Version dated: December 16, 2013

RH: Multi-locus species tree methods

Multi-locus species tree estimation methods and practices

Siavash Mirarab[1], Md Shamsuzzoha Bayzid[1], and Tandy Warnow[1]

[1]*Department of Computer Science, University of Texas at Austin, Austin, TX, 78712, USA;*

Corresponding author: Tandy Warnow, Department of Computer Science, University of Texas at Austin, Austin, TX, 78712, USA; E-mail: tandy@cs.utexas.edu.

Abstract.— Species tree estimation is complicated by processes, such as gene duplication and loss and incomplete lineage sorting, that cause discordance between gene trees and the species tree. Furthermore, while concatenation, a traditional approach to tree estimation, has excellent performance under many conditions, the expectation is that the best accuracy will be obtained through the use of species tree estimation methods that are specifically designed to address gene tree discordance. In this paper, we report on a study to evaluate one of the most promising species tree estimation methods – MP-EST, a pseudo-maximum likelihood estimator – as well as concatenation under maximum likelihood, the greedy consensus, and two supertree methods (MRP and MRL). Our study shows that several factors impact the absolute and relative accuracy of methods, including the number of gene trees, the accuracy of the estimated gene trees, and the amount of ILS. Concatenation can

outperform even the best summary methods in some cases (mostly when the gene trees had poor phylogenetic signal), but summary methods generally outperform concatenation when there is an adequate number of well estimated gene trees. The study suggests that coalescent-based species tree methods may be key to estimating highly accurate species trees from multiple loci.
(Keywords: incomplete lineage sorting, coalescent process, species tree estimation, supertree methods, consensus methods, concatenation, gene tree discordance, multilocus bootstrapping, MP-EST, MRP, MRL)

The estimation of species trees from multiple gene trees is necessary since true gene trees can differ from each other and from the true species tree due to multiple processes, including gene duplication and loss, horizontal gene transfer, and incomplete lineage sorting (ILS) (Maddison 1997; Degnan and Rosenberg 2009; Nakhleh 2013). Because ILS is considered to be a major challenge to species tree estimation (Edwards 2009), many methods have been developed to estimate species trees in the presence of ILS (surveyed in Degnan and Rosenberg (2009); Yang and Warnow (2011); Nakhleh (2013). Some of these methods are called "summary methods", because they operate by combining estimated gene trees. One of the earliest methods for estimating species trees in the presence of ILS was based on the MDC (minimize deep coalescence) optimization criterion (Maddison 1997; Maddison and Knowles 2006; Than and Nakhleh 2009) and later extended to address gene trees with missing taxa and estimation error (Yu et al. 2011; Bayzid and Warnow 2012). Subsequently other summary methods that used likelihood calculations (either in a

Bayesian or in a maximum likelihood framework) were also developed (discussed below). However, any summary method, even ones that do not consider any explicit cause for gene tree discord, can be used to estimate a species tree from a collection of gene trees.

Under the multi-species coalescent model (Rannala and Yang 2003), the rooted species tree and branch lengths (in coalescent units) defines a probability distribution on gene trees, and the model species tree is *identifiable* from this distribution (Allman et al. 2011). Hence, statistically consistent methods exist for estimating the species tree, which take advantage of this fact. A guarantee of statistical consistency for a summary method means that the method is proven to return the true species tree with high probability given a sufficiently large number of true gene trees sampled from the distribution defined by the species tree. MDC and standard consensus methods (e.g., the greedy consensus) have been proven to not be statistically consistent (Degnan et al. 2009) under the multi-species coalescent model, meaning that there are some model conditions in which they will not return the true tree with high probability, even given an unbounded number of true gene trees. However, new summary methods that are proven to be statistically consistent under the multi-species coalescence model have been developed, including the population tree from BUCKy (Larget et al. 2010), GLASS (Mossel and Roch 2010), MP-EST (Liu et al. 2010), STAR (Liu et al. 2009), STEAC (Liu et al. 2009), and STEM (Kubatko et al. 2009).

Alternative statistically consistent species tree estimation methods have been developed that co-estimate gene trees and species trees from a set of sequence alignments; this set includes methods such as BEST (Liu 2008) and *BEAST (Heled and Drummond 2010). Co-estimation methods can have outstanding accuracy, but are computationally much more expensive to run, and so far have not been able to be used with hundreds of genes (Bayzid and Warnow 2013). Therefore, they cannot be used within genome-scale analyses.

Simulation studies comparing summary methods on multi-locus datasets with gene

tree incongruence due to ILS (DeGiorgio and Degnan 2010; Yang and Warnow 2011; Bayzid and Warnow 2013; Yang and Warnow 2011; Leaché et al. 2013) have revealed differences in accuracy and computational requirements. Comparisons between summary methods and concatenation have shown mixed performance: while concatenation can definitely be less accurate in the presence of ILS, (Edwards et al. 2007; DeGiorgio and Degnan 2010; Heled and Drummond 2010; Liu et al. 2010; Bayzid and Warnow 2013), Bayzid and Warnow (2013) have also shown that concatenation can be more accurate than even some statistically consistent coalescent-based methods even in the presence of large amounts of ILS. In general, however, not enough is known about the relative accuracy of concatenation and summary methods under biologically realistic conditions, and the conditions that impact the relative and absolute performance of methods.

In this paper we explore the accuracy of species tree estimation methods, including concatenation and several summary methods, on biological and simulated datasets. We specifically seek to understand how certain model conditions (number of genes, amount of ILS, and gene tree estimation error) affect the absolute and relative accuracy of species trees estimated using these methods.

Performance Study Design

Methods.— We focus our attention on MP-EST, one of the summary methods that has good accuracy and is fast enough to run on hundreds to thousands of genes, and which has been used to analyze several multi-marker datasets (Song et al. 2012; Chiari et al. 2012; Zhong et al. 2013; Kumar et al. 2013). MP-EST is a summary method based on maximizing pseudo-likelihood that takes a set of rooted gene treesand computes an estimated species tree as well as branch lengths in coalescent units. The proof of statistical consistency for MP-EST establishes that if the input is a large enough set of true gene trees

(sampled from the distribution defined by multi-species coalescent process applied to the model species tree), then with high probability MP-EST will return the true species tree.

In addition to MP-EST, we also evaluate a consensus method, two supertree methods, and concatenation. The consensus method we explore is the greedy consensus, also called the extended majority consensus; this technique is popular, but known to be statistically inconsistent under some circumstances (Degnan et al. 2009). The supertree methods we consider are Matrix Representation with Parsimony (MRP) (Ragan 1992) and Matrix Representation with Likelihood (MRL) (Nguyen et al. 2012). We also include concatenation using unpartitioned maximum likelihood analysis (CA-ML). While it is unknown whether MRP, MRL, or CA-ML are statistically consistent under the multi-species coalescent model, simulation studies have suggested that MRP (Wang and Degnan 2011) and concatenation (Kubatko and Degnan 2007) can give incorrect trees with high support.

