-taxon networks there is a close correspondence between these concepts. However, on a larger network they diverge, with the following theorem giving a further tool for relating them.

Theorem 2 *Allman et al.* [21] (*B-quartet Inference Rule*) *Consider a rooted binary phylogenetic network $\mathcal{N}^+$ on n taxa, $n \geq 5$. Suppose that $\{a, b, c, d\}$ and $\{b, c, d, e\}$ are B-quartets on $\mathcal{N}^+$. If on the induced 4-taxon network any one of $\{a, b, c, e\}$, $\{a, b, d, e\}$, or $\{a, c, d, e\}$ is*

(*a*) *a T-quartet, with a, e not a cherry on the reduced unrooted tree of blobs for the induced 4-taxon network, or*
(*b*) *a B-quartet,*

then all of $\{a, b, c, e\}$, $\{a, b, d, e\}$, and $\{a, c, d, e\}$ are B-quartets on $\mathcal{N}^+$.

The previous two theorems lead to a powerful result for application.

Theorem 3 *Allman et al.* [21] *On an n-taxon rooted binary phylogenetic network $\mathcal{N}^+$ with generic numerical parameters, all B-quartets can be identified from the quartet CFs using CF-detectability* (*Theorem 1*) *and applications of the B-quartet Inference Rule* (*Theorem 2*).

In [21], these three theorems were the key to establishing that the tree of blobs of an arbitrary binary species network is identifiable from gene quartet concordance factors. In this work, they form the basis of an algorithm to infer that tree of blobs.

Statistical testing and estimation for cut *CF*s

A key component of the TINNiK algorithm for inference of the tree of blobs is testing gene tree data to determine which sets of four taxa are in accord with a cut *CF*. For this, we introduce a new hypothesis test.

Cut model testing and maximum likelihood inference

For any phylogenetic network, a *CF* is a point in the interior of the 2-dimensional probability simplex, $\Delta^2 = \{(p_1, p_2, p_3) \mid p_i > 0, \sum p_i = 1\}$.

Definition 7 The *cut model* comprises those points in Δ^2 representing cut *CF*s, that is $\{(p_1, p_2, p_3) \in \Delta^2 \mid p_i = p_j \text{ for some } i \neq j\}$, as depicted in Fig. 2 (L).

Data relevant to a *CF* is collected in the form of a *quartet count concordance factor (qcCF)* [27], a vector of counts

$$qcCF_{abcd} = (m_{ab|cd}, m_{ac|bd}, m_{ad|bc})$$

of the three resolved unrooted topological quartet gene trees, which for the taxon set $\{a, b, c, d\}$ are assumed to be independently drawn from the NMSC. In practice, these could be quartet trees individually inferred from sequence data for different genes, or quartet trees displayed on inferred gene trees on more taxa. However, our development of a statistical test assumes no inference error is present.

With total sample size $m = m_{ab|cd} + m_{ac|bd} + m_{ad|bc}$, the *empirical concordance factor*, which consistently estimates the concordance factor, is

$$\widehat{CF}_{abcd} = (\widehat{CF}_{ab|cd}, \widehat{CF}_{ac|bd}, \widehat{CF}_{ad|bc}) = qcCF_{abcd}/m.$$

Viewing $\widehat{CF}_{abcd}$ as a point in the simplex, closeness to the cut model lines lends informal support that the true CF_{abcd} is cut, while a greater distance supports that CF_{abcd} is non-cut. For judging closeness, however, one must take the sample size m into account.

To formulate a formal hypothesis test, fix four taxa a, b, c, d, and the data $qcCF_{abcd} = (m_{ab|cd}, m_{ac|bd}, m_{ad|bc})$. Assuming $qcCF_{abcd}$ arises as a trinomial sample from the distribution specified by some true CF_{abcd}, consider null and alternative hypotheses:

H_0: CF_{abcd} *is cut,*
H_1: CF_{abcd} *is non-cut.*

For a test statistic, we use the likelihood ratio statistic for the null and alternative models, with Appendix C.1 presenting the necessary calculations for the test.

Because the cut model has a singularity at (1/3, 1/3, 1/3) (Fig. 2 (L)), standard assumptions underlying the routine use of the χ^2_1 distribution for judging the test statistic are violated there. However, *CF* points near the centroid include those for trees and networks with short internal branches, and thus include some of those of the greatest interest to researchers. Building on work in [28], we thus develop an alternative testing distribution that takes into account this geometry of the cut model. Appendix C.2 presents its derivation and Appendix C.3 simulations illustrating its improved performance over the χ^2_1 distribution near and at the cut model singularity.

The *T*3 model and testing

An *anomalous quartet* $\{a, b, c, d\}$ is one whose *CF* is $ab|cd$ cut of the form (p, q, q) with $q \geq 1/3$. Graphically, this means the *CF* lies on the cut model depicted in Fig. 2 (L), but not on the *T*3 model shown in Fig. 2 (R). In [17], anomalous quartets for level-1 networks were investigated and shown to require a 3-cycle with two taxa descended from the hybrid node, and somewhat extreme numerical parameters which seem unlikely biologically. Further investigation by Ané et al. [29] suggests that anomalous quartets for more complex networks are also not likely to be common. For this reason, when inferring a tree of blobs it can be reasonable to assume that an unknown network has no anomalous quartets and use a *T*3 hypothesis test for *CF*s, rather than a cut test. The *T*3 test, developed in [28], is a backbone for the NANUQ method [17] for inferring level-1 topological networks under the NMSC.

Using the *T*3 test one might infer more B-quartets than using the cut test, as any $\widehat{CF}$s near the cut model line

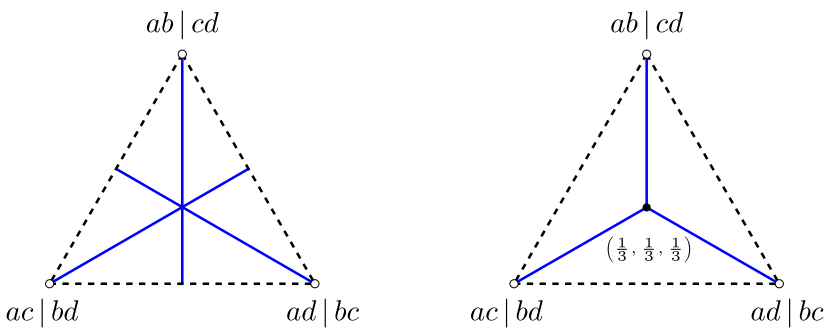

Fig. 2 Geometric view of *CF*s for 4-taxon network models, with dashed lines outlining the simplex Δ^2. Each point in Δ^2 arises as a *CF* under the NMSC, even when restricting to level-1 networks [26]. (L) The cut model consists of 3 blue line segments, with each formed by *CF*s arising from 4-networks with a specific resolved tree of blobs topology. *CF*s off the cut model arise only from networks with unresolved trees of blobs. (R) The *T*3 submodel is those cut *CF*s with smallest entry occurring exactly twice. Using it as the null hypothesis in the TINNiK algorithm may lead to more sets of 4 taxa initially judged as B-quartets than using the cut test

segments of the form (p, q, q) with $q > 1/3$ might support the null hypothesis for the cut test, but be rejected by the $T3$ test and flagged as indicating 4-blobs. Thus, using the $T3$ test can produce a less resolved tree of blobs than the cut test. While this is not a conservative approach in the sense of hypothesis testing (since it may lead to more rejections of a null hypothesis of a tree-like quartet relationship), the inferred tree of blobs it produces is a more cautious one that possibly avoids depicting erroneous resolution.

Although our implementation of TINNiK in `MSC_quartets` has a default option of the $T3$ test, we recommend performing data analysis with both tests. For many datasets we have found they give identical results, since they either infer the same initial B-quartets, or the B-quartet inference rule compensates for missing some of these initially using the cut test. When the results of using the two tests differ, investigating why that occurred may provide more insight into the data.

The TINNiK algorithm for inference of the tree of blobs

Building on the theorems and hypothesis tests from previous sections, we present a detailed algorithm, Tree of blobs INference for a Species NetworK, or TINNiK, for inferring the reduced unrooted topological tree of blobs of a network $\mathcal{N}^+$ from multigene data. We then analyze its running time and show its statistical consistency under the NMSC model for binary $\mathcal{N}^+$.

TINNiK applies hypothesis tests to classify empirical *CF*s as cut or non-cut, giving quartet trees of blobs, using Theorem 1. Then it repeatedly but efficiently uses the inference rule of Theorem 2 to infer all B-quartets for a network. TINNiK's next step is to use a quartet-based intertaxon distance formula [6] to convert B- and T-quartet information to a distance approximately fitting the topological tree of blobs. Then an inferred tree of blobs can be obtained by any of a number of well-known tree-building algorithms such as Neighbor-Joining [30], DescentTree [31], or FastME [32]. If the quality of the input data is unknown, or its fit to the NMSC model is doubted, we recommend the use of the Neighbor-Net algorithm [33] to confirm the distance reflects a strong tree signal before tree building.

One concern for algorithm design is how to handle empirical *CF*s that are near (1/3, 1/3, 1/3). These might arise from either a true multifurcation in a network (a hard polytomy), a "near multifurcation" of a resolved subnetwork with short internal edges (a soft polytomy), or from complex blobs with longer edges. TINNiK treats all *CF*s judged by a "star tree" hypothesis test to be close to (1/3, 1/3, 1/3) as B-quartets. While this is a natural approach, it does mean that the inferred tree of blob's structure may reflect both true blobs and further multifurcations due to data quality that is insufficient to resolve some cut edges. This may cause inferred blobs to be larger than true ones, but only when the data is inadequate to obtain greater resolution.

Algorithms

The B-quartet Inference algorithm takes as input a table of quartet count concordance factors (*qcCF*s). Working from a collection of m gene trees, all on the full set X of N taxa, this quartet table can be produced in time $\mathcal{O}(mN^4)$. Although under the NMSC model all gene trees are fully resolved, in applications some inferred trees may not be, but these can be handled either by discarding their unresolved quartets or assigning uniform "counts" of 1/3 to each resolved topology, as discussed in [27].

In order to access entries of this table rapidly, without scanning it in its entirety, we require that its rows, corresponding to sets of 4 taxa, be ordered so that the index for any set can be computed directly and quickly. We choose to order the sets of four taxa by *lex order*. In more detail, this means that if the taxa are designated by the numbers $1, 2, 3 \ldots, N$, and a set of 4 taxa is designated by a "word" of the four numbers in ascending order, then these words are ordered lexicographically using the usual order on natural numbers. Thus the first few sets are ordered as

$$(1,2,3,4), (1,2,3,5), \ldots, (1,2,3,N), (1,2,4,5), (1,2,4,6), \ldots.$$

In lex order, the index for a particular set of 4 taxa is given by the formula (Corollary 3.22, [34])

$$\rho(n_1, n_2, n_3, n_4) = \binom{N}{4} - \sum_{i=1}^{4} \binom{N - n_i}{5 - i}.$$

The computational simplicity of this formula allows for its rapid evaluation. Tabulating all binomial coefficients that might be needed in the formula in advance, so they are computed only once, requires time $\mathcal{O}(N)$. After this, however, the index for any set of 4 taxa can be computed in time $\mathcal{O}(1)$.

Two hypothesis tests are used in the algorithm. First, we use a star tree test to determine whether each *qcCF* is consistent with a 4-polytomy for each induced 4-taxon network. The null hypothesis is that the *CF* for the 4 taxa is (1/3, 1/3, 1/3), with the alternative its complement in the simplex. A standard χ^2_2 test at some level β on the likelihood ratio statistic is performed. Failure to reject the null suggests either a true 4-polytomy, or lack of sufficient information to infer a resolution.

In addition, one of the cut or $T3$ hypothesis tests described in Sect. Testing and Estimation is used in the algorithm to decide when a vector *qcCF* for four taxa is in accord with a cut relationship. More formally, rejecting the null hypothesis of this test at level α is interpreted as indicating a non-cut *CF*. By Theorem 1, this is evidence the 4 taxa form a B-quartet on the induced 4-taxon species network.

The following algorithm applies these tests and the inference rule for B-quartets of Theorem 2.

Algorithm (B-quartet Inference)

Input: An $\binom{N}{4} \times 3$ table of *qcCF*s for a set of N taxa with rows in lex order corresponding to subsets of 4 taxa, and columns corresponding to resolved quartet topologies, a choice of test "T3" or "cut," and significance levels $\alpha, \beta > 0$ for judging p-values in hypothesis tests on *qcCF*s.

Output: A vector B of length $\binom{N}{4}$ with entries corresponding to sets of 4 taxa in lex order, whose 1/0 entries indicate a set is/is not inferred to be a B-quartet.

1. *Initialization:*

   Create a $\binom{N}{4}$-element indicator vector B for the lex ordered sets of 4 taxa with all entries 0, indicating that no B-quartets are currently known. Create empty lists $L1$, $L2$ for iteratively storing indices of newly-found B-quartets. Compute binomial coefficients for use in indexing.
2. *Hypothesis testing:*

   (a) Apply a χ^2_2 test at level β as described above to each *qcCF* to decide which sets of 4 taxa are viewed as B-quartets because they are in accord with a *CF* of (1/3, 1/3, 1/3). Set the entries of B for these sets to be 1, and append the indices to $L1$.

