# Species Tree Estimation Using ASTRAL: Practical Considerations

Siavash Mirarab

**Abstract**

ASTRAL is a method for reconstructing species trees after inferring a set of gene trees and is increasingly used in phylogenomic analyses. It is statistically consistent under the multi-species coalescent model, is scalable, and has shown high accuracy in simulated and empirical studies. This chapter discusses practical considerations in using ASTRAL, starting with a review of published results and pointing to the strengths and weaknesses of species tree estimation using ASTRAL. It then continues to detail the best ways to prepare input gene trees, interpret ASTRAL outputs, and perform follow-up analyses.

## 1 Introduction

Understanding gene trees as entities evolving within species trees, the framework nicely summarized by Maddison (1997), has given statisticians a powerful model to approach genome-wide phylogenetic reconstruction. Genome evolution can be understood using a hierarchical generative model (Fig. 1a): gene trees are first sampled from a distribution defined by a model of gene evolution and parameterized by the species tree; then, sequences are sampled from distributions defined by a model of sequence evolution and parameterized by the gene trees and other necessary parameters. The choice of the exact model of sequence evolution and the model of gene tree evolution defines the exact hierarchical model.

A leading model of gene evolution is the multi-species coalescent (MSC) (see Degnan and Rosenberg, 2009). MSC models incomplete lineage sorting (ILS) and the resulting discordance between gene trees and the species tree (Fig. 1b). Note that, I use terms *gene* and *locus* interchangeably to refer to a recombination-

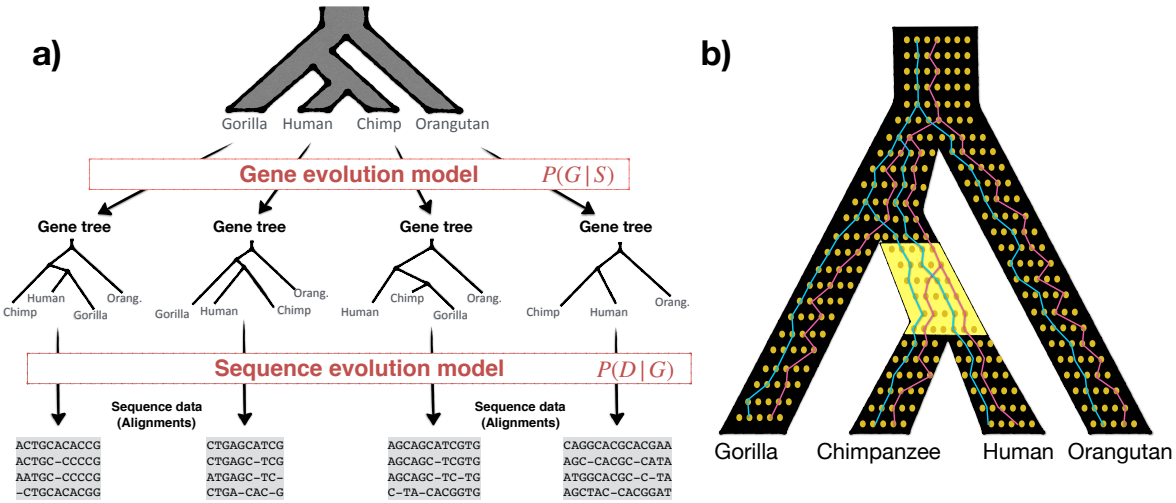


**Figure 1:** (a) The hierarchical model of genome evolution: the species tree parameterizes a model of gene tree evolution; gene trees, sampled from this model, parametrize a model of sequence evolution, which generates the sequences. (b) Tracing two lineages inside a species tree where each branch is a population. Pink lineages coalesce in ways that match the species tree topology. Cyan lineages fail to coalesce in the common ancestor of Human and Chimpanzee (yellow population), giving the cyan lineage from Chimpanzee a chance to coalesce with Gorilla before coalescing with Human – and creating Incomplete Lineage Sorting (ILS).

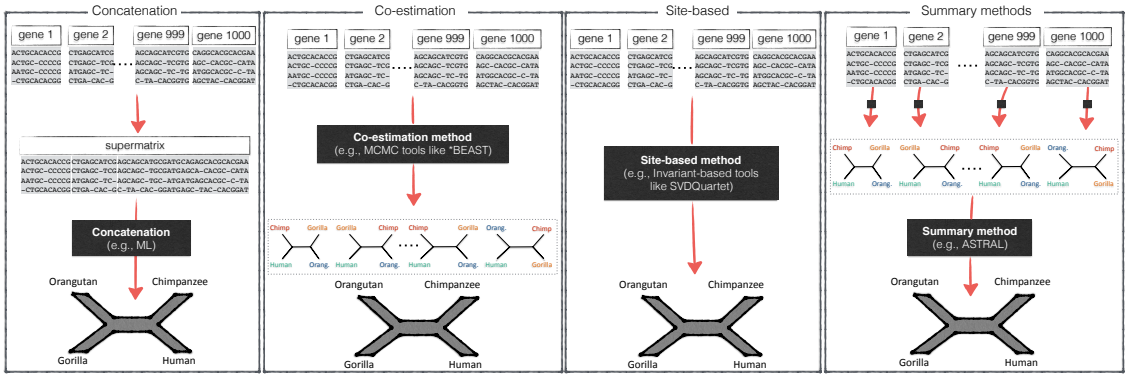


**Figure 2:** Four main approaches to species tree estimation.

free region of the genome (not functional genes). The MSC model is widely adopted due to its perceived biological realism and its mathematical convenience. Under MSC, the species tree is identifiable from a distribution of gene trees, giving us hope to recover the species tree from sampled gene trees.

Several approaches exist for inferring the species tree given multi-gene sequence data under MSC. Concatenating sequence from all loci and performing ML inference under a model of sequence evolution (Fig. 2) amounts to ignoring the gene evolution component of the hierarchical model (Fig. 1) and is proved by Roch and Steel (2015) not to be statistically consistent. This inconsistency, predicted earlier by Kubatko and Degnan (2007), has motivated the development of alternative MSC-based approaches.

Given the hierarchical nature of the model, the most statistically principled approach is to co-estimate gene trees and the species trees as part of one joint inference (Fig. 2). Methods of co-estimation have been developed, mostly using Bayesian MCMC to sample the distributions defined by the hierarchical model (e.g., Liu, 2008; Heled and Drummond, 2010) and have been shown in simulations to have good accuracy under the MSC model (Bayzid and Warnow, 2013; Ogilvie et al., 2016). These methods, however, need to sample a vast number of parameters: topologies of the species tree and all gene trees, their branch lengths, sequence evolution parameters (including rates of evolution), and population size. Due to the large parameter space, co-estimation methods have remained unable to scale to large or even moderate-size datasets despite recent progress (Ogilvie et al., 2017) and the use of divide-and-conquer (Zimmermann et al., 2014).

Scalable alternatives to co-estimation are of two types: summary methods and site-based methods. Site-based methods (e.g., Chifman and Kubatko, 2014; Bryant et al., 2012; De Maio et al., 2013) go directly from gene data to the species tree, without inferring gene trees, yet accounting for MSC. For example, SVDQuartets, a leading site-based method, uses invariants on site pattern matrices. Due to their reduced number of parameters, site-based methods are more scalable than co-estimation (Molloy and Warnow, 2019).

Summary methods divide the inference into two steps (Fig. 2); first, infer gene trees independently for all loci, then, combine these gene trees to get the species tree. Under the MSC model, sequence data from different genes are independent *conditioned* on gene trees but are not independent for unknown gene trees. Thus, summary methods can be understood as ignoring the dependence between gene loci in the gene tree inference step. Once gene trees are inferred, combining them to infer a species tree needs specific methods that are statistically consistent under the MSC model. Examples of such consistent methods include STAR (Liu et al., 2009), BUCKy-population (Larget et al., 2010), GLASS (Mossel and Roch, 2010), MP-EST (Liu et al., 2010), STELLS (Wu, 2012), DISTIQUE (Sayyari and Mirarab, 2016a), NJst (Liu and Yu, 2011), and a related method ASTRID (Vachaspati and Warnow, 2015). By breaking the analysis into many independent inferences, the summary approach can produce a very scalable pipeline (requires careful choices of methods). Perhaps because of their scalability, summary methods are widely used in biological analyses (see ). In particular, a summary method called ASTRAL (Mirarab et al., 2014b) has been used in many publications. In this Chapter, I focus on ASTRAL, intending to give practitioners guidelines for using it.

Section 2 overviews algorithmic details and theoretical properties of ASTRAL. Section 3 summarizes the literature on the performance of ASTRAL. The accuracy of ASTRAL depends on its input quality, and thus, Section 4 is dedicated to best practices in preparing the input gene trees. Sections 5 and 6 elaborate on the output of ASTRAL and follow-up analyses that can help researchers better understand the results.

# 2 ASTRAL Algorithm

## 2.1 Motivation and History

Computing the probability of a gene tree given a species tree is computationally challenging (Degnan and Salter, 2005), especially when the gene tree does not have branch lengths in coalescent units. Thus, developers of summary methods have looked beyond likelihood-based approaches. A helpful feature of MSC is that for rooted gene trees with three species (triplets) or unrooted gene trees with four species (quartets), the species tree topology is the most probable gene tree topology (Pamilo and Nei, 1988; Allman and Rhodes, 2003). Thus, on triplets/quartets of species, we can count the number of rooted/unrooted gene trees and pick the most frequent one as the species tree; it is trivial to show this method statistically consistent assuming gene trees are sampled from the distribution defined by MSC on a species tree. In contrast to triplets and quartets, in the general case of more species, the species trees can be discordant with the most likely gene trees (Degnan and Rosenberg, 2006, 2009), a condition known as the anomaly zone.

Several methods have extended the most-frequent-gene-tree method to more species by decomposing a dataset of $n$ species to all possible $\binom{n}{3}$ triplets or $\binom{n}{4}$ quartets. Larget et al. (2010) suggested using Bayesian concordance factors (Ané et al., 2007) to compute the most frequent quartet tree for all possible choices of quartets, and then, combining the quartets using a quartet-joining method (Ma et al., 2008). More recently, Sayyari and Mirarab (2016a) derived a consistent distance estimate between pairs of species based on how many times they are sisters among all possible quartets that include the two species of interest. Instead of finding the highest frequency gene tree, Liu et al. (2010) defined the pseudo-likelihood of the species tree by decomposing it into all possible triplets, computing the likelihood for each triplet, and combining the likelihoods by assuming independence. ASTRAL, too, decomposes gene trees to quartets.

