Anchored distances for quartet-based estimation of phylogenetic trees and applications to coalescent-based analyses

Erfan Sayyari[1] and Siavash Mirarab*[1]

[1]University of California, San Diego, Department of Electrical and Computer Engineering

Abstract

Inferring trees based on pairwise distances has utility in phylogenetic reconstruction broadly and specifically for estimation of species trees from gene trees under the coalescent model. In this paper, we introduce a new general approach for computing distances between a pair of leaves by first computing the topology and the internal branch length of quartet tree that includes that pair of leaves and two additional leaves, which we call "anchors". We show that using these anchored distances, a family of distance-based reconstruction methods can be designed, ranging in complexity between $\Theta(n^2)$ and $\Theta(n^4)$. We then describe how anchored distances can be used for statistically consistent estimation of species trees from discordant unrooted gene trees under the multi-species coalescent model. The resulting method, which we call DISTIQUE, is competitive with the best alternative methods. Supplementary material, code, and datasets are available at `http://esayyari.github.io/DISTIQUE`.

1 Introduction

Inferring trees using pairwise distances is a well-studied approach to phylogenetic reconstruction [1–4]. As maximum likelihood and Bayesian methods became faster [5–7], distance-based methods started to fall out of favor, but they still remain useful, for example, for quickly building guide trees for multiple sequence alignment [8–10]. Furthermore, many new distance-based methods have been introduced in recent years for coalescent-based analysis of gene trees [11–17].

The history of species and genes can be discordant [18], and Incomplete Lineage Sorting (ILS), modeled by the multi-species coalescent model [19], is one of the main causes of discordance. One approach for estimating the species phylogeny in the face of such discordances is first estimating a gene tree for each gene, and then summarizing them to build a species tree. The summary method, thus, takes as input a set of gene trees and returns a species tree. One desirable property of summary methods is statistical consistency (i.e., theoretical guarantees that it returns the correct species tree with an arbitrarily high probability as the number of (error-free) genes increases to infinity. Many statistically consistent summary methods have been developed (e.g., ASTRAL [14, 20], MP-EST [15], BUCKy-population [16]), including some that are distance-based (e.g., NJst [11], a new implementation of NJst called ASTRID [12], STAR [13], and GLASS [17]). Coalescent-based species tree estimation is a vibrant field of research, with rapid advances and many recent examples of successful biological analyses [21–23] (but see [24–28] for criticism of these methods).

A powerful general approach to phylogenetic reconstruction is analyzing quartets, which are subsets of four leaves in a tree. Unrooted quartet trees can have one of only three possible topologies, making them easy to analyze. Thus, one can first infer a set of quartet trees and then combine them to build a tree on the full dataset [2, 29, 30]. Relatedly, one can use induced quartets to estimate

*Corresponding author: smirarab@ucsd.edu

a so-called supertree [31] from a collection of input trees [30, 32–35]. Quartet-based supertree estimation has been revived in recent years [16, 20, 36–38] because of its connections to coalescent-based analyses [39–41]. Under the coalescent model, for unrooted species trees with four leaves, the most likely unrooted gene tree is identical to the species tree (but this is not true for larger trees [40,42]). Furthermore, the lengths of the internal quartet branches of species tree (in coalescent units) define the probabilities of the gene tree quartet topologies [40]. These observations are used in recent statistically consistent quartet-based species tree estimation methods. For example, ASTRAL seeks to find the species tree that shares the maximum number of quartet trees with input gene trees, an optimization problem that produces a statistically consistent estimator [20].

In this paper, we first introduce a general family of methods to compute pairwise leaf distances given a way to estimate quartet trees (their topology and the length of their single internal branch, called "quartet length" henceforth). We choose two arbitrary "anchor" leaves, and define pairwise distances between any other pair of leaves by computing the quartet tree topology and length for the quartet that consists of the two anchors and the two leaves of interest. We show that as long as the quartet estimator is consistent, our computed distance matrix is additive, and can be used for statistically consistent inference of tree *topologies* (but not branch lengths, except one variant that can compute *internal* branch lengths). These "double anchored" distances are defined for all leaves except the two anchors, but we introduce a family of methods that uses double anchored distances to reconstruct trees on the complete leaf-set. These methods can be adjusted to have running times that range between $\Theta(n^2)$ and $\Theta(n^4)$ for n species (assuming other factors are constant).

A motivation of our anchored quartet-based distance estimation is to enable pairwise distance calculation when estimating quartet topology and length is straightforward but estimating leaf distances is not. Coalescent-based analyses have these properties. We introduce a family of statistically consistent coalescent-based summary methods, called DISTIQUE, which use anchoring to estimate species trees from gene trees. We evaluate the accuracy of DISTIQUE on simulated and biological data, and show that its accuracy is competitive with the best alternative methods.

2 Theoretical results

We start with general results, and then describe their applications to coalescent-based analyses.

2.1 Anchored distance matrices

Definitions: We denote the leaf-set by $\mathcal{L}$ and let $n = |\mathcal{L}|$. We only consider unrooted trees, and for a tree T on $\mathcal{L}$, the set of quartet trees induced on all possible $\binom{n}{4}$ quartets of leaves is denoted by $\mathcal{Q}^T$. We use $ab.cd$ to note that in the quartet tree on $\{a, b, c, d\}$, a and b are sisters (there are only three possibilities). A tree T is equivalent to a distance matrix D^T, computed by summing lengths of the edges between pairs of leaves, and a distance matrix that corresponds to a tree is called additive [43]. We refer to the unique tree [43] associated with the additive distance matrix D as T^D (omitting superscripts when clear). $T|\mathcal{L}'$ denotes tree T restricted to the leaf-set $\mathcal{L}'$.

To test for the additivity of a distance matrix D, we can use the four point condition [43], which states, for quartet of leaves $\{a, b, c, d\} \in \mathcal{L}$, the median and the maximum of the following three values should be the same: $\{D[a, b] + D[c, d], D[a, c] + D[b, d], D[a, d] + D[b, c]\}$. When internal branch lengths are assumed positive, as we do throughout this paper, the minimum value is strictly smaller than the median. Let w.l.o.g. $D[a, b] + D[c, d]$ be the smallest value; the quartet topology $ab.cd$ is induced by T^D (i.e., $ab.cd \in \mathcal{Q}^T$). If we denote the length of the single internal branch in this quartet tree (which we call its "quartet length") by $d_D(a, b, c, d)$, it is easy to show that $d_D(a, b, c, d) = \frac{1}{2}[D[a, c] + D[b, d] - (D[a, b] + D[c, d])]$.

Proper setting: (β, α, f) is called a proper setting if $\beta, \alpha > 0$ are two constants and $f(x)$ is a monotonically increasing function, and $0 < f(x) < \beta$ for $x > 0$.

Anchored Distance Matrices: Given a tree T^D, a proper setting, and "anchor" leaves $u, v \in \mathcal{L}$,

$$D'_{uv}[a,b] = \begin{cases} \beta + \alpha.d_D(a,b,u,v) & ab.uv \notin \mathcal{Q}^T \\ \beta - f(d_D(a,b,u,v)) & ab.uv \in \mathcal{Q}^T \end{cases} \quad (1)$$

is a distance matrix on the leaf-set $\mathcal{L} - \{u, v\}$; we call D'_{uv} a "double anchored distance matrix", and say that D'_{uv} is induced from D anchored by u, v. Similarly, given a single anchor leaf $v \in \mathcal{L}$,

$$D'_v[a,b] = \sum_{u \in \mathcal{L} - \{a,b,v\}} D'_{uv}[a,b] \quad (2)$$

is a distance matrix on the leaf-set $\mathcal{L} - \{v\}$; we call D'_v a "single anchored distance matrix". And,

$$D'[a,b] = \sum_{v \in \mathcal{L} - \{a,b\}} D'_v[a,b] = \sum_{u,v \in \mathcal{L} - \{a,b\}} D'_{uv}[a,b] \quad (3)$$

is called an "all-pairs anchored distance matrix" and is defined on the complete leaf-set $\mathcal{L}$. Finally,

$$D''[a,b] = \max_{u,v \in \mathcal{L} - \{a,b\}} \max(0, \frac{D'_{uv}[a,b] - \beta}{\alpha}) \quad (4)$$

is called an "all-pairs anchored maximum distance matrix", and is defined on the complete $\mathcal{L}$. Thus, $D'[a,b]$ and $D''[a,b]$ are the sum and max of double anchored distances for all $\binom{n-2}{2}$ ways of choosing two anchors other than (a,b).

Theorem 1 *A double anchored distance matrix D'_{uv} induced from an additive distance matrix D^T anchored by an arbitrary pair of leaves $u, v \in \mathcal{L}$ is an additive distance matrix for the leaf-set $\mathcal{L}' = \mathcal{L} - \{u, v\}$ and corresponds to a tree that is topologically identical to $T|\mathcal{L}'$.*

Theorem 2 *A single anchored distance matrix D'_v induced from an additive distance matrix D^T anchored by an arbitrary $v \in \mathcal{L}$ is an additive distance matrix for the leaf-set $\mathcal{L}' = \mathcal{L} - \{v\}$ and corresponds to a tree that is topologically identical to $T|\mathcal{L}'$.*

Theorem 3 *An all-pairs anchored distance matrix D' induced from an additive distance matrix D^T is additive and corresponds to a tree with identical topology to T.*

Proofs of all three theorems are given in Appendix A, where we first show that the four point condition holds for an arbitrarily chosen quartet of leaves in all possible scenarios for adding u and v on a quartet tree. Proof of Theorem 2 is similar and involves adding one anchor to a quartet tree. Theorem 3 is proved by dividing Equation 3 to sums that resemble Theorems 1 and 2, and an extra case that can be easily analyzed.

Theorem 4 *An All-pairs anchored maximum distance matrix D'' induced from additive matrix D^T is additive and corresponds to a tree with identical topology and internal branch lengths to T.*

Proof. We prove that Equation 4 returns the sum of internal branch lengths on the path from a to b on T (we denote this by D^T_{ab}); the theorem automatically follows as D^T is additive and we are only interested in internal branch lengths. For simplicity, set $\alpha = 1$, but the proof holds for other α values. If (a,b) are not sisters, there is at least an anchor pair (u,v) with quartet topology $au.bv$ and $d_D(a,b,u,v) = D^T_{ab}$ (pick u from sister group of a after rooting T on b and vice-versa). With this choice, $D'_{uv}[a,b] - \beta = d_D(a,b,u,v) = D^T_{ab}$; moreover, $D'_{u'v'}[a,b] - \beta$ for other anchors are no bigger than D^T_{ab} (if they have $ab.u'v'$ topology, $D'_{u'v'}[a,b] < \beta$, and otherwise, it will give the length for a subset of the path from a to b). Thus, max returns D^T_{ab}, as desired. When (a,b) are sisters, $D^T_{ab} = 0$. For sister leaves, $D'_{uv}[a,b] < 0$ for any (u,v), and thus, $D''[a,b] = 0$, as desired. □

Algorithm 1 Anchored quartet-based algorithms. $d^{\mathcal{D}}(.)$ is a quartet estimator and returns the quartet topology and length. $(\beta, \alpha, f(x))$ gives a proper setting. Stopping criteria should *at least* require that every leaf is in at least a tree $T \in \mathcal{T}$ (i.e., no leaf is always an anchor). DAD, SAD, AAD, and AMD give algorithms for double, single, all-pairs, and all-pairs maximum anchored distance-based reconstruction, respectively.

function $D'_{u,v}(a,b)$
  $(t,d) \leftarrow d^{\mathcal{D}}(a,b,u,v)$
  if $t = ab.uv$ **then return** $\beta - f(d)$
  else return $\beta + \alpha.d$

function DAD($\mathcal{D}$)
  repeat for $\{u,v\} \subset \mathcal{L}$
    $D'_{uv} \leftarrow 0_{n-2 \times n-2}$
    for $\{a,b\} \subset \mathcal{L} - \{u,v\}$ **do**
      $D'[a,b] = D'_{u,v}(a,b)$
    add NeighborJ(D'_{uv}) to $\mathcal{T}$
  until stopping criteria
  return SuperFine($\mathcal{T}$)

function SAD($\mathcal{D}$)
  repeat for $\{u\} \subset \mathcal{L}$
    $D'_u \leftarrow 0_{n-1 \times n-1}$
    for $\{v\} \subset \mathcal{L} - \{u\}$ **do**
      for $\{a,b\} \subset \mathcal{L} - \{u,v\}$ **do**
        $D'[a,b] += D'_{u,v}(a,b)$
    add NeighborJ(D'_u) to $\mathcal{T}$
  until stopping criteria
  return SuperFine($\mathcal{T}$)

function AAD($\mathcal{D}$)
  $D' \leftarrow 0_{n \times n}$
  for $\{a,b\} \subset \mathcal{L}$ **do**
    for $\{u,v\} \subset \mathcal{L} - \{a,b\}$ **do**
      $D'[a,b] += D'_{u,v}(a,b)$
  return NeighborJ(D')

function AMD($\mathcal{D}$)
  $D'' \leftarrow 0_{n \times n}$
  for $\{a,b\} \subset \mathcal{L}$ **do**
    for $\{u,v\} \subset \mathcal{L} - \{a,b\}$ **do**
      $D''[a,b] = \max(D''[a,b], D'_{u,v}(a,b))$
  return NeighborJ(D'')

2.2 Phylogenetic reconstruction using anchored quartet-based distances

Theorems 1-4 can be used to design a general family of tree reconstruction methods based on quartet distances. Let $\mathcal{D}$ denote data from which we want to infer a tree; we do not make assumptions here about the nature of the data, but require a way to estimate quartet trees:

(Consistent) quartet estimator: A quartet estimator $d^{\mathcal{D}}(q)$ is a function that given a quartet of leaves $\{a,b,c,d\}$, uses $\mathcal{D}$ to estimate the quartet tree topology and the quartet length. A quartet estimator is statistically consistent if, as size of $\mathcal{D}$ increases, the estimated quartet tree topology and quartet length both converge in probability to correct values.