   (b) Apply the $T3$ or cut hypothesis test at level α to each *qcCF* in the table not already judged as B-quartet, to decide which sets of 4 taxa are viewed as *CF*-detectable B-quartets. Set the entries of B for these to 1, and append the indices to $L1$.

   (c) For those sets not inferred as B-quartets, infer a maximum likelihood estimate of a quartet tree topology for the 4-taxon tree of blobs, for use in 3(a)(i)(β). In case of a tie, choose uniformly at random.

3 *Inference rule:*

   (a) Loop over the entries of $L1$, each corresponding to a newly-determined B-quartet, say $\{a, b, c, d\}$.

   (i) Loop over the $4(N-1)$ sets of 4 taxa which have exactly 3 taxa in common with {a, b, c, d} . For concreteness, say such a set is {a, b, c, e} with e $\neq$ d. If {a, b, c, e} is a known B-quartet, then

      (α) Check B to see if any of {a, b, d, e}, {a, c, d, e}, {b, c, d, e} are B-quartets, and go to (γ) if one is found.

      (β) Check if any of the quartet trees $ea|bd$, $eb|ad$, $ea|cd$, $ec|ad$, $eb|cd$, $ec|bd$ was inferred in 2(c). If not, continue loop (i).

      (γ) Update the B vector to designate {a, b, d, e}, {a, c, d, e}, and {b, c, d, e} as B-quartets, and store indices of any newly-identified B-quartets in $L2$.

      Continue loop(i)

      Continue loop(a)

   (b) if $L2$ is not empty, store $L2$ into $L1$, void $L2$, and go to (a).

4. Return B.

In this algorithm, step 2 implements the theoretical *CF*-detectability result of Theorem 1, while step 3 implements the B-quartet inference rule of Theorem 2. The algorithm eventually considers every pair of B-quartets sharing three taxa, since any time a new one is discovered it is compared to all quartets that share three taxa with it. If one of these is a not-yet-inferred B-quartet, then this pair will be compared again later, once that quartet is inferred as a B-quartet. Theorem 3 therefore ensures the looping of step 3 can determine all B-quartets, assuming sufficient data in accord with the NMSC.

Steps 1 and 2 can each be accomplished in time $\mathcal{O}(N^4)$. Since there can be at most $\binom{N}{4}$ B-quartets that can appear in the lists L through all passes through

step 3a, and each is compared in step 3(a)i to $\mathcal{O}(N)$ other sets of 4-taxa, all applications of step 3 require time at most $\mathcal{O}(N^5)$. Thus the total time complexity is $\mathcal{O}(N^5)$.

To estimate the tree of blobs for a network, we seek a tree that displays unresolved quartet trees for all B-quartets and a resolved quartet tree with the topology estimated by maximum likelihood for all T-quartets. Note that the estimate of the topology is recorded when step 2c of the B-inference algorithm is performed, so we treat this as known.

Estimating the tree of blobs is now an instance of a supertree problem, with input all trees on 4 taxa, with the trees for B-quartets unresolved. To address this, we take the approach introduced in [6], in which an intertaxon distance is defined using quartet data – including that for unresolved quartets – to compute an intertaxon distance. Assuming perfect information, this distance would exactly fit the unknown tree. While inference may well lead to some incorrect quartets, distance-based tree construction methods that behave well under some noise can be used to return an inferred tree of blobs. Because of possible error in some quartets, and hence in the computed quartet distance, the inferred tree may not show exact polytomies, but rather some resolutions of them with short edges. It may thus be desirable to reduce to zero all edge lengths smaller than some cutoff δ. Theory behind such a cutoff will be discussed in the next subsection, in the proof of Theorem 4.

The full TINNiK algorithm we now outline takes as input a collection of gene trees on the taxa X, and returns an inferred topological tree of blobs for the network parameter which under the NMSC model produced those gene trees.

Algorithm (TINNiK)

Input: A collection of m unrooted topological gene trees, each on a taxon set X, with $|X| = N$, a choice of test "T3" or "cut," significance levels $\beta, \alpha > 0$ for judging p-values in hypothesis tests on *qcCF*s, and a minimum edge length $\delta \geq 0$.

Output: An estimate of the tree of blobs for the network parameter $\mathcal{N}^+$ producing the gene trees under the NMSC model.

1. Tabulate all *qcCF*s for the taxon set X across all gene trees.
2. Infer all B-quartets with the B-quartet Inference Algorithm with significance levels β, α, and the chosen test, retaining maximum likelihood topologies for all T-quartets.
3. Treating B-quartets as unresolved, and T-quartets as resolved with their inferred topologies, compute the quartet intertaxon distances of [6] for X.
4. Using a distance-based tree inference method suitable for non-ultrametric trees (e.g., NJ, FastME), infer a topological tree from the distance.
5. Set all edge lengths in the tree that are less than δ to 0.

The computational times for steps 1–5 using NJ in step 4 are, respectively, $\mathcal{O}(mN^4)$, $\mathcal{O}(N^5)$, $\mathcal{O}(N^4)$, $\mathcal{O}(N^3)$, $\mathcal{O}(N)$, for a combined $\mathcal{O}((m+N)N^4)$. We report computational times in practice below in Sect. Empirical Runtimes, when analyzing TINNiK on simulated and real data.

The TINNiK algorithm can also be applied when gene trees have missing taxa, provided each subset of 4 taxa occurs on at least one gene tree, so the qcCF is not the zero vector.

Statistical consistency

It is desirable that inference algorithms produce statistically consistent estimators. In this context, informally this means that given data (m gene trees) produced under the NMSC model on a species network, the probability of obtaining the correct tree of blobs approaches 1 as the amount of data approaches infinity. However, since the algorithms assume generic numerical parameters, and there are several other algorithm inputs, α, β, δ, a precise statement of an appropriate notion of consistency is more complicated. We proceed similarly to how consistency was addressed for the quartet-based NANUQ algorithm for inferring a level-1 species network in [17]. For simplicity, we also restrict to the case that the data is m gene trees, each on the full taxon set X, since generalizing from this is straightforward.

Before stating our formal consistency theorem, we describe explicitly what we mean by generic parameters. For a fixed topological binary species network, it is possible that the *CF* for an induced 4-taxon network with a 4-blob may be a cut *CF*. By Theorem 1, however, for each such topological 4-network the *CF* is non-cut for all parameters except those in a measure-0 subset of its numerical parameter space. Since for any full network there are only finitely many induced topological 4-networks and the finite union of measure-0 sets has measure zero, for generic parameters (*i.e.,* those outside this

measure-0 set) on the full network, all *CF*s for quartets inducing 4-blob networks will be non-cut.

We also need that for generic parameters the *CF* of an induced 4-taxon network is not (1/3, 1/3, 1/3). In the case of a 4-blob, this follows from the last paragraph. But since a 4-network without a 4-blob may have a *CF* with equal entries (e.g., a 3_2-cycle network [26]), more argument is needed that this does not occur generically. Note that if all hybridization parameters on a binary 4-network $\mathcal{Q}$ without a 4-blob are 0 or 1, then $\mathcal{Q}$ is essentially a resolved tree, for which $CF \neq (1/3, 1/3, 1/3)$. Analyticity of the parameterization then implies this inequality for generic parameters. Mimicking the argument above shows that no induced 4-taxon network has $CF = (1/3, 1/3, 1/3)$ for generic parameters on the full binary network.

These preliminary observations are used to prove the following:

Theorem 4 *For generic numerical parameters on a binary phylogenetic network $\mathcal{N}^+$, the TINNiK Algorithm using the cut test provides a statistically consistent estimate of the topological tree of blobs $\mathcal{T} = \mathcal{T}_{rd}(\mathcal{N}^-)$ under the NMSC. Specifically, there exists a sequence $\alpha_m \to 0$ such that for any $\beta > 0$ and $2 > \delta \geq 0$, the TINNiK algorithm on a set of m gene trees independently drawn from the NMSC model on a binary species network $\mathcal{N}^+$ will, with probability $\to 1$ as $m \to \infty$, infer $\mathcal{T}$.*

Proof We restrict to generic parameters ensuring that all induced 4-networks with 4-blobs have non-cut *CF*s, and no induced 4-taxon network has $CF = (1/3, 1/3, 1/3)$.

First consider step 2a of the B-quartet Inference algorithm. With generic parameters, for each set of 4 taxa the probability the χ^2_2 test with significance level β will reject the null hypothesis that a *CF* is (1/3, 1/3, 1/3) approaches 1 as $m \to \infty$. Since there are only finitely many 4-taxon subsets, the probability goes to 1 that this null hypothesis will be rejected for all. This holds regardless of the chosen value of $\beta > 0$.

In step 2b of the B-quartet Inference algorithm, the role of α in the cut test is more subtle, since if it is held fixed then we expect to erroneously reject the null model in a fraction α of all applications for each set of 4 taxa. To make such false negatives less common, we consider sequences of levels $\alpha_m \to 0$ as the number of gene trees $m \to \infty$.

The likelihood ratio statistic is judged using the distribution of Propositions 5 and 6 of Appendix C.2. If a true *CF* is cut, then as $m \to \infty$, the parameter μ_0 of that distribution goes to ∞ and the distribution converges to the χ^2_1. This holds even using the MLE in place of the true parameter. To ensure that the probability of failing to reject the null hypothesis approaches 1 as $m \to \infty$, it is enough to choose any sequence of significance levels with $\alpha_m \to 0$.

In contrast, if a true $CF = (p_1, p_2, p_3)$, is non-cut and hence not in the null model, let (m_1, m_2, m_3) denote a *qcCF* under the NMSC with sample size $m = m_1 + m_2 + m_3$, and $(\hat{p}_1, \hat{p}_2, \hat{p}_3)$ the MLE of the *CF* under the null model. Without loss of generality, assume the MLE is on the vertical line segment of Fig. 2 (L), so that

$$\hat{p}_1 = \frac{m_1}{m}, \;\; \hat{p}_2 = \hat{p}_3 = \frac{m - m_1}{2m}.$$

Using the formulas of Appendix C.1, the likelihood ratio statistic is then $\lambda = \lambda_m =$

$$\begin{aligned} &-2[m_1 \log m_1 + (m - m_1)\log((m - m_1)/2) - (m_1 \log m_1 + m_2 \log m_2 + m_3 \log m_3)] \\ &= m\left[-2\left(\left(1 - \frac{m_1}{m}\right)\left(\log\left(1 - \frac{m_1}{m}\right) - \log 2\right) - \left(\frac{m_2}{m}\log\frac{m_2}{m} + \frac{m_3}{m}\log\frac{m_3}{m}\right)\right)\right] \\ &= mY_m, \end{aligned}$$

where the random variable Y_m converges in probability to

$$d = -2\big((1-p_1)(\log(1-p_1) - \log 2) - (p_2 \log p_2 + p_3 \log p_3)\big).$$

Moreover, $d > 0$ since the unconstrained likelihood has a unique maximum at (p_1, p_2, p_3).

Now for any $\eta > 0$ there exists an M such that for $m > M$, $\mathbb{P}(Y_m > d/2) > 1 - \eta$ and thus that $\mathbb{P}(mY_m > md/2) > 1 - \eta$. Since η was arbitrary, as $m \to \infty$, $\mathbb{P}(\lambda_m > md/2) \to 1$.

Let α'_m be the probability that a χ^2_1-distributed random variable is greater than $md/2$, so $\alpha'_m \to 0$. Then since the test distribution converges to the χ^2_1, the probability of rejecting the null hypothesis is greater than $1 - \eta$ for sufficiently large m. Thus the probability of rejecting the null approaches 1.

While the value of d depended upon the particular 4-taxon set under consideration, since there are only finitely many such sets, by choosing α_m as the maximum of the α'_m we obtain a sequence of significance levels that with probability approaching 1 as $m \to \infty$ ensures the cut test will reject the null model for all induced 4-networks with a 4-blob and fail to reject it for all others. The argument so far has shown that with our choice of the α_m the hypothesis tests will lead us to correctly conclude that true *CF*s are cut or non-cut, with probability approaching 1 as $m \to \infty$.

When the true *CF* for 4 taxa is cut, its value is consistently inferred by maximum likelihood, and thus the topology of the 4-taxon reduced unrooted trees of blobs is as well. Thus as $m \to \infty$, with probability approaching 1, the remaining deterministic steps of the B-quartet Inference algorithm then correctly infer all B-quartets.

From this information on B-quartets and T-quartet topologies, TINNiK computes an intertaxon distance exactly fitting the network's tree of blobs. With no error in the distances, NJ or other tree-building algorithms recover the tree exactly.

Note that δ played no role in this argument so far, since its purpose in the algorithm is to suppress some error which, with probability approaching 1 as $m \to \infty$ is not present. We must however verify that δ has no detrimental effects in this asymptotic result. Reviewing [6], one sees that internal branches of a tree endowed with the quartet distance always have length of at least 2. Thus any $0 \leq \delta < 2$ will have no effect when all B- and T-quartets are properly determined. □

TINNiK test levels and graphical output

When TINNiK's hypothesis tests are applied, many sets of 4 taxa will overlap, so the *CF*s are not independent. Although a Bonferroni correction for multiple tests can be applied, controlling the family-wise error rate, we do not do so, as this is always equivalent to choosing a smaller significance level. Indeed, when the method is applied to inferred gene trees, which have unknown error, a fully justifiable formal correction is not known.