The main insight behind ASTRAL is to realize that the solution to the following optimization problem is a consistent estimator of the species tree (easy to prove based on results of Allman et al. (2011)). Let $\mathcal{Q}(T)$ be the set of all quartet tree topologies induced by a tree $T$.

**Maximum Quartet Support Species Tree (MQSST):** *Given a set of $k$ unrooted gene tree topologies $\mathcal{G}$ on (subsets of) $n$ species, find the species tree $T^*$ that shares the maximum total number of quartet trees with the set of gene trees. That is, find $T^* = \arg\max_T S(T)$ where*

$$S(T) = \sum_{G \in \mathcal{G}} |\mathcal{Q}(T) \cap \mathcal{Q}(G)| \,. \tag{1}$$

MQSST has been studied even before its connection to MSC was realized. The problem is NP-hard in several variations (Steel, 1992; Jiang et al., 2001; Lafond and Scornavacca, 2019), but heuristic solutions exist (e.g., Avni et al., 2015). One way to achieve scalability is to define a constrained version of the problem.

**Constrained MQSST:** *Solve the MQSST problem such that every branch (i.e., bipartition) of the species tree is drawn of a given set $\mathcal{X}$ of possible branches.*

Bryant and Steel (2001) were the first to define this problem (to my knowledge), which they solved using dynamic programming in time that grows as $O(n^5k + n^4|\mathcal{X}| + |\mathcal{X}|^2)$. ASTRAL uses a dynamic programming algorithm similar (but not identical) to that of Bryant and Steel (2001), with an improved running time (we were unaware of the method by Bryant and Steel (2001) in our original publication) and pointed out how solutions to constrained MQSST are consistent estimators under MSC model.

## 2.2 ASTRAL Algorithm

ASTRAL has three published versions: ASTRAL, ASTRAL-II, and ASTRAL-III, and most recently, a parallel implementation, ASTRAL-MP. Below, when not otherwise specified, we discuss ASTRAL-III. Readers not interested in mathematical and algorithmic details can skip this section.

### 2.2.1 Weight Calculation and Dynamic Programming

A node in a binary (or multifurcating) unrooted tree $T$ corresponds to a partition of leaves into three (or more) parts (Fig. 3a). Thus, a binary (multifurcating) tree can be represented as a set of tripartitions (multipartitions), one per node. The ASTRAL algorithm is based on three insights:

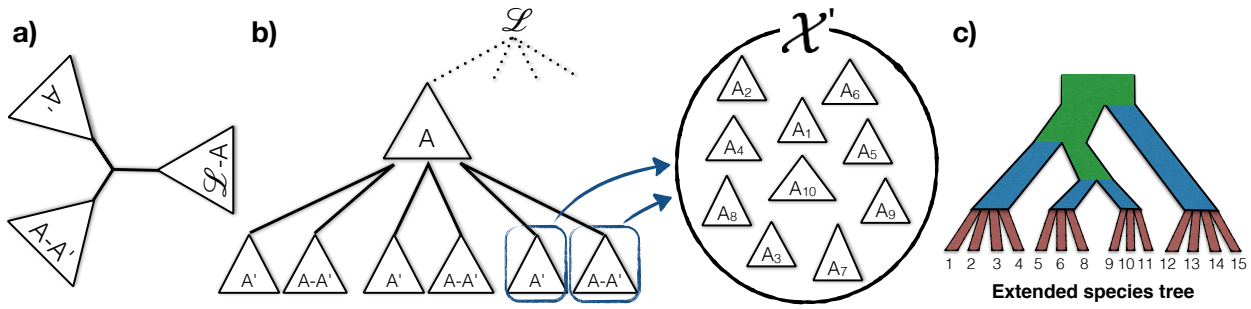


**Figure 3:** (a) An internal node in an unrooted tree create a (tri)partition of leaves. (b) The dynamic programming recursively divides each cluster $A$ into smaller cluster, drawing possible subsets from the set $\mathcal{X}$. (c) An extended species tree; species terminal branches are in blue; individuals are added as polytomies.

1. The number of quartet trees shared between two trees equals half the sum of the number of quartet topologies shared among all pairs of tripartitions/multipartitions, one from each tree.
2. The number of quartet topologies shared between a tripartition and a multipartition can be computed efficiently *without* listing all $\binom{n}{4}$ quartet topologies.
3. Dynamic programming can be used to find a set of tripartitions that can be combined into a fully binary tree and in total have the maximum possible number of shared quartets with gene trees.

Let the set of species be $\mathcal{L}$. Let also $\mathcal{N}(T)$ be the set of internal nodes in a tree $T$, represented as multipartitions. Any (species) tree that includes a tripartition $P$ as a node shares a certain number of quartet topologies with any (gene) tree that includes a multipartition $M$ as a node; we let $QI(P, M)$ denote this quantity. We define the *weight* of a tripartition as:

$$w(P) = \frac{1}{2} \sum_{G \in \mathcal{G}} \sum_{M \in \mathcal{N}(G)} QI(P, M) \; . \tag{2}$$

Insight 1 asserts that $S(T) = \sum_{P \in \mathcal{N}(P)} w(P)$. Thus, we need to $i$) compute $w(P)$ efficiently, and $ii$) find the tree with the maximum sum of $w(P)$ values. Zhang et al. (2018) derived an efficient formula for $QI$.

Given a multipartition $M = M_1|\ldots|M_d$ (representing an internal node in a gene tree) and a tripartition $P = P_1|P_2|P_3$ (an internal node in a species tree), for $1 \leq i \leq d$ and $1 \leq j, k \leq 3$, let $I(i, j) = |M_i \cap P_j|$, $S(j) = \sum_{i=1}^{d} I(i, j)$, and $R(j, k) = \sum_{i=1}^{d} I(i, j)I(i, k)$. Let $\left(\begin{smallmatrix} 2 & 1 & 1 \\ 3 & 3 & 2 \end{smallmatrix}\right) = (h_{i,j})$ be a constant matrix. Then,

$$QI(P, M) = \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{3} \binom{I(i,j)}{2} \Big( \big(S(h_{1,j}) - I(i, h_{1,j})\big)\big(S(h_{2,j}) - I(i, h_{2,j})\big) - R(h_{1,j}, h_{2,j}) + I(i, h_{1,j})I(i, h_{2,j}) \Big) \; . \tag{3}$$

Computing this equation requires $\Theta(d)$ time given $I(j, i)$ values. ASTRAL-III uses a polytree data structure to represent gene trees such that each $I(j, i)$ can be computed in constant time. The data structure also compresses nodes that appear in multiple gene trees. Zhang et al. (2018) showed how to compute $w(P)$ in $\Theta(D)$ where $D = O(nk)$ is the sum of the cardinalities of unique partitions observed in all gene trees.

To maximize $S(T)$, ASTRAL recursively divides $\mathcal{L}$ it into two smaller subsets (called *cluster*s) until it reaches leaves (Fig. 3b). Each cluster is divided such that the total sum of the weights below it is maximized among allowable divisions. This recursive method solves the MQSST problem optimally if all ways of dividing a cluster $A$ into $A' \subset A$ and $A - A'$ are examined. But that approach has an exponential running time, hence, the need for the constrained MQSST problem. Assume we have defined a set $\mathcal{X}$ of allowable bipartitions for the species tree. Let $\mathcal{X}' = \{A : A|\mathcal{L} - A \in \mathcal{X}\}$ and $Y = \{(C, D) : C \in \mathcal{X}', D \in \mathcal{X}', C \cap D = \emptyset, C \cup D \in \mathcal{X}'\}$. Kane and Tao (2017) showed $|Y| < |\mathcal{X}'|^{1.726}$. We restrict the dynamic programming such that $(A', A - A') \in Y$ (Fig. 3b). Let $S^*(A)$ be the score for an optimal subtree on cluster $A$. Then, the following dynamic programming solves the constrained MQSST problem optimally in time that scales in the worst case as $O(D|\mathcal{X}|^{1.726})$, spending the majority of time in computing Equ. 2 (Zhang et al., 2018):

$$S^*(A) = \max_{(A', A - A') \in Y} S^*(A') + S^*(A - A') + w(A'|A - A'|\mathcal{L} - A) \; . \tag{4}$$

#### 2.2.2 Constraint set

The sufficient condition for ASTRAL to be statistically consistent under the MSC model is to have all bipartitions from input gene trees in the set $\mathcal{X}$ (Mirarab et al., 2014b). However, Mirarab and Warnow (2015) showed that $\mathcal{X}$ might need to be expanded in order to obtain high accuracy in practice. ASTRAL-III expands $\mathcal{X}$ using rules summarized below, but users can also directly expanded the set. These heuristics rely on a similarity matrix computed based on how often a pair of species appear as sisters in gene tree quartets.

- When input gene trees are incomplete, first complete them before adding their bipartitions to $\mathcal{X}$ using the similarity matrix (Mirarab and Warnow, 2015). Similarly, when gene trees include polytomies, first resolve their polytomies in several ways before adding the bipartitions to $\mathcal{X}$ (Zhang et al., 2018).
- Compute greedy consensus trees of gene trees with various thresholds (the minimum required frequency for adding bipartitions to the consensus), resolve the polytomies in the greedy consensus trees, and add the resulting bipartitions to $\mathcal{X}$. To resolve polytomies, subsample one species from each side of the polytomy, and resolve it using two approaches: using the similarity matrix and by computing greedy consensus trees on the subsampled taxa.
- Ensure heuristics do not add more than $O(nk)$ bipartitions. With this rule, we get running times that increase as $O(D(nk)^{1.726}) = O((nk)^{2.726})$.

#### 2.2.3 Multiple individuals

ASTRAL can easily be extended to inputs where more than one species represent each species. Allman et al. (2011) have introduced the concept of an extended species tree: start with the species tree, and for each species, add all individuals sampled from that species under it, creating polytomies when needed (Fig. 3c). Rabiee et al. (2019) have extended dynamic programming of Equation 4 to compute the optimal extended species tree given gene trees with multiple individuals from some or all species. The dynamic programming is unchanged (treating individuals as taxon set $\mathcal{L}$), except for two modifications. ($i$) The boundary conditions need to change such that the algorithm stops as soon as a cluster equals the set of individuals of a species. ($ii$) Set $\mathcal{X}$ needs to change such that each cluster has either all or none of the individuals of each species. Satisfying this condition required new methods for building $\mathcal{X}$ (Rabiee et al., 2019).

