ANNUAL REVIEWS

Annual Review of Ecology, Evolution, and Systematics

Multispecies Coalescent: Theory and Applications in Phylogenetics

Siavash Mirarab,[1] Luay Nakhleh,[2] and Tandy Warnow[3]

[1]Electrical and Computer Engineering Department, University of California, San Diego, La Jolla, California 92093, USA; email: smirarab@ucsd.edu

[2]Department of Computer Science, Rice University, Houston, Texas 77005, USA; email: nakhleh@rice.edu

[3]Department of Computer Science, University of Illinois Urbana-Champaign, Urbana, Illinois 61801, USA; email: warnow@illinois.edu

Annu. Rev. Ecol. Evol. Syst. 2021. 52:247–68

First published as a Review in Advance on August 20, 2021

The *Annual Review of Ecology, Evolution, and Systematics* is online at ecolsys.annualreviews.org

https://doi.org/10.1146/annurev-ecolsys-012121-095340

Keywords

multi-species coalescent, incomplete lineage sorting, phylogenomics

Abstract

Species tree estimation is a basic part of many biological research projects, ranging from answering basic evolutionary questions (e.g., how did a group of species adapt to their environments?) to addressing questions in functional biology. Yet, species tree estimation is very challenging, due to processes such as incomplete lineage sorting, gene duplication and loss, horizontal gene transfer, and hybridization, which can make gene trees differ from each other and from the overall evolutionary history of the species. Over the last 10–20 years, there has been tremendous growth in methods and mathematical theory for estimating species trees and phylogenetic networks, and some of these methods are now in wide use. In this survey, we provide an overview of the current state of the art, identify the limitations of existing methods and theory, and propose additional research problems and directions.

1. INTRODUCTION

NP-hard: a problem is said to be NP-hard if every other problem in the class NP reduces to it; it is generally believed that no NP-hard problem can be solved exactly in polynomial time

Statistically identifiable: a parameter is statistically identifiable if the distribution over the data defined by the model uniquely determines the parameter

A species tree provides a context in which many biological questions can be addressed. For example, when the species tree is known, it is possible to learn how a given gene evolved through a sequence of duplications and losses (and in some cases horizontal transfers), estimate divergence dates, detect and understand adaptation, etc. Methods for these analyses typically depend on reconciling gene trees to the species tree, and some of these methods also enable improved gene tree estimation (for an entry into this literature, see Christensen et al. 2019, Doyon et al. 2011, Hahn 2007, Nakhleh 2013).

However, despite the increasing availability of genomic sequence data from across the tree of life, species tree estimation remains challenging for a variety of computational and statistical reasons. From the computational side, the most accurate approaches use statistical methods based on likelihood and are generally computationally intensive on large data sets. For example, maximum likelihood tree estimation methods, such as RAxML (Stamatakis 2014), attempt to find model trees that optimize the likelihood criterion, but this is a nondeterministic polynomial-time (NP)-hard optimization problem (Roch 2006), and hence, these methods employ local search heuristics to search for optimal trees [note that NP-hardness has the practical consequence that exact solutions are not guaranteed for polynomial-time methods (Garey & Johnson 1979)]. Another common type of method uses Bayesian MCMC (Markov chain Monte Carlo) to sample from the posterior distribution and needs to run until the chains converge to the stationary distribution. In contrast, while polynomial-time methods are also in use [e.g., neighbor joining (Saitou & Nei 1987)], these methods are not generally as accurate as likelihood-based methods.

Species tree estimation is also challenging because genome-scale evolution is very heterogeneous, with different genes having different evolutionary histories. For example, as illustrated in **Figure 1*a***, genes can evolve under one or multiple processes, including incomplete lineage sorting (ILS), gene duplication and loss (GDL), and horizontal gene transfer (HGT), that make their trees different from the species tree (Degnan & Rosenberg 2006, Maddison 1997).

In response to documented evidence of gene tree discordance (Burleigh et al. 2010, Jarvis et al. 2014, Sanderson & McMahon 2007, Smith et al. 2015) and reduced accuracy of standard species tree estimation methods (such as concatenated maximum likelihood) in the presence of discordance (Kubatko & Degnan 2007), substantial effort has been put toward developing theory and algorithms for estimating species trees and phylogenetic networks. Much of the theory is based on hierarchical statistical models where genes evolve within a species tree/network and then sequences evolve down the gene trees (**Figure 1*b***), thus treating phylogeny estimation (either from the sequences or from the estimated gene trees) as a statistical estimation problem. In all of the work we describe hereafter, it is assumed that there is no recombination within a locus (justifying a single gene tree per locus) and there is free recombination between loci (justifying independence among loci). A key statistical question is whether the species tree topology (generally in its unrooted form) is statistically identifiable from the probability distribution it defines over gene trees. Furthermore, for those models under which species trees are indeed identifiable, we may ask whether a given method for estimating the species tree is statistically consistent, which asserts that as the amount of data increases, the correct species tree is inferred with probability converging to 1.

There has been substantial development of theory addressing species tree estimation when the source of gene tree discordance is exclusively ILS, a population-level process modeled by the multispecies coalescent (MSC) (Maddison 1997, Rannala & Yang 2003, Takahata 1989). The MSC extends Kingman's coalescent process (Kingman 1982) to multiple species by assigning a fixed-size population to each branch of the species tree. Within the species tree, the root is at the top and the leaves at the bottom, which allows us to refer to the top and bottom of a branch. Given this, Kingman's coalescent process is run on each branch, and all lineages surviving at the top of

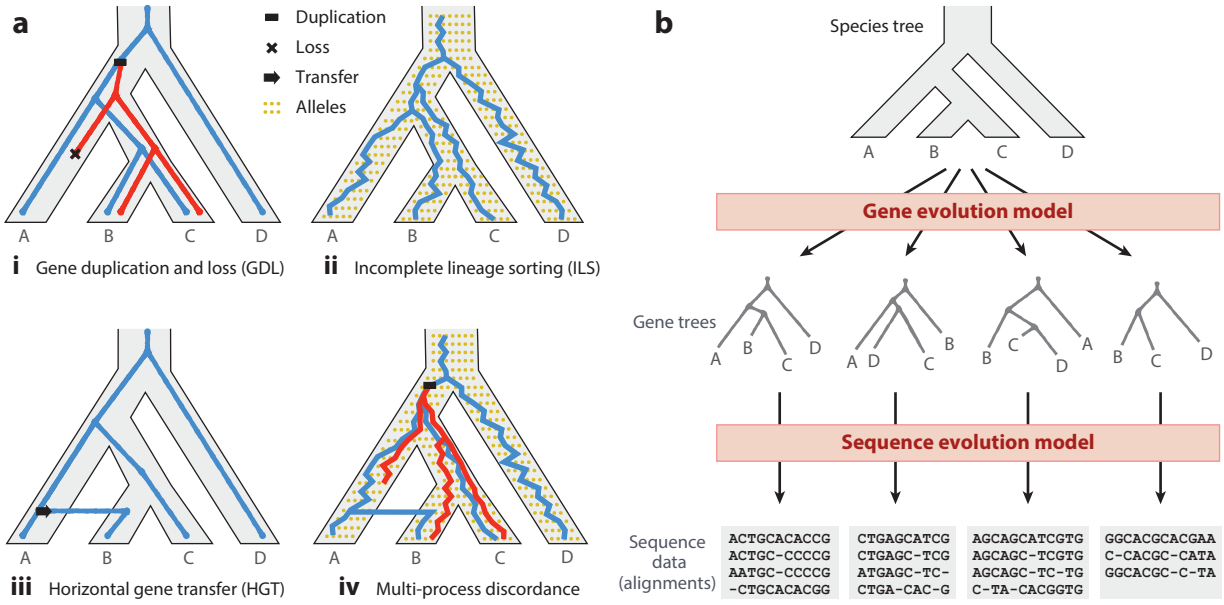

Figure 1

(*a*) A gene tree may differ from the species tree due to (*i*) gene duplication and loss, (*ii*) incomplete lineage sorting, (*iii*) horizontal gene transfer, or (*iv*) a combination of these processes. (*b*) A hierarchical model of evolution where gene trees evolve within or across the branches of a species tree according to processes illustrated in panel *a* and molecular sequences of individual loci evolve down their respective trees. A–D represent example species; two copies of a gene are shown in red and blue.

the branch (i.e., closer to the root) are moved to the bottom of the branch corresponding to the parent population. Lineages from two different species can first coalesce in the population of their most recent common ancestor, but if they fail to coalesce there, they coalesce further up the tree (creating deep coalescence).

Research over the last 10 years has established that the standard approach of concatenating the alignments and then running maximum likelihood on the combined superalignment is not statistically consistent under the MSC and can be positively misleading (Roch & Steel 2015). In response, new methods have been developed that are statistically consistent under the MSC, and several of these have good scalability and accuracy on simulated data sets. While there is still clear room for improvement, there are now several popular methods in wide use that have both good empirical accuracy and also strong theoretical guarantees when heterogeneity is due solely to ILS.

