Utilizing SVDquartets in Improving ASTRAL on High Gene Tree Estimation Error Model Conditions

Akhil Jakatdar*[1]

[1]Department of Computer Science, University of Illinois at Urbana-Champaign

Abstract

This paper proposes a new species tree estimation, Utilizing SVDquartets on ASTRAL (USA), that builds off of ASTRAL-III in attempting to improve on the ASTRAL family of methods on data with high instances of gene tree estimation error. USA builds off of ASTRAL-III in collapsing low support branch edges from this base method, and running SVDquartets on the generated polytomies. This paper compares USA utilizing different branch support threshold cut-off methods with both ASTRAL-III and SVDquartets in order to analyze its performance across different threshold values and in comparison with these established estimation methods. The results of this paper indicate that hard threshold cut-offs for branch support values are not reliable in improving performance on high gene tree estimation error datasets and that more context must be taken into account to generate signficant improvements.

1 Introduction

In the pursuit of estimating the phylogeny of different species, genomic discordance and estimation errors are two key obstacles in accurate reconstructing the species phylogeny. Genomic discordance, and more specifically gene tree discordance, define discrepancies in the topology of gene trees constructed from loci across the genome when compared to the true species tree topology, with Incomplete Lineage Sorting (ILS) being an example of this type of discordance. As the amount of genomic data available has increased in recent years, the problem of input gene tree estimation error (GTEE) has become a further factor to contend with in accurately reconstructing phylogeny data from gene trees. Many summary methods have been developed in the species tree estimation space to try and solve some of these problems, one being ASTRAL. ASTRAL [1] is one of the most well-known species tree estimation method, that infers said species trees from gene trees in a statistically consistent manner and accounting for gene tree discordance. ASTRAL aims to estimate species trees with the input set of gene trees by solving the Maximum Quartet Support Species Tree problem and has shown great results in the presence of most gene tree discordance.

However, recent studies in Molloy & Warnow [2] have pointed to a potential weakness in ASTRAL in comparison to other site-based coalescent methods on datasets that contain a high degree of GTEE (greater than 80%) as well as high incidents of ILS (greater than 75%). One of the site-based coalescent methods that show much greater levels of accuracy on these high GTEE and

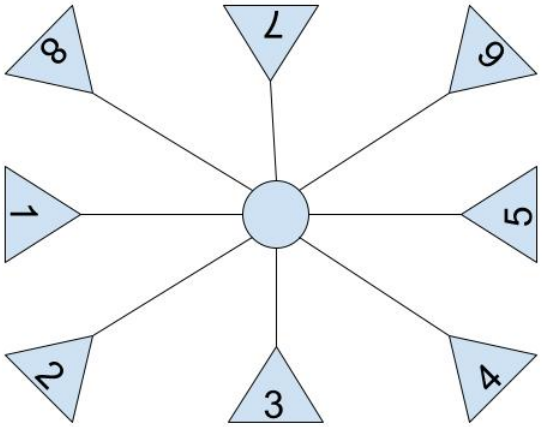
*akhilrj2@illinois.edu

high ILS model conditions is SVDquartets [3]. SVDquartets in PAUP* utilizes site-based pattern distribution in its estimation algorithm and utilizes a multitude of quartet amalgamation methods, Quartet MaxCut being the other used in this paper [4], in combining its set of inferred quartets. Thus the motivation to improve ASTRAL through utilizing SVDquartets in the aforementioned model conditions, specifically the high GTEE conditions, is explored through the contents of this paper.

2 Methods & Materials

This paper proposes a new method, Utilizing SVDquartets on ASTRAL (USA), that aims to improve ASTRAL-III [5] on the aforementioned datasets that contain high Gene Tree Estimation Error. The proposed method USA uses ASTRAL-III as a base method and build off of its initial tree estimation output. We can describe the input to USA to be a set of input gene trees as well as the alignments for all taxa contained in this gene trees and the output being an estimated unrooted resolved species trees. The decision to use ASTRAL-III of all the ASTRAL family species tree estimation methods stems from its two main improvements in comparison to ASTRAL-II. The first improvement involves constraining the bipartition set X in order to grow linearly with the number of species and input genes. Along with this runtime improvement, the second and more consequential improvement to our proposed method is shown through the fact that ASTRAL-III has a comparatively better ability to handle polytomies using techniques to constrain the optimal tree space and use similarities between gene trees [6]. With the improved runtime, ASTRAL-III has the feature of removing low-support branches in order to improve accuracy on estimation which is important for the second step of USA.