However, when a TINNiK analysis is reported for an empirical dataset, it should always include the values of α, β used, and whether the "cut" or "T3" test was used. Ideally, the gene trees that were used should be made publicly available, for reproducibility, as gene trees inferred by different methods might produce a different tree of blobs.

In regard to graphical output, the tree of blobs could be drawn in the usual way for phylogenies, with nodes rendered as points, but we recommend a modification. Depicting each internal node as a disk or ball gives visual emphasis that the nodes represent blobs with potentially complicated structures. Even degree-3 nodes should be shown this way, since non-node 3-blobs may exist. The implementation of TiNNiK in `MSCquartets` follows this graphical style, using red disks on an inferred tree of blobs.

Finally, although any planar drawing of the tree of blobs necessarily orders the edges emanating from a blob in some way, this circular order is essentially arbitrary. The true network may not even be embeddable in the plane without crossings, in which case no unique order is even determined. Although for certain networks (level-1, or, more generally, outer-labelled planar [19]) a unique circular order exists, TINNiK does not seek to find it, much less impose it on the tree of blobs. Viewers of a tree of blobs should keep this in mind when seeking biological insight.

Simulations and Applications

We present analyses of both simulated and empirical gene tree data, using the implementation of the TINNiK algorithm in the `MSCquartets 2.0` R package [22, 23]. Its primary functions, `TINNIK` and `TINNIK-dist`, utilize C++ code with the `Rcpp` package [35] for increased speed.

Datasets of gene trees were simulated under the NMSC on various networks using `PhyloCoalSimulations` [36]. As true samples under the NMSC, these do not have the gene tree inference error expected in empirical analyses. For analyses of empirical datasets, we used gene trees inferred and made publicly available by the researchers who originally analyzed them.

With a gene tree dataset available, TINNiK can be run quickly in R, for example using a single `MSCquartets` command:

`TINNIK(`*`gene_tree_file`*`)`

and default settings for arguments. See the vignette in the `MSCquartets` R package for an extended tutorial on using TINNiK, and Table 1 for timing information.

Simulations

A first set of simulations, analyzed in Sects. Analysis I: Varying α and Analysis II: Varying β, uses the model network $\mathcal{N}^+$ of Fig. 1. This network on 23 taxa with 7 hybrid nodes has some complicated features (*e.g.,* non-binary, non-tree-child [24]), with a tree-like cluster (*A* taxa), and three blobs (*B*s, *C*s, *D*s). The *B*-blob is descended from the *D*-blob, while the *C*- and *D*-blobs include more than one instance of gene flow.

Gene tree samples of size $n = 300, 500, 1000, 10000$ were produced, with branch lengths scaled by factors $k = 0.5, 1.0, 2.0$, for a total of 12 simulation parameter settings. These include cases where sampling error may be significant ($n = 300$), and when short branch lengths and the resulting high ILS ($k = 0.5$) may confound reticulation signal. The largest value of n should approximate asymptotic behavior. We adopt the terms 'high,' 'moderate,' and 'low' ILS for the scaling factors $k = 0.5, 1.0, 2.0$, respectively, as a convenience. Since non-matching gene tree quartets under the NMSC on a species tree with internal branch length 1 occur with probability approximately 0.25, our 'moderate ILS' is arguably 'moderately high.' Simulated gene tree datasets were analyzed using the `TINNIK` function with the default $T3$ test and varied values of α and β.

Since the $T3$ test and the star tree test have different foci (hybridization vs. lack of resolution), in Sects. Analysis I: Varying α and Analysis II: Varying β we investigate the effect of each test individually. For a general overview, Table 2 presents a summary of test levels α and simulation results for all parameter choices under the $T3$ test, with $\beta = 1$ fixed (so all network quartets are treated as resolved), and illustrates the effect that small sample size and/or high ILS may have.

To understand the effect of blob complexity, analyses in Sect. Analysis III: Varying blob complexity use a second set of simulations on the networks of Fig. 6. Network $\mathcal{N}_1^+$ has 10 taxa and a single 7-cycle, while network $\mathcal{N}_3^+$ is obtained from $\mathcal{N}_1^+$ by the addition of two hybrid edges cutting across the cycle, changing the blob from level-1 to level-3. Samples of size $n = 1000$ gene trees were simulated.

A final simulation, in Sect. Analysis IV: comparison to network inference, generated a sample of size $n = 10,000$ for the level-2 (2 overlapping cycles) network $\mathcal{N}^+$ of Fig. 7. Analyses were done with TINNiK and also SNaQ, which infers a level-1 network under the NMSC using pseudolikelihood on empirical quartet *CF*s [37]. SNaQ searches were done starting at the four level-1 networks displayed on $\mathcal{N}^+$ (obtained by deleting exactly one hybrid edge), with the user-defined maximum number of hybridizations, h_{max}, set to 1 and 2. Since theory justifies SNaQ's use only for level-1 networks, this modeling scenario violates its main assumption of network complexity, and it should not be expected to perform well. Our goal is not to point out any weakness of SNaQ, but to illustrate that TINNiK might help empiricists evaluate if an assumption made by another method is violated by contrasting its results to output from that method.

In accordance with Theorem 4, branch lengths in the TINNiK tree of blobs shorter than 2 were collapsed to zero in all analyses.

Results

We caution TINNiK users that one can rarely simply choose test levels $\alpha, \beta \in [0, 1]$ in advance (e.g., at the common level of 0.05) and obtain a strong analysis. Rather, a range of significance levels should be considered, in conjunction with viewing the resulting hypothesis test simplex plots and weighing one's understanding of the extent of noise present in inferred gene trees.

Varying α from small to large increases the number of *CF*s interpreted as signaling hybridization, potentially causing TINNiK's inferred tree of blobs to gain more or larger multifurcations. This can indicate which multifurcations have the strongest support. Even in simulated gene tree data, which has no model misspecification, the level of support can vary with network features such as blob complexity, hybridization parameter values, and location of a blob within the network. Simplex plots of test results, as discussed in [27], can help users choose values of α that give good separation of plotted *CF*s into tree-like and non-tree-like clusters.

Varying β from small to large decreases the number of *CF*s interpreted as indicating a star tree, potentially causing the inferred tree of blobs to be more resolved. If few *CF*s are plotted near the centroid (1/3,1/3,1/3) of the simplex, the value of β has little impact over a wide range. However, if many *CF*s are near the centroid, β's value can be quite impactful. In some empirical datasets that have been studied for signs of hybridization we have found *CF*s so tightly clustered near the centroid that whether any signal for hybridization exceeds likely gene tree inference error seems debatable. Again, simplex plots of test results for various β values can be a helpful guide.

Empirical runtimes

Representative runtimes are shown in Table 1. These were found using gene trees as input, and do not include the time to infer trees from sequence data. Runtimes for other α, β are similar, although the tallying of quartets need only be done once.

All times shown are a matter of seconds. In particular, TINNiK is much faster than SNaQ or PhyNEST which infer level-1 networks, and PhyloNet which seeks an arbitrary network, all of which perform pseudolikelihood optimization over (appropriate) network space. TINNiK's runtime is on par with NANUQ's for inferring a level-1 network topology. However, TINNiK gives meaningful (though coarse) output quickly without assumption on network level.

In conjunction with the theoretical time complexity given in Sect. Algorithms, these runtimes show that TINNiK easily scales to much larger datasets than are likely to be feasible by any current full network inference methods.

Table 1 Runtimes (averaged over ten runs) for the TINNiK algorithm in `MSCquartets`, on a 2020 Macbook 2 GHz Quad-Core i5, 32 GB RAM. Test levels are $\alpha = 0.001$ for the *T*3 test, and $\beta = 0.95$

Gene tree collection		Quartet tally (sec)	Hypothesis tests (sec)	Rest of TINNiK algorithm (sec)	Total (sec)
Simulation Analysis I:	10000 gts, $k = 1$	32.3	4.3	0.3	36.9
23 taxa	10000 gts, $k = 0.5$	31.1	4.6	0.3	35.9
	1000 gts, $k = 1$	3.3	5.6	0.3	9.2
	1000 gts, $k = 0.5$	3.0	6.6	0.3	9.9
Vanderpool [38]	1730 gts	16.5	11.5	0.4	28.4
29 primates					

Table 2 Blob detection using TINNiK for simulated data for $\mathcal{N}^+$ of Fig. 1 (L). Entries give ranges for α on which the full tree of blobs, and individual blobs, are correctly inferred with the *T*3 test. Interval endpoints are approximate, with dashes indicating the correct multifurcation is never inferred. For all analyses, $\beta = 1$

Number *n* of gene trees	Range of α values, for blob detection. ($\beta = 1$.)			
	Tree of Blobs			
	Fully correct	*B*-blob detected	*C*-blob detected	*D*-blob detected
	Low ILS: $k = 2$.			
10000	$[10^{-170}, 0.01]$	$[10^{-170}, 0.01]$	$[10^{-300}, 0.01]$	$[10^{-170}, 0.01]$
1000	$[10^{-15}, 10^{-4}]$	$[10^{-16}, 10^{-4}]$	$[10^{-49}, 0.01]$	$[10^{-15}, 10^{-4},]$
500	$[10^{-6}, 0.001]$	$[10^{-7}, 0.001]$	$[10^{-25}, 0.01]$	$[10^{-6}, 0.001]$
300	–	$[10^{-6}, 10^{-5}]$	$[10^{-17}, 0.01]$	–
300	–	–	–	–
	Moderate ILS: $k = 1$.			
10000	$[10^{-55}, 0.001]$	$[10^{-55}, 0.001]$	$[10^{-216}, 0.001]$	$[10^{-55}, 0.001]$
1000	$[10^{-7}, 0.001]$	$[10^{-7}, 0.001]$	$[10^{-23}, 0.01]$	$[10^{-7}, 0.001]$
500	$[10^{-4}, 0.005]$	$[10^{-4}, 0.008]$	$[1 \times 10^{-13}, 0.01]$	$[10^{-4}, 0.005]$
300	–	–	$[10^{-7}, 0.003]$	–
300	–	–	–	–
	High ILS: $k = 0.5$.			
10000	$[10^{-12}, 0.001]$	$[10^{-12}, 0.001]$	$[10^{-88}, 0.001]$	$[10^{-17}, 0.001]$
1000	–	–	$[10^{-8}, 0.001]$	–
500	–	–	$[10^{-4}, 0.001]$	–
300	–	–	–	–

Analysis I: Varying α

Our first analysis with TINNiK used a range of α values for the $T3$ test to detect quartet hybridization, but set $\beta = 1$ which, in effect, treats all quartets as resolved. Approximate ranges of α for which the full tree of blobs and individual blobs are detected are shown in Table 2.

With sample sizes $n \geq 500$ gene trees and ILS low or moderate, the tree of blobs is correctly inferred for a wide range of α. With 1000 gene trees sampled under low ILS condition, for example, the tree of blobs is correctly inferred for α ranging over eleven orders of magnitude. Even with high ILS, TINNiK returns the true tree of blobs from sample size 10,000.

A typical pattern of increasing resolution in the TINNiK tree of blobs as test level α is varied is shown in Fig. 3, for $n = 1000$ and $k = 1$. Smaller α sets a stricter criterion for a quartet to be judged non-tree-like, so the count of quartets initially flagged as B-quartets in the algorithm is decreased, and the number of B-quartets inferred using the inference rule of Theorem 2 may shrink as well. Proceeding from large α to small,

a) TINNiK first detects the tree-like A-group,
b) the C-blob is detected, and a cut edge separating the $\{C$-blob, A-group$\}$ from the B- and D-groups then appears,
c) the D-blob and then the B-blob are detected, so that the full tree of blobs is inferred for a range of $\alpha \in [10^{-7}, 0.001]$,
d) the B- and D-blobs become increasingly over-resolved, although the A-group and C-blob are correctly inferred even for very small values of α.

Several additional patterns from Table 2 and Figure 3 hold in a wide range of our experiments. First, detecting features of the tree of blobs by TINNiK is harder for some parts than others. For instance, decreasing α, the C-blob and the A-tree group are the first parts to be correctly detected by TINNiK, and remain correctly resolved for a large range of test levels. This suggests that the metric structure and topological complexity of a network may result in varying difficulty in correctly inferring specific parts of the tree of blobs. A single analysis may be insufficient to explore all hybridization in a large network.

Second, when ILS is present in anything other than low amounts, a gene tree sample of size 300 drawn from $\mathcal{N}^+$ appears too small to correctly infer the tree of blobs by TINNiK. Empiricists should be aware that the number of genes needed for accurate hybridization detection may be large. Whether these observations apply more generally to other data types and inference frameworks is unknown, as other tractable inference methods for non-level-1 networks are not yet available.

Analysis II: Varying β

Short branches on a network result in higher levels of ILS, which can cause *CF*s to be closer to (1/3, 1/3, 1/3). To study these effects on TINNiK's inference, a second analysis focused on the network $\mathcal{N}^+$ of Fig. 1 (L) with $k = 0.5$. We fixed $\alpha = 10^{-4}$ and varied the level β for the star tree test. Under this test, as β is decreased more quartets are taken to be star trees initially (and flagged as B-quartets) leading to more polytomies and less resolution in the TINNiK tree of blobs.

