Expediting Orthology Inference Through the Upfront Exclusion of Paralogous Sequences in Annotated Assemblies

---

LFSC 507 Final Project Report

Jackson Hoehn Turner
6 December 2024

INTRODUCTION

The modeling of evolutionary relationships between organisms is a foundational tenant of modern taxonomy and systematics. Such models were originally derived from comparisons of character states between selected organisms (Wiens et al. 2004). While potentially informative character-driven datasets lack the necessary depth to adequately represent groups of organisms to accurately model their relationships (Wiens et al. 2004; Young et al. 2020; Lozano-Fernandez et al. 2022). The advent of molecular methods allowed for the application of genetic information for evolutionary modeling. Methods such as Sanger sequencing were applied to recover new systematic relationships between organisms, especially those for which morphological differentiation is untenable (Shaibu et al. 2021). Despite the feasibility of such techniques, the high cost of sequencing per base pair limited the volume and relevance of genetic information accessible for phylogenetic applications (Young et al. 2020; Lozano-Fernandez et al. 2022). While more informative genes may be ubiquitous within groups of organisms, the high time and cost associated with their recovery via first generation sequencing limit their utility. Despite its successful application in quantifying biodiversity through DNA barcoding and limited utility for phylogenetics, Sanger sequencing lacks the ability to recover enough loci to be reliably useful for systematic review across a wide range of taxa (Young et al. 2020; Lozano-Fernandez et al. 2022).

The advent of modern sequencing technology has introduced an unprecedented volume of genomic data for which phylogenetic tools must adapt to effectively utilize. The low cost per base pair of next generation sequencing has resulted in the explosive increase of available genetic information (Young et al. 2020; Lozano-Fernandez et al. 2022). This advancement has culminated in an abundance of publicly available next generation sequencing reads Young et al. 2020; Lozano-Fernandez et al. 2022). The volume of loci recovered in this way introduces numerous single copy genes (orthologs) with phylogenetic utility (Young et al. 2020; Lozano-Fernandez et al. 2022). The retrieval of potentially thousands orthologs per sequenced organism through next generation sequencing presents a unique opportunity to propose and test taxonomic hypotheses (Young et al. 2020; Lozano-Fernandez et al. 2022; Yang et al. 2023). An ever-expanding assortment of bioinformatic tools has been designed to process sequencing data for analyses across the scientific gamut. Several of these tools, and pipelines leveraging them, have been applied for the recovery of phylogenetic reconstructions (Young et al. 2020; Lozano-Fernandez et al. 2022).

Upfront removal of paralogous sequences may streamline the generation of phylogenetic trees by reducing the time required for orthology inference. Orthology inference is a bottleneck for phylogenomics pipelines due to the computational complexity and resource demand of all-vs-all comparison of loci (Young et al. 2023). Taxa with high numbers of paralogous sequences, arising from factors such as large genome sizes, an abundance of repetitive elements,

or polyploidy, needlessly inflate the complexity of orthology inference. These genomic features present particular challenges in the orthology inference of some plants and fungi for this reason (Young et al. 2020; Yang et al. 2023). By reducing the analysis time required by orthology inference, excluding paralogous sequences from this analysis is expected to expedite phylogenetic reconstruction from genomic data.

METHODS

A dataset of 29 Gesneriaceae (Plantae:Lamiales) was selected for their complex and feature-rich genomic histories and previous inclusion in a state-of-the-art phylogenomics study for which to benchmark against the analysis performed here (Yang et al. 2023).

Full analyses of methods performed here were conducted on UTIA server centaur.ag.utk.edu while a test dataset was uploaded to and executed on the LFSC 507 course server (bioinfo.tk) to evaluate the performance of the custom python script representing the focal point of this project. Code for this project is publicly available at https://github.com/jacksonhturner/lfsc_507/tree/main/final_project.