We explored the performance of species tree methods under a wide range of model conditions, varying number of genes, amount of gene tree estimation error, and amount of ILS (43 model conditions in total). We also address the input given to a summary method. For example, MP-EST can be used with a single maximum likelihood (ML) tree estimate for each gene sequence alignment, or on a set of ML gene trees computed on bootstrap replicates of each gene sequence alignment. We refer to the first way of generating a set of gene trees as BestML, and hence we refer to the way of running MP-EST on the best found maximum likelihood gene trees as "MP-EST(BestML)". The second approach creates a set of estimated species trees (each computed using the summary method on some bootstrap replicate gene trees), from which a greedy consensus is then computed. We refer to this second type of approach as MLBS, for multi-locus bootstrapping (Seo 2008). Thus, MP-EST(MLBS) refers to the greedy consensus of the MP-EST species trees computed on the different sets of bootstrap replicates. Note our multi-locus bootstrapping procedure is

"site-only", meaning that genes are not re-sampled. Seo (2008) also introduced a mutli-locus bootstrapping procedure where first genes are re-sampled and then sites, and have shown it can sometimes work better than the site-only approach. However, re-sampling both genes and sites increases the computational burden. In this study, we will therefore focus on the site-only approach. We also explored the other summary methods using both BestML and MLBS gene trees; thus, Greedy(BestML), MRP(BestML), and MRL(BestML) refer to these summary methods applied to the best found maximum likelihood gene trees, and MRP(MLBS), MRL(MLBS), and Greedy(MLBS) refer to the greedy consensus of species tree estimations performed using the respective summary methods on multi-locus bootstrap replicate datasets.

Datasets.— We explore performance on both biological and simulated datasets. For biological datasets, we use the mammalian dataset from Song et al. (2012), three datasets studied by Salichos and Rokas (2013), and the amniota dataset of Chiari et al. (2012). Each of these datasets shows evidence for gene tree discord, and have a small enough number of species (at most 37) to be analyzed using MP-EST.

The simulation study is based on the mammalian dataset of Song et al. (2012), and contained 37 species and 447 loci. An MP-EST tree was calculated using ML gene trees that we estimated on the Song et al. (2012) dataset (see SOM Fig. S1), and we used this MP-EST tree as the basic model species tree. True gene trees were simulated down the model species tree according to the multi-species coalescent process (Rannala and Yang 2003). Branches in gene trees were rescaled to produce patterns corresponding to estimated gene trees of the real dataset, including divergence from the molecular clock. Sequence alignments were simulated down these gene trees according to the GTRGAMMA model. Thus, both true gene trees and gene trees estimated from sequence alignments were available. Different model conditions were created (see Table 1) to vary the number of

genes, the amount of gene tree estimation error, and the amount of ILS. To vary gene tree estimation error, the sequence length was adjusted, and to vary the amount of ILS, branch lengths in the model species tree were uniformly multiplied by two or five (to reduce ILS) or divided by half or one fifth (to increase ILS). See SOM Section S2 for more details.

Error Metric.— In our simulation studies, we quantify tree error using the missing branch rate (the percentage of branches in the true tree that do not appear in the estimated trees). In our experiments, all estimated species trees were fully resolved, and hence the missing branch rate, false positive branch rate, and normalized RF distance were all identical.

Results

Our simulation study is based on the mammal dataset of Song et al. (2012), which had 447 genes. Our first experiment explored methods on simulated datasets that were relatively close to this model tree – 25 to 800 genes, sequence lengths that varied from 500 to 1500, and branch lengths from five times to one-fifth the lengths of the basic model tree. The first experiment had the surpising result that MRL had better accuracy than MP-EST, a pattern we found difficult to believe would hold generally since MP-EST is statistically consistent and MRL is probably not. To explore this more carefully, we examined MP-EST, MRL, and Greedy on the very highest ILS condition we consider, and with up to 3200 true gene trees. On this high ILS model condition, MP-EST outperformed both MRL and Greedy on large numbers of true gene trees. Finally, we examined performance on biological datasets.

Experiment 1: Multi-locus Bootstrapping

Bootstrapping for multi-locus data can be done in many ways, for example, by re-sampling sites only, re-sampling both sites and genes, and employing strategies for

centering and pivoting support values, surveyed in Seo (2008). Here, we study a simple site-only re-sampling strategy that combines the i^{th} bootstrap replicate trees from all loci to produce each multiple bootstrap replicate input to the summary methods. The site-only re-sampling procedure has been used by some recent studies (Chiari et al. 2012; Kumar et al. 2013). The summary method is run separately on each bootstrap replicate input, and multiple bootstrap replicate species trees are obtained. This produces a set of estimated species trees, one for each set of bootstrap replicate gene trees, which can then be used to either estimate a species tree (for example, using a consensus method) or to draw support values on a given species tree estimate.

Experiment 1: MLBS vs. BestML.— We explored the impact of MLBS versus BestML on the datasets for Experiment 1. Figure 1 shows that for each summary method, and given amount of ILS and sequence length, MLBS gives better accuracy for small numbers of genes, and BestML is more accurate on larger numbers of genes. The actual transition point depends on the particular summary method and the gene tree accuracy (i.e., sequence length), and so can vary to some extent between model conditions. Also, although the differences in accuracy on small numbers of genes is somewhat small, it can be quite large for larger numbers of genes. The two approaches produced similar results when gene trees were accurately estimated (e.g. 1500bp), but the difference in species tree accuracy could be very large with less accurately estimated gene trees. For example, MP-EST(BestML) and MP-EST(MLBS) both had close to correct species tree constructions on 800 genes with 1500bp sequence alignments, but there were much larger differences on 800 genes with 500bp sequence alignments: MLBS had 6% error and BestML had 2.5% error. For this reason, we focus the remaining discussion of methods using BestML gene trees; results for MLBS analyses are provided in the supplementary material.

Experiment 1: Accuracy of branch support values.— Although species trees based on MLBS

gene trees typically were less accurate than species trees based on BestML gene trees, MLBS-based analyses have the potential to provide good estimations of branch support. We therefore evaluated whether branch support values computed using MLBS (but drawn on BestML topology) were indicative of the probability that a branch was correct.

We studied this question by binning all branch support values obtained throughout our entire set of experiments (e.g., various gene tree estimation error levels, number of genes, and ILS levels) into 19 different bins, and for each bin we computed the percentage of branches with support in that bin range that appeared in the model species tree. For example, all branches between 20% and 30% support were put into a bin, and we asked what percentage of those branches were correct; the ideal case is that this number is between 20% and 30%. Figure 2 shows support values computed using different summary methods on MLBS gene trees, and Pearson's correlation coefficients for each case.

Greedy had the best correlation at the higher support levels, but poorer correlation at the lower support levels. MRP and MRL were fairly good at most levels, but MRP was better than MRL. MP-EST correlated reasonably well at the highest support levels, and not very well at the lower support levels. Interestingly, all methods (and in particular MP-EST) tended to over-estimate the probability of being correct for the highest support levels, and under-estimate the probability of being correct at the lowest support levels. The under-estimation problem was very pronounced for MP-EST at the lowest support levels; for example, branches with support in the $[10, 20)$ range appeared in the true species tree 22% of the times for MRP, but 42% of the time for MP-EST.

