

Reply to Springer and Gatesy: The impact of long branches and misalignments on phylogenetic analysis is minimal

Shaoyuan Wu^{a,1}, Frank E. Rheindt^b, Jin Zhang^c, Jiajia Wang^a, Lei Zhang^a, Cheng Quan^d, Zhiheng Li^e, Min Wang^e, Feixiang Wu^e, Yanhua Qu^{f,1}, Scott V. Edwards^{g,1}, Zhonghe Zhou^{e,1}, and Liang Liu^{h,1}

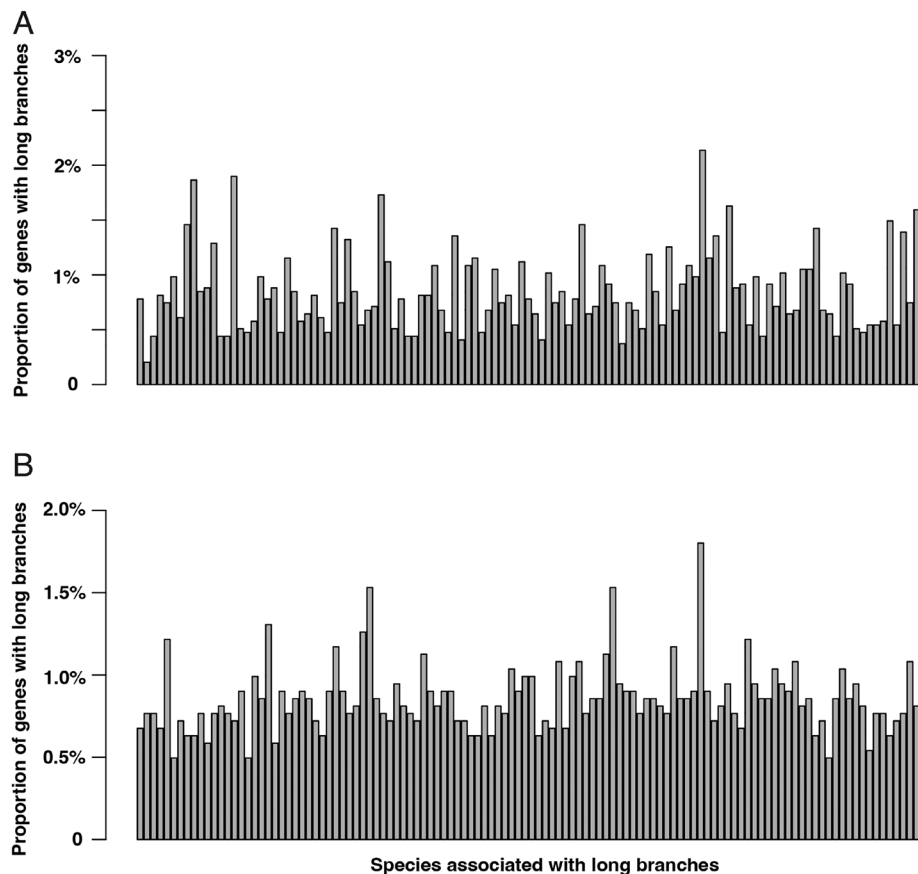


Fig. 1. (A) Random distribution of long branches across CDS gene trees; (B) Random distribution of long branches across intergenic gene trees.

Springer and Gatesy (1) suggested that our avian species tree, constructed using NJst (2), may be inaccurate due to observed disparities with the ASTRAL tree and potential misalignments. However, akin to their critiques of our other work, their criticisms stem from subjective opinions and misconceptions of phylogenetic analysis, rendering their arguments unfounded.

Springer and Gatesy argued that NJst is less accurate than ASTRAL, citing evidence from two simulation studies (3, 4). However, assessing phylogenetic methods solely through simulations cannot fully reflect real-world scenarios. Other studies have indeed suggested that NJst outperforms ASTRAL (5). Notably, NJst outcomes align closely with RAxML (*SI Appendix*, Figs. S1–S5 in ref. 2), indicating the robustness of NJst. In contrast, ASTRAL results differ significantly, warranting further validation.

Springer and Gatesy suggested that data misalignments could impact our phylogenetic results, citing 2,305 long branches (≥ 1.0 substitutions/site) across 1,824 gene trees: 1,626 CDS, 188 intergenic segments, and 10 introns. This

Author affiliations: ^aJiangsu Key Laboratory of Phylogenomics & Comparative Genomics, Jiangsu International Joint Center of Genomics, School of Life Sciences, Jiangsu Normal University, Xuzhou, Jiangsu 221116, China; ^bDepartment of Biological Sciences, National University of Singapore, Singapore 117543, Singapore; ^cSchool of Computer and Communication Engineering, Changsha University of Science & Technology, Changsha, Hunan 410114, China; ^dSchool of Earth Science and Resources, Chang'an University, Xi'an, Shaanxi 710054, China; ^eKey Laboratory of Vertebrate Evolution and Human Origins, Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing 100044, China; ^fKey Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China; ^gDepartment of Organismic and Evolutionary Biology, Museum of Comparative Zoology, Harvard University, Cambridge, MA 02138; and ^hDepartment of Statistics & Institute of Bioinformatics, University of Georgia, Athens, GA 30606

Author contributions: S.W., S.V.E., Z.Z., and L.L. designed research; S.W., F.E.R., J.Z., S.V.E., Z.Z., and L.L. performed research; Y.Q. contributed new reagents/analytic tools; S.W., F.E.R., J.Z., J.W., L.Z., C.Q., Z.L., M.W., F.W., Y.Q., S.V.E., Z.Z., and L.L. analyzed data; and S.W. wrote the paper.

The authors declare no competing interest.

