BIOINFORMATICS

Vol. 00 no. 00 2013
Pages 1–9

ASTRAL: Genome-Scale Coalescent-Based Species Tree Estimation from Bootstrap Gene Trees

R. Reaz [1], Md. S. Bayzid [1], M.S. Swenson [2], and T. Warnow [1*]

[1]Department of Computer Science, The University of Texas at Austin, Austin TX 78712
[2]Department of Electrical Engineering, The University of Southern California, Los Angeles CA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Species trees provide insight into basic biology, including the mechanisms of evolution and how it modifies biomolecular function and structure, biodiversity, and co-evolution between genes and species. Yet because gene trees often differ from species trees, estimating species trees accurately can require sophisticated techniques that account for gene tree incongruence. One of the most frequent causes for conflicting topologies between gene trees and species trees is incomplete lineage sorting (ILS), which is modelled by the multi-species coalescent. While many methods have been developed to estimate species trees from multiple genes, some which have statistical guarantees under the multi-species coalescent model, existing methods are too computationally intensive for use with genome-scale analyses or have shown to have poor accuracy under some realistic conditions.

Results: We present ASTRAL, a new method for estimating species trees from multiple genes. ASTRAL is statistically consistent, can run on datasets with thousands of genes, and has outstanding accuracy – improving upon three leading coalescent-based methods (*BEAST, MP-EST, and BUCKy-pop). ASTRAL also generally outperforms concatenation using maximum likelihood, even when the other coalescent-based methods are less accurate than maximum likelihood.

Availability: ASTRAL is available at http://www.cs.utexas.edu/users/bayzid/ASTRAL.zip.

Contact: tandy@cs.utexas.edu

1 INTRODUCTION

Phylogenies (evolutionary trees) provide insights into basic biology, including how life evolved, the mechanisms of evolution and how it modifies function and structure, and how organisms and genes co-evolve. However, species trees can differ from gene trees due to many biological processes, of which incomplete lineage sorting (ILS, modelled by the multi-species coalescent (Kingman, 1982)) is probably the most common. ILS is equivalent to "deep coalescence", which can lead to incongruence between gene trees and species trees, and occurs with high probability whenever the time between speciation events is short relative to the population size (Maddison, 1997). When ILS is present, gene trees can differ from each other and from the species tree, presenting substantial challenges to standard phylogeny estimation methods (Degnan and Rosenberg, 2009a; Edwards, 2009). For example, the standard approach, concatenation (which concatenates the multiple sequence alignments for different genes together into one super-alignment, and then estimates a tree on the super-alignment) can return incorrect trees with high confidence (Kubatko and Degnan, 2007; DeGiorgio and Degnan, 2010; Edwards *et al.*, 2007; Leaché and Rannala, 2011; Heled and Drummond, 2010; Larget *et al.*, 2010). Furthermore, under some conditions, even the most frequent gene tree topology may not be identical to the species tree topology (Degnan and Rosenberg, 2006, 2009b), a condition called "the anomaly zone".

Several species-tree estimation methods have been proven to be statistically consistent under the multi-species coalescent model, most of which estimate species trees by combining estimated gene trees, and so are called "summary methods". Statistical consistency for summary methods means that the method will return the true species tree with high probability, given a large enough number of true gene trees sampled from the distribution defined by the species tree. Some of these methods include STEM (Kubatko *et al.*, 2009), STAR (Liu *et al.*, 2009), GLASS (Mossel and Roch, 2011), MP-EST (Liu *et al.*, 2010), and BUCKy-pop (Larget *et al.*, 2010). Other statistically consistent species-tree estimation methods include BEST (Liu, 2008) and *BEAST (Heled and Drummond, 2010), which co-estimate gene trees and species trees from input sequence alignments.

Simulation studies have shown that *BEAST, generally considered the best of the co-estimation methods, has excellent accuracy, is computationally very intensive to run on datasets with 100 or more genes (Bayzid and Warnow, 2013); for example, a dataset with only 5 species and 166 genes took several weeks to analyze and required 1 billion MCMC iterations to reach convergence (Smith *et al.*, 2014). MP-EST and BUCKy-pop may be the most accurate of the summary methods; but MP-EST can run on large datasets containing thousands of genes, while BUCKy-pop is too computationally intensive to run on very large datasets (Yang and Warnow, 2011). MP-EST is in frequent use for biological dataset analysis (Song *et al.*, 2012; Leaché *et al.*, 2013; Zhao *et al.*, 2013; Zhong *et al.*, 2013), and is possibly the most popular of the coalescent-based methods for analysis of datasets with hundreds of genes.

However, even the best coalescent-based methods have not been shown to be consistently more accurate than concatenation, even in the presence of ILS (DeGiorgio and Degnan, 2010; Bayzid and Warnow, 2013) where concatenation is expected to have poor

*to whom correspondence should be addressed

accuracy. Thus, while coalescent-based approaches have good theoretical guarantees, performance in practice is unclear.

We present ASTRAL (Accurate Species TRee ALgorithm), a new coalescent-based species tree method. Like BUCKy-pop, ASTRAL operates by combining unrooted gene trees and is provably statistically consistent under the multi-species coalescent; however, ASTRAL uses very different techniques to combine the input gene trees into an estimated species tree. We evaluate ASTRAL in comparison to three statistically consistent species tree methods (BUCKy-pop, *BEAST, and MP-EST), and also to concatenation under maximum likelihood, on a collection of biological and simulated datasets. ASTRAL is consistently more accurate than all the coalescent-based methods under all the simulated model conditions we explore. Interestingly, ASTRAL is almost always more accurate than concatenation in these experiments, even when other coalescent-based methods (such as MP-EST) are less accurate.

2 APPROACH

ASTRAL uses a two-step technique in which we first use the input set of estimated gene trees to produce a set of weighted four-taxon trees (called "weighted quartet trees"), and then combine these weighted quartet trees into a tree on the full set of taxa using a heuristic aimed at finding a species tree of minimum distance to the set of weighted quartet trees (details below). Thus, ASTRAL is similar in overall structure to the population tree in BUCKy, and its proof of statistical consistency follows the same arguments as those provided for BUCKy-pop. To understand ASTRAL, we begin by defining a very simple approach, Combining Dominant Quartet Trees (CDQT), which is also statistically consistent. The proof that CDQT is statistically consistent explains why ASTRAL and BUCKy-pop are statistically consistent, and motivates their algorithmic designs.

Combining Dominant Quartet Trees (CDQT). The basic idea is to take the input set of gene trees, compute a "dominant quartet tree" (see below) for every four species, and then combine the dominant quartet trees into a supertree on the full set of species using a preferred quartet amalgamation technique. If the quartet amalgamation technique correctly computes the supertree when the dominant quartet trees are "compatible" (see below), then CDQT is a statistically consistent method under the multi-species coalescent. Thus, CDQT depends on the quartet amalgamation technique, and so is a general technique, and not a particular technique.