USA takes this initial tree estimation output and collapses all branches that contain a branch support value less than some threshold t and uses the Newick utilities [7] to assist in the collapsing process. ASTRAL branch support values represent the confidence that the method has in the quadripartition (four clusters around a branch) generated from removing the branch with higher support values indicating higher confidence. We define and vary this threshold t through our study to experimentally ascertain its ideal value. Once we have collapsed the branches based on said threshold, we arrive at an unresolved output tree T' from our originally estimated tree T. For each polytomy, our method resolves each polytomy by arbitrarily choosing a leaf from each branch in the unresolved polytomy, and running SVDquartets on this subset of leaves Q. As SVDquartets takes in alignments as an input, we create a mapping between alignments and their corresponding taxon and feed in the mapped alignments into SVDquartets for their corresponding taxon in he subset of taxons randomly picked from each subtree. We create this mapping between each leaf in Q with its associated subtree, and regraft the subtree back with its associated leaf in the resolved polytomy generated by SVDquartets. The resulting output estimation tree that contains the resolve polytomies is returned as the output of USA.



The example above shows a polytomy of degree $d = 8$. The USA algorithm chooses its subset of leaves from any polytomy by randomly selecting one leaf from each of the 8 subtrees shown and feeding the alignment data of each leaf into SVDquartets. The algorithm resolves the polytomy by using the output from SVDquartets where each leaf is replaced by its corresponding subtree.

We can now take a look at the dataset and the specific model conditions that we will be analyzing USA under and in comparison to both ASTRAL-III and SVDquartets. The dataset used in this study is the ASTRAL-II dataset generated in Molloy & Warnow as mimicking the original dataset conditions that found weaknesses in ASTRAL under high GTEE conditions is paramount in achieving results that are true to the original motivation of this paper. In order to prevent some unusual missing data issues with taxa found in some of the input gene trees, all gene trees found in the original dataset were regenerated from the alignment data using RAxML [8], in order to resolve this issue of missing taxon data in the subsequent running of the base ASTRAL-III method. We utilized the SimPhy [9] model to evolve genes down some species tree that falls under the Multi-species Coalescent (MSC) model. Each SimPhy model tree and collection of generated gene trees can be further parameterized by the following conditions in our study: number of taxa, number of genes, speciation rate, and species tree height. We set the number of taxa that we ran studies on to be 26 taxa to mimick the original Molloy & Warnow conditions. We further constrained the speciation rate to 10^{-6} and the species tree height to 500k. We varied the number of genes sequentially, (rather than randomly) from the original dataset, in the following orders of magnitude: 50, 100, 250, 500, 1000.

In designing the experiment, we decided on our performance metric being the average Robinson-Foulds (RF) error rate [10] between our estimated species tree and the true species tree. The RF error rate is calculated as the number of bipartitions found in one of the two trees, estimated and true, that does not appear in the other tree divided by the total number of possible bipartitions in both trees ($2n - 6$, where n is the number of leaves). We decided to run out experiment on the following replicates from the original ASTRAL-II dataset: 5, 11, 13, 16, 20 [6]. Further information on the scripts and datasets used in the studies shown in this paper can be found in the Supplementary Materials report.

3 Results & Discussion

The analysis of the result can be broken up into three main categories. The first category is comparing USA across three threshold values $t = 0.8, 0.4, 0.35$ with ASTRAL-III and SVDquartets across different number of genes. The next category compares USA across the three aforementioned threshold values with each other, and the final category compares the runtimes of the methods for their performance across the combined five different number of gene values. The RF error rates were calculated using DendroPy [11] and the resulting figures are shown below. Furthermore, when referring to performance in the analysis of the results and in the conclusion, unless otherwise stated, is referring to the RF error rate.

