Bioinformatics Advance Access published July 9, 2013

Naive Binning Improves Phylogenomic Analyses

Md. Shamsuzzoha Bayzid and Tandy Warnow[1]*

[1]Department of Computer Science, The University of Texas at Austin, Austin, Texas 78712, USA,

Associate Editor: Prof. David Posada

ABSTRACT

Motivation: Species tree estimation in the presence of incomplete lineage sorting (ILS) is a major challenge for phylogenomic analysis. Although many methods have been developed for this problem, little is understood about the relative performance of these methods when estimated gene trees are poorly estimated, due to inadequate phylogenetic signal.

Results: We explored the performance of some methods for estimating species trees from multiple markers on simulated datasets in which gene trees differed from the species tree due to ILS. We included *BEAST, concatenated analysis, and several "summary methods": BUCKy, MP-EST, MDC, MRP, and the greedy consensus. We found that *BEAST and concatenation gave excellent results, often with substantially improved accuracy over the other methods. We observed that *BEAST's accuracy is largely due to its ability to co-estimate the gene trees and species tree. However, *BEAST is computationally intensive, making it challenging to run on datasets with 100 or more genes or with more than 20 taxa. We propose a new approach to species tree estimation in which the genes are partitioned into sets, and the species tree is estimated from the resultant "supergenes." We show that this technique improves the scalability of *BEAST without affecting its accuracy and improves the accuracy of the summary methods. Thus, naive binning can improve phylogenomic analysis in the presence of ILS.

INTRODUCTION

Species tree estimation from multiple genes is often performed using concatenation (also called "combined analysis"): alignments are estimated for each gene and concatenated into a super-matrix, which is then used to estimate the species tree. When gene trees are concordant, concatenation can give very accurate results; however, this approach to species tree estimation is potentially problematic when gene trees differ from the species tree (and hence from each other) due to several biological factors, including gene duplication and loss, horizontal gene transfer, and incomplete lineage sorting.

The best studied of these problems is species tree estimation in the presence of incomplete lineage sorting (ILS), which is based on the multi-species coalescent (Wakeley, 2009). Many methods have been developed to estimate species trees in the presence of ILS, beginning with the MDC (minimize deep coalescence) approach suggested in Maddison (1997), and now including many different types of methods (see Degnan and Rosenberg (2009); Knowles (2009) for a discussion of some methods). Some of these new methods (for example, MP-EST (Liu *et al.*, 2010) and the population tree from BUCKy (Ané *et al.*, 2007; Larget *et al.*, 2010)) have been proven to be statistically consistent in the presence of ILS. In contrast, the greedy consensus, majority consensus, the concordance tree from BUCKy, and MDC (Degnan *et al.*, 2009; Than and Rosenberg, 2011; Ané *et al.*, 2007) can be inconsistent in the presence of ILS (i.e., there are some parameter settings under which these methods are inconsistent). The Bayesian method *BEAST (Heled and Drummond, 2010) may produce a statistically consistent point estimate (e.g., the MAP tree) of the species tree, but a formal proof has not yet been provided (however, see Steel (2010), which proves the statistical consistency of gene tree estimation using Bayesian MCMC methods). Simulations suggest that when true gene trees differ due to ILS, concatenated analysis can produce incorrect estimates of the species tree, sometimes with high confidence (Larget *et al.*, 2010; Liu *et al.*, 2010; Edwards *et al.*, 2007; DeGiorgio and Degnan, 2010; Kubatko and Degnan, 2007; Leaché and Rannala, 2011; Salichos and Rokas, 2013), leading to the conjecture that concatenated analyses are *not* statistically consistent. However, statistical consistency or inconsistency is a mathematical statement about performance in the limit and so requires a formal proof. Thus, while the evidence strongly suggests that, under conditions in which gene trees can differ due to ILS, concatenation can be statistically inconsistent but *BEAST will be statistically consistent, these are still open questions.

As a result of these studies and the growing awareness that ILS can be present in many phylogenomic datasets, there is great interest in using ILS-based estimation of species trees instead of concatenated analysis (Edwards, 2009; Degnan and Rosenberg, 2009; Huang *et al.*, 2010; Knowles, 2009). However, only a few studies have been published comparing ILS-based methods and even fewer have compared concatenated analyses to ILS-based methods. Performance in simulation has been mixed, with ILS-based methods outperforming concatenation in some cases but not all (DeGiorgio and Degnan, 2010; Leaché and Rannala, 2011; Liu *et al.*, 2010; Heled and Drummond, 2010; Edwards *et al.*, 2007). The performance of ILS-based methods on biological datasets has also been mixed, with concatenation often producing trees with very high bootstrap support that may not be completely correct, but ILS-based methods often producing trees with very low bootstrap support (Meredith *et al.*, 2011; Song *et al.*, 2012). Thus, we still do not know very much about the relative performance of ILS-based methods, how they compare to methods (such as concatenation) that do not take ILS into account, and what factors impact the absolute and relative performance of methods.

*Correspondence regarding this paper should be addressed to tandy@cs.utexas.edu

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of Texas at Austin on July 10, 2013

Published by Oxford University Press 2013.

In this paper, we report on a simulation study to evaluate a collection of methods for estimating species trees and gene trees in the presence of ILS. Our simulation study includes datasets generated under three model conditions from prior studies (Chung and Ané, 2011; Yu *et al.*, 2011b). One model condition has 17-taxon datasets that evolve under a strong molecular clock, and the other two model conditions have 11-taxon datasets that do not evolve under a clock. The amount of ILS varies between the three model conditions, ranging from relatively low amounts to very high amounts. Finally, estimated gene trees on these datasets have low average bootstrap support due to insufficient phylogenetic signal, reflecting conditions often encountered when sampling genes from throughout the genome. We study a wide range of methods for estimating species trees from multiple markers, including *BEAST (Heled and Drummond, 2010), both the population and concordance trees returned by BUCKy (Larget *et al.*, 2010; Ané *et al.*, 2007), MP-EST (Liu *et al.*, 2010), Phylonet-MDC (Than *et al.*, 2008; Yu *et al.*, 2011a), greedy consensus (GC) (also called the extended majority consensus), matrix representation with parsimony (MRP) (Baum and Ragan, 2004), and concatenation using maximum likelihood (CA-ML).

