"Correcting" Gene Trees to be More Like Species Trees Frequently Increases Topological Error

Zhi Yan *, Huw A. Ogilvie*, and Luay Nakhleh*

Department of Computer Science, Rice University, 6100 Main Street, Houston, TX 77005 USA

*Corresponding authors: E-mails: zhi.yan@rice.edu; huw.a.ogilvie@rice.edu; nakhleh@rice.edu.

Accepted: 21 May 2023

Abstract

The evolutionary histories of individual loci in a genome can be estimated independently, but this approach is error-prone due to the limited amount of sequence data available for each gene, which has led to the development of a diverse array of gene tree error correction methods which reduce the distance to the species tree. We investigate the performance of two representatives of these methods: TRACTION and TreeFix. We found that gene tree error correction frequently increases the level of error in gene tree topologies by "correcting" them to be closer to the species tree, even when the true gene and species trees are discordant. We confirm that full Bayesian inference of the gene trees under the multispecies coalescent model is more accurate than independent inference. Future gene tree correction approaches and methods should incorporate an adequately realistic model of evolution instead of relying on oversimplified heuristics.

Key words: gene tree error correction, gene tree inference, incomplete lineage sorting, anomaly zone, multispecies coalescent.

Significance

Gene tree information is essential for elucidating gene, genome, species, and phenotypic evolution, and a wide array of phylogenetic methods have been developed for gene tree estimation. Given that gene tree estimates are often inaccurate, several methods for "correcting" gene tree estimates have been devised. Here, we show that correction methods that are not based on an explicit statistical model of evolution such as the coalescent could produce poor results. To infer more accurate gene trees, one could use existing Bayesian methods that jointly estimate species and gene evolutionary histories, although additional work is needed to improve the scalability of that approach.

Introduction

Although genes evolve within the context of species, the evolutionary history of genes and gene families are unique and different from the species phylogeny because of processes such as incomplete lineage sorting (ILS), gene duplication and loss (GDL), and horizontal gene transfer (HGT) (Maddison 1997). Deciphering these individual histories of particular gene families is of great interest; to pick a few discoveries enabled by inferred gene trees, they have revealed effector and resistance genes in plant–pathogen interactions (Yang et al. 2013; McDonald et al. 2016), supported the importance of visual system changes to the adaptive radiation of cichlids (Torres-Dowdall et al. 2015) and identified orthologs of genes linked to human health and disease in model organisms (Maxwell et al. 2014; Waaijers et al. 2015).

Now that sequencing and assembly of eukaryotic genomes is relatively routine (Michael and VanBuren 2020; Rhie et al. 2021), in aggregate an enormous amount of data is available for phylogenetic analyses. Using the megabases or gigabases (Oliver et al. 2007) available in each genome, precise and accurate species histories can be inferred (Hahn and Nakhleh 2016). However, to infer the history of

individual gene families, the amount of information is much more limited with an average eukaryotic coding sequence length of roughly 1.3 kilobases (Xu et al. 2006). Fortunately, this limited sequence data can be augmented by information from the species phylogeny. When genes evolve following the multispecies coalescent (MSC) model, joint inference of the species and gene trees is substantially more accurate than inferring gene phylogenies independently (Szöllősi et al. 2014).

While joint inference methods are available for MSC (e.g., StarBEAST2; Ogilvie et al. 2017) or duplication and loss models (e.g., PHYLDOG; Boussau et al. 2013) of gene evolution, such methods are computationally intensive (Ogilvie et al. 2016). This has spurred the development of gene tree error correction tools intended to deal with GDL and HGT (Durand et al. 2005; Rasmussen and Kellis 2010; David and Alm 2011; Nguyen et al. 2012; Sjöstrand et al. 2012, 2014; Wu et al. 2013; Schreiber et al. 2014; Jacox et al. 2016; Noutahi et al. 2016; Lai et al. 2017; Bansal et al. 2018; Morel et al. 2020), which were developed to improve independently inferred gene trees through reconciliation with a given species tree. These approaches are more scalable and trivially parallelizable.