However, as described above, discordance between gene trees in biological data sets may also be due to other causes, including GDL, HGT, hybridization, etc. Estimation of species trees under any one of these processes presents challenges, and estimation under two or more is particularly difficult. Compared to species tree estimation under ILS-only scenarios, much less is understood about how to estimate species trees when GDL is present or when two or more processes that create gene tree discordance are present. Furthermore, under some scenarios (e.g., hybridization), a species tree is an inadequate model of the evolutionary history of a data set; in such cases, instead of a phylogenetic tree, we need a phylogenetic network.

In this review, we describe approaches for estimating species trees and phylogenetic networks that address gene tree discord due to ILS and multiple other causes. Rather than providing a survey of the different methods that can be used, we discuss the different techniques used in these methods and the computational and empirical challenges they address. Furthermore, because data sets for

Positively misleading: a phylogeny estimation method is positively misleading under a model if it converges to a tree other than the true tree with probability converging to 1 as the amount of data increases

Empirical accuracy: phylogeny estimation methods are tested with respect to accuracy on simulated data sets and on biological benchmark data sets; a method has good empirical accuracy when the estimation error is considered sufficiently low

species tree estimation are increasing in size, with hundreds to many thousands of loci and up to many hundreds or thousands of species (Jarvis et al. 2014, One Thousand Plant Transcriptomes Initiative 2019, Tarver et al. 2016), we focus our attention on methods that can analyze large data sets. Section 2 discusses methods for species tree estimation, mainly (but not exclusively) focused on methods that are proven to be statistically consistent under the MSC. In Section 3, we turn to methods for phylogenetic network estimation, specifically addressing estimation in the presence of hybridization and potentially also ILS. We finish in Section 4 with a discussion of lessons learned, remaining issues that need to be addressed, and thoughts about the future of phylogenomic estimation.

This review provides a theoretical framework within which to understand methods that attempt to estimate species trees and phylogenetic networks. However, theory is not everything when it comes to understanding methods. For example, the relative accuracy of concatenation analyses, especially when performed using good maximum likelihood methods, and the better methods that address ILS is mixed—even when restricted to data sets in which the only source of gene tree heterogeneity is ILS—and seems to depend on the model conditions (e.g., amount of gene tree heterogeneity, number of genes, etc.) (Molloy & Warnow 2018). Thus, although this review focuses on methods that explicitly address gene tree heterogeneity, we caution the reader that the relative performance of methods is hard to predict, and so careful review of the literature and examinations of both theoretical guarantees and empirical performance are needed to choose between methods.

2. SPECIES TREE ESTIMATION METHODS

2.1. Overview

There are essentially three types of methods for species tree estimation that explicitly address ILS and have been proven statistically consistent under the MSC. The input in all of these cases is a multi-locus data set, where for each locus we assume we have a multiple sequence alignment. The three categories are as follows:

- Summary methods, which operate by first estimating gene trees (one for each locus) and then using the information in these gene trees to estimate the species trees. The most well-known such method is ASTRAL (Mirarab et al. 2014b).
- Site-based methods, which calculate small trees (typically unrooted quartet trees or rooted triplet trees) from the site patterns and then combine these small trees into a tree on the full data set. The most well-known such method is SVDquartets (Chifman & Kubatko 2014), which is available through PAUP* (Swofford 2002).
- Coestimation methods, which coestimate the species tree and the set of gene trees. The most well-known such methods are StarBEAST (Heled & Drummond 2010) and its improved version, StarBEAST2 (Ogilvie et al. 2017).

While there are statistically consistent examples of each type of method, from an empirical standpoint, they have very different properties. We begin with the summary methods, as these are generally the fastest (although site-based methods can be even faster for small enough numbers of species), followed by the site-based methods, and then the coestimation methods, which are the slowest.

2.2. Summary Methods

Many methods that are proven statistically consistent estimators of the species tree under the MSC operate by using summary statistics in the input gene trees. For example, MP-EST (Liu et al. 2010) requires the gene trees to be rooted and then seeks a rooted species tree that maximizes

the pseudolikelihood of gene trees, which is the product of the likelihood of induced rooted triplet tree topologies in the gene trees. NJst (Liu & Yu 2011) and ASTRID (Vachaspati & Warnow 2015) operate by computing a distance matrix (the internode distance matrix) from the gene trees and then applying distance-based methods. ASTRAL (Mirarab et al. 2014b) and BUCKy-pop [the population tree in BUCKy (Larget et al. 2010)] operate by assigning scores to quartet trees for every four species and then using quartet amalgamation methods to construct the species tree from these weighted quartet trees. Yet other methods, such as GLASS (Mossel & Roch 2010), use as input not just tree topologies but also their branch lengths. However, methods that rely on branch lengths have been less accurate, perhaps because they do not account for deviations from a strict molecular clock (Degiorgio & Degnan 2013) and so are less widely used.

Thus, summary methods vary in terms of approach, with some based on NP-hard optimization criteria (and so employing heuristics that make them computationally intensive on large data sets) and others able to run in polynomial time. Moreover, many of these methods are able to analyze very large data sets, and some have shown excellent accuracy. Among these methods, the ASTRAL family of methods is now very commonly used, especially for data sets with large numbers of species. NJst and ASTRID, which are very similar in design, are not as commonly used but are among the few alternative methods that are very fast and scalable, and they have been able to provide accuracy that is competitive with ASTRAL under many model conditions.

2.3. Site-Based Methods

By design, site-based methods do not require the estimation of gene trees and can be used when only a very small number of sites per locus are available (and even when only a single site per locus is available). While several site-based methods exist (e.g., Bryant et al. 2012, Dasarathy et al. 2015, Richards & Kubatko 2020), we use SVDquartets as an example, since it is the most popular of the site-based methods.

SVDquartets uses linear algebra techniques to estimate the unrooted species tree on every set of four species, thus producing a set of quartet trees. Each quartet tree can be computed in polynomial time, so that the entire set of quartet trees also requires only polynomial time. Furthermore, the technique used in SVDquartets to estimate quartet trees is statistically consistent under the MSC (Wascher & Kubatko 2020).

Next, these quartet trees must be combined into a tree on the full data set, which requires the use of an amalgamation method (e.g., a supertree method specifically designed to assemble quartet trees). These amalgamation methods, such as Quartets MaxCut (Snir & Rao 2012), typically use heuristic search strategies that do not scale well to large data sets. [An alternative approach, based on constrained optimization using dynamic programming, is enabled in SVDquest (Vachaspati & Warnow 2018).]

Independent of the technique used to combine quartet trees, a major limitation of methods that use SVDquartets to compute quartet trees and then combine them is that the time to compute all of the quartet trees, although polynomial, increases quickly as the number of species increases. One approach to using SVDquartets on data sets with large numbers of species (or individuals, depending on the application) is to compute only a subset of the possible quartet trees. This approach improves running time, but it is not yet clear if accuracy is impacted by using only a subset of the quartet trees. As a result, the site-based methods are currently mainly suited to data sets with perhaps at most 100 species or individuals and many sites.

Despite the computational challenges, a significant advantage of site-based methods over other methods is that they do not rely directly (as summary methods do) or indirectly (as coestimation methods do) on the ability to estimate highly accurate gene trees. Given the substantial evidence

that summary methods are negatively impacted by gene tree estimation error (Mirarab et al. 2014a, Molloy & Warnow 2018, Patel et al. 2013), this may be a very important point.

Statistical consistency: a method is a statistically consistent estimator of a parameter under a model if, as the amount of data increases, the error in the estimated parameter converges to 0 with probability converging to 1

2.4. Coestimation Methods

The coestimation methods, which are the most computationally intensive, typically rely on a Bayesian approach to jointly sample the posterior distribution of parameters defined by a hierarchical combination of MSC and sequence evolution models such as GTR (**Figure 1*b***). The hierarchical models include many variables, including topology and branch lengths of all gene trees and the species tree and numerical parameters such as population sizes and rates of evolution. The combination of all of these discrete and continuous variables creates a huge space that is hard to fully sample, leading to the computational challenges faced by coestimation methods. For example, StarBEAST and StarBEAST2 use MCMC to sample from the joint distribution and so need to converge before their results are considered reliable. As a result, they are generally limited to small numbers of species (perhaps as many as 50) and loci (up to 50) but can take a long time to converge at the higher end of these ranges (e.g., weeks).

2.5. Scalable Methods

Of the methods currently available, only the more computationally efficient summary methods are able to analyze data sets with hundreds of species and thousands of loci. We have selected ASTRAL and ASTRID as examples of different types of summary methods that can analyze large data sets and have shown high accuracy under a range of conditions, including missing data (where some gene trees lack species).

2.5.1. ASTRAL. Here we describe the theoretical foundation and the methods that have been developed for the ASTRAL family of methods, which began in 2014 with ASTRAL (Mirarab et al. 2014b) and now includes also ASTRAL-II (Mirarab & Warnow 2015) and ASTRAL-III (Zhang et al. 2018b) (both for single individual data sets), ASTRAL-multi (Rabiee et al. 2019) (for multi-individual data sets), and ASTRAL-MP (Yin et al. 2019) (for multi-processor usage).