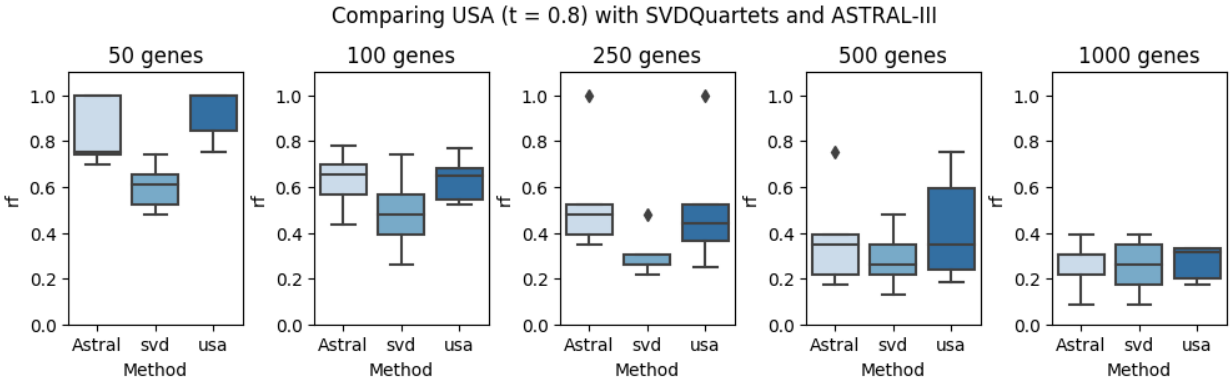


Figure 1: The above boxplot compares USA using a threshold cut-off of 0.8 with SVDquartets and ASTRAL across 50, 100, 250, 500 and 1000 genes.

There are interesting trends to note in Figure 1 when varying the number of genes for the three methods for USA with a threshold value of $t = 0.8$. SVDquartets performs considerably better than both ASTRAL-III and USA on 50 genes, with USA having a similar worse case estimation to ASTRAL-III. However, USA's average case on 50 genes is the worst of the tree methods. On 100 genes, SVDquartets shows the best results as well, although USA outperforms ASTRAL-III on its average case while ASTRAL-III shows a better best case situation. The gap between both methods with SVDquartets continues to reduce as the number of genes increases with USA considerably outperforming ASTRAL-III on all cases and being competitive with SVDquartets on the 250 gene conditions as well. Surprisingly, on 500 genes, USA seems to regress its performance although its average RF distance is still similar to ASTRAL-III. Finally, on the 1000 gene datasets, SVDquartets no longer shows the same strong performance on these model conditions and USA performs better than in its spread over the five replicates in comparison to SVDquartets, although ASTRAL slightly outperforms USA when it uses a threshold value of $t = 0.8$.

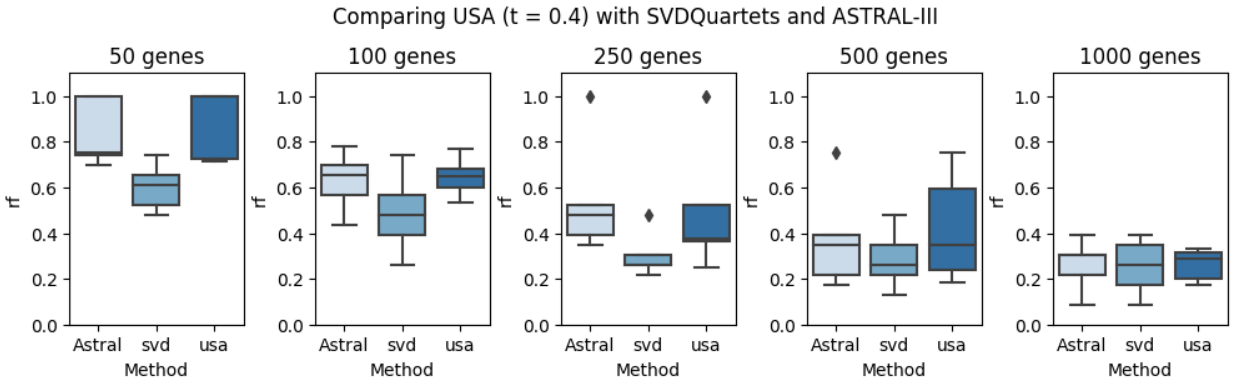


Figure 2: The above boxplot compares USA using a threshold cut-off of 0.4 with SVDquartets and ASTRAL across 50, 100, 250, 500 and 1000 genes.

A further analysis of the trends in Figure 2 can be compared with the USA method on a threshold value of $t = 0.4$. As the threshold value decreases for the 50 gene dataset, USA matches ASTRAL-III with a slightly better average RF distance across the five replicates. On the 100 gene dataset, USA performs similarly to it did in comparison to both SVDquartets and ASTRAL-III as it did with a threshold value of $t = 0.8$. Improvements for the lower threshold iteration of USA are visible in both the 250 gene dataset and the 1000 gene dataset where the average RF distance is much lower. In the 250 gene instance, USA outperforms ASTRAL-III on this metric.