Short reads representative of selected taxa were downloaded from NCBI SRA onto UTIA server centaur.ag.utk.edu using SRAtoolkit (Sayers et al. 2022). Raw reads were trimmed for adapter contamination and low quality sequences with cutadapt and assembled with Megahit under default parameters (Martin 2011; Li et al. 2015). Resulting assemblies were annotated using Augustus with a *S. lycosperum* reference. Annotated proteomes of selected taxa were queried against themselves with DIAMOND (Buchfink et al. 2021) under default parameters to identify potentially paralogous sequences. A custom python script was used to filter loci from both amino acid and nucleotide sequence proteomes matching to more than one other sequence in the prior DIAMOND search.

Run times for DIAMOND, OrthoFinder, and python script jobs were recorded alongside the percentage of sequences classified as orthologs for each sample (Buchfink et al. 2021; Emms & Kelly 2019). The correlation between recovered orthologs and python script run time was determined using the numpy package (Harris et al. 2020). Figures demonstrating the effectiveness of the python script and its relationship with assembly size were created using matplotlib (Hunter 2007). A t-test was performed using the scipy.stats package (Virtanen et al. 2020).

Orthology inference was conducted upon filtered orthologs for Gesneriaceae samples using OrthoFinder under default parameters and 15 cores (Emms & Kelly 2019). OrthoFinder run time for this pipeline was compared with an identical execution of the methods discussed here without paralog exclusion and 8 cores. Orthologs present in 90% of taxa were retained for

phylogenetic analysis using a custom python script (created prior to this course). Resulting orthologs were aligned with MAFFT (Katoh & Standley 2013) and masked with trimAl (Capella-Gutiérrez et al. 2009) at a threshold of 40% missing data. Maximum likelihood (ML) tree inference was conducted upon masked amino acid alignments using IQ-TREE2 with ModelFinder-Plus to identify the best phylogenetic model, 1000 ultrafast bootstraps, nearest neighbor interchange, and a relaxed hierarchical clustering algorithm (Minh et al. 2020; Kalyaanamoorthy et al. 2017). FigTree was used to visualize the ML reconstruction produced (Rambaut 2010). The resulting phylogeny was benchmarked for accuracy against a ML reconstruction from a study from which this dataset originated using Robinson-Foulds scores from the TreeDist R package (Smith 2020). The topology for the ground-truth tree was recovered using TreeSnatcher upon the reported phylogeny constructed from selected taxa (Laubach et al. 2012).

RESULTS

DIAMOND run time was significantly shorter than that of the custom python script ($p < 0.001$) irrespective of sample identity [Figure 1]. Most samples demonstrated a run time of less than 24 hours for the custom python script with one an outlier that required over 100 hours of wall time to complete [Figure 1]. The number of recovered orthologs has a significant and strongly positive association with the time required to run the custom python script for the provided Gesneriaceae dataset ($p < 0.05$; $r = 0.94$) [Figure 1].

OrthoFinder run time after paralog exclusion was 10 hours and 15 minutes compared to the 9 days, 18 hours, and 38 minutes required without the pipeline performed for this project [Figure 2]. The sum total wall time required for this pipeline, inclusive of DIAMOND and the custom python script, is 22 days, 11 hours, and 13 minutes [Figure 2]. Standardized for 8-core parallelization of the executed pipeline, its total wall time is reduced to 3 days, 4 hours, and 15 minutes [Figure 2]. Orthology inference for the pipeline performed recovered 11 single-copy genes that were, among OrthoFinder results, orthologous in 90% of taxa representing 95,111 bp across sequences inclusive of all taxa.

The ML reconstruction recovered from the pipeline performed here demonstrates a highly similar topology to the ground-truth tree from Yang et al (2023) [Figures 3 & 4]. Clade support throughout the phylogeny is weak, especially nodes deeply nested within *Henckelia* [Figure 3]. Monophyly of *Henckelia*, with the exception of *H. oblongifolia*, is conserved across both phylogenies with slight topological differences present in intergroup nodes [Figure 4]. Most clades recovered from the ML reconstruction generated here are consistent with the ground-truth tree, but some – especially those deep within *Henckelia*, display discordant topology [Figure 4].