The support values drawn on MP-EST trees showed some provocative patterns. Branches in the $[0, 10)$ support range had a much better chance of being correct than those in $[10, 20)$ (65% and 42% respectively). These odd patterns are not likely only a result of having limited samples. There were 74 branches that had support below 10% in MP-EST trees, and 48 of them were correct; this pattern points to a systematic bias in support

estimation using the site-only multi-locus bootstrapping procedure.

Pearson's correlation coefficients rank these methods as follows: MRP and MRL are the best (MRP correlation coefficient of 0.987 and MRL correlation coefficient of 0.985), Greedy behind these two (correlation coefficient of 0.964), and MP-EST in last place (correlation coefficient of 0.894). Despite clear biases, the support values obtained from the multi-locus bootstrapping procedure used with many different summary methods (Greedy, MRP, MRL, and MP-EST) may be reasonable enough in practice, in that when the purpose of branch support is to give a general idea of which branches are to be trusted, multi-locus bootstrapping seems to do a decent job. The most dramatic divergence from the ideal support values occurs on low support edges, especially those obtained using MP-EST. Thus, we believe multi-locus bootstrapping has practical value, but branches with low support should not be dismissed readily, and those with high support should be interpreted with care.

Experiment 1: MRL and MRP

MRL is an optimization problem that is derived from MRP (see Methods for full details). As with MRP, the set of input source trees is represented by a matrix over $\{0, 1, ?\}$ (with one column for each internal branch of each source tree). In the MRP analyses we performed, this matrix is then analyzed using maximum parsimony heuristics with all substitutions treated equally, using PAUP*. MRL uses the same input matrix, but analyzes it under a symmetric two-state maximum likelihood model (Nguyen et al. 2012).

In our experiments, MRL had better accuracy than the standard MRP technique in all cases, except with very high gene tree error (250bp alignments), where MRP was occasionally better than MRL (SOM Figure S2). Thus, MRL generally outperforms MRP. Hence, in the rest of this discussion we focus on comparing MRL to Greedy, MP-EST, and

concatenation. We focus on analyses using BestML gene trees here, but see SOM Figure S4 for results on MLBS gene trees.

In our simulations, the relative performance of methods depended on the model conditions (Fig. 3 and Fig. 4). We compare MRL to the other methods in turn.

Experiment 1: MRL vs. MP-EST.— When 800 gene trees with low estimation error (e.g. true gene trees, or those based on 1500bp alignments) were available, MRL and MP-EST both had perfect accuracy on all 5 replicates. As the amount of estimation error was raised by reducing alignment length, the error predictably went up for both methods; however, increases in gene tree estimation error produced larger increases in species tree estimation error for MP-EST than for MRL. When fewer than 800 genes were available, MRL performed better than MP-EST in almost all cases, regardless of the amount of gene tree error (the only exception was 100 genes with 1000bp alignments). We note that MRL had better accuracy than MP-EST even on true gene trees, except for the lowest ILS level where the two methods returned the true species tree; interestingly, the amount of ILS had only a small impact on the relative performance between these two methods. In general, the differences between MP-EST and MRL were lower with true gene trees, and higher with poorly estimated gene trees.

Experiment 1: MRL vs. Greedy.— The relative performance of MRL and Greedy depended on the number of genes. When only a few genes were available, MRL and Greedy had about the same accuracy. As the number of genes was increased, MRL started to consistently outperform Greedy. Furthermore, increasing ILS widened the gap between MRL and Greedy, but gene trees estimation error did not have a detectable impact on the relative performance of MRL and Greedy.

Experiment 1: MRL vs. Concatenation.— The relative performance of MRL and

concatenation depended on two factors: the gene tree resolution and amount of ILS. Given relatively accurate gene trees, MRL always outperformed concatenation. However, with reduced alignment length, and hence increased gene tree estimation error, concatenation outperformed MRL in many cases. Similarly, when the amount of ILS was reduced to its lowest level (5X branch length), concatenation was better than MRL. With increased ILS, the error went up for all methods, including MRL, but the error of concatenation trees increased faster than MRL trees.

Experiment 1: MP-EST

As we have seen, MRL typically outperformed MP-EST on these data for Experiment 1 (at most 800 genes, but varying amounts of ILS and gene sequence alignment lengths). We now compare MP-EST, which is the only statistically consistent method we studied, to Greedy, MRP, and concatenation.

The relative performance between Greedy and MP-EST depended on the model condition. On true gene trees, MP-EST was more accurate if there were many gene trees or the level of ILS was high enough, while Greedy was more accurate if the number of gene trees was very small. On estimated gene trees, MP-EST was more accurate if there were a large number of genes or if the ILS level was high; otherwise the methods were either close in performance or Greedy had an advantage over MP-EST. Relative performance of MRP and MP-EST (see SOM Fig. S3) similarly depended on the gene tree estimation error. With true or well-estimated gene trees, MP-EST was typically better, except when very few genes were available. As estimation error was increased, the pattern became more complicated, and the two methods each outperformed the other in some cases.

The comparison between MP-EST and concatenation is also interesting. Under the default ILS level, MP-EST was typically more accurate than concatenation except when gene tree error was high, or when few genes were available. Under low amounts of ILS,

concatenation was more accurate, and under high amounts of ILS MP-EST was more accurate.

In general, MP-EST did very well on large numbers of highly accurate (or true) gene trees, where it could be more accurate than most other methods (except MRL). However, under other conditions, simpler summary methods, including Greedy, could be more accurate. All these patterns are consistent with the fact that MP-EST is statistically consistent under the multi-species coalescence model used in our simulations, but Greedy is not, and concatenation and MRL may not be.

Experiment 2: High ILS and large numbers of genes

Results from the first experiment showed MRL being at least as accurate as MP-EST on all model conditions. This was an unexpected result, and one we felt was unlikely to be globally true - since MP-EST is proven statistically consistent, while MRL probably is not statistically consistent. Thus, we would expect MP-EST to have close to perfect accuracy for large enough numbers of accurate gene trees, and MRL to start to fail to recover the correct species tree under conditions of high ILS. Therefore, in this second experiment, we explicitly examined performance under the highest ILS condition (0.2X) in our study, allowed the number of genes to be very large (up to 3200), and considered only true gene trees; Figure 5 shows these results. Note that MRL has the best accuracy on up to 800 genes, but on larger numbers of genes MP-EST has the best accuracy (with up to 2% difference in accuracy). Thus, MP-EST can have the best accuracy of these methods, but the amount of true gene tree discordance, the number of genes, and their accuracy may all impact the relative accuracy of MP-EST and simpler summary methods.

Biological Dataset Results

We analyzed five biological datasets: the mammalian dataset analyzed by Song et al. (2012), three datasets analyzed by Salichos and Rokas (2013), and the amniota dataset studied by Chiari et al. (2012). We focus the discussion on MRL and MPEST.

Mammalian dataset.— Song et al. (2012) analyzed a dataset of 37 mammalian genes and 447 genes using MP-EST and concatenation. Their MP-EST analysis used a more extensive resampling multi-locus bootstrapping technique, where they also resampled genes as well as sites within sequence alignments. Therefore, their MP-EST analysis technique is not directly comparable to our analysis. They observed two interesting differences between concatenation and MP-EST using this more extensive version of MLBS: (1) concatenation put tree shrews (Scandentia) as sister to Glires (Rodentia/Lagomorpha), but MP-EST put them with primates with 99% support, and (2) concatenation put bats (Chiroptera) as sister to Cetartiodactyla, while MP-EST puts a Carnivora/Perissodactyla clade as sister to Cetartiodactyla, and bats as sister to this clade.