The input to CDQT is a set $\mathcal{G}$ of unrooted gene trees, one for each gene, and each gene tree has the set S of species for its leafset. The algorithm has the following steps:

1. For very four species a, b, c, d, we compute the set of induced four-leaf trees, one for each gene.
2. For every four leaves a, b, c, d, we determine which of the three possible unrooted trees on a, b, c, d occurs the most frequently; this is called the "dominant quartet tree on a, b, c, d".
3. We construct a tree T from the set of dominant quartet trees using a preferred quartet amalgamation technique.

Because the method depends on the choice of quartet amalgamation technique, we now discuss this issue. We say that a set of quartet trees is "compatible" if there is a tree T' such that every quartet tree is identical topologically to the subtree of T' induced on its leaf set. Furthermore, when the set of quartet trees is compatible, then there is a unique tree T' that induces all the quartet trees (called the "compatibility supertree"), and it can be computed in polynomial time using very simple techniques (e.g., the Naive Quartet Method, discussed in Erdos *et al.* (1999). Thus, while there are many quartet amalgamation techniques, most of them are able to return the compatibility supertree when the input set contains a tree on every four leaves and is compatible. We call such quartet amalgamation techniques "proper".

THEOREM 1. *If the quartet amalgamation technique is proper, then CDQT is statistically consistent under the multi-species coalescent model.*

PROOF. Let (T_0, Θ) be an unrooted model species tree in the multi-species coalescent model (and so T_0 is a binary species tree and Θ are the branch lengths in T_0 in coalescent units). Let Q be a set of k true gene trees sampled from the distribution on gene trees defined by (T_0, Θ). By Degnan (2013), for every four species a, b, c, d (leaves in T_0), the most probable unrooted gene tree on a, b, c, d is topologically identical to the unrooted tree induced by T_0 on a, b, c, d. Therefore, when k (the number of gene trees) is large enough, with high probability the dominant quartet tree on a, b, c, d will be equal to the most probable unrooted gene tree on a, b, c, d, and so also equal to the species tree on a, b, c, d. Hence, for large enough k, with high probability the set of dominant quartet trees will be compatible, and will uniquely identify the unrooted species tree; when this holds, CDQT will reconstruct the species tree. In other words, for every $\epsilon > 0$, there is some value K so that for $k > K$ and given k true gene trees sampled from the probability distribution on true gene trees defined by (T_0, Θ), with probability at least $1 - \epsilon$, the dominant quartet trees will be equal to the induced four-leaf species trees, and any proper quartet tree amalgamation technique will correctly reconstruct the species tree. Hence, CDQT is statistically consistent under the multi-species coalescent model.

The proof that CDQT is statistically consistent under the multi-species coalescent model provides a guarantee under idealized conditions – where all gene trees are correct and there are a sufficiently large number of them. However, in practice estimated gene trees have error and there may not be a sufficiently large number. Therefore, for good performance (and not just theoretical guarantees), species tree estimation methods need to work well with estimated gene trees – for which the dominant gene trees may not be identical to the most probable gene trees, and hence may not be compatible with each other. Therefore, heuristics for combining quartet trees, that can construct supertrees even when the quartet trees are incompatible, are valuable techniques for species tree estimation in the presence of ILS.

Related methods. BUCKy-pop (Larget *et al.*, 2010) (the population tree output by BUCKy) is one of the statistically consistent methods for species tree estimation under the multi-species coalescent model that uses a quartet-based approach. The input to BUCKy is a set of unrooted gene tree distributions, with one distribution per gene. (BUCKy was originally intended for use with posterior distributions computed using Bayesian MCMC methods, but has also been used with distributions computed using maximum likelihood bootstrapping; both approaches give similar results (Yang and Warnow, 2011).) In the first step, BUCKy-pop uses the gene

tree distributions to estimate a quartet tree every four species, and performs this estimation using sophisticated Bayesian techniques. In the second step, it combines these estimated quartet trees using a quartet tree amalgamation technique (Xin *et al.*, 2007). Because the quartet tree amalgamation technique will reconstruct the compatibility supertree if it exists, BUCKy-pop is statistically consistent under the multi-species coalescent model.

Another statistically consistent method for species tree estimation is MP-EST (Liu *et al.*, 2010). The input to MP-EST is a set of rooted gene trees, and the output is an estimated rooted species tree. Using somewhat more complex mathematical reasoning than was used in the proof of statistical consistency of CDQT and BUCKy-pop, Liu *et al.* (2010) proved that MP-EST is also statistically consistent under the multi-species coalescent.

ASTRAL. We present a new summary method, ASTRAL, which estimates species trees by combining gene trees. However, unlike BUCKy-pop and CDQT, which compute species trees by combining dominant quartet trees, ASTRAL computes weights for *every* possible four-leaf tree (and so for each of the three possible unrooted trees for every four leaves), and then combines this set of weighted quartet trees into a tree on the full set of species. The weights are computed using bootstrap gene tree samples, and so high weight suggests higher confidence in the quartet gene tree. ASTRAL uses a heuristic to combine the weighted quartet trees into a supertree, attempting to solve a version of the NP-hard "Maximum Quartet Compatibility" problem (Jiang *et al.*, 2001), where we set weights on the quartet trees. However, as we will see, this approach produces much more accurate species trees than BUCKy-pop (the most popular coalescent species tree method that uses quartet trees), and is also more accurate than MP-EST (the leading summary method for estimating species trees under the coalescent) and *BEAST (the coalescent-based method with the best accuracy to date). Also, ASTRAL typically improves on concatenation for model conditions with at least moderate levels of ILS, something that previous coalescent-based summary methods have not been able to do except when given large numbers of highly accurate gene trees.

3 ASTRAL

ASTRAL has two steps; in the first step, we generate a set of weighted quartet trees from the input, and in the second step we estimate the species tree from the set of weighted quartet trees. We now describe how we perform each step.

3.1 Step 1: Generate Weighted Quartet Trees

The input is a set of bootstrap gene trees on taxon set S, with one set for each of the N genes, $g_1, g_2, \ldots, g_N$. Given this set, we compute weights for every possible quartet tree $ab|cd$ of four leaves, where $ab|cd$ denotes the unrooted quartet tree with leaf set a, b, c, d in which the pair a, b is separated from the pair c, d by an edge. Thus, we compute a weight $w(q)$ for every possible (unrooted) quartet tree q. Here we describe how to apply the method for datasets that can have different numbers of bootstrap trees for different genes; the approach is a bit simpler to describe when the number of bootstrap replicates is identical across genes.

