TECHNICAL RESPONSE

AVIAN GENOMICS

Response to Comment on "Statistical binning enables an accurate coalescent-based estimation of the avian tree"

Siavash Mirarab,[1,2] Md. Shamsuzzoha Bayzid,[1] Bastien Boussau,[3] Tandy Warnow[1,4]*

Liu and Edwards argue against the use of weighted statistical binning within a species tree estimation pipeline. However, we show that their mathematical argument does not apply to weighted statistical binning. Furthermore, their simulation study does not follow the recommended statistical binning protocol and has data of unknown origin that bias the results against weighted statistical binning.

In (*1*), we introduced statistical binning, a method to improve species tree estimation from multiple loci when true gene trees can differ from the species tree due to incomplete lineage sorting (ILS) (*2*). When ILS is present, unpartitioned concatenation using maximum likelihood (ML) can be statistically inconsistent and fail to converge to the species tree as the number of loci increases (*3*). To address this challenge, statistically consistent coalescent-based "summary methods" have been developed [e.g., (*4*, *5*)]. However, all current proofs of statistical consistency for standard coalescent-based summary methods assume error-free gene trees. Furthermore, concatenated analyses can be more accurate than summary methods in the presence of substantial gene tree estimation error resulting from low phylogenetic signal (*1*, *5*–*11*), a problem that confronted the Avian Phylogenomics Consortium (*12*). Because simulations showed that species trees computed with statistical binning followed by the summary method maximum pseudo-likelihood estimation of species trees (MP-EST) (*4*) "produced more accurate estimated species trees compared to MP-EST applied to unbinned gene data sets that have low phylogenetic signal" (*12*), the Avian Phylogenomics Consortium (which included Liu and Edwards) decided to use statistical binning with MP-EST to compute a coalescent-based avian species tree (*12*).

Because pipelines using statistical binning are not statistically consistent (*11*), we developed weighted statistical binning (WSB) (see Fig. 1) and proved that as both the number of loci and sequence length per locus increase, WSB followed

[1]Department of Computer Science, University of Texas at Austin, Austin, TX, USA. [2]Department of Electrical and Computer Engineering, University of California at San Diego, San Diego, CA, USA. [3]Laboratoire de Biométrie Biologie Evolutive, Université de Lyon, France. [4]Departments of Bioengineering and Computer Science, The University of Illinois at Urbana-Champaign, Urbana, IL, USA.
*Corresponding author. E-mail: warnow@illinois.edu

