Rehabilitating the benefits of gene tree correction in the presence of incomplete lineage sorting

Abstract

Gene trees play an important role in various areas of phylogenomics. However, their reconstruction often relies on limited-length sequences and may not account for complex evolutionary events, such as gene duplications, losses, or incomplete lineage sorting (ILS), which are not modeled by standard phylogenetic methods. To address these challenges, it is common to first infer gene trees using conventional models, then refine them through species tree-aware correction methods. Recently, it has been argued that such corrections can lead to overfitting and force gene trees to resemble the species tree, thereby obscuring genuine gene-level variation caused by ILS. In this paper, we challenge and refute this hypothesis, and we demonstrate that, when applied carefully, correction methods can offer significant benefits, even in the presence of ILS.

1 Introduction

Gene trees are fundamental components in several downstream phylogenomics analyses. Deciphering the evolutionary histories of gene families provides insights into how genes diversify and adapt, and has applications ranging from the prediction of genes with similar functions, to enhancing our understanding of the processes of gain and loss and of the mechanisms that can lead to the expansion of gene families. Naturally, the accuracy of such biological inferences depends heavily on the accuracy of the gene trees themselves.

This contrasts with the fact that gene trees are difficult to infer with precision. Gene sequences are typically of limited length and thus convey only limited evolutionary signals [22, 19]. Furthermore, genes undergo evolutionary events that are not incorporated into typical phylogenetic reconstruction models. In particular, gene duplication events followed by divergent evolutionary fates can lead to heterogeneous mutation rates across the tree, with gene losses complicating the picture [16, 5, 8].

These challenges can be overcome by first reconstructing gene trees using standard models, then correcting them while accounting for such events. A common strategy is to collapse low-support branches of a gene tree and resolve the resulting polytomies using information from the species tree. For instance, the tool TRACTION finds a refinement of the gene tree with minimum Robinson-Foulds distance to the species

tree [2], whereas ecceTERA and profileNJ find the refinement with a minimum duplication and loss cost [6, 14] (see also [25, 26]). TreeFix uses a different approach and explores the space of trees around the input one, to find one with a smaller duplication and loss cost but a comparable likelihood [21]. Such correction methods were shown useful in joint inference of gene trees and species trees [1], detecting whole genome duplications [4], or predicting orthology and paralogy relationships [9], for example.

These correction algorithms only consider duplications, losses, and horizontal gene transfers (HGT, which exchange genes between co-existing species) even though incomplete lineage sorting (ILS) is also known to hinder the reconstruction process. Under ILS, alleles may persist through speciation events, producing gene trees that diverge from the species tree even in the absence of duplications or losses [3, 11, 24]. While there are methods that take ILS into account [10, 15], they are often computationally intensive and impractical at the genome scale, where thousands of gene families are analyzed. Instead, we often rely on fast but less accurate algorithms, possibly applying an efficient correction procedure on the resulting reconstructions.

In Yan et al. [23], the authors make a case against gene tree correction in the presence of ILS. One of the main arguments against current methods for correction is that ILS results in gene and species tree discordance, and thus that correcting the gene tree without accounting for ILS will make it resemble the species tree and inevitably produce errors. To support this, they perform simulations to compare TreeFix and TRACTION against the uncorrected predictions from IQ-TREE [13], and their results demonstrate that the two correction approaches often increase gene tree estimation error.

First, we note here that the tool ecceTERA was not evaluated in Yan et al.'s study, despite showing very good performances for non-ILS data sets [7]. Also, TreeFix, unlike several other correction algorithms, had already been shown to over-fit the data [18], so we do not consider it as representative method. Moreover, a single bootstrap support threshold of 75% was evaluated, which is arguably high and can lead to heavy modifications of the gene trees.

Thus we decided to re-evaluate the dataset from [23] and demonstrate that gene tree correction, when applied judiciously, does improve accuracy even in the presence of ILS.

Results

We first summarize the previous work and then proceed with the results of our reanalysis.

Previous simulations, runs and findings

In Yan et al. [23], a species tree S with 11 taxa (including one outgroup) was simulated using TreeSim [20]. The branch lengths were rescaled to make the root of the ingroup of height of 2, 5, and 10 in coalescent units, respectively, corresponding to low, medium, and high ILS rates. Then, coalescent simulations were performed

along S to generate datasets composed of 100 gene trees (each with 10 replicates and a single individual per species, with no duplications nor losses). Seq-Gen [17] was used on each gene tree to simulate a multiple sequence alignment. The authors tested two mutation rates $\theta \in \{0.001, 0.01\}$ and three possible number of sites in $\{200, 800, 2000\}$. IQ-TREE was used to reconstruct the gene trees with 100 bootstrap replicates. This produced 120 datasets of 100 gene trees, and each gene tree was corrected with TreeFix and TRACTION. For the latter, branches with bootstrap support below 75% were contracted. As a measure of gene tree estimation error, the uncorrected and corrected gene trees were compared to the true simulated tree using the Robinson-Foulds distance.