Figure 4 shows results for a sample size of $n = 1000$ gene trees. Proceeding from left to right, we see that for many values, $\beta > 10^{-6}$, the A-group and C-blob are correctly detected. As β is decreased, the A-, C-, and D-groups are correctly inferred, then the correct tree of blobs is found for $\beta \in [10^{-34}, 10^{-9}]$. Decreasing β further results in the D-blob collapsing incorrectly (for instance, in the bottom, right tree of Fig. 4, $\{D2, D3\}$ no longer form a cherry), and ultimately a star tree is produced.

Figure 5 shows typical simplex plots displaying the results of hypothesis tests (L) and application of the inference rule (R), for TINNiK test levels producing the true tree of blobs. The TINNiK algorithm first finds B-quartets corresponding to the red, green, and gold symbols displayed on the left. The increase in the number of B-quartets from the inference rule is visible in the gold symbols on the right.

That TINNiK can correctly infer the true tree of blobs when some B-quartets are found from the star tree test should be contrasted with results shown in Table 2 for this simulated data, where without B-quartets from the star tree test the tree of blobs was never correctly inferred. Using the star tree test to judge more quartets as unresolved (decreasing β) can thus help in obtaining the correct tree of blobs. High amounts of ILS from short branches can have the same qualitative impact on *CF*s as some blob structures, tending to equalize the entries, so that they are close to (1/3, 1/3, 1/3). The star tree test, by flagging such quartets as B-quartets regardless of the cause, helps prevent spurious resolution not strongly supported by the data.

Analysis III: Varying blob complexity

To investigate the effect that blob complexity might have on TINNiK's inference we considered a level-1 network $\mathcal{N}_1 = \mathcal{N}_1^+$ with a single 7-cycle, and then modified it by adding two additional hybridizations resulting in a level-3 network $\mathcal{N}_3 = \mathcal{N}_3^+$. Figure 6 (L) shows $\mathcal{N}_3$, with $\mathcal{N}_1$ composed of only the black and magenta edges. A simulated sample of $n = 1000$ gene trees was analyzed with TINNiK.

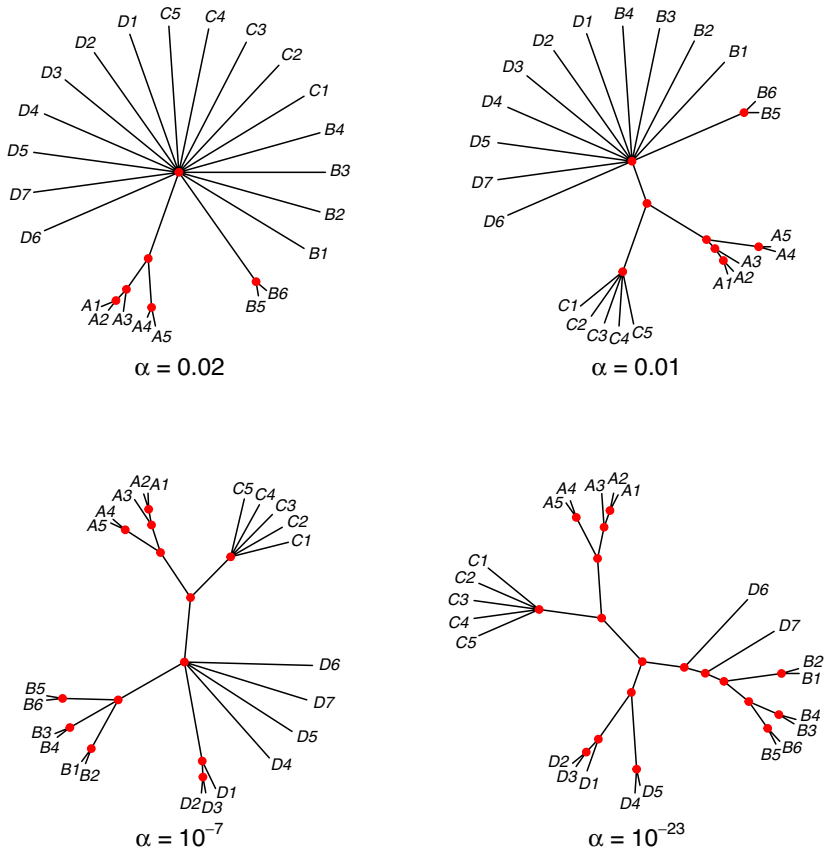

Fig. 3 Four TINNiK trees of blobs for a simulated sample of $n = 1000$ gene trees on network $\mathcal{N}^+$ of Fig. 1 (L) with $k = 1$ (moderate ILS), for $\beta = 1$ and $\alpha = 0.02, 0.01, 10^{-7}, 10^{-23}$. Increasing resolution as α is decreased is typical. The true tree of blobs (bottom, left) is inferred for a large range of test levels $\alpha \in [10^{-7}, 0.001]$. Although the (bottom, right) tree is over-resolved when $\alpha = 10^{-23}$, each split is compatible with a tree displayed on $\mathcal{N}^+$

For all $\binom{10}{4}$ 4-taxon sets the expected quartet concordance factors were computed with `QuartetNetworkGoodnessFit` [39] and plotted in Fig. 6 (R, top) ($\mathcal{N}_1$ magenta, $\mathcal{N}_3$ blue). Expected *CF*s not on the model lines in Fig. 2 correspond to B-quartets, while some of those on the model lines may be inferred as B-quartets using the inference rule. The effect of increasing topological complexity in this 7-blob is to "pull" many *CF*s closer to the centroid.

The pull of *CF*s toward the centroid with increasing topological complexity means that the signal for hybridization increasingly resembles that for lack of quartet resolution. Intuition for this is that each particular choice of lineage paths through a blob determines a *CF* in the simplex, with a convex sum of these giving the expected *CF*. But a convex sum of a collection of *CF*s will be their weighted center of gravity, and hence tend toward their "middle."

Since *CF*s computed from inferred gene trees in simulation studies have also been observed to be pulled toward the centroid from their expectation [27], the blurring of hybridization signal and lack of resolution may be very difficult to untangle. This suggests there may be practical limits on how complicated blob structure can be for reliable inference from *CF*s.

Table 3 shows a range of values for which the tree of blobs is correctly inferred when only one of the two test levels is varied. When $\beta = 1$, TINNiK infers the true tree of blobs for a much wider range of test levels α for $\mathcal{N}_1$ than $\mathcal{N}_3$. This is not surprising, since more of $\mathcal{N}_1$'s *CF*s are placed distant from the model lines than those for $\mathcal{N}_3$. Similarly, for fixed $\alpha = 10^{-25}$, TINNiK infers the true tree of blobs for a much wider range of β levels for the level-3 network $\mathcal{N}_3$.

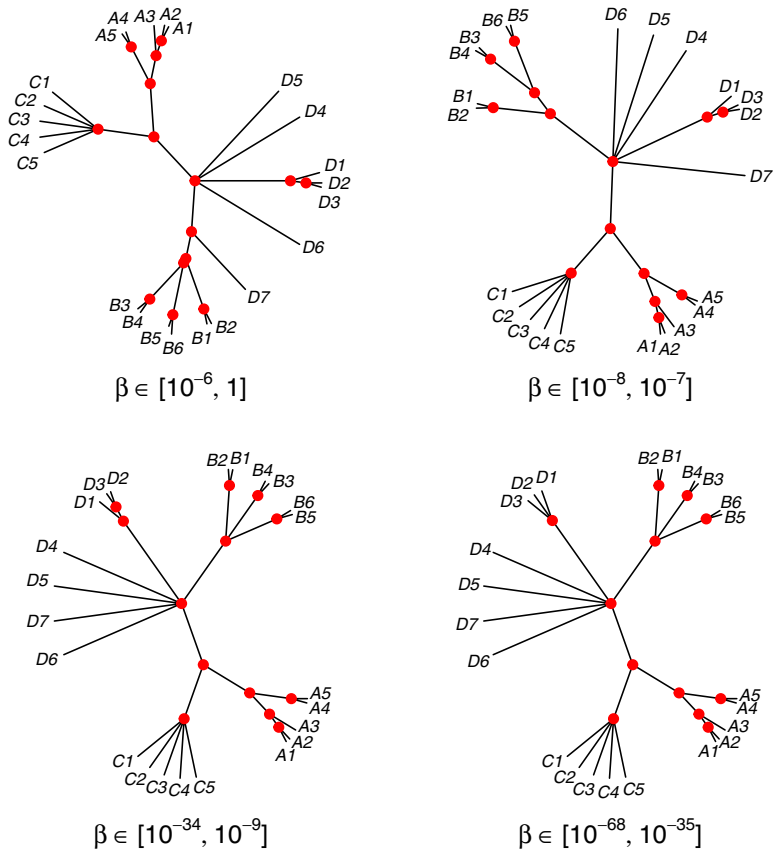

Fig. 4 The TINNiK tree of blobs for fixed $\alpha = 10^{-4}$ and various β for a simulated sample of $n = 1000$ gene trees for $\mathcal{N}^+$ with $k = 0.5$ (high ILS). Decreasing β results in less resolution in the tree of blobs. From left to right: (top) only the *A*- and *C*- groups are correct; all blobs except the *B*-blob are correct; (bottom) the TINNiK tree of blobs is correct for any $\beta \in [10^{-34}, 10^{-9}]$; the *D*-group lacks sufficient resolution. For even smaller β the TINNiK tree of blobs degrades to a star tree

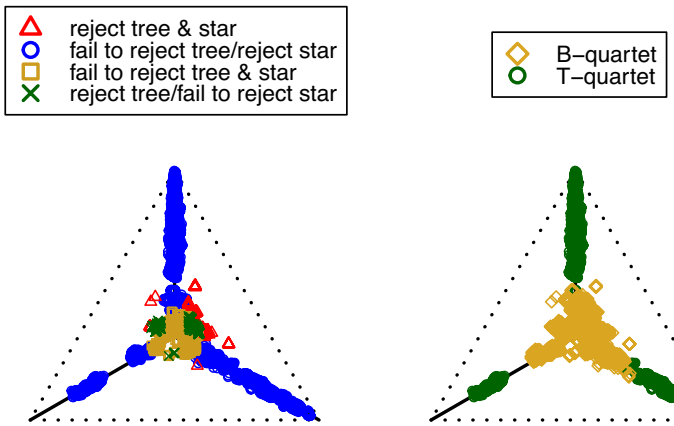

Fig. 5 Simplex plots showing the results of hypothesis tests for $\alpha = 10^{-4}, \beta = 10^{-10}$ (L) and after the application of the Inference Rule (R). B-quartet simplex plots for any $\beta \in [10^{-31}, 10^{-9}]$ are identical, although the hypothesis test results differ

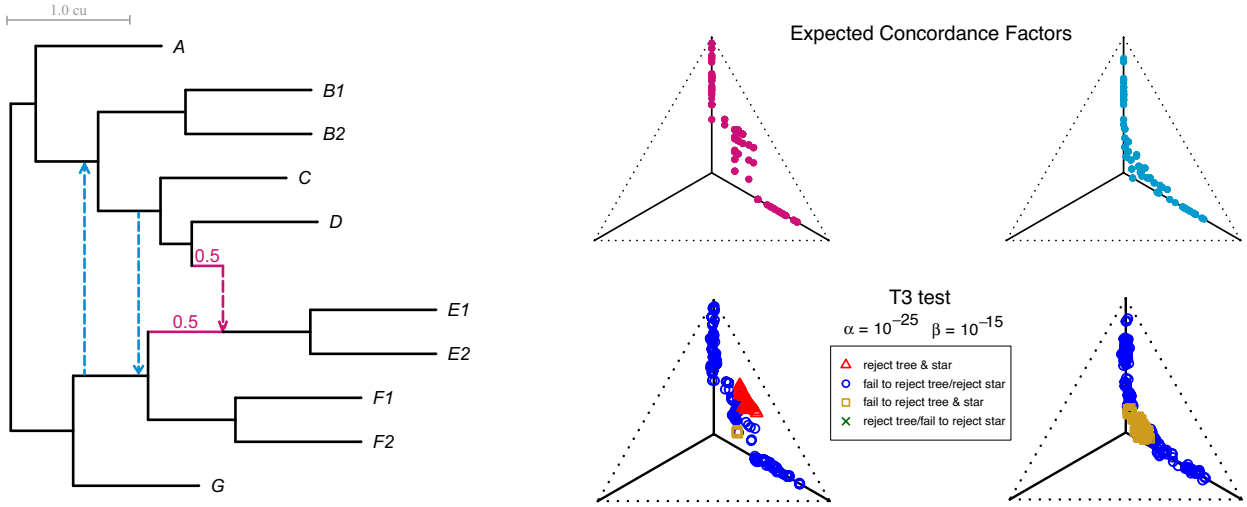

Fig. 6 (L) Model networks in Analysis III. $\mathcal{N}_1$ consists of black and magenta edges, and $\mathcal{N}_3$ all edges. All hybridization parameters are $\gamma = 0.5$, with Newick notation given in Appendix A. (Right, top) Expected *CF*s for $\mathcal{N}_1$ (magenta) and $\mathcal{N}_3$ (light blue). (Right, bottom) Typical simplex plots for $\mathcal{N}_1$ (left) and $\mathcal{N}_3$ (right) displaying hypothesis test results for $\alpha = 10^{-25}, \beta = 10^{-15}$. For these levels, TINNiK infers the true tree of blobs of both $\mathcal{N}_1$ and $\mathcal{N}_3$, but with different initial lists of B-quartets. For $\mathcal{N}_1$ most initial B-quartets are detected from non tree-like signal, but for $\mathcal{N}_3$ from star-like

Table 3 Range of α, β values, for correct blob detection in 7-blob networks for a sample of 1000 gene trees

	***T3* test:** α **interval,** $\beta = 1$	**Star tree test:** $\alpha = 10^{-25}$**,** β **interval**
$\mathcal{N}_1$	$[2 \times 10^{-10}, 0.03]$	$[2 \times 10^{-84}, 2 \times 10^{-13}]$
$\mathcal{N}_3$	$[0.002, 0.01]$	$[6 \times 10^{-109}, 1 \times 10^{-7}]$

Analysis IV: comparison to network inference

Some recent network inference methods seek to infer a level-1 network, yet offer no means of testing that assumption. One way that TINNiK might be helpful for this is by comparing its tree of blobs to an inferred level-1 structure. To test this possibility, we considered the level-2 network $\mathcal{N}^+$ of Fig. 7 (L), whose tree of blobs is a star tree. We analyzed simulated data of $n = 10,000$ gene trees from this network using SNaQ [37], which assumes the network is level-1. We thus knowingly violated SNaQ's assumptions, and did not expect its output to necessarily resemble the true network.

