Systematic Biology Advance Access published August 26, 2014

RH: Evaluating species tree methods for ILS

Evaluating summary methods for multi-locus species tree estimation in the presence of incomplete lineage sorting

Siavash Mirarab[1], Md Shamsuzzoha Bayzid[1], and Tandy Warnow[1,2]

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

[2] *Departments of Bioengineering and Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA.*

Corresponding author: Tandy Warnow, Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA; E-mail: warnow@illinois.edu

Abstract.— Species tree estimation is complicated by processes, such as gene duplication and loss and incomplete lineage sorting (ILS), 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 MP-EST – one of the most popular species tree estimation methods designed to address ILS – 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 be more accurate than the best summary methods in some cases (mostly when the gene trees have poor phylogenetic signal or when the level of ILS is low), but summary methods are generally more accurate than concatenation when there are an adequate number of sufficiently accurate gene trees. Our study suggests that coalescent-based species tree methods may be key to estimating highly accurate species trees from multiple loci.
(Keywords: incomplete lineage sorting, species tree estimation, supertree methods, consensus methods, concatenation, gene tree discordance, multilocus bootstrapping, MP-EST, MRP, MRL)

The estimation of species trees from multiple genes 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. Earlier summary methods were based on the MDC (minimize deep coalescence) optimization criterion (Maddison 1997; Maddison and Knowles 2006; Than and Nakhleh

2009; Yu et al. 2011; Bayzid and Warnow 2012), a method that is statistically inconsistent (Than and Rosenberg 2011). However, statistically consistent methods for estimating species trees under the multi-species coalescent model have been developed recently.

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 rooted gene trees, and the rooted model species tree is *identifiable* from the distribution of rooted or unrooted gene trees (for five taxa or more) (Allman et al. 2011). Hence, statistically consistent estimation of the species tree can be performed using summary methods, which operate by combining gene trees. Statistically consistent summary methods include ASTRAL (Mirarab et al. 2014), 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). Statistically consistent species tree estimation methods, such as BEST (Liu 2008) and *BEAST (Heled and Drummond 2010), have been developed that co-estimate the gene trees and species tree from a set of sequence alignments. These 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 (Knowles et al. 2012; Bayzid and Warnow 2013; Smith et al. 2014). Therefore, they cannot be used for 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; Leaché and Rannala 2011; Bayzid and Warnow 2013) have revealed differences in accuracy and computational requirements. Comparisons between summary methods and concatenation have shown mixed performance: while concatenation can 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), under different model conditions, concatenation can have lower missing branch rates than some statistically consistent coalescent-based methods, even in the presence of substantial amounts of ILS (DeGiorgio and Degnan 2010; Kimball et al. 2013; Bayzid and Warnow 2013; McCormack et al. 2013). 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 different techniques for estimating species trees from multiple loci. We focus on MP-EST, one of the leading statistically consistent summary methods, but also explore concatenation under maximum likelihood (CA-ML), the greedy consensus, and two supertree methods (MRP (Ragan 1992) and MRL (Nguyen et al. 2012)). We specifically seek to understand how the number of genes, amount of ILS, and gene tree estimation error (as impacted by sequence length) affect the absolute and relative accuracy of species trees estimated using these methods. We also evaluate the effectiveness of multi-locus bootstrapping procedures (Seo 2008).

Methods

Datasets

We explored performance on biological and simulated datasets. For biological datasets, we used the mammalian dataset from Song et al. (2012) and the Amniota dataset of Chiari et al. (2012), each of which show evidence for gene tree discord consistent with ILS.

We generated a collection of simulated datasets. The model species tree was based on the mammalian dataset of Song et al. (2012), containing 37 species and 447 orthologous loci, and was computed using MP-EST on ML gene trees we estimated on the Song et al. (2012) gene sequence alignments (see Fig. S1 in Supplementary Online Material (SOM) available at `http://dx.doi.org/10.5061/dryad.310q3`). Gene trees were simulated down the model species tree under the multi-species coalescent process (Rannala and Yang 2003). The branch lengths of these gene trees were rescaled to expected substitution units so that they produced patterns corresponding to estimated gene trees of the biological dataset, including divergence from the strict molecular clock. For terminal branches, we made their lengths equal to the corresponding branch from a randomly selected gene tree estimated on the biological dataset. We then ordered internal branches in the biological and simulated gene trees by their lengths, and rescaled the simulated gene tree branches in each percentile to match the length of those in the same branch length percentile of the biological gene trees. Sequences were evolved down these true gene trees under the GTRGAMMA model.

Different model conditions were created to vary the number of genes, the sequence length, and the amount of ILS. The sequence length was varied from 250bp to 1500bp to generate model conditions that had varying levels of average observed gene tree estimation error. Note that gene tree estimation error is a function of other factors besides sequence length (e.g., the rate of evolution and branch lengths); however, we used sequence length because it provides a straightforward mechanism for modifying gene tree estimation error. We measured the normalized Robinson-Foulds (RF) distance (Robinson and Foulds 1981) between true gene trees and the estimated gene trees, and used the overall average RF distance to refer to each model condition; since the average is over all replicates, it hides the inter- and intra-variance among replicates, which are shown in Figure S2.

Under the default (1X) ILS model conditions, the average observed gene tree estimation error (or AGE for short) was 42%, 27%, 16%, and 12%, respectively, for 250bp, 500bp, 1000bp, and 1500bp sequences (Table 2). Hence, the 250bp model condition is a 42% AGE condition, and the 0% AGE condition refers to the case where true gene trees were used. When we change the amount of ILS, the AGE value for 500bp alignments is always either 26% or 27% (Table 2); however, for simplicity, we always refer to the 500bp model condition as the 27% AGE level. See SOM Section S2 for more details.

To decrease or increase the amount of ILS, branch lengths in the model species tree were uniformly multiplied or divided by two or five; this resulted in five model conditions that varied from relatively low levels to very high levels of ILS (Table 1). For example, with 5X branches, true gene trees had only 9% average RF distance to true species trees, while, with 0.2X branches, the average RF distance was 79% (Table 1).

Multi-locus bootstrapping

Summary methods can be used with a single maximum likelihood (ML) tree estimate for each gene, or with a set of the ML gene trees estimated for the bootstrap replicates of each gene. We refer to the first way of generating a set of gene trees as BestML. We use MP-EST(BestML) to refer to MP-EST run on the set of best ML gene trees (one tree per gene), and we refer to other methods, similarly.

We refer to the second approach for generating estimated gene trees as MLBS, for multi-locus bootstrapping (Seo 2008). Given n genes, to create m replicate datasets, we perform the following steps. First, we use the non-parametric bootstrapping procedure (Felsenstein 1985) to create m pseudo-replicate datasets for each gene sequence alignment. We then estimate gene trees for each of these pseudo-replicate alignments using maximum likelihood, thus producing bootstrap gene trees $t_{i,j}$ for $1 \leq i \leq n$ and $1 \leq j \leq m$. Then, for $j = 1$ up to m, the preferred summary method (e.g., MP-EST, MRL, etc.) is run on the j^{th} bootstrap replicates of all n loci (i.e., $\{t_{1,j}, t_{2,j}, \ldots, t_{n,j}\}$ is one input to the summary method). This procedure produces m bootstrapped estimates of the species tree. A greedy consensus (also known as the extended majority consensus tree) of these m replicate trees can then be used as the estimate of the species tree. We refer to this greedy consensus tree as the MLBS tree, and we note that it depends on the summary method used to generate the bootstrap replicate species trees. (Thus, there is an MP-EST(MLBS) tree, a Greedy(MLBS) tree, etc.) See SOM Section S2.5 for additional details.

The MLBS approach we used only re-sampled sites. However, Seo (2008) introduced different approaches for multi-locus bootstrapping, including re-sampling both sites and genes (where first genes are re-sampled and then sites). Here, we focus on the site-only re-sampling strategy, which was also used in recent studies (Chiari et al. 2012; Kumar et al. 2013). However, we present an experiment on one model condition where we examine gene/site resampling in comparison to the site-only resampling technique, and we also show results for both strategies on the mammalian biological dataset.

Error metric

We quantified species tree error using the missing branch rate (the percentage of branches in the true species tree that do not appear in the estimated species tree). All estimated species trees in these experiments were fully resolved, and hence the missing branch rate, false positive branch rate, and normalized RF distance were all identical. In addition, we divided species tree branches into five distinct length categories and report the error separately for each category. Branch lengths below 0.1 coalescent units have been identified as susceptible to high levels of ILS (Kubatko and Degnan 2007). Accordingly, we used these branch length categories: very short (below 0.1 coalescent units), short (between 0.1 and 0.25), medium (between 0.25 and 0.625), long (between 0.625 and 1.5625), and very long (above 1.5625). The number of branches in each category depended on the model condition (see Table 1). In the 0.2X condition, we have adjacent branches that were very short (Fig. S1), creating conditions that may result in the anomaly zone (i.e., where the most frequent gene tree will not be topologically identical to the species tree) (Rosenberg 2013).

