Bioinformatics, 34, 2018, i697–i705
doi: 10.1093/bioinformatics/bty599
ECCB 2018

Towards an accurate and efficient heuristic for species/gene tree co-estimation

Yaxuan Wang[1,*] **and Luay Nakhleh**[1,2,*]

[1]Department of Computer Science and [2]Department of BioSciences, Rice University, Houston, TX 77005, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Species and gene trees represent how species and individual loci within their genomes evolve from their most recent common ancestors. These trees are central to addressing several questions in biology relating to, among other issues, species conservation, trait evolution and gene function. Consequently, their accurate inference from genomic data is a major endeavor. One approach to their inference is to co-estimate species and gene trees from genome-wide data. Indeed, Bayesian methods based on this approach already exist. However, these methods are very slow, limiting their applicability to datasets with small numbers of taxa. The more commonly used approach is to first infer gene trees individually, and then use gene tree estimates to infer the species tree. Methods in this category rely significantly on the accuracy of the gene trees which is often not high when the dataset includes closely related species.

Results: In this work, we introduce a simple, yet effective, iterative method for co-estimating gene and species trees from sequence data of multiple, unlinked loci. In every iteration, the method estimates a species tree, uses it as a generative process to simulate a collection of gene trees, and then selects gene trees for the individual loci from among the simulated gene trees by making use of the sequence data. We demonstrate the accuracy and efficiency of our method on simulated as well as biological data, and compare them to those of existing competing methods.

Availability and implementation: The method has been implemented in PhyloNet, which is publicly available at http://bioinfocs.rice.edu/phylonet.

Contact: yaxuan.wang@rice.edu or nakhleh@rice.edu

1 Introduction

Phylogenetic trees play a central role in almost all of biology. Species trees model how species split and diversify from their most recent common ancestors. Species trees are important not only for depicting how the species evolved, but also for mapping and understanding trait evolution, for species conservation efforts, and for genome sequencing efforts. Gene trees model how individual recombination-free loci within a set of genomes evolve from ancestral copies of the locus. Gene trees are crucial to understanding gene function, to elucidating evolutionary processes that acted on the genomes, and to inferring species evolutionary histories.

Species and gene trees have an intricate relationship governed by evolutionary processes acting on genomes within (and sometimes across) the branches of the species tree to give rise to genomes whose individual loci have heterogeneous gene trees (Maddison, 1997). In particular, given genomic data X, the likelihood of a species tree Ψ can be viewed as (Felsenstein, 1988)

$$\mathcal{L}(\Psi|X) \propto \int_G P(X|G)p(G|\Psi)dG, \tag{1}$$

where the integration is taken over all possible gene trees G, which are posited as a nuisance parameter in this formulation. When the data X consist of m independent loci, $X_1, \ldots, X_m$, the likelihood can be expressed as

$$\mathcal{L}(\Psi|X) \propto \int_G \prod_{i=1}^{m} P(X_i|g_i)p(g_i|\Psi)dG, \tag{2}$$

where g_i is the gene tree of the i-th locus.

However, the gene tree itself is a quantity of interest in many applications in biology. There are two existing approaches to obtaining the gene trees (Fig. 1). Given that the integration in the above equations is very hard to do numerically, Bayesian approaches for species tree inference end up sampling the gene trees as well. This, for example, is what the popular software package *BEAST (Heled and Drummond, 2010) implements. However, inference of species trees by sampling both the species tree and gene trees is computationally very demanding, thus limiting the

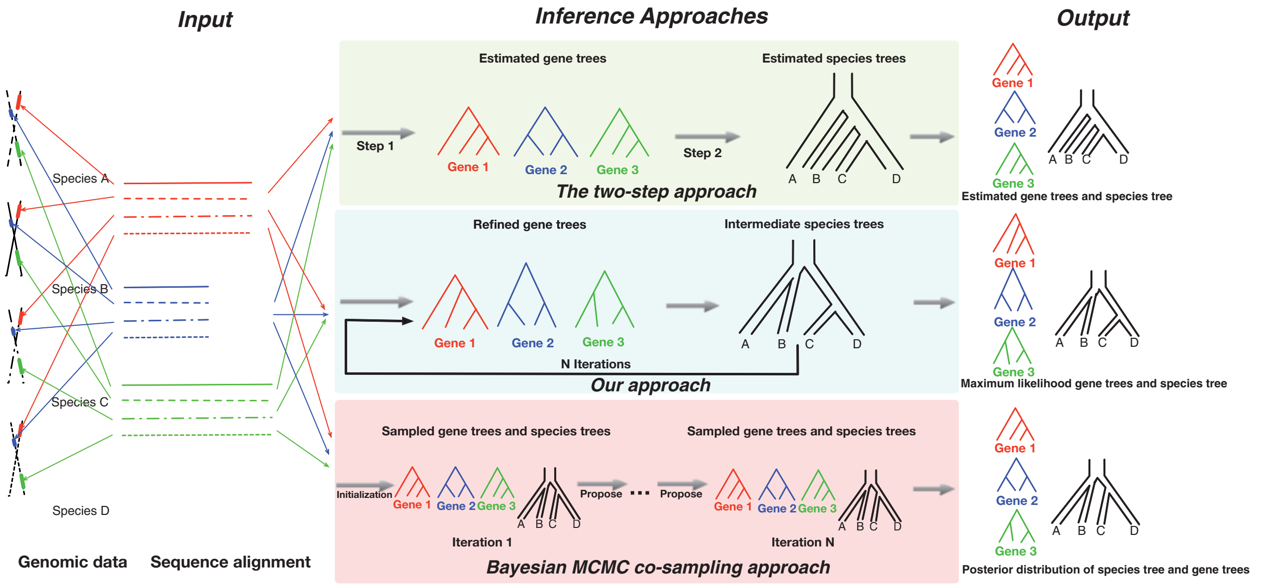


Fig. 1. Approaches to estimating gene trees and species trees. In the two-step approach, gene trees are first estimated for the individual loci, and a species tree is then inferred from the gene tree estimates. In the Bayesian approach, the posterior distribution of gene trees and species tree is sampled, where in each iteration all trees are sampled. Our approach is iterative: in each iteration the species tree is used to generate gene trees, gene trees for the individual loci are selected, and then a new species tree is estimated

applicability to relatively small datasets. An alternative approach that would yield both the gene trees and species trees is to first infer gene trees for the individual loci, and then use the assumed generative process that defines the probability distribution on gene trees to infer the species tree. Significant progress has been made on methods for inferring gene trees from sequence data. For example, RAxML (Stamatakis, 2006) scales up to tens of thousands of tips in the gene tree. In terms of generative processes, the multispecies coalescent, or MSC (Degnan and Rosenberg, 2009), has been widely assumed to explain gene tree heterogeneity and employed by a wide array of species tree inference methods (Nakhleh, 2013). In particular, the original ASTRAL method and its extension (Mirarab *et al.*, 2014; Mirarab and Warnow, 2015) provide a very fast and accurate approach to inferring species trees from gene tree estimates. However, it is important to highlight that the accuracy of the inferred gene trees could have a big impact in this two-step approach. Given the importance of accurate gene trees, independent of the task of species tree inference, methods, such as Treefix (Wu *et al.*, 2013), were developed to 'correct' errors in gene tree estimates by making use of a species tree and fixing the gene tree with respect to it. However, a method such as Treefix relies heavily also on the accuracy of the species tree used. Furthermore, the method could fix the gene tree by making it closer to the species tree, even when that should not necessarily be the case.