Statistically consistent quartet estimators can be designed for various types of data. For example, the four point condition gives a way to estimate quartet trees from sequence data [44], and the log-det method gives a model-based way of estimating corrected branch lengths [45]. As the next section shows, coalescent theory can also be used to design consistent quartet estimators.

Given a consistent quartet estimator $d^{\mathcal{D}}(.)$, Theorems 1-4 can be exploited to define a family of statistically consistent phylogenetic reconstruction methods that range in running time between $\Theta(n^2)$ and $\Theta(n^4)$ (fixing other parameters). Algorithm 1 shows general forms of these algorithms. All-pairs anchored distance-based (AAD) and all-pairs maximum distanced-based (AMD) are the simplest methods in this family, which use Equations 3 and 4 to compute the distance matrix (i.e., sum/max of double anchored distance over all possible $\binom{n-2}{2}$ anchors). They then compute the tree using a consistent distance-based method (here, neighbor joining [1]).

Theorem 5 *All-pairs anchored distanced-based (AAD) and all-pairs maximum distance-based (AMD) phylogenetic reconstruction algorithms shown in Algorithm 1 are statistically consistent.*

Proof (sketch). Since $d^{\mathcal{D}}(.)$ is assumed statically consistent, in limit, it will return the correct quartet topology with arbitrarily high probability, and its estimates of quartet lengths can be made arbitrarily close to true values with any desired probability. From Theorem 3 and 4, it follows that the distance matrices for AAD and AMD become arbitrarily close to additive, with high probability. The proof follows from statistical consistency of neighbor joining for distance matrices that are in limit arbitrarily close to additivity [1]. □

Using double or single anchored distance matrices would enable us to use only a subset of all $\binom{n-2}{2}$ anchors, but a difficulty is that these matrices do not include the complete leaf-set. To address

this difficulty, we compute a set of trees, each on $n-1$ or $n-2$ leaves, and then use a supertree method (i.e., SuperFine [46,47]) to combine them. General forms of algorithms for single anchored (SAD) or double anchored (DAD) distance-based tree inference are shown in Algorithms 1. A pair of anchors $\{u, v\}$ (or, a single anchor u for SAD) are selected with some criteria (e.g., random w/o replacement), a distance matrix is generated for $\mathcal{L} - \{u, v\}$ (or, $\mathcal{L} - \{u\}$ for SAD), and a distance method, neighbor joining [1], is used to estimate a tree on $\mathcal{L} - \{u, v\}$ (or, $\mathcal{L} - \{u\}$ for SAD). This process repeats until stopping criteria are met; the stopping criteria need to include that each leaf be at least in one estimated tree. Finally, the collection of trees is combined using a supertree method (e.g., SuperFine) to get a single tree. If SuperFine is used, we can show:

Theorem 6 *Single and double anchored distanced-based (SAD and DAD) phylogenetic reconstruction algorithms shown in Algorithm 1 are both statistically consistent.*

Proof (sketch). Assume the stopping criteria are reached after k rounds. By arguments similar to those used for Theorem 5, we can argue that each individual tree $T_i, 1 \leq i \leq k$ on $n-1$ (for SAD) or $n-2$ (for DAD) leaves is a statistically consistent estimate; thus, for any $\epsilon' < 1$ and i, there is a dataset size such that with probability $1-\epsilon'$, T_i is recovered correctly. Take the largest of these dataset sizes; with this size, with probability at least $(1-\epsilon')^k$, every T_i is correct. Thus, setting $\epsilon' < 1 - (1-\epsilon)^{\frac{1}{k}}$ allows us to argue that for any ϵ, there is a dataset size where all T_is are correct with probability higher than $1-\epsilon$. Then, by Theorem 1 of [46], the strict consensus merger (and by extension SuperFine) applied to the set of T_i trees returns the correct tree on $\mathcal{L}$. □

Branch lengths: Our anchored distances set the length of terminal branch to zero. Nevertheless, we proved these matrices additive for the true tree *topology*. A constant can be added to all distances for inference methods that expect positive terminal lengths. AMD returns statistically consistent estimates of *internal* branch lengths, but branch lengths from other methods should be ignored.

Running time analysis: AAD and AMD clearly require $\Theta(n^4)$ running time (assuming quartet estimator is constant time) to build the distance matrix; using the default neighbor joining algorithm, which requires $O(n^3)$, would result in $\Theta(n^4)$ asymptotic running time for AAD and AMD. The running times of SAD and DAD depend on the design of the stopping criteria, and also the exact distance method and SuperFine implementation used. To build the distance matrix, each iteration of DAD and SAD requires $\Theta(n^2)$ and $\Theta(n^3)$, respectively; if a fast neighbor joining algorithm is used (e.g., FNJ [48], or NINJA [4]), the running time of building each tree is $O(n^2)$. If the stopping criterion is limited by a constant, building the set of trees on $n-2$ or $n-1$ leaves using DAD and SAD becomes $\Theta(n^2)$ and $\Theta(n^3)$ respectively. However, the stopping criteria can limit the number of rounds to a function of n; for example, a stopping criteria that allows for up to $\sqrt{n}$ rounds would result in $O(n^{2.5})$ and $O(n^{3.5})$ respectively. Clearly, any function between $\Theta(n^2)$ and $\Theta(n^4)$ can be achieved by adjusting the stopping criteria. Finally, DAD and SAD use SuperFine to combine the set of k trees. The running time of SuperFine depends on the supertree method runs inside SuperFine; if used with a quadratic time algorithm, SuperFine becomes quadratic also, and thus, our DAD and SAD become $\Theta(n^2)$ and $\Theta(n^4)$ respectively.

2.3 DISTIQUE: anchored coalescent-based species tree estimation

Our anchored quartet-based methods enable calculation of pairwise distances, even when pairwise distances are not easy to estimate, as long as quartet distances can be computed. Species tree estimation under multi-species coalescent model [19] (which models ILS) is an example of such a scenario. Here, the input dataset $\mathcal{D}$ is a set of gene trees, and we seek to find a species tree. We show how our anchored quartet-bsaed distances can be used for statistically consistent species tree estimation. We assume in our theoretical analyses that the input gene trees are generated using the multi-species coalescent process, and that each gene includes exactly one individual per species.

Coalescent-based quartet estimator: For a given quartet q, let $p_1 \leq p_2 \leq p_3$ denote frequencies of the three quartet topologies induced by gene trees; we define the coalescence-based quartet estimator $d^{\mathcal{D}}(q)$ to return the topology that has frequency p_3, with quartet length set to $-\ln 3 - \ln p_1$.

Under the multi-species coalescent process, for four taxa, the species tree topology has higher probability than the two alternative frequencies [39]; moreover, for quartet length d in the species tree, the probability of the most probable quartet gene tree is $1 - \frac{2}{3}e^{-d}$, and the probability of the alternative topologies are both $\frac{1}{3}e^{-d}$ [39, 49]. It immediately follows that:

Corollary 7 *The coalescent-based quartet estimator is statistically consistent.*

DISTIQUE: We define $\beta = \ln 3$, $\alpha = 1$, and $f(x) = \ln(3 - 2e^{-x})$. For $x > 0$, $f(x)$ is clearly positive, monotonic, and bounded from above by $\beta = \ln 3$; thus, (β, α, f) is a proper setting. Moreover, it is easy to show that with these choices, the calculation of Equation 1 greatly simplifies to $D'_{uv}[a, b] = -\ln p(ab.uv)$. Thus, to define distances, we can simply find the frequency of all quartet topologies in gene trees and use methods shown in Algorithm 1 to reconstruct the species tree (replacing function $D'_{u,v}(a, b)$ with $-\ln p(ab.uv)$). We call the resulting family of methods DISTIQUE (Distance-based Inference of Species Trees from Induced QUartet Elements). Note that the all-pairs distance matrix simplifies to:

$$D'[a, b] = \sum_{u,v \in \mathcal{L} - \{a,b\}} -\ln p(ab.uv) \tag{5}$$

Theorem 8 *DISTIQUE methods are statistically consistent under multi-species coalescent model.*

Proof. By Corollary 7, our choice of quartet estimator is statistically consistent. Since $(\beta, \alpha, f(x))$ is a proper setting, Theorems 5 and 6 prove that DISTIQUE methods are statistical consistency. □

3 Experimental evaluation

We use simulated and real datasets to evaluate the accuracy of four versions of DISTIQUE. We evaluate species tree accuracy, and measure it using False Negative (FN) rate, which is equivalent to normalized RF distance [50] here because all estimated species trees were fully resolved.

3.1 Methods

We compare four versions of DISTIQUE, described below, against each other, and against ASTRAL-II [14], which is a quartet-based method, the NJst algorithm [11], implemented in ASTRID [12], which is a distance-based method, and concatenation using RAxML [5]. We use ASTRAL and ASTRID because of their connections to DISTIQUE, and because they have been found to be more accurate than alternative summary methods [11, 12, 14, 20, 51, 52].

DISTIQUE Variations In this paper, we evaluate only two variants of DISTIQUE: AAD and AMD (all-pairs sum and max). We chose the two all-pairs methods because they do not require a strategy for choosing anchors or stopping criteria, nor do they require the use of a supertree method. Thus, these are the simplest versions to test without extensive exploration of parameter choices. The two all-pairs DISTIQUE are the slowest variants, and future work needs to explore DISTIQUE family more broadly. Even these simple variants involved the following design choices.

Smoothing: The gene tree quartet probabilities are exponentially related to the species tree quartet lengths. Thus, for branches that are even moderately long, getting zero frequencies for some alternative quartet topologies becomes likely. For example, for a quartet species tree with length 8 (in coalescent units [49]), the probability of seeing no discordance among 1000 genes is 98%. Thus, using a simple empirical frequency estimator, we are likely to get zero probabilities, which

produce distances of infinity (Eq. 5). To avoid this problem, we use *Krichevsky-Trofimov (KT)* [53] or add-half estimator, which means a pseudo-count of 0.5 is added for each quartet topology. This estimator was shown to reach the min-max cumulative loss for KL divergence asymptotically [53].

Consensus resolution: As Section 4.1 will show, the naive application of DISTIQUE, while competitively accurate on model conditions with high ILS, has poor accuracy for low ILS. This is surprising because the low discordance between gene trees under low ILS should make inference easy. However, with low discordance, true probabilities of many quartet trees become close to zero and difficult to estimate (hence the need for smoothing). With adjacent long branches and quartet probabilities close to zero, it becomes impossible to compute distances that reflect the true topology from limited data (resembling the saturation problem in traditional distance-based methods [2]). For example, branch lengths of 8, 16, or 32 are all likely to result in no gene tree discordance given 1000 genes, making it impossible to distinguish between these different distances. We can construct examples when all gene trees are likely identical, yet AAD with smoothing is mislead (Fig. S8).

Long branches with low discordance are easy to recover, since they appear in most gene trees. A simple majority rule (50%) consensus of gene trees would return all the long branches. The 50% consensus can be unresolved, but is proved *not* positively misleading under the coalescent model [54]. Thus, we can simply compute the majority consensus and then use DISTIQUE to resolve polytomies in the consensus tree, and this method would remain statistically consistent. To resolve a polytomy, we first assign a label to each branch pendant to it (defining a mapping from species to labels), and then build a tree using DISTIQUE with the labels as leaves; this tree defines a resolution of the polytomy. To use DISTIQUE with labels, we need to find empirical frequencies for quartets of labels. To do so, we select all quartets of species such that each species belongs to a different label (this will be a non-empty set), and set the frequency of the label quartet to the arithmetic mean of the frequencies of corresponding species quartets (we also tried geometric mean and root mean square, and observed no differences in results). We apply DISTIQUE to these label frequencies, which produce a tree on the label set and therefore a resolution of the polytomy.

Polytomies in input: DISTIQUE-AAD can seamlessly handle polytomies in input gene trees; unresolved nodes in gene trees result in unresolved quartet trees for that gene, which we exclude from our calculation of the empirical quartet frequencies.

DISTIQUE is implemented in python, using external libraries for manipulating trees (Dendropy [55]) and for recovering quartet trees from gene trees [56]. We used FastME [57] as our distance-based method, and used the same tool inside ASTRID (see Supplementary material for details); we also tried other distance-based method PhyD* [58] and observed no meaningful differences. Running time of DISTIQUE and other methods is given in Figures S4 and S5.