Our focus in this paper is the accuracy of the species tree topology, rather than other aspects – such as species tree branch length estimation or branch support values. In particular, the missing branch rate metric we used ignores branch support. However, we also evaluated whether branch support values computed using MLBS but drawn on the

BestML tree topology were indicative of the probability that a branch was correct. We studied this question by binning all branch support values 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%. For each method, we calculated Pearson's correlation between support values and frequency of correctness.

Species tree methods explored

We evaluated MP-EST, two supertree methods (MRP and MRL), a consensus method (Greedy), and concatenation (CA-ML). For biological and simulated mammalian datasets, we used RAxML version 7.3.5 to estimate all the gene trees under the GTRGAMMA model, with 20 runs for BestML and 200 replicates of bootstrapping (see SOM Section S2). For the biological datasets of Chiari et al. (2012), we used the set of 100 replicates of bootstrapped gene trees provided to us by the authors.

MP-EST.— MP-EST uses a heuristic search to find a species tree that optimizes a maximum pseudo-likelihood score, given a set of rooted gene trees. MP-EST is statistically consistent under the multi-species coalescent model, has good accuracy, and is fast enough to run on hundreds to thousands of genes; for these reasons, it 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). We used version 1.3 of MP-EST for all the analyses, and ran MP-EST 10 times, returning the species tree with the best pseudo-likelihood score.

Greedy.— We included the greedy consensus (implemented in Dendropy (Sukumaran and Holder 2010)) as a summary method. Although the greedy consensus of gene trees is not statistically consistent under the multi-species coalescent model (Degnan et al. 2009), it is a simple and fast method.

MRP.— Matrix Representation with Parsimony (Ragan 1992), or MRP, is a standard supertree method that operates by encoding its set of input trees as a matrix over $\{0, 1, ?\}$ characters. Each bipartition in each tree is encoded as a column in this matrix by assigning 0 to all taxa on one side, 1 to taxa on the other side, and ? to the missing taxa. This matrix is then analyzed using maximum parsimony treating all substitutions equally. Although it is not yet known if MRP is statistically consistent, simulations suggest it is not (Wang and Degnan 2011). We used custom Java code, available at `https://github.com/smirarab/mrpmatrix`, to compute the MRP matrix from the gene trees, and used PAUP* (Swofford 2003) for the parsimony analysis with the standard heuristic search.

MRL.— Matrix Representation with Likelihood (MRL) (Nguyen et al. 2012) is similar to MRP except that once the MRP 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 justification, but a previous simulation study (Nguyen et al. 2012) explored MRL and MRP for supertree estimation (without any ILS) and showed that MRL can produce more accurate trees than MRP. We built the data matrix for MRL using the same Java code used for MRP, randomly choosing for each bipartition 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.

CA-ML.— Concatenation using maximum likelihood (CA-ML) is a widely used method; however, simulation studies have strongly suggested that CA-ML is statistically inconsistent and can give incorrect trees with high support (Kubatko and Degnan 2007).

All CA-ML analyses were performed using an unpartitioned GTRGAMMA analysis using RAxML (best of 10 runs). We did not perform bootstrapping for CA-ML on the simulated data due to computational challenges.

Research Questions

We explored the following questions in our simulation study:

Question 1 Which set of input gene trees (BestML or MLBS) produces species trees with lower missing branch rates?

Question 2 Are support values obtained from the multi-locus bootstrapping approach predictive of accuracy?

Question 3 How do methods compare to each other under different model conditions?

Question 4 How well do summary methods perform in estimating rapid radiations, given very large numbers of genes?

By definition, the MLBS tree includes all the branches in the BestML tree that have support greater than 50%, but highly supported branches in the MLBS tree could be absent from the BestML tree. Therefore, it is interesting to ask which approach gives a better point estimate of the species tree (Question 1), and whether support values drawn on BestML trees using the MLBS approach are reliable (Question 2). However, the main objective of this study is comparing the topological accuracy of different methods under various model conditions (Question 3). One of the specific questions we addressed is how accurately different methods estimate the species tree in the presence of rapid radiations, which are considered particularly difficult, partly due to the very high levels of ILS they generate (Whitfield and Lockhart 2007). Furthermore, genome-scale data may be required to reconstruct rapid radiations with high accuracy (Rokas et al. 2003; Wolf et al. 2002). Thus, we also specifically examined conditions where the amount of ILS is very high (e.g., as a result of rapid radiation) and thousands of genes are also available (Question 4).

We built three collections of simulated datasets. For the first collection (used for Questions 1-3), we fixed ILS to the 1X level and varied the number of genes (25 to 800) and gene tree AGE levels (0% to 42%). In the second collection (used for Questions 2 and 3), we used 200 genes and varied the amount of ILS (5X to 0.2X), using both true gene trees and estimated gene trees with 27% AGE level. Finally, for the third collection (used for Question 4), we fixed ILS to the highest level (0.2X) and varied the number of genes from 100 to 3200, using both true gene trees and estimated gene trees with 27% AGE level. In the first two collections, we created 20, 10 and 5 replicates, respectively, for datasets with up to 200, 400, and 800 genes, while for the last collection we created 20, 10 and 5 replicates for datasets with up to 800, 1600, and 3200 genes, respectively.

We explored performance of all summary methods on the simulated datasets, but we explored CA-ML only in Question 3, as it was prohibitively costly for the other questions. For results on biological datasets, we examine MRL, MP-EST, and CA-ML.

Results

Question 1: MLBS vs. BestML

Under the 1X model condition and for each summary method and AGE level, MLBS had lower missing branch rates for small numbers of genes, and BestML had lower error rates on larger numbers of genes (Fig. 1), but the actual transition point depended on the particular summary method and the AGE level. The difference in missing branch rate between these two approaches was generally small for small numbers of genes, but was sometimes substantial for larger numbers of genes or when performed with less accurately estimated gene trees. For example, MP-EST(BestML) and MP-EST(MLBS) both produced very accurate species trees given 800 genes with 12% AGE, but there were much larger differences on 800 genes with 27% AGE, where MLBS had 6% error and BestML had 2.5% error.

Gene resampling While our main interest was evaluating the site-only resampling strategy, we also evaluated the effect of using gene and site resampling instead of site-only resampling on one model condition where BestML was more accurate than MLBS (400 genes of 27% AGE and default ILS level with 10 replicates). Due to computational constraints, we limited the number of bootstrap replicates (for both ways of running MLBS) to 100 for this experiment. We found that site-only and gene/site resampling strategies produced very similar results (Fig. S3), with no statistically significant differences for any of the four methods (p-value > 0.7 according to a Wilcoxon rank sum test (Bauer 1972)). In the case of MP-EST, gene/site and site-only MLBS returned the same consensus tree for all 10 replicates. In addition, bootstrap support values obtained using gene/site resampling were generally similar to those obtained using site-only resampling (see Fig. S4). The remaining discussion focuses on BestML gene trees (see SOM for MLBS).

Question 2: Accuracy of branch support values

Pearson's correlation coefficients between support values and frequency of correctness were 0.99 for MRP, 0.99 for MRL, 0.96 for Greedy, and 0.89 for MP-EST. Greedy had better correlation at the higher support levels, but poorer correlation at the lower support levels (Fig. 2). Similarly, MP-EST correlated reasonably well at the highest support levels, and not very well at the lower support levels (Fig. 2). Interestingly, all methods tended to slightly 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 particularly 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.

We also performed gene/site resampling for one model condition; even with this limited sample, it was clear that the problem with under-estimating support values continued. For example, in total across the 10 replicates of the gene/site resampling experiment, there were 13 branches in MP-EST trees that had support below 30%, and all but one appeared in the true species tree.

Question 3: Relative topological accuracy of species tree estimations

We first fixed the amount of ILS to the 1X level, and varied the number of genes and the AGE level (Fig. 3). Next, we fixed the number of genes to 200, and varied ILS level with both true and estimated gene trees (Fig. 4). All p-values reported in this section are results of two-way ANOVA tests comparing pairs of methods overall and also testing the impact of the number of genes, AGE levels, or ILS levels on the relative performance of the two methods (with FDR correction for multiple tests; n= 23); see Table S3 for p-values and details of the ANOVA tests.

The general patterns were consistent with our expectations: for all methods, the species tree estimation accuracy was improved by increasing the number of genes (Fig. 3), but was reduced by increasing the AGE level (see Fig. S5 for correlation of average gene tree error and the species tree error) or increasing ILS level (Fig. 4).

We next compared pairs of methods, focusing on BestML analyses (see Fig. S6 for results using MLBS). As expected, the relative performance of methods was affected by the model condition.

MRL vs. MRP.— MRL had better accuracy than MRP in nearly every case, except for the 42% AGE model conditions where MRP was occasionally better than MRL (Fig. S7). We omit MRP from the rest of this discussion.

MRL vs. MP-EST.— Overall, MRL was more accurate than MP-EST, and the difference was statistically significant ($p < 0.0001$). However, there were conditions where both had the same or nearly the same accuracy (Fig. S8). For example, under the default (1X) ILS level with 400 or 800 true or highly accurate gene trees, both methods had excellent accuracy. MP-EST and MRL also had excellent accuracy on smaller numbers of true gene trees with reduced ILS (Figs. 4 and S8c). The impact of the number of genes or the AGE level on the relative performance of MRL and MP-EST was not overall statistically significant ($p = 0.5$ and $p = 0.2$, respectively). However, with 200 genes or more, as the AGE level increased, the difference between MRL and MP-EST significantally increased ($p = 0.03$; see Table 3 and Fig. S8b). The amount of ILS had no statistically significant impact ($p = 0.5$) on the relative performance between these two methods (Fig. 4).