Note that on every set of four species, there are three possible unrooted quartet trees (simply called "quartet trees"). Also note that every gene tree on the set S of taxa induces a single quartet tree on a, b, c, d. Thus, given a set X_i of bootstrap replicate gene trees for a single gene g_i we define the support for quartet tree $ab|cd$ to be the fraction of the trees in X that induce $ab|cd$ on set a, b, c, d. Note that if the set X contains only fully resolved trees, then the support for the three different unrooted trees on a, b, c, d adds up to 1. Thus, for each gene g_i we obtain the support of each quartet tree on every four species in S. For a given quartet tree $q = ab|cd$, we compute the average of the support of the quartet tree for the different genes, and denote this value by $w(q)$. Note that $0 \leq w(ab|cd) + w(ac|bd) + w(ad|bc) \leq 1$, but that $w(ab|cd)+w(ac|bd)+w(ad|bc) = 1$ under the assumption that all the bootstrap replicate gene trees are fully resolved (which they are when RAxML bootstrapping (Stamatakis, 2006) is used, as in these experiments).

3.2 Step 2: Construct supertree

We use a technique called "WQFM" to combine the quartet trees into a tree on the full set of taxa. WQFM is the weighted version of the QFM technique, developed in Reaz *et al.* (2013). We briefly describe WQFM, and provide more details in the supplementary materials.

Terminology. For an unrooted tree T on taxon set $P \subset S$, we let $L(T)$ denote the leaf set of T. Every edge in T defines a bipartition of its leaf set (defined by deleting the edge but not its endpoints from T), which is denoted by π_e. However, we can also refer to an arbitrary bipartition on set P, whether or not it is present in a given tree T; thus, we let (A, B) be a bipartition with A on one side and B on the other (note that the order of A and B does not matter).

Under the assumption that all bootstrap replicate trees in the input are fully resolved, then given a bipartition (X, Y) on set $P \subseteq S$, we partition the quartet trees defined by the input bootstrap replicates trees as follows:

- quartet trees that are *satisfied* by (X, Y): those quartet trees $ab|cd$ where $\{a, b\} \subseteq X$ and $\{c, d\} \subseteq Y$, or $\{a, b\} \subseteq Y$ and $\{c, d\} \subseteq X$ (i.e., the bipartition (X, Y) separates the two sibling leaf pairs with the quartet tree from each other),
- quartet trees that are *violated* by (X, Y): those quartet trees q whose taxa are fully contained in $X \cup Y$, and where X and Y each contains exactly two of the four taxa in q but q is not satisfied by (X, Y), and
- quartet trees that are *deferred* by (X, Y): those quartet trees q so that $X \cup Y$ misses at least one of the four taxa in q.

In fact, we can partition all possible quartet trees using any given bipartition, whether or not they appear in any bootstrap replicate tree.

We will refer to a pair (X, Y) with $X \cap Y = \emptyset$ and $X \cup Y \subseteq S$ as either a *full bipartition* (or simply a *bipartition*). A non-trivial bipartition is one that has at least two taxa on each side.

Divide-and-conquer approach Let Q be a set of weighted quartet trees over a taxon set $P \subseteq S$, where the quartet trees in Q may not have all their taxa in P.

The divide-and-conquer approach takes the pair (Q, P) as input. The basic algorithm operates in a top-down manner: a good non-trivial bipartition is produced, rooted trees are calculated on the two parts of the bipartition, and then combined together into a rooted tree on the full dataset by making them both subtrees of a common root. Then the tree is unrooted. The key to the algorithm is therefore finding the bipartition, and showing how to recurse on the

subproblems so as to produce rooted trees. Otherwise it is a simple recursive algorithm.

This is the same basic top-down technique as used in Quartets MaxCut (QMC) (Snir and Rao, 2010, 2012), so the only difference in the two methods is how the good non-trivial bipartition is produced. One of the important differences between WQFM and QMC is that QMC can only be run on unweighted quartets, and as we will show, supertrees constructed from weighted quartet trees, whose weights are based on bootstrap support, are *much* more accurate than trees constructed from unweighted quartet trees. Thus, despite the similarity between QMC and WQFM, the technique we use to construct a tree from a set of weighted quartet trees, the differences in algorithm design are important to the accuracy of the resultant tree.

We now briefly describe the technique used to find a good bipartition. Recall that we score a bipartition with respect to the set Q of quartet trees by the total weight of all satisfied quartets minus the total weight of all violated quartets. The technique to find the bipartition uses a heuristic iterative strategy, in which each iteration begins with the bipartition from the previous iteration, and tries to improve it. If the search strategy within this iteration finds a better bipartition, then a new iteration begins with the new bipartition. Thus, the strategy continues until it reaches a local optimum. The search within each iteration, however, allows for bipartitions with poorer scores to be computed, and hence the overall strategy is not purely hill-climbing. The running time of each iteration is polynomial, but the number of iterations depends on the search. We provide more details the supplementary material.

Given the final bipartition (A, B), we use it to define two inputs to WQFM. By running WQFM recursively, we construct two rooted trees, one on A and one on B. We then create a rooted tree on $A \cup B = S$ (the full set of taxa), and then ignore the rooting to obtain an unrooted tree on S. Thus, the rest of the algorithm depends on how we define these two inputs, and how we use WQFM to obtain rooted trees.

Letting (A, B) denote the bipartition that is produced in the divide step, we divide Q into three sets, as follows. The first set contains all quartet trees that are either satisfied or violated by $(A|B)$. The other two sets are Q_A and Q_B, where $Q_A = \{q \in Q : |q \cap A| \geq 3\}$, and Q_B is defined similarly. Note that all quartet trees in Q_A and Q_B are deferred by $(A|B)$.

For each quartet tree $q \in Q_A$ with $|q \cap A| = 3$, we label the taxon that is not in A by a new dummy taxon b^* We similarly relabel one leaf in the relevant quartet trees in Q_B with a new dummy taxon a^*. This produces sets Q'_A and Q'_B, which are on sets $A' = A \cup \{b^*\}$ and $B' = B \cup \{a^*\}$, respectively. We then recurse on each pair (Q'_A, A') and (Q'_B, B'), producing trees that we combine by identifying leaves a^* and b^*, and suppressing nodes of degree two. The base case is obtained when the taxon set has three or fewer leaves, in which case we return a star.

THEOREM 2. *ASTRAL is statistically consistent under the multi-species coalescent model.*

PROOF. Statistical consistency for a summary method follows if the method will return the true species tree given a sufficiently large number of true gene trees sampled from the distribution defined by the model species tree. So suppose we are given a large number of true gene trees so that the most probable gene tree is also the dominant gene tree. With true gene trees, all branch supports are 100%. Therefore, the weights on quartet trees will be equal to the proportion of the gene trees that induce that particular quartet tree. Note that the weight of the dominant quartet tree will be greater than all other quartet trees. Because that the dominant quartet tree (one with the highest frequency) is the most probable gene tree, and hence also the true species tree for its leaf set (since there are no anomalous 4-leaf unrooted gene trees), the best score is obtained by the true species tree.