### 2.3 Summary of known theoretical results related to ASTRAL

**Consistency - general.** All versions of ASTRAL give a statistically consistent estimator of the species tree if input gene trees are sampled randomly under the multi-species coalescent model (i.e., with no gene tree error, no sampling bias, and no model violations).

**Consistency - missing data.** Nute et al. (2018) showed that ASTRAL remains statistically consistent when species are allowed to be missing from gene trees. Key assumptions required for the exact ASTRAL are that the presence of a gene for a species should be independent of the gene tree topology and presence of other genes for that species. The default (constrained) version is also consistent if each clade of the species tree has a *non-zero* chance of having no missing data in each gene.

**Inconsistency - estimated gene trees.** Roch et al. (2019) have proved that ASTRAL and other "reasonable" summary methods that use gene tree topology are statistically inconsistent if each gene has limited length and gene trees are computed using ML. Under specific conditions, they show ASTRAL and even partitioned ML fail due to long branch attraction even if there is *no gene tree incongruence*.

**Inconsistency - Reticulation.** Solís-Lemus et al. (2016) have shown that ASTRAL can be statistically inconsistent under certain conditions when gene trees evolve on a phylogenetic network (thus, with a combination of ILS and gene flow).

**Sample complexity.** Shekhar et al. (2018) have shown that the number of genes required by the exact version of ASTRAL to compute the correct species tree with high probability grows quadratically with the inverse of the shortest branch length and grows logarithmically with the number of species.

# 3 Accuracy and Scalability

## 3.1 Accuracy

**Versus concatenation.** Simulation studies have indicated that the relative performance of concatenation and ASTRAL depends at least on two factors: the amount to true gene tree discordance (ILS) and the amount of gene tree estimation error (e.g., Mirarab et al., 2014b; Mirarab and Warnow, 2015; Giarla and Esselstyn, 2015; Molloy and Warnow, 2018). For example, Mirarab and Warnow (2015) found ASTRAL to be more accurate than concatenation using ML (CA-ML) when true gene tree discordance (i.e., level of ILS) was high (Fig. 4a) or when gene tree error was relatively low (Fig. 4b). In contrast, CA-ML was more accurate when either the true gene tree discordance was low or when the discordance was moderate or high, but the gene tree error was also high. The two methods had similar accuracy when ILS levels were moderate, *and* gene tree error was also moderate; e.g., when normalized Robinson and Foulds (1981) (RF) distance between true and estimate gene trees was between 20% to 40% (Fig. 4b). Moreover, Davidson et al. (2015) found ASTRAL outperforms CA-ML in the presence of both ILS and HGT (Fig. 4f).

In practical terms, when gene tree discordance is very low or when gene tree error is expected to be high, CA-ML may be preferable to ASTRAL whereas in other scenarios ASTRAL is preferable. Because neither method universally dominates the other, it seems wise to use both methods and compare the results. One way to decide is to simulate data emulating real data and comparing methods. For example, Giarla and Esselstyn (2015) found that for their dataset of tree shrews, both ASTRAL and CA-ML produced one wrong branch in simulations that emulated the real data, but only CA-ML had high bootstrap support for the wrong branch. Ballesteros and Sharma (2019) showed in simulations that ASTRAL could recover the correct branch even when a branch does not appear in *any* of the input gene trees.

**Versus other summary methods.** Several simulations studies compare summary method (e.g., Mirarab et al., 2014b; Mirarab and Warnow, 2015; Sayyari and Mirarab, 2016a; Molloy and Warnow, 2018; Vachaspati and Warnow, 2015), including some that do not involve developers of ASTRAL (Giarla and Esselstyn, 2015; Ballesteros and Sharma, 2019). Overall, the accuracy of ASTRAL has compared favorably to alternative ILS-based summary methods such as NJst (Liu and Yu, 2011), ASTRID (Vachaspati and Warnow, 2015), MP-EST (Liu et al., 2010), wQMC (Avni et al., 2015), as well as consensus and supertree methods such as greedy consensus, MulRF (Chaudhary et al., 2013), and MRP (Ragan, 1992). For example, ASTRAL outperformed NJst by small but consistent margins in simulations by Mirarab and Warnow (2015) (Fig. 4a-c) and in simulations with HGT done by Davidson et al. (2015) (Fig. 4f) and was essentially tied with ASTRID in Molloy and Warnow (2018) and Vachaspati and Warnow (2015). DISTIQUE was close to ASTRAL but not any better in Sayyari and Mirarab (2016a). ASTRAL dominated MP-EST (Mirarab et al., 2014b; Mirarab and Warnow, 2015), especially with large numbers of species (Fig. 4c). Giarla and Esselstyn (2015) showed ASTRAL outperforms MulRF in terms of topological accuracy on simulations that match their real dataset of tree shrews. Beyond accuracy on conditions that seek to emulate real data, Shekhar et al. (2018) have compared ASTRAL with NJst in idealized cases in terms of data requirements: the number of error-free genes required to recover the correct species tree with high accuracy. Their simulations showed mixed patterns: out of three true species trees tested, they found ASTRAL required fewer genes in two cases (with only short branches), and NJst required fewer genes in the third case (with long basal branches).

**Other results.** Simulation studies have tested the robustness of ASTRAL to factors such as HGT (Fig. 4f), gene flow (Solís-Lemus et al., 2016), gene tree error (Bayzid et al., 2015), and missing data (Molloy and Warnow, 2018) (discussed later). Rabiee et al. (2019) studied the relative impact of increasing the number of loci or the number of individuals per species on ASTRAL accuracy and found more loci to be far more beneficial. Beyond simulations, researchers have also compared ASTRAL to other methods on empirical data (e.g., Giarla and Esselstyn, 2015; Simmons et al., 2016; Edwards et al., 2016; Meiklejohn et al., 2016; Streicher et al., 2016; Shen et al., 2017). Since the ground truth is not known on real data, these results are harder to interpret and cannot be easily summarized without loss of important nuance. Referring the reader to these publications, I note that overall, the performance of ASTRAL on real data has been positive.

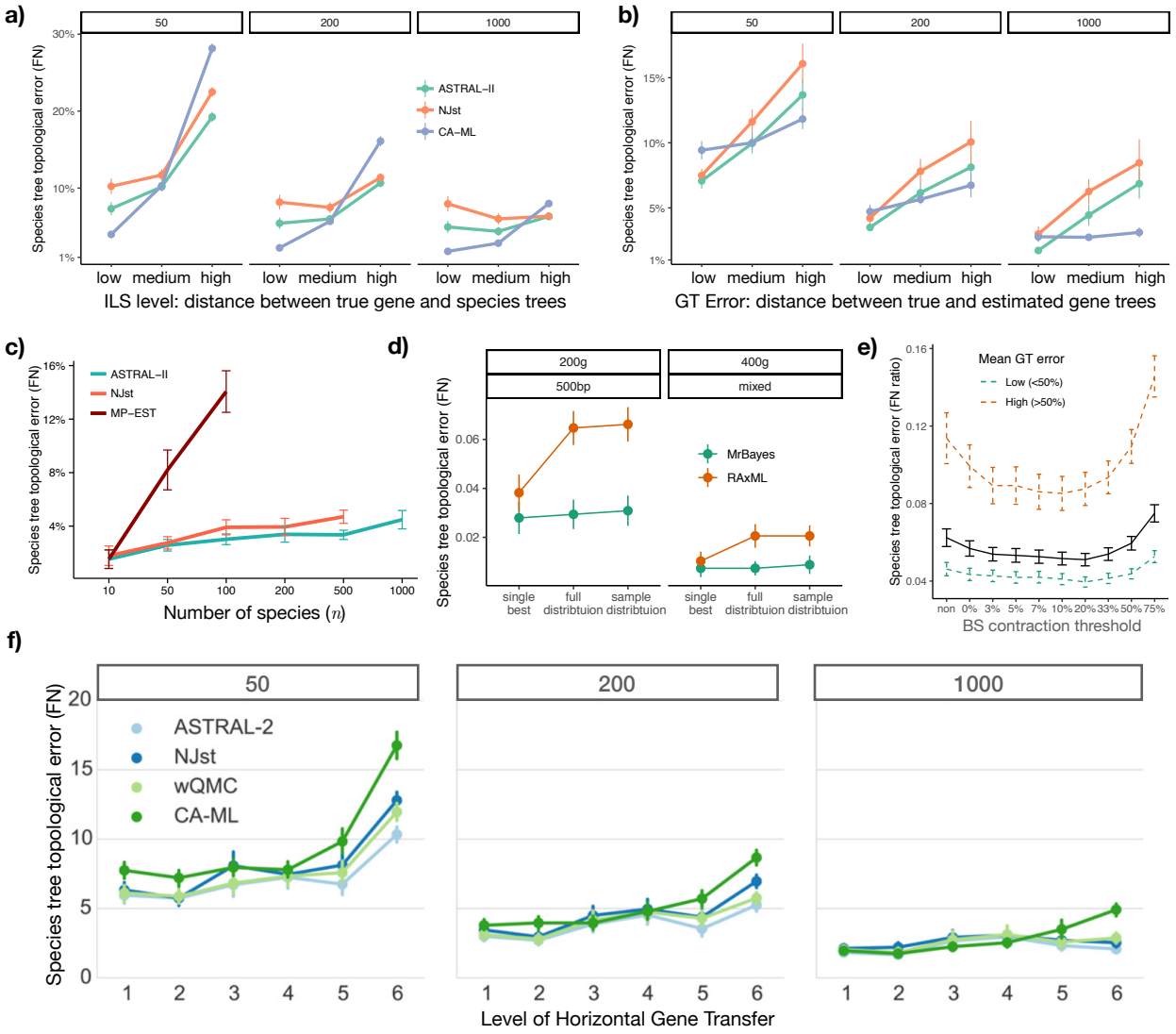


**Figure 4: Accuracy of ASTRAL in simulations**. Shown is the species tree error, defined as the proportion of branches in the true tree missing from the estimated trees; i.e., False Negative (FN) rate. (a-c) Simphy simulations by Mirarab and Warnow (2015) comparing concatenation (CA-ML) and summary methods. (a) The level of ILS is set to low, medium, or high by adjusting tree height ($10^7$, $2\times10^6$, $5\times10^5$) and affects mean RF distance between true species trees and true gene trees (9%, 34%, 68%). Speciation rate: $10^{-7}$; number of genes: 50, 200, or 1000 (boxes); 201 species. (b) Replicates in (a) are categorized into three sets based on the mean normalized RF distance between true gene trees and gene trees estimated by FastTree: $[0, 25]$%, $(25, 40]$%, $(40, 100]$%, corresponding to low, medium, and high gene tree error. (c) Error versus the numbers of species for medium levels of ILS ($2 \times 10^6$ height) and speciation rate $10^{-6}$ with 1000 genes. (d) Mammalian-like simulations from Mirarab et al. (2014b) with 37 species with 200 gene trees (500bp), or 400 gene trees (mix of 500bp and 1000bp). Gene trees estimated using RAxML or MrBayes. Input: *single best* tree per gene (ML for RAxML or maximum credibility for MrBayes); *full distribution* per gene (200 BS replicates for RAxML or a sample of 200 trees for MrBayes); *sample distributions* to get a single tree per gene and repeat 200 times to report their consensus (a.k.a MLBS). MrBayes results are unpublished. (e) Simulations with 100 species and moderately high ILS (46% RF between true gene trees and species trees) by Zhang et al. (2018). Contracting branches with $\leq 20$% BS support (threshold of contraction shown on x-axis) from best ML trees reduces error. Dividing replicates into high and low gene tree error shows that contraction helps mostly in the high error case. (f) 50 species simulations by Davidson et al. (2015), comparing accuracy of methods in presence of HGT. All six model conditions (x-axis) have ILS but differ in level of HGT, ranging from no HGT (1) to very high (6). Thus, true gene tree discordance varies: ∼33% (1-3), ∼45% (4), ∼55% (5), and ∼70% (6). CA-ML is the least robust and ASTRAL is the most robust to HGT.