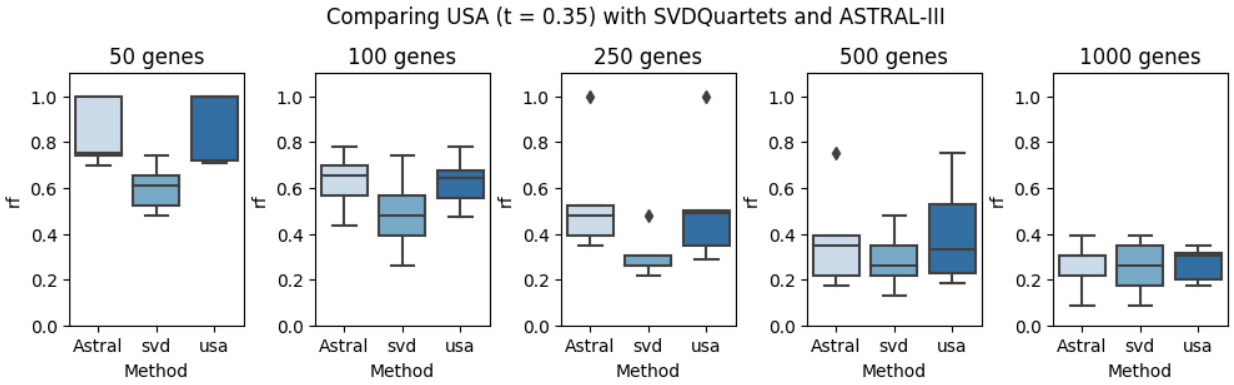


Figure 3: The above boxplot compares USA using a threshold cut-off of 0.35 with SVDquartets and ASTRAL across 50, 100, 250, 500 and 1000 genes.

A similar analysis on Figure 3 can be done on USA with a threshold value of $t = 0.35$. The lower threshold cut-off does not seem to affect the performance of USA on the 50 gene dataset. However, we can see improvements of USA on the 100 gene dataset, specifically with improvements on its best and average case making it slightly better than ASTRAL-III on these replicates. An interesting trend shown on the 250 gene dataset is the regression on performance when the threshold is further lowered from $t = 0.4$ to $t = 0.35$ on the average performance of USA. Both 500 and 100 gene datasets show similar performance for USA with a threshold value of 0.35 as well when compared to the $t = 0.4$ iteration.

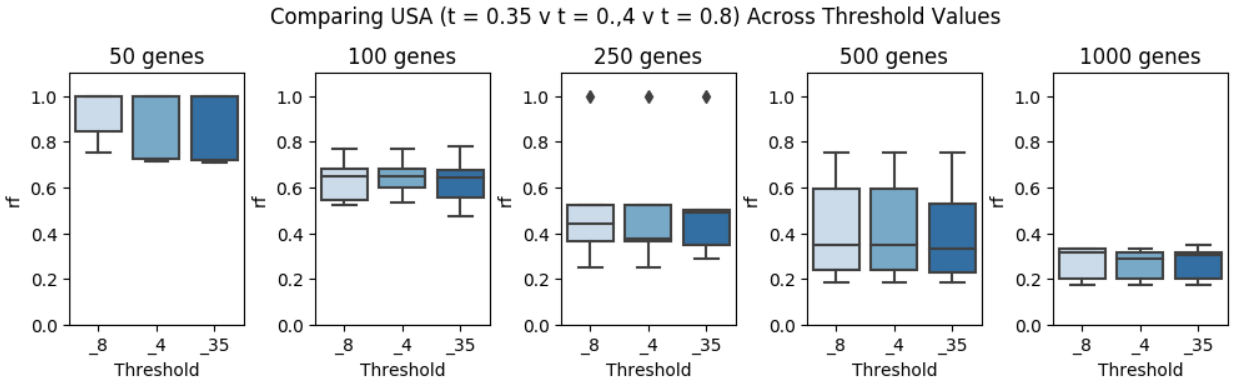


Figure 4: The above boxplot compares USA across the three aforementioned threshold values $t = 0.8, 0.4, 0.35$ across 50, 100, 250, 500 and 1000 genes.

We can then compare the three versions of USA across the same five model conditions to get a better gauge of the importance of threshold cut-offs in their performance on the Molloy & Warnow ASTRAL-II dataset as shown in Figure 4. As the number of genes increase, the performance of our method across all three threshold cut-offs improves. On the 50 gene model condition, as the threshold increases, the performance of USA suffers. This general trend can be shown to some extent in all two other conditions, with both 250 and 500 gene dataset performance results showing clear improvements as the threshold cut-off value decreases.