Our study revealed that many methods *have poor accuracy when the individual gene sequence alignments have low phylogenetic signal*. This vulnerability to poor signal affects all methods, but especially those that combine estimated gene trees; by comparison, *BEAST and CA-ML are relatively less impacted.

We developed an approach to address the vulnerability of species tree methods to low phylogenetic signal. We randomly partitioned the genes into subsets (which we call "supergenes"), estimated trees from these supergene alignments, and then used methods to estimate the species tree from the supergene trees. This approach did not produce statistically significant changes in accuracy on the 17-taxon datasets, but improved the accuracy of the trees estimated by combining estimated gene trees, often very substantially, on the 11-taxon datasets. Running *BEAST on the binned supergene alignments did not impact its accuracy, but did improve its scalability. Furthermore, when used with binning, several methods came close to being as accurate as *BEAST, while being orders of magnitude faster than *BEAST. Thus, this study suggests that highly accurate large-scale phylogenomic analyses may be achievable through a naive binning technique.

1 MATERIALS AND METHODS

See the Supplementary Materials for details.

Datasets: We used simulated 11-taxon (Chung and Ané, 2011) and 17-taxon (Yu *et al.*, 2011a,b) multi-gene datasets. The 11-taxon datasets have 100 genes, and the 17-taxon datasets have 32 genes. The protocols used in the two studies were fairly similar, however, the 11-taxon datasets reflect more heterogeneity, and hence are less idealized than the 17-taxon datasets. In each case, a model species tree was generated and a set of gene trees within each species tree (with one haploid individual sampled per species) produced under a coalescent process. This produces gene trees that can differ topologically from their associated species tree due to ILS. DNA sequences were then simulated down each gene tree under the Jukes-Cantor model. 100 replicates were generated for each model condition, and each replicate consisted of a set of true sequence alignments (i.e., one alignment for each gene).

The 11-taxon and 17-taxon datasets differ in some regards. First, the 17-taxon datasets evolved under a molecular clock, but the 11-taxon datasets did not. Second, the 11-taxon datasets have very short sequences (only 500 nucleotides), but the 17-taxon datasets have long sequences (2000 nucleotides). In the 11-taxon model conditions, there is substantial rate variation between the gene trees and species tree, but this is not true for the 17-taxon model conditions. Finally, the model conditions also varied in the amount of ILS, as we now discuss.

We calculated two statistics to evaluate the level of ILS in each model condition: the "average clade distance" between the true species tree and the true gene trees and the percentage of the true gene trees that have the same topology as the true species tree. The clade distance between two rooted trees (i.e., the rooted analog of the bipartition distance) is the total number of unique non-trivial clades (in one tree but not in both) divided by $2n - 4$. Thus, if two rooted trees on 7 leaves share exactly 2 clades in common, the clade distance is 60%. Using this metric, the 17-taxon datasets have the highest amount of ILS (avg. clade distance 25.7%). The 11-taxon datasets came in two forms, one with somewhat lower (but still high) amounts of ILS (avg. clade distance 14.8%), and one with very low amounts of ILS (avg. clade distance 2.9%). We refer to the two 11-taxon models as strongILS and weakILS, accordingly. The percent of gene trees that match the species tree also fits with this relative ranking: 73.1% for the 11-taxon weakILS datasets, 21.3% for the 11-taxon strongILS datasets, and only 1.7% for the 17-taxon datasets. Thus, the 17-taxon datasets have extremely high levels of ILS, but the 11-taxon strongILS also have a high level of ILS.

Selecting subsets of genes. For the 17-taxon datasets, we used the provided 8-gene and 32-gene datasets; for the 11-taxon datasets, we sampled from the 100 genes to produce subsets with the desired number of genes.

Gene Tree Estimation: We compared *BEAST, RAxML v. 7.3.1 (Stamatakis, 2006), and FastTree-2 (Price *et al.*, 2010), as gene tree estimators. We used 20 runs of RAxML on each of the alignments, and retained the tree with the best ML score; for FastTree-2, we used it with only one run (since it is deterministic, it is not improved by multiple runs). For *BEAST, we ran it as described below. We used RAxML with 400 bootstrap replicates for BUCKy and for Phylonet.

Species tree methods: We include *BEAST, MP-EST, BUCKy-pop, BUCKy-con, CA-ML, the greedy consensus (GC), Phylonet-MDC, and matrix representation with parsimony (MRP); see below for details. With the exception of *BEAST and CA-ML, these methods estimate the species tree by combining estimated gene trees; we refer to these as "summary methods". For MP-EST, MRP, and GC, we use the binary gene trees as input (these methods either require binary gene trees or have not been shown to improve by contracting low support branches (Yang and Warnow, 2011)).

We used *BEAST v. 1.6.2 (Heled and Drummond, 2010) in its default setting, and used the default point estimates for the gene trees and species tree. For a given *BEAST analysis, we discarded the first 10% of the trees returned by the analysis, and then sampled one (1) out of each 1000 of the remaining trees. We ran *BEAST long enough to return ESS values that were large enough to suggest possible convergence. Even after 150 hours of analysis, the ESS statistics for *BEAST on the 11-taxon 100-gene

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of Texas at Austin on July 10, 2013

strongILS datasets were very poor, suggesting that *BEAST had not converged; therefore, we omit results of *BEAST for these datasets.

We used MP-EST (Liu *et al.*, 2010) in its default setting, using MAXROUND=100000, and with RAxML gene trees rooted at the provided outgroup.