To summarize the findings of [23] on these datasets, TreeFix generally improves accuracy on short sequences of length 200, has mixed results on sequences of length 800, and has decreased accuracy on longer sequences of length 2000, especially on higher mutation rates $\theta = 0.01$. As for TRACTION, in most cases its accuracy is close but slightly worse than the uncorrected gene trees, with a more pronounced decrease under high levels of ILS. See [23, Figure 2] for detailed results.

Our runs and findings

For rehabilitating the benefits of gene tree correction, we choose to rerun tests using ecceTERA, a program that implements a fast parsimony reconciliation algorithm accounting for gene duplication, gene loss and HGTs [6].

We corrected gene trees via ecceTERA under two different settings: the first considers only duplications and losses as the macro-evolutionary events influencing gene histories, while the second additionally incorporates gene transfer events. All other ecceTERA parameters are set to their default values. For both settings, we run ecceTERA with three different support thresholds: 50, 70 and 90. Branches with a bootstrap support lower than the given threshold are collapsed by ecceTERA, and rearranged by finding a refinement with a minimum duplication and loss cost, or duplication, loss and transfer cost [7].

We then computed the Robinson-Foulds distance between the simulated gene trees and the ones reconstructed with IQ-TREE, and between the simulated gene trees and the IQ-TREE trees corrected by ecceTERA. Figure 1 shows the result when correcting using duplications and losses, while Figure 2 also considers HGTs.

Looking at Figure 1, the initial observation to highlight is that implementing corrections generally proves beneficial across all parameters examined. Under every single setting, the box plots using threshold 50 exhibit a lower or equal estimation error box than the uncorrected version, being strictly better in most cases. This becomes particularly noticeable for the combination of a mutation rate of 0.001 and 200 sites, where reconstruction errors tend to occur more frequently.

Additionally, aside from the scenario with a mutation rate of 0.001, 200 sites, and high levels of ILS, using a threshold of 50 appears to be more beneficial than the other two thresholds. Furthermore, the advantages of correction become even more pronounced when ILS is high compared to when it is low. This observation stands in sharp contrast to the hypothesis proposed in [23].