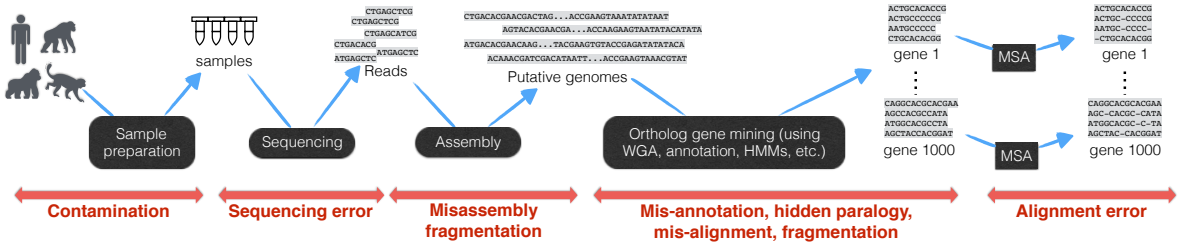


**Figure 5:** The phylogenomics pipeline. To get the typical input to coalescent-based methods, a set of gene alignments, many steps have to be taken, and each step is prone to errors that can propagate.

### 3.2 Running time

The running time of ASTRAL has improved through its three versions, both in theoretical guarantees of worst-case asymptotic running time and in empirical measures. ASTRAL-I had guaranteed polynomial running time but is the slowest version. ASTRAL-II did not guarantee polynomial running time but was faster than ASTRAL-I. The current version, ASTRAL-III, has an asymptotic worst-case running time of $O(D(nk)^{1.726})$, which itself is $O((nk)^{2.726})$ (recall that $D$ is the sum of degrees of *unique* nodes in input gene trees). The theoretical running time, thus, is a function of $n$, $k$, and the amount of gene tree discordance, which controls both the search space and $D$. In practice, on datasets tested by Zhang et al. (2018), the empirical running time of ASTRAL-III seems to increase with $n^2k^2$. Thus, for example, a researcher planning to double the number of gene trees can expect a four folds increase in the running time. Similarly, other things equal, increasing the number of species should roughly quadruple the running time.

Example numbers may be instructive. Zhang et al. (2018) report that ASTRAL-III took roughly 16 hours on average for a dataset with very high ILS, 48 species, and 16,000 genes (so, a large-$k$ scenario), and roughly 9 hours on a dataset with moderate ILS, 5000 species, and 1000 genes (so, a large-$n$ dataset).

For a comparison of the running time of ASTRAL to other methods, see Molloy and Warnow (2019).

**ASTRAL-MP.** Yin et al. (2019) have recently developed a parallel version of ASTRAL for CPU (multi-core and vectorization) and GPU that can analyze very large datasets. This version, called ASTRAL-MP, speeds up runs by up to 100X compared to ASTRAL-III, especially for datasets with large numbers of gene trees. On a dataset with 10,000 species, 1000 gene trees, and moderate ILS, ASTRAL-MP takes between 5 to 32 hours (11 hours on average) given a single GPU and 24 cores with AVX2. On a real insect transcriptomic dataset with 144 taxa and 1478 genes, each with 100 bootstrapped gene trees ($k = 147800$ in total), ASTRAL-MP with four GPUs and 24 cores finished in 35 hours.

## 4 Input to ASTRAL: Practical Considerations

The ideal input to coalescent-based methods is a set of perfectly aligned orthologous regions present in all genomes, with each region small enough to avoid recombination but large enough to have a strong phylogenetic signal and with regions distributed randomly across the genome and placed far enough to make them fully unlinked. For summary methods, gene trees are ideally estimated under models of sequence evolution that are correct (but not overly parameterized) using consistent methods that utilize the data efficiently (i.e., have optimal sample complexity). Satisfying all these requirements is hard, if not impossible. Thus, phylogenomic projects seeking to use summary methods like ASTRAL face many practical choices.

In practice, Phylogenomic analyses include many steps before arriving at gene tree and species tree estimation (Fig. 5). From sample preparation to sequencing, assembly, and annotation, to orthology detection and multiple sequence alignment (MSA), the pipeline includes steps that are far from trivial (Philippe et al., 2017). Each step is error-prone, and some steps (e.g., orthology detection and MSA) seek to solve computational problems that are incidentally best solved with the knowledge of the phylogeny.

Several groups have discussed error propagation through steps of the pipeline and the impact on the species tree (e.g., Patel, 2013; Mirarab et al., 2014a; Gatesy and Springer, 2014; Philippe et al., 2017; Molloy and Warnow, 2018). We can aspire to move away from a pipeline approach and towards a unified statistical

inference, with full join modeling of uncertainty (Szöllősi et al., 2014). Since this end-to-end co-estimation remains unavailable currently and likely impractical in the near future, we are left having to deal with pipelines, which requires awareness of errors and making an effort to mitigate their impact. Below, I discuss best practices that have emerged from published work in preparing the input to ASTRAL.

## 4.1 Gene tree estimation

### 4.1.1 Gene tree uncertainty

The standard input to ASTRAL is ML gene trees inferred under standard models of sequence evolution. Restricting ourselves to ML, several options are available.

**bestML:** The most straightforward choice is to use the tree with the best likelihood found by a heuristic ML method. The bestML input is the most natural approach but ignores gene tree uncertainty.

**Contracted bestML:** Each gene tree is bootstrapped, and support values are computed for bestML trees. Then, branches with extremely low support in bestML trees are contracted, and the resulting multifurcating trees are used as input to ASTRAL (one per gene).

**MLBS:** Multi-locus Bootstrapping (MLBS) seeks to model uncertainty by performing bootstrapping for each gene. Then, these bootstrapped gene trees are used to create several inputs to ASTRAL (with or without gene resampling); running ASTRAL on each input set produces a set of outputs, which are then summarized using methods such as greedy consensus to generate a final consensus result.

**ALLBS:** All replicate bootstrapped gene trees are combined to form a single input to ASTRAL.

At least two simulation studies (Mirarab et al., 2016, 2014b) have shown that MLBS or ALLBS have lower accuracy than simply using bestML, except perhaps when only a small number of genes are available (see Fig. 4d). Sayyari and Mirarab (2016b) have provided an explanation. The set of bootstrapped gene trees show a higher level of gene tree discordance than the set of bestML gene trees. The increased discordance is not biological but is a result of the lowered phylogenetic signal in bootstrapped gene alignments. This increased level of gene tree error, as we previously saw (Fig. 4b), can reduce the accuracy of ASTRAL.

Contracted bestML, in contrast to MLBS, can improve accuracy. The important (if somewhat counter-intuitive) point to remember is that only collapsing branches with *extremely low* support improves accuracy, and contracting other branches can *increase* error. Zhang et al. (2018) have shown that collapsing branches with BS below 5-20% can improve accuracy by a substantial margin (Fig. 4d) and that the improvements in accuracy are higher when input gene trees have higher levels of error and when more gene trees are available. For example, for input gene trees with the mean error above 50%, ASTRAL tree enjoys a 25% reduction in error (from 0.114 to 0.085) after contracting gene tree branches with support below 10%. Thus, contracting *very* low support branches can increase accuracy substantially. Aggressively collapsing branches with support $< 50\%$ or $< 75\%$ (i.e., only keeping high support branches) can substantially reduce the accuracy (Fig. 4d). Future research should explore smarter algorithms for collapsing low support branches.

Arcila et al. (2017) have suggested inferring gene trees constrained to include a set of predefined undisputed clades chosen by the researcher. They show promising results on several datasets using ASTRAL applied with such constrained gene trees. However, the method can also remove some of the real discordances among gene trees (as opposed to noise). As pointed out elsewhere (Mirarab, 2017), this approach runs several theoretical risks, including biasing result in unexpected ways.

### 4.1.2 Inference tools and models

The choice of the gene tree inference tool may be consequential, and published simulations have compared FastTree (Price et al., 2010) and RAxML (Stamatakis, 2014). Despite earlier results (Liu et al., 2011), Sayyari et al. (2017) have found using simulated and empirical data that FastTree *can* be less accurate than RAxML in inferring gene trees (under limited conditions they test), and the increased gene tree error leads to less accurate ASTRAL trees. Thus, using best available ML methods (ideally with multiple starting trees) is preferable to faster tools in biological analyses. Moreover, effects of misspecified sequence evolution models

have been discussed for phylogenomics in general (e.g., Phillips et al., 2004; Jeffroy et al., 2006), but to my knowledge, have not been studied for ASTRAL. We should expect that systematic model misspecification can lead to biases in estimated gene tree distributions, which can lead to errors in the species tree.

In unpublished simulations, I have compared the ML method RAxML and the Bayesian method MrBayes (Huelsenbeck and Ronquist, 2001) on a mammalian-like simulated dataset (Mirarab et al., 2014b). Like ML, the output of MrBayes is used in three ways: using a single maximum credibility tree per gene, using a large sample of trees per gene (akin to ALLBS), and repeatedly sampling single trees from gene tree distributions produced by MrBayes (akin to MLBS). Interestingly, unlike ML and BS, the use of Bayesian distributions removes the sensitivity to the mode of input so that all three types of input perform similarly (Fig. 4d). These results also show a small advantage in using MrBayes compared to RAxML when both are using in their best setting (i.e., a single tree per gene). These preliminary results warrant more studies in the future.