The design of ASTRAL, and its proof of statistical consistency under the MSC, is based on the observation that there are no anomalous unrooted quartet trees; that is, although for every four species the three possible unrooted gene trees each have strictly positive probability, the most probable such quartet tree is topologically identical to the species tree (Allman et al. 2011, Degnan 2013). As a result, given a set of unrooted gene trees, the unrooted species tree topology that has the highest quartet support from the input gene trees is likely to be a good estimate of the true species tree topology. Furthermore, as the number of genes increases, then this estimate converges to the true species tree topology with probability converging to 1. In other words, maximizing the Quartet Support score is a statistically consistent estimator of the true species tree topology under the MSC (Mirarab et al. 2014b). We formalize this maximum quartet support species tree (MQSST) problem as follows. Given a set of k unrooted gene tree topologies, 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

$$S(T) = \sum_{g \in \mathrm{G}} |\mathrm{Q}(T) \cap \mathrm{Q}(g)|,$$

and $\mathcal{Q}(T)$ gives the set of all quartet tree topologies induced by a tree T.

Bryant & Steel (2001) were the first to study the MQSST problem, and it was proved NP-hard by Lafond & Scornavacca (2019). Although exact algorithms and heuristics exist [e.g.,

Avni et al. (2015)], these do not run in polynomial time. The key to ASTRAL's approach is that it addresses the challenge of solving MQSST by constraining the search space. Specifically, ASTRAL specifies an allowed set, X, of bipartitions (i.e., splits of the set of species into two sets) and requires that the species tree that is produced draw all of its bipartitions (one for each branch in the tree) from this set, X. As proven in Bryant & Steel (2001) and Mirarab et al. (2014b), by constraining the set of allowed bipartitions, the MQSST problem becomes solvable in polynomial time. (Specifically, these methods use a dynamic programming formulation that allows the problem to be solved exactly without explicitly exploring the entire set of feasible species trees that satisfy the constraints.) Moreover, since every bipartition of the species tree has a strictly positive probability of appearing in the gene trees, an exact solution to the constrained MQSST problem is a statistically consistent estimator of the species tree under the MSC whenever X is guaranteed to have all of the bipartitions from the input gene trees (Mirarab & Warnow 2015). All of the versions of ASTRAL are statistically consistent under the MSC because they are able to define a constraint set, X, that satisfies this condition.

Here we describe some of the changes to ASTRAL since it was first introduced in Mirarab et al. (2014b). In the first version of ASTRAL (Mirarab et al. 2014b), which we refer to here as ASTRAL-I, X was simply the set of bipartitions in the input gene trees. The second version, introduced in Mirarab & Warnow (2015) as ASTRAL-II, differed from ASTRAL-I by expanding X using several heuristics and also added heuristic methods to handle missing data and polytomies in input trees. The third version, ASTRAL-III (Zhang et al. 2018b), further tweaked these rules to ensure that the size of X was bounded from above by a constant times nk [where n is the number of species and k is the number of genes; this is denoted by $|\mathcal{X}| = O(nk)$]. In addition, Zhang et al. (2018b) proved that ASTRAL-III has a worst-case running time of $O((nk)^{2.726})$. However, Zhang et al. (2018b) found empirically that the running-time growth is closer to quadratic with the values of n and k found in typical data sets.

Several features were added to ASTRAL after the initial publication. Sayyari & Mirarab (2016) added branch length estimation (in coalescent units) and introduced a notion of support called the local posterior probability based on quartet scores. This measure was later extended to provide a test of whether species branches should be collapsed to a polytomy (Sayyari & Mirarab 2018). This polytomy test is also useful for species delimitation: When multiple individuals are available, but their delimitation is not known, a method called SODA can use ASTRAL results and the polytomy test to suggest a delimitation (Rabiee & Mirarab 2020). For cases in which species delimitation is known for multiple individuals, Rabiee et al. (2019) developed ASTRAL-multi to allow multiple individuals (or multiple alleles) to be used as inputs. The algorithm is similar, except the constraints are set up in a way that ensures all individuals of the same species are monophyletic, and the dynamic programming stops recursion as soon as all individuals below a node all belong to the same species. Heuristics for building the set X are also changed to allow multiple individuals.

While ASTRAL-III is sufficiently fast for most data, for very large data sets, one can use a vectorized, randomized, and parallelized [both for central processing units (CPUs) and graphics processing units (GPUs)] version of it called ASTRAL-MP (Yin et al. 2019). ASTRAL-MP can speed runs by two orders of magnitude compared to ASTRAL-III, and its running-time advantages are most visible for data sets with large numbers of input gene trees.

As we have noted, ASTRAL is statistically consistent when the input gene trees are sampled from the distribution defined by the MSC model. In addition, ASTRAL is statistically consistent under an identically and independently distributed model of species missing from gene trees, as well as under a clade-based model (Nute et al. 2018). Beyond these positive results, two negative results regarding the consistency of ASTRAL have also been proved. First, Roch et al. (2019) have shown that ASTRAL (or any reasonable summary methods that use gene tree topologies as input)

is statistically inconsistent when gene trees are computed using maximum likelihood from gene sequence alignments of arbitrarily bounded length, even under conditions without any gene tree incongruence. This negative result was established for a genome-wide variant of the long-branch attraction (LBA) phenomenon that makes some methods (e.g., maximum parsimony) inconsistent and also applies to partitioned concatenation using maximum likelihood (i.e., when a single tree topology is sought but the numeric parameters can differ across the loci). Practically, this negative result should lead to caution in estimating a species tree using partitioned concatenation or ASTRAL (or other summary methods) under LBA conditions and more generally when the different loci have low phylogenetic signal (e.g., very short gene alignments or very low rates of evolution). The second negative result is from Solís-Lemus et al. (2016), who proved ASTRAL can be statistically inconsistent when gene trees evolve on a phylogenetic network through a combination of ILS and gene flow. Thus, when gene flow is expected, phylogenetic network estimation methods (see Section 3) should be considered instead of ASTRAL or other species tree estimation methods.

2.5.2. Distance-based methods. Another type of summary method for species tree estimation operates by first computing a distance matrix from the input gene trees and then constructing a tree from the distance matrix. Examples of this type of approach include STAR (Liu et al. 2009) (which assumes the input gene trees are rooted) and NJst (Liu & Yu 2011) and ASTRID (Vachaspati & Warnow 2015) (which can be used with unrooted gene trees). Here we focus on NJst and ASTRID, which use the same distance matrix but then compute trees using different methods. The distance matrix is the average internode distance matrix, which uses the average number of nodes (across all of the gene trees) between a given pair of species as the distance between the species. As shown in Allman et al. (2016), this distance matrix converges, as the number of genes increases, to a matrix that is additive for the true species tree (i.e., for which there are branch lengths on the true species tree that realize the matrix of leaf-to-leaf distances). Given this distance matrix, NJst computes a tree using neighbor joining (Saitou & Nei 1987), while ASTRID computes a tree using FastME (Lefort et al. 2015). NJst and ASTRID have the same theoretical guarantees: They are statistically consistent under the MSC, run in polynomial time, and have been extended to allow for multiple individuals in each species through a minor adjustment to the internode distance matrix (and are statistically consistent in this setting).

By design, ASTRID and NJst are nearly identical, and in many cases the differences between the methods are negligible; however, FastME is typically at least as fast as neighbor joining, making ASTRID the more scalable of the two methods. Several studies have compared ASTRID and ASTRAL (e.g., Molloy & Warnow 2018, Vachaspati & Warnow 2015) and found the two methods to have close accuracy, with neither generally outperforming the other. However, ASTRID can be much faster than ASTRAL [e.g., by an order of magnitude (Vachaspati & Warnow 2015)], especially on data sets with large numbers of species and loci and also high ILS.

2.6. Species Tree Estimation under Other Models

Section 2.5 discussed species tree estimation in the presence of heterogeneity due only to ILS and presented many methods that have strong theoretical guarantees (e.g., statistical consistency under the MSC) and excellent empirical performance on both biological and simulated data sets.

However, these methods do not address other sources of gene tree heterogeneity, such as GDL, HGT, etc. Furthermore, most can handle only single-copy input gene trees. As a result of this limitation, most biological analyses restrict themselves to genes that happen to be single-copy. For some groups, like plants, this approach can discard the vast majority of the data. For example, two studies on plant transcriptomes had to discard thousands of available multi-copy

genes (Wickett et al. 2014, Leebens-Mack et al. 2019) and use only the 400–800 single-copy gene trees. As the number of genes greatly impacts accuracy (Mirarab et al. 2014b, Mirarab & Warnow 2015), it has been argued (Smith & Hahn 2020) and shown (Zhang et al. 2020) that restricting analyses to single-copy genes can dramatically decrease the accuracy of the species tree estimation.