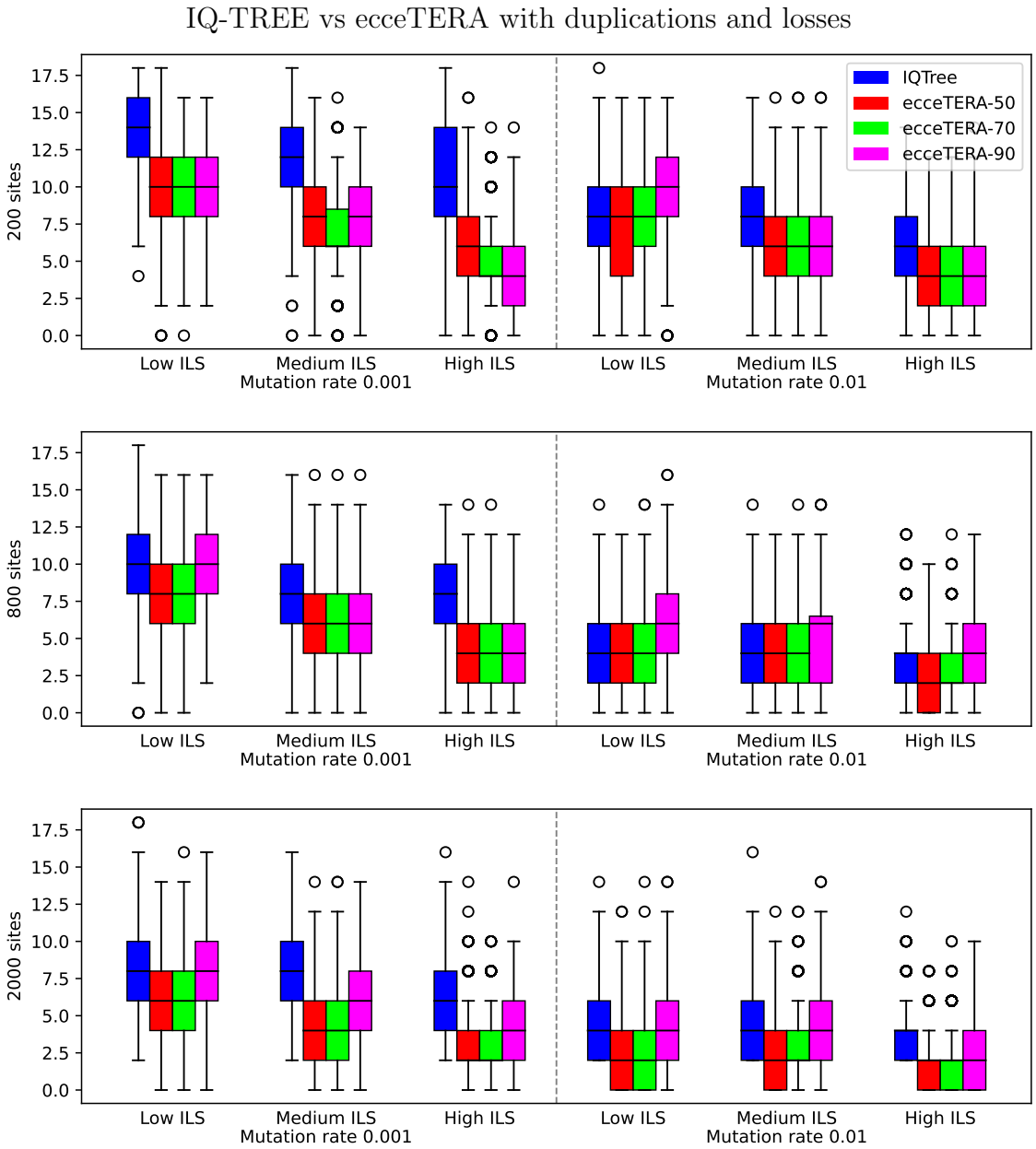

Figure 1: Gene tree error estimation (via the Robinson-Foulds distance) of IQ-TREE and ecceTERA under thresholds 50, 70, and 90, while considering horizontal gene transfer under the DL model.

In Figure 2, we can see that correcting taking into account gene transfer is still beneficial, even if less pronouncedly, when the mutation rate is low (0.001). When the mutation is higher (0.01), the correction starts becoming deleterious for low and medium ILS, and 200 and 800 sites. For those cases, additionally considering gene transfer as possible macro-evolutionary events seems to add too much liberty to the correction of gene topologies, leading to overfitting.