DISCUSSION

The methods performed here were successful in dramatically reducing the time required for othology inference within the tested dataset. While the filtering of paralogous sequences likely was a substantial contributor to this observation, including only amino acid ortholog sequences and excluding nucleotide sequences are expected to have impacted OrthoFinder run time. A greater number of cores devoted to OrthoFinder for the orthology inference of the pipeline with paralogs excluded likely also impacted run time and is reflective upon the time constraints imposed by the deadline of this assignment. The inclusion of corresponding nucleotide sequences and the implementation of an equal number of cores is expected to provide a more comprehensive evaluation of filtering paralogous sequences prior to orthology inference.

The ML reconstruction created from the Gesneriaceae dataset captured a reasonably accurate topology relative to the study upon which it was benchmarked given the low number of genes acquired. Future selection of genes orthologous in a lower percentage of taxa is expected to substantially raise the number of loci recovered for ML reconstruction (Young et al. 2020). Monophyly of the genus *Henckelia* was recovered from the reconstruction generated from the pipeline performed here with minor topological discordance between intergroup nodes, with the exception of *H. oblongifolia*, agreeing for conventional systematic findings for this group (Yang et al. 2023). *H. oblongifolia* is a known systematic outlier regarded ancestral to other Henckelia, and reflects the ongoing systematic development for this group (Yang et al. 2023). Low observed bootstrap support and extended terminal branches in the resultant ML reconstruction is likely driven by the low number of genes selected for analysis. The limited genetic information leveraged here is suspected to have generated the observed topological discordance from the ground-truth tree.

Despite the restrictions of data quality and analysis time imposed by the submission deadline for this assignment, the methods performed here achieved the primary goal of expediting orthology inference while producing a ML reconstruction with an acceptably accurate topology (Yang et al. 2023). The pronounced reduction in OrthoFinder run time is likely at least in part due to the upfront exclusion of paralogs upon which this project is predicated. The exponentially increasing demand of computational resources of all-vs-all orthology inference with the inclusion of additional taxa foreshadows the benefit of upfront paralog exclusion in taxa with large volumes of genetic content, like many plants (Young et al. 2020; Yang et al. 2023). These results incentivize the future exploration of this methodology to streamline phylogenetic reconstruction across a wide range of taxonomic groups, especially those with large genome sizes.

TABLE & FIGURES

Table 1: Table of Gesneriaceae sample species and SRA accession numbers. This table is available at https://github.com/jacksonhturner/lfsc_507/blob/main/final_project/final_submission/LFSC_final_project_table_1.csv.