We used BUCKy (Ané *et al.*, 2007) with the default setting to compute two species tree estimations - the population tree (BUCKy-pop) and the concordance tree (BUCKy-con). We computed gene tree distributions using RAxML with bootstrapping and also using *BEAST as input to BUCKy. On each model condition and number of genes, we ran BUCKy using a sufficiently large number of MCMC iterations to reach sufficiently low standard deviations for the concordance factors to suggest possible convergence.

We used Phylonet v. 2.4 (Than *et al.*, 2008) for a version of the NP-hard MDC (Minimize Deep Coalescence) problem that takes gene tree branch support values into consideration. Although MDC is not statistically consistent (Than and Rosenberg, 2011), Phylonet-MDC can produce highly accurate species trees (Yang and Warnow, 2011) when applied to gene trees in which all the low support branches are collapsed. Phylonet provides a technique to solve this version of MDC exactly, even for unrooted gene trees (Bayzid and Warnow, 2012; Yu *et al.*, 2011a,b), which can be used on datasets with a small enough number of taxa; we used this exact method for MDC for the 11-taxon datasets, and Phylonet's heuristic method (which restricts the solution space to those trees all of whose bipartitions come from the input set of trees) for the 17-taxon datasets. We used Phylonet on the ML gene trees with all branches having bootstrap support less than 75% collapsed.

We used PAUP* to estimate MRP (matrix representation with parsimony), using the standard heuristic search, and also to compute a greedy consensus (GC) (also called the "extended majority consensus") of the estimated gene trees. Both of these analyses are performed on the binary gene trees estimated by maximum likelihood. We also studied CA-ML, using RAxML to infer a species trees from the super-alignment (without partitioning), and using 10 independent runs (-N 10).

Criteria: We report tree error using the missing branch rate (also known as the FN or "false negative" rate), which is the proportion of internal branches in the true tree defining bipartitions that are missing in the estimated tree. The use of FN rates rather than Robinson-Foulds (RF) rates is due to the observation that some of the methods for estimating trees produce unresolved trees, and the RF rates would be biased in favor of these methods (Rannala *et al.*, 1998). We tested for statistical significance using the Wilcoxon signed rank test.

Experiments: The first experiment compared the "fast" methods (all methods except *BEAST and BUCKy) on 100 replicates of the 11-taxon and 17-taxon datasets, varying the number of genes, using RAxML to estimate gene trees. The second experiment compared the full set of methods on 20 replicates of these model conditions, again using RAxML to estimate gene trees. We explored the accuracy of gene trees estimated by RAxML, FastTree, and *BEAST in the third experiment. The fourth experiment evaluated the accuracy of species trees computed for gene trees estimated by *BEAST. The fifth experiment then examined the impact of binning genes into supergenes, using a simple "naive" binning technique.

2 RESULTS

We show results evaluating computational aspects of the different methods, and then results of the five basic experiments exploring accuracy. See the Supplementary Materials for additional details.

Computational Issues. The phylogenomic pipelines we studied differed dramatically in terms of their running times, making some methods infeasible to use on some datasets within the limits of this study. Due to space limitations, we present a brief discussion of the computational requirements of the different methods, and direct the interested reader to the Supplementary Materials for full results.

Pipelines that used *BEAST took the most time, with running times of 80-150 hours for the 50-gene datasets with 11 taxa; analyses of the 100-gene datasets with 11 taxa did not converge, even in 150 hours. The pipelines with BUCKy, when used with distributions computed using RAxML bootstrapping, took up to 5 hours, but were able to be run on even the 100-gene 11-taxon datasets. Pipelines with Phylonet when used with the RAxML bootstrap trees (restricted to the high support edges) took up to 2 hours per dataset (almost all of that for running RAxML). Pipelines with MRP, GC, and CA-ML took just a few minutes per dataset.

Because of these computational issues, we only ran BEAST on unbinned datasets with at most 50 genes (and even these were very computationally intensive). We also did not run *BEAST or unbinned BUCKy on more than 20 replicates for any model condition. Therefore, in the remaining study, we show results for the "fast" methods (everything but *BEAST and BUCKy) on 100 replicates of the model conditions, and we examine *BEAST and BUCKy on only 20 replicates of the model conditions. We do, however, show results using BUCKy with binning on 100 replicates of some model conditions. In total, we estimate that we used at least 5000 CPU hours, just for the *BEAST runs.

Experiment 1: The first experiment explored the accuracy of the "fast" methods for estimating species trees, i.e., CA-ML, MP-EST, MRP, Phylonet, and GC; see Figures 1, 2, and 3. CA-ML had the best accuracy, with very large improvements over other methods on the 11-taxon datasets and small improvements on the 17-taxon datasets. All the improvements are statistically significant: $p < 0.003$ for the 11-taxon strongILS with up to 100 genes and the 11-taxon weakILS with up to 25 genes, and $p \leq 0.04$ for the 17-taxon datasets.

Experiment 2: We then evaluated BUCKy-pop, MP-EST, *BEAST, CA-ML, and BUCKy-con. Because *BEAST is computationally intensive, the analyses were limited to 20 replicates per datapoint. See Figures 4, 5, and 6.

Note that *BEAST and CA-ML are the two most accurate methods on these data, with the greatest improvement over the other methods on the 11-taxon weakILS datasets and the least improvement on the 17-taxon datasets. The relative performance between CA-ML and *BEAST varied, with CA-ML better in some cases and worse in others, and often the difference was small.

BUCKy-pop is in third place, and even matched the accuracy of *BEAST on the 11-taxon strongILS datasets with 25 genes. A comparison between BUCKy-pop and BUCKy-con shows that they had very close accuracy in most cases, but that BUCKy-pop was sometimes more accurate than BUCKy-con (e.g., on the 11-taxon strongILS datasets with 25 or 50 genes), with statistically significant differences ($p = 0.003$ and $p = 0.035$, respectively).