4 EXPERIMENTS

We explore performance on a collection of biological and simulated dataasets. For the biological datasets, we only examine MP-EST and ASTRAL, the main focus of this study. The simulated datasets include some small datasets with only 11 taxa, on which we can run concatenated analysis with maximum likelihood (CA-ML), MP-EST, *BEAST, BUCKy-pop, and ASTRAL (and its variants). The larger simulated datasets are based on a mammalian dataset, and have 37 taxa; these are too large for *BEAST.

We compare the estimated species trees to the model species tree (for the simulated datasets) or to the scientific literature (for the biological datasets), to evaluate accuracy. For the case of the biological datasets, we examine the bootstrap support as well. Tree error is measured using the Robinson-Foulds (RF) (Robinson and Foulds, 1981) rate; because all trees estimated here are completely bifurcating, this is the same as the missing branch rate (percentage of internal edges in the model tree missing in the estimated tree).

4.1 Simulated datasets

We used datasets generated in other studies to explore performance of coalescent-based methods for estimating species trees. All datasets consist of gene sequence alignments generated under a multi-stage simulation process that begins with a species tree, simulates gene trees down the species tree under the multi-species coalescent model (and so can differ topologically from the species tree), and then simulates gene sequence alignments down the gene trees.

We used a collection of 11-taxon datasets generated by Chung and Ané (2011) with high levels of ILS (the "strongILS" datasets studied in Bayzid and Warnow (2013)), and a collection of 37-taxon datasets based on a biological dataset analysis, also with high ILS, studied in Mirarab *et al.* (2013).

11-taxon datasets. The 11-taxon datasets had 100 genes, which we subsampled to produce smaller datasets, and we analyzed 20 replicates for every number of genes. Every gene has sequence length 500 which results in estimated gene trees with average bootstrap support (BS) of 53%. The evolutionary process is highly heterogeneous between genes in the presence of ILS, and is under a relaxed (not strong) molecular clock. We created 400 bootstrap replicates per gene.

Mammalian simulated datasets. This dataset was simulated by taking the species tree estimated by MP-EST on the biological dataset studied in Song *et al.* (2012). This species tree had branch lengths in coalescent units, which we used to produce a set of gene trees under the coalescent model. Thus, the mammalian simulation model tree has an ILS level based on a coalescent analysis of the biological mammalian dataset, and other properties

of the simulation that are set to reflect the biological sequences they studied.

The simulation models for each gene tree are also under a relaxed molecular clock, as with the 11-taxon datasets. Sequences were simulated down each gene tree under the GTRGAMMA model, and maximum likelihood gene trees were estimated on each sequence alignment using RAxML (Stamatakis, 2006).

Variants of the basic model species tree were generated by varying the amount of ILS, the number of genes, and the sequence length for each gene; these modifications also impact the amount of gene tree estimation error and the average bootstrap support in the estimated gene trees, and so can be modified to produce datasets that resemble the biological data. The default model tree conditions (including the number of genes, sequence length distribution, and amount of ILS) were set to produce a dataset called the "mixed condition" that most resembled the mammalian dataset studied in Song *et al.* (2012).

The amount of ILS was varied by adjusting the branch length (shorter branches increases ILS). A model condition with reduced ILS was created by uniformly doubling (2X) the branch lengths, and two model conditions with higher ILS were generated by uniformly dividing the branch lengths by two (0.5X) and five (0.2X). The amount of ILS obtained without adjusting the branch lengths is referred to as "moderate ILS", and was estimated by MP-EST on the biological data.

The average BS in the biological data was 71%, and so we generated sequence lengths that produced estimated gene trees with bootstrap support bracketing that value – 500bp alignments produced estimated gene trees with 63% average BS and 1000bp alignments produced estimated gene trees with 79% BS. The "mixed dataset" of 400 genes was produced using 200 genes with 63% BS and 200 genes with 79% BS, and had average BS of 71% - like the biological data.

We varied the number of genes from 50 to 800 to explore both smaller and larger numbers of genes than the full biological dataset (which had roughly 400 genes).

For each model condition (specified by the ILS level, the number of genes, and the sequence length), we created 20 replicates, except for the 400- and 800-gene model conditions where we created 10 and five replicates respectively. 200 ML bootstrap replicates were generated for each gene.

4.2 Biological datasets

We analyzed three biological datasets: the mammalian dataset from Song *et al.* (2012), containing 37 species and 447 genes, the amniota (turtle) dataset from Chiari *et al.* (2012) with 248 genes across 16 species, and the tree shrew dataset of Kumar *et al.* (2013) with 19 species and 1006 gene trees.

4.3 Methods

We compared ASTRAL and its variants (described below) with MP-EST, *BEAST, BUCKy-pop (the population tree estimated by BUCKy), and concatenated analysis (CA-ML). We provide details of how each method was run as supplementary materials.

Bootstrap gene trees were used as input to the summary methods (MP-EST, *BEAST, BUCKy-pop, ASTRAL, and the ASTRAL variants). MP-EST was run using multi-locus bootstrapping (MLBS), where inputs are created by selecting one bootstrap replicate tree per gene from the set of bootstrap replicates available for each gene. Next, species trees were estimated by MP-EST followed by computing the greedy consensus (extended majority consensus) of these estimated species trees.

We computed branch support on trees computed using ASTRAL as follows. We used a variant of ASTRAL with MLBS, just as described above for MP-EST; this produces a collection of estimated species trees (one on each collection of bootstrap replicate gene trees). The frequency with which each bipartition appears in this collection is the "bootstrap support" for the bipartition, and is used to provide branch support in the ASTRAL tree. We use this MLBS branch support technique to provide branch support for the biological dataset analyses.

5 RESULTS

Comparing ASTRAL variants. We explored three variants of ASTRAL. The default one, which we call "ASTRAL", runs WQFM on the set of all possible quartet trees, using average bootstrap support for the weight of each quartet tree. We also explored results for a version of ASTRAL we call "ASTRAL(one)", which runs WQFM on a set of quartet trees with only the best supported quartet tree(s) for every four taxa, using the same weight. The final one is "ASTRAL(one,no-weight)", which uses the same set of quartet trees as ASTRAL(one), but considers them all equally weighted.

Table 1 presents the comparison between these variants on the mammalian dataset with 50 to 800 genes. In all cases, ASTRAL(one) is more accurate than ASTRAL(one,no-weight), indicating that considering the bootstrap support for each quartet is helpful. Furthermore, the differences are large (e.g., ASTRAL(one) had RF rate of 7.5% and ASTRAL(one,no-weight) had RF rate of 9.4% for the 50-gene mammalian dataset). The comparison between ASTRAL and ASTRAL(one) indicates that the two methods have very similar accuracy, and neither is clearly more accurate than the other. Hence, the biggest advantage is obtained by using weights, rather than by using all three quartet trees.