2.6.1. Species tree estimation under gene duplication and loss. Species tree estimation in the presence of GDL presents a different set of challenges that require new mathematics and techniques. Because of gene duplication, each gene can have multiple copies of some species, so that each gene has a gene family tree. Since it is often very difficult to determine orthologous groups (i.e., a set of genes that all deviated from one another in speciation events), the inference of a species tree may require the ability to combine these gene family trees (also called MUL-trees) into a species tree. Several methods for combining these MUL-trees were developed [e.g., gene tree parsimony, which attempts to minimize the total number of GDL events (Chaudhary et al. 2010), and MulRF (Chaudhary et al. 2014), which is an adaptation of the Robinson-Foulds (Robinson & Foulds 1981) supertree problem to the MUL-tree setting]. These methods are based on heuristics for NP-hard optimization problems, and some are slow on large data sets. Furthermore, none of these methods have been proven statistically consistent under any GDL model. Statistical methods based on parametric GDL models [e.g., PHYLDOG (Boussau et al. 2013) and MixTreEM-DLRS (Ullah et al. 2015)] are highly appealing but are much more computationally intensive than MulRF and the gene tree parsimony methods.

Several recent method developments have occurred that are relevant to species tree estimation when GDL is present. ASTRAL-multi (Rabiee et al. 2019), a variant of ASTRAL designed for multiple individuals per species, was proven to be statistically consistent under an established model of GDL if gene copies are encoded as different individuals (Legried et al. 2020). Subsequently, FastMulRFS (Molloy & Warnow 2020) was developed, a method for the Robinson-Foulds Supertree problem adapted to the MUL-tree setting (Chaudhary et al. 2014) and proven to be statistically consistent under some restricted conditions (i.e., when no adversarial GDL occurs). In addition, FastMulRFS was shown to be very fast and more accurate than ASTRAL-multi (Molloy & Warnow 2020).

Another recent development is ASTRAL-Pro (Zhang et al. 2020), a new variant of ASTRAL developed specifically to address GDL and ILS. ASTRAL-Pro depends on a technique to tag the internal nodes of the gene trees as either duplication nodes or speciation nodes, and when this tagging is correct, then ASTRAL-Pro is statistically consistent under GDL models. In addition, ASTRAL-Pro has been found both to have better accuracy than ASTRAL-multi and to match or improve on other methods on simulated data sets where GDL and ILS are both present (Zhang et al. 2020).

2.6.2. Species tree estimation under horizontal gene transfer. While HGT implicitly suggests that a phylogenetic network (rather than a tree) is needed, it can be argued that when HGT is sufficiently random, there can still be a well-defined underlying species tree, on top of which the HGT events occur. Therefore, we ask, What is known about the estimation of an underlying species tree given random HGT?

Here, theoretical results established in Roch & Snir (2013) and Daskalakis & Roch (2016) show that when the amount of random HGT is not too large, then the underlying species tree is statistically identifiable from the gene tree distribution, in much the same way that the species tree is identifiable under the MSC, given the gene tree distribution. Most interestingly Daskalakis & Roch (2016) show that under these conditions, for every four species, the most probable quartet tree topology is the true species tree. Hence, methods like ASTRAL that seek to optimize the

MQSST optimization problem are provably statistically consistent under these models (Davidson et al. 2015).

Furthermore, simulation studies have shown that when HGT is present, the empirical performance of ASTRAL is superior to other methods, including concatenated maximum likelihood, and that this advantage increases with the level of HGT (Davidson et al. 2015). Thus, from both an empirical and a theoretical perspective, methods that combine gene trees and are based on quartet tree frequencies can provide improved accuracy over concatenation analyses when HGT is present.

2.6.3. Species tree estimation under combined models. So far, our discussion has addressed only species tree estimation when heterogeneity is due to a single process—ILS, GDL, or HGT. Yet, most empirical data sets exhibit heterogeneity due to multiple causes, so understanding their performance when two or more processes are at play is important. Models of gene tree evolution that capture multiple causes of discordance have been designed [e.g., DLCoal by Rasmussen & Kellis (2012)], and developing methods that are based on these models has been a long-standing challenge (Szöllsi et al. 2014).

A recent advance toward methods that are provably statistically consistent under two or more processes is the (unpublished) result that ASTRAL-multi is statistically consistent under the DLCoal model of both GDL and ILS (Markin & Eulenstein 2020). Beyond theoretical considerations, however, simulation studies have shown that ASTRAL, ASTRAL-Pro, and other quartet-tree methods have low topological errors under conditions that explored multiple concurrent sources of gene tree discord (e.g., Davidson et al. 2015, Yan et al. 2020b, Zhang et al. 2020), and this suggests that these quartet-tree approaches may be statistically consistent under other combinations.

3. PHYLOGENETIC NETWORK ESTIMATION

In the previous section, we presented methods that can be used to estimate species trees when ILS is present. Even when HGT is present, the discussion in the previous section focused on species tree estimation despite HGT, rather than on estimating HGT itself. Here we extend this work to the case of phylogenetic network estimation when ILS and hybridization are both present.

We note that there are two types of phylogenetic networks used in biology: explicit phylogenetic networks (which are graphical models of evolutionary events that include reticulations) and implicit phylogenetic networks (which are also graphical models but aim to represent the input data and so are better seen as data-display networks). Here we focus the discussion on methods for explicit phylogenetic networks and the challenges in developing improved methods of this type, as the MSC has been extended to this type of networks. For further discussion about the two types of methods, see Morrison (2011).

3.1. The Multi-Species Network Coalescent Model

Hybridization is the process of reproduction between individuals in different species (Barton & Hewitt 1985). When hybridization occurs, the evolutionary history of the set of species is best represented by a phylogenetic network. A phylogenetic network includes nodes with two parents to capture hybridizations between pairs of species (including ancestral species). Note that the path that an allele takes going backward in time toward the root is not unique (which is the case if the evolutionary history of the species is modeled by a tree). In hybridization between two divergent species, reproduction between two individuals from the two species produces an offspring, referred to as an F1 hybrid, whose genome is inherited, in equal quantities, from the two parents. However,

after several rounds of interbreeding and backcrossing, the genetic materials in descendants of the F1 hybrid do not necessarily trace back in equal proportions to the original parents. In other words, the number of loci in descendants of the F1 hybrid that trace back their lineages to one species could very well be different from the number of loci that trace back their lineages to the other. Indeed, the number of introgressed alleles in the genome of a descendant of a hybrid individual could be very rare after a large number of generations has passed since the hybridization event (Folk et al. 2018). This fact is very important to take into account when developing models and algorithms for inference of evolutionary histories that include hybrid species, as the signal of hybridization in the genomes of extant species could be too small to infer the hybridization events, particularly ancient ones.

Figure 2 illustrates the evolutionary histories of two loci in the genomes of six individuals, two from each of the three species A, B, and C. **Figure 2*a*** and ***b*** together illustrate the evolutionary