As has been appreciated for decades, ancestral polymorphism can persist through speciation events, leading to ILS which is one of the major sources of gene tree heterogeneity (Suh et al. 2015; Wang et al. 2018; Alda et al. 2019). It is even possible that there are regions where the most probable gene tree topology differs from the species tree (Degnan and Rosenberg 2006). Failure to account for common outcomes of evolutionary processes, like ILS as an outcome of population genetics, is likely to yield misinterpretations of evolutionary history. Note that although there are existing gene tree error correction approaches allowing for ILS, they are either parsimony-based or nonparametric, not incorporating the coalescent process probabilistically (Stolzer et al. 2012; Christensen et al. 2019, 2020).

In this study, we picked TreeFix (Wu et al. 2013) and TRACTION (Christensen et al. 2019, 2020) as two representative methods for species tree attraction-based methods of gene tree error correction, hereafter species tree attraction methods. The former is a popular method utilizing the information from the species tree and sequence data based on a GDL model. The latter is a very recent nonparametric method that improves the uncertain branches by solving the RF-Optimal Tree Refinement problem, which resolves polytomies in an input tree t such that the refined tree t has the minimum Robinson–Foulds distance (RF) (Robinson and Foulds 1981) to a given binary tree T, and it has been shown to be accurate when applied to simulated data with ILS. However, in our results, TRACTION actually worsened the accuracy of gene trees under higher but realistic levels of ILS, and TreeFix did the same when mutation rates were faster. We suggest this is due to an approach to error correction that reduces distance between gene and species trees based on heuristics which in effect removes outlier nodes, compared with statistical models that are able to impute gene tree times and topology appropriately where there is a lack of data in the corresponding sequence alignment.

Results

We defined and quantified gene tree estimation error (GTEE) as the unrooted normalized RF distance between inferred gene trees—either inferred using the joint inference method StarBEAST2, or independently inferred using the maximum likelihood (ML) software IQ-TREE and the Bayesian software MrBayes—and the true simulated gene trees. The performance of gene tree inference methods, along with the ability of species tree attraction methods to correct independently inferred gene trees, was studied under four levels of informativeness (signal) in the multiple sequence alignments, which were effected by varying the number of sites and the population mutation rate $\theta = 4N\mu$, where N is the diploid population size and μ is the number of mutations per site per generation. Equal numbers of replicates were simulated under low, medium, and high levels of ILS by varying the scale of the species tree in coalescent units. The properties and performance, averaged across all three levels of ILS, of the simulated data and species tree attraction methods were calculated for each level of informativeness (table 1). We also characterized the effect of the interaction between informativeness and ILS on GTEE for all approaches we tested for gene tree estimation and error correction (fig. 1).

Gene tree correction methods aim to reduce GTEE by modifying the gene trees based on information in the species tree. We judged their efficacy by whether they changed the GTEE distribution from the uncorrected IQ-TREE distribution; if they shift the bulk of that distribution towards lower RF distances, they are improving the accuracy of the gene trees; conversely, if they shift it towards higher RF distances, they are reducing the accuracy of the gene trees. Our analysis shows that TRACTION actually *increased* GTEE under our simulation settings except when the signal is high and ILS is low or moderate (fig. 2). TreeFix could also reduce the accuracy of gene trees in cases where the signal in the sequence data is high ($\theta = 0.01$ with 800 sites). TreeFix performed better than TRACTION, and when the signal is low, it did improve the ML tree inferred by IQ-TREE. Still, gene trees inferred under MSC using StarBEAST2 were substantially more accurate (fig. 2). We observed similar results on 5-taxon anomaly-zone data, except that TreeFix had the worst performance; see supplementary fig. S3, Supplementary Material online. To make the latter case a fair comparison with TRACTION and TreeFix, the StarBEAST2 species topology was fixed to be the same as the true simulated topology, and the true topology was also used as input for TRACTION and TreeFix.

Table 1.
Levels and Trends of Gene Tree Estimation Error

Population mutation rate θ	0.001			0.01		
No. of sites	**200**	**800**	**2000**	**200**	**800**	**2000**
Avg. no. of parsimony-informative sites	1.57	6.31	15.9	16.8	67.6	168
Avg. GTEE[a]	0.794	0.559	0.367	0.497	0.249	0.135
[RF(TRACTION, ST) < RF(GT, ST)] %[b]	2.57%	11.7%	24%	37.9%	60.8%	55.3%
[RF(TRACTION, ST) < RF($\widehat{GT}$, ST)] %[c]	5.17%	20.1%	30.2%	67.4%	73.5%	59.4%
[RF(TreeFix, ST) < RF(GT, ST)] %[d]	92.7%	96.6%	95.7%	95.8%	93.4%	85.5%
[RF(TreeFix, ST) < RF($\widehat{GT}$, ST)] %[e]	99.6%	99.8%	99%	99.2%	96.9%	87.7%
[RF(TRACTION, GT) < RF($\widehat{GT}$, GT)] %[f]	0.485%	0.485%	3.4%	12.6%	18.4%	11.7%
[RF(TreeFix, GT) < RF($\widehat{GT}$, GT)] %[g]	80.6%	55.8%	26.2%	32.5%	14.1%	5.34%