Table 1. Comparing ASTRAL variants on the mammalian datasets. We present tree error (Robinson-Foulds error rates, expressed as a percentage, out of 100) for three variants of ASTRAL: the default technique ("ASTRAL"), the technique in which WQFM is applied to the subset of gene trees containing only the best supported of the three topologies on each gene (ASTRAL-one), and the technique where the unweighted QFM method is applied to the same set of gene trees (ASTRAL-one,no-weight). For a given number of genes (column), the best score is in **bold**. Data for all model conditions presented here were generated under moderate levels of ILS (1X) with sequence lengths of 500.

number of genes	50	100	200	400	800
ASTRAL	**7.3**	**6.3**	**4.4**	4.4	**3.5**
ASTRAL(one)	7.5	6.5	4.7	**4.1**	**3.5**
ASTRAL(one,no-weight)	9.4	8.1	7.4	5.6	4.7

Comparison on the mixed mammalian model dataset. We begin with results under the mixed model dataset, shown in Figure 1. These are the most relevant to the biological dataset in Song *et al.* (2012) in terms of number of genes, average bootstrap support per estimated gene tree, and amount of ILS. ASTRAL has the best accuracy, with only 1.6% RF rate, followed by BUCKy-pop (2.6% RF), CA-ML (3.6% RF), and then by MP-EST (4.5% RF).

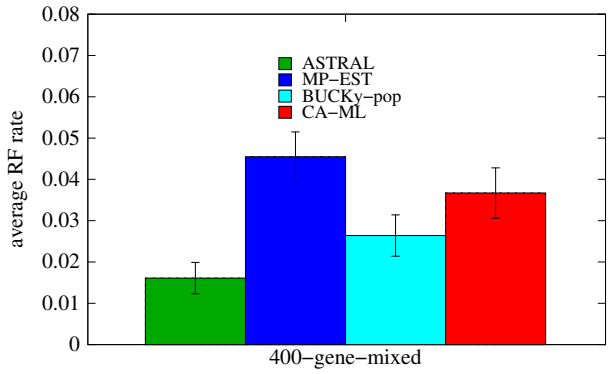

Fig. 1. Mixed mammalian model condition. We show average RF rates (over 20 replicates) with error bars for coalescent methods and CA-ML on the mixed dataset of of 400 mammalian genes under default (1X) ILS, produced using 200 genes of length 500 and 200 genes with of length 1000. This dataset is generated under the model tree constructed by MP-EST for the dataset in Song *et al.* (2012).

Thus, this model condition has enough ILS that coalescent-based methods can be more accurate than concatenation (CA-ML), but not all coalescent-based methods are more accurate than concatenation.

ASTRAL vs. coalescent-based methods. On the 11-taxon datasets (Fig 2), ASTRAL and MP-EST had very close performance, and they were more accurate than *BEAST and BUCKy-pop. On the mammalian datasets (Figs. 3-4), we have comparisons between ASTRAL, MP-EST, and BUCKy-pop (*BEAST could not be run on these data due to computational challenges). Here we see that ASTRAL is more accurate than MP-EST on every model condition tested. The comparison between ASTRAL and BUCKy-pop is interesting: for most conditions, ASTRAL is more accurate, but for some conditions the two methods have the same accuracy. Specifically, ASTRAL was always more accurate than BUCKy-pop on 200 genes (for all ILS levels and sequence lengths tested), but matched the accuracy at 400 genes (tested only under the default ILS level model condition) with both sequence lengths (Figs 3-4). Thus, the relative performance between ASTRAL and BUCKy-pop is impacted by the number of genes, so that with up to 200 mammalian genes ASTRAL is more accurate, and then the two methods have the same accuracy for large enough numbers of genes.

Thus, overall ASTRAL has the best accuracy, and on every tested model condition ASTRAL either is the most accurate of all methods, or it ties for best accuracy. Furthermore, while MP-EST matched ASTRAL on the 11-taxon datasets, it was less accurate on the mammalian datasets (and similarly although BUCKy-pop matched ASTRAL for some mammalian model conditions, it was less accurate on other mammalian model conditions and all 11-taxon model conditions). The comparison between ASTRAL and *BEAST shows that ASTRAL has better accuracy on all 11-taxon model conditions. Thus, ASTRAL has the best general performance of the coalescent-based methods on these datasets.

ASTRAL vs. concatenation. The comparison between ASTRAL and concatenation using maximum likelihood (CA-ML) shows that ASTRAL was often more accurate than concatenation, even for cases where CA-ML was more accurate than many other coalescent-based methods. For example, Figure 2 shows results on mammalian genes with 500bp with default ILS level (1X), and varies the numbers of genes. For these data, ASTRAL is more accurate than CA-ML for 50 to 200 genes, and then CA-ML becomes more accurate (by at most 1.1%) at 400 and 800 genes. Figure 4 also shows results on mammalian genes with default ILS level (1X), and varies the sequence length (500bp and 1000bp) and number of genes (200 and 400); here we see that ASTRAL is more accurate than CA-ML for both numbers of genes when the genes have 1000bp, and is only less accurate than CA-ML for 400 genes with 500bp. Thus, the sequence length for the genes impacts the relative performance between CA-ML and ASTRAL.

The case with 200 genes with 500bp is very interesting: here, CA-ML is more accurate than both MP-EST and BUCKy-pop, but ASTRAL is more accurate than CA-ML; thus, the relative performance of concatenation and coalescent-based methods depends not only on the model condition, but the choice of coalescent-based method.

Impact of ILS. Changes in the ILS level (Fig. 3) for the mammalian datasets with 500bp sequences and 200 genes show that error rates increase for all methods as the ILS level increases (2X is the lowest ILS level, because the branches are twice as long as the default condition and this reduces ILS, and 0.2X is the highest ILS level since the branches are one-fifth the length of the default level and this increases ILS). As expected, the increase in ILS has a bigger impact on CA-ML than it has on the coalescent-based methods. Thus, while CA-ML has the best accuracy of all methods under the lowest ILS condition (2X), it has the worst accuracy under the highest ILS condition (0.2X).

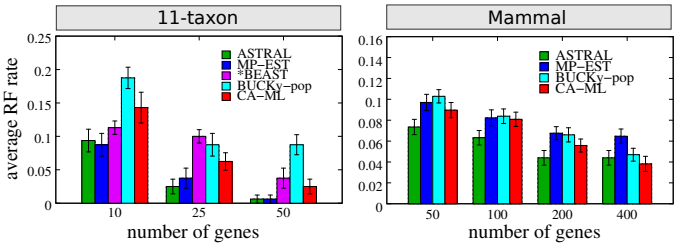

Fig. 2. Impact of number of genes. We show average RF rates with error bars for coalescent methods and CA-ML for both the 11-taxon datasets (left) and mammalian datasets with 500bp under the default ILS (1X) level (right), each with different numbers of genes.