MRL vs. Greedy.— MRL was overall significantly more accurate than greedy ($p = 0.008$), and the relative performance of methods was not impacted by the AGE level or the number of genes ($p = 1.0$ and $p = 0.2$ respectively). Increasing ILS significantly widened the gap ($p = 0.0001$) between MRL and Greedy (Fig. 4).

MRL vs. CA-ML.— MRL was overall significantly more accurate than CA-ML ($p < 0.0001$). The relative performance depended on the AGE level ($p = 0.001$) and amount of ILS ($p = 0.03$), but not on the number of genes ($p = 0.9$). Given relatively accurate gene trees, MRL was always more accurate than CA-ML (Figs. 3 and S9). However, when the AGE level was 27% or more (i.e., on 500bp or shorter alignments), there were some cases where CA-ML was more accurate than MRL (Fig. 3). Similarly, when the amount of ILS was reduced to its lowest level (5X branch length), CA-ML and MRL both had very low error, but CA-ML was slightly more accurate (Fig. 4). As the ILS level increased, the error for CA-ML increased faster than for MRL (Fig. 4).

MP-EST vs. Greedy.— The relative performance between Greedy and MP-EST depended on the level of ILS ($p < 0.0001$): Greedy was at least as accurate as MP-EST with lower ILS, while MP-EST was more accurate with increased ILS (Fig. 4). The number of genes may impact the relative performance ($p = 0.055$): Greedy was more accurate with small numbers of genes, but with large numbers of genes, MP-EST was at least as accurate as Greedy (Fig. 3). AGE level did not impact the relative performance ($p = 0.2$).

MP-EST vs. CA-ML.— CA-ML was more accurate than MP-EST under low levels of ILS, and was less accurate under high levels of ILS (Fig. 4), and the impact of ILS on their relative accuracy was statistically significant ($p = 0.001$). Interestingly, under the default 1X ILS condition, the AGE level also impacted the relative performance of MP-EST and CA-ML ($p < 0.0001$), so that MP-EST was typically more accurate whenever the AGE level was not too high (Figs. 3 and S9).

Branch length breakdown.— Generally, recovering long and very long branches was relatively easy for all methods (Fig. 3), even from relatively few genes (e.g., 50). However, with 42% AGE level, MP-EST failed to recover 8-14% of the long branches (even with 800 genes), while the other summary methods missed 5-8% of the long branches, and CA-ML missed only 0-5% of the long branches (Fig. 3). Similarly, with lowered ILS (2X) and 27% AGE, MP-EST failed to correctly recover 10% of the long branches, whereas CA-ML and MRL missed 0% and 5% of branches, respectively (Fig. 4b).

Given less than 200 genes, none of the methods were able to recover most of the short branches, and CA-ML and Greedy had the highest error rates for these branches.

However, with a large enough number of genes, all methods (other than Greedy) did very well at recovering the short branches (Fig. 3). Recovery of the very shortest branches was the most difficult for all methods, with CA-ML having the highest missing branch rates, followed by MP-EST and Greedy, and finally by MRL (Figs. 3 and 4b).

Question 4: High ILS and large numbers of genes

For the highest level of ILS (0.2X) and 3200 genes, MP-EST had the best accuracy of all methods, with the largest improvement on true gene trees. Not surprisingly, the number of genes impacted accuracy, and MRL, MRP, and MP-EST consistently improved as the number of genes increased (Fig. 5). Interestingly, Greedy stopped improving at 400 genes, even on true gene trees. Overall, Greedy had the highest error of all four methods on these data, and MRP had the second highest error. The differences between MRL and MP-EST were generally small (and the standard error bars mostly overlap), but MRL was more accurate for the smaller numbers of genes and MP-EST was more accurate for larger numbers of genes.

Biological Dataset Results

We analyzed two biological datasets: the mammalian dataset analyzed by Song et al. (2012) and the Amniota dataset studied by Chiari et al. (2012). Table S2 summarizes the main discussion points presented below.

Mammalian dataset.— Song et al. (2012) observed two interesting differences between CA-ML and MP-EST on their dataset of 37 mammalian species and 447 genes: (1) concatenation put tree shrews *(Tupaia belangeri)* as sister to Glires (Rodentia/Lagomorpha), but MP-EST put them with primates with high support, and (2) concatenation put bats (Chiroptera) as sister to Cetartiodactyla, while MP-EST put a

Carnivora/Perissodactyla clade as sister to Cetartiodactyla, and bats as sister to the clade containing Cetartiodactyla, Carnivora, and Perissodactyla. Both of these relationships are of great interest and the alternative relationships have been observed in other studies (see Hu et al. (2012) for a comprehensive review of the placement of bats, and see Janecka et al. (2007); Boussau et al. (2013); Kumar et al. (2013) for the placement of tree shrews).

We found 21 genes in this dataset that had mislabeled species (since confirmed by the authors), and also identified two outlier genes that had unusually high levels of highly supported gene tree incongruence with the remaining genes (see Fig. S10). We removed these 23 genes, and reanalyzed the remaining 424 genes using MP-EST (using both gene/site and site-only resampling), MRL, and CA-ML (Fig. 6). MRL(BestML), MRL(MLBS), and MP-EST(BestML) were all topologically identical, and put tree shrews as sister to Glires. Support in the MRL trees for this placement was 84% using site-only resampling and 77% using gene/site resampling, both of which were reasonably high. MP-EST(MLBS), on the other hand, placed tree shrews as sister to primates with 62% support using both site-only resampling and gene/site resampling (note that the same relationship had 99% support in Song et al. (2012); the exact cause of this difference is not clear to us).

All MRL and MP-EST trees differed from the CA-ML tree with respect to the position of bats (*Myotis lucifugus* and *Pteropus vampyrus*); CA-ML put bats with Cetartiodactyla, and the MP-EST and MRL trees had high support ($>$ 82%) for placing bats as sister to a (Cetartiodactyla,(Perissodactyla,Carnivora)) clade. Thus, this discordance between summary methods and the CA-ML method was robust to the choice of the summary method (unlike the position of tree shrews). Overall, our analyses strongly support the placement of bats as sister to a clade containing Cetartiodactyla, Perissodactyla, and Carnivora.

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 (Zardoya and Meyer 1998; Iwabe et al. 2005; Hugall et al. 2007) 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 site-only MLBS procedure on two sets of gene trees – one based on AA alignments, and the other based on DNA alignments. Their MP-EST analyses resolved bird/turtle/crocodile differently, depending on whether AA and DNA gene trees were used. The AA MP-EST tree, just like concatenation (using Bayesian analyses) on either AA or DNA, put turtles as sister to archosaurs with 99% support; however, the DNA MP-EST tree put turtles 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 (Fig. 7). Our MP-EST trees were all topologically identical to those obtained by Chiari et al. (2012) and had very similar bootstrap support values (MLBS and BestML gave identical results). However, 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 the MRL analyses did not change with the type of gene trees (AA or DNA).

Discussion

This study shows various trends, which we summarize. The first observation is that the point estimates produced by summary methods were impacted by the choice of input gene tree distribution (e.g., either BestML or MLBS), and that using BestML produced more accurate species tree topologies when the number of genes was large enough, but the MLBS approach was more accurate when the number of genes was small (Fig. 1). The fact that BestML topologies were more accurate than MLBS topologies also means that some correct branches in the BestML trees have low support (at most 50%). Reassuringly, for all summary methods, the highly supported branches in species trees computed using BestML gene trees tended to be correct with probabilities proportional to their support (Fig. 2). We did not perform concatenation with bootstrapping due to computational requirements, and therefore cannot comment on the reliability of support for concatenation; however, others have shown that concatenation can result in high support for wrong relationships in the presence of enough ILS (Kubatko and Degnan 2007).

There were many cases where differences between summary methods were quite small (e.g., on true gene trees with low ILS), representing a difference of one or two edges in a species tree; whether these differences would be of scientific importance would depend on the particular biological question. However, the differences between summary methods increased with the amount of ILS and with average gene tree estimation error (AGE) (which seems to affect MP-EST more than other summary methods); see correlations in Fig. S5 and ANOVA analyses in Table S3.

One of the unexpected trends is that MRL produced more accurate species trees than other summary methods under nearly all model conditions; the only observed exceptions in this study had very large numbers of gene trees and the highest ILS level (0.2X), where MP-EST was more accurate.

CA-ML implicitly assumes that there is no ILS and that all true gene trees have identical topologies, and so it is not surprising that CA-ML can have better accuracy than summary methods when there is low ILS, especially given limited amounts of data (Patel et al. 2013). Our results are consistent with these observations: under very low ILS conditions, CA-ML was often more accurate than summary methods, while under moderate to high ILS conditions (except in the presence of high AGE levels), CA-ML was generally not as accurate as MP-EST or MRL. Finally, we note that CA-ML analyses we performed were unpartitioned, and it is possible that using a partitioning scheme could improve the accuracy of concatenation (Brandley et al. 2005) (our simulated gene trees used the same 4x4 substitution matrix and alpha shape parameter for the gamma distribution of rates-across-sites but had different branch lengths and topology).