In this work, we introduce a new method that is inspired by both the accuracy of the Bayesian methods that co-estimate the species and gene trees and the speed of the two-step methods that make use of already estimated gene trees to infer species trees. More specifically, our method proceeds in an iterative manner where in each iteration it uses a fast method such as ASTRAL to obtain a species tree estimate, parameterizes it and uses it to generate a set of gene trees, and finally, for each locus, it selects an optimal gene tree from among the simulated gene trees. The iterations are repeated for either a fixed number of times or until some convergence is reached. The use of a method like ASTRAL and sidestepping the construction of gene trees 'from scratch' ensures the speed of the method. The accuracy of the inferred species tree and its use as a generative process of gene trees explain the method's accuracy.

We implemented our method in the publicly available software package PhyloNet (Than *et al.*, 2008; Wen *et al.*, 2018) and tested its performance on simulated data as well as the multi-locus rattlesnake dataset of Kubatko *et al.* (2011). On the simulated data, our results show that our method improves on the gene tree accuracy of existing methods, especially for datasets with low signal-to-noise ratio, and obtain species trees that are comparable in accuracy to ASTRAL. For the biological dataset, our method obtains the same species tree as the Bayesian MCMC method of Rannala and Yang (2017). However, our method obtains the result in less than 50 min, whereas the Bayesian method took about 10 h (it is important to note, though, that the Bayesian method infers various parameters of interest beyond the tree topologies, and our method does not). These results indicate that fast heuristics that co-estimate gene and species trees could be developed without resorting to the full strength, yet high computational requirements, of fully Bayesian methods.

2 Background

In the multilocus phylogeny inference problem, the input data X consist of m sequence alignments $X_1, X_2, \ldots, X_m$. These sequence alignments are obtained from m unlinked regions in the genomes of a set of species or populations of interest. Two important assumptions about the m loci are:

- There is no recombination within any locus. This assumption means each sequence alignment X_i evolved down a tree g_i.
- There is free recombination between every pair of loci. This assumption means that loci are independent.

The output of the inference problem is a phylogenetic tree Ψ that represents the evolutionary history of species, and a set of gene tree $G = \{g_1, g_2, \ldots, g_m\}$, where g_i represents the evolutionary history of locus i.

As discussed above, dealing with the uncertainty (due to lack of signal) in the gene trees of the individual loci can be handled in a principled manner by treating the gene trees as nuisance parameters and integrating over them, as presented in Eq. (2). The posterior distribution of species trees and their parameters is then given by

$$f(\Psi|X) \propto \mathcal{L}(\Psi|X)P(\Psi) = P(\Psi)\prod_{i=1}^{m}\int_{G} P(X_i|g_i)P(g_i|\Psi)dg_i. \quad (3)$$

This is the Bayesian formulation employed by *BEAST (Heled and Drummond, 2010), BEST (Liu *et al.*, 2009) and the SNAPP method (Bryant *et al.*, 2012). A major difference among these methods is that the SNAPP method utilizes bi-allelic markers (each X_i is only one site and each genome takes one of two possible states for that site) and, consequently, employs an exact algorithm for numerical integration over all gene trees; that is, it does not sample gene trees. The other two methods provide a sample of the gene trees in addition to the posterior sample of the species trees. These methods tend to provide accurate results but suffer from the prohibitive computational requirements that analyzing large datasets entails.

The two-step approach for inferring species trees is by far the most commonly used approach. Here, gene trees for the individual loci are first inferred. The species tree is then inferred from the gene trees themselves. This is for example the approach taken by several methods that focus solely on inferring the species tree from a collection of gene trees (Kubatko *et al.*, 2009; Mirarab *et al.*, 2014; Mirarab and Warnow, 2015; Than and Nakhleh, 2009; Wu, 2012). These methods are very fast, but rely heavily on accurate estimates of the gene trees.

Algorithm 1 Iterative Co-Estimation (iCE) Algorithm

```
1: procedure iCE($X = (X_1, \ldots, X_m)$, $NumIter$, $NumSamples$)
2:   $G \leftarrow GeneTrees(X)$;
3:   $G_{best} \leftarrow G$;
4:   $\Psi^{(0)} \leftarrow SpeciesTree(G)$;
5:   $\Psi_{best} \leftarrow \Psi^{(0)}$;
6:   $\Psi_{local} \leftarrow \Psi^{(0)}$;
7:   for $t \leftarrow 1$ to $NumIter$ do
8:     $G_t \leftarrow Simulate(\Psi_{local}, NumSamples)$;
9:     $G'_t \leftarrow \{\}$;
10:      for each locus $i$ do
11:        $g_i \leftarrow \text{argmax}_{g \in G_t}[p(X_i|g) \cdot p(g|\Psi^{(t-1)})]$;
12:        $G'_t \leftarrow G'_t \cup \{g_i\}$;
13:      end for
14:      $\Psi^{(t)} = SpeciesTree(G'_t)$
15:      if $\mathcal{L}(\Psi^{(t)}|X) > \mathcal{L}(\Psi_{best}|X)$ then
16:        $G_{best} \leftarrow G'_t$;
17:        $\Psi_{best} \leftarrow \Psi^{(t)}$;
18:        $\Psi_{local} \leftarrow \Psi^{(t)}$;
19:      else if $\mathcal{L}(\Psi^{(t)}|X) > \mathcal{L}(\Psi_{local}|X)$ then
20:        $\Psi_{local} \leftarrow \Psi^{(t)}$;
21:      else if $RandomJump(G'_t, \Psi^{(t)}, \Psi_{local})$ then
22:        $\Psi_{local} \leftarrow \Psi^{(t)}$;
23:      end if
24:    end for
25: return $G_{best}, \Psi_{best}$
26: end procedure
```

3 Materials and methods

We now describe our method, which takes an iterative approach to co-estimating the species tree and gene trees from the sequence data of m independent loci. In each iteration, the method uses the species tree from the previous iteration to simulate a set of gene trees. Then, for each locus it identifies the best tree from among the simulated ones. Finally, the last step of the iteration is to generate a new species tree from the set of gene trees that were identified for the individual loci. The pseudo-code of the method is given in Algorithm 1.