We re-analyzed this dataset, and found 21 genes that had mis-labeled species (since confirmed by the authors). We also identified two outlier genes that we decided to remove from the collection; thus 424 genes remained. We reanalyzed this subset of the dataset using MP-EST, MRL, and concatenation. Resulting trees were congruent in most branches, but also had interesting differences.

MRL(BestML), MRL(MLBS), and MP-EST(BestML) were all topologically identical, but differed from MP-EST(MLBS) in the position of tree shrews. In our analyses, MP-EST(MLBS) placed tree shrews *(Tupaia belangeri)* as sister to primates with 62% support. In the Song et al. (2012) MP-EST analysis, tree shrews were sister to primates with 99% support. Thus, our MP-EST(MLBS) tree differs from the Song et al. (2012) MP-EST tree. We confirmed the differences were not due to the 23 filtered genes. The most likely explanation is the differences in how the two studies used MLBS: we did

not resample genes, and they did.

We did not look at the relative accuracy of MLBS in which genes are also resampled, so we cannot say anything about how using MLBS in this more extensive way might impact accuracy. However, our study shows that that using BestML instead of MLBS resulted in improved species trees when the number of genes was sufficiently large for the model condition. Therefore, we think the MP-EST(BestML) analysis may be more accurate than MP-EST(MLBS) (as we performed it) and possibly more accurate than MP-EST as performed by Song et al. (2012). Furthermore, as noted above, our simulation studies have shown that multi-locus bootstrapping has a tendency to under-estimate support for some correct branches. Based on these observations, we hypothesize that tree shrews should be sister with Glires. This placement is consistent with our analyses using MP-EST(BestML), MRL(BestML), and MRL(MLBS), and also with the Song et al. (2012) analysis using concatenation. Support in the MRL(BestML) tree for this placement is 84%, which is reasonably high. Also, although the support for this placement is only 38% for MP-EST(BestML), our study showed that low support levels are often substantially underestimated by MP-EST. We conclude that the Song et al. (2012) dataset is more suggestive of tree shrews being sister to Glires rather than to primates, because this resolution is found with high support using concatenation and MRL, and with low support using MP-EST(BestML). However, this conclusion has to be interpreted with regards to this particular dataset. We note that tree shrews were placed as sister to primates in other studies that had better taxon sampling (Kumar et al. 2013), studies based on gene duplication and loss (Boussau et al. 2013), and studies that used rare genomic events (Janecka et al. 2007).

All MRL and MPEST trees differ from the concatenation tree in the position of bats (*Myotis lucifugus* and *Pteropus vampyrus*); while concatenation puts bats with Cetartiodactyla, both species tree methods have high support (95% and 92%) for placing

bats as sister to a (Cetartiodactyla,(Perissodactyla,Carnivora)) clade. In other words, this discordance between summary methods and the concatenation method is robust to the choice of the summary method (unlike the position of tree shrews). Thus, we claim our analyses as a whole prefer the placement of bats as sister to a clade containing Cetartiodactyla, Perissodactyla, and Carnivora.

Metazoa dataset.— Salichos and Rokas (2013) studied a metazoan dataset using concatenation. This dataset has very sparse taxon sampling, including only 20 species across the entire metazoa and one outgroup. In addition, relatively few genes (only 255) are available, and the gene trees have relatively low phylogenetic signal (average bootstrap support is 47%). This lack of support can be caused by poor phylogenetic signal in orthologs recovered for these distantly related taxa, or by problems associated with low taxon sampling. Given the small number of genes, many of which have low phylogenetic signal, it is possible that species trees estimated using MLBS instead of BestML may be appropriate. With all these caveats considered, we reanalyzed this dataset (see Fig. 7) and observed some interesting patterns.

The bootstrap support values were quite low for all the summary methods, making it clear that the summary methods were not able to resolve the metazoan phylogeny. The concatenation tree, in contrast, had high support on almost all branches. The concatenation tree was broadly congruent with the current knowledge about metazoan phylogeny, but also had some issues, discussed below.

Bilateria were recovered as a clade in all five trees reconstructed here, but interestingly, MP-EST and MRL trees had low support for this clade (respectively 52% and 62%). The sister to Bilateria was *N. vectensis* (representing Cnidaria) in the concatenation tree, but *N. vectensis* was grouped with *T. adhaerens* (representing Placozoa), and this clade, which itself had high support in MP-EST and MRL trees, was sister to Bilateria.

The relationship recovered in the concatenation tree is more in line with most molecular studies with better taxon sampling (Hejnol et al. 2009; Sperling et al. 2009; Philippe et al. 2009; Ryan et al. 2013). However, the relationship offered by summary method has also been previously implied in the literature (Schierwater et al. 2009).

Inside Bilateria, summary methods had difficulty recovering Chordates as a monophyletic clade; for example, *S. purpuratus* (representing Echinodermata) was placed inside Chordates in some analyses. The only summary method that recovered Chordates was MP-EST(MLBS), which had 72% support. The other summary methods did not recover Chordates, but had low support for its non-monophyly (28%, 29% and 55% respectively in MP-EST(BestML), MRL(BestML), and MRL(MLBS)). Note that concatenation recovers Chordates with only 59% support.

Inside Chordates, all methods easily recovered vertebrates as a clade, but again had trouble in identifying the sister to vertebrates. Based on the prior literature (Bourlat et al. 2006; Delsuc et al. 2006; Singh et al. 2009; Edgecombe et al. 2011), Urochordates (represented here by *C.intestinalis*) is strongly believed to be sister to the vertebrates, and this relationship was recovered only by the concatenation tree and MRL(MLBS). The remaining summary methods put *B. floridae* (representing Cephalochordates) as sister to Chordates. MRL(BestML) had only 36% support for this incorrect relationship, but both MP-EST trees had 100% support.

Another problematic taxon is *C.elegans* (representing Nematoda). The MP-EST(BestML) tree placed *C.elegans* outside Protostomia and as the sister to all Bilateria with 16% support - a placement that is clearly incorrect. MP-EST(MLBS) puts *C.elegans* at the base of Protostomia, which is better than its position in MP-EST(BestML); however, this placement is also likely incorrect, as the current best molecular evidence puts Nematoda and Arthropoda closer to each other than either to Platyhelminthes (represented here by *S. Mansoni*) (Lartillot and Philippe 2008;

Edgecombe et al. 2011). None of the remaining trees obtained here, not even the concatenation tree, put *C. elegans* as sister to Arthropoda; instead the concatenation and both MRL trees put it as sister to *S. Mansoni*.

To summarize, concatenation seems to produce more accurate species trees than other methods, and MRL seems to be more accurate than MP-EST. On this dataset, results of analyses using MLBS gene trees seem to be generally more reliable than analyses using BestML gene trees. However, our overall conclusion is that all methods, including concatenation, are unable to give a good resolution of the metazoan phylogeny using this dataset. We hypothesize that the challenges on this dataset result from the sparse taxon sampling, but that the relatively small number of genes and low phylogenetic signal per gene contributes to these challenges. Furthermore, the improved reliability of trees based on MLBS gene trees rather than BestML gene trees is also consistent with the observation that small numbers of genes (especially when there is poor phylogenetic signal) can be better analyzed using MLBS.