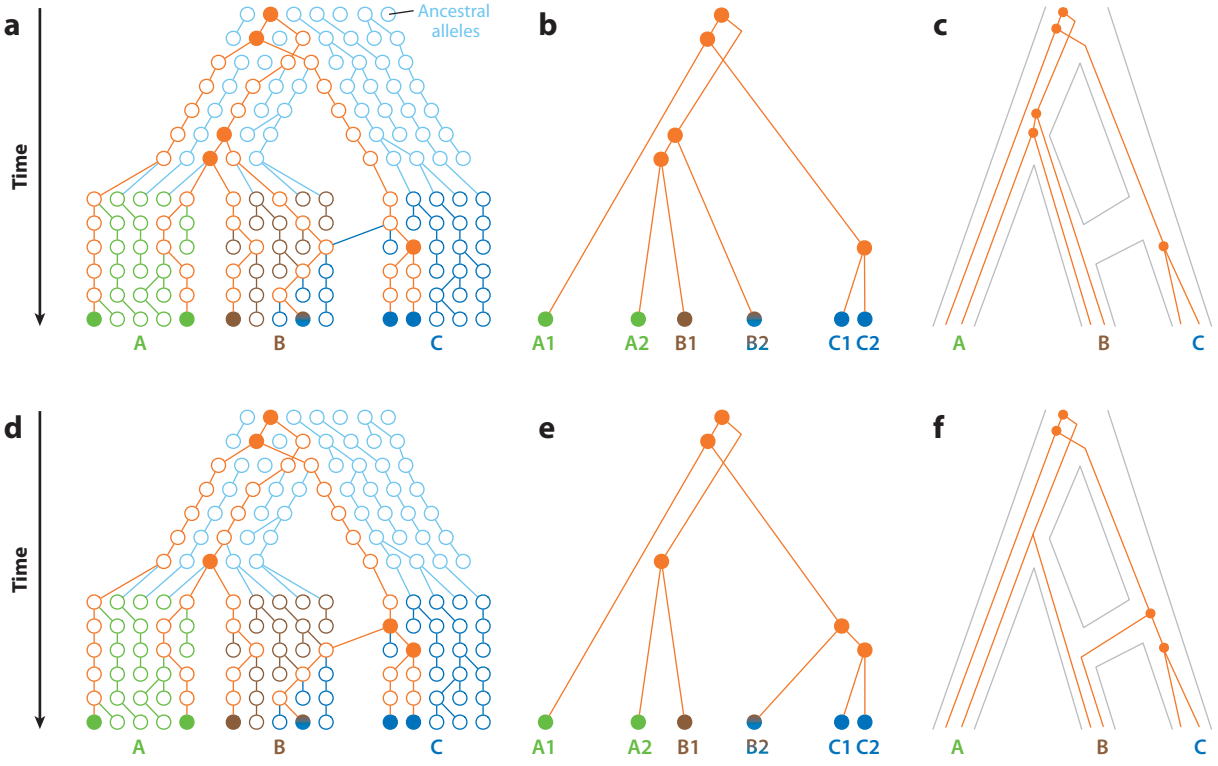

Figure 2

Hybridization and the multispecies network coalescent. (*a*) The evolutionary history of three species (populations)—A (*green*), B (*brown*), and C (*dark blue*)—that includes hybridization between B and C. Solid circles in the extant populations correspond to sampled individuals (two per species), and solid circles of ancestral nodes correspond to coalescence events. Orange circles and lines denote the genealogy of the six sampled individuals. (*b*) The gene tree of a locus whose coalescence events are shown in panel *a*. A1, A2, B1, B2, C1, and C2 correspond respectively to the six solid circles from left to right in the extant populations in panel *a*. The solid circle of individual B2 is colored both brown and blue to indicate that its genomic material is inherited from both B ancestors and C ancestors due to hybridization. (*c*) The gene tree within the branches of the phylogenetic network corresponding to the species evolutionary history in panel *a*. (*d*) The same as panel *a* but showing the evolution of an introgressed locus. (*e*) The gene tree of the introgressed locus. (*f*) The gene tree of the introgressed locus within the branches of the phylogenetic network. The gray lines in panels *c* and *f* visually define the boundary of the phylogenetic network. Light blue circles in panels *a* and *d* correspond to ancestral alleles before the split of A and B from their common ancestor.

history of a locus in individual B2 whose genealogy reflects the species tree in that the locus in B2 coalesces with a locus from individual A2 in the ancestral population of A and B, with the genealogy of this locus shown in **Figure 2*b***. **Figure 2*c*** shows the genealogy of a locus that does not involve an introgression signal within the branches of the phylogenetic network. **Figure 2*d*** illustrates the evolutionary history of a locus in B2 that was inherited from an individual in species C through introgression. The genealogy of this locus is shown in **Figure 2*e***. **Figure 2*f*** shows the genealogy of an introgressed locus.

It is important to note that hybridization and ILS could cooccur, as **Figure 2** illustrates. For example, in the case of both loci, an allele from individual A2 coalesces with an allele from individual B1 before it coalesces with the allele from individual A1, which is an instance of ILS.

More generally, the topology of a phylogenetic network is a directed, acyclic graph that has a unique node as the root (all other nodes are descendants of the root) and two different types of nodes: nodes that have single parents (which denote speciation events) and nodes that have two parents (which denote hybridization or other reticulation events). As in the case of species trees, the leaves of a phylogenetic network are labeled uniquely by the species of interest.

To model evolutionary histories of species that potentially include species of hybrid ancestry while accounting also for the stochasticity of the coalescent process, the MSC was extended to operate within the branches of a phylogenetic network, rather than a phylogenetic tree, giving rise to the multispecies network coalescent (MSNC) (Wen et al. 2016; Yu et al. 2012, 2014). This process is illustrated in **Figure 2*a*** and ***d***. Just like inference of species trees under the MSC, there are three categories of methods for inferring phylogenetic networks under the MSNC (for a detailed survey of recent advances in this area, see Elworth et al. 2019):

- Summary methods, which estimate the phylogenetic network from gene tree estimates.
- Site-based methods, which estimate the phylogenetic network from bi-allelic markers.
- Coestimation methods, which coestimate the phylogenetic network and gene trees of the individual loci from sequence alignment data.

Before we discuss these three categories, we briefly discuss work on estimating phylogenetic network topologies by summarizing gene trees.

3.2. Inferring Network Topologies Using Discrete Optimization

The earliest work on phylogenetic network inference viewed a network as a structure that summarizes the potentially conflicting signals in a set of gene trees that are assumed to correspond to different loci. Furthermore, to have a well-defined problem formulation, networks with the smallest number of hybridizations were always sought by inference methods. More formally, a phylogenetic network displays a set of phylogenetic trees, where each tree is obtained by removing some of the reticulation edges in the network; see **Figure 3** for an example. This notion of displayed trees was the basis for maximum parsimony and maximum likelihood inference methods of phylogenetic networks from sequence alignment data without accounting for ILS (Jin et al. 2006, Nakhleh et al. 2005). Furthermore, methods for inferring phylogenetic networks from gene trees without accounting for ILS sought to solve the following problem:

- Input: A set of gene trees, $\mathcal{G} = \{g_1, g_2, \ldots, g_m\}$, where g_i is a gene tree for locus i.
- Output: A phylogenetic network, Ψ, with the smallest number of reticulation nodes such that each of the m trees in the input is displayed by Ψ.

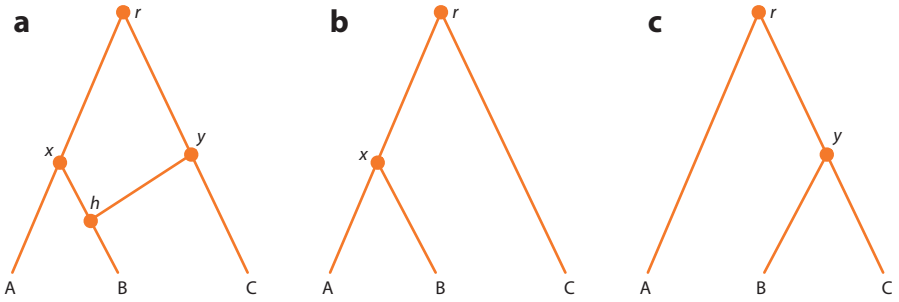

Figure 3

A phylogenetic network and its displayed trees. (*a*) A phylogenetic network whose leaves are labeled by taxa A, B, and C. The root of the network is labeled *r*, the node corresponding to hybridization is node *h*, and its parents are nodes *x* and *y*. The edges connecting *x* to *h* and *y* to *h* are called reticulation edges. (*b*) One of the two trees displayed by the network. This tree captures gene genealogies of loci that are not introgressed and would often be referred to as the species tree. (*c*) The second displayed tree, which captures gene genealogies of introgressed loci.

This problem is NP-hard (Wang et al. 2001), and several methods were developed to solve it, mostly heuristically (Park & Nakhleh 2012; Van Iersel et al. 2010; Wu 2010, 2013). However, since these methods did not account for ILS, all incongruence among gene trees was attributed to hybridization, which would lead to overly complex, and often wrong, networks when ILS is a factor behind gene tree incongruence. To account for ILS in addition to hybridization while staying within the framework of parsimonious inference, a modified criterion of embedding gene trees inside phylogenetic networks was introduced (Yu et al. 2013).

3.3. Statistical Inference of Phylogenetic Networks

The body of work discussed in the previous section has at least four limitations. First, minimizing the number of reticulations is not necessarily biologically plausible in all scenarios. Second, this work allows only for inferring the phylogenetic network topologies without the ability to estimate other parameters of interest, such as divergence times. Third, the work does not allow for incorporating the stochasticity of the coalescent process in a principled manner—the work of Yu et al. (2013) could still result in arbitrarily complex phylogenetic networks. In particular, except for very simple cases, the set of trees displayed by a network does not capture all possible gene genealogies (Zhu et al. 2016). To address these challenges, statistical inference methods under the MSNC (Wen et al. 2016) were introduced.

To enable statistical inference, a phylogenetic network is parameterized as follows. First, associated with each reticulation node is a reticulation time, and with each other node, a divergence time (all leaves have time 0). Furthermore, associated with each branch in the network is a population mutation rate parameter. Time here could be measured in years, generations, or coalescent units. Finally, associated with the reticulation edges are inheritance probabilities that, under the MSNC, correspond to an intermixture model (Long 1991) and indicate the ratios of genetic materials of a hybrid coming from its two parents. A phylogenetic network—topology and parameters—now defines the MSNC and a distribution over gene trees. When the phylogenetic network is a tree, i.e., it has no reticulation nodes, the MSNC reduces to the MSC. The probability mass and density functions of gene genealogies under the MSNC have been derived (Yu et al. 2012, Wen & Nakhleh 2018), which has allowed for developing several inference methods under the MSNC.

As discussed in Section 3.1, these methods fall into three categories, depending on the input type: sequence alignments, biallelic markers, or gene tree estimates.