We now describe the details of the main steps of the algorithm.

Line 2 builds an initial set of gene trees, one for each locus. Here, any method for building gene trees could be used (e.g. maximum likelihood as implemented by RAxML). In order to obtain initial ultrametric trees, we used the UPGMA method in our implementation. On Line 4, an initial species tree is inferred from the initial set of gene trees. Similarly, any method for building a species tree from gene tree estimates (e.g. ASTRAL) could be used to implement this step.

Line 8 uses the species tree obtained in iteration $(t-1)$ to simulate *NumSamples* gene trees for iteration t. Since we employ the multispecies coalescent model, we use the program ms (Hudson, 2002) for this step. It is important to explain here how the species tree branch lengths are set, since those, along with the tree's topology, govern the gene tree distribution from which the samples are generated by ms. For the very first iteration $(t=1)$, the species tree $\Psi^{(0)}$ is inferred using the method GLASS (Mossel and Roch, 2010). In this case, the divergence time of node u in the species tree is set to the maximum coalescence time of any alleles from any pair of species under the node u. This guarantees that the initial set of gene trees is consistent with the species tree.

The setting of the species tree branch lengths is different in iterations $t > 1$. In this case, the species tree and its internal branch lengths (in coalescent units) are inferred by ASTRAL. For the external branches, we set their lengths as follows. In addition to the ASTRAL tree, we keep track of a GLASS tree (for temporal constraints). We consider the leaves of the ASTRAL tree in descending order of their depths (where the depth of a leaf is the number of branches on the path from the root to that leaf). We first consider the two deepest leaves x and y (if more than two exist at the largest depth, two of them are selected arbitrarily) and compute the shortest distance d between x and y across all gene trees and the GLASS tree. The length of each of the external branches incident with x and y is set to $d/2$. We repeat this process of all leaves, continuing according to the decreasing order of leaf depth. Note that for leaf node z, if z's sibling is an internal node I, the height of I would have already been calculated. Then the external branch length of z can be calculated treating I as a leaf node.

Lines 10–13 select of each locus the optimal gene tree from the set of gene trees simulated in the current iteration. There are at least two ways to make the selection here. The pseudo-code shows one way. However, Line 11 could also be replaced by

$$g_i \leftarrow \text{argmax}_{g \in G_t} p(X_i|g). \quad (4)$$

In this case, only the sequence data are used to select a gene tree each individual locus from the set of simulated trees. The benefit from this alternative evaluation would be speed, since the algorithm would not need to evaluate the probabilities of gene trees. However, we show below, in the Results section, that using both $p(X|g)$ and $p(g|\Psi)$ in evaluating the gene trees results in more accurate results. It is important to note here that one of the strengths of our method is that it completely circumvents heuristic searches for gene trees (except maybe for the initial set of gene trees obtained on Line 2).

On Line 11, when a gene tree g is evaluated, its branch lengths are re-estimated so as to maximize the likelihood of the gene tree given that sequence alignment for the relevant locus. This re-estimation is done by feeding g's topology and the relevant sequence alignment to RAxML so that g's branch lengths are optimized.

After new gene trees are selected for the individual loci, a new species tree is estimated on Line 14. Once again, any accurate species tree inference method (e.g. ASTRAL) could be used here. If the new species tree has a higher likelihood than the best one seen so far, the new species tree, along with the gene trees from the current iteration, is marked as such (Lines 16–20).

To ameliorate the problem of getting stuck in local maxima, the algorithm allows for random jumps on Line 21 as follows. If the likelihood of the new species tree Ψ_t is worse than the current best, the algorithm may still accept it using the following procedure. If the algorithm jumps to a species tree Ψ_t whose likelihood is lower than the current best species tree, Ψ_t will be stored in Ψ_{local}. Then in the next iteration, the algorithm first compares the likelihood of new species tree, say Ψ_{t+1}, with Ψ_{best} on Line 15. If the likelihood of Ψ_{t+1} is larger than that of Ψ_{best}, Ψ_{t+1} and G'_t will be accepted as current optimal candidates for species tree and gene trees, respectively. If the likelihood of Ψ_{t+1} is smaller than or equal to that of Ψ_{best}, the algorithm compares Ψ_{t+1} with Ψ_{local} in terms of their likelihoods on Line 19. If the likelihood of Ψ_{t+1} is larger, the algorithm accepts it and stores it in Ψ_{local} since it is the best species tree after the last random jump. However, if the likelihood of Ψ_{t+1} is smaller, the method accepts Ψ_{t+1} with some probability on Line 21. In other words, Ψ_{local} represents the 'best species tree since last random jump.' Note that the likelihood of Ψ_{local} is never larger than that of Ψ_{best} by design of the algorithm. Finally, Ψ_{best} and G'_t are returned as the species tree and gene tree estimates, respectively, of the algorithm.

The algorithm makes use of two pre-defined parameters: *NumIter*, which is the number of iterations that the algorithm executes, and *NumSamples*, which is the number of gene trees simulated under a given species tree. These two parameters play the role of knobs that control tradeoff between accuracy and speed. Intuitively, setting these two parameters to high values ensures better accuracy. However, both contribute to added running time. Similarly, setting both parameters to low values ensures faster completion of the analysis, yet not necessarily with high accuracy. In the Results section below, we discuss the impact of different settings to these parameters.

4 Results

We evaluated the performance of our method on both simulated data and an empirical dataset.

4.1 Performance on simulated data

4.1.1 Simulation setup

For species trees, we used the two topologies shown in Figure 2. The length of every internal branch in each of the two trees was set to 1.0 coalescent units. Furthermore, for the symmetric tree, the lengths of all external branches were set to 1.0. For the asymmetric tree, the length of each of the external branches incident with leaves 15 and 16 were set to 1.0, and the length of every other external branch was set so that the tree is ultrametric (i.e. the lengths of all paths from the root to leaves are equal). For example, the length of the external branch incident with leaf 7 is 9.0. These two tree topologies along with the branch lengths follow the same settings of Rannala and Yang (2017).