Vertebrates dataset.— We also studied the vertebrates dataset analyzed by Salichos and Rokas (2013). Similar to the metazoan dataset, this dataset had sparse and uneven taxon sampling, but includes 1087 genes, which is many times more than metazoa, and gene trees had on average 75% branch support, which is also much better than the metazoan dataset. Figure 8 shows trees obtained using MP-EST and MRL on the vertebrates dataset. The concatenation tree reported by Salichos and Rokas (2013) (not shown) was identical to MP-EST(BestML) and MRL(BestML) trees in terms of topology, but has 100% support everywhere. This topology did not violate any of the well-established clades in the vertebrate phylogeny, and was thus acceptable. The support values for both BestML trees were high for many but not all branches. In both trees, the resolution inside the *E. caballus*/*B. taurus*/*C.familiaris* (horse/cow/dog) clade had very low support, but this

relationship was uncertain and has been resolved in different ways even with much better taxon sampling (see Hu et al. (2012) for a comprehensive review). Thus lack of support might be appropriate for this particular relationship on this dataset. In general, MRL(BestML) had higher support values than MP-EST(BestML).

The only tree that was clearly problematic was MP-EST(MLBS), which contradicted established groupings in two ways: instead of orangutans (*P. pygmaues*), rhesus monkey (*M. Mulatta*) was sister to the human/gorilla clade, and also the well-established clade Euarchontoglires was not recovered. Also, the Perciformes/Tetraodontiformes clade has received support in a recent molecular study (Friedman et al. 2012), but the MP-EST(MLBS) tree has *O. latipes* as sister to *G. aculeatus* (representing Perciformes) with 84% support and hence contradicts (Friedman et al. 2012) while the other trees put *G. aculeatus* as sister to Tetraodontiformes and agree with (Friedman et al. 2012). Thus, on this dataset, summary methods used with BestML gene trees produced consistent and reliable results, but analyses using MLBS gene trees were not as reliable.

Yeast dataset.— We also analyzed the yeast dataset by Salichos and Rokas (2013), which had 23 species and 1070 loci. Here, most analyses were largely congruent and had high support (SOM Fig. S6). Trees obtained by both summary methods differed from each other and from the concatenation tree in one edge. MP-EST tree put *C. lusitaniae* with a *C. guiliermondii*/*D. hansenii* clade, but in the concatenation tree, *C. lusitaniae* was at the base of the so-called CTG group (Dujon 2010). MRL also put *C. lusitaniae* at the base of CTG group, but, unlike concatenation and MP-EST, did not recover *C. guiliermondii* and *D. hansenii* as sister taxa. All three trees differed from the topology given by Dujon (2010), which is meant to summarize the state of the knowledge about yeast evolution. In particular, no tree recovered the proposed Clavispora clade (*C. guiliermondii*,*C. lusitaniae*).

Thus, all trees were similar, and where they differed, none was more accurate than others.

Amniota dataset.— Chiari et al. (2012) assembled a dataset of 248 genes across 16 amniota taxa, with the goal of resolving the position of turtles relative to birds and crocodiles. Most recent molecular studies (Hugall et al. 2007; Iwabe et al. 2005; Zardoya and Meyer 1998) have recovered birds and crocodiles as sister groups (forming Archosaurs) and turtles as sister to this clade. Chiari et al. (2012) used MP-EST with a multi-locus bootstrapping procedure (identical to what we have studied) and on two sets of gene trees, one based on AA alignments, and the other based on nucleotides. Their analyses using MP-EST, somewhat disappointingly, resolved bird/turtle/crocodile differently when AA and DNA gene trees were used. The AA MP-EST tree, just like concatenation on either AA or DNA, put turtles as sister to Archosaurs with 99% support. The DNA MP-EST tree put turtle as sister to crocodiles with 90% support.

We obtained the gene trees from the authors and re-analyzed the dataset using MP-EST and MRL (see Fig. 9). Our MP-EST trees were all topologically identical to those obtained by Chiari et al. (2012) and had very similar bootstrap support values. Thus, MLBS and BestML gave identical results. However, MRL results were interesting in that both AA and DNA MRL trees put turtles as sister to Archosaurs (with 100% and 89% support respectively). Thus unlike MP-EST, the results of an MRL analysis of this dataset is not dependent on which set of gene trees are used.

The set of gene trees analyzed by Chiari et al. (2012) had mediocre support values (50% for AA and 65% for DNA, where support is drawn on the greedy consensus of 100 bootstrapped gene trees replicates). Therefore they had medium numbers of genes with mediocre support, a condition that based on our simulation studies does not predispose to high accuracy with MP-EST. Instead, MRL was found to have better accuracy under these conditions. Thus simulations favor MRL. Furthermore, the amino-acid and nucleotide

MRL analyses, the amino-acid MP-EST analyses, and concatenation are all in agreement with most of the previous literature in putting turtles as sister to Archosaurs. We therefore believe the (turtles,(birds,crocodiles)) hypothesis is much stronger, both based on the prior literature and these analyses of the dataset of Chiari et al. (2012).

Discussion

This study shows various trends, which we summarize. The first observation is that summary methods could be impacted by the choice of input gene tree distribution (i.e., either BestML or MLBS). Furthermore, using BestML produced more accurate trees if there was a sufficiently large number of genes, but MLBS was better if there was only a small number of genes, and the transition point between MLBS and BestML depended on the model condition. There were also cases where the choice between BestML and MLBS had little impact, including very highly accurate gene trees, or when the number of genes was in the "moderate" range. Nevertheless, the data does suggest that when the analysis is based on very large numbers of genes (and in particular, in a genome-scale study), then BestML may produce more accurate results compared to MLBS.

While there were many cases where gaps between summary methods were quite small (e.g., on true gene trees under relatively low ILS conditions), generally the difference between methods increased with the amount of ILS, and with estimation error in the gene trees. Also, although our evaluation on estimated gene trees was limited to 800 genes, on datasets with estimated gene trees, MRL typically outperformed other summary methods regardless of the amount of ILS. However, MP-EST outperformed all summary methods when analyzing large numbers of true gene trees under high ILS conditions, suggesting the possibility that MP-EST might also outperform summary methods when analyzing large numbers of very highly accurate estimated gene trees under very high ILS conditions.

The performance of concatenation is also interesting. Under very low ILS conditions, concatenation was often more accurate than summary methods. In other conditions, except for cases when gene trees are poorly estimated (due to short sequences), concatenation was typically not as accurate as MP-EST or MRL, though it might be more accurate than Greedy.

Thus, under conditions with moderate to high levels of ILS and a large enough number of sufficiently well estimated gene trees, the best summary methods produced more accurate species tree estimates than concatenation.