There are some other interesting trends for the remaining two model conditions. Both 100 and 1000 gene datasets show performance regression as the the threshold cut-off decreases, which points to inconclusive results regarding improvement trends based on hard cut-off values. As the threshold cut-off decreases, USA only utilizes SVDquartets on branches with a much lower confidence bar, which tend to be areas in which ASTRAL-III performs poorly and can accurately assess this low confidence. However, as the cut-off increases, the polytomies generated by collapsing even higher support branches can lead to unnecessary regrafting caused by running SVDquartets which can lead to greater inaccuracy especially if those polytomies have been generated closer to the leaves. Datasets that show performance improvements as USA decreases its threshold cut-off tend to see higher occurrences of sub-threshold support branches near the root rather than the leaves, and thus resolves these polytomies where performance improvements are much more apparent and stark in contrast with lower level potential changes that the base method performs well on.

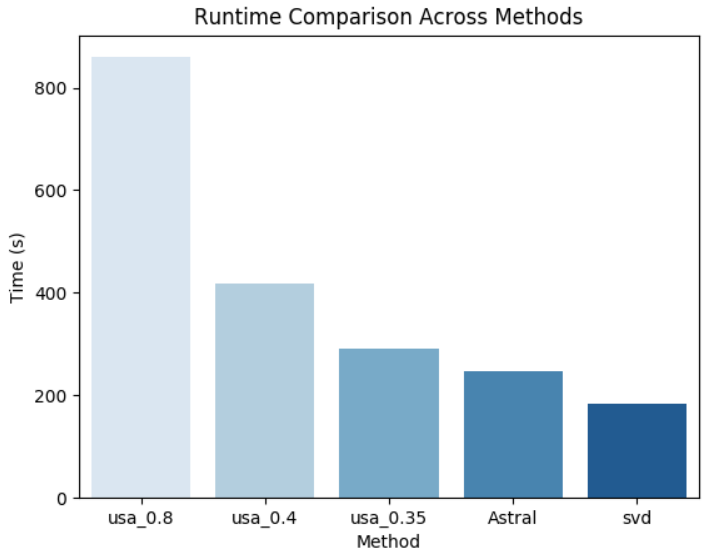


Figure 5: The above bar chart compares the runtime of five methods across all five model conditions.

The analysis of the runtime of the three instances of USA coupled with ASTRAL-III and SVDquartets gives results as expected based on the construction of each method as shown in Figure 5. The runtime comparison takes into account the total time that each method took on all five gene number values as mentioned in earlier results. SVDquartets as shown in the above figure takes the least amount of time to run on all five model conditions followed by ASTRAL-III. When looking at USA, a method that aggregates both SVDquartets and ASTRAL-III in its algorithmic construction, as the threshold value increases, the runtime increases at a linear rate. As the threshold increases, USA begins to collapse branches that contain higher levels of support, and thus as the threshold value t approaches 1, the number of collapsed branches approaches the total number of branches in the ASTRAL-III output tree T. Therefore, the number of unresolved polytomies for which USA runs SVDquartets on to resolve approaches the entirety of T which accounts for the increase in runtime.

In analyzing the linear rate of increase across increasing threshold values, there are two potential factor that can be taken into account for this occurrence. Firstly, the runtime data supports the assertion that the distribution of support values take a distribution on the ASTRAL-II dataset that favors branches with low confidence support branches that can be found in clusters along the initial ASTRAL-III output tree T. Thus the linear rate of runtime growth can be shown to exist as lower threshold values will lead to more additional collapsed branches found in specific polytomies while higher threshold values will have more additional total polytomies on T to resolve.

4 Conclusion

The results of the study show promise in improving the ASTRAL family of methods through points of low confidence shown through low branch support values. However, there are some open questions that are key in better understanding how low confidence threshold values can be calculated. The results of our study theorize that hard cut-offs that do not take into context other factors and

structures of the data and thus performance based on these hard threshold cut-offs vary widely and thus are not a proper fit for improving the performance of ASTRAL in a statistically significant manner. Further areas of improvement in giving contextual threshold markers can relate to using statistical criteria such as the Z-score in order to better score the uncertainty of support branch values while taking into account the support value distribution across the entire base estimation tree as well as creating heuristics regarding the leaf selection step when running SVDquartets on an unresolved polytomy. The current arbitrary selection can provide varied results and thus improvements on this step can be vital in selecting better leaf candidates to run SVDquartets on.