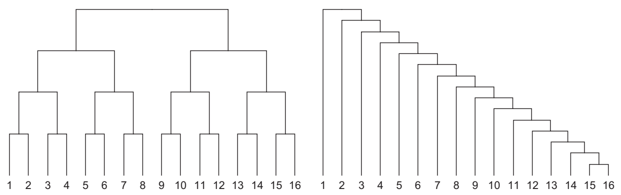


Fig. 2. The two species tree topologies used in the simulations. Left: symmetric tree topology; Right: asymmetric tree topology. These are the same topologies used in the simulation study of Rannala and Yang (2017). For the symmetric tree, all branch lengths are set to 1.0 coalescent units. For the asymmetric tree, all internal branch lengths, as well as the lengths of the two external branches incident with leaves 15 and 16, are set to 1.0; all other branch lengths are set so that the tree is ultrametric (all path lengths from the root to any leaf are equal)

We then used each of the two species trees to generate 10 datasets with 2 gene trees, 10 datasets with 5 gene trees, 10 datasets with 10 gene trees, 10 datasets with 20 gene trees and 10 datasets with 50 gene trees. For generating *ng* gene trees, we used the following ms commands:

- For the symmetric species tree:
  ```
  ms 16 ng -T -I 16 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 -ej 0.5 16
  15 -ej 0.5 14 13 -ej 0.5 12 11 -ej 0.5 10 9 -ej 0.5 8 7
  -ej 0.5 6 5 -ej 0.5 4 3 -ej 0.5 2 1 -ej 1.0 15 13 -ej
  1.0 11 9 -ej 1.0 7 5 -ej 1.0 3 1 -ej 1.5 13 9 -ej 1.5 5 1
  -ej 2.0 9 1
  ```
- For the asymmetric species trees:
  ```
  ms 16 ng -T -I 16 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 -ej 0.5 2 1
  -ej 1.0 3 1 -ej 1.5 4 1 -ej 2.0 5 1 -ej 2.5 6 1 -ej 3.0 7 1
  -ej 3.5 8 1 -ej 4.0 9 1 -ej 4.5 10 1 -ej 5.0 11 1 -ej 5.5
  12 1 -ej 6.0 13 1 -ej 6.5 14 1 -ej 7.0 15 1 -ej 7.5 16 1
  ```

We then used each of the gene trees to simulate three sequence alignments using Seq-gen (Rambaut and Grass, 1997) with sequence lengths 200, 600 and 1000 sites under the HKY model, and using two different population mutation rates ($\theta = 4N\mu$, where N is the effective population size and μ is the mutation rate per site per generation): $\theta = 0.01$ for a fast rate of mutation and $\theta = 0.001$ for a slow rate of mutation. The Seq-gen command used for sequence length *sl* and population mutation rate θ is:

- `seq-gen -mHKY -s` $\theta/2$ `-l sl -on < TrueGT`

where `TrueGT` is the true (model) gene tree used to simulate the sequences.

It is important to point out here that the choice of θ makes a big impact on the accuracy of the inferred trees, in particular the gene trees. For $\theta = 0.001$, the rate of mutation is very slow, resulting in low signal in each sequence alignment for a gene tree to be accurately inferred. Of course, the sequence length also plays an important role since longer sequences contain more signal for the tree inference.

So, in total we generated 2 (tree topologies) × 10 (datasets) × 5 (numbers of loci) × 3 (sequence lengths) × 2 (population mutation rates) = 600 datasets, each consisting of sequence alignments for 2, 5, 10, 20, or 50 loci.

4.1.2 Methods

In addition to our method, we also ran other methods as follows:

- For building gene trees from sequence alignments, we used the FastTree (Price *et al.*, 2010) and RAxML (Stamatakis, 2014) software.

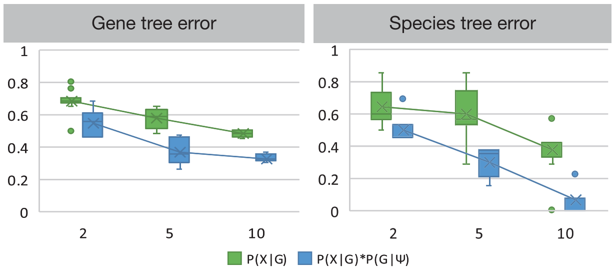


Fig. 3. Error rates of our method based on the two different ways of selecting gene trees for the individual loci. The x-axis shows the three different numbers of loci used in this experiment

- For species tree inference from gene trees, we used ASTRAL-II (Mirarab and Warnow, 2015). ASTRAL-II was run on gene trees obtained by FastTree.
- For correcting gene trees with respect to a species tree, we used the Treefix program (Wu *et al.*, 2013). For the species tree, we used the true species tree (this would be the ideal situation that, in most cases, not attainable in practice), as well as the species tree inferred by ASTRAL-II.

4.1.3 Choice of the gene tree likelihood function

As we discussed above, two ways of selecting a gene tree for locus i from among the gene trees simulated by ms are based on $P(X_i|g) \cdot p(g|\Psi)$, which is what is given in the pseudo-code above, or based solely on $P(X_i|g)$ as given by Eq. (4). To study which of these two methods yields better results, we used the 2-, 5- and 10-locus, 1000-site datasets generated on the 16-taxon symmetric tree with $\theta = 0.001$. We ran our method for 50 iterations and reported the accuracy of the gene trees and species trees inferred. The accuracy in this case was quantified as the Robinson-Foulds (RF) distance (Robinson and Foulds, 1981) between the inferred trees and the true trees. We normalize the RF distances to values in $[0, 1]$ by dividing the size of the symmetric difference between two trees by the number of internal edges in each of the trees. The results are shown in Figure 3.

As the results show, the use of both terms—$P(X_i|g)$ and $p(g|\Psi)$—in selecting the gene trees for the individual loci results in more accurate results in terms of the accuracy of both the gene trees and the species tree. In general, both ways of selecting the gene trees result in improving accuracy as the number of loci increases. However, the improvement, especially in terms of the species tree accuracy, is much more significant when both terms are used. In particular, when 10 loci are used, the species tree error drops down to 0, on average, even when the gene trees still have upwards of 30% error in them.

Based on these results, we decided to use the gene tree selection given in the pseudo-code of Algorithm 1 above. All the results henceforth are based on this setting.

4.1.4 Impact of *NumSamples* and *NumIter*

As we discussed above, two important parameters that control the performance, in terms of accuracy and computational requirements, are *NumSamples*, the number of trees simulated on a given species tree, and *NumIter*, the number of iterations the method performs. While setting both parameters to high values guarantees better results, such a setting also results in high computational costs. Therefore, an important question pertains to reasonable choice of values for both parameters. To address this question, we set out to study the performance of the method in terms of accuracy and running time as we vary the setting of these two parameters.