3.3.1. Coestimation methods. As discussed in Section 2.4, StarBEAST (Ogilvie et al. 2017) is the most commonly used tool for coestimating species and gene trees from sequence alignment data under the MSC. StarBEAST implements a Bayesian MCMC that estimates the posteriors over the trees and their various parameters. Bayesian coestimation under the MSNC was developed independently by Wen & Nakhleh (2018) and Zhang et al. (2018a), who implemented it in the PhyloNet software package (Than et al. 2008) via the command MCMC_SEQ and as the BEAST2 package SpeciesNetwork, respectively. These two methods estimate the posterior distribution over phylogenetic networks, gene trees, and their parameters. Mostly recently, Flouri et al. (2020) implemented a similar inference method in the software package BPP.

Coestimation under the MSC is computationally very challenging, which has limited the applicability of the method to small data sets. In the case of the MSNC, coestimation is even more challenging, as the space is transdimensional, with the Markov chains having to consider models with differing numbers of parameters due to changes in the number of reticulation nodes considered. A heuristic method for coestimating species trees and gene trees was introduced by Wang & Nakhleh (2018), and the same method could in principle apply to phylogenetic networks as well. Furthermore, the efficiency of Bayesian MCMC could be improved by carefully restricting the space of the gene trees (Wang et al. 2020).

3.3.2. Biallelic marker methods. Bryant et al. (2012) introduced a novel method for inferring species trees under the MSC when the data consists of biallelic markers. Assuming independence among the markers, the authors provided an algorithm for analytically integrating over all possible gene histories, thus bypassing the need for sampling gene trees in Bayesian MCMC inference of species trees. Zhu et al. (2018) extended the work of Bryant et al. (2012) so that the inference is done under the MSNC while still analytically integrating over, rather than sampling, gene histories. Given the computational complexity of this method, Zhu & Nakhleh (2018) introduced a method for maximum pseudolikelihood inference of phylogenetic networks from biallelic markers under the MSNC that scales to larger data sets. More recently, Rabier et al. (2020) derived a faster algorithm for computing the full likelihood than the algorithm of Zhu et al. (2018).

In addition to inference of phylogenetic networks, biallelic markers have been used to determine the presence of hybridization by testing deviation of site pattern frequencies from those expected under the MSC. The most widely known and used test in this category is the D-statistic (Durand et al. 2011), which was then extended to larger data sets (Elworth et al. 2018, Pease & Hahn 2015). Furthermore, phylogenetic invariants using biallelic markers were derived to test the presence of hybridization by Kubatko & Chifman (2019) and implemented in HyDe (Blischak et al. 2018).

3.3.3. Summary methods. Yu et al. (2014) introduced two methods for maximum likelihood inference of phylogenetic networks under the MSNC, one that uses rooted gene tree topologies alone and another that also uses gene tree branch lengths. Wen et al. (2016) extended this work into a fully Bayesian framework. Given the prohibitive likelihood computations under the MSNC using gene tree topologies alone, Yu & Nakhleh (2015) and Solís-Lemus & Ané (2016) introduced maximum pseudolikelihood methods that assume rooted and unrooted gene trees, respectively. These pseudolikelihood methods can be viewed as the network counterparts of the MP-EST method (Liu et al. 2010) for species tree inference and can be applied to data consisting of gene tree topologies alone.

3.4. Gene Duplication and Loss and Polyploidy

The models and methods described thus far in this section assume diploid hybridization. The model of Rasmussen & Kellis (2012) was recently extended to allow for modeling GDL, in addition to diploid hybridization, on phylogenetic networks (Du et al. 2019). Furthermore, a few methods were developed for inferring phylogenetic networks that model allopolyploidy; these include parsimony methods (Huber & Moulton 2006, Thomas et al. 2017, Yan et al. 2020a) along the lines of methods described in Section 3.2 and statistical inference methods that explicitly extend the MSC to handle polyploidy (Jones et al. 2013, Oxelman et al. 2017). Yet, much remains to be done in this area.

4. SUMMARY AND FUTURE DIRECTIONS

This review has described methods for species tree and phylogenetic network estimation that address gene tree heterogeneity due to various causes, beginning with ILS but also including GDL, HGT, and hybridization. Of particular concern have been the challenges involved in using these methods on data sets that contain either many loci or many species or individuals.

Most of the focus has been on ILS-only scenarios, where there is the greatest theoretical understanding and the most advanced methods. In this setting, coestimation methods, such as StarBEAST, may provide the best accuracy but are the most computationally intensive. The summary methods are the most popular because of their scalability to large numbers of both loci and species, but their accuracy is generally reduced when gene tree estimation error is high (a common occurrence in biological data sets). Site-based methods are likely to be the most helpful for data sets where the loci lack sufficient phylogenetic signal for summary methods to provide good accuracy; however, these are computationally intensive on data sets with large numbers of species or individuals.

In contrast, method development for GDL-only scenarios or for scenarios involving two or more processes (e.g., ILS and GDL, or ILS and hybridization) is much less advanced. Given the importance of addressing multiple processes, there is an urgent need to extend the mathematical modeling of evolution to address these model combinations and then to establish the theoretical properties (e.g., statistical consistency) of species tree and phylogenetic network estimation under these models. However, as we have noted, computational issues are significant when selecting between methods for species tree estimation under ILS-only scenarios, and the importance of these issues will increase as model complexity (through incorporation of additional processes) increases.

To conclude, it is worth noting that all of the methods discussed in this review assume that species have already been delimited. Species delimitation is a very challenging problem, and the MSC has been used as one model for solving it (Yang & Rannala 2010). However, the accuracy of this approach is also debated (Sukumaran & Knowles 2017), so the general problem of species delimitation is best considered an open one.

SUMMARY POINTS

1. Heterogeneity in evolutionary histories between loci is to be expected and arises from multiple biological processes, including incomplete lineage sorting (ILS), gene duplication and loss (GDL), horizontal gene transfer (HGT), and hybridization. While some combinations of processes can be adequately modeled by a phylogenetic tree, other combinations require phylogenetic networks. Therefore, phylogeny estimation methods should be evaluated under a range of conditions that include heterogeneity.

2. Methodological developments over approximately the last 10 years have resulted in many methods for species tree estimation that have been proven statistically consistent under the multispecies coalescent (MSC) model. In contrast, research has also established that concatenation analyses (where the multiple sequence alignments for the different loci are combined into one superalignment and a tree is then estimated on the superalignment using maximum likelihood or other such methods) are not statistically consistent under the MSC.
3. Relative accuracy on biological or simulated data sets may not reflect whether a method is statistically consistent or not; for example, concatenation analysis using maximum likelihood, although not statistically consistent under the MSC, may be more accurate in some conditions than methods that are proven to be statistically consistent.
4. Species tree estimation under other evolutionary scenarios, including GDL or HGT, is much less advanced than species tree estimation that addresses ILS. For example, only ASTRAL-multi has been proven statistically consistent under a GDL model to date.
5. Phylogenetic networks are necessary models for evolution under scenarios where hybridization is present; yet, very few methods are available to estimate such networks.
6. When preparing to estimate a phylogeny on a particular data set, it is best to consider the properties of your data set in choosing the method (or methods) to use. For those data sets where heterogeneity across the genome seems very low (as suggested by comparisons of the estimated gene trees to each other and/or to an estimated species tree), concatenation may be sufficient to provide high accuracy. When heterogeneity is present but is consistent with ILS alone, then using several different methods (e.g., the better summary methods, site-based methods, and even coestimation methods) can be helpful, and even concatenation can be useful if ILS is not too high. When heterogeneity exists but is suggestive of causes other than ILS, then other approaches may be needed, ranging from methods that address GDL to phylogenomic network methods. By using several different methods, it becomes possible to discover those aspects of the phylogeny that are consistent across different analyses and then focus attention on the areas where methods differ.

FUTURE ISSUES

1. Phylogenetic network estimation methods are available that provide excellent accuracy on small data sets, and some can address multiple sources of discord. However, the limitation to very small data sets is a very significant challenge. New methods for scaling these network estimation methods to larger data sets are needed.
2. Many methods for phylogenomic estimation are restricted to single-copy gene trees and so depend on accurate prediction of orthology. Since orthology prediction is challenging, advances in detecting orthology would be very helpful. Alternatively, methods that estimate the species tree from gene family trees (which can contain multiple copies of each species) would obviate the need for orthology detection and represent a substantial advance. While some such methods have been developed that have been established as

statistically consistent under GDL models, this approach is generally in its infancy, and further work is needed.