3.2 Datasets

We use three sets of simulated datasets from the literature in our studies: the mammalian dataset of [24], the avian dataset of [27,59], and the 11-taxon dataset of [51]. The first two datasets are based on biological data and have a single species tree topology, whereas the 11-taxon dataset is simulated using SimPhy [60] and has a different species tree per replicate. For the 11-taxon dataset, we have four levels of ILS, ranging from low (M1) to very high (M4), and for each case, we vary both the number of genes (100, 500, 1000) and the number of sites per gene (from 10 to 200). For the mammalian and avian datasets, we create two different collections, one where we fix the number of genes (to 200 for mammalian and to 1000 for avian) and vary the amount of ILS, and a second collection, where we fix the amount of ILS (to 0.2X for mammalian and 1X for avian) and vary the number of genes (100 to 3200 for mammalian and 200 to 2000 for avian). The amount of ILS is changed by multiplying or dividing branch lengths by 2 or 5; shorter branches (0.2X and

Table 1: Statistical significance of species tree topological error, comparing DISTIQUE-AAD-Cons versus ASTRAL and ASTRID, run on estimated gene trees, tested using ANOVA, with multiple test correction using Benjamini-Hochberg [44] ($n = 24$; $\alpha = 0.05$). The columns labelled *method* show p-values of the comparison between methods; those labelled *m:ILS*, *m:sites*, or *m:genes* show significance of the interaction between *method* and other parameters. †: DISTIQUE is outperformed by ASTRAL or ASTRID; ∗: DISTIQUE outperforms ASTRAL or ASTRID.

	DISTIQUE versus ASTRAL				DISTIQUE versus ASTRID			
	method	m:ILS	m:sites	m:genes	method	m:ILS	m:sites	m:genes
11-taxon	0.9861	0.9861	0.9861	0.9861	$< \mathbf{10^{-4}}^{*}$	$< \mathbf{10^{-4}}$	0.9861	$< \mathbf{10^{-4}}$
avian-1X	0.0856	–	–	0.7613	$\mathbf{0.0001}^{\dagger}$	–	–	0.9929
avian-1000g	0.0511	$< \mathbf{10^{-4}}$	–	–	$< \mathbf{10^{-4}}^{\dagger}$	0.8667	–	–
mammalian-0.2X	0.0856	–	–	0.6096	0.7613	–	–	0.8667
mammalian-200g	0.0856	0.0868	–	–	0.9861	**0.0161**	–	–

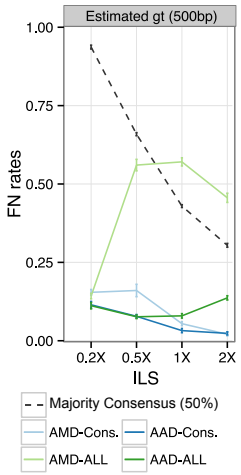

Figure 1: DISTIQUE variants on mammalian datasets with 200 genes.

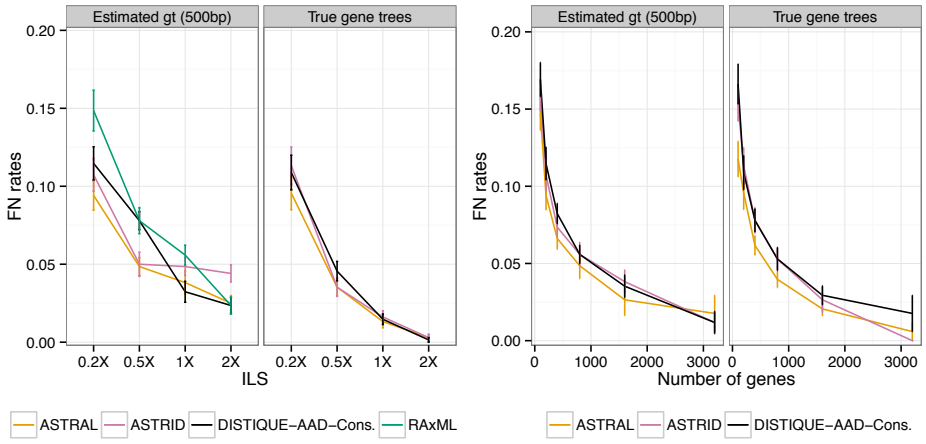

Figure 2: DISTIQUE-AAD-Cons versus other methods on the mammalian dataset. (left) number of genes: 200; (right) ILS: 0.2X. Mean and standard error of species tree error is shown for true and estimated gene trees (500bp alignments).

0.5X) produce more ILS and longer branches reduce ILS See Table S1. For the biological analyses, we re-analyzed a dataset of 2022 supergene trees from an avian phylogenomic dataset [22, 27].

4 Results

4.1 Comparison between DISTIQUE variants

We compare DISTIQUE-AAD and DISTIQUE-AMD (sum and maximum of all-pairs), each applied to either the entire dataset (ALL) or to polytomies of a 50% majority rule consensus (Cons). Figure 1 compares these four versions of DISTIQUE on the mammalian dataset as we vary the amount of ILS, and also shows the missing branch rate for the majority consensus tree. With very high ILS (0.2X), all methods have similar accuracy, and both AAD variants are slightly better than AMD. As ILS decreases, a surprising pattern emerges. When DISTIQUE is applied to the entire dataset, the error unexpectedly goes up with decreased ILS, a pattern that is more pronounced for AMD. As discussed before, we attribute this pattern to difficulties of estimating long quartet lengths. When DISTIQUE is used to resolve polytomies in the consensus tree, the accuracy improves with decreased ILS, as expected. Note that even with reduced ILS, the consensus tree misses more than 25% of branches, and leaves some polytomies for DISTIQUE to resolve. Similar results were obtained on other datasets (Supplementary Figs. S1-S3). Since DISTIQUE-ADD-Cons was consistently our best method, we use this variant for comparisons to other methods.

4.2 DISTIQUE versus other methods

Mammalian: Figure 2 compares DISTIQUE against other summary methods as we change the level of ILS or the number of genes. As expected, reducing ILS and increasing the number of genes result in improved accuracy for all methods, and error is somewhat higher when estimated gene trees are used instead of true gene trees. With true gene trees, summary methods have very similar accuracy. When estimated gene trees are used, overall, DISTIQUE and ASTRID (the two distance-based methods) have similar accuracies (with no statistically significance differences according to an ANOVA test with FDR correction [44]; see Table 1). However, their relative accuracy depends on the level of ILS (the interaction effect is statistically significant - Table 1), where DISTIQUE is better with lower ILS and ASTRID with higher ILS. Comparing the two quartet-based methods, DISTIQUE and ASTRAL, it seems ASTRAL has a slight advantage but the differences are not statistically significant ($p = 0.086$). On the 500bp dataset with 200 genes, DISTIQUE was always at least as good as concatenation (RAxML was not run for 0.2X in [24] due to running time).

Avian: Figure 3 compares DISTIQUE against other methods on the avian datasets, when the amounts of ILS or the number of genes are varied. With true gene trees, the distance-based methods ASTRID and DISTIQUE slightly outperform ASTRAL, but differences tend to be small. On estimated gene trees, ASTRID and ASTRAL both seem to outperform DISTIQUE, but differences are statistically significant only for ASTRID (Table 1). The relative performance of DISTIQUE and ASTRAL is not significantly impacted by the number of genes, but the amount of ILS does have a statistically significant impact (Table 1); with more ILS, DISTIQUE has better accuracy and with lower ILS, ASTRAL is more accurate. The relative performance of DISTIQUE and ASTRID is impacted by neither the amount of ILS nor the number of genes. Interestingly, on this dataset, the species tree error based on estimated gene trees is much higher than the error with true gene trees (twice the error or more in most cases). Concatenation was the least accurate method in general, except on the dataset with reduced ILS (2X), where it was a bit more accurate than DISTIQUE.

11-taxon: Figure 4 shows results on the 11-taxon datasets, with varying amount of ILS, number of genes, and the numbers of sites per gene (which controls the amount of gene tree error). Overall, there are no statistically significant differences between DISTIQUE and ASTRAL, which have extremely close accuracies for most conditions. DISTIQUE is significantly more accurate than ASTRID on this dataset, and the magnitude of the difference depends on both the amount of ILS and number of genes, but not the number of sites (Table 1). Thus, differences become larger as the amount of ILS increases and the number of genes decreases. Finally, the relative performance of concatenation (RAxML) and other methods seems to be impacted by the amount of ILS and the number of sites. With low ILS, concatenation has much better accuracy than summary methods, including DISTIQUE, but as the amount of ILS increases, its error increases rapidly. Also, with few sits per gene, differences between concatenation and summary methods are small, but as the number of sites per gene increases and gene trees become more accurate, summary methods improve rapidly, while the accuracy of concatenation fails to improve as much.

4.3 Biological results

On the avian supergene trees, we ran ASTRAL, ASTRID, and DISTIQUE with multi-locus bootstrapping [61]; all three methods generated trees with relatively high bootstrap support (on average 95% for ASTRAL, 91% for NJst, and 97% for DISTIQUE), and were all different from the published tree generated using MP-EST [22] (trees shown in Figs. S6, S7). DISTIQUE and ASTRID differed on two branches, which both had low support in ASTRID. DISTIQUE and ASTRAL differed in five branches, and two of these in DISTIQUE and three of them in ASTRAL had high support. The presence of highly supported conflict between coalescent-based methods indicates a need for skeptical and careful analysis of results of any single method on real data.

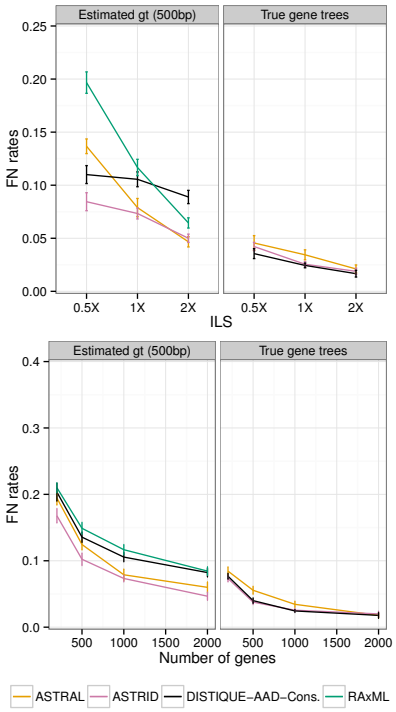

Figure 3: Results on avian datasets. (top) 1000 genes; (bottom) 1X ILS. Mean and standard error of species tree error is shown for true and estimated gene trees (20 replicates).

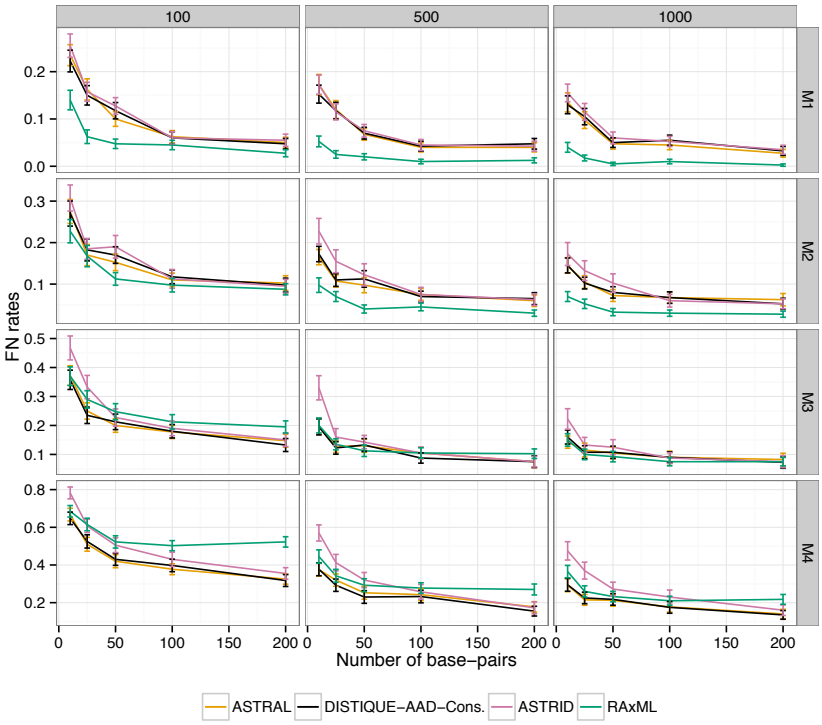

Figure 4: DISTIQUE versus other methods on the 11-taxon dataset. Columns show different number of genes, and rows show different levels of ILS. Mean and standard error of species tree error is shown for estimated gene trees with varying number of sites per gene (50 replicates).

5 Discussions and future work

We introduced a new distance-based approach for tree reconstruction by inferring topologies and internal branch lengths for quartets of leaves. Our new anchoring technique computes additive pairwise distances from quartet trees (potentially using a subset of all $\Theta(n^4)$ quartets of leaves), a technique that is useful especially when quartet trees are easy to calculate but pairwise distances are not. Coalescent-based analyses have this property. We used anchoring to design DISTIQUE, a new statistically consistent summary method for species tree estimation.

DISTIQUE was not consistently better than other summary methods in our experiments, but it was always close to the best method, and was never outperformed by more than 5% of branches. When we used true gene trees, DISTIQUE tended to match or outperform other summary methods, but it was slightly more sensitive to gene tree estimation error than ASTRAL and ASTRID were. It is possible that these differences reflect the fact that DISTIQUE, unlike ASTRAL or ASTRID (i.e., NJst), uses the expected quartet probabilities according to the coalescence model, and as a result is more sensitive to errors in gene trees, which distort these probabilities. Using methods beyond simple empirical frequency [62] might further improve the accuracy of DISTIQUE.