SNaQ's optimal level-1 networks with 1 and 2 hybridizations are shown in Fig. 7 (C,R). Note that the network $\widehat{\mathcal{N}_1}$ returned by SNaQ when $h_{max} = 1$ can not be obtained from $\mathcal{N}^+$ by removing a single hybrid edge, nor is its tree of blobs $\mathcal{T}(\widehat{\mathcal{N}_1})$ a star tree. Much of the inferred metric information also has little relationship to the true network's branch lengths. When $h_{max} = 2$, the inferred SNaQ network $\widehat{\mathcal{N}_2}$ has two cycles joined with a branch of length zero. While the tree of blobs for $\widehat{\mathcal{N}_2}$ would be a star tree if the zero branch length were collapsed, the inferred blob structure is misleading. For instance, the close hybrid relationship between *D* and *E* is inferred as a more distant non-hybrid one.

The tree of blobs inferred by TINNiK is a (correct) star tree for any $\alpha > 10^{-199}$ and $\beta > 10^{-109}$. For no values of α does TINNiK obtain a tree of blobs reflecting any of the individual cycles that SNaQ infers. Since both SNaQ and TINNiK base their inference on the same quartet *CF*s, the conflict is even more striking.

Empirical data

We apply TINNiK to infer trees of blobs from several empirical datasets: Hawaiian flowering plants [40] and primates [38]. These have been analyzed for hybridization previously, with conflicting results depending on the method used.

Hawaiian Cyrtandra

A recent study by Kleinskopf et. al. [40] investigated hybridization and introgression in the Hawaiian *Cyrtandra*. Although samples were collected across the islands, network analyses by PhyloNet and SNaQ were restricted to single island subsamples. The dataset consists of 569 gene trees, a few with missing taxa. Most of the gene trees are poorly resolved, with a majority of gene

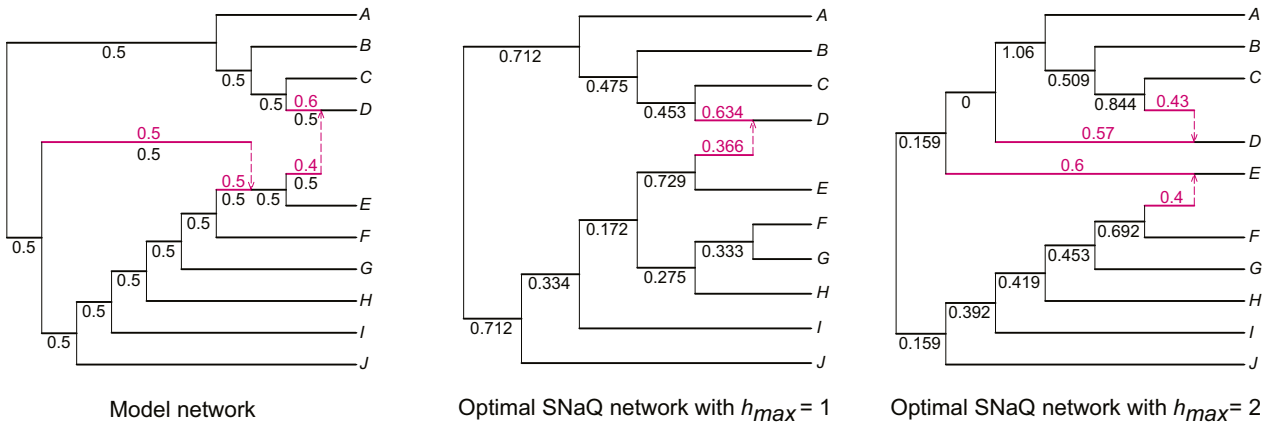

Fig. 7 (L) A level-2 model network with star tree of blobs. Hybrid edges and hybridization parameters are in magenta, with Newick notation given in Appendix A. (C) The optimal inferred network from SNaQ with the maximal number of hybridizations constrained to $h_{max} = 1$, and (R) with $h_{max} = 2$. In both (C,R), terminal and hybrid branch lengths are absent since they are not identifiable under the NMSC from *CF*s when only a single lineage is sampled from the descendant population, and therefore are not inferred by SNaQ

quartet trees unresolved for almost all sets of 4 taxa. For the Kauai island group of 7 taxa, for example, the star tree topology is a majority for all but one gene quartet (34 of 35).

Networks inferred by PhyloNet and SNaQ (with $h_{max} = 1$) for the Kauai group agreed [40, Fig. 4]. Since the lack of gene tree resolution indicated information content might be low, TINNiK analyses were performed both with unresolved gene quartets omitted, and with unresolved quartets apportioned uniformly among the three resolved topologies. TINNiK's analyses support the PhyloNet/SNaQ underlying tree of blobs (Fig. 8 (L)) over a wide range of test levels with both methods for handling unresolved gene quartets. Specifically, when unresolved quartets are included in the analysis, the supporting TINNiK tree of blobs shown in Fig. 8 (L) is obtained for any $\alpha \in [0.03, 0.14]$ with $\beta = 1$, and for $\alpha = 0.05$ with $\beta \in [0.3, 1]$.

In contrast, there is considerable discrepancy between the PhyloNet and SNaQ analyses for the 8-species Oahu group [40]. PhyloNet infers a tree, while SNaQ infers a level-1 network with 2 cycles. We found that TINNiK inferred exactly four topologies as α, β, and the treatment of polytomies were varied: (1) a binary tree agreeing with that inferred by PhyloNet, (2) a tree of blobs $\mathcal{T}$ pictured in Fig. 8 (C) with exactly two cut edges, (3) a tree with three cut edges differing from (C) by moving the attachment for *C. calpidicarpa* from the multifurcation, and 4) a star tree. When α is small and β large, so that there are no initial B-quartets, the inferred TINNiK tree agrees with that of PhyloNet (and `MSCquartets`' QDC tree [6]). This supports PhyloNet's analysis in that signal for

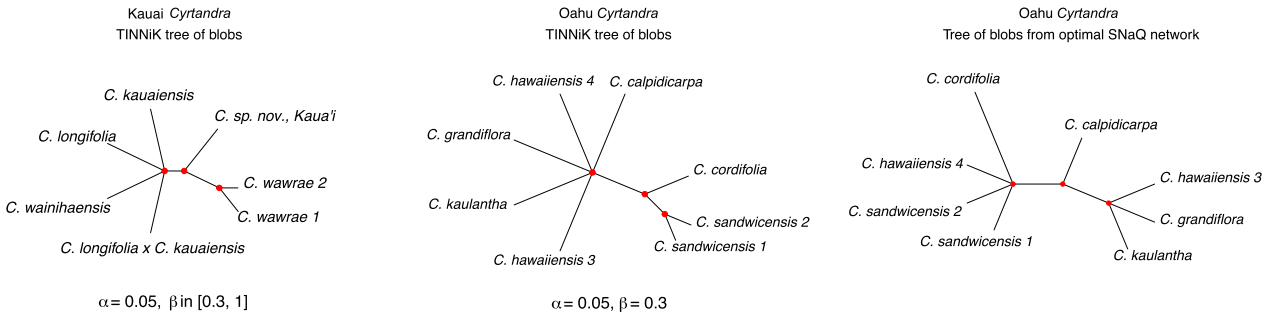

Fig. 8 (L) For the Kauai dataset, TINNiK's tree of blobs supports the PhyloNet and SNaQ analyses for a wide range of test levels α, β, regardless of handling of unresolved quartet topologies. (C) The TINNiK tree of blobs for the Oahu dataset when $\alpha = 0.05$ and $\beta = 0.3$ with either treatment of polytomies conflicts with both SNaQ and PhyloNet analyses. It supports some blob structure, but not that inferred by SNaQ. (R) The tree obtained by contracting cycles in the SNaQ network inferred from the Oahu dataset. PhyloNet infers a resolved tree for these data

hybridization in the Oahu data may be weak. Moreover, the inferred TINNiK tree $\mathcal{T}$ of Fig. 8 (C) does not agree with the tree of Fig. 8 (R) obtained by contracting cycles in SNaQ's optimal network, nor does its variant with *C. calpidicarpa* moved. Possible reasons for this conflict might again be signal too weak for these analyses, or that the underlying network is not level-1. Regardless of the cause, TINNiK illuminates that further investigation is needed to understand relationships in this group.

Primate data

A recent study of primates by Vanderpool et al. [38] used full genome data to investigate phylogenetic relationships between 26 primates. Multiple analyses were performed, but we focus on two investigations, into resolution of a clade of New World Monkeys (NWMs), and of possible introgression within a subset of 7 taxa, the Papionini group. These data were also studied in [18] using PhyNEST. Input for our analyses were the 1730 gene trees estimated in [38].

The placement on the primate tree of some NWMs is uncertain, with one analysis supporting that *A. nancymaae* and *C. jacchus* form a clade sister to the {*S. boliviensis, C. Capucinis imitator*} clade, and a second that *A. nancymaae* is sister to the {*S. boliviensis, C. Capucinis imitator*} clade with *C. jacchus* an outgroup [38]. Using `MSCquartets` to compute empirical *CF*s and to perform hypothesis tests, we found that quartet *CF*s that clustered near the centroid are exactly those that might resolve this issue. In Fig. 9 (L) for any $\beta < 0.1$ (shown with $\alpha = 10^{-7}$), the golden squares clustered around the centroid where the star tree hypothesis is not rejected for any alternative resolved topology, are those involving {*C. jacchus*, *A. nancymaae*}, exactly one of {*S. boliviensis*, *C. capucinus imitator*} and a fourth taxon. As seen in Fig. 9 (R), the TINNiK tree of blobs has a degree 4 node for any $\beta < 0.1$, which does not support further resolution. Note that our choice of α ensured no putative 4-blob quartets, so this multifurcation arose solely due to support for star-like quartets.

A subset of four Asian Papionini (*Cercocebus atys*, *Mandrillus leucophaeus*, *Papio anubis*, *Theropithecus gelada*) and three African Papionini (*Macaca asciculari*s, *Macaca mulatta*, *Macaca nemestrina*) were also analyzed by Vanderpool et. al., with multiple introgression events found between and among these groups [38, Fig. 4] using the Δ method of [41]. Specifically, seven introgression events were inferred, with four crossing continental boundaries.

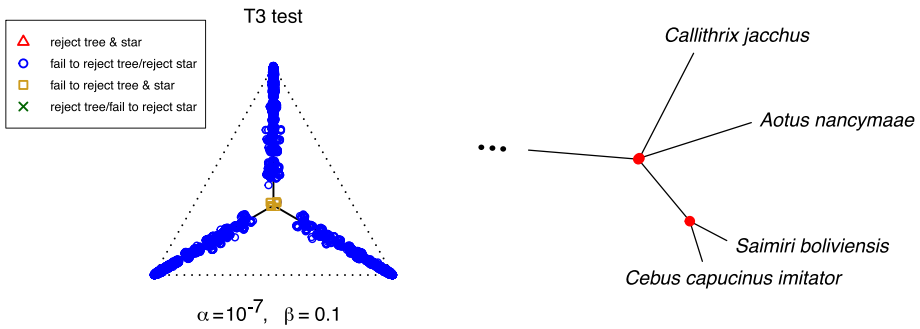

Fig. 9 (L) Simplex plots illustrate that hypothesis test results support the star tree topology for quartets with *A. nancymaae*, *C. jacchus* and one of the other two NWMs, and (R) close up of tree of blobs for the NWMs for any $\beta \leq 0.1$

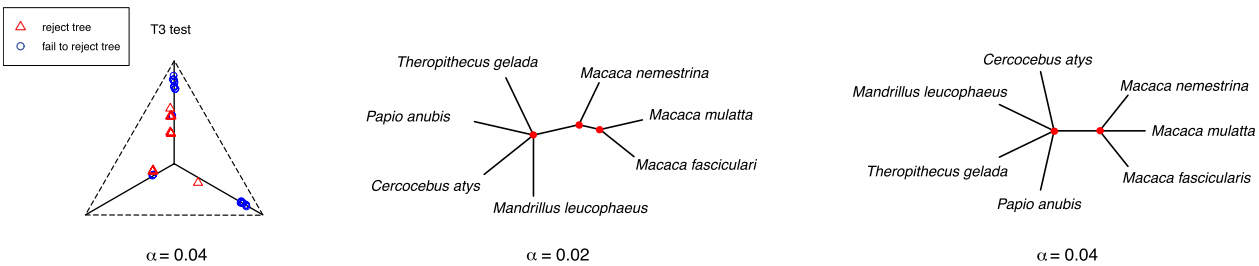

Fig. 10 (L) Results of *T*3 hypothesis tests for $\alpha = 0.04$, $\beta = 1$; (C) for $\alpha \in [0.015, 0.032]$, TINNiK's tree of blobs supports hybridization among the African Papionini; and (R) for $\alpha \in [0.033, 0.093]$ TINNiK supports hybridization within the African Papionini and within the Asian macaques. Only for larger values of α Does TINNiK return a star tree of blobs

The sequence data for these taxa were reanalyzed by Kong et. al. using PhyNEST to infer level-1 networks with $h_{max} = 1, 2$ hybridizations [18]. *Theropithecus gelada* was found to be a hybrid of *Papio anubis* and *Mandrillus leucophaeus* when $h_{max} = 1$, with an additional hybridization among the Macaques when $h_{max} = 2$. These level-1 hybridization cycles do not cross continental boundaries.