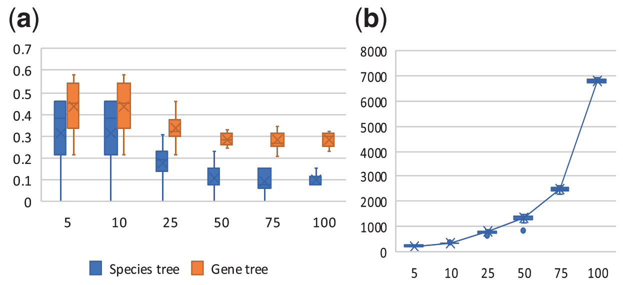


Fig. 4. Error rate of inferred gene trees and species tree and running time as the value of *NumSamples* is varied. Values on the x-axis are the values to which *NumSamples* was set. (**a**) Error rates of inferred gene trees and species tree. (**b**) Running time (seconds)

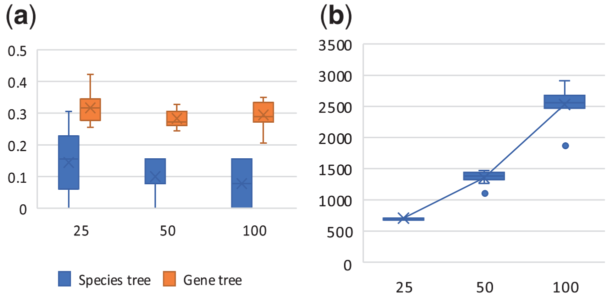


Fig. 5. The error rate of inferred gene trees and species tree, and the running time as the values of NumIter is varied. Values on the x-axis are the values to which *NumIter* was set. (**a**) Error rate of inferred gene trees and species tree. (**b**) Running time (seconds)

In the experiments here, we used the 10-locus, 1000-site datasets generated on the 16-taxon symmetric tree with $\theta = 0.001$. Figure 4 shows the results for varying the value of *NumSamples*.

As the results show, the accuracy of both the gene trees and species tree improve as the number of trees simulated by ms increases. However, the improvement plateaus after $NumSamples = 50$. Coupled with the fact the running time of the method increases significantly when going from 50 to 75 simulated trees, a good choice for the value of *NumSamples* was 50, which is the value we used in all subsequent analyses.

Using the same datasets, and setting *NumSamples* to 50, we now varied the number of iterations that method ran on each dataset. We inspected the gene tree and species tree accuracy, as well as the running time of the method. The results are shown in Figure 5.

As the results show, the accuracy of both gene trees and the species tree improve as the number of iterations is increased from 25 to 50, and then from 50 to 100. However, the improvement is not a significant beyond 50 iterations. Given the significant increase in the running time when going from 50 iterations to 100 iterations, we chose 50 for the number of iterations, which is the setting used henceforth.

4.1.5 The method's accuracy and comparison to other methods

To study the performance of our method in terms of the accuracy of the gene trees and species trees it estimates an compare it to that of other methods, we varied the values of the parameters that have an impact on these methods' performance: the number of loci used in the inference, the sequence length of the individual loci, and the population mutation rate θ. Figure 6 shows the box plot of the error (measured by the Robinson-Foulds distance) in the inferred gene

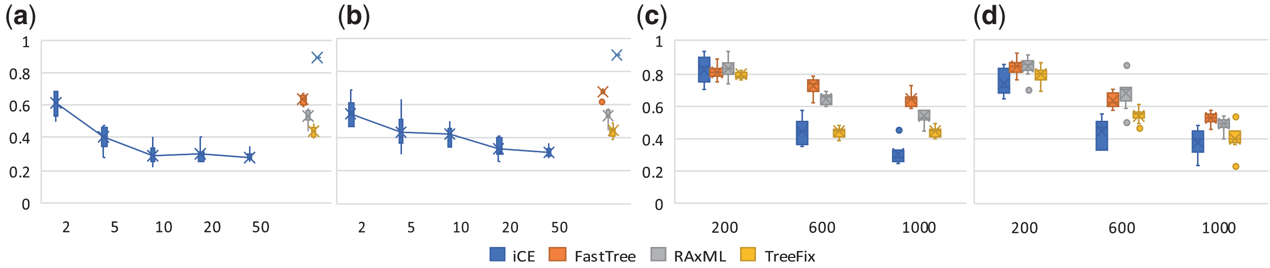


Fig. 6. Gene tree estimation error (RF distance) of the various methods. (**a**) Varying the number of loci on the symmetric species tree. (**b**) Varying the number of loci on the asymmetric species tree. (**c**) Varying the sequence length on the symmetric species tree. (**d**) Varying the sequence length on the asymmetric species tree. Sequence length used for panels (a) and (b) is 1000, and number of loci used for panels (c) and (d) is 10. All results are based on $\theta = 0.001$

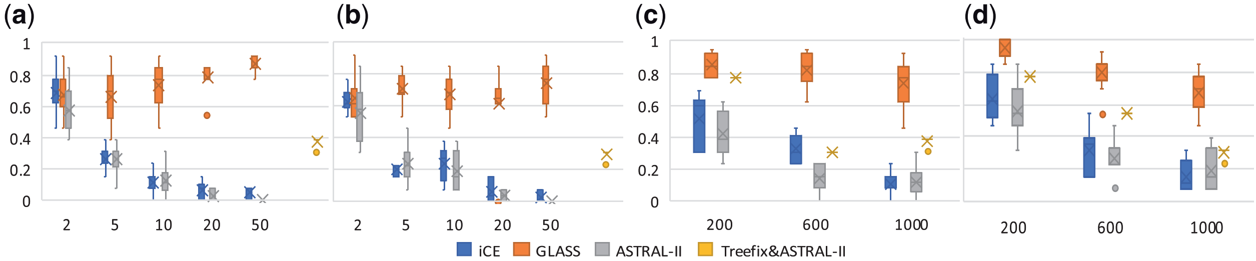


Fig. 7. Species tree estimation error (RF distance) of the various methods. (**a**) Varying the number of loci on the symmetric species tree. (**b**) Varying the number of loci on the asymmetric species tree. (**c**) Varying the sequence length on the symmetric species tree. (**d**) Varying the sequence length on the asymmetric species tree. Sequence length used for panels (a) and (b) is 1000, and number of loci used for panels (c) and (d) is 10. All results are based on $\theta = 0.001$

trees as a function of varying the number of loci (while keeping the sequence length fixed) and varying the sequence length (while keeping the number of loci fixed).