Despite having strong competition in ASTRAL and ASTRID (and possibly other methods we did not test), we believe there are reasons to further develop DISTIQUE in various ways. The accuracies of double or single anchored versions of DISTIQUE with $O(n^2)$ to $O(n^4)$ need to be tested to see if they can achieve similar levels of accuracy as the current version. Moreover, the accuracy of DISTIQUE may improve using other techniques such as estimating quartet lengths using parametric methods, or weighting various quartets according to coalescent expectations. Finally, we may be able to use DISTIQUE to co-estimate gene trees and the species tree using distances.

References

[1] N Saitou and Masatoshi Nei. The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.

[2] Peter Erdos, Mike Steel, L Szekely, and T Warnow. A few logs suffice to build (almost) all trees: Part II. *Theoretical Computer Science*, 221(1-2):77–118, 1999.

[3] William J Bruno, Nicholas D Socci, and Aaron L Halpern. Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Molecular Biology and Evolution*, 17(1):189–197, 2000.

[4] Travis J Wheeler. Large-scale neighbor-joining with NINJA. In *Algorithms in Bioinformatics*, pages 375–389. Springer, 2009.

[5] Alexandros Stamatakis. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.

[6] Fredrik Ronquist, Maxim Teslenko, Paul van der Mark, Daniel L Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc A Suchard, and John P Huelsenbeck. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, 61(3):539–542, 2012.

[7] Morgan N. Price, P S Dehal, and Adam P. Arkin. FastTree-2 Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, 5(3):e9490, 2010.

[8] K. Katoh, K. Kuma, H. Toh, and T. Miyata. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Research*, 33(2):511–518, 2005.

[9] Robert C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004.

[10] T. Wheeler and J. Kececioglu. Multiple alignment by aligning alignments. In *Proceedings of the 15th ISCB Conference on Intelligent Systems for Molecular Biology*, pages 559–568, 2007.

[11] Liang Liu and Lili Yu. Estimating species trees from unrooted gene trees. *Systematic Biology*, 60:661–667, 2011.

[12] Pranjal Vachaspati and Tandy Warnow. ASTRID: Accurate Species TRees from Internode Distances. *BMC Genomics*, 16(Suppl 10):S3, 2015.

[13] Liang Liu, Lili Yu, Dennis K Pearl, and Scott V Edwards. Estimating species phylogenies using coalescence times among sequences. *Systematic Biology*, 58(5):468–477, 2009.

[14] Siavash Mirarab and Tandy Warnow. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, 31(12):i44–i52, 2015.

[15] Liang Liu, Lili Yu, and Scott V Edwards. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, 10(1):302, 2010.

[16] Bret R Larget, Satish K Kotha, Colin N Dewey, and Cecile Ané. BUCKy: Gene tree/species tree reconciliation with the Bayesian concordance analysis. *Bioinformatics*, 26(22):2910–2911, 2010.

[17] Elchanan Mossel and Sebastien Roch. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(1):166–171, 2010.

[18] Wayne P. Maddison and Stable Url. Gene Trees in Species Trees. *Systematic Biology*, 46(3):523, 1997.

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

[20] Siavash Mirarab, Rezwana Reaz, Md Shamsuzzoha Bayzid, Théo Zimmermann, M Shel Swenson, and Tandy Warnow. ASTRAL: Genome-Scale Coalescent-Based Species Tree. *Bioinformatics*, 30(17):i541–i548, 2014.

[21] Sen Song, Liang Liu, Scott V Edwards, and Shaoyuan Wu. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences*, 109(37):14942–7, 2012.

[22] Erich D Jarvis, Siavash Mirarab, Andre J Aberer, Bo Li, Peter Houde, Cai Li, Simon Y W Ho, Brant C. Faircloth, Benoit Nabholz, Jason T Howard, Alexander Suh, Claudia C Weber, Rute R da Fonseca, Jianwen Li, Fang Zhang, Hui Li, Long Zhou, Nitish Narula, Liang Liu, Ganeshkumar Ganapathy, Bastien Boussau, Md Shamsuzzoha Bayzid, Volodymyr Zavidovych, Sankar Subramanian, Toni Gabaldón, Salvador Capella-Gutiérrez, Jaime Huerta-Cepas, Bhanu Rekepalli, Kasper Munch, Mikkel H. Schierup, Bent Lindow, Wesley C Warren, David Ray, Richard E Green, Michael W Bruford, Xiangjiang Zhan, Andrew Dixon, Shengbin Li, Ning Li, Yinhua Huang, Elizabeth P Derryberry, Mads Frost Bertelsen, Frederick H Sheldon, Robb T. Brumfield, Claudio V Mello, Peter V Lovell, Morgan Wirthlin, Maria Paula Cruz Schneider, Francisco Prosdocimi, José Alfredo Samaniego, Amhed Missael Vargas Velazquez, Alonzo Alfaro-Núñez, Paula F Campos, Bent Petersen, Thomas Sicheritz-Ponten, An Pas, Tom Bailey, Paul Scofield, Michael Bunce, David M Lambert, Qi Zhou, Polina Perelman, Amy C. Driskell, Beth Shapiro, Zijun Xiong, Yongli Zeng, Shiping Liu, Zhenyu Li, Binghang Liu, Kui Wu, Jin Xiao, Xiong Yinqi, Qiuemei Zheng, Yong Zhang, Huanming Yang, Jian Wang, Linnea Smeds, Frank E Rheindt, Michael J Braun, Jon Fjeldsa, Ludovic Orlando, F Keith Barker, Knud Andreas Jø nsson, Warren Johnson, Klaus-Peter Koepfli, Stephen OBrien, David Haussler, Oliver A Ryder, Carsten Rahbek, Eske Willerslev, Gary R Graves, Travis C. Glenn, John E McCormack, Dave W Burt, Hans Ellegren, Per Alström, Scott V Edwards, Alexandros Stamatakis, David P Mindell, Joel Cracraft, Edward L Braun, Tandy Warnow, Wang Jun, M Thomas P Gilbert, and Guojie Zhang. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215):1320–1331, 2014.

[23] Norman J. Wickett, Siavash Mirarab, Nam Nguyen, Tandy Warnow, Eric Carpenter, Naim Matasci, Saravanaraj Ayyampalayam, Michael S. Barker, J. Gordon Burleigh, Matthew A. Gitzendanner, Brad R. Ruhfel, Eric Wafula, Joshua P. Der, Sean W. Graham, Sarah Mathews, Michael Melkonian, Douglas E. Soltis, Pamela S. Soltis, Nicholas W. Miles, Carl J. Rothfels, Lisa Pokorny, A. Jonathan Shaw, Lisa DeGironimo, Dennis W. Stevenson, Barbara Surek, Juan Carlos Villarreal, Béatrice Roure, Hervé Philippe, Claude W. dePamphilis, Tao Chen, Michael K. Deyholos, Regina S. Baucom, Toni M. Kutchan, Megan M. Augustin, Jun Wang, Yong Zhang, Zhijian Tian, Zhixiang Yan, Xiaolei Wu, Xiao Sun, Gane Ka-Shu Wong, and James Leebens-Mack. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences*, 111(45):E4859–E4868, 2014.

[24] Siavash Mirarab, Md Shamsuzzoha Bayzid, and Tandy Warnow. Evaluating summary methods for multi-locus species tree estimation in the presence of incomplete lineage sorting. *Systematic Biology*, page syu063, 2014.

[25] John Gatesy and Mark S Springer. Phylogenetic Analysis at Deep Timescales: Unreliable Gene Trees, Bypassed Hidden Support, and the Coalescence/Concatalescence Conundrum. *Molecular Phylogenetics and Evolution*, 80:231–266, 2014.

[26] Mark S. Springer and John Gatesy. The gene tree delusion. *Molecular Phylogenetics and Evolution*, 2015.

[27] Siavash Mirarab, Md Shamsuzzoha Bayzid, Bastien Boussau, and Tandy Warnow. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science*, 346(6215), 2014.

[28] Swati Patel, Rebecca T Kimball, and Edward L Braun. Error in phylogenetic estimation for bushes in the tree of life. *Journal of Phylogenetics and Evolutionary Biology*, 1(2):110, 2013.

[29] K Strimmer and a von Haeseler. Quartet puzzling - a quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.*, 13:964–969, 1996.

[30] Sagi Snir and Satish Rao. Quartets MaxCut: A divide and conquer quartets algorithm. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(4):704–718, 2010.

[31] O R P Bininda-Emonds, editor. *Phylogenetic Supertrees: combining information to reveal the tree of life*, volume 4. Kluwer Academic Publishers, 2004.

[32] Eliran Avni, Reuven Cohen, and Sagi Snir. Weighted Quartets Phylogenetics. *Systematic Biology*, 64(2):233–242, 2015.

[33] David Bryant and Michael Steel. Constructing Optimal Trees from Quartets. *Journal of Algorithms*, 38:237–259, 2001.

[34] Tao Jiang, Paul Kearney, and Ming Li. A Polynomial Time Approximation Scheme for Inferring Evolutionary Trees from Quartet Topologies and Its Application, 2001.

[35] Raul Piaggio-Talice, J.Gordon Burleigh, and Oliver Eulenstein. Quartet Supertrees. In OlafR.P. Bininda-Emonds, editor, *Phylogenetic Supertrees SE - 9*, volume 4 of *Computational Biology*, pages 173–191. Springer Netherlands, 2004.

[36] Sebastien Roch and Sagi Snir. Recovering the Treelike Trend of Evolution Despite Extensive Lateral Genetic Transfer: A Probabilistic Analysis. *Journal of Computational Biology*, 20(2):93–112, 2013.

[37] Noah WM Stenz, Bret Larget, David A Baum, and Cécile Ané. Exploring tree-like and non-tree-like patterns using genome sequences: An example using the inbreeding plant species arabidopsis thaliana (l.) heynh. *Systematic biology*, 64(5):809–823, 2015.

[38] Sebastien Roch and Tandy Warnow. On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. *Systematic Biology*, page syv016, 2015.

[39] E S Allman, James Degnan, and J A Rhodes. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *Journal of Mathematical biology*, 62:833–862, 2011.

[40] J Degnan. Anomalous unrooted gene trees. *Systematic Biology*, 62:574–590, 2013.

[41] Julia Chifman and Laura S Kubatko. Quartet Inference from SNP Data Under the Coalescent Model. *Bioinformatics*, 30(23):3317–3324, 2014.

[42] Noah A. Rosenberg. Discordance of species trees with their most likely gene trees: a unifying principle. *Molecular Biology and Evolution*, 30(12):2709–2713, 2013.

[43] Peter Buneman. A note on the metric properties of trees. *Journal of Combinatorial Theory, Series B*, 17(1):48–50, 1974.

[44] Y Benjamini and Y Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57:289–300, 1995.

[45] Michael Steel. Recovering a tree from the leaf colourations it generates under a Markov model, 1994.

[46] M Shel Swenson, Rahul Suri, C Randal Linder, and Tandy Warnow. SuperFine: fast and accurate supertree estimation. *Systematic biology*, 61(2):214–27, 2012.

[47] Nam Nguyen, Siavash Mirarab, and Tandy Warnow. MRL and SuperFine+MRL: new supertree methods. *Algorithms for Molecular Biology*, 7(1), 2012.

[48] Isaac Elias and Jens Lagergren. Fast neighbor joining. *Theoretical Computer Science*, 410:1993–2000, 2009.

[49] James Degnan and Noah A Rosenberg. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology and Evolution*, 24(6):332–340, 2009.

[50] DF Robinson and LR Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2):131–147, 1981.

[51] Jed Chou, Ashu Gupta, Shashank Yaduvanshi, Ruth Davidson, Mike Nute, Siavash Mirarab, and Tandy Warnow. A comparative study of SVDquartets and other coalescent-based species tree estimation methods. *BMC Genomics*, 16(Suppl 10):S2, 2015.

[52] Mark P Simmons and John Gatesy. Coalescence vs. concatenation: sophisticated analyses vs. first principles applied to rooting the angiosperms. *Molecular Phylogenetics and Evolution*, 2015.

[53] Raphail E Krichevsky and Victor K Trofimov. The performance of universal encoding. *Information Theory, IEEE Transactions on*, 27(2):199–207, 1981.

[54] James H. Degnan, Michael DeGiorgio, David Bryant, and Noah A. Rosenberg. Properties of Consensus Methods for Inferring Species Trees from Gene Trees. *Systematic Biology*, 58(1):35–54, February 2009.

[55] J Sukumaran and MT Holder. Dendropy: a Python library for phylogenetic computing. *Bioinformatics*, 26(12):1569–71, 2010.

[56] Andreas Sand, Morten K Holt, Jens Johansen, Rolf Fagerberg, Gerth Stølting Brodal, Christian NS Pedersen, and Thomas Mailund. Algorithms for computing the triplet and quartet distances for binary and general trees. *Biology*, 2(4):1189–209, 2013.

[57] Vincent Lefort, Richard Desper, and Olivier Gascuel. FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program: Table 1. *Molecular Biology and Evolution*, 32(10):2798–2800, 2015.

[58] Alexis Criscuolo and Olivier Gascuel. Fast nj-like algorithms to deal with incomplete distance matrices. *BMC bioinformatics*, 9(1):166, 2008.

[59] Md Shamsuzzoha Bayzid, Siavash Mirarab, Bastien Boussau, and Tandy Warnow. Weighted Statistical Binning: Enabling Statistically Consistent Genome-Scale Phylogenetic Analyses. *PLoS ONE*, 10(6):e0129183, 2015.