The results on biological datasets are consistent with these observations. On datasets with a sufficiently large number of gene trees (i.e., mammals, vertebrates, and yeast), summary methods applied to BestML gene trees produce trees that show the most concordance with established clades. We also observed differences in some analyses on these three datasets, but these were consistent with preferences for BestML gene trees over MLBS and (in some cases) with MRL over MP-EST. The metazoan dataset has a smaller number of genes relative to the model condition, and analyses based on MLBS were more compatible with the literature than analyses based on BestML; this is also consistent with our study.

Results on the amniota dataset are also interesting. Here, the choice of gene tree distribution (BestML or MLBS) has little impact on either MRL and MP-EST; this is consistent with an intermediate number of gene trees with low support. Most interesting, however, is that MRL gave more reliable results than MP-EST: MRL was independent of whether the dataset was nucleotide or amino-acid (in contrast with MP-EST), and was consistent with concatenation and the preferred hypothesis based on the previous literature regarding the placement of turtles relative to birds and alligators, while MP-EST on nucleotide gene trees contradicted this placement.

Taken together, the results on biological and simulated data demonstrate that the

best summary methods can produce highly accurate estimates of species trees under conditions with high ILS, given a large enough number of well estimated gene trees. However, these results also demonstrate that the summary methods we tested have some vulnerability to gene tree estimation error, which results in reduced accuracy when gene tree estimation error is high. Furthermore, MP-EST seems possibly more vulnerable to gene tree estimation error than MRL. That is, while MP-EST has outstanding performance when a large number of true gene trees is available - a result that follows from being statistically consistent - it can have poorer accuracy than MRL on estimated gene trees. We offer the following possible explanation for this hypothesis. MP-EST estimates species trees by finding parameters for the species tree that are likely to produce the observed (estimated) gene trees. Hence, when a substantial amount of the gene tree discordance is due to estimation error, the species tree estimated by MP-EST is more likely to have topological errors and shorter branches (in coalescent units). In other words, gene tree discord due to estimation error violates key assumptions of the MP-EST model, which assumes all discord is due to ILS. As a result, MP-EST under-estimated branch lengths and hence over-estimates the amount of ILS (see SOM Fig. S5), as observed in other simulation studies (Leaché et al. 2013). In contrast, MRL, since it is agnostic to the causes of gene tree discord, may not be misled by gene tree estimation error as much as MP-EST. The observations here regarding MP-EST suggest that other statistically consistent summary methods that use likelihood calculations under the coalescent model are also likely to have the same vulnerabilities. This study should therefore not be taken as a criticism of MP-EST, which has shown improved accuracy relative to other statistically consistent summary methods (STAR, STEAC, etc.). Instead, the study suggests a general vulnerability for summary methods, and which may be particularly significant for methods that use likelihood calculations based on an input set of estimated gene trees.

Recommendation

Taking all these observations into consideration, we make the following recommendations. First, gene tree estimation accuracy and resolution is important and has a large impact on species tree estimation accuracy. Therefore, every attempt should be made to obtain estimated gene trees that are highly accurate. It is also possible (although we did not consider this approach in this study) that screening genes and restricting only to the most reliable gene trees might improve summary methods - especially if they are vulnerable to gene tree estimation error. In particular, therefore, screening might be beneficial for MP-EST. However, screening data presents methodological challenges, since it can bias the results (the gene tree sample may no longer be drawn from the same distribution).

Second, because the choice between MLBS and BestML gene trees impacts accuracy, and this study suggests that the accuracy may be improved by using BestML gene trees (especially for large numbers of genes), we recommend that both types of analysis be employed and the resultant species tree estimates compared.

Third, we recommend that many approaches to species tree estimation be considered, including methods such as MP-EST that are statistically consistent, but also considering concatenation and simple summary methods such as MRL.

Summary

This study showed that all the methods we studied (including the simple summary methods and concatenation) could give reasonably accurate species tree estimations even under conditions with high amounts of ILS. However, there were significant differences between methods that revealed certain trends about how the model condition (number of genes, accuracy of gene trees, and amount of ILS) impacted the relative performance of methods. Thus, while concatenation was typically less accurate than the best summary

methods on these data, there were conditions (e.g., low ILS levels, or small numbers of poorly estimated gene trees under most ILS levels) where concatenation gave more accurate results than all summary methods. Other trends we observed included the surprising result that MRL, a simple summary method that has no statistical guarantees as far as we know, typically gave more accurate trees than MP-EST, and that MP-EST only was more accurate than MRL when there was a very large number of gene trees that evolved with very high levels of ILS.

In some basic sense what this study suggests is that summary methods are vulnerable to gene tree error and small numbers of genes. This vulnerability seems to be true of all summary methods, but may be particularly true for likelihood-based summary methods. On the other hand, when gene trees are highly accurate and ILS levels are sufficiently high, then the best summary methods are typically more accurate than concatenation. Therefore, the choice of whether to use a summary method or concatenation, and which summary to use, will depend on the dataset conditions.

Thus, this study confirms the hypothesis that the best species tree accuracy can be obtained using methods that can account for gene tree heterogeneity rather than through standard techniques like concatenation. However, it also suggets the advantage of statistically consistent coalescent-based species tree estimation relative to concatenation may not be obtained on datasets with small numbers of estimated gene trees – instead, the real promise of these methods may lie in genome-scale datasets.

Methods

Gene tree estimation

We used RAxML version 7.3.5 to estimate all the gene trees under the GTRGAMMA model, with 200 replicates of bootstrapping. For BestML gene trees, we ran RAxML 20

times, and took the best scoring tree. For biological datasets of Salichos and Rokas (2013) and Chiari et al. (2012), 100 replicates of bootstrapped gene trees were provided to us by the authors, which we used. For the mammalian dataset, we re-estimated gene trees using the same procedure as our simulation study (see SOM Section S2).

Greedy

The greedy (or extended majority) consensus (referred to as Greedy) orders all the bipartitions in the source trees according to their frequency, and adds them one by one to the output tree if they do not have conflicts with already added bipartitions. We used Dendropy (Sukumaran and Holder 2010) to implement the greedy consensus.

MRP

MRP, or Matrix Representation with Parsimony (Ragan 1992), is a standard supertree method that operates by encoding its set of input trees as a matrix over $\{0, 1, ?\}$ characters. Thus, each bipartition in each tree is encoded as column by assigning 0 to all taxa in one side and 1 to taxa on the other side (? is used for missing taxa). This matrix is then analyzed using maximum parsimony, treating all substitutions equally. We used a custom Java code, available at `https://github.com/smirarab/mrpmatrix` for creation of the MRP matrix from gene trees, and used PAUP* for the parsimony analysis, using the standard heuristic search (see SOM Section S2).

MRL

Matrix Representation with Likelihood (MRL) (Nguyen et al. 2012) is similar to MRP, except, once the matrix is built, it is analyzed using maximum likelihood under a symmetric binary model of sequence evolution. This application of the likelihood model to

MRP matrices lacks any theoretical justifications, but a previous simulation study showed that MRL can outperform MRP (Nguyen et al. 2012). We build the data matrix for MRL using the same Java code used for MRP, noting that it is important to randomize the matrix representation in order to achieve good accuracy using MRL. That is, for each bipartition, we randomly decide which side should be 0, and which side should be 1. We analyze this data matrix using RAxML version 7.3.5 under the BINCAT model.