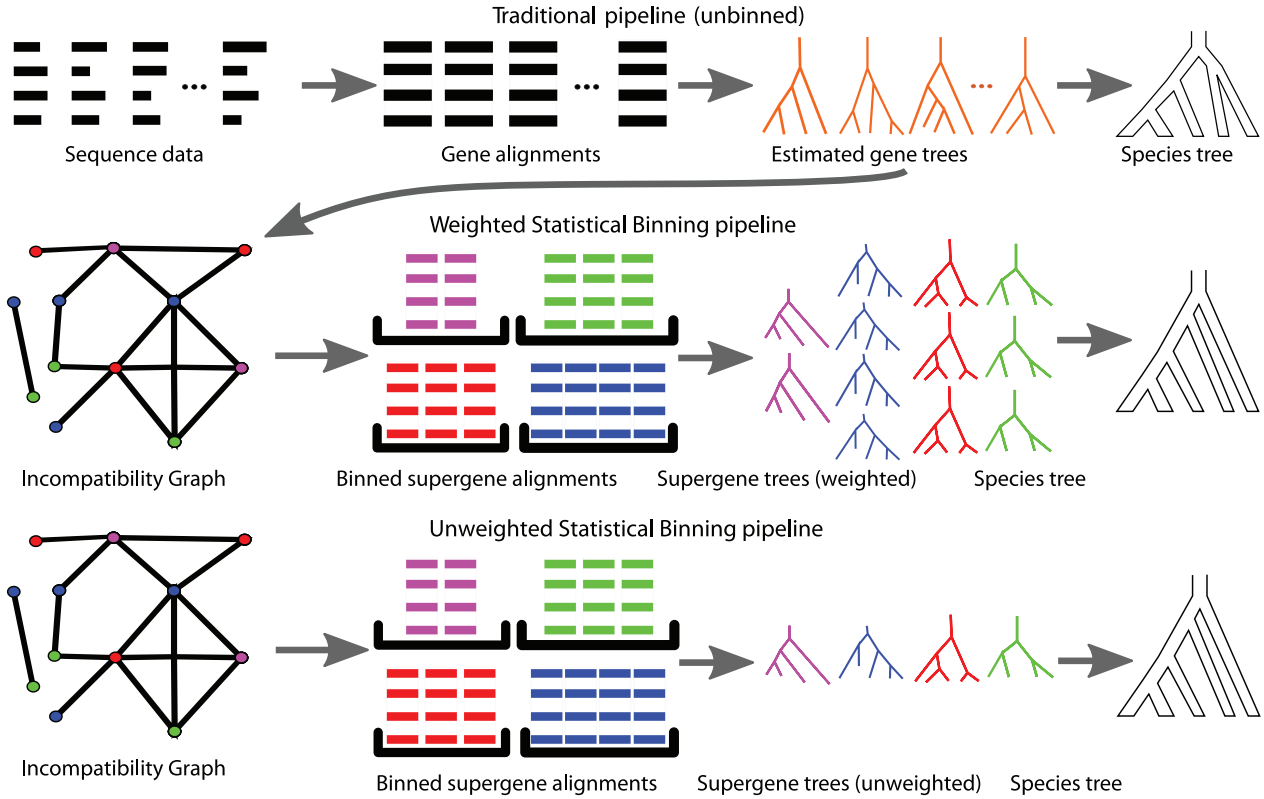

Fig. 1. Phylogenomic pipelines: unbinned (top), weighted statistical binning (middle), and unweighted statistical binning (bottom). Statistical binning divides the genes into bins that have no highly supported conflicts, estimates supergene trees on each bin, and then combines the supergene trees using the selected summary method. The WSB method differs from unweighted binning by replicating each supergene tree by the number of genes within its bin.