The results on the two biological datasets are consistent with the simulation study. On the mammalian dataset (with a relatively large number of genes), all analyses, except for MP-EST(MLBS), placed tree shrews as sister to Glires. Our simulations showed that using BestML instead of MLBS resulted in improved point estimates of the species trees when the number of genes was sufficiently large, suggesting that the topology obtained using MP-EST(BestML) may be more accurate than the topology obtained using MP-EST(MLBS). Neither tree had high support for the placement of tree shrews, but our results suggest that the multi-locus bootstrapping procedure may under-estimate support for some correct branches in the BestML tree (Fig. 2). Based on these observations, our analyses of the Song et al. (2012) data provide some support for the placement of tree shrew as sister to Glires, but should not be considered definitive.

We also observed that MRL gave reasonable results on the Amniota datasets, possibly more so than MP-EST, in that MRL gave the same results for both nucleotide or amino-acid data while MP-EST did not. Chiari et al. (2012) had moderate numbers of genes with moderate support (on average 50% for AA and 65% for DNA), a condition that favors MRL over MP-EST, according to our study. Furthermore, the amino-acid and nucleotide MRL analyses, the amino-acid MP-EST analyses, and CA-ML, are all in agreement with most of the previous literature (Iwabe et al. 2005; Hugall et al. 2007) in putting turtles as sister to archosaurs. Based on prior literature and our analyses, we consider the (turtles,(birds,crocodiles)) hypothesis to be stronger than the alternatives.

Taken together, the results on biological and simulated data demonstrate that the best summary methods can produce highly accurate estimates of species trees given a large enough number of sufficiently accurate gene trees. However, these results also demonstrate that the summary methods we tested are vulnerable to gene tree estimation error, as has been observed in prior studies (Leaché and Rannala 2011; DeGiorgio and Degnan 2013; Patel et al. 2013; Bayzid and Warnow 2013). In addition, our study suggests that 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.

We offer the following possible explanation for why MP-EST did not produce highly accurate species tree topologies under conditions with high gene tree estimation error. MP-EST estimates species trees by finding a model species tree (rooted tree topology and branch lengths in coalescent units) that is likely to produce the observed (estimated) distribution on rooted three-taxon gene trees. Hence, given substantial gene tree discordance, MP-EST will produce species trees with short branches. However, gene tree estimation error increases the discordance between gene trees, which results in MP-EST producing species trees with shorter branch lengths (in coalescent units) than the true species tree branch lengths (as we saw in our study; see Fig. S11). High levels of gene tree estimation error generally impact the estimation of underlying model parameters used by MP-EST, and can result in increased estimation error for the species tree topology.

Consequently, although branch lengths in species trees estimated using coalescent-based methods *can* be interpreted as indications of ILS levels, these interpretations should be performed with great care. Specifically, if the gene trees are likely to have high estimation error (suggested possibly by low bootstrap support in estimated gene trees), then the branch lengths in the estimated species tree may be under-estimated (relative to the true species tree branch lengths), so that the implied amount of ILS may be over-estimated relative to the true amount. The observations here regarding MP-EST suggest that other statistically consistent summary methods that use likelihood calculations under the coalescent model, and that operate by combining estimated gene trees, may also have the same vulnerabilities.

Taking all these observations into consideration, we make the following recommendations. First, since gene tree estimation accuracy can have a large impact on species tree estimation accuracy, 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 (e.g., those with highest support or highest levels of phylogenetic signal, as suggested by Salichos and Rokas (2013)) might improve summary methods. 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), and it is not always clear how gene tree reliability can be measured. Moreover, filtering reduces the number of genes, and the number of genes strongly impacted accuracy in our studies. Understanding the effects of screening and developing effective filtering methods therefore need further study.

Second, because the choice between MLBS and BestML gene trees impacts the final tree, 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. Relationships that have high support in MLBS but are absent from BestML analyses should be treated with caution.

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. When analyses do not agree, consideration for the causes for the disagreement may indicate which analyses is likely to be more reliable, or suggest the need for additional data (e.g., more genes or taxa) or better data (e.g., more accurate alignments and more accurate gene trees).

In this study, we used sequence length to vary gene tree estimation error; however, the sequence length needed to provide a given level of gene tree estimation error is impacted by substitution rates (Kuhner and Felsenstein 1994), branch lengths, missing data (Lemmon et al. 2009), and number of taxa. For example in rapid radiations, gene tree branch lengths tend to be very short, which increases the difficulty in producing highly accurate gene trees - even given long sequences. Also, recombination-free sequences can be quite short, and increasing the length of each marker increases the risk of running into recombination events, and so has the potential to reduce accuracy (Edwards 2009) (but also see Lanier and Knowles (2012) for a contrasting point of view).

Although this study explored performance on large numbers of genes, we did not explore performance on large numbers of taxa, nor under conditions with missing data (i.e., gene trees with some missing taxa), nor on datasets where the gene trees cannot be rooted. Thus, the results of this study should be interpreted with some care, and it is possible that relative performance of methods could change under these more challenging conditions. For example, MP-EST (since it requires rooted gene trees) can be impacted by missing data, at least when the missing data make it difficult to root the estimated gene trees correctly (Springer and Gatesy 2014). In addition, the limitation in this study to datasets with at most 37 species means that the relative performance of MP-EST, MRL, Greedy, and CA-ML on datasets with substantially larger numbers of species cannot be predicted. Indeed, there is a possibility that MP-EST, since it uses a heuristic to search for maximum pseudo-likelihood species trees, may be impacted in terms of computation time as well as its ability to find good solutions to its optimality criterion. This issue will need to be explored using additional simulated and biological datasets with larger numbers of species. Finally, we did not explore conditions where gene tree discordance was partially due to other biological causes, such as duplication and loss, gene flow (Leaché et al. 2014), horizontal gene transfer, or hybridization (Nakhleh 2013), and the relative performance of coalescent-based methods and concatenation might change under those more complicated conditions.

SUPPLEMENTARY MATERIAL: Supplementary material, including data files and online-only appendices, can be found in the Dryad data repository at `http://dx.doi.org/10.5061/dryad.310q3`.

Funding

This work was supported by National Science Foundation (grant number DBI-1062335) to TW, a Howard Hughes International Predoctoral Fellowship to SM, and by the University of Alberta through a subaward to the University of Texas to TW.

Acknowledgments

The authors thank Bastien Boussau who allowed us to use his simulated datasets in this paper. The authors also thank the editors and anonymous reviewers for their valuable feedback.

*

References

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

Bauer, D. F. 1972. Constructing confidence sets using rank statistics. Journal of the American Statistical Association 67:687–690.

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.

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

Brandley, M. C., A. Schmitz, and T. W. Reeder. 2005. Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. Systematic Biology 54:373–390.

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.

DeGiorgio, M. and J. H. Degnan. 2013. Robustness to divergence time underestimation when inferring species trees from estimated gene trees. Systematic Biology 63:66–82.

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 and Evolution 24:332–340.

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 104:5936–5941.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–791.

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. Zhang, and L. Yu. 2012. Summary of laurasiatheria (mammalia) phylogeny. Zoological Research 33:65–74.

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 318:792–794.

Kimball, R. T., N. Wang, V. Heimer-McGinn, C. Ferguson, and E. L. Braun. 2013. Identifying localized biases in large datasets: A case study using the avian tree of life. Molecular Phylogenetics and Evolution 69:1021–1032.

Knowles, L. L., H. C. Lanier, P. B. Klimov, and Q. He. 2012. Full modeling versus summarizing gene-tree uncertainty: method choice and species-tree accuracy. Molecular Phylogenetics and Evolution 65:501–509.

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.

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

Kuhner, M. K. and J. Felsenstein. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Molecular Biology and Evolution 11:459–468.

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

Lanier, H. C. and L. L. Knowles. 2012. Is recombination a problem for species-tree analyses? Systematic Biology 61:691–701.

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.

Leaché, A. D., R. B. Harris, B. Rannala, and Z. Yang. 2014. The influence of gene flow on species tree estimation: a simulation study. Systematic Biology 63:17–30.

Leaché, A. D. and B. Rannala. 2011. The accuracy of species tree estimation under simulation: a comparison of methods. Systematic Biology 60:126–37.

Lemmon, A. R., J. M. Brown, K. Stanger-Hall, and E. M. Lemmon. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Systematic Biology 58:130–45.

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.

McCormack, J. E., M. G. Harvey, B. C. Faircloth, N. G. Crawford, T. C. Glenn, and R. T. Brumfield. 2013. A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. PLoS One 8:e54848.

Mirarab, S., R. Reaz, M. Bayzid, T. Zimmermann, M. Swenson, and T. Warnow. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics (in press).

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.

Patel, S., R. T. Kimball, and E. L. Braun. 2013. Error in phylogenetic estimation for bushes in the tree of life. Journal of Phylogenetics and Evolutionary Biology 1:110.

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.

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

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

Rosenberg, N. A. 2013. Discordance of species trees with their most likely gene trees: A unifying principle. Molecular Biology and Evolution 30:2709–2713.