A TINNiK analysis of these seven taxa was performed with $\beta = 1$, since the simplex plot of Fig. 10 (L) shows no *CF*s close to that of the star tree. In Fig. 10 (C) the TINNiK tree of blobs for test levels $\alpha \in [0.008, 0.032]$ agrees with the PhyNEST analysis when $h_{max} = 1$. For larger $\alpha \in [0.033, 0.093]$ the tree of blobs is shown (R), with multifurcations for each continent, consistent with the PhyNEST analysis when $h_{max} = 2$. For test levels $\alpha > 0.093$, TINNiK returns the star tree, consistent with the analysis using Δ. However, such a large value of alpha indicates weak support for additional hybridization spanning continents.

Conclusions

The implementation of the TINNiK algorithm in `MSC-quartets` provides the first software tool for statistically-justified inference of the tree-like parts of a species network. With input of gene trees inferred from multilocus sequence data, it quickly returns an inferred tree of blobs of the network under the NMSC, without restrictive assumptions on the reticulation structure within the blobs. Because TINNiK is a quartet-based method, employing quartet-based hypothesis tests and quartet-based combinatorial rules followed by a fast distance method for constructing the tree of blobs, it is designed to scale to large numbers of taxa.

In some cases, the tree of blobs, perhaps with partial information on individual blob structure, may represent the most we can tell about a species network from biological data. While the theoretical limits to inference of complex blob structure are still unknown, recent work [19] has shown that different blob structures are indistinguishable from certain types of commonly-used gene tree summary data. Even in cases where theoretical identifiability holds, practical identifiability may not, as the signal distinguishing the precise structure may be obscured by even small levels of noise. Learning a blob is present, or only part of its structure, may be the strongest practical inference that can be performed for some data.

When more can be inferred, the tree of blobs for a large group of species can provide a good starting point for a more targeted investigation into the unknown relationships represented by its multifurcations. Its inference might be either a first step in an exploratory data analysis, or form the basis for a divide-and-conquer approach, although the more demanding statistical inference of internal blob structure requires further theoretical and practical development.

Several recently-proposed network inference methods assume a level-1 structure (SNaQ, PhyNEST, NANUQ), but their performance under level misspecification has not been studied in published work. (Note, however, that NANUQ's splits graph can suggest such model violation.) Since TINNiK does not assume a particular level or other special blob structure, it can provide an important alternative perspective. Inference of a full network that is incompatible with TINNiK's tree of blobs can suggest possible model violations and a need for further analysis. If gene trees have already been inferred, TINNiK's speed means its use in this way requires little additional computational effort.

An inferred tree of blobs may also be useful for the heuristic searches performed by methods attempting to find complete networks. When a starting network is needed, a TINNiK tree of blobs is a natural candidate so the search may spend less time finding the tree-like parts of the network. Even if TINNiK produces an over-resolved tree, in our experiments this is often a tree displayed on the network, so that as new hybrid edges are introduced the search may still soon focus on good candidate networks. Finally, for those methods requiring an *a priori* upper bound on the number of reticulations, TINNiK can again be helpful by suggesting the number of blobs, and thus a minimum number of reticulations needed.

Although our justification of the TINNiK algorithm in this work has emphasized the NMSC model, its essential ideas could be applied to other models of gene tree formation. For instance, recent work [19] considered two other models, one in which gene trees must be displayed on the species network so coalescence is immediate, and a common-inheritance coalescent model in which the standard coalescent applies but only inside displayed trees. In both these cases it is possible to identify B-quartets for 4-taxon networks from certain data types, and thus follow the outline of our algorithm.

The introduction of TINNiK for inferring the tree of blobs of a species network from biological data should encourage the development of other algorithms for this problem. Network inference remains difficult for both theoretical and practical reasons, and phylogenomics will benefit from an expanding array of approaches.

Appendix A

Newick for model networks

The Newick string for the model network shown in Fig. 1 (L) used in simulations in Sects. Analysis I: Varying and Analysis II: Varying is:

(((((A1:0.4,A2:0.4)a1:1.0,A3:0.2)a2:1.5,(A4:0.5,A5:0.5)a3:1.2)aa:1.0, ((((C1:0.5,#H5:0.5::0.3)c1:0.6,#H6:0.5::0.4)c2:0.4,#H7:0.4::0.25)c3:0.3, ((C2:0.4)#H7:0.55::0.75,((C3:0.6)#H6:0.75::0.6,((C4:0.2)#H5:0.6::0.7, C5:0.3)c4:0.5)c5:0.4)c6:0.4)cc:1.0)ac:1.0,(((D1:0.6,(D2:0.3,D3:0.3)d10:0.6) d11:1.2,#H4:0.5::0.8)d1:0.4,(((D4:0.6,(D5:0.25)#H2:0.2::0.6)d9:0.2) #H4:0.1::0.2,#H3:0.1::0.4)d3:0.8,((((#H2:0.2::0.4,D6:0.65)d2:0.2)#H3:0.1::0.6,(((B1:0.3,B2:0.2)b1:1.0,#H1:0.4::0.3)b2:0.5,(((B3:0.4,B4:0.4) b5:1.0)#H1:0.4::0.7,(B5:0.4,B6:0.5)b3:1.0)b4:0.6)bb:1.2)d5:0.2,D7:0.3) d6:0.6)jj:0.5)r;

The Newick string for the model network shown in Fig. 6 (L) used in of Sect. Analysis III: Varying blob complexity is:

((G:1.0,((((F2:1.0,F1:1.0):0.7,((E2:1.0,E1:1.0):0.7)#H1:0.6::0.5):0.1)#H2:0.4::0.5, #H3:0.0::0.5):0.1):0.5,((((#H2:0.0::0.5,((#H1:0.25::0.5,D:1.0):0.25,C:1.0):0.2):0.3, (B2:1.0,B1:1.0):0.7):0.1)#H3:0.4::0.5,A:1.0):0.2)r;

The Newick string for the model network shown in Fig. 7 (L) used in of Sect. Analysis IV: comparison to network inference is:

((((I:1.0,(H:1.0,(G:1.0,(F:1.0,((E:1.0,#H2:0.5::0.4):0.5)#H1:0.5::0.5):0.5):0.5):0.5):0.5):0.5,J:1.0):0.5,#H1:0.5::0.5):0.5,(A:1.0,(B:1.0,(C:1.0,(D:1.0)#H2:0.5::0.6):0.5):0.5):0.5);

Appendix B

B- and T-quartets on a sunlet network

See Fig. 11

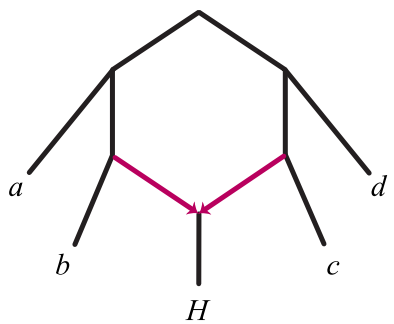

Fig. 11 The quartet $\mathcal{Q} = \{a, b, c, d, \}$ is a B-quartet on $\mathcal{N}^+$, but a *T*-quartet on the induced network $\mathcal{N}^+_{\mathcal{Q}}$

Appendix C

Cut hypothesis test and simulations

We use the notation of Sect. Cut model testing and maximum likelihood inference.

C.1 Cut model topology estimation and LR statistic

For the cut model, the maximum likelihood parameter estimate for data

$$qcCF_{abcd} = (m_{ab|cd}, m_{ac|bd}, m_{ad|bc})$$

with $m = m_{ab|cd} + m_{ac|bd} + m_{ad|bc}$ is found by computing the maximum of the trinomial likelihood constrained to each of the 3 line segments of Fig. 2 (L) and then choosing the largest. (Ties broken at random.) For the vertical line the maximizer is

$$\left(\frac{m_{ab|cd}}{m}, \frac{m_{ac|bd} + m_{ad|bc}}{2m}, \frac{m_{ac|bd} + m_{ad|bc}}{2m}\right) = \left(\frac{m_{ab|cd}}{m}, \frac{m - m_{ab|cd}}{2m}, \frac{m - m_{ab|cd}}{2m}\right),$$

the projection of $\widehat{CF}$ (the normalized *qcCF*) orthogonally to the line. A comparison of the likelihood at the maximizers on the three lines leads to the three regions shown in Fig. 12 (L) for which normalized *qcCF*s lead to cut model MLEs on the model lines in each region. For use in TINNiK, when the cut model is not rejected, we need only the topology of the MLE, which is determined solely by the color of the region in which $\widehat{CF}$ lies.

The likelihood ratio statistic for the hypothesis test described in Sect. Cut model testing and maximum likelihood inference requires the maximum log-likelihoods under the cut (null) and unconstrained (alternative) trinomial models. For the vertical line of the cut model, the maximum log-likelihood is

$$m_{ab|cd} \log m_{ab|cd} + (m - m_{ab|cd})(\log(m - m_{ab|cd}) - \log 2) - m \log m + C,$$

while for the unconstrained model, with MLE $\widehat{CF}$, the maximum log-likelihood is

$$m_{ab|cd} \log m_{ab|cd} + m_{ac|bd} \log m_{ac|bd} + m_{ad|bc} \log m_{ad|bc} - m \log m + C,$$

with C a constant.

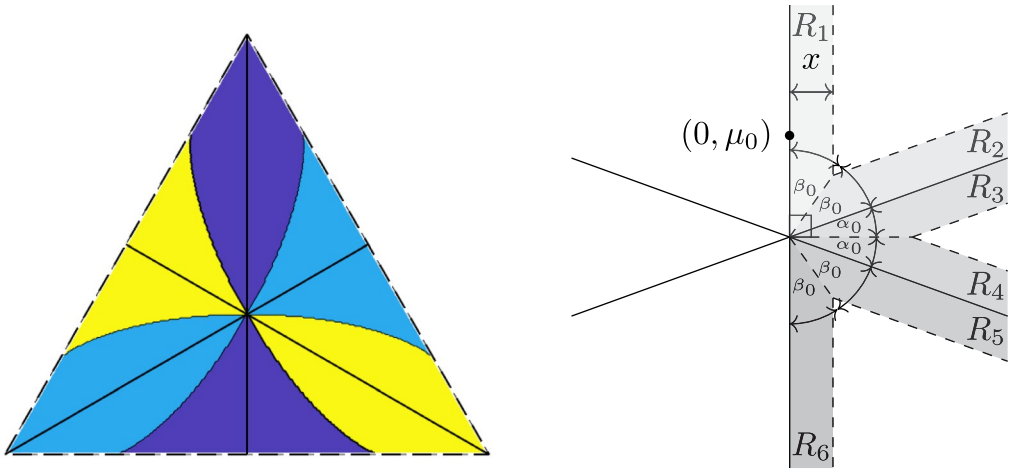

Fig. 12 (L) Regions for which data with an empirical $\widehat{CF}$ gives an MLE in the cut model on each of the three cut model lines. The MLE is obtained by moving orthogonally to the model line in the same colored region as $\widehat{CF}$. (R) The image of the cut model in Δ_2 under a linear transformation to the plane, with $\left(1-\frac{2}{3}\phi_0, \frac{1}{3}\phi_0, \frac{1}{3}\phi_0\right)$ mapped to $(0, \mu_0)$, and the 3 model lines mapped to the three lines shown. The region of integration for $G(x)$ in the proof of Proposition 6 is all shaded regions R_i and their reflections about the vertical line

C.2 Cut test distributions

To derive an asymptotic distribution for the likelihood ratio (LR) statistic for the cut model hypothesis test of Sect. Cut model testing and maximum likelihood inference, we follow similar derivations for $T1$ and $T3$ tests using Theorem 3.1 of [28]. That work also provides more discussion of why model singularities, such as the cut model's (1/3, 1/3, 1/3), make the use of a standard distribution inappropriate.