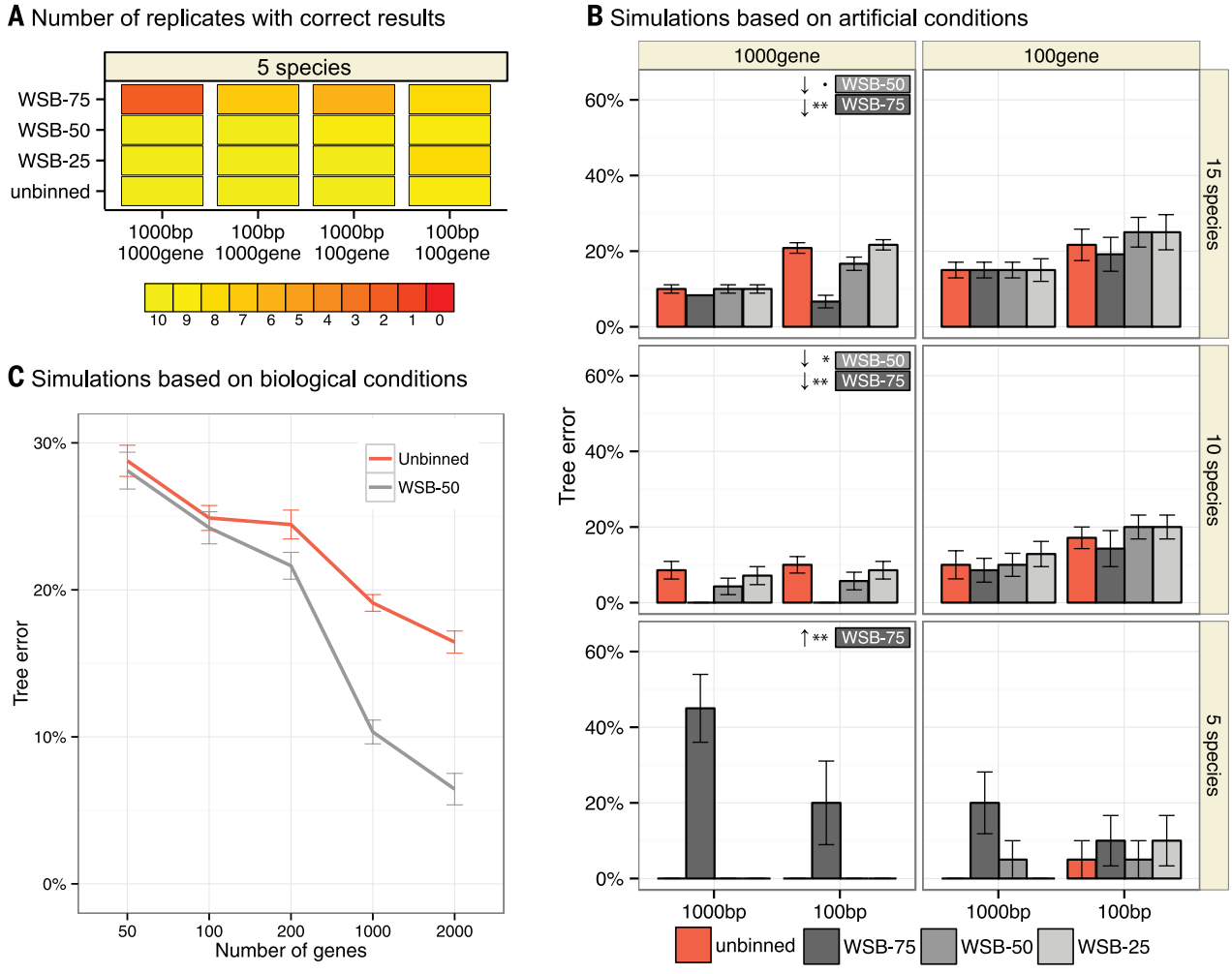

Fig. 2. Simulation studies evaluating the impact of WSB on MP-EST analyses. (**A**) Tree error rates (percentage of missing branches) for species trees estimated with WSB (MLBS gene trees, fully partitioned ML analyses) and unbinned analyses on the five-species data sets studied by Liu and Edwards, and similar model conditions with 10 and 15 species, all with 1000 genes. Symbols: ↑ indicates that using WSB increases species tree estimation error; ↓ indicates that using WSB decreases error. (**B**) Number of replicates for which the true species tree is recovered on the five-species data sets studied by Liu and Edwards. (**C**) Results on simulated avian data sets with 1X branch lengths and 500-bp sequences per locus, based on the avian tree from (*12*). MP-EST analyses are based on multilocus bootstrapping, and unpartitioned ML analyses are used to compute supergene trees.

by any consistent summary method converges in probability to the true species tree (i.e., pipelines using WSB are statistically consistent) (*11*). However, it is not known whether WSB, fully partitioned concatenation, and standard summary methods are consistent or inconsistent, given bounded length sequences for each locus (*6*, *14*).

Liu and Edwards attempt to prove that WSB will not converge in probability to the species tree when the number of loci increases but the sequence length per locus is bounded (*13*). Their mathematical argument does not apply to WSB, because it relies on a theorem in (*3*) that unpartitioned ML is inconsistent in the presence of ILS, whereas WSB uses fully partitioned ML to estimate supergene trees. Furthermore, the theorem in (*3*) cannot be applied to fully partitioned ML, and it is unknown whether fully partitioned ML is consistent in the presence of ILS (*14*). Therefore, we reject the statement in (*13*) that "all forms of binning...are statistically inconsistent in large regions of parameter space." Statistical inconsistency requires a mathematical proof, so Liu and Edwards have not established statistical inconsistency for WSB.

Liu and Edwards present a five-species simulation study in which they explored the impact of both weighted and unweighted statistical binning (*13*). However, many of the supergene alignments for parameter set 2 [1000 genes, 100 base pairs (bp)] contain sequences that did not come from any of the individual gene sequence alignments generated by Liu and Edwards. Importantly, these extraneous sequences in the supergene alignments reduce the accuracy of binned analyses but have no effect on unbinned analyses and thus introduce bias into the experiment.

Our analysis of their data (Fig. 2A) produced more accurate results for WSB than they reported. The differences on parameter set 2 are mostly due to extraneous data in the supergene alignments they generated. However, we also determined that they used unpartitioned ML to compute supergene trees, which also reduced accuracy of WSB, whereas we used fully partitioned ML, as required for WSB's theoretical guarantees.