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of Texas at Austin on July 10, 2013

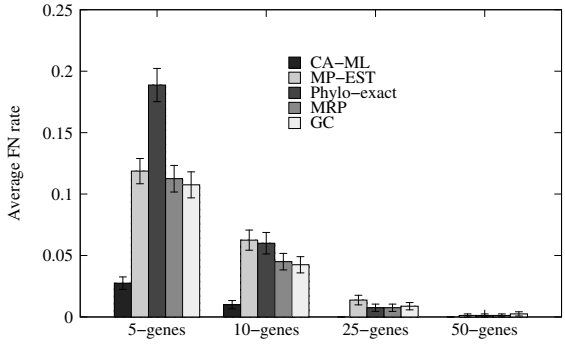

Fig. 1. Results for "fast" methods on 100 replicate 11-taxon weakILS models. CA-ML uses just the input alignments, and the other methods use gene trees estimated using RAxML. We show means with standard error bars.

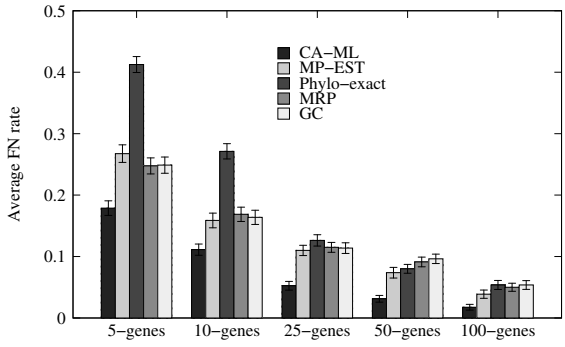

Fig. 2. Results for "fast" methods on 100 replicate 11-taxon strongILS models. We show results (means with standard error bars) for up to 100 genes.

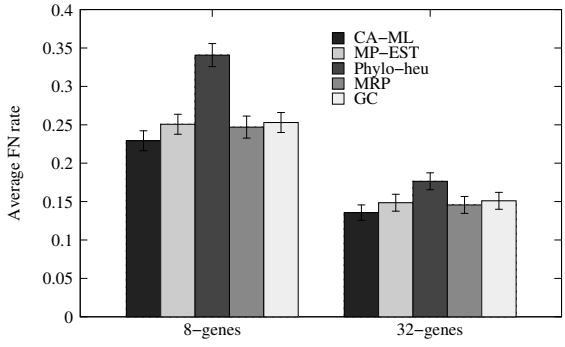

Fig. 3. Results for "fast" methods on 100 replicate 17-taxon models. We show means with standard error bars.

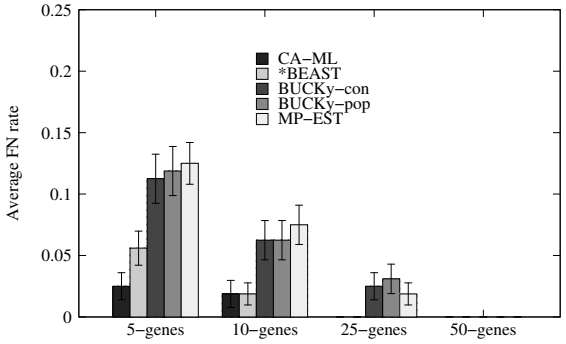

Fig. 4. Results for *BEAST, BUCKy, MP-EST, and CA-ML on 20 replicate 11-taxon weakILS datasets. We show means with standard error bars. CA-ML and *BEAST return the true tree on the 25-gene case, and all methods shown return correct trees on the 50-gene case. Therefore, no results are shown for datasets with 100 genes.

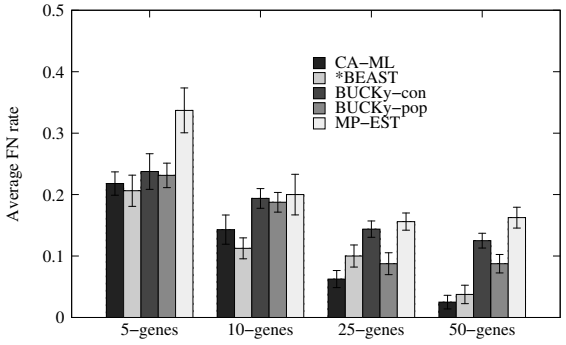

Fig. 5. Results for *BEAST, BUCKy, MP-EST, and CA-ML on 20 replicate 11-taxon strongILS datasets. We show means with standard error bars.

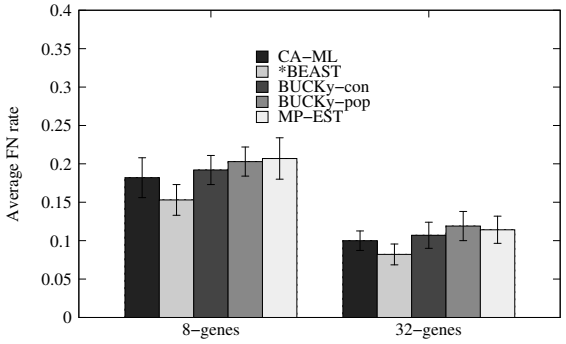

Fig. 6. Results for *BEAST, BUCKy, MP-EST, and CA-ML on 20 replicate 17-taxon datasets.

Of these various observations, the most important here are the following: CA-ML and *BEAST had the best accuracy on these data; the gap between methods was least on the 17-taxon datasets, and greatest on the 11-taxon weakILS datasets; and all methods became less accurate with increases in the amount of ILS. It is easy to understand why the methods that are not statistically consistent under ILS increase in error with the degree of ILS, but not that easy to to understand why *BEAST, MP-EST, and BUCKy-pop decrease in accuracy with increases in ILS. Here we offer a possible explanation for this trend.