Assume the generating parameter in the cut model is $\theta_0 = (1-2\phi_0/3, \phi_0/3, \phi_0/3)$. Applying a linear transformation dependent on the sample size n and the Fisher information matrix $\mathcal{I}$ as in [28], the simplex is mapped to $\mathbb{R}^2$, and the cut model to lines crossing at the origin, with the vertical line segment mapped to the y-axis. Then $\theta_0 \mapsto (0, \mu_0)$, with $\mu_0 = \mu_0(n) := \sqrt{2n}(1-\phi_0)(\phi_0(3-2\phi_0))^{-1/2}$, and the information matrix becomes the identity. This transformation is not conformal (unless $\phi_0 = 1$) and the two transformed model lines not containing the generating parameter form angles $\alpha_0 = \arctan\left((3(3-2\phi_0))^{-1/2}\right)$ with the horizontal. See Fig. 12 (R).

Proposition 5 *The likelihood ratio statistic for testing H_0 versus H_1 for a true parameter point $\theta_0 = (1-2\phi_0/3, \phi_0/3, \phi_0/3)$, $\phi_0 \in (0, 3/2)$, of the cut model, with sample size n is asymptotically distributed as the random variable*

$$\Lambda_n = \min\left(Z^2, \left(\sin\alpha_0 Z + \cos\alpha_0 \bar{Z}\right)^2, \left(\sin\alpha_0 Z - \cos\alpha_0 \bar{Z}\right)^2\right),$$

where $Z \sim \mathcal{N}(0,1)$, $\bar{Z} \sim \mathcal{N}(\mu_0, 1)$, $\mu_0(n) := \sqrt{2n}(1-\phi_0)(\phi_0(3-2\phi_0))^{-1/2}$, and $\alpha_0 = \arctan\left((3(3-2\phi_0))^{-1/2}\right)$.

Here "asymptotically distributed" means that the likelihood ratio statistic and this random variable converge in distribution to the same limit as $n \to \infty$.

Proof Letting $\gamma_0 = \tan\alpha_0$, in the transformed space the image of Θ_0 is contained in the union of the lines $x = 0$, $y = \gamma_0 x$ and $y = -\gamma_0 x$.

By Theorem 3.1 of [28], the approximate distribution of the likelihood ratio statistic is the distribution of the minimum squared Euclidean distance between a normal sample, $\mathcal{N}((0, \mu_0), I)$, and the three lines in the transformed space. Assuming that θ_0 is not too close to the boundary of the simplex, in a sense dependent on the sample size, little of the mass of $\mathcal{N}((0, \mu_0), I)$ is outside the image of the simplex. Thus, for the remainder of the argument, we replace these line segments with lines intersecting at the singularity $(0, 0)$.

Denote the marginal probability distributions of the bivariate normal sample by $Z \sim \mathcal{N}(0,1)$ and $\bar{Z} \sim \mathcal{N}(\mu_0, 1)$. We next determine the squared distance of a sample point $(Z, \bar{Z})$ to each of the three lines.

Considering the first line, $x = 0$, the squared Euclidean distance is Z^2. For the line $y = \gamma_0 x$, the closest point $(X, \gamma_0 X)$ to $(Z, \bar{Z})$ has $X = (Z + \gamma_0\bar{Z})/(1+\gamma_0^2)$, so the squared distance is

$$\frac{\gamma_0^2}{1+\gamma_0^2}\left(Z - \frac{1}{\gamma_0}\bar{Z}\right)^2 = \left(\sin\alpha_0 Z - \cos\alpha_0 \bar{Z}\right)^2.$$

Similarly, for the line $y = -\gamma_0 x$ the squared distance is $\left(\sin\alpha_0 Z + \cos\alpha_0 \bar{Z}\right)^2$. The claim follows by taking the minimum of these squared distances. □

For testing purposes, we characterize this distribution further.

Proposition 6 *The probability density function for the random variable $\widetilde{\Lambda}_n$ of Proposition 5 is, for $\lambda > 0$, is $f_{\widetilde{\Lambda}_n}(\lambda)$ with*

$$\begin{aligned} f_{\widetilde{\Lambda}_n}(\lambda) = &\frac{1}{2\sqrt{2\pi\lambda}}\Bigg[\exp\left(-\frac{1}{2}\lambda\right)\left(2-\operatorname{erf}\left(\frac{\sqrt{\lambda}\cot\beta_0+\mu_0}{\sqrt{2}}\right)-\operatorname{erf}\left(\frac{\sqrt{\lambda}\cot\beta_0-\mu_0}{\sqrt{2}}\right)\right)\\ &+\exp\left(-\frac{1}{2}\left(\sqrt{\lambda}-\mu_0\cos\alpha_0\right)^2\right)\left(2-\operatorname{erf}\left(\frac{\sqrt{\lambda}\cot\alpha_0+\mu_0\sin\alpha_0}{\sqrt{2}}\right)-\operatorname{erf}\left(\frac{\sqrt{\lambda}\cot\beta_0-\mu_0\sin\alpha_0}{\sqrt{2}}\right)\right)\\ &+\exp\left(-\frac{1}{2}\left(\sqrt{\lambda}+\mu_0\cos\alpha_0\right)^2\right)\left(2-\operatorname{erf}\left(\frac{\sqrt{\lambda}\cot\alpha_0-\mu_0\sin\alpha_0}{\sqrt{2}}\right)-\operatorname{erf}\left(\frac{\sqrt{\lambda}\cot\beta_0+\mu_0\sin\alpha_0}{\sqrt{2}}\right)\right)\Bigg], \end{aligned}$$

where $\alpha_0 = \arctan\left((3(3-2\phi_0))^{-1/2}\right)$ *and* $\beta_0 = \frac{1}{2}\left(\frac{\pi}{2}-\alpha_0\right)$.

Proof To determine the probability density function for the distribution of Proposition 5, let $G(x)$ denote the cumulative distribution function of the (non-squared) Euclidean distance. This is found by integrating the distribution $\mathcal{N}((0,\mu_0),I)$ over the tube of points within distance x from the image of Θ_0, with simplifications using the symmetry of the region and normal distribution as shown in Fig. 12 (R). Although the generating parameter $(0,\mu_0)$ is shown above the origin, for $\phi_0 \in \left(1,\frac{3}{2}\right)$ it may be below. In fact, α_0 is an increasing function of ϕ_0, with $\alpha_0(0) \approx 0.322$ and $\lim_{\phi_0 \to \frac{3}{2}^-} \alpha_0(\phi_0) = \frac{\pi}{2}$.

Then $G(x) = 2\sum_{i=1}^{6} G_i(x)$, where G_i is the integral over the shaded strip R_i, and the density of the Euclidean distance is $g(x) = 2\sum_{i=1}^{6} \frac{d}{dx}G_i(x)$.

Considering $\frac{d}{dx}G_1(x)$ first:

$$\begin{aligned} \frac{d}{dx}G_1(x) &= \int_{\alpha_0+\beta_0}^{\frac{\pi}{2}} \frac{d}{dx}\int_0^{\frac{x}{\cos\beta}} \frac{1}{2\pi}\exp\left(-\frac{1}{2}\left(r^2-2\mu_0 r\sin\beta+\mu_0^2\right)\right) r\,dr\,d\beta\\ &= \int_{\alpha_0+\beta_0}^{\frac{\pi}{2}} \frac{1}{2\pi}\exp\left(-\frac{1}{2}\left(\frac{x^2}{\cos^2\beta}-2\mu_0\frac{x}{\cos\beta}\sin\beta+\mu_0^2\right)\right)\frac{x}{\cos^2\beta}\,d\beta\\ &= \frac{1}{2\pi}\exp\left(-\frac{1}{2}x^2\right)\int_{\alpha_0+\beta_0}^{\frac{\pi}{2}} \exp\left(-\frac{1}{2}\left(x^2\tan^2\beta-2\mu_0 x\tan\beta+\mu_0^2\right)\right)\frac{x}{\cos^2\beta}\,d\beta. \end{aligned}$$

Substituting $y = x\tan\beta$ gives

$$\begin{aligned} \frac{d}{dx}G_1(x) &= \frac{1}{2\pi}\exp\left(-\frac{1}{2}x^2\right)\int_{x\tan(\alpha_0+\beta_0)}^{\infty}\exp\left(-\frac{1}{2}(y-\mu_0)^2\right)dy\\ &= \frac{1}{2\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)\left(1-\operatorname{erf}\left(\frac{1}{\sqrt{2}}(x\tan(\alpha_0+\beta_0)-\mu_0)\right)\right). \end{aligned}$$

Next we consider $\frac{d}{dx}G_2(x)$ and $\frac{d}{dx}G_3(x)$. Rotating the figure $2\beta_0$ counter-clockwise to simplify the computation, the generating parameter is mapped as $(0,\mu_0) \mapsto (-\mu_0\cos\alpha_0, \mu_0\sin\alpha_0)$. Then

$$\begin{aligned}\frac{d}{dx}G_2(x) &= \int_{\frac{\pi}{2}}^{\frac{\pi}{2}+\beta_0} \frac{d}{dx}\int_0^{-\frac{x}{\cos\beta}} \frac{1}{2\pi}\exp\left(-\frac{1}{2}\Big(r^2+2\mu_0 r\cos(\alpha_0+\beta)+\mu_0^2\Big)\right) r\,dr\,d\beta \\ &= \int_{\frac{\pi}{2}}^{\frac{\pi}{2}+\beta_0} \frac{1}{2\pi}\exp\left(-\frac{1}{2}\left(\frac{x^2}{\cos^2\beta}-2\mu_0 x\frac{\cos(\alpha_0+\beta)}{\cos\beta}+\mu_0^2\right)\right)\frac{x}{\cos^2\beta}\,d\beta \\ &= \frac{1}{2\sqrt{2\pi}}\exp\left(-\frac{1}{2}(x-\mu_0\cos\alpha_0)^2\right)\left(1-\operatorname{erf}\left(\frac{1}{\sqrt{2}}(x\cot\beta_0-\mu_0\sin\alpha_0)\right)\right)\end{aligned}$$

and

$$\begin{aligned}\frac{d}{dx}G_3(x) &= \int_{\frac{\pi}{2}-\alpha_0}^{\frac{\pi}{2}} \frac{d}{dx}\int_0^{\frac{x}{\cos\beta}} \frac{1}{2\pi}\exp\left(-\frac{1}{2}\Big(r^2+2\mu_0 r\cos(\alpha_0+\beta)+\mu_0^2\Big)\right) r\,dr\,d\beta \\ &= \int_{\frac{\pi}{2}-\alpha_0}^{\frac{\pi}{2}} \frac{1}{2\pi}\exp\left(-\frac{1}{2}\left(\frac{x^2}{\cos^2\beta}+2\mu_0 x\frac{\cos(\alpha_0+\beta)}{\cos\beta}+\mu_0^2\right)\right)\frac{x}{\cos^2\beta}\,d\beta \\ &= \frac{1}{2\sqrt{2\pi}}\exp\left(-\frac{1}{2}(x+\mu_0\cos\alpha_0)^2\right)\left(1-\operatorname{erf}\left(\frac{1}{\sqrt{2}}(x\cot\alpha_0-\mu_0\sin\alpha_0)\right)\right).\end{aligned}$$

Next we consider $\frac{d}{dx}G_4(x)$ and $\frac{d}{dx}G_5(x)$. Rotating the figure $\frac{\pi}{2}+\alpha_0$ counter-clockwise, the generating parameter becomes $(-\mu_0\cos\alpha_0, -\mu_0\sin\alpha_0)$. Then

$$\begin{aligned}\frac{d}{dx}G_4(x) &= \int_{\frac{\pi}{2}}^{\frac{\pi}{2}+\alpha_0} \frac{d}{dx}\int_0^{-\frac{x}{\cos\beta}} \frac{1}{2\pi}\exp\left(-\frac{1}{2}\Big(r^2+2\mu_0 r\cos(\alpha_0-\beta)+\mu_0^2\Big)\right) r\,dr\,d\beta \\ &= \int_{\frac{\pi}{2}}^{\frac{\pi}{2}+\alpha_0} \frac{1}{2\pi}\exp\left(-\frac{1}{2}\left(\frac{x^2}{\cos^2\beta}-2\mu_0 x\frac{\cos(\alpha_0-\beta)}{\cos\beta}+\mu_0^2\right)\right)\frac{x}{\cos^2\beta}\,d\beta \\ &= \frac{1}{2\sqrt{2\pi}}\exp\left(-\frac{1}{2}(x-\mu_0\cos\alpha_0)^2\right)\left(1-\operatorname{erf}\left(\frac{1}{\sqrt{2}}(x\cot\alpha_0+\mu_0\sin\alpha_0)\right)\right)\end{aligned}$$

and

$$\begin{aligned}\frac{d}{dx}G_5(x) &= \int_{\frac{\pi}{2}-\beta_0}^{\frac{\pi}{2}} \frac{d}{dx}\int_0^{\frac{x}{\cos\beta}} \frac{1}{2\pi}\exp\left(-\frac{1}{2}\Big(r^2+2\mu_0 r\cos(\alpha_0-\beta)+\mu_0^2\Big)\right) r\,dr\,d\beta \\ &= \int_{\frac{\pi}{2}-\beta_0}^{\frac{\pi}{2}} \frac{1}{2\pi}\exp\left(-\frac{1}{2}\left(\frac{x^2}{\cos^2\beta}+2\mu_0 x\frac{\cos(\alpha_0-\beta)}{\cos\beta}+\mu_0^2\right)\right)\frac{x}{\cos^2\beta}\,d\beta \\ &= \frac{1}{2\sqrt{2\pi}}\exp\left(-\frac{1}{2}(x+\mu_0\cos\alpha_0)^2\right)\left(1-\operatorname{erf}\left(\frac{1}{\sqrt{2}}(x\cot\beta_0+\mu_0\sin\alpha_0)\right)\right).\end{aligned}$$

Finally, we consider $\frac{d}{dx}G_6(x)$, which is identical to $\frac{d}{dx}G_1(x)$, but after mapping the generating parameter as $(0, \mu_0) \mapsto (0, -\mu_0)$. Then

$$\frac{d}{dx}G_6(x) = \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)\left(1 - \operatorname{erf}\left(\frac{1}{\sqrt{2}}(x \tan(\alpha_0 + \beta_0) + \mu_0)\right)\right).$$

The claim follows after noting that $\tan(\alpha_0 + \beta_0) = \cot \beta_0$ and performing a change of variable to the squared Euclidean distance. □

Using the density of Proposition 6 for judging likelihood ratio statistics is still complicated by its dependence on the unknown true parameter ϕ_0. While several approaches to deal with this are discussed in [28], the simplest is to replace ϕ_0 by its MLE under the cut model. Although theory does not guarantee the good performance of this approach, in the next section we investigate performance through simulation.