As the figure shows, FastTree infers gene trees with error rate higher than 60%. The gene trees inferred by RAxML are slightly better with error rates around 45%. Varying the number of loci has no effect on the performance of either of these two methods since they work on individual loci independently (they do not make use of the multiple loci). Therefore, we only provide the average gene tree error of all datasets rather than showing individual results for each condition in Figure 6. The accuracy of the gene trees inferred by our method improves as the number of loci increases. Only when two loci are used is the accuracy of the gene trees worse than that of RAxML's gene trees. However, when five or more loci are used, our method infers more accurate gene trees, with error rates lower than 30%.

As for Treefix, we applied it as follows. For a given number of loci, we used the gene trees estimated by FastTree as input to ASTRAL to infer a species tree. We then used this species tree to correct (using Treefix) the gene trees for the corresponding dataset. For example, for a dataset with 20 loci, 20 gene trees were used by ASTRAL to infer the species tree, and each of the 20 gene trees was then corrected with respect to the inferred species tree. As the results show, the accuracy of the gene trees improves slightly but it is still lower than that of RAxML.

When we fix the number of loci but vary the sequence length, we still find that our method performs at least as well, and sometimes better than, the other methods. As the figure shows, the sequence length has a large impact on the accuracy of the inferred gene trees. For very short sequences (200 sites), all methods have a similar performance. However, as the sequence length increases, our method starts outperforming the other methods, especially when 1000 sites are used.

Two final points are in order here. First, the shape of the model species tree does not seem to impact the results. Second, a very low mutation rate was used to generate the synthetic data that underlies the results in Figure 6. This low mutation rate makes it hard to estimate gene trees. The improvement obtained by our method highlights the significance of co-estimating gene trees and species tree in obtaining more accurate gene trees.

In terms of the accuracy of the species tree obtained by our method, we find that it is comparable to that of ASTRAL, which is one of the most accurate methods for species tree estimation. Results are shown in Figure 7. As the results show, the performance of both our method and ASTRAL improves with increasing the number of loci. The performance of the GLASS method is poor due to its extreme sensitivity to coalescent time estimates from the individual loci. One other method that we ran was ASTRAL applied to the gene trees fixed by Treefix. As the results show, the performance of ASTRAL in this case is worse than applying ASTRAL to the 'uncorrected' gene trees. One hypothesis for this is that Treefix is biasing the gene trees towards the species tree it uses, thus affecting the distribution of gene trees, which is the signal that species tree estimation methods use. To explore this hypothesis, we quantified the topological differences between the estimated gene trees and the true gene trees as well as against the true species tree. The results are shown in Figure 8.

As the results show, the RF distance between the gene tree inferred by RAxML and the true gene tree is higher than that between the RAxML tree after correction by Treefix and the true gene tree. In other words, Treefix seemingly improved the gene trees when their quality is taken in terms of their distance from the true tree. However, the results also show that Treefix makes the gene trees closer to the species tree. That is, Treefix biases the gene trees towards the species tree it uses. In particular, when using the true species tree (which is not attainable in most cases in practice), Treefix made the gene trees even closer to it. As we discussed above, this is not a desirable property when it comes to species tree inference, since species tree inference methods make use of the distribution of the gene trees. In other words, Treefix could be completely changing the gene tree distribution, even though it is making individual gene trees closer to their true counterparts.

Finally, we set out to study the performance of the methods as a function of varying the population mutation rate (θ). Figure 9 shows the results on the symmetric species tree; similar results were observed for the asymmetric one. As the figure shows, our method consistently outperforms GLASS and ASTRAL in terms of its ability to recover the individual internal nodes of the true species tree. This outperformance is stronger in the case of the lower mutation rate, which makes sense given that the gene trees and coalescent time estimates that ASTRAL and GLASS, respectively, make use of, are poor in quality. Consistent with the above results, GLASS, in general, has the worst performance among the three methods, and this is especially pronounced for the lower of the two mutation rates. It is important to note here that this analysis is more refined than reporting the RF distances, as done above, and helps shed more light on what nodes in the true species tree present more challenges for the methods. In particular, the results show that ASTRAL does better with the shallower internal nodes than the deeper ones, whereas our method's performance seems to be less sensitive to the depth of the internal nodes.

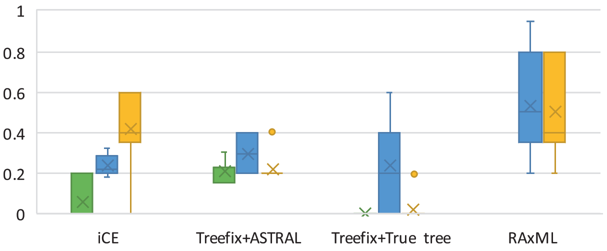


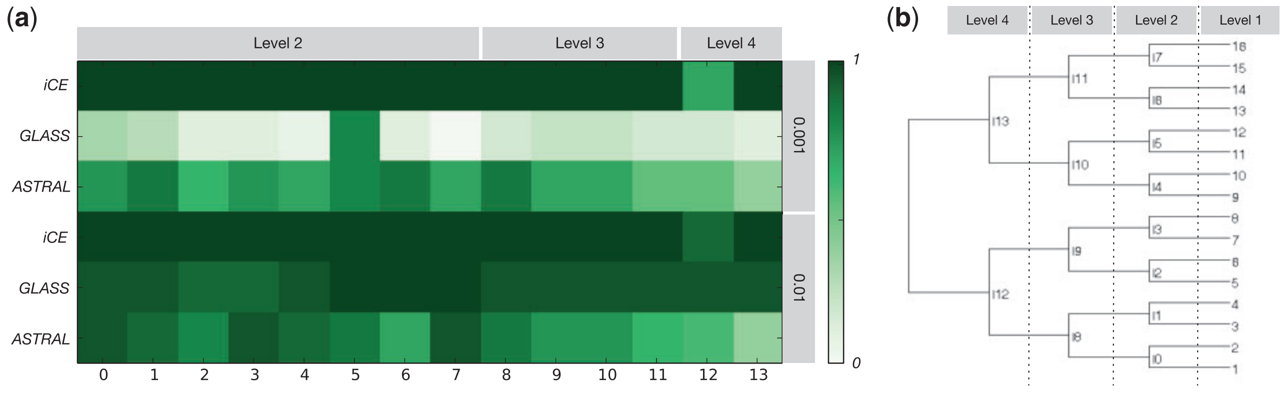
Fig. 8. Error rates between different trees. Green: the RF distance between the true and inferred species tree given by our method (iCE) or the tree fed to Treefix. For 'Treefix + ASTRAL,' the species tree inferred by ASTRAL was used to guide the Treefix correction. For 'Treefix + True tree,' the true species tree was used to guide the Treefix correction. Blue: the RF distance between the true gene tree and inferred gene tree. Yellow: the RF distance between the inferred gene tree and true species tree. Results are based on the 10-locus, 1000-site datasets with $\theta = 0.001$