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

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

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

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 109:14942–7.

Springer, M. S. M. and J. Gatesy. 2014. Land plant origins and coalescence confusion. Trends in Plant Science 19:267–9.

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

Swofford, D. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Massachussets.

Than, C. and L. Nakhleh. 2009. Species tree inference by minimizing deep coalescences. PLoS Computational Biology 5:e1000501.

Than, C. and N. A. Rosenberg. 2011. Consistency Properties of Species Tree Inference by Minimizing Deep Coalescences. Journal of Computational Biology 18:1–15.

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.

Whitfield, J. and P. Lockhart. 2007. Deciphering ancient rapid radiations. Trends in Ecology and Evolution 22:258–265.

Wolf, Y., I. Rogozin, N. Grishin, and E. Koonin. 2002. Genome trees and the tree of life. Trends in Genetics 18:472–479.

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.

Table 1: Levels of simulated ILS, breakdown of branches by length, and true gene tree discordance.

ILS[a]	very short [b]	short	medium	long	very long	avg. RF[c]	min. RF[d]
0.2X	9	8	16	1	0	79%	50% (2)
0.5X	5	4	8	16	1	54%	29% (12)
1X	3	2	6	10	13	32%	3% (1)
2X	0	4	2	7	21	18%	0% (1)
5X	0	0	4	2	28	9%	0% (115)

[a] 1X is the default ILS condition, corresponding to the species tree estimated on the mammalian dataset using MP-EST; rescaled lengths result in reduced ILS (2X and 5X) and increased ILS (0.5X and 0.2X).

[b] The number of branches in the model species tree that fall into various categories of length (in coalescent units): very short ($bl < 0.1$), short ($0.1 \leq bl < 0.25$), medium ($0.25 \leq bl < 0.625$), long ($0.625 \leq bl < 1.5625$), and very long ($1.5625 \leq bl$).

[c] The average RF distance between the true model species tree and true gene trees (over 4000 gene trees).

[d] The minimum RF distance; the number of genes with the minimum distance is shown parenthetically.

Table 2: Simulated model conditions, gene tree estimation error (AGE), and bootstrap support (BS).

ILS	Sequence length	Number of genes	AGE[a]	avg. BS[b]
0.2X	500bp	100,200,400,800,1600,3200	26%	64%
0.5X	500bp	200	26%	64%
1X	250bp	25,50,100,200,400,800	42%	46%
1X	500bp	25,50,100,200,400,800	27%	63%
1X	1000bp	25,50,100,200,400,800	16%	79%
1X	1500bp	25,50,100,200,400,800	12%	84%
2X	500bp	200	27%	64%
5X	500bp	200	26%	63%

[a] Measured as the average RF distance between true gene trees and estimated gene trees.

[b] Average bootstrap support over all branches of all the gene trees. Bootstrap support values are measurable on biological datasets and are correlated with gene tree estimation error. The mammalian dataset had average BS of 71%, putting it in between the 1X 500bp and 1000bp model conditions.

Figure 1: **MLBS vs. BestML gene tree strategies on the simulated 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. Rows correspond to various levels of gene tree estimation error (see Table 2).

Figure 2: **Relationship between bootstrap support calculations (obtained using multi-locus bootstrapping drawn on BestML trees) and frequency of correct branches.** Each box shows the aggregated results on all model conditions of the mammalian simulated datasets. Support values are binned with the breaks $0, 10, 20, 30, 40, 50, 55, 65, 75, 80, 85, 90, 92, 94, 96, 98, 100,$ and with each bin including the right value and excluding the left value (e.g., the first bin is [0,10), and the last bin only includes branches with 100% support). For each bin, the figures show the percentage of branches in the estimated species trees that were correctly estimated. The diagonal $y = x$ lines show the ideal scenario.

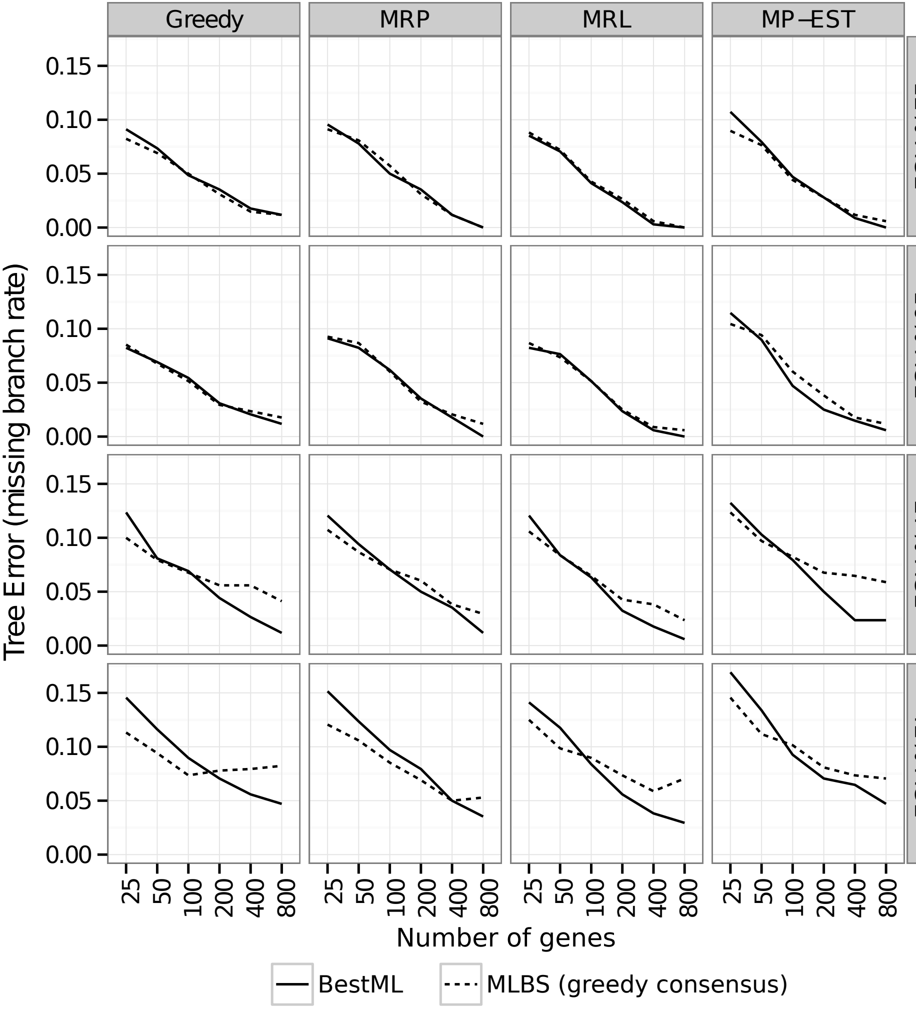
Figure 3: **Comparing methods on mammalian simulated datasets.** We compare three summary methods (MRL, Greedy, and MP-EST) on BestML gene trees and also CA-ML. The impact of changing the number of genes and average gene tree estimation error (AGE) is shown. The amount of ILS is fixed at the 1X level. Tree estimation error is computed using the missing branch rate with respect to all branches of the model species (top row), and also with respect to various categories of branches (remaining rows). Average error over multiple replicate runs is shown and error bars show standard error (see 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.