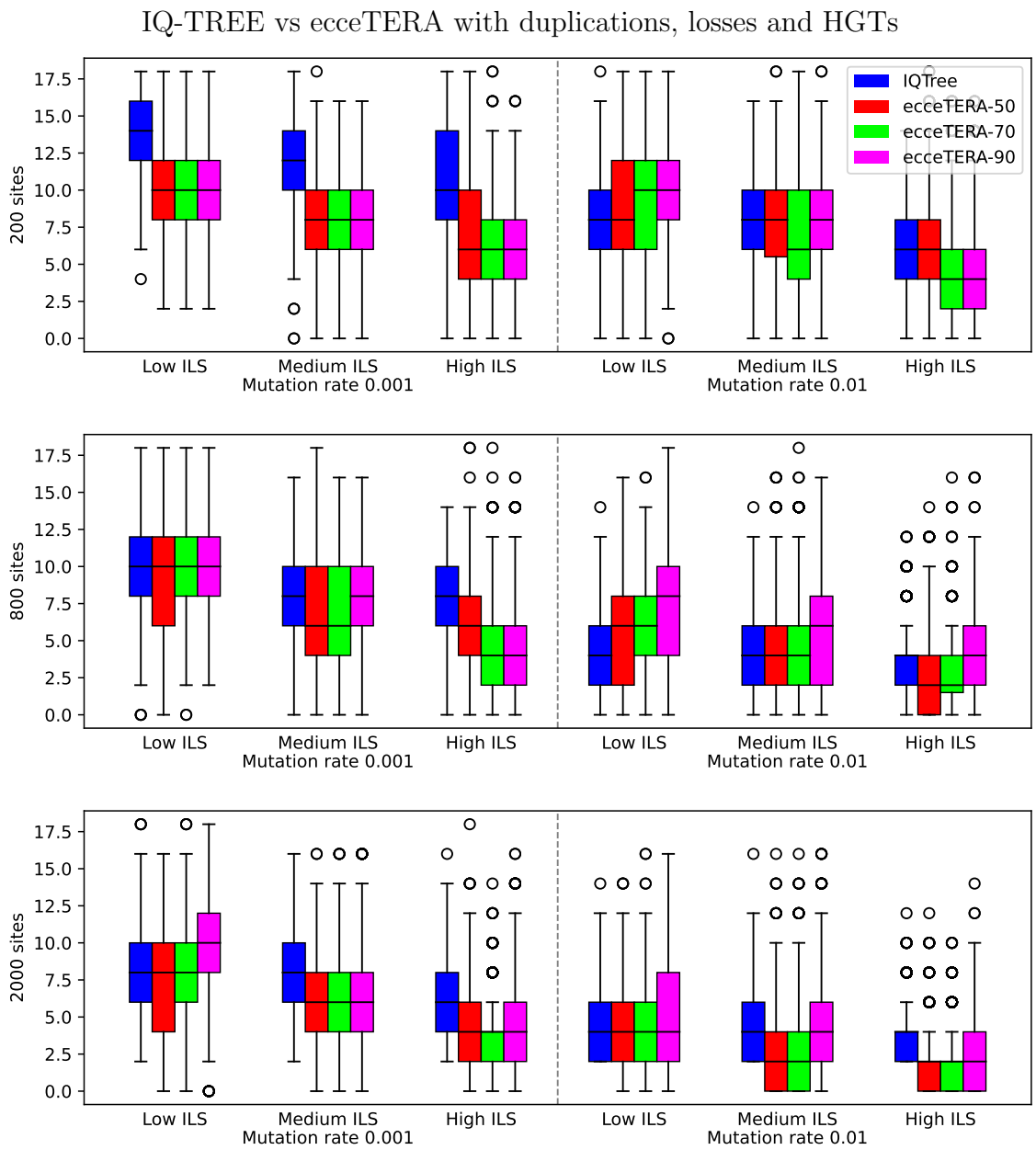

Figure 2: Gene tree error estimation (via the Robinson-Foulds distance) of IQ-TREE and ecceTERA under thresholds 50, 70, and 90 under the DTL model.

Discussion

The results shown in the previous section contrast with the findings of Yan et al. We believe that this is due to several key factors.

First, while it is true that branches of gene trees discordant with the species tree may result from ILS and should not always be corrected, such regions are often well-supported by bootstrap analyses. A conservative threshold thus preserves these branches. On the other hand, poorly supported branches may reflect errors due to

duplication and loss, which methods such as ecceTERA are specifically designed to correct. The analysis in [23] was limited to a support threshold of 75%, which probably explains how this was overlooked. Second, the methods chosen by the authors may not be the most appropriate for gene tree correction. TRACTION optimizes the Robinson-Foulds distance, which is not directly related to evolutionary events that affect genes (although it may build accurate species trees from gene trees). TreeFix, while more sophisticated, explores a vast tree space and may overfit gene trees to the species tree, as shown in [18]. In contrast, methods such as ecceTERA address both concerns by explicitly modeling duplications and losses and only modifying poorly supported regions.

Let us also note here that the dataset that we re-analyzed contains gene trees that evolve exclusively under ILS –without duplications, HGTs, or losses. While this is sufficient to rehabilitate correction methods when ILS is present, future experiments should focus on gene trees that also include these events in addition to ILS. In this case, the gene trees are probably much more difficult to reconstruct using sequence-based methods, and we expect correction methods to provide even more meaningful improvements. Future experiments should also question the impact of ILS on gene and species tree co-estimation methods such as AleRax [12].

While we agree that caution is required when correcting gene trees in the presence of ILS, our results show that applying correction methods appropriately has clear benefits. Rather than discarding these approaches, they should be used with care to improve the reliability of phylogenomic inferences, even when ILS is present in the data.

References

[1] Bastien Boussau, Gergely J Szöllősi, Laurent Duret, Manolo Gouy, Eric Tannier, and Vincent Daubin. Genome-scale coestimation of species and gene trees. *Genome research*, 23(2):323–330, 2013.

[2] Sarah Christensen, Erin K Molloy, Pranjal Vachaspati, Ananya Yammanuru, and Tandy Warnow. Non-parametric correction of estimated gene trees using traction. *Algorithms for Molecular Biology*, 15:1–18, 2020.

[3] James H Degnan and Noah A Rosenberg. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in ecology & evolution*, 24(6):332–340, 2009.

[4] Riccardo Dondi, Manuel Lafond, and Celine Scornavacca. Reconciling multiple genes trees via segmental duplications and losses. *Algorithms for Molecular Biology*, 14:1–19, 2019.