## 4.2 Filtering data

A vexing problem in phylogenomics is that curating sequence data using visual inspection is impossible. Thus, methods for *detecting* errors automatically, perhaps in downstream steps, are also needed. Empirical studies often employ several mostly *ad hoc* methods for filtering erroneous data (Philippe et al., 2017), typically relying on a mix of visual inspection of (parts of) data and automatic error detection tools. While the extent of the negative impact of errors in input on the output ASTRAL tree is not fully understood, efforts to minimize such errors seem necessary. However, tampering with data to remove error can also remove signal and introduce bias – and thus warrants caution and careful study.

### 4.2.1 Filtering leaves from gene trees

One way of detecting abnormalities is to examine the estimated gene trees that show unexpected patterns. For example, Wickett et al. (2014) rooted gene trees and detected and removed branches from gene trees with extremely long root to tip distance compared to the other species. Visual inspection of gene trees is also what Gatesy, Springer, and colleagues have used in several of their published criticism of previous phylogenomic studies (Gatesy and Springer, 2014; Springer and Gatesy, 2014, 2016, 2017). A well-studied automated approach for detecting species with unstable positions in individual gene trees is rogue taxon detection (e.g., Aberer et al., 2013; Westover et al., 2013). Rogue taxon detection methods tend to identify the same species (usually those on long branches) on many genes (Mai and Mirarab, 2018). Since removing the same taxon from many genes reduces taxon occupancy, rogue taxon removal may prove problematic. The effect of rogue taxon removal on ASTRAL, to my knowledge, has not been tested.

**TreeShrink.** Mai and Mirarab (2018) developed an automatic method called TreeShrink to find suspicious patterns of branch length in gene trees (Fig. 6a). TreeShrink tries to successively shrink the diameter (i.e., the maximum total branch length between any two leaves) of each gene tree by removing species. Then, for each species in each gene tree, TreeShrink computes a *signature* (Fig. 6b), which quantifies its impact on the diameter of the gene tree. Finally, it examines the *distribution* of signatures of each species across all gene trees and detects outliers in these distributions using a simple heuristic. Outliers is a case when a species has an uncharacteristically large signature (e.g., is on a long branch) in a gene tree compared to the rest of gene trees. Since outliers are defined using a distribution across genes for a single species, a taxon with high signatures in most genes will not be detected as an outlier in those genes (outgroups tend to be like this). In contrast, a taxon with low signatures in most genes but high signatures in a handful of genes will be detected as an outlier for those high-signature genes. Once the abnormally long branches are detected, TreeShrink removes specific species from specific genes but does not remove the entire gene. Mai and Mirarab (2018) showed on several biological datasets that TreeShrink reduces the pairwise discordance among gene trees beyond random removal of taxa (Fig. 6c) and also often beyond methods such as RogueNaRok (Aberer et al., 2013). Moreover, it avoids removing any specific species from too many genes.

### 4.2.2 Filtering entire gene trees

A more extreme form of filtering is to remove entire gene trees. Several criteria for removing entire genes have been proposed and tested (e.g., Meiklejohn et al., 2016; Hosner et al., 2016; Chen et al., 2015; Huang

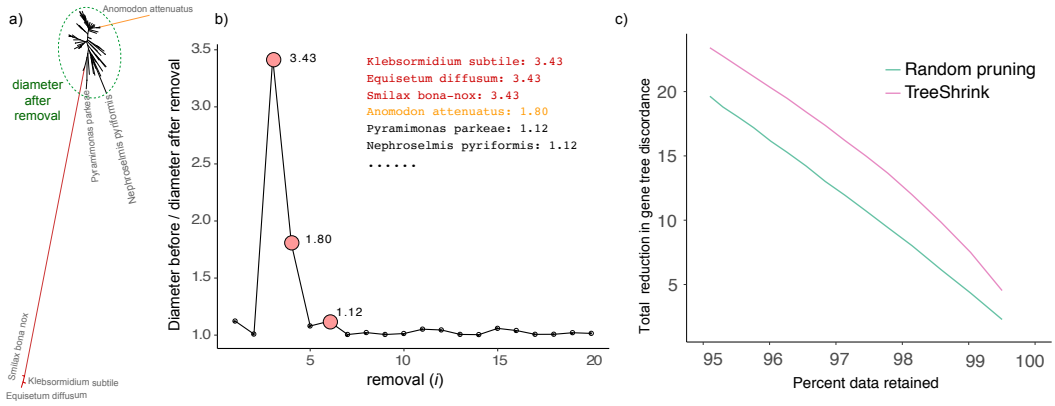


**Figure 6: TreeShrink** on Wickett et al. (2014) dataset. (a) A gene tree with abnormally long branches. (b) TreeShrink identifies three (red) tips that increase the diameter 3.4X and are likely erroneous, followed by another (orange) that may also be problematic. (c) TreeShrink reduces discordance substantially more than random filtering. y-axis: *reduction* in gene tree discordance defined as mean pairwise matching splits distance between gene trees.

et al., 2016; Longo et al., 2017; Blom et al., 2017). Some of the criteria have to do with missing data and will be discussed later, while others relate to error and uncertainty in gene trees. Molloy and Warnow (2018) give a recent summary of the literature and also, to my knowledge, give the only study that evaluates the accuracy of ASTRAL in simulation in response to removing full genes. Their simulations show that removing genes with high gene tree estimation error can improve accuracy for low levels of ILS but *reduces* accuracy for moderate to high levels of ILS. Nevertheless, even for ILS levels where removing genes helps accuracy, on real data, we do not have direct access to gene tree error to decide what genes to remove. Instead, we have to resort to filtering by other factors, such as bootstrap support. Since these proxies do not perfectly correlate with gene tree error, the positive impact of filtering may further diminish. Neither Molloy and Warnow (2018) nor any other simulation study tested filtering by proxies for ASTRAL. Lanier and Knowles (2015) and Liu et al. (2015) have studied the question for STEM and MP-EST and have observed little or no reason for filtering. On empirical datasets, the conclusions have been mixed, with some studies (e.g., Hosner et al., 2016; Meiklejohn et al., 2016; Longo et al., 2017) recommending removal of gene trees and others finding no evidence that filtering helps (Chen et al., 2015; Blom et al., 2017). Overall, there is little evidence in the literature suggesting that removing entire genes because of lack of support is helpful to ASTRAL analyses.

### 4.2.3 Filtering for missing data

The most common type of filtering is in response to missing data. For summary methods, two types of missing data exist – *missing genes* and *fragmentary data* (*type I* and *type II* in the parlance of Hosner et al. (2016)). The two types have different consequences and have inspired two types of filtering.

**Type I (missing genes).** This type of missing data occurs when a gene is entirely missing for some of the species but is present in others. The only suggested filtering for these types of missing data is to remove them. Molloy and Warnow (2018) found no evidence in simulations that removing genes with this type of missing data helps accuracy. Their results are in agreement with several empirical studies (e.g., Chen et al., 2015; Hosner et al., 2016) that also saw no benefit in filtering genes.

**Type II (fragmentation).** When a species includes a gene, but only partially, it introduces missing data in the gene tree inference step. Both Hosner et al. (2016) and Sayyari et al. (2017) showed that the presence of fragmentary data can be problematic, confirming earlier observations (e.g., Wickett et al., 2014; Springer and Gatesy, 2016). They show fragmentary sequences increase gene tree error, which translates to increased species tree error. Sayyari et al. (2017) suggested a simple yet effective solution: remove species with fragmentary sequences from gene alignments before inferring the gene tree. The optimal level of filtering depends on the dataset; Sayyari et al. (2017) defined species with less than half of the total alignment length as fragmentary while Wickett et al. (2014) used a one-third threshold.

To summarize, removing loci because of missing species is not recommended, but removing specific species from loci because of fragmentation is recommended. Note that filtering fragmentary data replaces type II missing data with type I. That such a trade-off helps accuracy, once again, underscores the negative impact of gene tree error and the benefit in reducing error – even if this reduction adds to missing data. Note that the distinction between types I and II is not relevant for concatenation. There is no reason to think that removing fragmentary data from a concatenation analysis could help accuracy, as it only adds missing data.

Results showing that removing type I missing data fails to help accuracy do not imply missing genes are harmless. As shown in simulations by Nute et al. (2018), missing genes can increase error in ASTRAL trees, especially when the number of genes is low, the amount of ILS is very high, or when entire clades tend to be missing. Thus, missing data can hurt accuracy, but filtering low occupancy loci is not a solution.

Recently, Gatesy et al. (2018) added a twist. For a set of empirical data, given two alternative species trees, they computed the difference in quartet score of the two trees *for each gene* and called it partitioned coalescence support (PCS). Genes with extremely high PCS for either alternative tree *tend to* be a lot more complete than other genes; in some cases, ASTRAL trees change if only a couple of these high occupancy genes are removed. They arrived at the opposite conclusion of the common wisdom, suggesting that when several genes have much more occupancy than others, we should remove genes with high occupancy, only keeping genes with fewer species so that all genes have about the same number of species. Future work should further test this idea. More broadly, the creative PCS approach may prove useful in detecting genes with overtly high impact on the species tree due to missing data or other problems.

# 5 ASTRAL Output

## 5.1 Species tree topology and its quartet score

ASTRAL outputs the tree with the maximum quartet score among all trees within its search space (defined by $\mathcal{X}$). Since ASTRAL limits the search space (unless run with `-x`), it is possible that other trees with better quartet scores exist. In simulations, increasing the search space beyond the default used in ASTRAL-II or -III (e.g., by adding all bipartitions from the true tree) rarely even improves accuracy, though it occasionally improves the quartet score slightly (Mirarab and Warnow, 2015; Zhang et al., 2018).

ASTRAL is a statistically consistent estimator under the MSC model given gene trees sampled from the true distribution defined under MSC. However, ASTRAL does not use a parametric model and is not tied to likelihood under the MSC model. As such, ASTRAL can be considered a non-parametric estimator. As it has been long argued (Holmes, 2003), absent access to the correct model, reliance on non-parametric methods can be beneficial. Thus, ASTRAL (and other non-parametric methods like NJst/ASTRID) may be more robust than parametric methods (e.g., MP-EST), especially given the limited reach of the MSC model. After all, MSC ignores *both* gene tree error *and* biological sources of discordance other than ILS. ASTRAL is a natural estimator for any model where the most likely gene tree matches the species tree for quartets, as is the case for some HGT regimes (Roch and Snir, 2013). Further, ASTRAL has even been used as a supertree method outside the phylogenomics context, with reasonable results (Vachaspati and Warnow, 2017).