NOTE.—For both TRACTION and TreeFix, the output topology was consistently either closer to the species tree than the uncorrected input topology, or equidistant from it.
[a]Gene tree (GT) estimation error (GTEE), measured by the normalized Robinson–Foulds (RF) distance between the true and IQ-TREE-inferred gene trees ($\widehat{GT}$).
[b]Percentage of TRACTION-corrected gene trees that are closer in distance than corresponding true gene trees to the species tree (ST).
[c]Percentage of TRACTION-corrected gene trees that are closer in distance than corresponding IQ-TREE-inferred gene trees to the species tree.
[d]Percentage of TreeFix-corrected gene trees that are closer in distance than corresponding true gene trees to the species tree.
[e]Percentage of TreeFix-corrected gene trees that are closer in distance than corresponding IQ-TREE-inferred gene trees to the species tree.
[f]Percentage of TRACTION-corrected gene trees that are closer in distance than corresponding IQ-TREE-inferred gene trees to the true gene trees.
[g]Percentage of TreeFix-corrected gene trees that are closer in distance than corresponding IQ-TREE-inferred gene trees to the true gene trees.

We next examined the degree to which the TRACTION and TreeFix corrected gene trees towards the reference species tree, either away from the uncorrected gene trees inferred from sequences simulated along true gene trees simulated under the MSC, or away from the true gene trees themselves. The corrected gene tree topologies were always either closer to the species tree than the originally inferred topology, or the same distance. TreeFix was particularly aggressive in altering gene tree topologies, with (when averaged across all conditions) over 96% of topologies altered to be closer to the species tree under all conditions we studied. Unfortunately, these methods were often "correcting" gene trees to be closer to the species tree than the true gene trees were, again particularly so in the case of TreeFix (table 1). Finally, we assessed the computational performance of each method in terms of running time (supplementary figs. S1 and S4, Supplementary Material online). On datasets with 100 loci, IQ-TREE always took less than 33 s to complete, whereas MrBayes typically required 0.4 to 2 h to finish. StarBEAST2 was the slowest method; it accumulated an effective sample size (ESS) at average rates of 22.9, 13.3, and 5.34 per hour for alignments of length 200, 800, and 2000 sites respectively. At those average rates, StarBEAST2 would require 8.7, 15, and 37.5 h to accumulate 200 ESS.

Discussion

The proliferation of gene tree error correction methods demonstrates the intense level of interest in the problem of improving gene tree accuracy without resorting to joint inference of species and gene phylogenies. However, we have demonstrated here that species tree attraction methods should be used with extreme caution when ILS causes the true gene tree histories to be highly discordant from the history of corresponding species. This has the potential to increase the estimation error in gene trees, which may have cascading effects on the accuracy and reliability of downstream analyses.

A previous investigation into gene tree error correction found that species tree attraction methods work well when uncorrected GTEE is high and gene tree discordance is much lower than in our analysis (Christensen et al. 2019, 2020). This is compatible with our finding that these methods essentially modify gene trees to be closer in distance to the species tree, since when ILS is relatively low the true gene trees will be more congruent with the species tree. So when GTEE is high simply reducing discordance by altering them to be more similar to the species tree will increase accuracy, as most of the inferred discordance will be random error rather than deriving from biological processes. However, as we have shown, when ILS is higher, species tree attraction methods will increase GTEE regardless of the level of error in the originally inferred gene trees, as genuine discordance is being removed. This is analogous to removing outlier measurements for a smoother fit.