3. Current coestimation and site-based methods have limited scalability, but future research can seek to develop alternative methods that eliminate some of those limitations. While existing coestimation methods use Bayesian approaches, there is no reason other forms of coestimation that do not need a full sampling of the posterior distributions cannot be developed. Similarly, while current site-based methods rely on explicitly dividing the set of species into quartets and triplets, there is no inherent reason site-based methods cannot operate on the entire alignment directly, eliminating the need to iterate through all triplets or quartets. Future work should further explore such possibilities.
4. All types of methods used in practice assume the multiple sequence alignments and/or gene trees that comprise their input are free of errors. Yet, neither multiple sequence alignment nor gene tree estimation is reliably accurate, and this can be for many different reasons (including heterotachy and other violations of model assumptions, even within a single gene). In addition, many forms of error such as incorrect annotations, misalignment, and undetected paralogy can creep into the data sets during various steps of the long pipeline used to prepare data. Better understanding of how to detect and eliminate such errors, perhaps by relying on expectations under the MSC model to detect outliers, also requires further research.
5. Modeling of gene tree discordance across the genome can be further complicated by biases caused by misspecified models of sequence evolution. In particular, if biases appear consistently across genes, species tree estimation methods may be misled. Future work should further characterize the effects of model misspecification on existing methods and seek to develop models and methods to alleviate the negative impacts.
6. In addition to the species tree topology, we are often interested in estimating branch support, branch length, and perhaps dating of internal nodes. While Bayesian coestimation methods can produce these quantities, much less is known about how to estimate support or date internal nodes under complex evolutionary histories including gene tree heterogeneity using other types of methods.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was supported in part by the US National Science Foundation (Grant III-1845967 to S.M. and Grants DBI-2030604, CCF-1514177, and CCF-1800723 to L.N.) and by the Grainger Foundation (to T.W.).

LITERATURE CITED

Allman ES, Degnan JH, Rhodes JA. 2011. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *J. Math. Biol.* 62:833–62

Allman ES, Degnan JH, Rhodes JA. 2016. Species tree inference from gene splits by unrooted STAR methods. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 15(1):337–42

Avni E, Cohen R, Snir S. 2015. Weighted quartets phylogenetics. *Syst. Biol.* 64(2):233–42

Barton NH, Hewitt GM. 1985. Analysis of hybrid zones. *Annu. Rev. Ecol. Syst.* 16:113–48

Blischak PD, Chifman J, Wolfe AD, Kubatko LS. 2018. HyDe: a python package for genome-scale hybridization detection. *Syst. Biol.* 67(5):821–29

Boussau B, Szöllősi GJ, Duret L, Gouy M, Tannier E, Daubin V. 2013. Genome-scale coestimation of species and gene trees. *Genome Res.* 23(2):323–30

Bryant D, Bouckaert R, Felsenstein J, Rosenberg NA, RoyChoudhury A. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.* 29(8):1917–32

Bryant D, Steel M. 2001. Constructing optimal trees from quartets. *J. Algorithms* 38(1):237–59

Burleigh JG, Bansal MS, Eulenstein O, Hartmann S, Wehe A, Vision TJ. 2010. Genome-scale phylogenetics: inferring the plant tree of life from 18,896 gene trees. *Syst. Biol.* 60(2):117–25

Chaudhary R, Bansal MS, Wehe A, Fernández-Baca D, Eulenstein O. 2010. iGTP: a software package for large-scale gene tree parsimony analysis. *BMC Bioinform.* 11(1):574

Chaudhary R, Fernández-Baca D, Burleigh JG. 2014. MulRF: a software package for phylogenetic analysis using multi-copy gene trees. *Bioinformatics* 31(3):432–33

Chifman J, Kubatko LS. 2014. Quartet inference from SNP data under the coalescent model. *Bioinformatics* 30(23):3317–24

Christensen S, Molloy E, Vachaspati P, Yammanuru A, Warnow T. 2019. Non-parametric correction of estimated gene trees using TRACTION. *Algorithms Mol. Biol.* 15:1

Dasarathy G, Nowak R, Roch S. 2015. Data requirement for phylogenetic inference from multiple loci: a new distance method. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 12(2):422–32

Daskalakis C, Roch S. 2016. Species trees from gene trees despite a high rate of lateral genetic transfer: a tight bound. In *Proceedings of the 2016 Annual ACM-SIAM Symposium on Discrete Algorithms*, ed. R Kraughgamer, pp. 1621–30. Philadelphia: Soc. Ind. Appl. Math.

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

Degiorgio M, Degnan JH. 2013. Robustness to divergence time underestimation when inferring species trees from estimated gene trees. *Syst. Biol.* 63(1):66–82

Degnan JH. 2013. Anomalous unrooted gene trees. *Syst. Biol.* 62(4):574–90

Degnan JH, Rosenberg NA. 2006. Discordance of species trees with their most likely gene trees. *PLOS Genet.* 2(5):e68

Doyon JP, Ranwez V, Daubin V, Berry V. 2011. Models, algorithms and programs for phylogeny reconciliation. *Briefings Bioinform.* 12(5):392–400

Du P, Ogilvie HA, Nakhleh L. 2019. Unifying gene duplication, loss, and coalescence on phylogenetic networks. In *Bioinformatics Research and Applications: 15th International Symposium, ISBRA 2019, Barcelona, Spain, June 3–6, 2019, Proceedings*, ed. Z Cai, P Skums, M Li, pp. 40–51. New York: Springer

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* 28(8):2239–52

Elworth RAL, Allen C, Benedict T, Dulworth P, Nakhleh L. 2018. D_{GEN}: A test statistic for detection of general introgression scenarios. In *Proceedings of the 18th Workshop on Algorithms in Bioinformatics (WABI)*, ed. L Parina, E Ukkonen, pp. 1–13. Dagstuhl, Germ.: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik

Elworth RL, Ogilvie HA, Zhu J, Nakhleh L. 2019. Advances in computational methods for phylogenetic networks in the presence of hybridization. In *Bioinformatics and Phylogenetics*, ed. T Warnow, pp. 317–60. New York: Springer

Flouri T, Jiao X, Rannala B, Yang Z. 2020. A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. *Mol. Biol. Evol.* 37(4):1211–23

Folk RA, Soltis PS, Soltis DE, Guralnick R. 2018. New prospects in the detection and comparative analysis of hybridization in the tree of life. *Am. J. Bot.* 105(3):364–75

Garey MR, Johnson DS. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York: Freeman
Hahn MW. 2007. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol.* 8(7):R141
Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.* 27(3):570–80
Huber KT, Moulton V. 2006. Phylogenetic networks from multi-labelled trees. *J. Math. Biol.* 52(5):613–32
Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346(6215):1320–31
Jin G, Nakhleh L, Snir S, Tuller T. 2006. Maximum likelihood of phylogenetic networks. *Bioinformatics* 22(21):2604–11
Jones G, Sagitov S, Oxelman B. 2013. Statistical inference of allopolyploid species networks in the presence of incomplete lineage sorting. *Syst. Biol.* 62(3):467–78
Kingman JFC. 1982. The coalescent. *Stoch. Process. Their Appl.* 13(3):235–48
Kubatko LS, Chifman J. 2019. An invariants-based method for efficient identification of hybrid species from large-scale genomic data. *BMC Evol. Biol.* 19(1):112
Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst. Biol.* 56:17–24
Lafond M, Scornavacca C. 2019. On the weighted quartet consensus problem. *Theor. Comput. Sci.* 769:1–17
Larget BR, Kotha SK, Dewey CN, Ané C. 2010. BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics* 26(22):2910–11
Leebens-Mack JH, Barker MS, Carpenter EJ, Deyholos MK, Gitzendanner MA, et al. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574(7780):679–85
Lefort V, Desper R, Gascuel O. 2015. FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol. Biol. Evol.* 32(10):2798–800
Legried B, Molloy EK, Warnow T, Roch S. 2020. Polynomial-time statistical estimation of species trees under gene duplication and loss. In *International Conference on Research in Computational Molecular Biology*, pp. 120–35. New York: Springer
Liu L, Yu L. 2011. Estimating species trees from unrooted gene trees. *Syst. Biol.* 60(5):661–67
Liu L, Yu L, Edwards SV. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol. Biol.* 10(1):302
Liu L, Yu L, Pearl DK, Edwards SV. 2009. Estimating species phylogenies using coalescence times among sequences. *Syst. Biol.* 58(5):468–77
Long JC. 1991. The genetic structure of admixed populations. *Genetics* 127:417–28
Maddison W. 1997. Gene trees in species trees. *Syst. Biol.* 46(3):523–36
Markin A, Eulenstein O. 2020. Quartet-based inference methods are statistically consistent under the unified duplication-loss-coalescence model. arXiv:2004.04299 [q-bio.PE]
Mirarab S, Bayzid MS, Boussau B, Warnow T. 2014a. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science* 346(6215):1250463
Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. 2014b. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30(17):i541–48
Mirarab S, Warnow T. 2015. ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31(12):i44–52
Molloy EK, Warnow T. 2018. To include or not to include: the impact of gene filtering on species tree estimation methods. *Syst. Biol.* 67(2):285–303
Molloy EK, Warnow T. 2020. FastMulRFS: fast and accurate species tree estimation under generic gene duplication and loss models. *Bioinformatics* 36(Suppl. 1):i57–65
Morrison DA. 2011. *An Introduction to Phylogenetic Networks*. Uppsala, Swed.: RJR Productions
Mossel E, Roch S. 2010. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 7(1):166–71
Nakhleh L. 2013. Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends Ecol. Evol.* 28(12):719–28

Nakhleh L, Jin G, Zhao F, Mellor-Crummey J. 2005. Reconstructing phylogenetic networks using maximum parsimony. In *Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference*, pp. 93–102. New York: IEEE