If an expected *qcCF* has some small counts, say less than 5, and hence its normalization lies near the boundary of the simplex, an alternative testing procedure is necessary. In the case of only 2 small counts, the geometry of the parameter space far from a vertex can be ignored, giving essentially the same situation for the $T1$ and $T3$ tests already implemented in `MSC-quartets`. Either parametric bootstrapping from $\widehat{\theta}_0$, or a much faster precomputed approximation, can be used.

If only one expected count is small, the normalization lies near an edge of the simplex. Under the cut model, the other two counts should be approximately equal and approximately binomially distributed. Then a standard binomial test can be applied.

C.3 Cut test simulation

Figure 13 shows results of simulations comparing p-values for the likelihood ratio statistic for simulated *qcCFs* from the cut model, using the distribution of Proposition 6 with the MLE for ϕ_0, and a standard χ^2_1 distribution. While neither distribution produces the desired cumulative distribution of p-values, that of Proposition 6 comes closer when μ_0 has smaller magnitude. The value $\mu_0 = 0$ corresponds to the model singularity (1/3, 1/3, 1/3), where both tests will perform very conservatively for small significance levels, seldom rejecting the null model. As μ_0 is varied away from 0, performance improves.

While μ_0 depends on both the true model parameter and the sample size n, it has a simple interpretation: if σ_y is the standard deviation of the y-coordinate of the random observations, then $|\mu_0|\sigma_y$ is the distance between the generating parameter θ_0 and the model singularity.

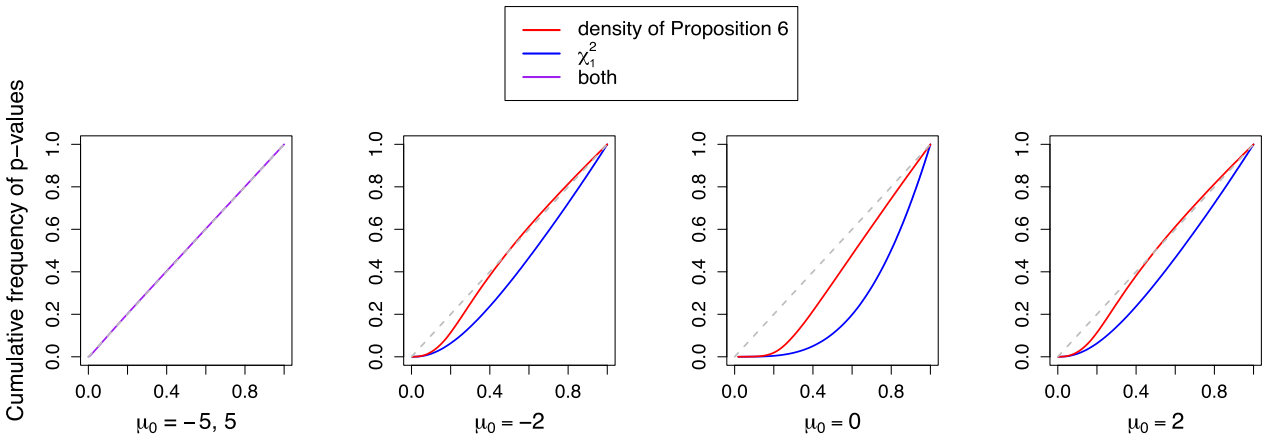

Fig. 13 Cumulative distributions of p-values computed from simulations, for the distribution of the likelihood ratio statistic given in Proposition 6 using maximum likelihood estimates of ϕ_0 (red), and for the χ^2_1 distribution (blue) for sample size $n = 10^6$. The cdf plots are indistinguishable for $|\mu_0|$ large. The diagonal line represents ideal behavior. At and near the model singularity, $\mu_0 = 0$, the distribution of Theorem 6 performs better than a χ^2_1

Acknowledgements
We thank Kristina Wicke for many helpful conversations, and for testing the software implementation extensively.

Funding
ESA and JAR were partially supported by NSF grant DMS 2051760 and NIH grant P20GM103395. HB was supported by NSF grant DMS 2331660. JDM was funded by The Australian Research Council Centre of Excellence for Plant Success in Nature and Agriculture (CE200100015).

Data availability
Empirical data, all generated by other researchers in previous publications, are publicly available.

Code availability
TINNiK is implemented as part of the `MSCquartets 2.0` R package, freely available on CRAN.

Declarations

Competing interests
The authors declare that they have no Competing interests.

Received: 20 April 2024 Accepted: 22 August 2024
Published online: 05 November 2024

References

1. Liu L, Yu L. Estimating species trees from unrooted gene trees. Syst Biol. 2011;60(5):661–7.
2. Liu L, Yu L, Kubatko L, Pearl DK, Edwards SV. Coalescent methods for estimating phylogenetic trees. Mol Phylogenetics Evol. 2009;53(1):320–8.
3. Flouri XT, Jiao Rannala B, Yang Z. Species tree inference with BPP using genomic sequences and the multispecies coalescent. MolBiol Evol. 2018;35(10):2585–93. https://doi.org/10.1093/molbev/msy147.
4. Zhang C, Rabiee M, Sayyari E, Mirarab S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinform. 2018;19(6):153. https://doi.org/10.1186/s12859-018-2129-y.
5. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, De Maio N, et al. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 2019;15(4):1006650.
6. Rhodes JA. Topological metrizations of trees, and new quartet methods of tree inference. IEEE/ACM Trans Comput Biol Bioinform. 2020;17(6):2107–18. https://doi.org/10.1109/TCBB.2019.2917204.
7. Yourdkhani S, Rhodes JA. Inferring metric trees from weighted quartets via an intertaxon distance. Bull Math Biol. 2020;82(7):97. https://doi.org/10.1007/s11538-020-00773-4.
8. Meng C, Kubatko LS. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. Theor Popul Biol. 2009;75(1):35–45. https://doi.org/10.1016/j.tpb.2008.10.004.
9. Yu Y, Than C, Degnan JH, Nakhleh L. Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. Syst Biol. 2011;60(2):138–49.
10. Yu Y, Degnan JH, Nakhleh L. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS Genet. 2012;8:1002660.
11. Degnan JH. Modeling hybridization under the network multispecies coalescent. Syst Biol. 2018;67(5):786–99. https://doi.org/10.1093/sysbio/syy040.
12. Zhu J, Yu Y, Nakhleh L. In the light of deep coalescence: revisiting trees within networks. BMC Bioinform. 2016;5:271–82.
13. Zhang C, Ogilvie HA, Drummond AJ, Stadler T. Bayesian inference of species networks from multilocus sequence data. Mol Biol Evol. 2017;35(2):504–17.
14. Zhu J, Wen D, Yu Y, Meudt HM, Nakhleh L. Bayesian inference of phylogenetic networks from bi-allelic genetic markers. PLoS Comput Biol. 2018;14(1):1005932.
15. Yu Y, Nakhleh L. A maximum pseudo-likelihood approach for phylogenetic networks. BMC Genom. 2015;16:10.
16. Solís-Lemus C, Ané C. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLoS Genet. 2016;12(3):1005896.
17. Allman ES, Baños H, Rhodes JA. NANUQ: a method for inferring species networks from gene trees under the coalescent model. Algorithms Mol Biol. 2019;14(24):1–25.
18. Kong S, Swofford DL, Kubatko LS. Inference of phylogenetic networks from sequence data using composite likelihood. Syst. Biol. 2024:syae054. https://doi.org/10.1093/sysbio/syaeo54.
19. Rhodes JA, Banos H, Xu J, Ané C. Identifying circular orders for blobs in phylogenetic networks. 2024. arXiv:2402.11693.
20. Gusfield D, Bansal V, Bafna V, Song YS. A decomposition theory for phylogenetic networks and incompatible characters. J Comput Biol. 2007;14(10):1247–72. https://doi.org/10.1089/cmb.2006.0137.
21. Allman ES, Baños H, Mitchell JD, Rhodes JA. The tree of blobs of a species network: identifiability under the coalescent. J Math Biol. 2023;86(1):10. https://doi.org/10.1007/s00285-022-01838-9.
22. Rhodes JA, Baños H, Mitchell JD, Allman ES. MSCquartets 1.0: quartet methods for species trees and networks under the multispecies coalescent model in R. Bioinformatics. 2020;37(12):1766–8. https://doi.org/10.1093/bioinformatics/btaa868.
23. Rhodes JA, Baños H, Mitchell JD, Allman ES. MSCquartets: An R package for Analyzing Gene Tree Quartets under the Multi-Species Coalescent. v2.0. 2023. https://cran.r-project.org/web/packages/MSCquartets/index.html.
24. Kong S, Pons JC, Kubatko L, Wicke K. Classes of explicit phylogenetic networks and their biological and mathematical significance. J Math Biol. 2022;84(6):47. https://doi.org/10.1007/s00285-022-01746-y.
25. Semple C, Steel M. Phylogenetics. Oxford: Oxford University Press; 2005.
26. Baños H. Identifying species network features from gene tree quartets. Bull Math Biol. 2019;81:494–534.
27. Allman ES, Mitchell JD, Rhodes JA. Gene tree discord, simplex plots, and statistical tests under the coalescent. Syst Biol. 2021;71(4):929–42. https://doi.org/10.1093/sysbio/syab008.
28. Mitchell JD, Allman ES, Rhodes JA. Hypothesis testing near singularities and boundaries. Electron J Statist. 2019;13(1):2150–93.
29. Ané C, Fogg J, Allman ES, Baños H, Rhodes JA. Anomalous networks under the multispecies coalescent: theory and prevalence. J Math Biol. 2024;88:29.
30. Studier J, Kepler K. A note on the neighbor-joining algorithm of Saitou and Nei. Mol Biol Evol. 1988;5:729–31.
31. Wang W, Barbetti J, Wong T, Thornlow B, Corbett-Detig R, Turakhia Y, Lanfear R, Minh BQ. DecentTree: scalable neighbour-joining for the genomic era. Bioinformatics. 2023;39(9):536. https://doi.org/10.1093/bioinformatics/btad536.
32. Lefort V, Desper R, Gascuel O. FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. Mol Biol Evol. 2015;32(10):2798–800.
33. Bryant D, Moulton V. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol. 2004;21:255–65.
34. Williamson SG. Combinatorics for computer science. Rockville: Computer Science Press; 1985.
35. Eddelbuettel D, Francois R. Rcpp: seamless R and C++ integration. J Stat Softw. 2011;40(8):1–18. https://doi.org/10.18637/jss.v040.i08.
36. Fogg J, Allman ES, Ané C. PhyloCoalSimulations: a simulator for network multispecies coalescent models, including a new extension for the inheritance of gene flow. Syst Biol. 2023;72(5):1171–9. https://doi.org/10.1093/sysbio/syad030.
37. Solís-Lemus C, Ané C. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLOS Genet. 2016;12(3):1–21. https://doi.org/10.1371/journal.pgen.1005896.
38. Vanderpool D, Minh BQ, Lanfear R, Hughes D, Murali S, Harris RA, Raveendran M, Muzny DM, Hibbins MS, Williamson RJ, Gibbs RA, Worley KC, Rogers J, Hahn MW. Primate phylogenomics uncovers multiple rapid radiations and ancient interspecific introgression. PLOS Biol. 2020;18(12):1–27. https://doi.org/10.1371/journal.pbio.3000954.
39. Ané C. QuartetNetworkGoodnessFit: a Julia package for phylogenetic networks analyses using four-taxon subsets. 2023. https://github.com/cecileane/QuartetNetworkGoodnessFit.jl. v0.5.0.

40. Kleinkopf J, Roberts W, Wagner W, Roalson E. Diversification of hawaiian cyrtandra (gesneriaceae) under the influence of incomplete lineage sorting and hybridization. J Syst Evol. 2019. https://doi.org/10.1111/jse.12519.
41. Huson D, Kloepper T, Lockhart P, Steel M. Reconstruction of reticulate networks from gene trees. Berlin: Springer; 2005. p. 233–49.

Publisher's Note