We tested whether maximum clade credibility summary trees from Bayesian posterior distributions of individually inferred gene trees would help, but it was only minimally more accurate than simple ML inference due to the limited amount of information in each locus. The only approach we found that substantially decreased GTEE was full Bayesian inference under the MSC, as has previously been reported (Szöllősi et al. 2014). Given the poor scalability of full Bayesian MSC inference, both in terms of the increase in time required to finish analyses as the amount of loci or species is increased, and the difficulty in parallelizing those analyses, we believe it is still worth pursuing gene tree error correction methods.

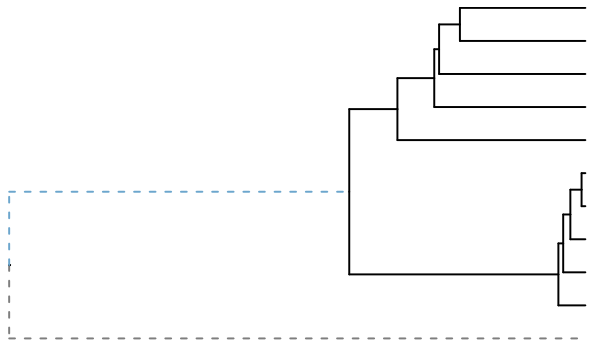

Fig. 1.—Model species tree. The 10-taxon ingroup (solid lines), generated under a birth–death process, was scaled to a crown height of 2, 5, and 10 coalescent units which corresponded to high, medium, and low ILS, respectively. To root gene trees an outgroup was added (bottom black dashed line), and the distance from the ingroup to the root node was fixed at 7 coalescent units in all analyses to avoid ILS past the root node without saturating the sequence alignments (top blue dashed line).

We suggest that gene tree correction should be carried out using model-based methods that are suitable for the system of interest. Future methods should only alter gene tree nodes when there is greater support for a different topology given a model of genealogical inheritance incorporating the species history, than there is support for the original topology in the sequence data given a substitution model. This is analogous to imputing missing data, rather than removing outliers. If improved methods for gene tree correction are developed which reliably reduce gene tree error under perfect conditions where the species tree is known, they should also be evaluated under imperfect conditions where the species tree is also estimated.

Materials and Methods

We simulated an 11-taxon tree as a model species history, which was scaled to three different levels (fig. 1). First, we drew one birth rate λ from a uniform distribution on [0.5, 1], then one death rate μ from [0, λ]. Second, we simulated a 10-taxon tree under a birth–death process with those birth- and death-rates using TreeSim (Stadler 2011) as the ingroup. Third, we rescaled the branch lengths of the resulting tree to obtain three species trees with root heights of 2, 5, and 10 in coalescent units, corresponding to the scenarios of low, medium, and high levels of ILS, respectively. Fourth, in order to accurately root the gene trees, we added an outgroup to the simulated ingroup such that the outgroup has 7 distance in coalescent units from the ingroup. We performed coalescent simulations to generate datasets with 100 gene trees, each with 10 replicates and a single individual per species.

For each gene tree, we employed Seq-Gen (Rambaut and Grassly 1997), and used two different population mutation rates ($\theta = 4N\mu$): 0.001 and 0.01, to simulate sequence data of length 200, 800, and 2000 nucleotides under the HKY model (Hasegawa et al. 1985) together with among-site rate variation. For each replicate, base frequencies were drawn from a flat Dirichlet distribution, the transition/transversion ratio κ was drawn from a log-normal distribution with a mean of 1 and a standard deviation of 1.25, and the shape α value of the four rate category discrete gamma model was drawn from a log-normal distribution with a mean of −1 and a standard deviation of 1.5. Scaling the gene trees in this manner only changes the substitution rate μ and not the population size component N of θ.

In total, $3 \times 10 \times 2 \times 2 = 120$ datasets were generated, each with 100 gene trees and corresponding multiple sequence alignments. We then utilized IQ-TREE version 2.1.3 (Minh et al. 2020) with 100 bootstrap replicates under the Jukes–Cantor model to reconstruct ML gene trees from the simulated alignments, and rooted them using the outgroup taxon. We also used the Bayesian method MrBayes version 3.2.7a (Ronquist et al. 2012) to estimate gene trees independently. MrBayes was configured to use a Jukes–Cantor substitution model, running 3 chains for each gene tree with 1 million iterations each, sampling every 1000 iterations, using the automatic stopping rule, and a 20% burn-in. Maximum clade credibility tree topologies were summarized from the posterior distributions (Heled and Bouckaert 2013).