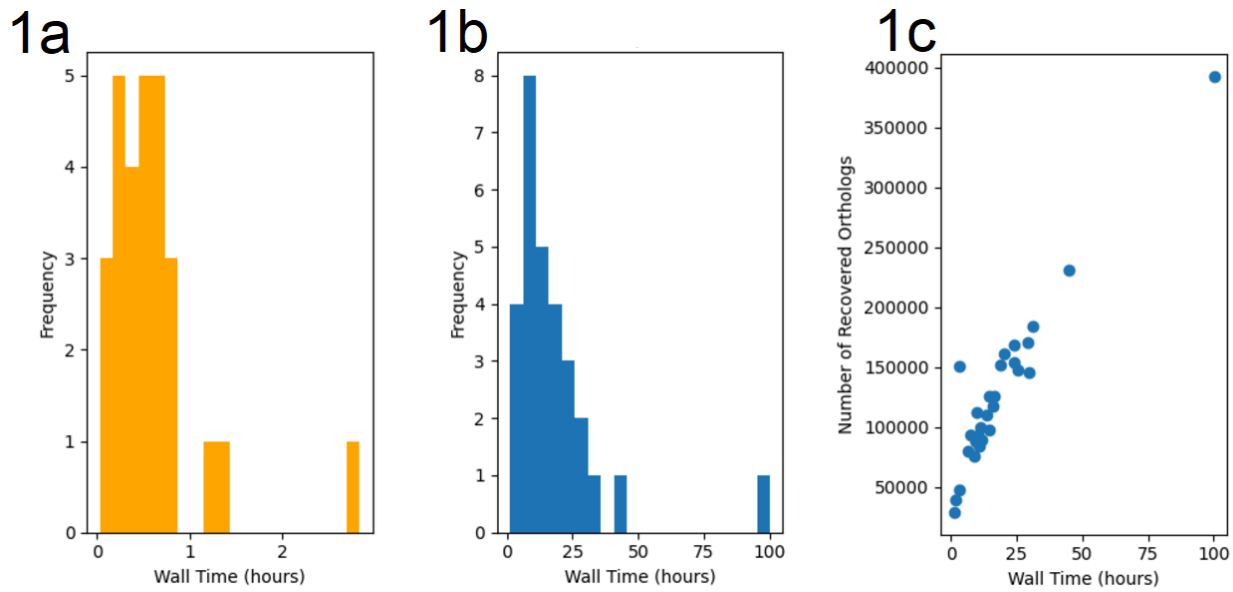

Figure 1: Histogram of run time of DIAMOND for Gesneriaceae samples (1a), histogram of run times of the custom python script used to identify and recover orthologs from Gesneriaceae DIAMOND results (1b), and scatterplot demonstrating the relationship between the run time of the custom python script for Gesneriaceae samples and the number of orthologs recovered (1c). A full-size image of this figure is available at https://github.com/jacksonhturner/lfsc_507/blob/main/final_project/final_submission/LFSC_final_project_figure_1.png.

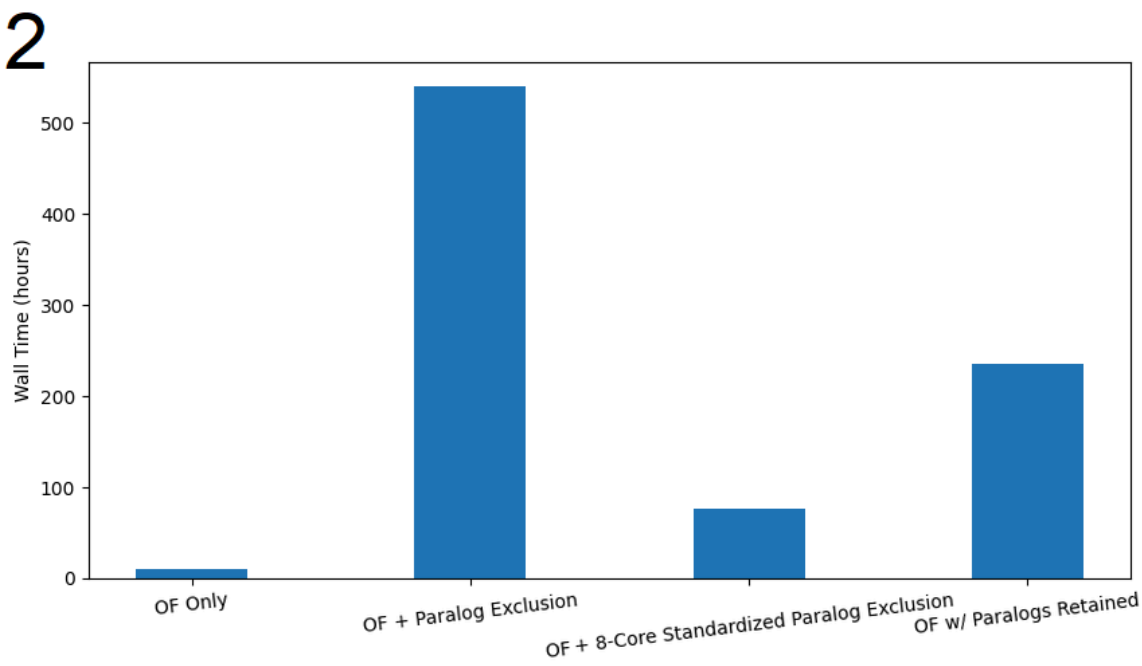

Figure 2: Bar plot of OrthoFinder benchmarking run times in hours for tested procedures. From left to right, tested procedures include OrthoFinder run time only after paralog exclusion ("OF only"); the sum total run time of OrthoFinder, DIAMOND, and the custom python script ("OF + Paralog Exclusion); the sum total run time of OrthoFinder, DIAMOND, and the custom python script with DIAMOND and custom python script run times divided among 8 cores ("OF + 8-Core Standardized Paralog Exclusion"); and OrthoFinder with paralogous sequences retained ("OF w/ Paralogs Retained"). A full-size image of this figure is available at https://github.com/jacksonhturner/lfsc_507/blob/main/final_project/final_submission/LFSC_final_project_figure_2.png.