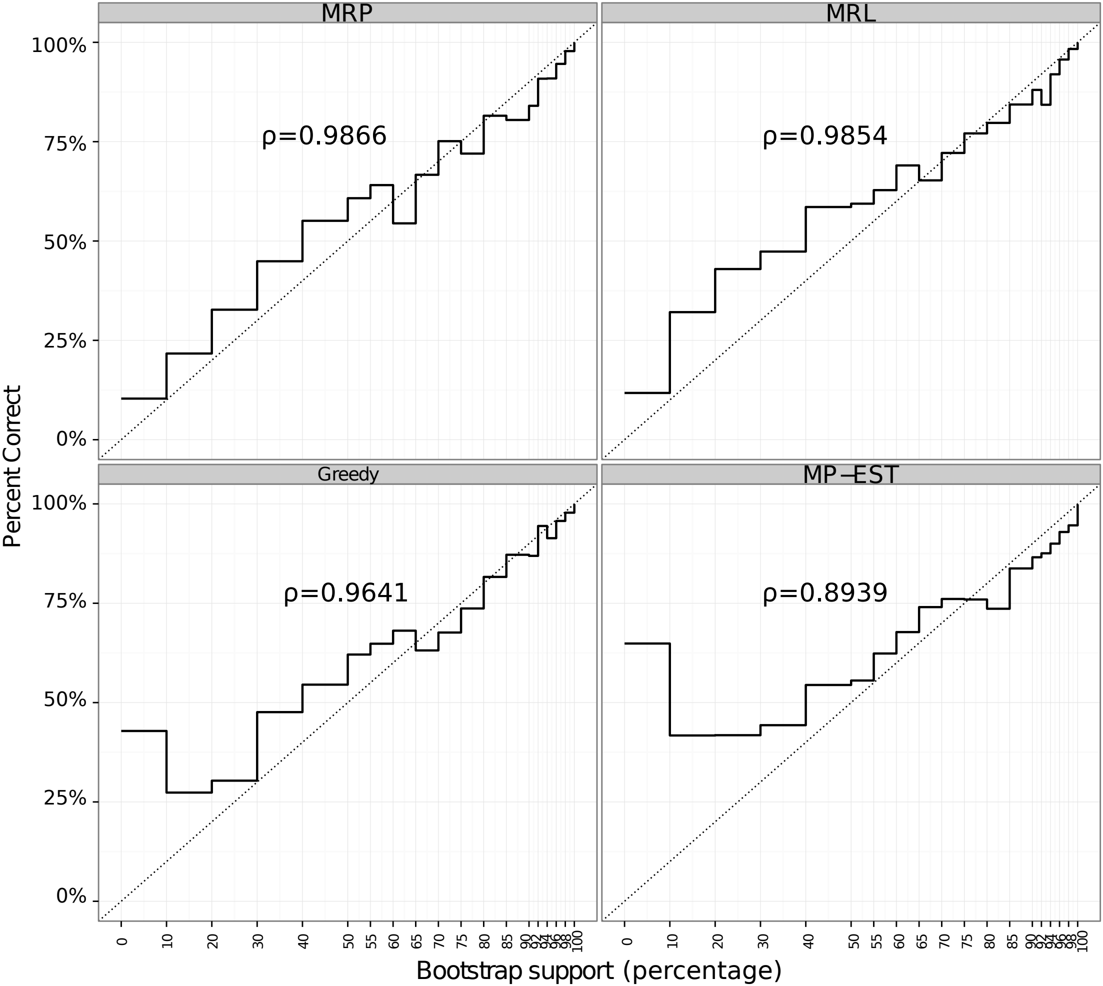
Figure 4: **Impact of ILS level on species tree estimation.** We compare the performance of MRL with Greedy, MP-EST, and CA-ML on the simulated mammalian dataset using true and estimated (BestML) gene trees with 26-27% estimation error Missing branch rate is computed with respect to (a) all branches of the model species tree or (b) various categories of branch length. We show the average tree error (bars indicate standard error) over 20 replicates of 200 genes (see Table S1 for standard deviation).

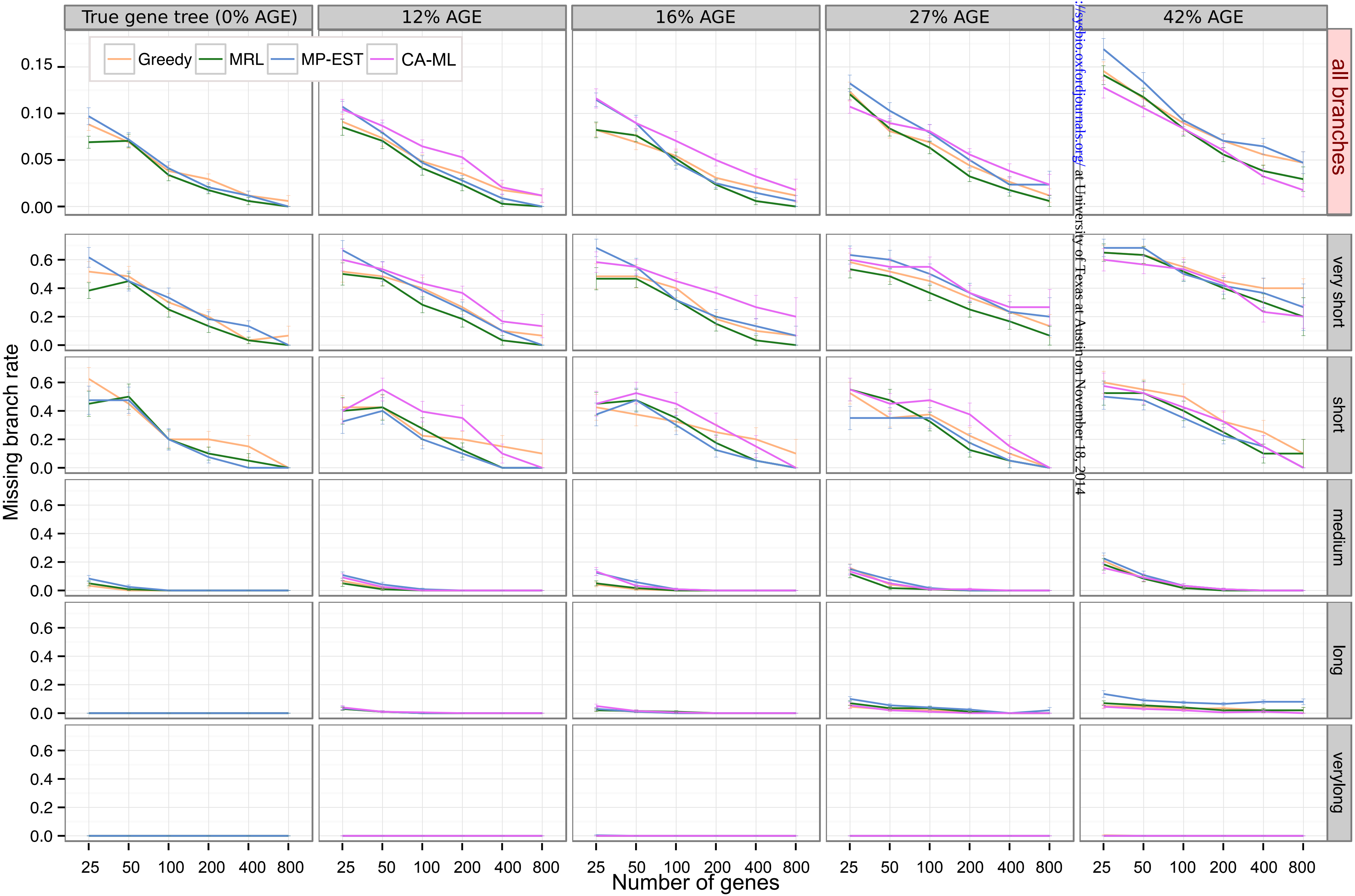
Figure 5: **Comparing performance of various methods on up to 3200 gene trees, simulated under 0.2X ILS level (i.e. the highest levels of ILS)**. Results for 100 to 800 genes are over 20 replicates, 1600 genes over 10 replicates, and 3200 genes over 5 replicates. Lines show average error and error bars show standard error (see Table S1 for standard deviation).

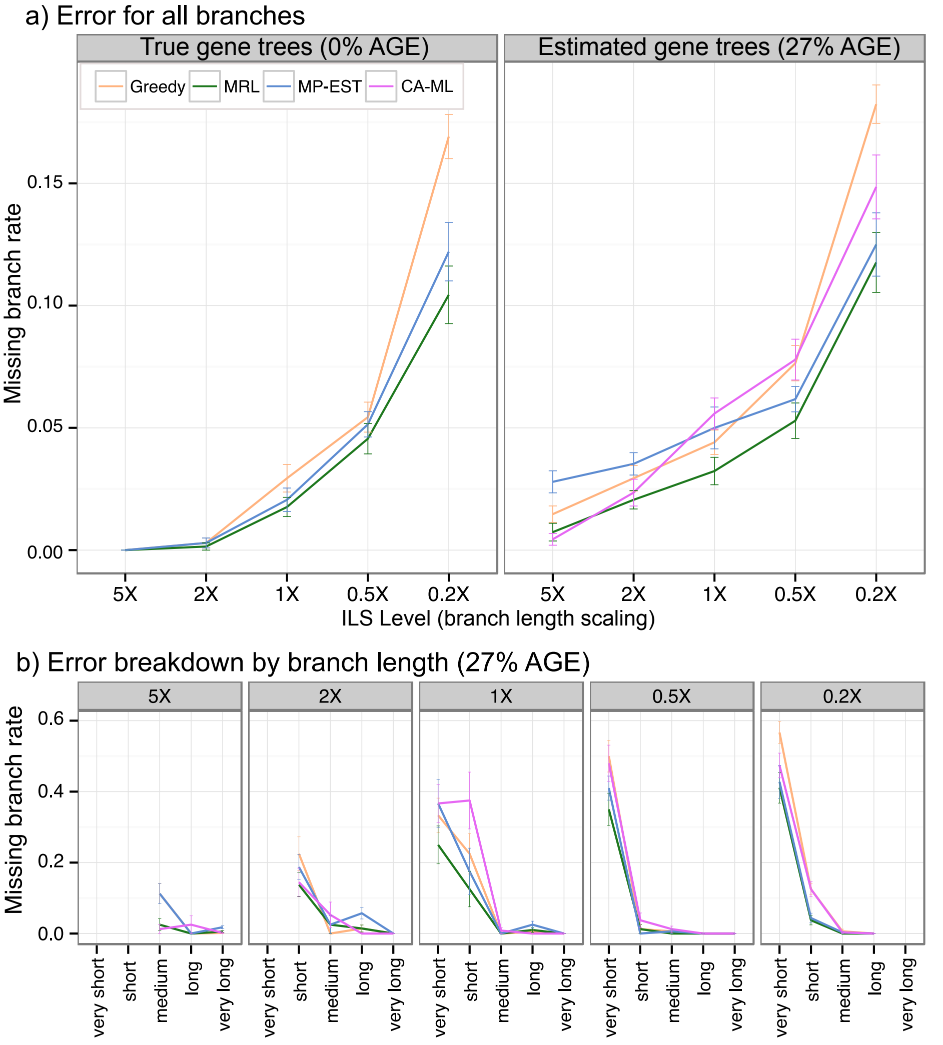
Figure 6: **Mammals biological dataset analyses.** We analyzed the dataset in Song et al. (2012) using MRL, MP-EST, and CA-ML. Edges with no label have 100% support. Dashed branches are those that differ across various analyses. For summary methods, two support values are shown: the first value is based on site-only resampling bootstrapping and the second value is based on both gene and site resampling.