[60] Diego Mallo, Leonardo de Oliveira Martins, and David Posada. SimPhy: Phylogenomic Simulation of Gene, Locus and Species Trees. *bioRxiv*, 2015.

[61] Tae Kun Seo. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. *Molecular Biology and Evolution*, 25(5):960–971, 2008.

[62] Sudeep Kamath, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In *Proceedings of The 28th Conference on Learning Theory*, pages 1066–1100, 2015.

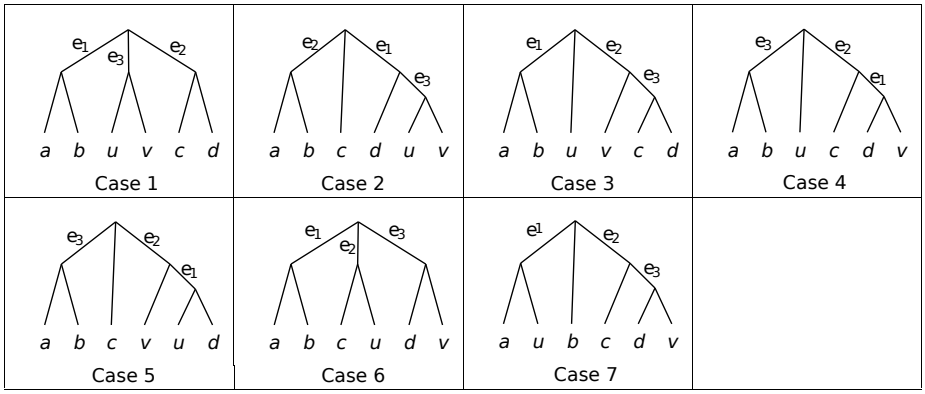

Figure 5: Placements of two anchors u and v on a given quartet topology $ab.cd$.

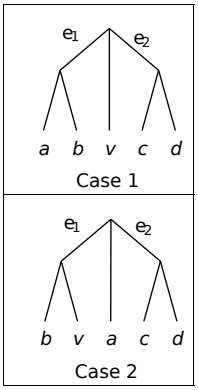

Figure 6: Placing anchor v on quartet tree $ab.cd$.

Appendices

A Theorem proofs

Proof of Theorem 1: We show that for arbitrary $\{a, b, c, d\} \subset \mathcal{L}$, the four point conditions holds for D'_{uv} and that the conditions are compatible with T. Thus, assuming $ab.cd \in \mathcal{Q}^T$, we show

$$D'_{uv}[a, b] + D'_{uv}[c, d] < D'_{uv}[a, c] + D'_{uv}[b, d] = D'_{uv}[a, d] + D'_{uv}[b, c] \tag{6}$$

Figure 5 shows all ways of placing anchors (u and v) on the quartet tree $ab.cd$. Anchors u and v can be sisters, in which case they can be placed on the internal branch (Case 1) or on one of the tip branches (Case 2; w.l.o.g, we use the branch leading to d). When u and v are not sisters, they can be both placed on the internal branch (Case 3), or one on the internal branch and the other on a tip branch (Case 4), or they can be both on terminal branches, which can be done in three ways: u and v can be on the same terminal branch (Case 5), on different but adjacent branches (Case 6), or on two non-adjacent branches (Case 7).

In Table 2, for each of the seven cases, we compute all parts of Equation 6; $LHS = D'_{uv}[a, b] + D'_{uv}[c, d]$, $RHS1 = D'_{uv}[a, d] + D'_{uv}[b, c]$, and $RHS2 = D'_{uv}[a, c] + D'_{uv}[b, d]$. Computation of values in Table 2 is straight-forward; each $D'_{uv}[x, y]$ should be compute using Equation 1; thus, where $xy.uv$ is induced by the tree shown in Figure 5, we use $[\beta - f(d)]$ and otherwise we use $[\beta + \alpha d]$, where $d = d_D(x, y, u, v)$ is the quartet length (length of the internal branch) for the quartet tree induced by $\{x, y, u, v\}$. For example, for Case 1, $D'_{uv}[a, b]$ is $[\beta - f(e_1 + e_3)]$ because $ab.uv$ is induced by the tree, and the distance of the edge on the $ab.uv$ quartet is $e_1 + e_3$; in Case 7, we set $D'_{uv}[a, b]$ to $\beta + \alpha e_1$ because $ab.uv$ is *not* induced by the tree and $d_D(a, b, u, v) = e_1$.

Table 2: Four point condition for all 7 cases shown in Figure 5. For each case, $LHS < RHS1$ and $RHS1 = RHS2$.

	LHD $D'_{uv}[a, b] + D'_{uv}[c, d]$	RHS1 $D'_{uv}[a, d] + D'_{uv}[b, c]$	RHS2 $D'_{uv}[a, c] + D'_{uv}[b, d]$
Case 1	$[\beta - f(e_1 + e_3)] + [\beta - f(e_2 + e_3)]$	$[\beta - f(e_3)] + [\beta - f(e_3)]$	$[\beta - f(e_3)] + [\beta - f(e_3)]$
Case 2	$[\beta - f(e_1 + e_2 + e_3)] + [\beta - f(e_3)]$	$[\beta - f(e_3)] + [\beta - f(e_1 + e_3)]$	$[\beta - f(e_1 + e_3)] + [\beta - f(e_3)]$
Case 3	$[\beta - f(e_1)] + [\beta - f(e_3)]$	$[\beta + \alpha e_2] + [\beta + \alpha e_2]$	$[\beta + \alpha e_2] + [\beta + \alpha e_2]$
Case 4	$[\beta - f(e_3)] + [\beta + \alpha e_1]$	$[\beta + \alpha(e_1 + e_2)] + [\beta + \alpha e_2]$	$[\beta + \alpha e_2] + [\beta + \alpha(e_1 + e_2)]$
Case 5	$[\beta - f(e_2 + e_3)] + [\beta + \alpha e_1]$	$[\beta - f(e_2)] + [\beta + \alpha e_1]$	$[\beta - f(e_2)] + [\beta + \alpha e_1]$
Case 6	$[\beta - f(e_1)] + [\beta + \alpha(e_2 + e_3)]$	$[\beta + \alpha e_3] + [\beta + \alpha e_2]$	$[\beta + \alpha e_2] + [\beta + \alpha e_3]$
Case 7	$[\beta + \alpha e_1] + [\beta + \alpha e_3]$	$[\beta + \alpha(e_1 + e_2 + e_3)] + [\beta + \alpha e_2]$	$[\beta + \alpha(e_1 + e_2)] + [\beta + \alpha(e_2 + e_3)]$

We need to show that $LHS < RHS1$ and that $RHS1 = RHS2$. We remind the reader that all branches are assumed to be strictly positive, and that f is a positive and monotonically increasing function bounded from above by β. In all cases, the equality of $RHS1$ and $RHS2$ is immediately clear from Table 2. The inequality ($LHS < RHS1$) follows directly from the fact that $f(x)$ is monotonically increasing in Cases 1, 2, and 5. For Case 3, because of positivity of branch lengths and $f(x)$, we have $LHS < 2\beta < RHS$. Similarly, for Case 4, $LHS < 2\beta+\alpha e_1 < 2\beta+\alpha e_1+2\alpha e_2 = RHS$. Case 6 follows from the positivity of f, and Case 7 is trivially correct for positive branch lengths. □

Proof of Theorem 2: For distances defined using Equation 2, we show that the four point condition holds for an arbitrary $\{a, b, c, d\} \subset \mathcal{L}$, assuming w.l.o.g that $ab.cd \in \mathcal{Q}^T$:

$$\sum_{u\notin\{a,b\}} D'_{uv}[a,b] + \sum_{u\notin\{c,d\}} D'_{uv}[c,d] < \sum_{u\notin\{a,d\}} D'_{uv}[a,d] + \sum_{u\notin\{a,b\}} D'_{uv}[b,c] = \sum_{u\notin\{a,c\}} D'_{uv}[a,c] + \sum_{u\notin\{b,d\}} D'_{uv}[b,d]$$

Let $\mathcal{L}' = \mathcal{L} - \{a, b, c, d\}$. Each sum can be broken into cases where $u \in \{a, b, c, d\}$ or $u \in \mathcal{L}'$:

$$D'_{cv}[a,b] + D'_{dv}[a,b] + D'_{av}[c,d] + D'_{bv}[c,d] + \sum_{u\in\mathcal{L}'} D'_{uv}[a,b] + D'_{uv}[c,d] <$$
$$D'_{bv}[a,d] + D'_{cv}[a,d] + D'_{av}[b,c] + D'_{dv}[b,c] + \sum_{u\in\mathcal{L}'} D'_{uv}[a,d] + D'_{uv}[b,c] =$$
$$D'_{bv}[a,c] + D'_{dv}[a,c] + D'_{av}[b,d] + D'_{cv}[b,d] + \sum_{u\in\mathcal{L}'} D'_{uv}[a,c] + D'_{uv}[b,d]$$

For terms where $u \in \mathcal{L}'$, the additivity is proved in Theorem 1. Thus, we need to prove inequalities and qualities of terms where $u \in \{a, b, c, d\}$. The single anchor v can be placed either on the internal branch (Case 1) or on a terminal branches (Case 2), as Figure 6 shows. In Case 1, for each term of the sum, we have:

$$LHS = [\beta-f(e_1)]+[\beta-f(e_1)]+[\beta-f(e_2)]+[\beta-f(e_2)] < 4\beta < [\beta+\alpha e_1]+[\beta+\alpha e_2]+[\beta+\alpha e_1]+[\beta+\alpha e_2] = RHS$$

and for the second case we have:

$$\begin{aligned} LHS = [\beta+\alpha e_1] + [\beta+\alpha e_1] + [\beta - f(e_2)] + [\beta - f(e_1+e_2)] &< 4\beta + 2\alpha e_1 - f(e_1+e_2) < \\ [\beta+\alpha e_2] + [\beta - f(e_1)] + [\beta+\alpha e_1] + [\beta+\alpha(e_1+e_2)] &= RHS1 = RHS2 \end{aligned}$$

Since addition of additive matrices are additive, the proof follows. □

Proof of Theorem 3: Similar to proof of Theorem 2, terms in Equation 3 can be of three types: $\{u, v\}$ and $\{a, b, c, d\}$ intersect in either (I) two elements, (II) one element, or (III) none. Thus, the four point condition can be written out as ($\mathcal{L}' = \mathcal{L} - \{a, b, c, d\}$):

$$2D'_{ab}[c,d] + \sum_{v\in\mathcal{L}'} D'_{cv}[a,b] + D'_{dv}[a,b] + D'_{av}[c,d] + D'_{bv}[c,d] + \sum_{u,v\in\mathcal{L}'} D'_{uv}[a,b] + D'_{uv}[c,d] <$$
$$2D'_{ad}[b,c] + \sum_{v\in\mathcal{L}'} D'_{cv}[a,d] + D'_{bv}[a,d] + D'_{av}[b,c] + D'_{dv}[b,c] + \sum_{u,v\in\mathcal{L}'} D'_{uv}[a,d] + D'_{uv}[b,c] =$$
$$2D'_{ac}[b,d] + \sum_{v\in\mathcal{L}'} D'_{bv}[a,c] + D'_{dv}[a,c] + D'_{av}[b,d] + D'_{cv}[b,d] + \sum_{u,v\in\mathcal{L}'} D'_{uv}[a,c] + D'_{uv}[b,d]$$

For terms of type (III) and (II), the additivity is proved in Theorems 1 and 2, respectively. Thus, we need to prove additivity only for terms of type (I), which have no anchors; we have:

$$LHS = 2D'_{ab}[c,d] = 2[\beta-f(d)] < 2\beta < 2[\beta+\alpha d] = 2D'_{ad}[b,c] = 2D'_{ac}[b,d] = RHS;\ d = d_D(a,b,c,d)$$

Thus, for all three types, inequalities and qualities needed to prove additivity hold. □

Anchored distances for quartet-based estimation of phylogenetic trees and applications to coalescent-based analyses (Supplementary Matrial)

Erfan Sayyari[1] and Siavash Mirarab*[1]

[1]University of California, San Diego, Department of Electrical and Computer Engineering

Contents

List of Figures

List of Tables

*Corresponding author: smirarab@ucsd.edu

1 Supplementary Figures and Tables

1.1 Supplementary Tables

Table S1: **Empirical statistics of simulated Avian and Mammalian datasets.** Model condition $2X$ corresponds to the case where ILS is reduced by increasing the branch lengths (2 times longer), and $0.5X$ represents the case where ILS is increased by reducing the branch lengths (2 times shorter). In the same way, the model condition with $0.2X$ corresponds to the case where ILS is reduced by dividing the branch lengths by five. Average Robinso-Foulds (RF) distances between true gene trees and the model species tree are provided in *AD to species tree*. *# gene trees* shows number of gene trees that are available for the corresponding dataset and ILS. *#base pairs* represents number of base pairs, and *# replicates* shows number of replicates for the corresponding dataset and ILS. In column *Ref.*, the reference paper for each dataset is provided. For the Mammalian with ILS level $0.2X$, *# replicates* 5 and 10 are for the model conditions where *# gene trees* is 3200, and 1600 respectively. Also for the Avian dataset with ILS level $1X$, *# replicates* 10 is only for the model condition with *# gene trees* 2000.