3

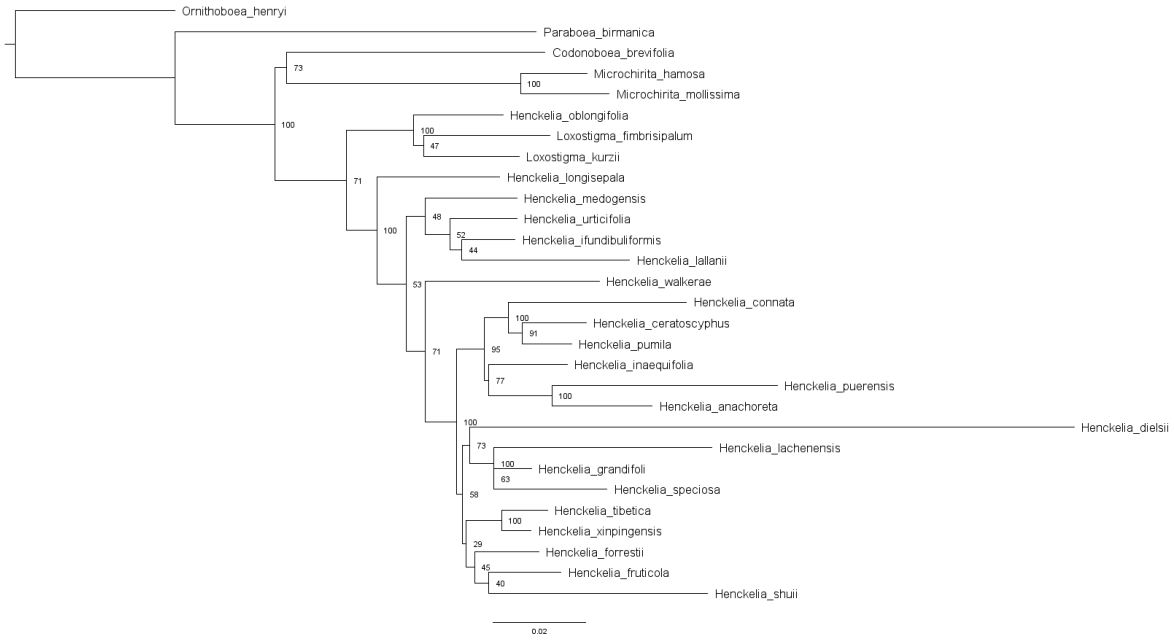

Figure 3: ML reconstruction of 29 Gesneriaceae samples rooted upon *O. henryi* due to its relative outgroup status. The scale bar represents the number of base pair substitutions per site as a measure of branch length. Enumerated nodes report bootstrap support percentages. A full-size image of this figure is available at https://github.com/jacksonhturner/lfsc_507/blob/main/final_project/final_submission/LFSC_final_project_figure_3.png.