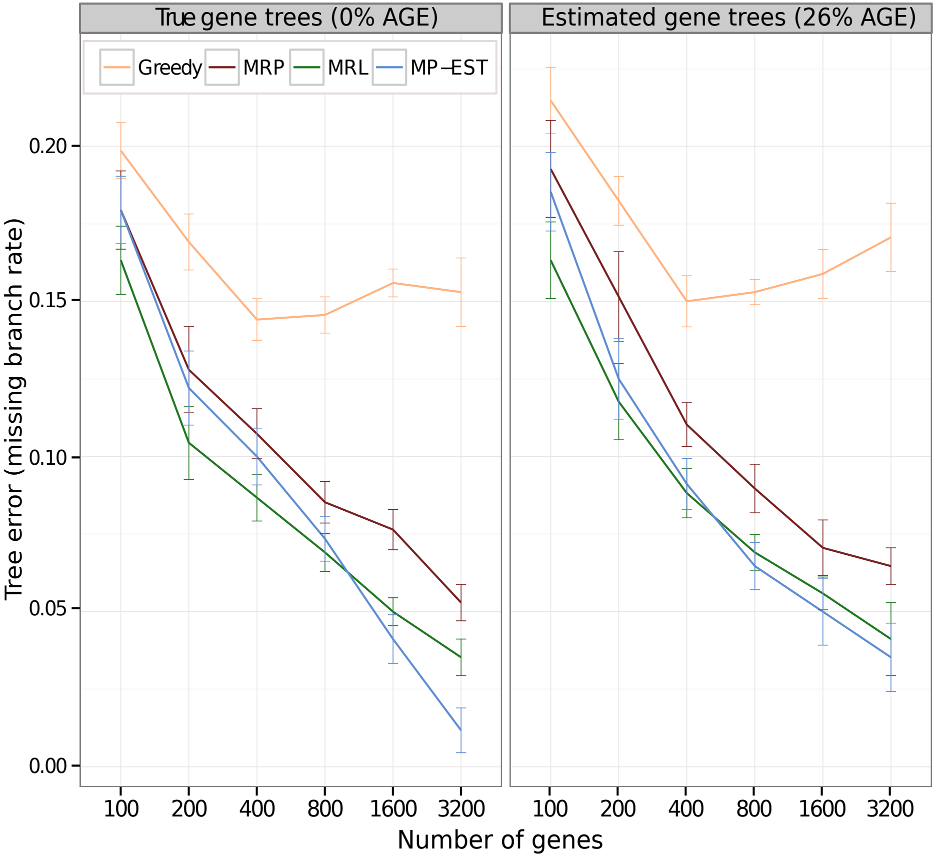
Figure 7: **Amniota biological dataset analyses.** Trees estimated from DNA and AA gene trees are shown for two summary methods: MP-EST (a and b) and MRL (c and d); in addition, Bayesian concatenation trees estimated by Chiari et al. (2012) are also reproduced here for comparison (e and f). Branches resolving the birds/crocodiles/turtles relationship are shown as dashed.

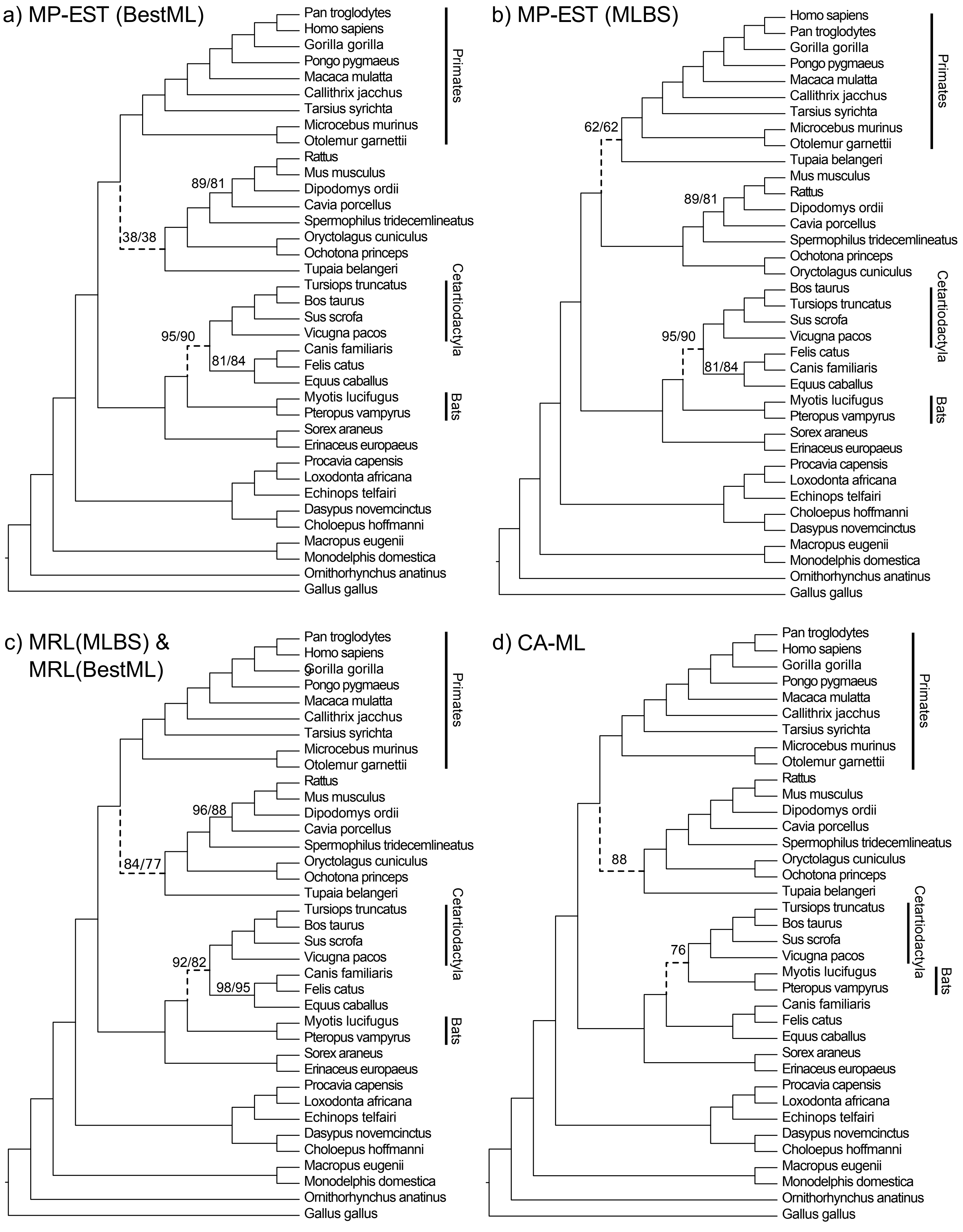

Greedy
MRP
MRL
MP−EST
12% AGE
16% AGE
27% AGE
42% AGE
0.15
0.10
0.05
0.00
Tree Error (missing branch rate)
25
50
100
200
400
800
Number of genes
BestML
MLBS (greedy consensus)