Along with the tree topology, ASTRAL outputs its quartet score, which is the number of quartet trees in gene trees that are present in the species tree. We normalize the absolute value by the total number of quartet trees in input gene trees (e.g., $k\binom{n}{4}$ if there is no missing data) to give a more interpretable score. For example, a quartet score of 0.8 means that 80% of quartet trees in input gene trees are in the output tree. Thus, the normalized quartet score can be used as a measure of the amount of gene tree discordance. However, the score has to be interpreted with care as gene tree error is likely to reduce the quartet score.

## 5.2 Branch Lengths in Coalescent Units

ASTRAL estimate coalescent units (CU) lengths of all internal branches and of terminal branches corresponding to species with multiple individuals. True CU branch lengths are proportional to the number of generations spanned by the branch and inversely proportional to the population size (Degnan and Rosenberg, 2009). CU is important in MSC modeling because branch CU length is what identifies the amount of topological discordance. Shorter branches lead to more discordance, especially when adjacent to each other. For a quartet, if the length of the only internal branch is $d$ in CU, the probability of a gene tree matching

the species tree is $1 - \frac{2}{3}e^{-d}$ and the probability of each of the two alternative topologies is $\frac{1}{e}e^{-d}$ (Fig 7a). Thus, when a quartet gene tree appears $f > \frac{1}{3}$ times, $-\ln \frac{3}{2}(1-f)$ gives an estimate of the branch length.

ASTRAL exploits this observation to compute CU branch lengths (with simplifying assumptions) using a fast algorithm for computing mean quartet frequencies "around" each branch (Fig 7b). Sayyari and Mirarab (2016b) have shown that despite assumptions, ASTRAL CU branch lengths are accurate *when* a sufficiently large number of true gene trees are used (Fig 7d). ASTRAL CU branch lengths, however, suffer from two issues, which limit their usability in practice. An obvious shortcoming is that terminal branches for single-individual species lack an estimated length, limiting the utility of the computed branch lengths.

The second difficulty is the lack of robustness to gene tree error. Gene tree error tends to increase gene tree discordance; as ASTRAL branch lengths are only a function of discordance (and nothing else), gene tree error results in under-estimation of branch lengths. For example, Sayyari and Mirarab (2016b) observed lengths that were close to an order of magnitude underestimated for the least strong gene trees they tested (Fig 7d). In conditions where gene trees had even moderately high resolution (e.g., 60% mean BS corresponding to 1500bp genes), estimated lengths were relatively accurate.

### 5.3 Branch Support using Local Posterior Probability (localPP)

The traditional method for obtaining branch support for species trees is multi-locus bootstrapping (MLBS) (Seo, 2008). MLBS first bootstraps gene trees and then runs the summary method in replicate runs using bootstrapped gene trees as input. In the end, support values are computed by counting how often a branch appears in this collection of bootstrapped species trees. The MLBS method has turned out to have severe limitations (e.g., Simmons et al., 2019). For example, Mirarab et al. (2016) showed in simulations that MLBS tends to both overestimate and underestimate support. The heart of the problem is the increased discordance among bootstrapped gene trees, compared to ML gene trees (which themselves tend to overestimate conflict). Recall that each locus can be relatively short and lacking in informative sites, a condition that is not conducive to accurate bootstrapping (Felsenstein, 1985). To address limitations of MLBS, Sayyari and Mirarab (2016b) designed a way to compute branch support for ASTRAL trees, without bootstrapping.

If a branch in an estimated species tree is correct, for every quartet selected around it, the probability of observing the species tree should be at least 1/3. Thus, asking whether a branch is correct is akin to asking whether the true probability of all quartet trees around the branch appearing in gene trees is higher than 1/3. Given the distribution of quartet frequencies in an error-free sample of gene trees, the Bayes's rule can be used to compute the probability that the probability of observing the quartet tree is above 1/3. This probability will give use the posterior probability (PP) of the branch being correct, given the input gene trees. For four species, the PP can be computed analytically (with a choice of a convenient prior distribution on branch lengths, which corresponds to assuming the species tree is generated under the Yule model).

Exact calculation of PP is difficult for more than four species. However, with several simplifying assumptions, we can fall back to the case of four species. A main assumption is *locality*: in computing the support for a branch, we assume that all four branches around it are correct, enabling us to only consider three rearrangements around the branch (Fig. 7b). Because of this assumption, this measure of support is called localPP. Sayyari and Mirarab (2016b) show in simulations that localPP is more accurate than MLBS when gene trees are inferred using short loci (i.e., from gene trees with relatively high error) and matches MLBS when gene trees are highly accurate (Fig. 7e). Moreover, localPP does not require bootstrapping gene trees and therefore is much faster than MLBS. Since its introduction, localPP has been adopted by many studies.

## 6 Followup Analyses and Visualization

Several analyses can follow the species tree inference to gain additional insights. These followup analyses can be performed on any tree, whether computed by ASTRAL or not. ASTRAL can perform the following analyses and compute localPP for any tree given to it using the `-q` option.

### 6.1 Testing for Polytomies

A central question in systematics is whether a particular branch is resolvable given the present data or more broadly, at all. Branches that cannot be resolved are removed, resulting in polytomies. Polytomies are

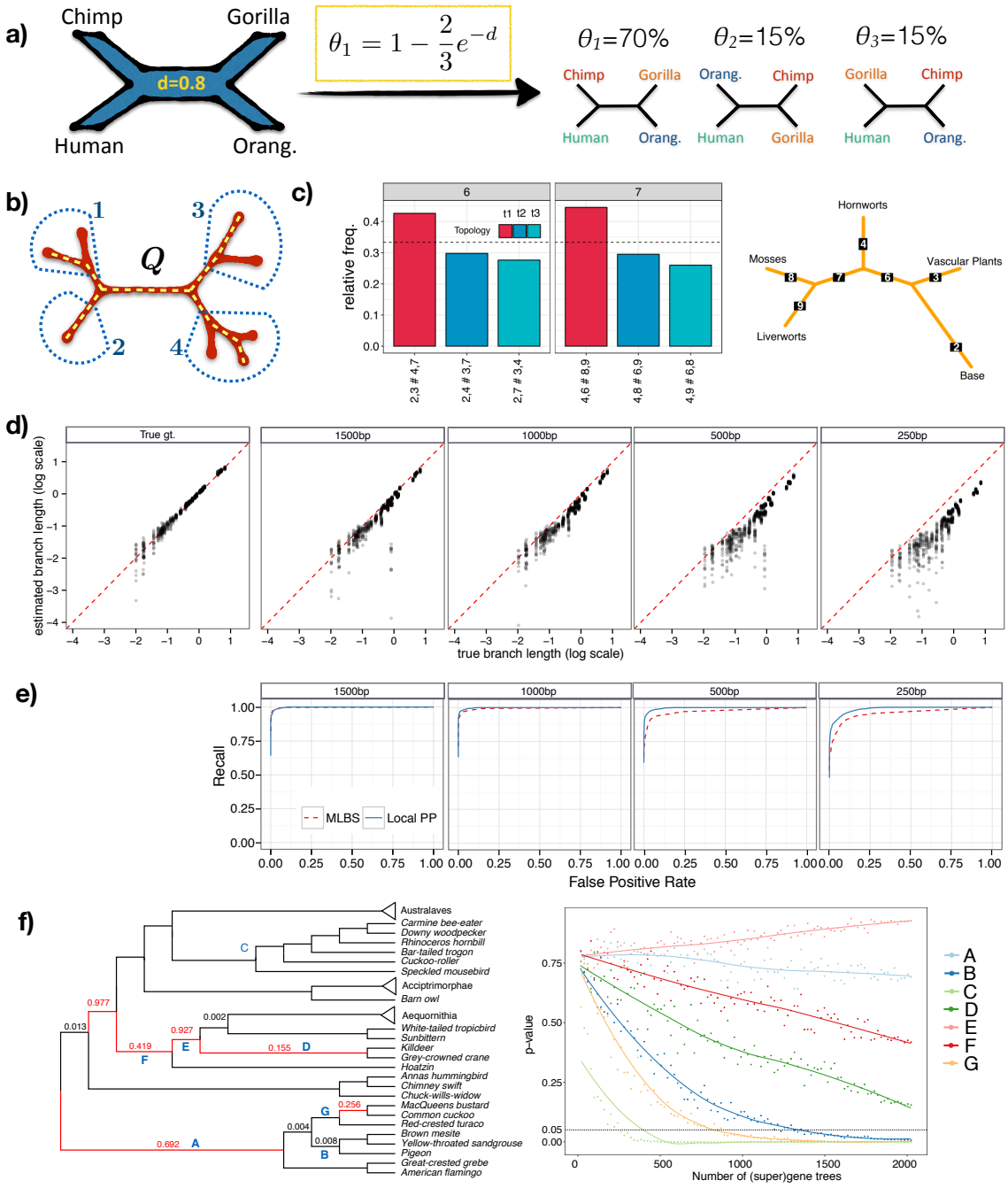


**Figure 7: Per-branch quartet frequencies**. (a) The distribution of quartet gene tree topologies ($\theta_1 \dots \theta_3$) as a function of the CU branch length $d$. (b) Branch $Q = 12|34$ can be rearranged in two ways: $13|24$ and $14|24$. Choosing a leaf from each group $1 \dots 4$ gives a quartet *around* branch $Q$ (e.g., yellow dashed). (c) Visualizing quartet support using DiscoVista. (d) Accuracy of ASTRAL branch length in simulations (Sayyari and Mirarab, 2016b) on true and estimated gene trees (alignments length: 250bp to 1500bp). (e) Accuracy of localPP in the same simulations. For a threshold $p$, branches of the ASTRAL tree and the two rearrangements around each of them ($3(n-3)$ in total) are categorized into true positives (correct, support $\geq p$), true negatives (correct, support $< p$), false positives (incorrect, support $\geq p$), false negatives (incorrect, support $< p$), and recall and false positive rates are computed. Exploring $p$ produces the ROC curve where a higher line means more true positives for each false positive rate. LocalPP dominates MLBS. (f) Testing the polytomy null hypothesis for 7 recalcitrant branches (A-F) in the ASTRAL species tree computed from 2022 supergene trees of an avian dataset. With more genes, p-values drop for some branches but not for others (Sayyari and Mirarab, 2018). Branch E (position of enigmatic species Hoatzin) seems to be best explained by a hard polytomy.

called *hard* when the multifurcation is biological, and no amount of data should be able to resolve it, and or *soft* when the present data cannot resolve the relationships due to lack of power. Sayyari and Mirarab (2018) introduced a frequentist approach for testing the following null hypothesis: a given branch in the tree has length zero and should be contracted. Note that under the null hypothesis, the quartet frequencies around a branch are 1/3 for all three resolutions around the branch. A simple Chi-squared test can be used to test this null hypothesis. However, the failure to reject the null is not the acceptance of null; thus, when the null hypothesis is not rejected, we replace the branch with a polytomy, but we cannot say if it is a soft or a hard polytomy. Sayyari and Mirarab (2018) showed in simulations that the method successfully controls the false positive rate and is powerful in rejecting the null, given sufficient genes. The method also showed intriguing patterns when applied to the base on Neoaves (Fig. 7f), indicating that some (but not all) recalcitrant branches should perhaps be replaced with polytomies. The test for polytomies is implemented in ASTRAL and can be invoked using the option `-t 12` (see ASTRAL documentations).