[5] Jean-Francois Gout, Yue Hao, Parul Johri, Olivier Arnaiz, Thomas G Doak, Simran Bhullar, Arnaud Couloux, Fréderic Guérin, Sophie Malinsky, Alexey

Potekhin, et al. Dynamics of gene loss following ancient whole-genome duplication in the cryptic paramecium complex. *Molecular biology and evolution*, 40(5):msad107, 2023.

[6] Edwin Jacox, Cedric Chauve, Gergely J Szöllősi, Yann Ponty, and Celine Scornavacca. eccetera: comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics*, 32(13):2056–2058, 2016.

[7] Edwin Jacox, Mathias Weller, Eric Tannier, and Celine Scornavacca. Resolution and reconciliation of non-binary gene trees with transfers, duplications and losses. *Bioinformatics*, 33(7):980–987, 2017.

[8] Reza Kalhor, Guillaume Beslon, Manuel Lafond, and Celine Scornavacca. A rigorous framework to classify the postduplication fate of paralogous genes. *Journal of Computational Biology*, 31(9):815–833, 2024.

[9] Manuel Lafond, Krister M Swenson, and Nadia El-Mabrouk. Error detection and correction of gene trees. In *Models and algorithms for genome evolution*, pages 261–285. Springer, 2013.

[10] Wayne P Maddison and L Lacey Knowles. Inferring phylogeny despite incomplete lineage sorting. *Systematic biology*, 55(1):21–30, 2006.

[11] Chen Meng and Laura Salter Kubatko. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theoretical population biology*, 75(1):35–45, 2009.

[12] Benoit Morel, Tom A Williams, Alexandros Stamatakis, and Gergely J Szöllősi. Alerax: a tool for gene and species tree co-estimation and reconciliation under a probabilistic model of gene duplication, transfer, and loss. *Bioinformatics*, 40(4):btae162, 03 2024.

[13] Lam-Tung Nguyen, Heiko A Schmidt, Arndt Von Haeseler, and Bui Quang Minh. Iq-tree: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, 32(1):268–274, 2015.

[14] Emmanuel Noutahi, Magali Semeria, Manuel Lafond, Jonathan Seguin, Bastien Boussau, Laurent Guéguen, Nadia El-Mabrouk, and Eric Tannier. Efficient gene tree correction guided by genome evolution. *PLoS One*, 11(8):e0159559, 2016.

[15] Huw A Ogilvie, Remco R Bouckaert, and Alexei J Drummond. Starbeast2 brings faster species tree inference and accurate estimates of substitution rates. *Molecular biology and evolution*, 34(8):2101–2114, 2017.

[16] Susumu Ohno. *Evolution by gene duplication*. Springer Science & Business Media, 2013.

[17] Andrew Rambaut and Nicholas C Grass. Seq-gen: an application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3):235–238, 1997.

[18] Celine Scornavacca, Edwin Jacox, and Gergely J Szöllősi. Joint amalgamation of most parsimonious reconciled gene trees. *Bioinformatics*, 31(6):841–848, 2015.

[19] Paul Simion, Frédéric Delsuc, and Herve Philippe. To what extent current limits of phylogenomics can be overcome? *Phylogenetics in the genomic era*, pages 2–1, 2020.

[20] Tanja Stadler. Simulating trees with a fixed number of extant species. *Systematic biology*, 60(5):676–684, 2011.

[21] Yi-Chieh Wu, Matthew D Rasmussen, Mukul S Bansal, and Manolis Kellis. Treefix: statistically informed gene tree error correction using species trees. *Systematic biology*, 62(1):110–120, 2013.

[22] Lin Xu, Hong Chen, Xiaohua Hu, Rongmei Zhang, Ze Zhang, and ZW Luo. Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms. *Molecular biology and evolution*, 23(6):1107–1108, 2006.

[23] Zhi Yan, Huw A Ogilvie, and Luay Nakhleh. "correcting" gene trees to be more like species trees frequently increases topological error. *Genome Biology and Evolution*, 15(6):evad094, 2023.

[24] Louxin Zhang. From gene trees to species trees ii: Species tree inference by minimizing deep coalescence events. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(6):1685–1691, 2011.

[25] Yu Zheng, Taoyang Wu, and Louxin Zhang. A linear-time algorithm for reconciliation of non-binary gene tree and binary species tree. In *Combinatorial Optimization and Applications: 7th International Conference, COCOA 2013, Chengdu, China, December 12-14, 2013, Proceedings*, pages 190–201. Springer, 2013.

[26] Yu Zheng and Louxin Zhang. Reconciliation with nonbinary gene trees revisited. *Journal of the ACM (JACM)*, 64(4):1–28, 2017.