	ILS	AD to species tree	# gene trees	# base pairs	# replicates	Ref.
Mammalian	$2X$	18%	200	500,true	20	[3]
	$1X$	32%	200	500,true	20	[3]
	$0.5X$	54%	200	500,true	20	[3]
	$0.2X$	79%	$100, 200, 400, 800, 1600, 3200$	500,true	$5, 10, 20$	[3]
Avian	$2X$	35%	1000	500,true	20	[4]
	$1X$	47%	$200, 500, 1000, 2000$	500,true	$10, 20$	[4]
	$0.5X$	59%	1000	500,true	20	[4]

Table S2: **Empirical statistics of simulated 11-taxon dataset [2]**. Model condition M1 corresponds to the very low ILS, model condition M2 corresponds to low ILS, model condition M3 shows high ILS, and model condition M4 for very high ILS. *AD* represents average bipartition distance between true gene trees and true species trees, expressed as a percentage. The rest of columns are the same as Table S1

dataset	AD	# base pairs	# gene trees	# replicates	Reference
11-taxon M1	15.5%	$10, 25, 50, 100, 200$	$100, 500, 1000$	50	[2]
11-taxon M2	38.3%	$10, 25, 50, 100, 200$	$100, 500, 1000$	50	[5]
11-taxon M3	66.3%	$10, 25, 50, 100, 200$	$100, 500, 1000$	50	[2]
11-taxon M4	85.0%	$10, 25, 50, 100, 200$	$100, 500, 1000$	50	[5]

1.2 Supplementary Figures

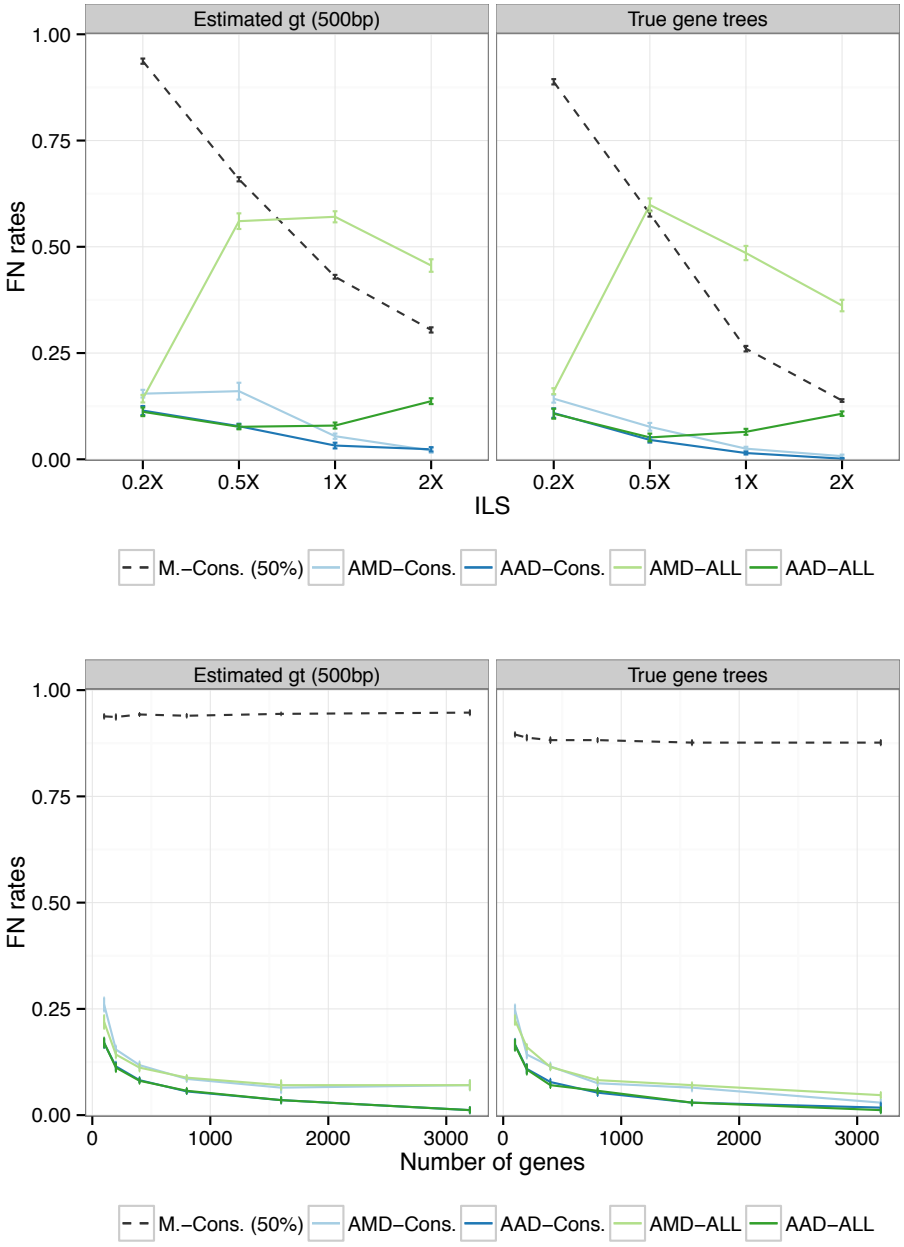

Figure S1: **Accuracy of different implementations of DISTIQUE and Majority Consensus method (50%) for Mammalian dataset.** This figure compares four versions of DISTIUQE on the Mammalian dataset as we vary the amount of ILS, and also shows the missing branch rate for the majority consensus tree. (top) number of genes: 200; (below) ILS: 0.2X. Mean and standard error of species tree error is shown for true and estimated gene trees (500bp alignments). With very high ILS (0.2X), the accuracy for all of the implementations of DISTIQUE are close. With high ILS (0.5X), AAD-Cons, and AAD-ALL have similar accuracy, and AMD-Cons is the next best one. As ILS decreases, when DISTIQUE is applied to the entire dataset, the error goes up, which is more pronounced for AMD-ALL. The results of AMD-ALL for true gene trees is worse than simple Majority Consensus (50%). As discussed before, we attribute this pattern to difficulties of estimating long quartet lengths. When DISTIQUE is used to resolve polytomies in the consensus tree, the accuracy improves with decreased ILS, as expected. Note that even with reduced ILS, the consensus tree on estimated gene trees misses more than 25% of branches, and leaves some polytomies for DISTIQUE to resolve.

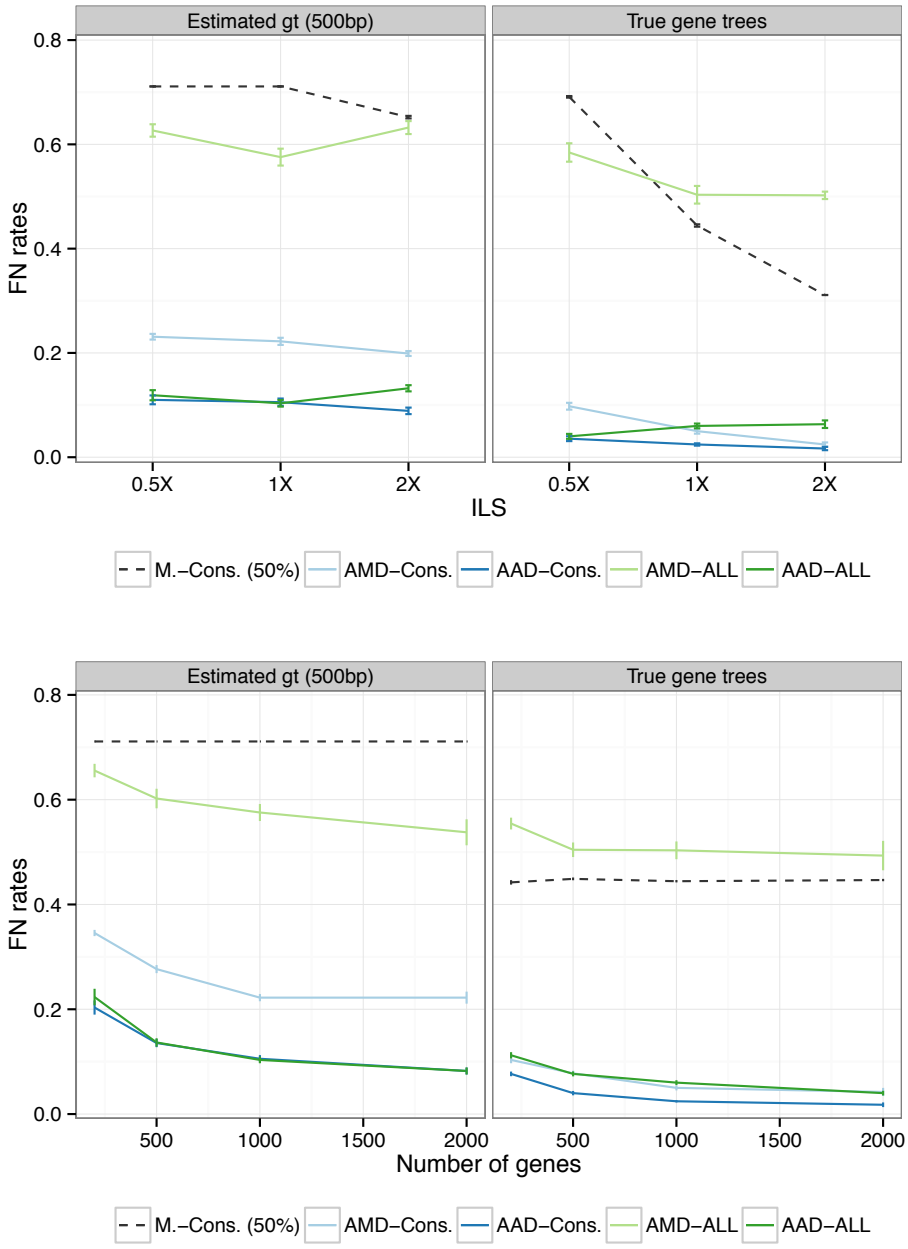

Figure S2: **Accuracy of different implementations of DISTIQUE and Majority Consensus method (50%) for Avian dataset.** This figure compares four versions of DISTIUQE on the Avian dataset as we vary the amount of ILS, and also shows the missing branch rate for the majority consensus tree. (top) number of genes: 1000; (below) ILS: 1X. Mean and standard error of species tree error is shown for true and estimated gene trees (500bp alignments). With high ILS (0.5X), AAD-Cons, and AAD-ALL have similar accuracy, and AMD-Cons is the next best one. As ILS decreases, when DISTIQUE is applied to the entire dataset, the error goes up. The results of AMD-ALL for true gene trees is worse than simple Majority Consensus (50%). As discussed before, we attribute this pattern to difficulties of estimating long quartet lengths. When DISTIQUE is used to resolve polytomies in the consensus tree, the accuracy improves with decreased ILS, as expected. Note that even with reduced ILS, the consensus tree misses more than 30% of branches, and leaves some polytomies for DISTIQUE to resolve.

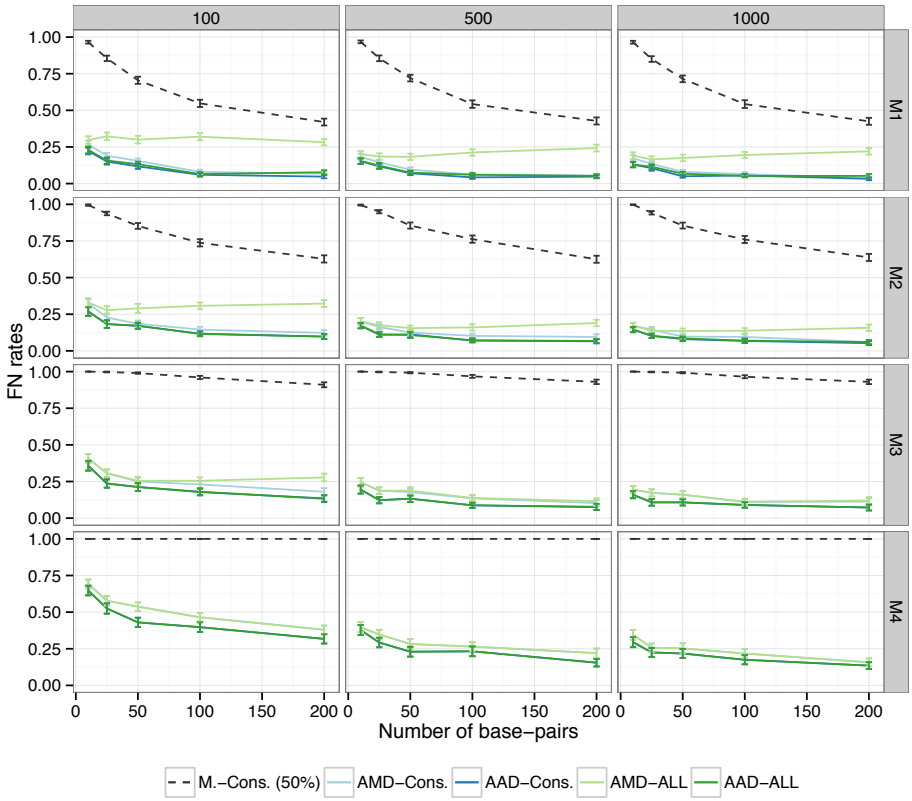

Figure S3: **Accuracy of different implementations of DISTIQUE and Majority Consensus method (50%) for 11-taxon dataset.** This figure compares four versions of DISTIUQE on the 11-taxon dataset as we vary the amount of ILS (rows), and number of genes (columns), and also shows the missing branch rate for the Majority Consensus tree. Mean and standard error of species tree error is shown for estimated gene trees (with number of base pairs varying from 10 to 200). With low and very low ILS (M1 and M2), as number of base pairs increases the missing branch rate of Majority Consensus goes down, but missing branch rate is always above 40%. In high and very high ILS (M3 and M4) Majority Consensus has missing branch rate above 75%, which shows there are big polytomies.