MP-EST

We used version 1.3 of MP-EST for all the analyses. We run MP-EST 10 times, and selected the species tree with the best likelihood score.

Multi-locus bootstrapping procedure

Given n genes, and with the objective of creating m replicate datasets, we perform the following steps. First, gene trees are estimated for each of the n loci using maximum likelihood. We use the non-parametric bootstrapping procedure Felsenstein (1985) to create m replicate datasets for each gene sequence alignments. This produces bootstrap replicates trees $t_{i,j}$ for $1 \leq i \leq n$ and $1 \leq j \leq m$. Then, the summary method is run m times, each time on all j^{th} bootstrap replicates of all n loci. Thus, the set $\{t_{1,1}, t_{2,1}, \ldots, t_{n,1}\}$ is one input, $\{t_{1,2}, t_{2,2}, \ldots, t_{n,2}\}$ is another input, and so on. This procedure produces m bootstrapped estimates of the species tree, and can be used to draw support values on a given species tree. Other ways of performing multi-locus bootstrapping include re-sampling genes as well as sites, and thus can be more extensive; see Seo (2008) for more discussion.

Concatenation

All concatenation analyses on simulated datasets were performed using an unpartitioned GTRGAMMA analysis using RAxML. RAxML was run 10 times and the best likelihood tree was used (see SOM Section S2). We did not perform bootstrapping for concatenation runs of the simulated datasets. For biological datasets, concatenation results were available from respective publications (also without partitioning). We re-ran an unpartitioned concatenation for the mammalian dataset after removing the 23 problematic genes, and saw no real differences with the concatenation tree reported by Song et al. (2012).

Acknowledgments

The authors thank Bastien Boussau who allowed us to use gene trees and sequence alignments he simulated in this paper. TW was supported by NSF grant DEB DBI-1062335. SM was supported by a Howard Hughes International Predoctoral Fellowship. MSB was supported by the University of Alberta through a subaward to the University of Texas.

*

References

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

Bayzid, M. S. and T. Warnow. 2012. Estimating optimal species trees from incomplete gene trees under deep coalescence. Journal of Computational Biology 19:591–605.

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

Bourlat, S. J., T. Juliusdottir, C. J. Lowe, R. Freeman, J. Aronowicz, M. Kirschner, E. S. Lander, M. Thorndyke, H. Nakano, A. B. Kohn, A. Heyland, L. L. Moroz, R. R. Copley, and M. J. Telford. 2006. Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. Nature 444:85–88.

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

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

DeGiorgio, M. and J. H. Degnan. 2010. Fast and consistent estimation of species trees using supermatrix rooted triples. Molecular Biology and Evolution 27:552–69.

Degnan, J., M. DeGiorgio, D. Bryant, and N. Rosenberg. 2009. Properties of consensus methods for inferring species trees from gene trees. Systematic Biology 58:35–54.

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

Delsuc, F., H. Brinkmann, D. Chourrout, and H. Philippe. 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature 439:965–968.

Dujon, B. 2010. Yeast evolutionary genomics. Nature Reviews: Genetics 11:512–524.

Edgecombe, G. D., G. Giribet, C. W. Dunn, A. Hejnol, R. M. Kristensen, R. C. Neves, G. W. Rouse, K. Worsaae, and M. V. Sørensen. 2011. Higher-level metazoan relationships: recent progress and remaining questions. Organisms Diversity & Evolution 11:151–172.

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

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

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

Friedman, M., K. L. Kuhn, J. A. Moore, W. L. Smith, P. C. Wainwright, M. P. Davis, R. I. Eytan, T. J. Near, and A. Dornburg. 2012. Resolution of ray-finned fish phylogeny and timing of diversification. Proceedings of the National Academy of Sciences 109:13698–13703.

Hejnol, A., M. Obst, A. Stamatakis, M. Ott, G. W. Rouse, G. D. Edgecombe, P. Martinez, J. Baguñà, X. Bailly, U. Jondelius, et al. 2009. Assessing the root of bilaterian animals

with scalable phylogenomic methods. Proceedings of the Royal Society B: Biological Sciences 276:4261–4270.

Heled, J. and A. J. Drummond. 2010. Bayesian inference of species trees from multilocus data. Molecular Biology and Evolution 27:570–80.

Hu, J.-Y., Y.-P. Zhang, and L. Yu. 2012. Summary of Laurasiatheria (mammalia) phylogeny. Dongwuxue Yanjiu 33:E65–74.

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

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

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

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

Kubatko, L. S., B. C. Carstens, and L. L. Knowles. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25:971–973.

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

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

Lartillot, N. and H. Philippe. 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Philosophical Transactions of the Royal Society of London. Series B, Biological sciences 363:1463–1472.

Leaché, A. D., R. B. Harris, B. Rannala, and Z. Yang. 2013. The Influence of Gene Flow on Species Tree Estimation: A Simulation Study. Systematic Biology 0:1–14.

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

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

Liu, L., L. Yu, D. K. Pearl, and S. V. Edwards. 2009. Estimating species phylogenies using coalescence times among sequences. Systematic Biology 58:468–77.

Maddison, W. and L. Knowles. 2006. Inferring phylogeny despite incomplete lineage sorting. Systematic Biology 55:21–30.

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

Mossel, E. and S. Roch. 2010. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. IEEE/ACM Transactions on Computational Biology and Bioinformatics 7:166–71.

Nakhleh, L. 2013. Computational approaches to species phylogeny inference and gene tree reconciliation. Trends in Ecology and Evolution 28:719–728.

Nguyen, N., S. Mirarab, and T. Warnow. 2012. MRL and SuperFine+ MRL: new supertree methods. Algorithms for Molecular Biology 7:3.

Philippe, H., R. Derelle, P. Lopez, K. Pick, C. Borchiellini, N. Boury-Esnault, J. Vacelet, E. Renard, E. Houliston, E. Quéinnec, et al. 2009. Phylogenomics revives traditional views on deep animal relationships. Current Biology 19:706–712.

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

Rannala, B. and Z. Yang. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645–1656.

Ryan, J. F., K. Pang, C. E. Schnitzler, A.-D. Nguyen, R. T. Moreland, D. K. Simmons, B. J. Koch, W. R. Francis, P. Havlak, S. A. Smith, N. H. Putnam, S. H. D. Haddock, C. W. Dunn, T. G. Wolfsberg, J. C. Mullikin, M. Q. Martindale, and A. D. Baxevanis. 2013. The genome of the ctenophore mnemiopsis leidyi and its implications for cell type evolution. Science 342:1242592+.

Salichos, L. and A. Rokas. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327–31.

Schierwater, B., M. Eitel, W. Jakob, H.-J. Osigus, H. Hadrys, S. L. Dellaporta, S.-O. Kolokotronis, and R. DeSalle. 2009. Concatenated analysis sheds light on early metazoan evolution and fuels a modern urmetazoon hypothesis. PLoS biology 7:e1000020.

Seo, T.-K. 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. Molecular Biology and Evolution 25:960–971.