Recall that the conditions that favor ILS are very short branches in the species tree. Thus, the conditions that increase the amount of ILS (i.e., short branches) also make it challenging to estimate the gene trees. In fact, the weakILS model trees have long branches (and are called "LB" in Chung and Ané (2011)), and the strongILS model

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of Texas at Austin on July 10, 2013

trees have short branches (and are called "SB" in Chung and Ané (2011)), and gene trees estimated using RAxML have lower error on the 11-taxon weakILS model conditions than on the strongILS model conditions (30% vs. 40%, respectively). Therefore, it's not at all surprising that species trees estimated by combining gene trees under the highILS model conditions would have higher error than species trees estimated by combining gene trees under the lowILS model condition. Finally, the 17-taxon datasets had the highest level of ILS, and on these data the summary methods perform the worst. Note that this vulnerability applies to all summary methods, even to the statistically consistent methods like MP-EST and BUCKy-pop.

Experiments 3 and 4: The next two experiments attempted to understand why *BEAST was so much more accurate than the summary methods. In Experiment 3, we evaluated the accuracy of the gene trees estimated by *BEAST, FastTree-2, and RAxML for all three model conditions, and observed that *BEAST produces substantially more accurate gene trees than FastTree-2 and RAxML. For example, under the 11-taxon weakILS model condition with 50 genes, gene trees estimated by *BEAST had only 3.3% error while gene trees estimated by RAxML had 31.9% error - a reduction of roughly 90%. More generally, the greatest improvement was for the model condition with the lowest rate of ILS (11-taxon weakILS), and the least improvement was for the model condition with the highest rate of ILS (17-taxon datasets). However, even on the 17-taxon datasets, the reduction was at least 50%. Results for the 17-taxon datasets are given in Figure 7; see the Supplementary Materials for the other results. These analyses also show that RAxML has a small but statistically significant advantage over FastTree (differences in missing branch rate of at most 1.7% on the 11-taxon weakILS conditions, 2.5% on the 11-taxon strongILS conditions, and 1.1% on the 17-taxon conditions).

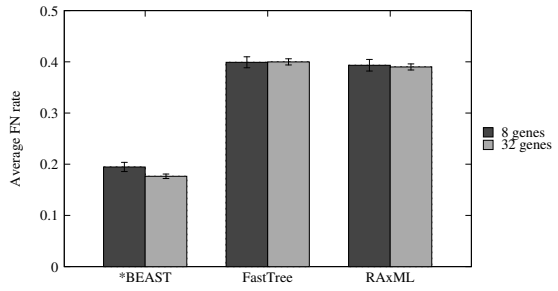

Fig. 7. Gene tree estimation error rates on 17-taxon datasets. Average and standard error bars (over 20 replicates) of *BEAST, RAxML, and FastTree-2.

In Experiment 4 (see Supplementary Materials), we examine the results of using the summary methods (i.e., BUCKy-con, BUCKy-pop, Phylo-MDC, MP-EST, MRP, and GC) on inputs of gene trees estimated by *BEAST. These experiments show that species trees estimated by combining gene trees estimated by *BEAST are essentially as accurate as the species trees estimated by *BEAST, and there are no statistically significant differences. This suggests that the accuracy obtained by *BEAST is primarily due to its improved gene tree accuracy, rather than to some sophisticated way of combining accurate gene trees.

Experiment 5: Since reduced phylogenetic signal in individual gene sequence alignments impacts the summary methods, we considered the following approach.

- Step 1: Partition the genes into bins,
- Step 2: Within each bin, compute a "supergene" alignment, by concatenating the alignments for the genes in the bin,
- Step 3: Compute a "supergene tree" using ML on each supergene alignment, and
- Step 4: Estimate the species tree from the set of supergene trees (using one of the "summary" methods), or from the set of supergene alignments (using *BEAST, for example).

Since this binning technique can put genes into the same bin that may not share the same history, this approach is a blend of CA-ML and the species tree estimation technique used in Step 4.

Our motivation for this approach is empirical. The hope is that since each supergene has more sites, ML trees estimated on each supergene might be more resolved than ML trees estimated on the individual genes. If the genes placed in the same bin have the same gene tree topology, then this approach could potentially lead to higher accuracy gene trees. If the genes placed in the same bin have different gene tree topologies, then they may not represent any gene tree that appears in the dataset, but may be closer to the species tree. In either case, summary methods applied to these supergene trees might be more accurate than summary methods applied to the individual gene trees.

Evaluating binning on fast methods. In our initial experiment, we explored the impact of binning on the fast methods on 100 replicate datasets of each model condition. We used bins with 5 genes each for the 11-taxon datasets, and bins with 4 genes each for the 17-taxon datasets. We do not present results for *BEAST (unbinned or binned) or BUCKy on unbinned datasets due to computational issues; however, we do show results for BUCKy on binned datasets. Note also that because we ran CA-ML *without partitioning*, binning has no impact on CA-ML.

Results for the 11-taxon strongILS datasets are shown in Figures 8, 9, and 10. See the Supplementary Materials for results on the 11-taxon weakILS datasets and 17-taxon 32-gene datasets. Binning improved accuracy for all methods for the 11-taxon datasets (both weakILS and strongILS), but not always statistically significantly. Results on the 17-taxon datasets showed that binning did not have any statistically significant impact on any method ($p > 0.22$).

On the 11-taxon weakILS datasets with 25 genes, all methods improved in accuracy. These improvements were statistically significant for MP-EST and Phylonet ($p = 0.002$ and $p = 0.016$, respectively), but not for the other summary methods. However, all methods were already highly accurate without binning.

Binning produced large reductions in error for many methods on the 11-taxon strongILS datasets with 25 genes. Phylonet-MDC showed the largest improvement (reduction from 12.6% to 9.6%, $p = 0.002$), MP-EST showed the second largest improvement (reduction from 11.0% to 8.8%, $p = 0.021$), and GC and MRP showed the least improvement (reductions of at least 1%, but not statistically significant, $p = 0.16$ and $p = 0.18$, respectively).