Rannala and Yang (2017) introduced an improved Bayesian MCMC method for inferring species trees. As we mentioned above, such a method estimates more parameters than just the topologies of the gene and species trees. However, here we focused on comparing these topologies with the ones estimated by our method to examine how the speed of our method, as compared to the much more demanding method of Rannala and Yang (2017), interplays with the accuracy. Results are shown in Figure 10. As the results show, both methods perform similarly when 10 loci are used. In the case when only two loci were used, both methods performed better on datasets with the higher mutation rate, with our method performing better on the shallower internal nodes and the Bayesian MCMC method performing better on the deeper nodes. In the case of the very low mutation rate, both methods had comparable, low accuracy, yet with the Bayesian MCMC method outperforming our method on the deeper nodes still.

4.2 Performance on an empirical dataset

For the empirical dataset, we analyzed the *Sistrurus rattlesnakes* dataset of Kubatko *et al.* (2011). We analyzed data from six subspecies, three from *Sistrurus catenatus*— *S.c. catenatus* (Scc), *S.c. tergeminus* (Sct) and *S.c. edwardsii* (Sce)—and three from Sistrurus miliarius— *S.m. miliarius* (Smm), *S.m. barbouri* (Smb) and *S.m. streckeri* (Sms). Twenty individuals from these six subspecies were used: eight from Scc, three from Sce, four from Sct, two from Sms, one from Smm and two from Smb. The data consist of 18 nuclear loci and an mtDNA (mitochondrial) gene fragment from these species. For each locus, the number of sequences was between 44 and 48 (due to the availability of data from multiple individuals per species) and the lengths of sequences range from 194 to 849. We set $\theta = 0.0015$ following the setting of Kubatko *et al.* (2011).

The implementation of our method currently has support for one individual per species. Therefore, we analyzed the 20 individuals as if they were 20 sub-populations, each with one individual sampled from it. Kubatko *et al.* (2011) concatenated the loci and inferred a tree using BEAST. Both trees are shown in Figure 11. As the figure shows, both methods agree on six clades and differ only in the placement of the Smm individuals, where our method grouped it with the



Frequency matrix of internal nodes under different mutation levels

The true species tree.

Fig. 9. Accuracy of the species tree inferred by various methods on two different settings of the population mutation rate (θ). (**a**) Frequency matrix of each internal node. Numbers on the x-axis correspond to internal nodes in the species tree (e.g. 5 corresponds to node I5). The frequency, which ranges from 0 to 1, is the proportion of all datasets that a given internal node was recovered correctly in the inferred species tree. (**b**) The model species tree used in the simulations. Results are based on 10-locus, 1000-site datasets. (a) Frequency matrix of internal nodes under different mutation levels. (b) The true species tree

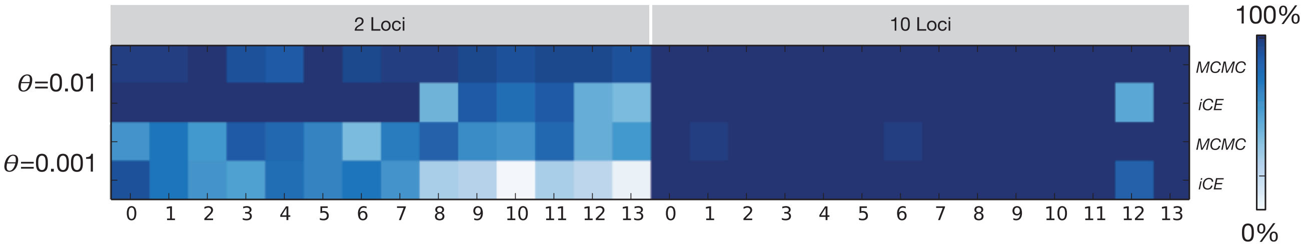


Fig. 10. Accuracy of the species tree inferred by our method (iCE) and the Bayesian MCMC method of Rannala and Yang (2017). Frequency matrix of each internal node of the model species tree of Figure 9(b) is shown. 1000-site datasets were used here

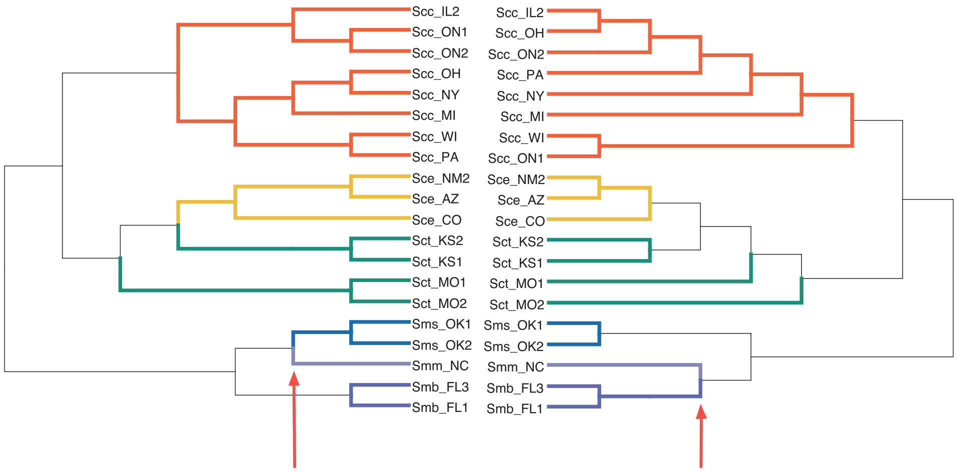


Fig. 11. Trees inferred on all 20 individuals by two different methods. Left: the maximum clade credibility tree inferred by BEAST. Right: the tree inferred by our method (individuals are not mapped to species). The red arrows point to the disagreement in the results obtained by the two methods

two other individuals from the same subspecies and BEAST split them.

If we collapse all individuals from the same sub-species into one leaf so that a species tree has only six leaves, we obtain the tree in Figure 12. The figure also shows trees inferred by *BEAST (Drummond *et al.*, 2012) and the method of Rannala and Yang (2017). While the tree obtained by our method differs from the analysis that used *BEAST, it is identical to that obtained by the method of Rannala and Yang (2017), though the tree in Figure 11 demonstrates that there is incomplete lineage sorting involving Sct (two of the Sct individuals coalesce with the Sce individuals before they coalesce with the other two Sct individuals).