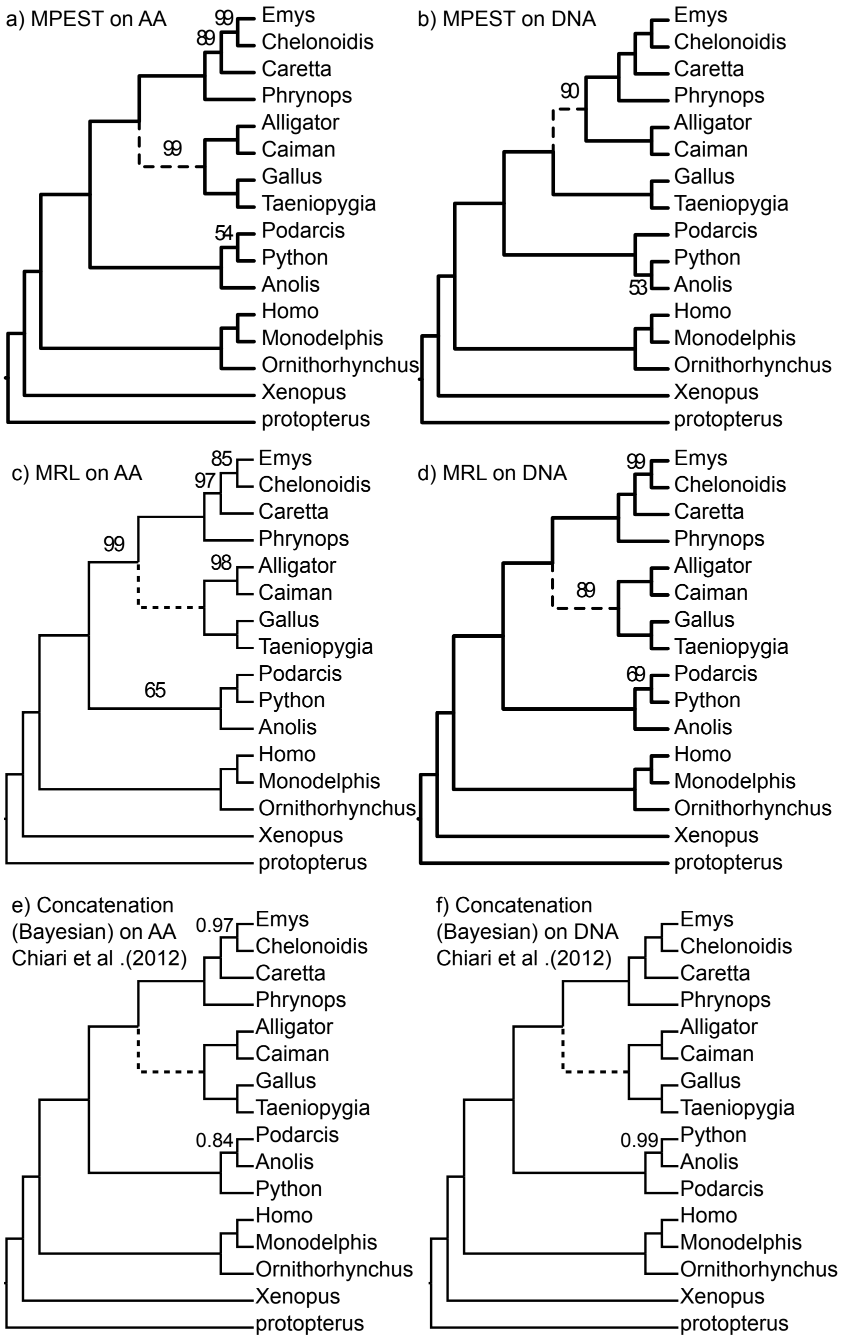

MRP
MRL
Greedy
MP−EST
ρ=0.9866
ρ=0.9854
ρ=0.9641
ρ=0.8939
Percent Correct
100%
75%
50%
25%
0%
0
10
20
30
40
50
55
60
65
70
75
80
85
90
92
94
96
98
100
Bootstrap support (percentage)


True gene tree (0% AGE)
12% AGE
16% AGE
27% AGE
42% AGE
Greedy
MRL
MP-EST
CA-ML
all branches
very short
short
medium
long
verylong
Missing branch rate
0.00
0.05
0.10
0.15
0.0
0.2
0.4
0.6
Number of genes
25
50
100
200
400
800


a) Error for all branches
True gene trees (0% AGE)
Estimated gene trees (27% AGE)
Greedy
MRL
MP-EST
CA-ML
Missing branch rate
0.00
0.05
0.10
0.15
5X
2X
1X
0.5X
0.2X
ILS Level (branch length scaling)
b) Error breakdown by branch length (27% AGE)
5X
2X
1X
0.5X
0.2X
Missing branch rate
0.0
0.2
0.4
0.6
very short
short
medium
long
very long


True gene trees (0% AGE)
Estimated gene trees (26% AGE)
Greedy
MRP
MRL
MP−EST
Tree error (missing branch rate)
0.00
0.05
0.10
0.15
0.20
100
200
400
800
1600
3200
Number of genes


a) MP-EST (BestML)
Pan troglodytes
Homo sapiens
Gorilla gorilla
Pongo pygmaeus
Macaca mulatta
Callithrix jacchus
Tarsius syrichta
Microcebus murinus
Otolemur garnettii
Primates
Rattus
Mus musculus
Dipodomys ordii
Cavia porcellus
Spermophilus tridecemlineatus
Oryctolagus cuniculus
Ochotona princeps
Tupaia belangeri
89/81
38/38
Tursiops truncatus
Bos taurus
Sus scrofa
Vicugna pacos
Canis familiaris
Felis catus
Equus caballus
Cetartiodactyla
95/90
81/84
Myotis lucifugus
Pteropus vampyrus
Bats
Sorex araneus
Erinaceus europaeus
Procavia capensis
Loxodonta africana
Echinops telfairi
Dasypus novemcinctus
Choloepus hoffmanni
Macropus eugenii
Monodelphis domestica
Ornithorhynchus anatinus
Gallus gallus
b) MP-EST (MLBS)
Homo sapiens
Pan troglodytes
Gorilla gorilla
Pongo pygmaeus
Macaca mulatta
Callithrix jacchus
Tarsius syrichta
Microcebus murinus
Otolemur garnettii
Primates
62/62
Tupaia belangeri
Mus musculus
Rattus
Dipodomys ordii
Cavia porcellus
Spermophilus tridecemlineatus
Ochotona princeps
Oryctolagus cuniculus
89/81
Bos taurus
Tursiops truncatus
Sus scrofa
Vicugna pacos
Cetartiodactyla
95/90
Felis catus
Canis familiaris
Equus caballus
81/84
Myotis lucifugus
Pteropus vampyrus
Bats
Sorex araneus
Erinaceus europaeus
Procavia capensis
Loxodonta africana
Echinops telfairi
Choloepus hoffmanni
Dasypus novemcinctus
Macropus eugenii
Monodelphis domestica
Ornithorhynchus anatinus
Gallus gallus
c) MRL(MLBS) &
MRL(BestML)
Pan troglodytes
Homo sapiens
Gorilla gorilla
Pongo pygmaeus
Macaca mulatta
Callithrix jacchus
Tarsius syrichta
Microcebus murinus
Otolemur garnettii
Primates
Rattus
Mus musculus
Dipodomys ordii
Cavia porcellus
Spermophilus tridecemlineatus
Oryctolagus cuniculus
Ochotona princeps
Tupaia belangeri
96/88
84/77
Tursiops truncatus
Bos taurus
Sus scrofa
Vicugna pacos
Canis familiaris
Felis catus
Equus caballus
Cetartiodactyla
92/82
98/95
Myotis lucifugus
Pteropus vampyrus
Bats
Sorex araneus
Erinaceus europaeus
Procavia capensis
Loxodonta africana
Echinops telfairi
Dasypus novemcinctus
Choloepus hoffmanni
Macropus eugenii
Monodelphis domestica
Ornithorhynchus anatinus
Gallus gallus
d) CA-ML
Pan troglodytes
Homo sapiens
Gorilla gorilla
Pongo pygmaeus
Macaca mulatta
Callithrix jacchus
Tarsius syrichta
Microcebus murinus
Otolemur garnettii
Primates
Rattus
Mus musculus
Dipodomys ordii
Cavia porcellus
Spermophilus tridecemlineatus
Oryctolagus cuniculus
Ochotona princeps
Tupaia belangeri
88
Tursiops truncatus
Bos taurus
Sus scrofa
Vicugna pacos
Cetartiodactyla
76
Myotis lucifugus
Pteropus vampyrus
Bats
Canis familiaris
Felis catus
Equus caballus
Sorex araneus
Erinaceus europaeus
Procavia capensis
Loxodonta africana
Echinops telfairi
Dasypus novemcinctus
Choloepus hoffmanni
Macropus eugenii
Monodelphis domestica
Ornithorhynchus anatinus
Gallus gallus


a) MPEST on AA
99
89
99
54
Emys
Chelonoidis
Caretta
Phrynops
Alligator
Caiman
Gallus
Taeniopygia
Podarcis
Python
Anolis
Homo
Monodelphis
Ornithorhynchus
Xenopus
protopterus
b) MPEST on DNA
90
53
Emys
Chelonoidis
Caretta
Phrynops
Alligator
Caiman
Gallus
Taeniopygia
Podarcis
Python
Anolis
Homo
Monodelphis
Ornithorhynchus
Xenopus
protopterus
c) MRL on AA
85
97
99
98
65
Emys
Chelonoidis
Caretta
Phrynops
Alligator
Caiman
Gallus
Taeniopygia
Podarcis
Python
Anolis
Homo
Monodelphis
Ornithorhynchus
Xenopus
protopterus
d) MRL on DNA
99
89
69
Emys
Chelonoidis
Caretta
Phrynops
Alligator
Caiman
Gallus
Taeniopygia
Podarcis
Python
Anolis
Homo
Monodelphis
Ornithorhynchus
Xenopus
protopterus
e) Concatenation
(Bayesian) on AA
Chiari et al .(2012)
0.97
0.84
Emys
Chelonoidis
Caretta
Phrynops
Alligator
Caiman
Gallus
Taeniopygia
Podarcis
Anolis
Python
Homo
Monodelphis
Ornithorhynchus
Xenopus
protopterus
f) Concatenation
(Bayesian) on DNA
Chiari et al .(2012)
0.99
Emys
Chelonoidis
Caretta
Phrynops
Alligator
Caiman
Gallus
Taeniopygia
Python
Anolis
Podarcis
Homo
Monodelphis
Ornithorhynchus
Xenopus
protopterus