On 50 genes, all methods had reductions in error, with Phylonet-MDC showing the largest improvement (reduction from 8.9% to

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of Texas at Austin on July 10, 2013

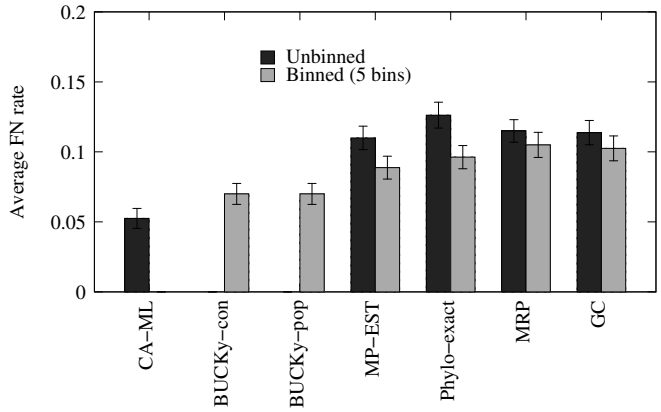

Fig. 8. Results of binning experiments of the fast methods on 100 replicates of the 11-taxon 25-gene strongILS datasets. Each bin contains 5 genes. We omit BUCKy on unbinned genes and *BEAST (binned or unbinned) because these are too slow to run on all 100 replicates within our time limits. CA-ML is not impacted by binning because it uses an unpartitioned analysis.

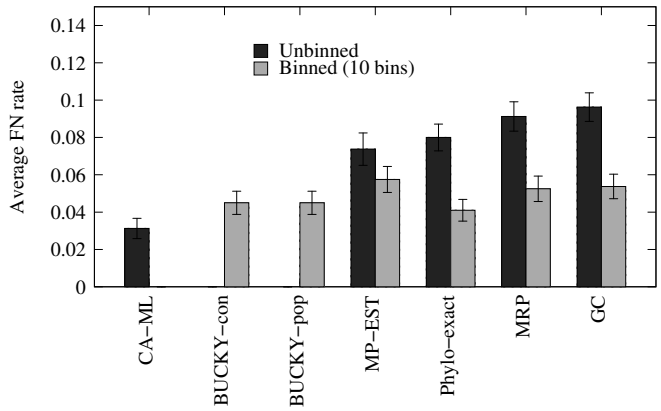

Fig. 9. Results of the binning experiment for the fast methods on 100 replicates of the 11-taxon 50-gene strongILS datasets. Each bin contains 5 genes. We omit BUCKy on unbinned genes and *BEAST (whether binned or unbinned) because these are too slow to run on all 100 replicates within our time limits. CA-ML is not impacted by binning, because it uses an unpartitioned analysis.

4.1%, $p < 10^{-5}$), GC showing the next largest improvement (reduction from 9.6% to 5.4%, $p < 10^{-4}$), MRP the next largest improvement (reduction from 9.1% to 5.3%, $p < 10^{-4}$), and MP-EST with the smallest improvement (reduction from 7.3% to 5.7%, but not statistically significant, $p = 0.057$).

On 100 genes, all methods had reductions in error, and again Phylonet-MDC had the largest improvement (reduction from 5.4% to 2.4%, $p < 10^{-3}$), GC had the next largest (reduction from 5.4% to 3.4%, $p = 0.007$), and MRP and MP-EST showing smaller improvements that were not statistically significant ($p > 0.07$).

Thus, binning improved the accuracy of all methods on the 11-taxon model conditions, with large reductions for the strongILS conditions and smaller (but still significant) reductions on the weakILS conditions. The greatest improvements were for intermediate numbers of genes, in which the methods used without

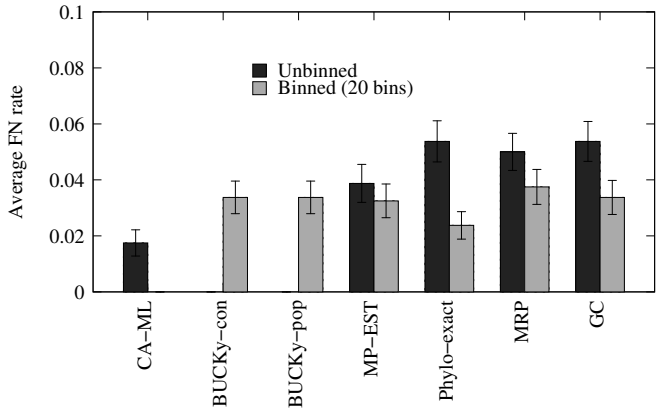

Fig. 10. Results of the binning experiment for the fast methods on 100 replicates of the 11-taxon 100-gene strongILS datasets. Each bin contains 5 genes. We omit BUCKy on unbinned genes and *BEAST (whether binned or unbinned) because these are too slow to run on all 100 replicates within our time limits. CA-ML is not impacted by binning because it uses an unpartitioned analysis.

binning still had some error (and hence could be improved), but had enough genes so that binning produced a reasonable number of supergenes. Binning had no statistically significant impact on the 17-taxon model conditions with 100 replicates ($p > 0.22$). CA-ML was still the most accurate of all tested methods, but some methods came close to the accuracy of CA-ML when used with binning.

Evaluating binning on all methods. Due to the computational effort in using *BEAST, we limited the analysis to only 20 replicates of each model condition. We limit this discussion to the impact of binning on *BEAST and BUCKy, since the analysis on 100 replicate datasets allowed us to evaluate binning on the other methods with a higher number of replicates. Results on the 11-taxon weakILS datasets with 25 genes are shown in the Supplementary Materials; all methods improved, but the improvement was statistically significant only for BUCKy-pop (reduction from 3.1% to 0.0%, $p = 0.03$). Results on the 20-replicate 17-taxon datasets (see Supplementary Materials) show no statistically significant differences ($p > 0.3$) for BUCKy-pop, BUCKy-con, and *BEAST, and all differences were very small (at most 0.5%). Results on 11-taxon strongILS datasets are shown in Figures 11 and 12. BUCKy-pop generally improved with binning, but the results were not statistically significant ($p > 0.06$). BUCKy-con also improved using binning (reduction in error from 14.3% to 9.4% on 25 genes, from 12.5% to 5% on 50 genes, and from 5.6% to 2.5% on 100 genes), and the changes on 25 and 50 genes were statistically significant ($p = 0.018$ and $p = 0.005$, respectively).