In Figure 13 we show a gene tree of one of the 19 loci (for space limitations, we omit the trees of the other 18 loci, and use only one for illustrative purposes). The gene tree was inferred for the *Clone 41* locus, where the alignment had 274 sites. As the figure clearly shows, our method infers a tree that is mostly compatible with the species tree and the RAxML tree, yet fully resolved while the RAxML tree has very little resolution.

5 Conclusions

In this paper, we introduced a novel heuristic for co-estimating species tree and gene tree from datasets of multiple, unlinked loci. The method is inspired by the accuracy of Bayesian MCMC sampling methods and the speed of Approximate Bayesian Computation (ABC). The former category of methods inspired the iterative nature where in each iteration a species tree and gene trees from the individual loci are estimated from the sequence data directly. From the

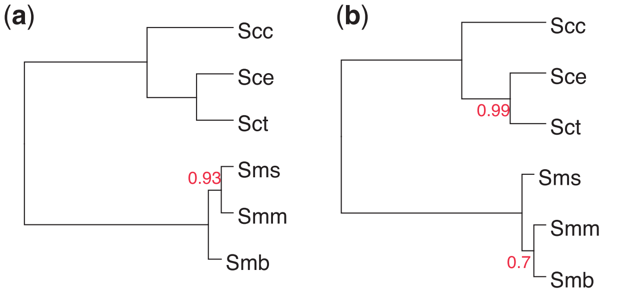


Fig. 12. Species trees on the six subspecies. (**a**) Tree inferred by *BEAST. (**b**) Tree inferred by our method as well as the method of Rannala and Yang (2017). Numbers in red are support values (posterior probabilities) as reported by Rannala and Yang (2017)

ABC method we made use of the concept of treating the species tree as a generative model to simulate gene trees from which to select, rather than search the tree space, individual gene trees for the loci.

We implemented our method and studied its performance on simulated data and an empirical dataset. For the simulated data, we studied the effect of the sequence length, number of loci and mutation rate on the performance of the method, and compared that performance to competing methods. Our results show that for low mutation rate where the signal is very weak for inferring gene trees independently, our method performs much better. The key message is that co-estimating both types of trees (species and gene trees) provides a powerful signal for inferring accurate, resolved gene trees even when the signal in individual sequence alignments is low.

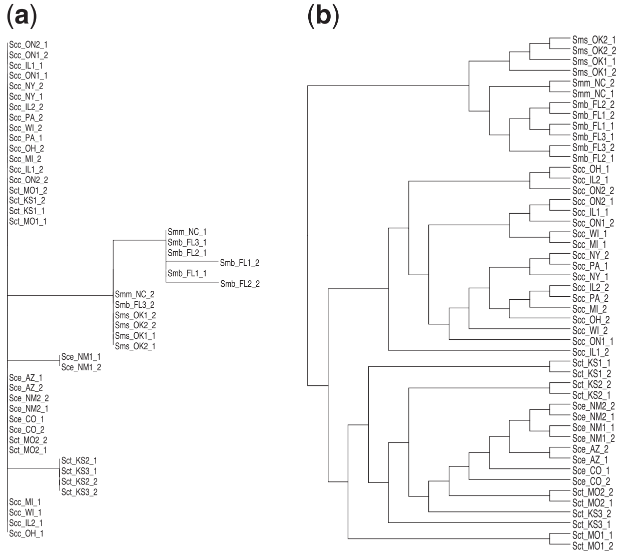


Fig. 13. Inferred gene tree for the *Clone 41* locus in the rattlesnake data. (**a**) Gene tree inferred by RAxML. (**b**) Gene tree inferred by our method

Furthermore, despite its speed and heuristic nature, our method was able to infer a species tree on the empirical dataset that matches that inferred by the more detailed and more computationally demanding Bayesian MCMC method of Rannala and Yang (2017).

One potential direction for future research is to develop methods that sample parameters of interest (e.g. divergence times, etc.) conditional on species and gene tree topologies estimated by our method. The benefit from such an approach is that our method would produce accurate topologies very efficiently, thus obfuscating the need to search the space of topologies and continuous parameters at the same time.

Funding

This work was supported by grants DBI-1355998, CCF-1302179, CCF-1514177 and DMS-1547433 from the National Science Foundation.

Conflict of Interest: none declared.

References

Bryant,D. *et al*. (2012) Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.*, **29**, 1917–1932.

Degnan,J.H. and Rosenberg,N.A. (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.*, **24**, 332–340.

Drummond,A.J. *et al*. (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.*, **29**, 1969–1973.

Felsenstein,J. (1988) Phylogenies from molecular sequences: inference and reliability. *Annu. Rev. Genet.*, **22**, 521–565.

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

Hudson,R.R. (2002) Generating samples under a wright–fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Kubatko,L.S. *et al*. (2009) Stem: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*, **25**, 971–973.

Kubatko,L.S. *et al*. (2011) Inferring species-level phylogenies and taxonomic distinctiveness using multilocus data in sistrurus rattlesnakes. *Syst. Biol.*, **60**, 393–409.

Liu,L. *et al*. (2009) Estimating species phylogenies using coalescence times among sequences. *Syst. Biol.*, **58**, 468–477.

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

Mirarab,S. *et al*. (2014) Astral: genome-scale coalescent-based species tree estimation. *Bioinformatics*, **30**, i541–i548.

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

Mossel,E. and Roch,S. (2010) Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *IEEE/ACM Trans. Comput. Biol. Bioinf. (TCBB)*, **7**, 166–171.

Nakhleh,L. (2013) Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends Ecol. Evol.*, **28**, 719–728.

Price,M.N. *et al*. (2010) Fasttree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.

Rambaut,A. and Grass,N.C. (1997) Seq-gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, **13**, 235–238.

Rannala,B. and Yang,Z., (2017) Efficient bayesian species tree inference under the multispecies coalescent. *Syst. Biol.*, **66**, 823–842.

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

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

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

Than,C. and Nakhleh,L. (2009) Species tree inference by minimizing deep coalescences. *PLoS Comput. Biol.*, **5**, e1000501.

Than,C. *et al*. (2008) PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*, **9**, 322.

Wen,D. *et al*. (2018) Inferring phylogenetic networks using PhyloNet. *Syst. Biol.*, **67**, 735–740.

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

Wu,Y.C. *et al*. (2013) Treefix: statistically informed gene tree error correction using species trees. *Syst. Biol.*, **62**, 110–120.