Impact of amount of data. Figure 2 and 4 allows us to evaluate the impact on methods by increasing the amount of data in two ways: increasing the number of genes, or increasing the sequence length per gene. Note that increasing the data improves all methods, but increases in the sequence length have a bigger impact on the coalescent-based methods than on CA-ML. The most striking result is that the coalescent-based methods improve more with increases in the sequence length than with increases in the number of genes, but this is not necessarily the case for CA-ML. Note that coalescent-based methods can be less accurate than CA-ML on gene trees

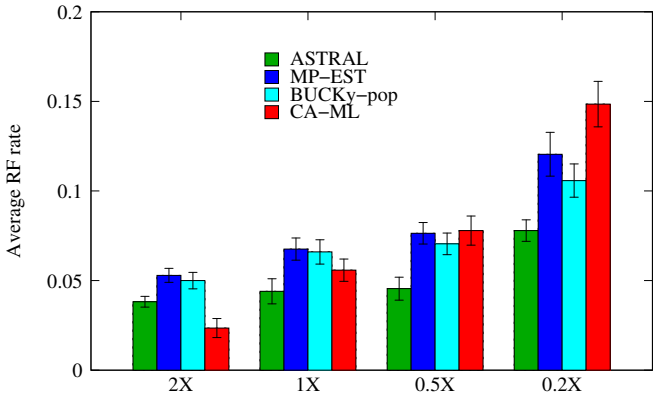

Fig. 3. Impact of ILS. We show average RF rates with error bars for coalescent methods and CA-ML on 200 mammalian genes with 500bp, under different levels of ILS. 2X represents a reduced ILS level, 1X is the default, and 0.5X and 0.2X represent increased ILS levels.

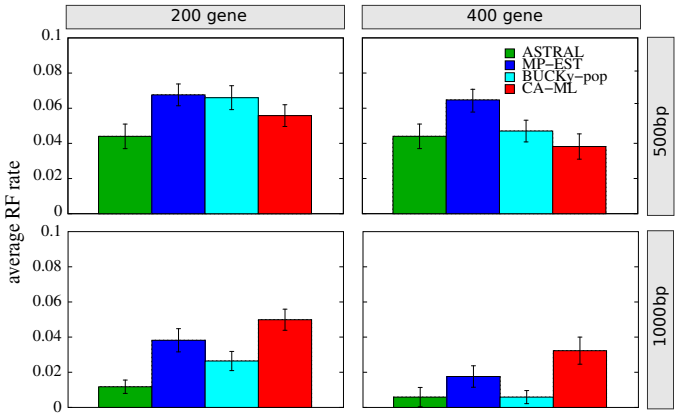

Fig. 4. Impact of amount of data. We show average RF rates with error bars for coalescent methods and CA-ML on on datasets with different numbers of genes (columns) and sequence length per gene (rows) on 1X simulated mammalian genes.

with reduced accuracy (the 500bp case) and then more accurate than CA-ML on gene trees with improved accuracy (the 1000bp case); thus, doubling the gene sequence length changes not only the absolute performance but the relative performance. MP-EST seems the most vulnerable to sequence length and overall amount of data, and ASTRAL the least.

5.1 Biological datasets

We analyzed three biological datasets which have been previously studied in the literature using coalescent-based methods: a dataset for turtle (Amniota) evolution from Chiari *et al.* (2012), a dataset for mammalian evolution from Song *et al.* (2012), and a dataset for Euarchontoglires from Kumar *et al.* (2013). We present detailed results for the turtle and mammalian datasets here; the results on the Euarchontoglires dataset were topologically identical to the previous analyses published on the data and are less interesting.

Turtle (Amniota) datasets. We analyzed data for 248 genes on 16 amniota species from Chiari *et al.* (2012). Previous studies had placed turtles as the sister to birds and crocodiles (Archosaurs) (Hugall *et al.*, 2007; Iwabe *et al.*, 2005; Zardoya and Meyer, 1998). Chiari *et al.* (2012) used concatenation and MP-EST with multi-locus bootstrapping on two sets of gene trees – one based on amino acid (AA) and the other based on nucleotide (NT) alignments. Concatenation and MP-EST on the AA gene trees resolved the clade as (turtles,(birds,crocodiles)) (i.e., birds and crocodiles were considered sister taxa, consistent with the earlier studies) while MP-EST on the NT data produced (birds,(turtles,crocodiles)) (i.e., that contradicted the previous studies). The different resolutions of the two MP-EST trees was considered problematic for MP-EST. Because the concatenation tree and the MP-EST(AA) tree agreed and were consistent with previous studies, the resolution with turtles as sister to birds and crocodiles was considered more likely to be correct.

We re-analyzed both datasets using ASTRAL and obtained identical topologies in both analyses, supporting the standard resolution (turtles,(birds,crocodiles)). However, the trees differed in the branch support for the bird/crocodile clade – 99% support in the ASTRAL(AA) tree and 67% support in the ASTRAL(NT) tree.

Figure 5 shows the ASTRAL and MP-EST trees on the nucleotide gene trees. Note the topological differences and changes in branch support between the two trees. (Branches without support values all have 100% support.)

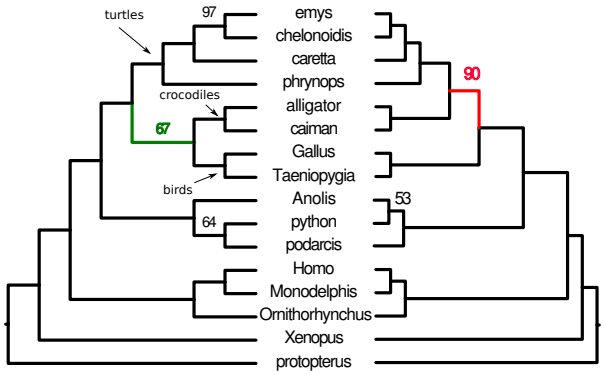

Fig. 5. ASTRAL and MP-EST analyses of the Amniota (turtle) nucleotide dataset from Chiari *et al.* (2012). We show ASTRAL (left) and MP-EST (right) trees on 248 nucleotide gene trees across 16 species. 100% support values are not shown. ASTRAL on AA gene trees returned the same tree as ASTRAL on nucleotide gene trees, but had 99% support for turtles being sister to Archosaurs (birds and crocodiles).