The impact of binning on *BEAST is interesting. On the 100-gene datasets, we were unable to run *BEAST to convergence without binning even with 150 hours of analysis; however, *BEAST was able to reach acceptable ESS values in only 10 hours using 4 threads when run on 20 bins with 5 genes each. Thus, the use of binning did not impact the accuracy of *BEAST, but it made it feasible to use *BEAST on datasets with large numbers of genes.

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of Texas at Austin on July 10, 2013

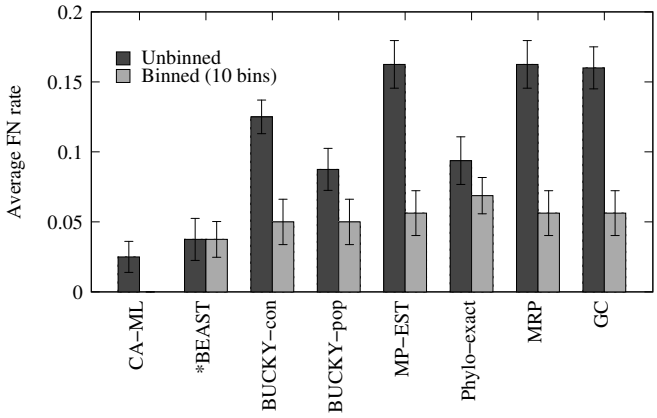

Fig. 11. Results of the binning experiment for all methods on 20 replicate 11-taxon 50-gene strongILS datasets. CA-ML is not impacted by binning since it is an unpartitioned analysis. Each bin contains 5 genes. Average and standard error bars shown.

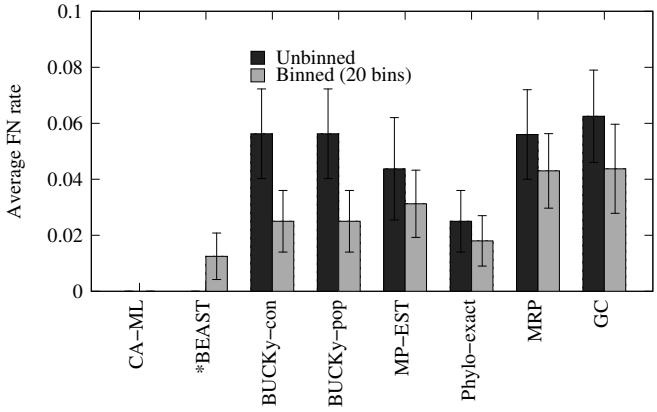

Fig. 12. Results of the binning experiment for all methods (except *BEAST) on 20 replicate 11-taxon 100-gene strongILS datasets. Each bin contains 5 genes. Average and standard error bars shown. We omit *BEAST on unbinned genes because it could not run to convergence on this dataset within the time limit; however, we show results for *BEAST on the binned datasets. CA-ML returns the true tree on these data.

3 DISCUSSION

The main purpose of this study was to evaluate methods for estimating species trees in the presence of ILS under realistic conditions. Since many real world phylogenomic analyses have to contend with genes with poor phylogenetic signal (Salichos and Rokas, 2013), we specifically examined conditions in which estimated gene trees were only partially resolved. As expected, the number of genes and amount of ILS impacted the accuracy of the methods we tested, so that all methods returned more accurate trees with increasing numbers of genes and decreasing levels of ILS. However, in addition to these expected results, we make the following observations:

First, all the summary methods we studied were impacted by gene tree estimation error. In contrast, although *BEAST and CA-ML were also affected by the amount of phylogenetic signal in the multiple sequence alignments, the impact was generally less.

Second, CA-ML and *BEAST had similar accuracy, and were generally more accurate than the summary methods we tested.

Third, *BEAST produced dramatically more accurate gene trees than ML analyses on the alignments, and summary methods on these gene trees produced species trees as accurate as *BEAST species trees, explaining why *BEAST produces more accurate species trees than other methods.

Fourth, the naive binning technique we tested generally improved coalescent-based methods. It improved the scalability of *BEAST without impacting its accuracy, making it feasible to use *BEAST on datasets with many genes. Binning also improved the accuracy of species trees estimated using the summary methods we tested on the 11-taxon conditions, although the degree of impact depended on the number of genes and the level of ILS. Finally, binning had no statistically significant impact on the 17-taxon conditions.

The observation that summary methods are vulnerable to poor phylogenetic signal in the gene sequences is consistent with the empirical studies reported by Salichos and Rokas (2013) and Meredith *et al.* (2011), and this study would seem to suggest that naive binning would be helpful for species tree estimation under these circumstances. However, naive binning could have unforeseen negative consequences if it puts genes with different histories into the same bin. The good performance of naive binning under the 11-taxon strongILS condition we explored suggests that it may be somewhat robust in practice, even under relatively high rates of ILS. However, since naive binning did not improve the accuracy of summary methods on the 17-taxon datasets (which had the highest rate of ILS), this suggests that naive binning could reduce accuracy when the amount of ILS is very large. There is also a possibility that binning will only be helpful when concatenation is more accurate than the coalescent-based methods. Therefore, further research is needed in order to assess the conditions in which binning improves or reduces accuracy. See Supplementary Materials, Section 4, for additional discussion about this issue.