In addition to the five-species data sets studied by Liu and Edwards, we explored similar model conditions with 10 and 15 species (Fig. 2B). Ten replicate data sets were generated under each model condition (number of taxa, number of genes, and sequence length). We evaluated the statistical significance of differences between methods, correcting for multiple tests (using false discovery rate correction, n = 18 statistical tests). There are no statistically significant differences for analyses with 100 genes (Fig. 2B). Results obtained on 1000 genes (Fig. 2B) show that WSB with bootstrap support (BS) threshold of 50% and 75% produced statistically significant reductions

in the species tree error rate for 15- and 10-species data sets ($P < 0.007$ for BS threshold of 75% on both 10- and 15-species data sets, $P = 0.044$ for BS threshold of 50% on 10-species data sets, and $P = 0.096$ for BS threshold of 50% on 15-species data sets). On the five-species, 1000-gene data sets (Fig. 2B), WSB and unbinned analyses were identical except when the BS threshold was 75%, which led to a statistically significant increase in the species tree estimation error ($P = 0.0069$). Thus, the effect of WSB depends on the model condition and BS threshold but was neutral to highly beneficial for all 10- and 15-taxon data sets that we analyzed.

The simulation condition explored by Liu and Edwards has a model species tree with only five species and very high ILS (i.e., the average topological distance between true gene trees and species trees is 82%), and evolves sequences under a strict molecular clock. The reduction in accuracy produced by using WSB on these data is consistent with a trend we reported in (*11*), where we observed that WSB can reduce accuracy on data sets with very high ILS and small numbers of species. However, for larger data sets, WSB almost always improved the accuracy of species tree topologies and branch lengths, and reduced the incidence of strongly supported false positive branches (*11*). For example, WSB led to substantial improvements on the 48-taxon avian simulated data sets, which have a fairly high ILS level (average distance between true gene trees and species tree of 47%) (Fig. 2C).

Liu and Edwards only examined five-taxon data sets. Although performance on very small numbers of species is of interest for some analyses, the avian phylogenomics project (*12*) and many other phylogenomic data sets have substantially larger taxon sets. Thus, the research community needs species tree estimation methods that are highly accurate for large taxon data sets with gene tree estimation error. Although there is progress in the development of summary methods with good accuracy under these conditions (*5*), all current summary methods are affected by gene tree estimation error (*1*, *5*–*11*). Hence, WSB provides a useful tool for species tree estimation in modern phylogenomic analysis.

REFERENCES AND NOTES

1. S. Mirarab, M. S. Bayzid, B. Boussau, T. Warnow, *Science* **346**, 1250463 (2014).
2. W. P. Maddison, *Syst. Biol.* **46**, 523–536 (1997).
3. S. Roch, M. Steel, *Theor. Popul. Biol.* **100**, 56–62 (2015).
4. L. Liu, L. Yu, S. V. Edwards, *BMC Evol. Biol.* **10**, 302 (2010).
5. S. Mirarab, T. Warnow, *Bioinformatics* **31**, i44–i52 (2015).
6. S. Roch, T. Warnow, *Syst. Biol.* **64**, 663–676 (2015).
7. S. Patel, R. Kimball, E. Braun, *J. Phylogenet. Evol. Biol.* **1**, 2 (2013).
8. S. Mirarab, M. S. Bayzid, T. Warnow, *Syst. Biol.* 10.1093/sysbio/syu063 (2014).
9. M. P. Simmons, J. Gatesy, *Mol. Phylogenet. Evol.* **91**, 98–122 (2015).
10. J. Gatesy, M. S. Springer, *Mol. Phylogenet. Evol.* **80**, 231–266 (2014).
11. M. S. Bayzid, B. Boussau, S. Mirarab, T. Warnow, *PLOS ONE* 10.1371/journal.pone.0129183 (2015).
12. E. D. Jarvis *et al*., *Science* **346**, 1320–1331 (2014).
13. L. Liu, S. V. Edwards, *Science* **350**, 171 (2015).
14. T. Warnow, *PLOS Currents* 10.1371/currents.to1.8d41ac0f13d1abedf4c4a59f5d17b1f7 (2015).

ACKNOWLEDGMENTS

All simulated data, commands, and details about the simulation study are available at www.cs.utexas.edu/users/phylo/datasets/binning-response. Funding for this research was provided by the U.S. National Science Foundation (DBI-1461364) to T.W., by the Howard Hughes Medical Institute to S.M., and by the CNRS and Agence Nationale de la Recherche (ANR) through grant ANR-10-BINF-01-01 Ancestrome to B.B.

18 February 2015; accepted 28 July 2015
10.1126/science.aaa7719

Science
AAAS

Response to Comment on "Statistical binning enables an accurate coalescent-based estimation of the avian tree"

Siavash Mirarab *et al.*

Science **350**, 171 (2015);
DOI: 10.1126/science.aaa7719