### 6.2 Per Branch Quartet Support (Measure of Discordance)

Phylogenomic studies are often interested in the amount of discordance *per branch* of the species tree. The computation of branch length and localPP in ASTRAL is contingent on first computing, for each branch, its quartet support. Note that around each branch of an unrooted tree, many quartets are defined (Fig. 7b) that map to that branch and only that branch (there are $n-3$ to $(n/4)^4$ such quartets per branch).

The quartet support of a species tree branch with respect to gene trees (with no missing leaves) is the proportion of times that quartets around the branch are resolved identically to the species tree in the gene trees. When there are missing data, the definition becomes more tricky because several normalization schemes become possible. We use the following definition. First, we discard all genes that do not fully include *any* of the quartets around the branch. Then, for each gene, we compute what portion of its quartets supports each topology, and we compute the mean of these values over all genes. Thus, we get a number between 0 and 1 for each quartet topology around the branch.

The quartet score of a branch can be used as a measure of discordance around the branch. ASTRAL can output quartet scores for all branches of a given tree (`-t 1` and `-t 8`). Several points are worth mentioning:

- Values close to 1/3 point to very high levels of discordance. However, given a large number of gene trees, quartet scores that have relatively small divergences from 1/3 (e.g., 40%) can lead to high localPP. Also, remember that discordance includes both true discordance and the effects of gene tree error.
- Under ILS, one expects quartet scores of the second and third topologies to be identical. When the two frequencies diverge substantially, ILS assumptions are violated, either during gene tree estimation (e.g., due long branch attraction) or because other biological sources of discordance (e.g., paralogy) also exist. Both cases warrant extra caution in interpretation.
- In rare occasions, a branch of the ASTRAL tree has a quartet score below 1/3. This can happen for several reasons, but in all cases, the branch should be considered unresolved (will have a localPP of 0).

**DisvoVista.** The best way to interrogate the quartet scores produced by ASTRAL is to visualize them for branches of interest. Sayyari et al. (2018) have developed a tool called DiscoVista to visualize quartet scores around important branches (and also produce other visualizations of discordance). For example, in Figure 7c, we summarize quartet scores around two focal branches of the plant tree from Wickett et al. (2014); helpfully, DiscoVista collapses large groups into individual nodes for better visualization.

## 7 Conclusion

I reviewed the relatively substantial body of knowledge available in the literature on the ASTRAL method, including best practices for using it. I hope the reader comes away with these messages:

- ASTRAL is a statistically consistent method of species tree estimation given inputs sampled with no error under the MSC model.

- ASTRAL is extremely scalable and can analyze many thousands of species.
- ASTRAL, like other summary methods, can be sensitive to gene tree estimation error, a problem that is alleviated by not eliminated if *extremely* low support branches in gene trees are contracted.
- ASTRAL has performed well in terms of accuracy in simulation analyses compared to other summary methods. The performance with respect to concatenation depends on the amount of discordance and phylogenetic signal in input genes.
- ASTRAL's native localPP is a better method of computing support than multi-locus bootstrapping.
- On real data, care is needed for preparing the input to ASTRAL, in particular, to avoid negative impacts of fragmentary data, but extensive gene tree filtering is not recommended.
- ASTRAL is statistically inconsistent under models of gene evolution that include gene flow. However, it has shown high accuracy under simulations with high levels of (randomly distributed) HGT.

# References


Aberer, A. J., Krompass, D., and Stamatakis, A. (2013). Pruning rogue taxa improves phylogenetic accuracy: An efficient algorithm and webservice. *Systematic Biology*, 62(1):162–166.

Allman, E. S., Degnan, J. H., and Rhodes, J. A. (2011). Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *J. Math. Biol.*, 62:833–862.

Allman, E. S. and Rhodes, J. A. (2003). Phylogenetic invariants for the general Markov model of sequence mutation. *Mathematical biosciences*, 186(2):113–144.

Ané, C., Larget, B. R., Baum, D. A., Smith, S. D., and Rokas, A. (2007). Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution*, 24(2):412–426.

Arcila, D., Ortí, G., Vari, R., Armbruster, J. W., Stiassny, M. L. J., Ko, K. D., Sabaj, M. H., Lundberg, J., Revell, L. J., and Betancur-R., R. (2017). Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. *Nature Ecology & Evolution*, 1(January):0020.

Avni, E., Cohen, R., and Snir, S. (2015). Weighted Quartets Phylogenetics. *Systematic Biology*, 64(2):233–242.

Ballesteros, J. A. and Sharma, P. P. (2019). A Critical Appraisal of the Placement of Xiphosura (Chelicerata) with Account of Known Sources of Phylogenetic Error. *Systematic Biology*, pages 1–62.

Bayzid, M. S., Mirarab, S., Boussau, B., and Warnow, T. (2015). Weighted Statistical Binning: Enabling Statistically Consistent Genome-Scale Phylogenetic Analyses. *PLoS ONE*, 10(6):e0129183.

Bayzid, M. S. and Warnow, T. (2013). Naive binning improves phylogenomic analyses. *Bioinformatics*, 29(18):2277–84.

Blom, M. P. K., Bragg, J. G., Potter, S., and Moritz, C. (2017). Accounting for Uncertainty in Gene Tree Estimation: Summary-Coalescent Species Tree Inference in a Challenging Radiation of Australian Lizards. *Systematic Biology*, 66(3):352–366.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., and Roychoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution*, 29(8):1917–1932.

Bryant, D. and Steel, M. (2001). Constructing Optimal Trees from Quartets. *Journal of Algorithms*, 38(1):237–259.

Chaudhary, R., Burleigh, J. G., and Fernández-Baca, D. (2013). Inferring species trees from incongruent multi-copy gene trees using the Robinson-Foulds distance. *Algorithms for Molecular Biology*, 8:28.

Chen, M.-Y., Liang, D., and Zhang, P. (2015). Selecting Question-Specific Genes to Reduce Incongruence in Phylogenomics: A Case Study of Jawed Vertebrate Backbone Phylogeny. *Systematic Biology*, 64(6):1104–1120.

Chifman, J. and Kubatko, L. S. (2014). Quartet Inference from SNP Data Under the Coalescent Model. *Bioinformatics*, 30(23):3317–3324.

Davidson, R., Vachaspati, P., Mirarab, S., and Warnow, T. (2015). Phylogenomic species tree estimation in the presence of incomplete lineage sorting and horizontal gene transfer. *BMC Genomics*, 16(Suppl 10):S1.

De Maio, N., Schlötterer, C., and Kosiol, C. (2013). Linking Great Apes Genome Evolution across Time Scales Using Polymorphism-Aware Phylogenetic Models. *Molecular Biology and Evolution*, 30(10):2249–2262.

Degnan, J. H. and Rosenberg, N. A. (2006). Discordance of Species Trees with Their Most Likely Gene Trees. *PLoS Genetics*, 2(5).

Degnan, J. H. and Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology and Evolution*, 24(6):332–340.

Degnan, J. H. and Salter, L. A. (2005). Gene tree distributions under the coalescent process. *Evolution*, 59(1):24–37.

Edwards, S. V., Xi, Z., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., Zhong, B., Wu, S., Lemmon, E. M., Lemmon, A. R., Leaché, A. D., Liu, L., and Davis, C. C. (2016). Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Molecular Phylogenetics and Evolution*, 94:447–462.

Felsenstein, J. (1985). Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution*, 39(4):783–791.

Gatesy, J., Sloan, D., Warren, J. M., Baker, R. H., Simmons, M. P., and Springer, M. S. (2018). Partitioned coalescence support reveals biases in species-tree methods and detects gene trees that determine phylogenomic conflicts. *bioRxiv*, page 461699.

Gatesy, J. and Springer, M. S. (2014). Phylogenetic Analysis at Deep Timescales: Unreliable Gene Trees, Bypassed Hidden Support, and the Coalescence/Concatalescence Conundrum. *Molecular phylogenetics and evolution*, 80:231–266.

Giarla, T. C. and Esselstyn, J. A. (2015). The Challenges of Resolving a Rapid, Recent Radiation: Empirical and Simulated Phylogenomics of Philippine Shrews. *Systematic Biology*, 64(5):727–740.

Heled, J. and Drummond, A. J. (2010). Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, 27(3):570–580.

Holmes, S. (2003). Statistics for phylogenetic trees. *Theoretical Population Biology*, 63(1):17–32.

Hosner, P. A., Faircloth, B. C., Glenn, T. C., Braun, E. L., and Kimball, R. T. (2016). Avoiding Missing Data Biases in Phylogenomic Inference: An Empirical Study in the Landfowl (Aves: Galliformes). *Molecular Biology and Evolution*, 33(4):1110–1125.

Huang, C.-H., Sun, R., Hu, Y., Zeng, L., Zhang, N., Cai, L., Zhang, Q., Koch, M. A., Al-Shehbaz, I., and Edger, P. P. (2016). Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Molecular biology and evolution*, 33(2):394–412.

Huelsenbeck, J. P. and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755.

Jeffroy, O., Brinkmann, H., Delsuc, F., and Philippe, H. (2006). Phylogenomics: the beginning of incongruence? *Trends in Genetics*, 22(4):225–231.

Jiang, T., Kearney, P., and Li, M. (2001). A Polynomial Time Approximation Scheme for Inferring Evolutionary Trees from Quartet Topologies and Its Application. *SIAM Journal on Computing*, 30(6):1942–1961.