One of the most interesting results in this paper is the observation that CA-ML outperformed all the ILS-based methods that operate by combining estimated gene trees. This observation would seem to run counter to other simulation studies that have shown that concatenation can return incorrect trees with high confidence and can also produce trees that are less accurate than trees estimated by ILS-based methods (Larget *et al.*, 2010; Liu *et al.*, 2010; Edwards *et al.*, 2007; DeGiorgio and Degnan, 2010; Leaché and Rannala, 2011; Heled and Drummond, 2010). However, these studies used simulated datasets that evolve under a strong molecular clock (a condition that may benefit some coalescent-based methods more than concatenation (DeGiorgio and Degnan, 2010)), few taxa, and generally had many genes relative to the number of taxa (and estimated gene trees on these alignments may have been fairly accurate). In contrast, our study had 11- and 17-taxon datasets, at most 100 genes, and poorly estimated gene trees. Thus, it seems that there are conditions under which some ILS-based methods might outperform CA-ML, and other conditions under which CA-ML might outperform the ILS-based methods. In particular, it is possible that the critical issue is the number of genes, and that ILS-based methods will have better accuracy than concatenation when the number of genes is large enough. Clearly, further research is needed in order to understand which conditions favor each type of approach. See Section 4 in Supplementary Materials for more discussion of these issues.

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of Texas at Austin on July 10, 2013

4 CONCLUSIONS AND FUTURE WORK

Under the conditions of our experiments (at least 11 taxa, at most 100 genes, and low signal per gene sequence alignment) we observed relatively poor species tree estimations using standard summary methods, and more accurate results from concatenation or from *BEAST, a method that co-estimates gene trees and species trees. However, the current co-estimation methods (including *BEAST) are computationally intensive and may not be feasible for use with more than 100 genes or more than 20 species. This study showed that a simple binning technique was able to make dramatic improvements in scalability for *BEAST, and generally improve the accuracy of summary methods, thus making some of these methods nearly as accurate as *BEAST.

This study should not be interpreted as recommending the use of naive binning, but instead as an indication of the potential for binning techniques to improve species tree estimation. For example, statistical techniques could be used to estimate whether a set of genes is likely to have a common tree, so that bins would only include genes expected to have a common history. Also, while concatenation performed well in this study, we conjecture that new techniques designed to handle markers with limited phylogenetic signal, might outperform concatenation even under these model conditions. Whether these new techniques will employ binning, or other ways of working with poorly estimated gene trees, the potential for substantial advances in species tree estimation could be great.

5 FUNDING AND ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation [grants DBI-1062335, DEB 0733029, and DBI-0735191] and the Fulbright Foundation [International Fulbright Science and Technology PhD Award to MSB]. We wish to thank the anonymous reviewers, and C. Ané, J. Degnan, B. Faircloth, M. Hahn, S. Mirarab, and J. Pease for their helpful comments. We also thank Y. Chung, C. Ané, Y. Yu, and L. Nakhleh for the use of their datasets.

REFERENCES

Ané, C., Larget, B., Baum, D. A., Smith, S. D., and Rokas, A. (2007). Bayesian estimation of concordance among gene trees. *Mol Biol Evol*, **24**, 412–426.

Baum, B. and Ragan, M. A. (2004). The MRP method. In O. R. P. Bininda-Emonds, editor, *Phylogenetic Supertrees: combining information to reveal The Tree Of Life*, pages 17–34. Kluwer Academic, Dordrecht, the Netherlands.

Bayzid, M. S. and Warnow, T. (2012). Estimating optimal species trees from incomplete gene trees under deep coalescence. *J Comput Biol*, **19(6)**, 591–605.

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

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

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

Degnan, J. H., DeGiorgio, M., Bryant, D., and Rosenberg, N. A. (2009). Properties of consensus methods for inferring species trees from gene trees. *Syst Biol*, **58**, 35–54.

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

Edwards, S. V., Liu, L., and Pearl, D. K. (2007). High-resolution species trees without concatenation. *Proc Natl Acad Sci*, **104**(14), 5936–5941.

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

Huang, H., He, Q., Kubatko, L., and Knowles, L. L. (2010). Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. *Syst Biol*, **59**(5), 573–583.

Knowles, L. L. (2009). Estimating species trees: Methods of phylogenetic analysis when there is incongruence across genes. *Syst Biol*, **58**(5), 463–367.

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

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

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

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

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

Meredith, R. W., Janecka, J. E., Gatesy, J., Ryder, O. A., Fisher, C. A., Teeling, E. C., Goodbla, A., Eizirik, E., Simo, T. L. L., Stadler, T., et al. (2011). Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science*, **334**(6055), 521–524.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE*, **5**(3), e9490.

Rannala, B., Huelsenbeck, J., Yang, Z., and Nielsen, R. (1998). Taxon sampling and the accuracy of large phylogenies. *Syst Biol*, **47**, 702–710.

Salichos, L. and Rokas, A. (2013). Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, **497**(7449), 3627–331.

Song, S., Liu, L., Edwards, S. V., and Wu, S. (2012). Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc Natl Acad Sciences*, **109**(37), 14942–14947.

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

Steel, M. (2010). Consistency of Bayesian inference of resolved phylogenetic trees. ArXiv: 1001.2864.

Than, C. V. and Rosenberg, N. A. (2011). Consistency properties of species tree inference by minimizing deep coalescences. *J Comp Biol*, **18**, 1–15.

Than, C. V., Ruths, D., and Nakhleh, L. (2008). PhyloNet: A software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinf*, **9**, 322.

Wakeley, J. (2009). *Coalescent Theory*. Roberts.

Yang, J. and Warnow, T. (2011). Fast and accurate methods for phylogenomic analyses. *BMC Bioinf*, **12(Suppl 9:S4)**.

Yu, Y., Warnow, T., and Nakhleh, L. (2011a). Algorithms for MDC-based multi-locus phylogeny inference. In *Proc RECOMB 2011*, 531–545.

Yu, Y., Warnow, T., and Nakhleh, L. (2011b). Algorithms for MDC-based multi-locus phylogeny inference: Beyond rooted binary gene trees on single alleles. *J Comp Biol*, **18**(11), 1543–1559.

Downloaded from http://bioinformatics.oxfordjournals.org/ at University of Texas at Austin on July 10, 2013