We applied TreeFix v1.1.10 (Wu et al. 2013) and TRACTION v1.0 (Christensen et al. 2019, 2020) to correct gene trees inferred by IQ-TREE. TRACTION requires a threshold value for contracting a low support branch, so we adopted a support threshold of 75% as used in Christensen et al. (2019). Additionally, we ran StarBEAST2 (Ogilvie et al. 2017), a method for joint Bayesian inference of species and gene trees with fixed reference species tree, to sample posterior distributions of gene phylogenies from simulated sequences. We performed StarBEAST2 analyses under a strict clock model with a Yule prior on the species tree. For the population model, we assumed a single constant population size $N\mu$ for the entire tree, with a Gamma prior which had a shape of 1.5 and a scale of 0.003333. This is a broad prior with a mode of 0.0016665 corresponding to $\theta = 4N\mu = 0.0066666$, in between the true values of θ used for simulation. The site models were linked and configured to be the same as the model used for simulation (HKY with 4 gamma categories). The substitution rate variation shape, transition/transversion ratio κ, and base frequencies were all estimated. Posterior distributions of gene trees were also summarized using maximum clade credibility trees. We recorded the elapsed wall-clock time for each replicate analysis. All analyses were executed on a compute cluster (powered by AMD EPYC 7642 CPUs), using a single thread per replicate across all methods.

We also conducted an experiment on simulated 5-taxon data where the model species tree was in the anomaly zone. These datasets include 30 different model conditions,