Kane, D. and Tao, T. (2017). A bound on partitioning clusters. *Electr. J. Comb.*, 24:P2.31.

Kubatko, L. S. and Degnan, J. H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*, 56:17–24.

Lafond, M. and Scornavacca, C. (2019). On the Weighted Quartet Consensus problem. *Theoretical Computer Science*, 769:1–17.

Lanier, H. C. and Knowles, L. L. (2015). Applying species-tree analyses to deep phylogenetic histories: Challenges and potential suggested from a survey of empirical phylogenetic studies. *Molecular Phylogenetics and Evolution*, 83:191–199.

Larget, B. R., Kotha, S. K., Dewey, C. N., and Ané, C. (2010). BUCKy: Gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics*, 26(22):2910–2911.

Liu, K., Linder, C. R., and Warnow, T. (2011). RAxML and FastTree: Comparing Two Methods for Large-Scale Maximum Likelihood Phylogeny Estimation. *PLoS ONE*, 6(11):e27731.

Liu, L. (2008). BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics*, 24(21):2542–2543.

Liu, L., Xi, Z., Wu, S., Davis, C. C., and Edwards, S. V. (2015). Estimating phylogenetic trees from genome-scale data. *Annals of the New York Academy of Sciences*, 1360(1):36–53.

Liu, L. and Yu, L. (2011). Estimating Species Trees from Unrooted Gene Trees. *Systematic Biology*, 60(5):661–667.

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

Liu, L., Yu, L., Pearl, D. K., and Edwards, S. V. (2009). Estimating species phylogenies using coalescence times among sequences. *Systematic Biology*, 58(5):468–477.

Longo, S., Faircloth, B., Meyer, A., Westneat, M., Alfaro, M., and Wainwright, P. (2017). Phylogenomic analysis of a rapid radiation of misfit fishes (Syngnathiformes) using ultraconserved elements. *Molecular Phylogenetics and Evolution*, 113:33–48.

Ma, B., Xin, L., and Zhang, K. (2008). A new quartet approach for reconstructing phylogenetic trees: quartet joining method. *Journal of Combinatorial Optimization*, 16(3):293–306.

Maddison, W. P. (1997). Gene Trees in Species Trees. *Systematic Biology*, 46(3):523–536.

Mai, U. and Mirarab, S. (2018). TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics*, 19(S5):272.

Meiklejohn, K. A., Faircloth, B. C., Glenn, T. C., Kimball, R. T., and Braun, E. L. (2016). Analysis of a Rapid Evolutionary Radiation Using Ultraconserved Elements: Evidence for a Bias in Some Multispecies Coalescent Methods. *Systematic Biology*, 65(4):612–627.

Mirarab, S. (2017). Phylogenomics: Constrained gene tree inference. *Nature Ecology & Evolution*, 1(2):0056.

Mirarab, S., Bayzid, M. S., Boussau, B., and Warnow, T. (2014a). Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science*, 346(6215):1250463–1250463.

Mirarab, S., Bayzid, M. S., and Warnow, T. (2016). Evaluating Summary Methods for Multilocus Species Tree Estimation in the Presence of Incomplete Lineage Sorting. *Systematic Biology*, 65(3):366–380.

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

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

Molloy, E. and Warnow, T. (2019). Large-scale Species Tree Estimation. *ArXiv preprint: 1904.02600*.

Molloy, E. K. and Warnow, T. (2018). To Include or Not to Include: The Impact of Gene Filtering on Species Tree Estimation Methods. *Systematic Biology*, 67(2):285–303.

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

Nute, M., Chou, J., Molloy, E. K., and Warnow, T. (2018). The performance of coalescent-based species tree estimation methods under models of missing data. *BMC Genomics*, 19(S5):286.

Ogilvie, H. A., Bouckaert, R. R., and Drummond, A. J. (2017). StarBEAST2 Brings Faster Species Tree Inference and Accurate Estimates of Substitution Rates. *Molecular Biology and Evolution*, 34(8):2101–2114.

Ogilvie, H. A., Heled, J., Xie, D., and Drummond, A. J. (2016). Computational Performance and Statistical Accuracy of *BEAST and Comparisons with Other Methods. *Systematic Biology*, 65(3):381–396.

Pamilo, P. and Nei, M. (1988). Relationships between gene trees and species trees. *Molecular biology and evolution*, 5(5):568–583.

Patel, S. (2013). Error in Phylogenetic Estimation for Bushes in the Tree of Life. *Journal of Phylogenetics & Evolutionary Biology*, 01(02):110.

Philippe, H., Vienne, D. M. d., Ranwez, V., Roure, B., Baurain, D., and Delsuc, F. (2017). Pitfalls in supermatrix phylogenomics. *European Journal of Taxonomy*.

Phillips, M. J., Delsuc, F., and Penny, D. (2004). Genome-scale phylogeny and the detection of systematic biases. *Molecular Biology and Evolution*.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree-2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, 5(3):e9490.

Rabiee, M., Sayyari, E., and Mirarab, S. (2019). Multi-allele species reconstruction using ASTRAL. *Molecular Phylogenetics and Evolution*, 130:286–296.

Ragan, M. a. (1992). Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution*, 1(1):53–58.

Robinson, D. and Foulds, L. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2):131–147.

Roch, S., Nute, M., and Warnow, T. (2019). Long-Branch Attraction in Species Tree Estimation: Inconsistency of Partitioned Likelihood and Topology-Based Summary Methods. *Systematic Biology*, 68(2):281–297.

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

Roch, S. and Steel, M. (2015). Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theoretical Population Biology*, 100:56–62.

Sayyari, E. and Mirarab, S. (2016a). Anchoring quartet-based phylogenetic distances and applications to species tree reconstruction. *BMC Genomics*, 17(S10):101–113.

Sayyari, E. and Mirarab, S. (2016b). Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. *Molecular biology and evolution*, 33(7):1654–1668.

Sayyari, E. and Mirarab, S. (2018). Testing for Polytomies in Phylogenetic Species Trees Using Quartet Frequencies. *Genes*, 9(3):132.

Sayyari, E., Whitfield, J. B., and Mirarab, S. (2017). Fragmentary Gene Sequences Negatively Impact Gene Tree and Species Tree Reconstruction. *Molecular Biology and Evolution*, 34(12):3279–3291.

Sayyari, E., Whitfield, J. B., and Mirarab, S. (2018). DiscoVista: Interpretable visualizations of gene tree discordance. *Molecular Phylogenetics and Evolution*, 122:110–115.

Seo, T.-K. (2008). Calculating Bootstrap Probabilities of Phylogeny Using Multilocus Sequence Data. *Molecular Biology and Evolution*, 25(5):960–971.

Shekhar, S., Roch, S., and Mirarab, S. (2018). Species Tree Estimation Using ASTRAL: How Many Genes Are Enough? *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(5):1738–1747.

Shen, X.-x., Hittinger, C. T., and Rokas, A. (2017). Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology & Evolution*, 1(5):0126.

Simmons, M. P., Sloan, D. B., and Gatesy, J. (2016). The effects of subsampling gene trees on coalescent methods applied to ancient divergences. *Molecular Phylogenetics and Evolution*, 97(January):76–89.

Simmons, M. P., Sloan, D. B., Springer, M. S., and Gatesy, J. (2019). Gene-wise resampling outperforms site-wise resampling in phylogenetic coalescence analyses. *Molecular Phylogenetics and Evolution*, 131(November 2018):80–92.

Solís-Lemus, C., Yang, M., and Ané, C. (2016). Inconsistency of Species Tree Methods under Gene Flow. *Systematic Biology*, 65(5):843–851.

Springer, M. S. and Gatesy, J. (2014). Land plant origins and coalescence confusion. *Trends in plant science*, 19(5):267–9.

Springer, M. S. and Gatesy, J. (2016). The gene tree delusion. *Molecular Phylogenetics and Evolution*, 94(Part A):1–33.

Springer, M. S. and Gatesy, J. (2017). On the importance of homology in the age of phylogenomics. *Systematics and Biodiversity*, pages 1–19.

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

Steel, M. (1992). The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification*, 9(1):91–116.

Streicher, J. W., Schulte, J. A., and Wiens, J. J. (2016). How Should Genes and Taxa be Sampled for Phylogenomic Analyses with Missing Data? An Empirical Study in Iguanian Lizards. *Systematic Biology*, 65(1):128–145.

Szöllõsi, G. J., Tannier, E., Daubin, V., and Boussau, B. (2014). The inference of gene trees with species trees. *Systematic Biology*, 64(1):e42–e62.

Vachaspati, P. and Warnow, T. (2015). ASTRID: Accurate Species TRees from Internode Distances. *BMC Genomics*, 16(Suppl 10):S3.

Vachaspati, P. and Warnow, T. (2017). FastRFS: fast and accurate Robinson-Foulds Supertrees using constrained exact optimization. *Bioinformatics*, 33(5):631–639.

Westover, K. M., Rusinko, J. P., Hoin, J., and Neal, M. (2013). Rogue taxa phenomenon: A biological companion to simulation analysis. *Molecular Phylogenetics and Evolution*, 69(1):1–3.

Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E. J., Matasci, N., Ayyampalayam, S., Barker, M. S., Burleigh, J. G., Gitzendanner, M. A., Ruhfel, B. R., Wafula, E., Der, J. P., Graham, S. W., Mathews, S., Melkonian, M., Soltis, D. E., Soltis, P. S., Miles, N. W., Rothfels, C. J., Pokorny, L., Shaw, A. J., DeGironimo, L., Stevenson, D. W., Surek, B., Villarreal, J. C., Roure, B., Philippe, H., DePamphilis, C. W., Chen, T., Deyholos, M. K., Baucom, R. S., Kutchan, T. M., Augustin, M. M., Wang, J. J., Zhang, Y., Tian, Z., Yan, Z., Wu, X., Sun, X., Wong, G. K.-S., and Leebens-Mack, J. J. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences*, 111(45):4859–4868.

Wu, Y. (2012). Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution*, 66(3):763–775.

Yin, J., Zhang, C., and Mirarab, S. (2019). ASTRAL-MP: scaling ASTRAL to very large datasets using randomization and parallelization. *Bioinformatics*, btz211.

Zhang, C., Rabiee, M., Sayyari, E., and Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, 19(S6):153.

Zimmermann, T., Mirarab, S., and Warnow, T. (2014). BBCA: Improving the scalability of *BEAST using random binning. *BMC genomics*, 15(Suppl 6):S11.