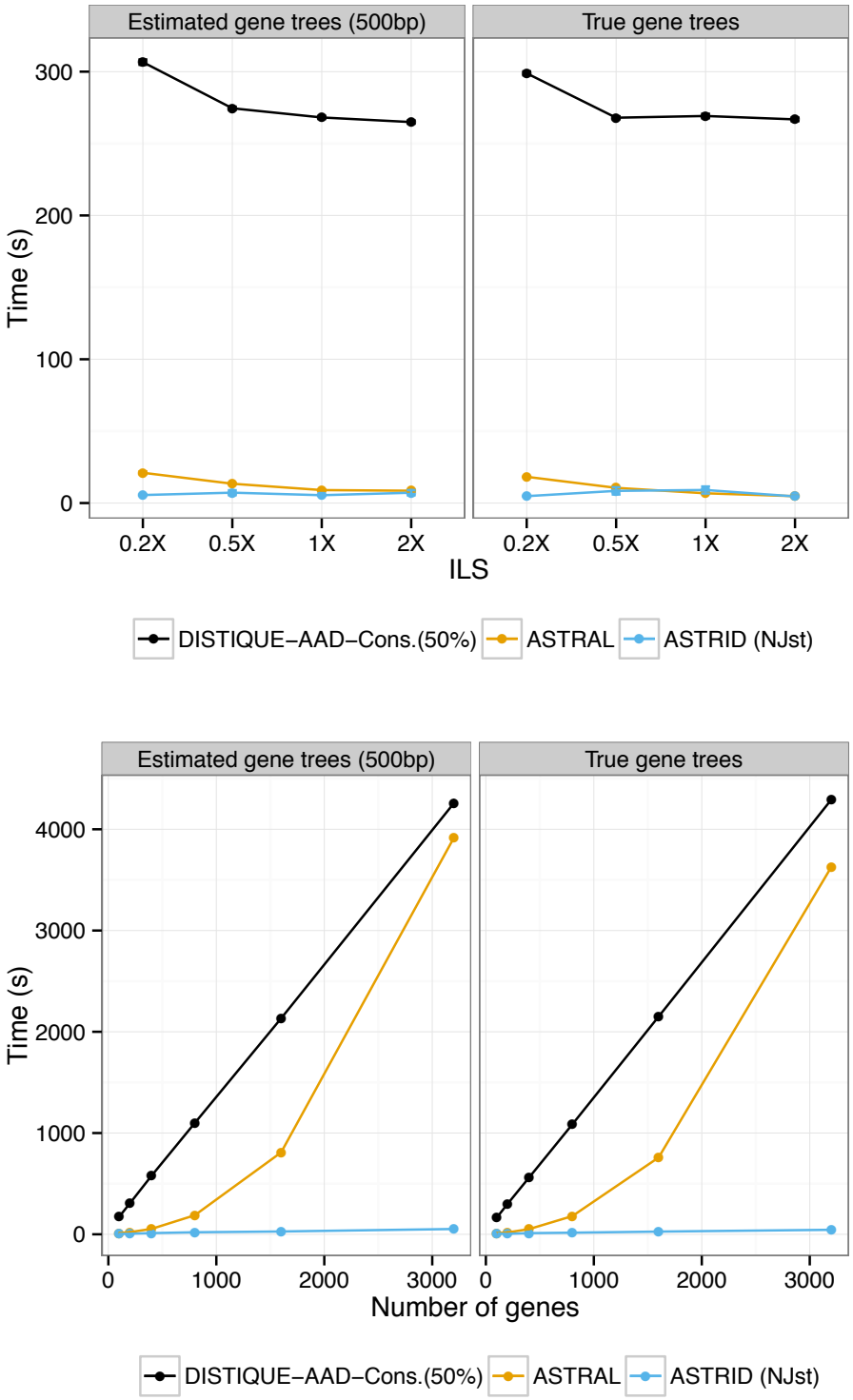

Figure S4: **Running times of DISTIQUE versus other methods for Mammalian dataset.** (top) number of genes: 200; (below) ILS: 0.2X, with 500bp alignments. The time complexity of DISTIQUE and ASTRID (NJst) are almost linear with respect to number of genes, while ASTRAL running time increases with respect to number of genes super-linearly.

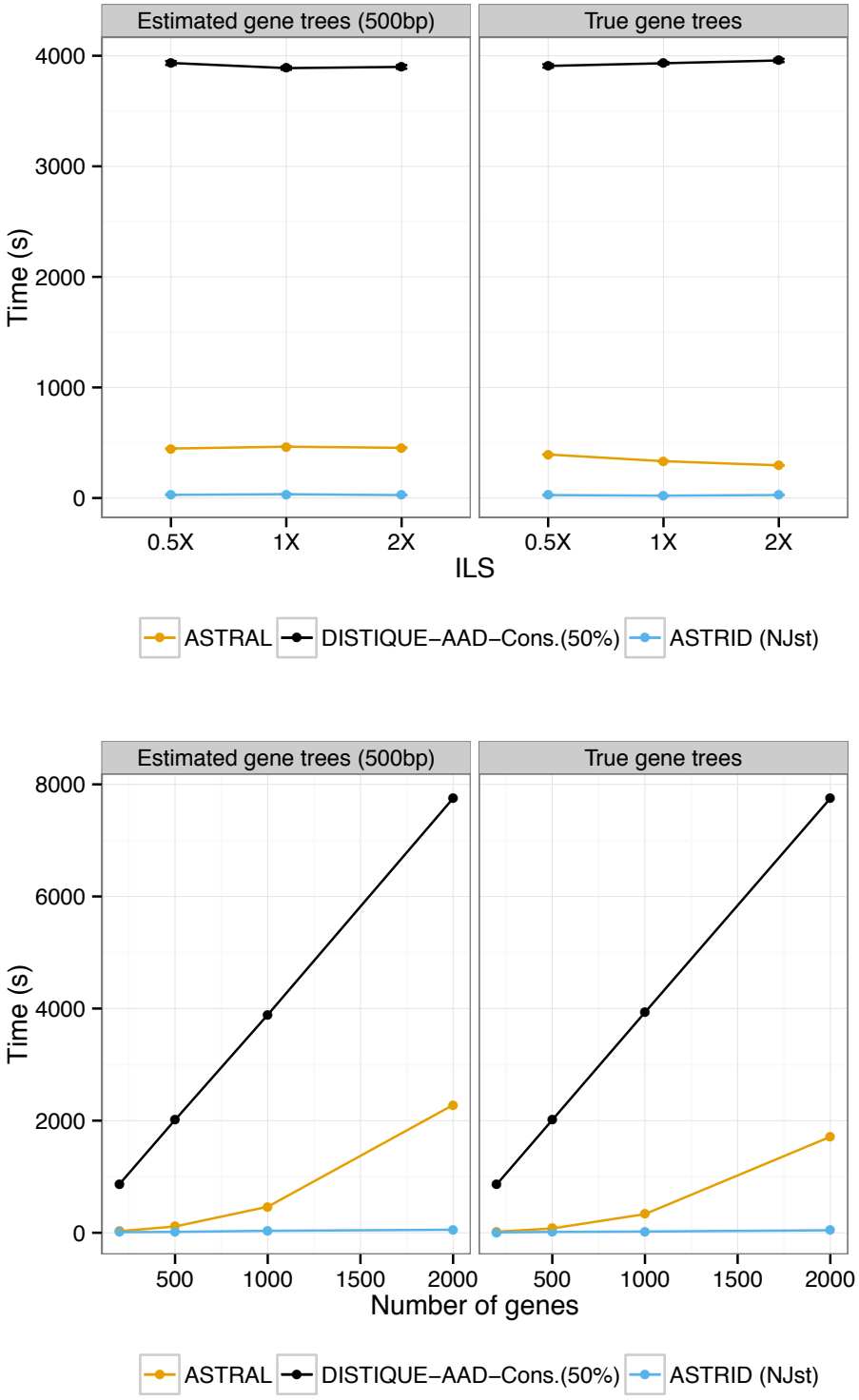

Figure S5: **Running times of DISTIQUE versus other methods for Avian dataset.**. (top) number of genes: 1000; (below) ILS: 1X, with 500bp alignments. This figure shows that the running time of none of the methods depends on ILS level for 1000 gene trees. The running time of DISTIQUE and ASTRID (NJst) are almost linear with respect to number of genes, while ASTRAL running time increases with respect to number of genes super-linearly.

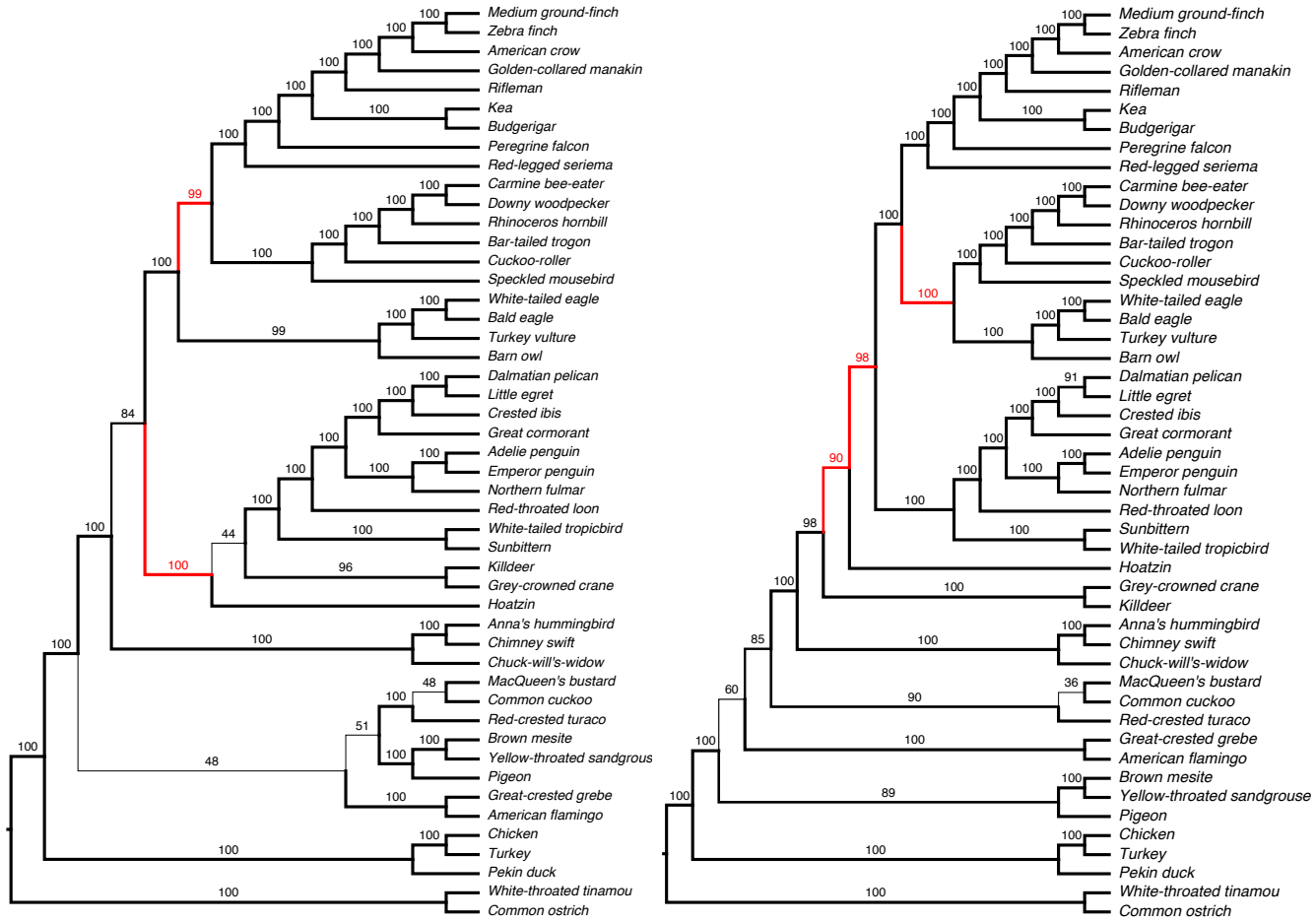

Figure S6: **Species trees generated based on AAD-Cons and ASTRAL using Avian biological dataset [1]**. (left) ASTRAL, (right) AAD-Cons differed in 5 branches, and two of these in DISTIQUE and three of them in ASTRAL had high support, which are represented with red. The presence of highly supported conflict between various coalescent-based methods indicates a need for skeptical and careful analysis of results of any single method on real data.