Further improvements to the experiment design can be described as follows. Increasing the number of experiments run and the number of replicates can provide greater clarity into some of the trends visualized in this study. Furthermore, running USA on biological datasets that contain many of the same degrees of ILS such as the Jarvis et al. Avian biological datasets [12] can be further avenues of research into testing the performance of USA.

Another potential area of improvement on the method as a whole can be from dynamic reconstructions of the algorithm that utilize different ways of capturing the output of both ASTRAL and SVDquartets. One possible method that utilizes both outputs through a different lense stems from running both methods on subsets of the total input gene tree/alignment set, and generating a consensus species tree estimation from the output of both methods. Then running some form of Maximum Likelihood estimation, like FastTree2 [13], on the unresolved polytomies can be sufficient in resolving them accurately while not sacrificing scalability. Future research through any of the avenues mentioned are key in further improvements of USA and species tree estimation in general. For any further questions on the contents of this study, feel free to email the author at akhilrj2@illinois.edu.

References

[1] Mirarab, S. & Warnow, T. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* **31**, i44–i52 (2015). URL `https://doi.org/10.1093/bioinformatics/btv234`. `https://academic.oup.com/bioinformatics/article-pdf/31/12/i44/17101439/btv234.pdf`.

[2] Molloy, E. K. & Warnow, T. To Include or Not to Include: The Impact of Gene Filtering on Species Tree Estimation Methods. *Systematic Biology* **67**, 285–303 (2017). URL `https://doi.org/10.1093/sysbio/syx077`. `https://academic.oup.com/sysbio/article-pdf/67/2/285/24105597/syx077.pdf`.

[3] Chou, J. *et al.* A comparative study of SVDquartets and other coalescent-based species tree estimation methods. *BMC Genomics* **16**, S2 (2015). URL `https://doi.org/10.1186/1471-2164-16-S10-S2`.

[4] Snir, S. & Rao, S. Quartet maxcut: A fast algorithm for amalgamating quartet trees. *Molecular Phylogenetics and Evolution* **62**, 1–8 (2012). URL `https://www.sciencedirect.com/science/article/pii/S1055790311003101`.

[5] Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**, 153 (2018). URL `https://doi.org/10.1186/s12859-018-2129-y`.

[6] Yin, J., Zhang, C. & Mirarab, S. ASTRAL-MP: scaling ASTRAL to very large datasets using randomization and parallelization. *Bioinformatics* **35**, 3961–3969 (2019). URL https://doi.org/10.1093/bioinformatics/btz211. https://academic.oup.com/bioinformatics/article-pdf/35/20/3961/30149137/btz211.pdf.

[7] Junier, T. & Zdobnov, E. M. The Newick utilities: high-throughput phylogenetic tree processing in the Unix shell. *Bioinformatics* **26**, 1669–1670 (2010). URL https://doi.org/10.1093/bioinformatics/btq243. https://academic.oup.com/bioinformatics/article-pdf/26/13/1669/512992/btq243.pdf.

[8] Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006). URL https://doi.org/10.1093/bioinformatics/btl446. https://academic.oup.com/bioinformatics/article-pdf/22/21/2688/16851699/btl446.pdf.

[9] Mallo, D., De Oliveira Martins, L. & Posada, D. SimPhy : Phylogenomic Simulation of Gene, Locus, and Species Trees . *Systematic Biology* **65**, 334–344 (2015). URL https://doi.org/10.1093/sysbio/syv082. https://academic.oup.com/sysbio/article-pdf/65/2/334/24589042/syv082.pdf.

[10] Robinson, D. & Foulds, L. Comparison of phylogenetic trees. *Mathematical Biosciences* **53**, 131–147 (1981). URL https://www.sciencedirect.com/science/article/pii/0025556481900432.

[11] Sukumaran, J. & Holder, M. T. DendroPy: a Python library for phylogenetic computing. *Bioinformatics* **26**, 1569–1571 (2010).

[12] Jarvis, E. D. *et al.* Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320–1331 (2014). URL https://www.science.org/doi/abs/10.1126/science.1253451. https://www.science.org/doi/pdf/10.1126/science.1253451.

[13] Price, M. N., Dehal, P. S. & Arkin, A. P. Fasttree 2 – approximately maximum-likelihood trees for large alignments. *PLOS ONE* **5**, 1–10 (2010). URL https://doi.org/10.1371/journal.pone.0009490.