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

Ogilvie HA, Bouckaert RR, Drummond AJ. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Mol. Biol. Evol.* 34(8):2101–14

One Thousand Plant Transcriptomes Initiative. 2019. One thousand plant transcriptomes and phylogenomics of green plants. *Nature* 574:679–85

Oxelman B, Brysting AK, Jones GR, Marcussen T, Oberprieler C, Pfeil BE. 2017. Phylogenetics of allopolyploids. *Annu. Rev. Ecol. Evol. Syst.* 48:543–57

Park H, Nakhleh L. 2012. MURPAR: a fast heuristic for inferring parsimonious phylogenetic networks from multiple gene trees. In *Bioinformatics Research and Applications: 8th International Symposium, ISBRA 2012, Dallas, TX, USA, May 21–23, 2012. Proceedings*, ed. L Bleris, I Măndoiu, R Schwartz, J Wang, pp. 213–24. New York: Springer

Patel S, Kimball R, Braun E. 2013. Error in phylogenetic estimation for bushes in the tree of life. *J. Phylogenetics Evol. Biol.* 1(2):110

Pease JB, Hahn MW. 2015. Detection and polarization of introgression in a five-taxon phylogeny. *Syst. Biol.* 64(4):651–62

Rabiee M, Mirarab S. 2020. SODA: multi-locus species delimitation using quartet frequencies. *Bioinformatics* 36:5623–31

Rabiee M, Sayyari E, Mirarab S. 2019. Multi-allele species reconstruction using ASTRAL. *Mol. Phylogenetics Evol.* 130:286–96

Rabier CE, Berry V, Glaszmann JC, Pardi F, Scornavacca C. 2020. On the inference of complex phylogenetic networks by Markov Chain Monte-Carlo. bioRxiv 2020.10.07.329425. **https://doi.org/10.1101/2020.10.07.329425**

Rannala B, Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* 164(4):1645–56

Rasmussen M, Kellis M. 2012. Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Res.* 22(4):755–65

Richards A, Kubatko L. 2020. Bayesian weighted triplet and quartet methods for species tree inference. arXiv:2010.06063 [q-bio.PE]

Robinson D, Foulds L. 1981. Comparison of phylogenetic trees. *Math. Biosci.* 53(1–2):131–47

Roch S. 2006. A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 3(1):92–94

Roch S, Nute M, Warnow T. 2019. Long-branch attraction in species tree estimation: inconsistency of partitioned likelihood and topology-based summary methods. *Syst. Biol.* 68(2):281–97

Roch S, Snir S. 2013. Recovering the treelike trend of evolution despite extensive lateral genetic transfer: a probabilistic analysis. *J. Comput. Biol.* 20(2):93–112

Roch S, Steel M. 2015. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theor. Popul. Biol.* 100:56–62

Saitou N, Nei M. 1987. The neighbour-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4(4):406–25

Sanderson MJ, McMahon MM. 2007. Inferring angiosperm phylogeny from EST data with widespread gene duplication. *BMC Evol. Biol.* 7(1):S3

Sayyari E, Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. *Mol. Biol. Evol.* 33(7):1654–68

Sayyari E, Mirarab S. 2018. Testing for polytomies in phylogenetic species trees using quartet frequencies. *Genes* 9(3):132

Smith ML, Hahn MW. 2020. New approaches for inferring phylogenies in the presence of paralogs. *Trends Genet.* 37:P174–87

Smith SA, Moore MJ, Brown JW, Yang Y. 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evol. Biol.* 15(1):150

Snir S, Rao S. 2012. Quartet MaxCut: a fast algorithm for amalgamating quartet trees. *Mol. Phylogenetics Evol.* 62(1):1–8
Solís-Lemus C, Ané C. 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLOS Genet.* 12(3):e1005896
Solís-Lemus C, Yang M, Ané C. 2016. Inconsistency of species tree methods under gene flow. *Syst. Biol.* 65(5):843–51
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–13
Sukumaran J, Knowles LL. 2017. Multispecies coalescent delimits structure, not species. *PNAS* 114(7):1607–12
Swofford DL. 2002. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4. *Software Package*. **https://paup.phylosolutions.com/**
Szöllsi GJ, Tannier E, Daubin V, Boussau B. 2014. The inference of gene trees with species trees. *Syst. Biol.* 64(1):e42–62
Takahata N. 1989. Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics* 122(4):957–66
Tarver JE, Dos Reis M, Mirarab S, Moran RJ, Parker S, et al. 2016. The interrelationships of placental mammals and the limits of phylogenetic inference. *Genome Biol. Evol.* 8(2):330–44
Than C, Ruths D, Nakhleh L. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinform.* 9(1):322
Thomas GW, Ather SH, Hahn MW. 2017. Gene-tree reconciliation with MUL-trees to resolve polyploidy events. *Syst. Biol.* 66(6):1007–18
Ullah I, Parviainen P, Lagergren J. 2015. Species tree inference using a mixture model. *Mol. Biol. Evol.* 32(9):2469–82
Vachaspati P, Warnow T. 2015. ASTRID: accurate species trees from internode distances. *BMC Genom.* 16(Suppl. 10):S3
Vachaspati P, Warnow T. 2018. SVDquest: Improving SVDquartets species tree estimation using exact optimization within a constrained search space. *Mol. Phylogenetics Evol.* 124:122–36
Van Iersel L, Kelk S, Rupp R, Huson D. 2010. Phylogenetic networks do not need to be complex: using fewer reticulations to represent conflicting clusters. *Bioinformatics* 26(12):i124–31
Wang L, Zhang K, Zhang L. 2001. Perfect phylogenetic networks with recombination. *J. Comput. Biol.* 8(1):69–78
Wang Y, Nakhleh L. 2018. Towards an accurate and efficient heuristic for species/gene tree co-estimation. *Bioinformatics* 34(17):i697–705
Wang Y, Ogilvie HA, Nakhleh L. 2020. Practical speedup of Bayesian inference of species phylogenies by restricting the space of gene trees. *Mol. Biol. Evol.* 37(6):1809–18
Wascher M, Kubatko L. 2020. Consistency of SVDQuartets and maximum likelihood for coalescent-based species tree estimation. *Syst. Biol.* 70(1):33–48
Wen D, Nakhleh L. 2018. Co-estimating reticulate phylogenies and gene trees from multi-locus sequence data. *Syst. Biol.* 67(3):439–57
Wen D, Yu Y, Nakhleh L. 2016. Bayesian inference of reticulate phylogenies under the multispecies network coalescent. *PLOS Genet.* 12(5):e1006006
Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter EJ, et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *PNAS* 111(45):4859–68
Wu Y. 2010. Close lower and upper bounds for the minimum reticulate network of multiple phylogenetic trees. *Bioinformatics* 26(12):i140–48
Wu Y. 2013. An algorithm for constructing parsimonious hybridization networks with multiple phylogenetic trees. *J. Comput. Biol.* 20(10):792–804
Yan Z, Cao Z, Liu Y, Nakhleh L. 2020a. Maximum parsimony inference of phylogenetic networks in the presence of polyploid complexes. bioRxiv 2020.09.28.317651. **https://doi.org/10.1101/2020.09.28.317651**
Yan Z, Du P, Hahn MW, Nakhleh L. 2020b. Species tree inference under the multispecies coalescent on data with paralogs is accurate. bioRxiv 498378. **https://doi.org/10.1101/498378**

Yang Z, Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. *PNAS* 107(20):9264–69

Yin J, Zhang C, Mirarab S. 2019. ASTRAL-MP: scaling ASTRAL to very large datasets using randomization and parallelization. *Bioinformatics* 35(20):3961–69

Yu Y, Barnett R, Nakhleh L. 2013. Parsimonious inference of hybridization in the presence of incomplete lineage sorting. *Syst. Biol.* 62(5):738–51

Yu Y, Degnan J, Nakhleh L. 2012. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLOS Genet*. 8:e1002660

Yu Y, Dong J, Liu K, Nakhleh L. 2014. Maximum likelihood inference of reticulate evolutionary histories. *PNAS* 111(46):16448–53

Yu Y, Nakhleh L. 2015. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genom*. 16:S10

Zhang C, Ogilvie HA, Drummond AJ, Stadler T. 2018a. Bayesian inference of species networks from multilocus sequence data. *Mol. Biol. Evol.* 35(2):504–17

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

Zhang C, Scornavacca C, Molloy EK, Mirarab S. 2020. ASTRAL-pro: quartet-based species-tree inference despite paralogy. *Mol. Biol. Evol.* 37(11):3292–307

Zhu J, Nakhleh L. 2018. Inference of species phylogenies from bi-allelic markers using pseudo-likelihood. *Bioinformatics* 34(13):i376–85

Zhu J, Wen D, Yu Y, Meudt HM, Nakhleh L. 2018. Bayesian inference of phylogenetic networks from bi-allelic genetic markers. *PLOS Comput. Biol.* 14(1):e1005932

Zhu J, Yu Y, Nakhleh L. 2016. In the light of deep coalescence: revisiting trees within networks. *BMC Bioinform.* 17(14):415