Mammalian dataset. Song *et al.* (2012) analyzed a dataset with 447 genes across 37 mammalian species using MP-EST and combined analyses. In our analysis of this data we detected 21 genes with mislabelled sequences (incorrect taxon names, confirmed by the authors) which we removed from the dataset. We also identified two additional gene trees with potential problems, that were clearly topologically very different from all other gene trees, and removed these as well. We re-analyzed the reduced dataset using MP-EST,

ASTRAL, and CA-ML (the MP-EST and CA-ML analyses were performed for another study that is under submission, but the ASTRAL analysis is new).

The CA-ML on the full and reduced datasets are topologically identical, and the same is true for MP-EST on the full and reduced datasets. However, the CA-ML and MP-EST trees on the reduced dataset differ in the placement of Scandentia (tree shrews) and bats (Chiroptera): CA-ML has Scandentia as sister to Glires (rodents and lagomorphs, which contains rabbits, hares, and pikas) with 88% support, while MP-EST has Scandentia as sister to Primates with 62% support.

Figure 6 shows the ASTRAL tree on the Song *et al.* (2012) dataset. Both ASTRAL and MP-EST trees place bats (Chiroptera) as the sister to all other Laurasiatheria except Eulipotyphla, while concatenation places bats as the sister to Cetartiodactyla. So, unlike the position of tree shrews, MP-EST and ASTRAL resolve the position of bats identically.

The ASTRAL tree places Scandentia as sister to Glires with 77% support, and thus agrees with the CA-ML tree but differs from the MP-EST tree. The resolution of this relationship is clearly controversial and of great interest, however many previous studies (e.g., Boussau *et al.* (2013); Janecka *et al.* (2007)) did not take ILS into account.

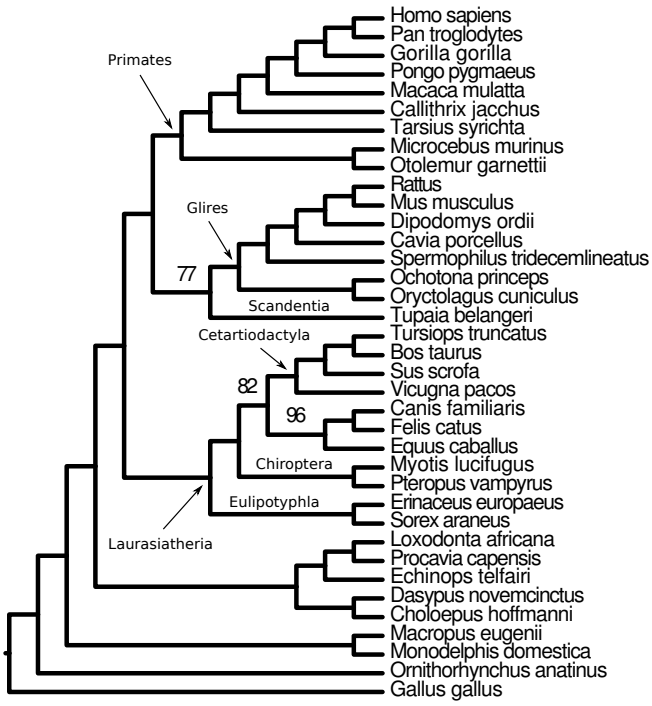

Fig. 6. ASTRAL analysis of the Song *et al.* (2012) mammalian dataset. We show the ASTRAL tree for the Song *et al.* (2012) dataset on 424 gene trees across 37 mammals (100% support values are not shown).

5.2 Running Time

Comparisons between coalescent-based methods reveal substantial differences in running time. Here we focus on the largest datasets (mammalian simulated datasets with 37 taxa), and explore performance on model conditions with moderate ILS, gene sequences of length 500bp, and with 400 and 800 genes.

The first stage of ASTRAL, which computes the quartet tree weights, has running time that increases linearly with the number of genes, but is trivially parallelized and fast for an individual gene. Each gene (set of bootstrap trees) can be processed in around 35 seconds, thus the serial running time for this first stage depends linearly on the number of genes, and is approximately 4 hours for 400 genes, and under 8 hours for 800 genes. However, because this can be trivially parallelized, with a sufficient number of processors it can complete in under a minute. The second stage, which uses WQFM to combine weighted quartet trees, is not impacted by the number of genes, and is not parallelized. In our analyses, it took under 25 minutes on the 37-taxon mammalian datasets. Thus, ASTRAL, run as a sequential program, completed in approximately 4.5 hours for 400 genes and 8.5 hours for 800 genes. However, using distributed computation on the UTCS Condor suite, ASTRAL completed in under an hour for both numbers of genes.

BUCKy-pop strictly runs in serial and uses a Bayesian MCMC technique to estimate gene trees, and then a simple deterministic algorithm (Xin *et al.*, 2007) to combine gene trees. The first step can take a long time and substantial memory to reach convergence, and is impacted by memory. On the 37-taxon mammalian simulated datasets, BUCKy-pop ran to completion for datasets with up to 400 genes (where it took approximately 5 hours), but ran out of memory on the 800-gene dataset on the UTCS Condor cluster. Thus, BUCKy-pop and ASTRAL have comparable serial running time on the 400-gene dataset, but ASTRAL was able to complete on the 800-gene dataset on the Condor cluster, while BUCKy-pop was not. Furthermore, ASTRAL is easily parallelized, but BUCKy-pop is not.

As with BUCKy-pop and ASTRAL, MP-EST performs some preprocessing that is impacted by the number of genes, but afterwards it is only impacted by the number of taxa. However, MP-EST run with multi-locus bootstrapping (as we did here, and how Song *et al.* (2012) and others have run MP-EST) is very much impacted by the number of bootstrap replicates. That is, for 200 bootstrap replicates, we ran MP-EST 200 times, each on a different set of bootstrap replicate gene trees, and then computed the greedy consensus of the resulting MP-EST species tree estimates. Each individual analysis on the 37-taxon dataset took on average 100 minutes. These analyses can be run in parallel, but if run sequentially, using 200 bootstrap replicates would imply a sequential running time of 200*100 = 20000 minutes, or 333 hours. Because the greedy consensus is extremely fast (seconds on 200 37-taxon trees), given a sufficient number of processors MP-EST can complete in about 100 minutes.

Thus, comparing these methods purely as sequential algorithms on the 400-gene dataset, ASTRAL and BUCKy-pop had similar running times (4-5 hours), and MP-EST was much slower (333 hours). On the 800-gene dataset, ASTRAL increased to approximately 8.5 hours, MP-EST stayed at 333 hours, and BUCKy-pop was unable to complete using the available memory. In addition the first phase of, ASTRAL and MP-EST are implemented

to exploit parallelism and complete in under 2 hours on these datasets, but BUCKy-pop is not. Overall, although not extremely fast, ASTRAL is faster than the other methods, and able to run on these phylogenomic datasets in reasonable time frames, without the need for supercomputers.