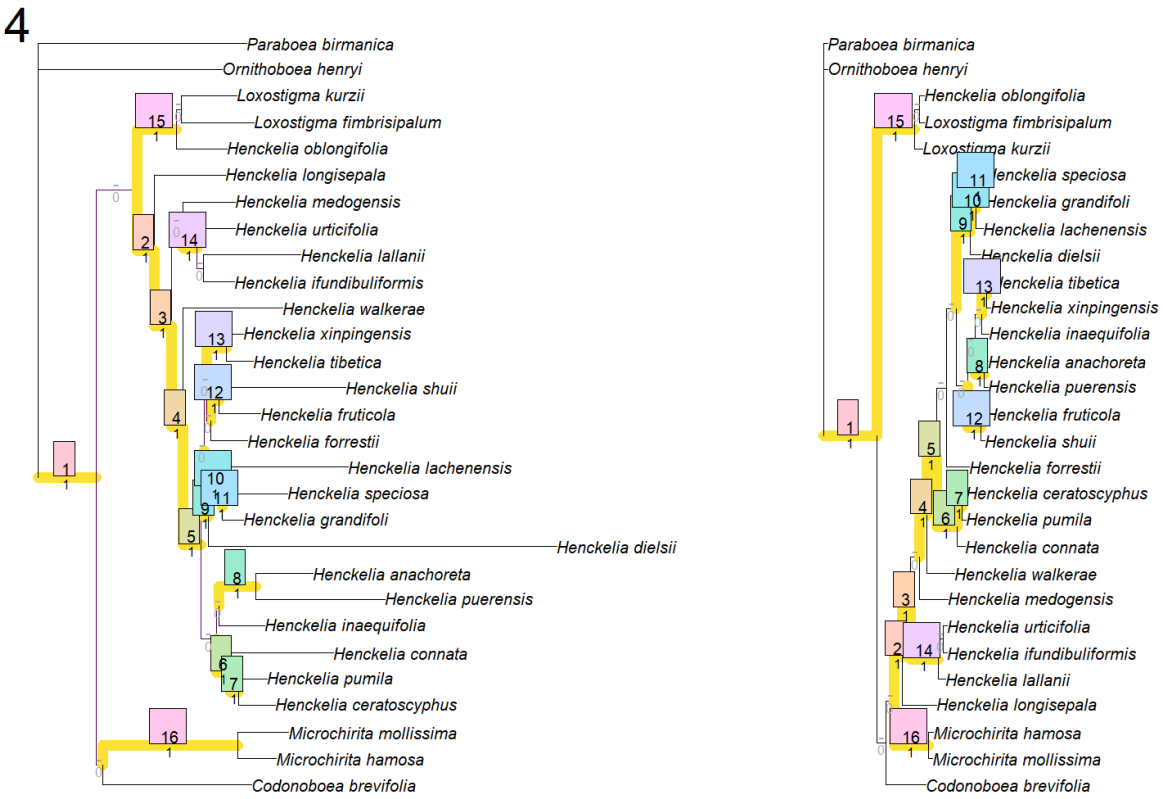

Figure 4: Side-by-side comparison of ML phylogenetic reconstruction from the executed pipeline (left) and the ground-truth phylogeny upon which this exercise was benchmarked (right). Interclade nodes are denoted by colored, enumerated boxes while shared clades are highlighted in yellow. A full-size image of this figure is available at https://github.com/jacksonhturner/lfsc_507/blob/main/final_project/final_submission/LFSC_final_project_figure_4.png.

REFERENCES

Buchfink B, Reuter K, Drost HG, "Sensitive protein alignments at tree-of-life scale using DIAMOND", Nature Methods 18, 366–368 (2021). doi:10.1038/s41592-021-01101-x

Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. “trimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses.” Bioinformatics 25, no. 15 (August 1, 2009): 1972–73. https://doi.org/10.1093/bioinformatics/btp348.

Emms, David M., and Steven Kelly. “OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics.” Genome Biology 20, no. 1 (December 2019): 238. https://doi.org/10.1186/s13059-019-1832-y.

Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. Nature 585, 357–362 (2020). https://doi.org/10.1038/s41586-020-2649-2

Hunter, J.D. "Matplotlib: A 2D Graphics Environment," in Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, May-June 2007, doi: 10.1109/MCSE.2007.55.

Kalyaanamoorthy, Subha, Bui Quang Minh, Thomas K F Wong, Arndt Von Haeseler, and Lars S Jermiin. “ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates.” Nature Methods 14, no. 6 (June 2017): 587–89. https://doi.org/10.1038/nmeth.4285.