Singh, T. R., G. Tsagkogeorga, F. Delsuc, S. Blanquart, N. Shenkar, Y. Loya, E. J. Douzery, and D. Huchon. 2009. Tunicate mitogenomics and phylogenetics: peculiarities of the Herdmania momus mitochondrial genome and support for the new chordate phylogeny. BMC Genomics 10:534.

Song, S., L. Liu, S. V. Edwards, and S. Wu. 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 109:14942–7.

Sperling, E. A., K. J. Peterson, and D. Pisani. 2009. Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa. Molecular Biology and Evolution 26:2261–74.

Sukumaran, J. and M. Holder. 2010. Dendropy: a Python library for phylogenetic computing. Bioinformatics 26:1569–71.

Than, C. and L. Nakhleh. 2009. Species tree inference by minimizing deep coalescences. PLoS Computational Biology 5.

Wang, Y. and J. Degnan. 2011. Performance of Matrix Representation with Parsimony for inferring species from gene trees. Statistical Applications in Genetics and Molecular Biology 10:1–39.

Yang, J. and T. Warnow. 2011. Fast and accurate methods for phylogenomic analyses. BMC Bioinformatics 12:S4.

Yu, Y., T. Warnow, and L. Nakhleh. 2011. Algorithms for MDC-based multi-locus phylogeny inference: beyond rooted binary gene trees on single alleles. Journal of Computational Biology 18:1543–59.

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

Zhong, B., L. Liu, Z. Yan, and D. Penny. 2013. Origin of land plants using the multispecies coalescent model. Trends in plant science 18:492–495.

ILS	Sequence length	Number of genes	GT error	BS	GT discordance
0.2X	true gt	100,200,400,800,1600,3200	0%	NA	79%
0.2X	500bp	200	26%	64%	79%
0.5X	truegt	200	0%	NA	54%
0.5X	500bp	200	26%	64%	54%
1X	true gt	25,50,100,200,400,800	0%	NA	32%
1X	250bp	25,50,100,200,400,800	42%	46%	32%
1X	500bp	25,50,100,200,400,800	27%	63%	32%
1X	1000bp	25,50,100,200,400,800	16%	79%	32%
1X	1500bp	25,50,100,200,400,800	12%	84%	32%
2X	true gt	200	0%	NA	18%
2X	500bp	200	27%	64%	18%
5X	true gt	200	0%	NA	9%
5X	500bp	200	26%	63%	9%

Table 1: **Empirical statistics of the simulated mammalian datasets.** Model conditions with varying levels of ILS (first column), varying sequence length (second column), and varying number of genes (third column) are simulated. Sequence length is varied to produce different levels of gene tree estimation error, shown in the fourth column, measured as average RF distance between estimated gene trees and true gene trees. 1X is unscaled branch length condition, corresponding to the species tree estimated on the mammalian dataset using MP-EST; rescaled lengths that are larger reduce ILS, and shorter lengths increase ILS. Various levels of gene tree estimation error, which are not measurable on biological datasets, result in various levels of bootstrap support (BS, shown in the fifth column), which are measurable on the real data. The mammalian dataset had mean BS of 71%, putting it in between the 1X 500bp and 1000bp model conditions. ILS results in true gene trees that differ from the species tree, and changing the ILS level changes the amount of this discordance, shown in the sixth column, measured as the RF distance between the true model species tree and true gene trees.

Figure 1

Figure 1: **MLBS vs. BestML gene tree strategies on the mammalian datasets.** We compare species trees estimated using the two types of gene tree inputs: best maximum likelihood estimates of the gene sequence alignments (BestML) analyzed using the summary method, or the greedy consensus of the species trees estimated using the summary methods on the multi-locus bootstrap replicate datasets (MLBS). Tree error is measured using the missing branch rate (i.e. false negative rate), which was always identical to the RF rate on these data. Average error is shown over 20 replicates for model conditions with 25-200 genes, 10 replicates for 400 genes, and 5 replicates for 800 genes (See SOM Table S1 for standard deviation).

Figure 2

Figure 2: **Accuracy of multi-locus bootstrapping in calculating branch support values**. Each box shows the aggregated results on all model conditions of the mammalian simulated dataset. Support values are binned with the following breaks: $0, 10, 20, 30, 40, 50, 55, 65, 75, 80, 85, 90, 92, 94, 96, 98, 100$; with each bin including the right value and excluding the left value (and thus the last bin includes only branches with 100% support). For each bin, figures show the percentage of branches in the estimated species trees that were correctly estimated. The diagonal $y = x$ lines show the ideal scenario.

Figure 3

Figure 3: **Comparing performance of concatenation and summary methods on mammalian simulated datasets.** We compare three summary methods (MRL, Greedy, and MP-EST) on BestML gene trees and also concatenation using RAxML. The impact of changing the number of genes and the alignment length (and hence gene tree estimation error) is shown. The amount of ILS is fixed at the 1X level. (See SOM Fig. S4 for similar results on MLBS trees.) Tree estimation error is computed using the missing branch rate with respect to the model species tree. Average error over multiple replicate runs is shown (See SOM Table S1 for standard deviation). We had 20 replicates for model conditions with 25 to 200 genes, 10 replicates for 400 genes, and 5 replicates for 800 genes.

Figure 4

Figure 4: **Impact of ILS level on species tree estimation.** We compare the performance of MRL with Greedy, MP-EST, and concatenation using RAxML on the simulated mammalian dataset using true gene trees or BestML gene trees. (See SOM Fig. S4 for similar results on MLBS trees.) Tree estimation error is computed using the missing branch rate with respect to the model species tree. We show the average error over 20 replicates of 200 genes using true gene trees and estimated gene trees from 500bp alignments (See SOM Table S1 for standard deviation).

Figure 5

Figure 5: **Comparing performance of various methods on up to 3200 true gene trees, simulated under 0.2X ILS level (i.e. very increased levels of ILS)**. Results for 100 and 200 genes are over 20 replicates, 400 genes over 10 replicates, and the rest over 5 replicates (See SOM Table S1 for standard deviation). With very large numbers of true gene trees, MP-EST is more accurate than MRL.

Figure 6

Figure 6: **Mammals biological dataset analyses.** Edges with no label have 100% support. Edges that differ across various analyses are in pink.

Figure 7

Figure 7: **Metazoa biological dataset analyses.** Edges with no label have 100% support. Edges that are in violation of the concatenation tree are in pink.

Figure 8

Figure 8: **Vertebrates biological dataset analyses.** Edges with no label have 100% support. Concatenation (not shown) is identical to MP-EST(BestML) and MRL(BestML) in terms of topology, but has 100% support everywhere. Edges that are in violation of the concatenation tree are in pink.

Figure 9

Figure 9: **Amniota biological dataset analyses.** The edges concerning the resolution of birds/crocodiles/turtles are given in pink. MP-EST trees are topologically identical to results of Chiari et al. (2012); note that the MP-EST trees on AA and NT have different topologies. MRL gives topologically identical trees under all variations (NT or AA, BestML or MLBS) on these data, with changes only in the bootstrap support. Turtles are placed differently in the MP-EST tree on the nucleotides than in the other trees. All MRL trees, and the MP-EST tree on the amino-acid data, put turtles as sister to Archosaurs (birds and crocodiles). In contrast, MP-EST on the nucleotide data puts them as sister to crocodiles (alligator and caiman).