6 CONCLUSIONS

This study introduced ASTRAL, a method for estimating species trees from a set of estimated gene trees that is provably statistically consistent under the multi-species coalescent model. ASTRAL has excellent accuracy, with substantial improvements over leading coalescent-based methods (*BEAST, MP-EST, and BUCKy-pop) under most conditions. ASTRAL also improved or matched the accuracy of CA-ML (concatenation under maximum likelihood) under many conditions, including cases where CA-ML was more accurate than all the other coalescent-based methods we tested.

Funding: This research was supported by NSF DBI 1062335 and by a grant from the University of Alberta.

REFERENCES

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

Boussau, B., Szöllsi, G., and Duret, L. (2013). Genome-scale coestimation of species and gene trees. *Genome Research*, **23**(2), 323–330.

Chiari, Y., Cahais, V., Galtier, N., and Delsuc, F. (2012). Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (archosauria). *BMC Biology*, **10**(1), 65.

Chung, Y. and Ané, C. (2011). Comparing two Bayesian methods for gene tree/species tree reconstruction: A simulation with incomplete lineage sorting and horizontal gene transfer. *Syst Biol*, **60**(3), 261–275.

DeGiorgio, M. and Degnan, J. H. (2010). Fast and consistent estimation of species trees using supermatrix rooted triples. *Mol Biol Evol*, **27**(3), 552–569.

Degnan, J. H. (2013). Anomalous unrooted gene trees. *Syst Biol*, **62**(4), 574–590.

Degnan, J. H. and Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genetics*, **2**, 762 – 768.

Degnan, J. H. and Rosenberg, N. A. (2009a). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecology Evolution*, **26**(6).

Degnan, J. H. and Rosenberg, N. A. (2009b). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecology Evolution*, **26**(6).

Edwards, S. V. (2009). Is a new and general theory of molecular systematics emerging? *Evolution*, **63**(1), 1–19.

Edwards, S. V., Liu, L., and Pearl, D. K. (2007). High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences of the United States of America*, **104**(14), 5936–5941.

Erdos, P., Steel, M., Szekely, L., and Warnow, T. (1999). A few logs suffice to build (almost) all trees (ii). *Theoretical Computer Science*, **221**, 77–118.

Heled, J. and Drummond, A. J. (2010). Bayesian inference of species trees from multilocus data. *Mol Biol Evol*, **27**, 570–580.

Hugall, A. F., Foster, R., and Lee, M. S. (2007). Calibration choice, rate smoothing, and the pattern of tetrapod diversification according to the long nuclear gene rag-1. *Systematic Biology*, **56**(4), 543–563.

Iwabe, N., Hara, Y., Kumazawa, Y., Shibamoto, K., Saito, Y., Miyata, T., and Katoh, K. (2005). Sister group relationship of turtles to the bird-crocodilian clade revealed by nuclear DNA–coded proteins. *Molecular Biology and Evolution*, **22**(4), 810–813.

Janecka, J. E., Miller, W., Pringle, T. H., Wiens, F., Zitzmann, A., Helgen, K. M., Springer, M. S., and Murphy, W. J. (2007). Molecular and genomic data identify the closest living relative of primates. *Science (New York, N.Y.)*, **318**(5851), 792–794.

Jiang, T., Kearney, P., and Li, M. (2001). A polynomial-time approximation scheme for inferring evolutionary trees from quartet topologies and its applications. *SIAM J. Comput.*, **30**(6), 1924–1961.

Kingman, J. F. C. (1982). The coalescent. *Stoch Proc Appl*, **13**, 235–248.

Kubatko, L. S. and Degnan, J. H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst Biol*, **56**, 17.

Kubatko, L. S., Carstens, B. C., and Knowles, L. L. (2009). STEM: Species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinf*, **25**, 971–973.

Kumar, V., Hallström, B. M., and Janke, A. (2013). Coalescent-based genome analyses resolve the early branches of the euarchontoglires. *PloS One*, **8**(4), e60019.

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

Leaché, A. D. and Rannala, B. (2011). The accuracy of species tree estimation under simulation: a comparison of methods. *Syst Biol*, **60**(2), 126–137.

Leaché, A. D., Harris, R. B., Rannala, B., and Yang, Z. (2013). The Influence of Gene Flow on Species Tree Estimation: A Simulation Study. *Syst Biol*, pages 1–14, doi= 10.1093/sysbio/syt049.

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

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.

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

Maddison, W. (1997). Gene trees in species trees. *Syst. Bio.*, **46**(3), 523–536.

Mirarab, S., Bayzid, M. S., and Warnow, T. (2013). Multi-locus species tree estimation methods and practices. *Systematic Biology*. (under review).

Mossel, E. and Roch, S. (2011). Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *IEEE Comp Biol Bioinform*, **7**(1), 166–171.

Reaz, R., Bayzid, M. S., and Rahman, M. S. (2013). Accurate phylogenetic tree reconstruction from quartets: A heuristic approach. *PLoS Computational Biology*. (under review).

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

Smith, B. T., Harvey, M. G., Faircloth, B. C., Glenn, T. C., and Brumfield, R. T. (2014). Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. *Systematic biology*, **63**(1), 83–95.

Snir, S. and Rao, S. (2010). Quartets maxcut: A divide and conquer quartets algorithm. *IEEE/ACM Transaction of Computational Biology and Bioinformatics*, **7**(4), 704–718.

Snir, S. and Rao, S. (2012). Quartet maxcut: A fast algorithm for amalgamating quartet trees. *Journal of Molecular Phylogenetics and Evolution*, **62**, 1–8.

Song, S., Liu, L., Edwards, S. V., and Wu, S. (2012). Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences of the United States of America*, **109**(37), 14942–7.

Stamatakis, A. (2006). RAxML-NI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinf*, **22**, 2688–2690.

Xin, L., Ma, B., and Zhang, K. (2007). A new quartet approach for reconstructing phylogenetic trees: Quartet joining method. In *Proceedings of Computing and Combinatorics (COCOON 2007*, LNCS 4598, pages 40–50. Springer.

Yang, J. and Warnow, T. (2011). Fast and accurate methods for phylogenomic analyses. *BMC Bioinformatics*. to appear.

Zardoya, R. and Meyer, A. (1998). Complete mitochondrial genome suggests diapsid affinities of turtles. *Proceedings of the National Academy of Sciences*, **95**(24), 14226–14231.

Zhao, L., Zhang, N., Ma, P.-F., Liu, Q., Li, D.-Z., and Guo, Z.-H. (2013). Phylogenomic analyses of nuclear genes reveal the evolutionary relationships within the BEP clade and the evidence of positive selection in poaceae. *PLoS ONE*, **8**(5), e64642.

Zhong, B., Liu, L., Yan, Z., and Penny, D. (2013). Origin of land plants using the multispecies coalescent model. *Trends in Plant Science*, **18**(9), 492–495.