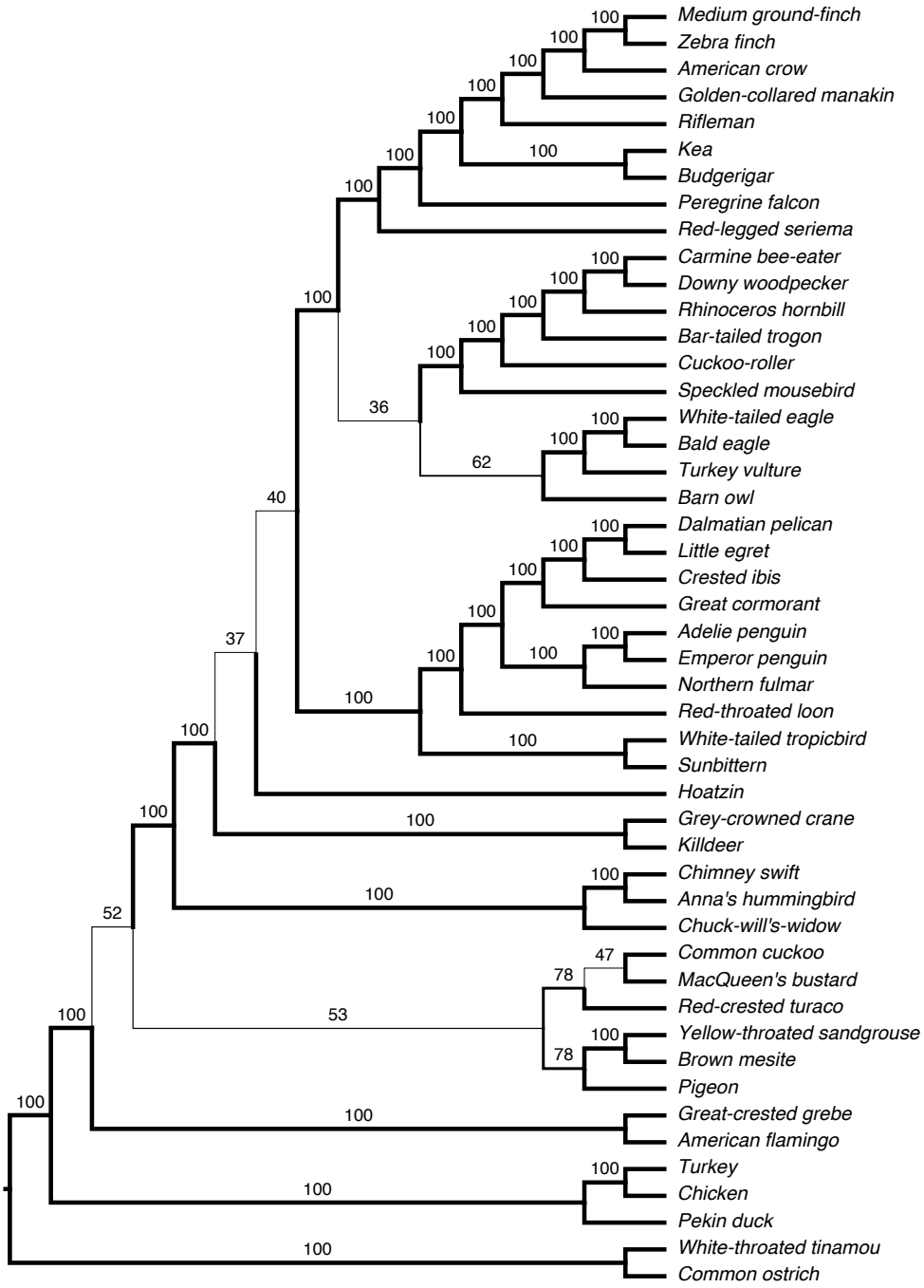

Figure S7: **Species trees generated based on ASTRID (NJst) using Avian biological dataset** DISTIQUE and ASTRID differed on two branches, which both had low support in ASTRID.

2 Difficulties with long branches and need for majority consensus

In this section, we will explain a problematic case that leads us to combine AAD with majority consensus. Figure S8 shows an unrooted species tree, with many long branches, with length L. We assume L is long enough that for our given number of gene trees, with high probability, all gene trees will be topologically identical to the species tree. Note that for any number of genes, there are branches long enough where discordance is highly unlikely (we give one example below). Thus, we assume that for branches of length L, we have zero quartet trees that conflict with them.

In our example, the distance of a to c is L and a to b is $3L$ (note that in our analyses, we only calculate branch length for internal branches). We will show that AAD can be mislead to give a smaller distance for a to b than a to c.

Recall that distances are defined as:

$$D'[a,b] = \sum_{v \in \mathcal{L}-\{a,b\}} D'_v[a,b] = \sum_{u,v \in \mathcal{L}-\{a,b\}} D'_{uv}[a,b] \tag{1}$$

where

$$D'_{uv}[a,b] = \begin{cases} \beta + \alpha.d_D(a,b,u,v) & ab.uv \notin \mathcal{Q}^T \\ \beta - f(d_D(a,b,u,v)) & ab.uv \in \mathcal{Q}^T \end{cases} \tag{2}$$

Also recall that for coalescent-based analyses, $\beta = \ln 3$, $\alpha = 1$, $f(x) = \ln(3 - 2e^{-x})$, and with our use of add-half smoothing, Equation 2 simplifies to

$$D'_{uv}[a,b] = -\ln p(ab.uv) = -\ln\left(\frac{freq(ab.uv) + 0.5}{n + 1.5}\right) \tag{3}$$

where n is the number of genes, and $freq(ab.uv)$ is the number of genes with induced quartet topology $ab.uv$. Using Equations 1 and 3 and considering all selections of anchors u and v, we have

$$\begin{aligned} D'[a,b] =& -\binom{3}{1}.\binom{|X|+1}{1}.\ln\left(\frac{0.5}{n+1.5}\right) - \binom{|X|+1}{2}\ln\left(\frac{n+0.5}{n+1.5}\right) \quad -\binom{3}{2}\ln\left(\frac{0.5}{n+1.5}\right) \\ \approx& -3(|X|+2)\ln\left(\frac{0.5}{n+1.5}\right) \end{aligned} \tag{4}$$

The first term comes from choosing one anchor from $\{u_1, u_2, u_3\}$, and choosing the other anchors from $\{c\} \cup X$. In these cases, a and b are further from each other than the anchors, and because of our assumption about L, the frequency of $ab.uv$ is expected to be zero. The second term comes from choosing both anchors from $\{c\} \cup X$; for these, the frequency of $ab.uv$ is expected to be n. Finally, the last term comes from choosing both anchors from the set $\{u_1, u_2, u_3\}$, where once again, the expected frequency of $ab.uv$ is zero for long enough L, leading to the use of the pseudo count. For large enough n, we can approximate $\ln\left(\frac{n+0.5}{n+1.5}\right) \approx 0$.

The same thing could be written for a and c:

$$\begin{aligned} D'[a,c] =& -\binom{4}{1}.\binom{|X|}{1}\ln\left(\frac{0.5}{n+1.5}\right) - \binom{|X|}{2}\ln\left(\frac{n+0.5}{n+1.5}\right) \quad -\binom{4}{2}\ln\left(\frac{n+0.5}{n+1.5}\right) \\ \approx& -4(|X|)\ln\left(\frac{0.5}{n+1.5}\right) \end{aligned} \tag{5}$$

The first term comes from choosing an anchor from of the $\{u_1, u_2, u_3, b\}$, and the other anchor from X; here, for long enough L, frequency of $ab.uv$ would be expected to be zero. The second term comes from choosing both anchors from X, and the last term comes from choosing both anchors from $\{u_1, u_2, u_3, b\}$; in these cases we expect the frequency of $ab.uv$ to be n.

Comparing the Equations 4 and 5, for $|X| > 2$, the distance between a and b would be smaller than the distance between a and c. This clearly is in contradiction to our tree, so AAD in this case becomes misleading. However, note that all branches with length L are assumed to generate no discordance, and thus will be in the final tree.

Large L: Assume that $L = 16$ (in coalescent units) and we have 1000 gene trees. The probability of having only topologies that agree with the species tree in all 1000 gene trees is $(1-2/3e^{-16})^{1000} = 0.99993$. Thus we expect all n gene trees to have that topology with very high probability. The probability of having only the species tree topology for branches of length $2L$ and $3L$ is even larger.

2.1 Computing pseudo-counts for AMD

In AMD and DISTIQUE-AMD method, in case of zero frequencies for some quartet topologies, without changing the definition of pseudo-count, the relative information about quartets would be lost. For example, assume we have topology $ab.cz$ with long internal branch length, like 16 as mentioned, and $ab.dz$ with internal branch length of 20 (longer than previous length), and no sample for none of the topologies that contradict the species tree. In this case the distance between a, and c is equal to distance of a, and d which is $\ln \frac{0.5}{n+1.5}$, where n is the number of samples. So the relative information is lost. In order to avoid this problem, the definition of pseudo count in AMD and DISTIQUE-AMD is slightly changed. In order to have a pseudo count that could capture the relative distances, first the number of zero quartet topologies are counted. This is called n^0_{ab}. The pseudo count in this case is defined as:

$$\ln \prod_{i=1}^{n^0_{ab}} \frac{0.5}{k_i + 1.5} \tag{6}$$

Where 0.5 comes from our add half estimator, is probability of zero frequencies, and k_i is the number of samples for quartet topology of i.

3 Commands and version numbers

We used ASTRAL version 4.7.8 to find the species trees from gene trees:

```
java -Xmx2000M -jar astral.4.7.8.jar -i [GENE TREES] -o [OUTPUT SPECIES TREE]
```

The ASTRID (NJst) results were produced using the following command:

```
python ASTRID.py -i [GENE TREES] -m [fastme2] -o [OUTPUT SPECIES TREE] -c
[CACHE]
```

CACHE is the distance matrix produced by ASTRID.

For DISTIQUE-AAD we used the following command:

```
python distique-2.py -a [mean] -g [GENE TREES] -m [prod] -o [OUTPUT DIRECTORY]
```

For DISTIQUE-AMD we used the following command:

```
python distique-2.py -a [mean] -g [GENE TREES] -m [min] -o [OUTPUT DIRECTORY]
```

Here the flag a specifies which averaging method to use the partial quartet tables from complete quartet tables around each polytomy.

For comparison and computing false negative missing branchs we used the command available at `https://github.com/smirarab/global/tree/master/src/shell`:

```
compareTrees.missingBranch [-s SPECIES TREE] [-g ESTIMATED SPECIES TREE ]
```

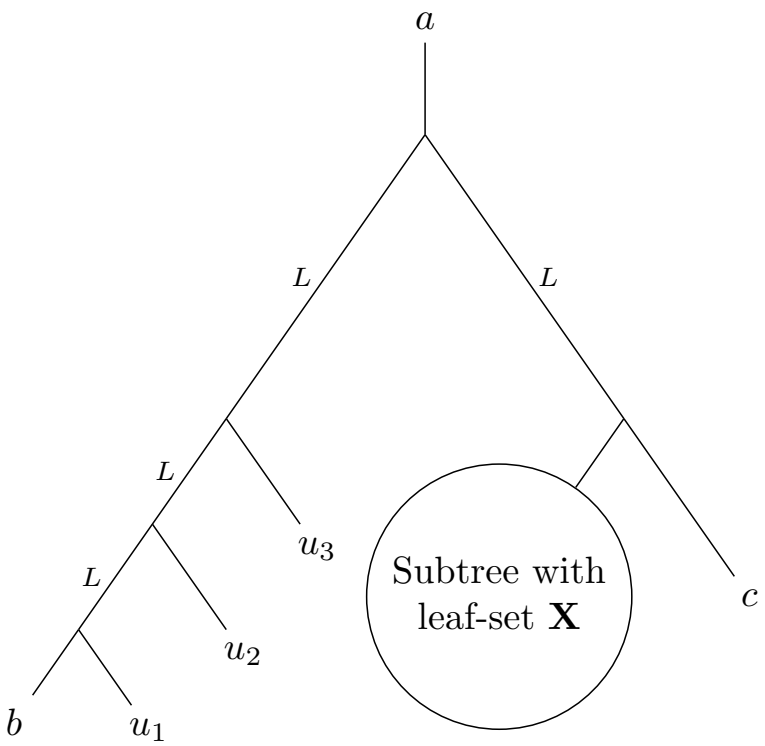

Figure S8: **And example where the AAD method might be misleading for long branches.**

References

[1] Tae Kun Seo. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. *Molecular Biology and Evolution*, 25(5):960–971, 2008.

[2] Jed Chou, Ashu Gupta, Shashank Yaduvanshi, Ruth Davidson, Mike Nute, Siavash Mirarab, and Tandy Warnow. A comparative study of SVDquartets and other coalescent-based species tree estimation methods. *BMC Genomics*, 16(Suppl 10):S2, 2015.

[3] Siavash Mirarab, Md Shamsuzzoha Bayzid, and Tandy Warnow. Evaluating summary methods for multi-locus species tree estimation in the presence of incomplete lineage sorting. *Systematic Biology*, page syu063, 2014.

[4] Siavash Mirarab, Md Shamsuzzoha Bayzid, Bastien Boussau, and Tandy Warnow. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science*, 346(6215), 2014.

[5] Md Shamsuzzoha Bayzid, Siavash Mirarab, Bastien Boussau, and Tandy Warnow. Weighted Statistical Binning: Enabling Statistically Consistent Genome-Scale Phylogenetic Analyses. *PLoS ONE*, 10(6):e0129183, 2015.