Katoh, K., and D. M. Standley. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” Molecular Biology and Evolution 30, no. 4 (April 1, 2013): 772–80. https://doi.org/10.1093/molbev/mst010.

Laubach, Thomas, Arndt Von Haeseler, and Martin J Lercher. “TreeSnatcher plus: Capturing Phylogenetic Trees from Images.” BMC Bioinformatics 13, no. 1 (December 2012): 110. https://doi.org/10.1186/1471-2105-13-110.

Li, Dinghua, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. “MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph.” Bioinformatics 31, no. 10 (May 15, 2015): 1674–76. https://doi.org/10.1093/bioinformatics/btv033.

Lozano-Fernandez, Jesus. “A Practical Guide to Design and Assess a Phylogenomic Study.” Genome Biology and Evolution 14, no. 9 (September 1, 2022): evac129. https://doi.org/10.1093/gbe/evac129.

Martin, Marcel. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” EMBnet.Journal 17, no. 1 (May 2, 2011): 10. https://doi.org/10.14806/ej.17.1.200.

Minh, Bui Quang, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt Von Haeseler, and Robert Lanfear. “IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era.” Edited by Emma Teeling. Molecular Biology and Evolution 37, no. 5 (May 1, 2020): 1530–34. https://doi.org/10.1093/molbev/msaa015.

Rambaut, A. (2010) FigTree v1.3.1. Institute of Evolutionary Biology, University of Edinburgh, Edinburgh. http://tree.bio.ed.ac.uk/software/figtree/

Sayers E.W., Bolton E.E., Brister J.R., Canese K., Chan J., Comeau D.C., Connor R., Funk K., Kelly C., Kim S., Madej T., Marchler-Bauer A., Lanczycki C., Lathrop S., Lu Z., Thibaud-Nissen F., Murphy T., Phan L., Skripchenko Y., Tse T., Wang J., Williams R., Trawick B.W., Pruitt K.D., Sherry S.T. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2022 Jan 7;50(D1):D20-D26. doi: 10.1093/nar/gkab1112.

Smith, Martin R. “Information Theoretic Generalized Robinson–Foulds Metrics for Comparing Phylogenetic Trees,” 2020.

Stanke, M., O. Keller, I. Gunduz, A. Hayes, S. Waack, and B. Morgenstern. “AUGUSTUS: Ab Initio Prediction of Alternative Transcripts.” Nucleic Acids Research 34, no. Web Server (July 1, 2006): W435–39. https://doi.org/10.1093/nar/gkl200.

Shaibu J.O., Onwuamah C.K., James A.B., Okwuraiwe A.P., Amoo O.S., Salu O.B., Ige F.A., Liboro G., Odewale E., Okoli L.C., Ahmed R.A., Achanya D., Adesesan A., Amuda O.A., Sokei J, Oyefolu B.A.O., Salako B.L., Omilabu S.A., Audu R.A. Full length genomic sanger sequencing and phylogenetic analysis of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) in Nigeria. PLoS One. 2021 Jan 11;16(1):e0243271. doi: 10.1371/journal.pone.0243271. PMID: 33428634; PMCID: PMC7799769.

Wiens, John J. “The Role of Morphological Data in Phylogeny Reconstruction.” Edited by Tim Collins. Systematic Biology 53, no. 4 (August 1, 2004): 653–61. https://doi.org/10.1080/10635150490472959.

Virtanen, P., Gommers, R., Oliphant, T.E. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17, 261–272 (2020). https://doi.org/10.1038/s41592-019-0686-2

Yang, Li-Hua, Xi-Zuo Shi, Fang Wen, and Ming Kang. “Phylogenomics Reveals Widespread Hybridization and Polyploidization in Henckelia (Gesneriaceae).” Annals of Botany 131, no. 6 (July 10, 2023): 953–66. https://doi.org/10.1093/aob/mcad047.

Young, Andrew D., and Jessica P. Gillung. “Phylogenomics — Principles, Opportunities and Pitfalls of Big‐data Phylogenetics.” Systematic Entomology 45, no. 2 (April 2020): 225–47. https://doi.org/10.1111/syen.12406.