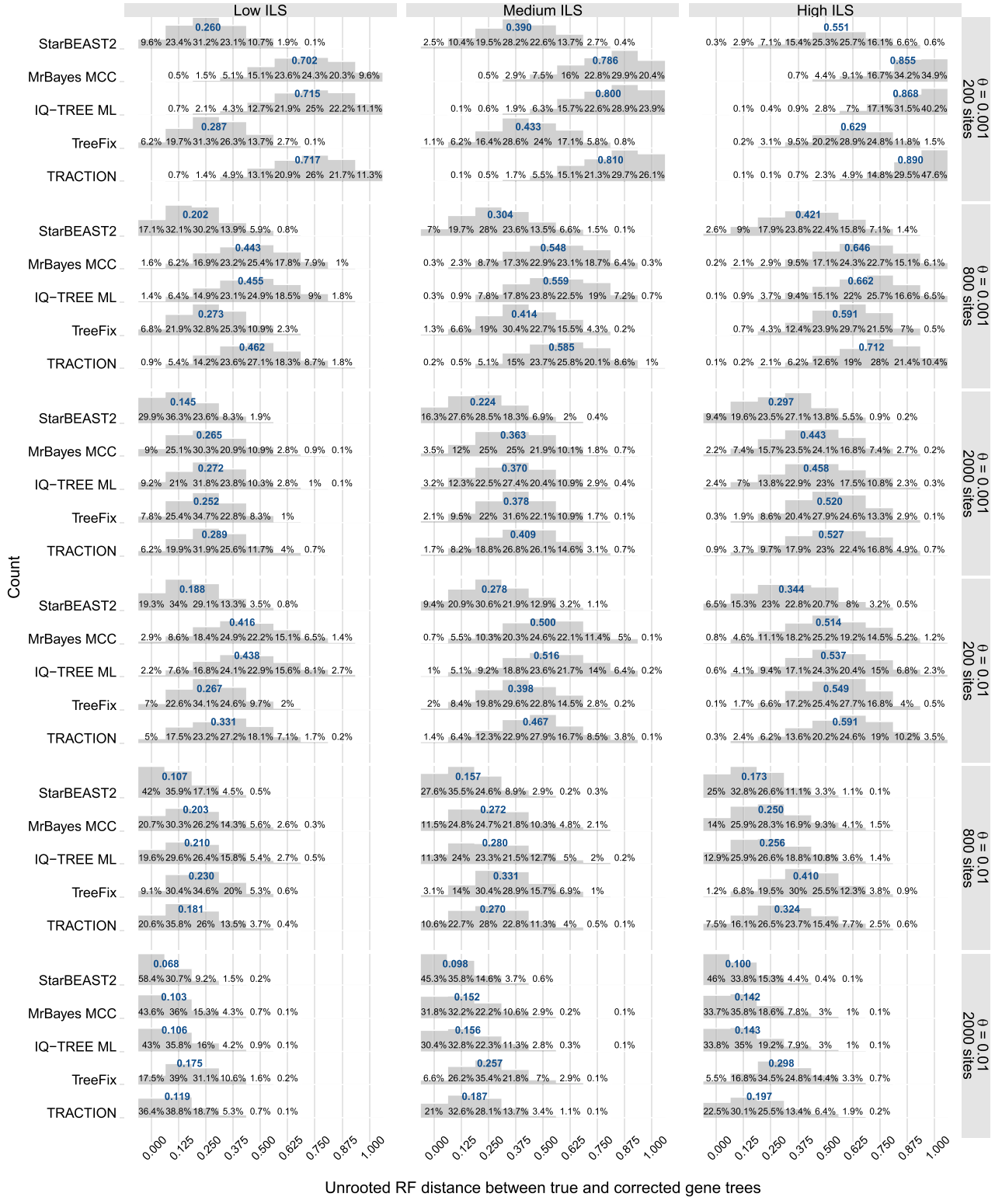

Fig. 2.—Error distributions of StarBEAST2-inferred, uncorrected, and corrected gene trees under the model conditions with 11 species and 100 genes. Uncorrected gene trees were either the maximum clade credibility topologies summarized from posterior tree samples from MrBayes or the ML topologies inferred using IQ-TREE. ML gene trees were corrected using either TRACTION or TreeFix. GTEE is defined as the normalized unrooted RF distance between the estimated uncorrected or corrected gene trees and the true simulated gene trees. Numbers in bold blue text are the mean of each distribution.

each with 10 replicates. The simulation setup for generating these datasets can be found in the supplementary Section S2, Supplementary Material online.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution online*.

Acknowledgments

This work was supported in part by NSF grants CCF-1514177, CCF-1800723, and DBI-2030604 to L.N.

Data Availability

The data underlying this article are available in the GitHub Repository, at https://github.com/Moerz/Gene-Tree-Fixing.

Literature Cited

Alda F, et al. 2019. Resolving deep nodes in an ancient radiation of neotropical fishes in the presence of conflicting signals from incomplete lineage sorting. Syst Biol. 68(4):573–593.

Bansal MS, Kellis M, Kordi M, Kundu S. 2018. RANGER-DTL 2.0: rigorous reconstruction of gene-family evolution by duplication, transfer and loss. Bioinformatics. 34(18):3214–3216.

Boussau B, et al. 2013. Genome-scale coestimation of species and gene trees. Genome Res. 23(2):323–330.

Christensen S, Molloy EK, Vachaspati P, Warnow T. 2019. TRACTION: fast non-parametric improvement of estimated gene trees. In: Huber KT, Gusfield D, editors. 19th International Workshop on Algorithms in Bioinformatics (WABI 2019). Volume 143 of Leibniz International Proceedings in Informatics (LIPIcs). Germany: Dagstuhl. p. 4:1–4:16. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Christensen S, Molloy EK, Vachaspati P, Yammanuru A, Warnow T. 2020. Non-parametric correction of estimated gene trees using TRACTION. Algorithms Mol Biol. 15(1):1–18.

David LA, Alm EJ. 2011. Rapid evolutionary innovation during an Archaean genetic expansion. Nature 469(7328):93–96.

Degnan JH, Rosenberg NA. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2(5):1–7.

Durand D, Halldórsson BV, Vernot B. 2005. A hybrid micro-macroevolutionary approach to gene tree reconstruction. In: Annual International Conference on Research in Computational Molecular Biology. Berlin, Heidelberg: Springer. p. 250–264.

Hahn MW, Nakhleh L. 2016. Irrational exuberance for resolved species trees. Evolution 70(1):7–17.

Hasegawa M, Kishino H, Yano T-A. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 22(2):160–174.

Heled J, Bouckaert RR. 2013. Looking for trees in the forest: summary tree from posterior samples. BMC Evol Biol. 13(1):221.

Jacox E, Chauve C, Szöllősi GJ, Ponty Y, Scornavacca C. 2016. ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony. Bioinformatics 32(13):2056–2058.

Lai H, Stolzer M, Durand D. 2017. Fast heuristics for resolving weakly supported branches using duplication, transfers, and losses. In: Meidanis J, Nakhleh L, editors. Comparative genomics. Cham: Springer International Publishing. p. 298–320.

Maddison WP. 1997. Gene trees in species trees. Syst Biol. 46(3):523–536.

Maxwell EK, et al. 2014. Evolutionary profiling reveals the heterogeneous origins of classes of human disease genes: implications for modeling disease genetics in animals. BMC Evol Biol. 14(1):212.

McDonald MC, et al. 2016. Utilizing gene tree variation to identify candidate effector genes in *Zymoseptoria tritici*. G3 6(4):779–791.

Michael TP, VanBuren R. 2020. Building near-complete plant genomes. Curr Opin Plant Biol. 54:26–33. Genome studies and molecular genetics.

Minh BQ, et al. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 37(5):1530–1534.

Morel B, Kozlov AM, Stamatakis A, Szöllősi GJ. 2020. GeneRax: a tool for species-tree-aware maximum likelihood-based gene family tree inference under gene duplication, transfer, and loss. Mol Biol Evol. 37(9):2763–2774.

Nguyen TH, et al. 2012. Accounting for gene tree uncertainties improves gene trees and reconciliation inference. In: International Workshop on Algorithms in Bioinformatics. Berlin, Heidelberg: Springer. p. 123–134.

Noutahi E, et al. 2016. Efficient gene tree correction guided by genome evolution. PLoS ONE. 11(8):1–22.

Ogilvie HA, Bouckaert RR, Drummond AJ. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Mol Biol Evol. 34(8):2101–2114.

Ogilvie HA, Heled J, Xie D, Drummond AJ. 2016. Computational performance and statistical accuracy of *BEAST and comparisons with other methods. Syst Biol. 65(3):381–396.

Oliver MJ, Petrov D, Ackerly D, Falkowski P, Schofield OM. 2007. The mode and tempo of genome size evolution in eukaryotes. Genome Res. 17(5):594–601.

Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Bioinformatics 13(3):235–238.

Rasmussen MD, Kellis M. 2010. A Bayesian approach for fast and accurate gene tree reconstruction. Mol Biol Evol. 28(1):273–290.

Rhie A, et al. 2021. Towards complete and error-free genome assemblies of all vertebrate species. Nature 592(7856):737–746.

Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. Math Biosci. 53(1–2):131–147.

Ronquist F, et al. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 61(3):539–542.

Schreiber F, Patricio M, Muffato M, Pignatelli M, Bateman A. 2014. TreeFam v9: a new website, more species and orthology-on-the-fly. Nucleic Acids Res. 42(D1):D922–D925.

Sjöstrand J, et al. 2014. A Bayesian method for analyzing lateral gene transfer. Syst Biol. 63(3):409–420.

Sjöstrand J, Sennblad B, Arvestad L, Lagergren J. 2012. DLRS: gene tree evolution in light of a species tree. Bioinformatics 28(22):2994–2995.

Stadler T. 2011. Simulating trees with a fixed number of extant species. Syst Biol. 60(5):676–684.

Stolzer M, et al. 2012. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. Bioinformatics 28(18):i409–i415.

Suh A, Smeds L, Ellegren H. 2015. The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. PLoS Biol. 13(8):e1002224.

Szöllősi GJ, Tannier E, Daubin V, Boussau B. 2014. The inference of gene trees with species trees. Syst Biol. 64(1):e42–e62.

Torres-Dowdall J, Henning F, Elmer KR, Meyer A. 2015. Ecological and lineage-specific factors drive the molecular evolution of rhodopsin in cichlid fishes. Mol Biol Evol. 32(11):2876–2882.

Waaijers S, Ramalho JJ, Koorman T, Kruse E, Boxem M. 2015. The *C. elegans* Crumbs family contains a CRB3 homolog and is not essential for viability. Biol Open. 4(3):276–284.
Wang K, et al. 2018. Incomplete lineage sorting rather than hybridization explains the inconsistent phylogeny of the wisent. Commun Biol. 1(1):1–9.
Wu Y-C, Rasmussen MD, Bansal MS, Kellis M. 2013. TreeFix: statistically informed gene tree error correction using species trees. Syst Biol. 62(1):110–120.
Xu L, et al. 2006. Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms. Mol Biol Evol. 23(6):1107–1108.
Yang S, et al. 2013. Rapidly evolving *R* genes in diverse grass species confer resistance to rice blast disease. Proc Natl Acad Sci USA. 110(46):18572–18577.

Associate editor: Barbara Holland