Copyright © 2024 the Author(s). Published by PNAS. This article is distributed under [Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 \(CC BY-NC-ND\)](#).

¹To whom correspondence may be addressed. Email: shaoyuanwu@post.harvard.edu, quyh@ioz.ac.cn, sedwards@fas.harvard.edu, zhouzhonghe@ivpp.ac.cn, or lliu@uga.edu.

Published July 8, 2024.

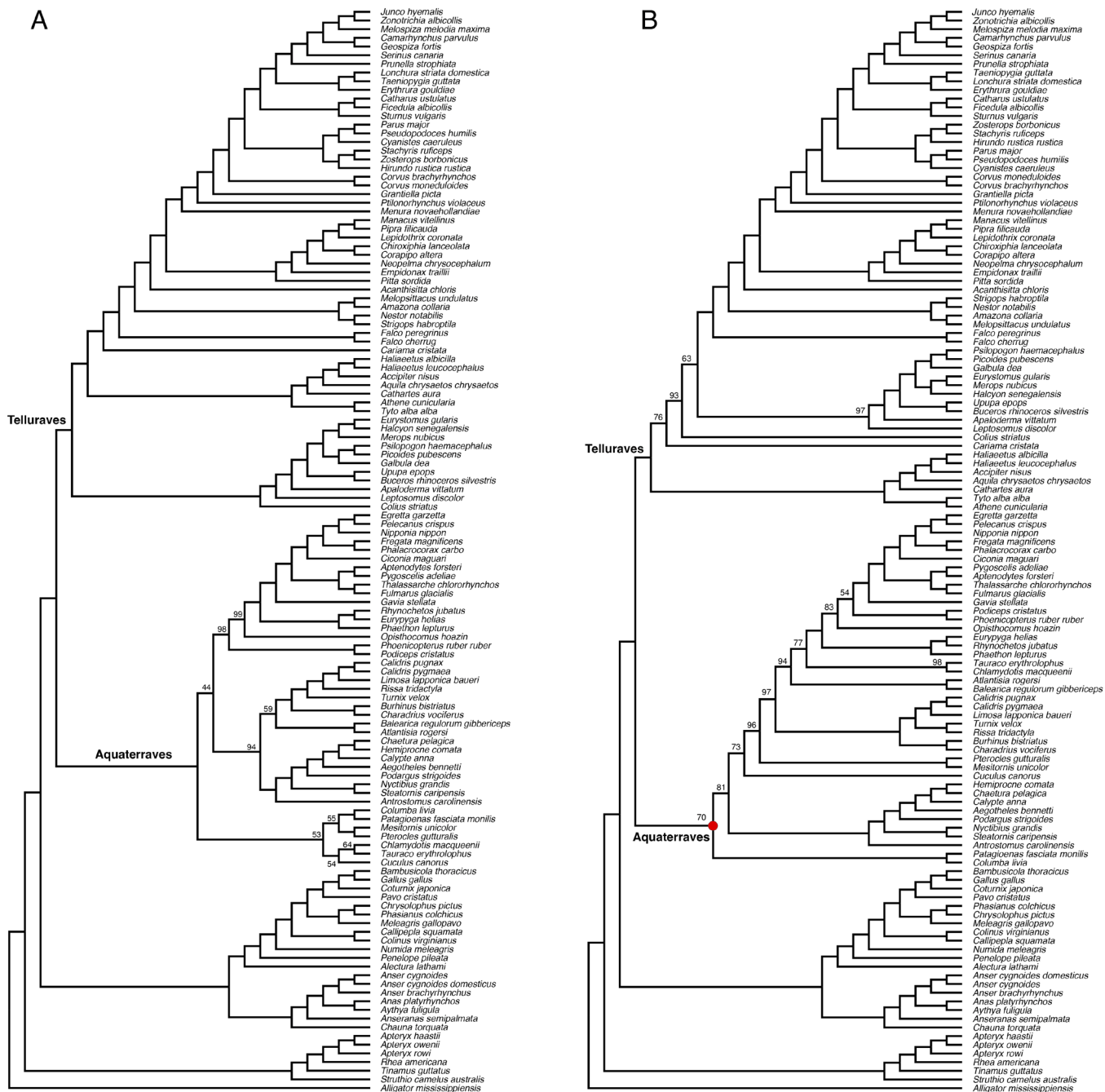


Fig. 2. (A) Njst species tree after excluding the 991 CDS and 180 intergenic gene trees with long branch (≥ 1 substitutions/site); (B) ASTRAL species tree after excluding the 991 CDS and 180 intergenic gene trees with long branch (≥ 1 substitutions/site). Numbers on nodes indicate the bootstrap support values, and nodes with maximum bootstrap support are not shown. Note that in the resulting ASTRAL tree, Columbiformes is classified within Aquaterraves (highlighted in red), diverging from Springer and Gatesy's ASTRAL tree, which placed Columbiformes within Telluraves. In contrast, the resulting Njst tree remains identical to the Njst tree presented in ref. 2.

finding indicates an average of only 1.26 long branches per gene tree. Our analysis revealed a random distribution of these long branches across gene trees. In CDS gene trees, 634 showed long branches exclusively linked to an outgroup; in the remaining 991, long branches were uniformly distributed across ingroup taxa (Fig. 1A). Among intergenic gene trees, eight displayed long branches solely linked to an outgroup; in the remaining 180, long branches were uniformly distributed across ingroup taxa (Fig. 1B). Only one out of the ten intron gene trees exhibited long branches involving non-outgroup species. After excluding 991 CDS and 180 intergenic

gene trees, we reconstructed Njst and ASTRAL trees. The resulting Njst tree remained identical to the tree including all gene trees (Fig. 2A), whereas the ASTRAL tree changed to support Aquaterraves (Fig. 2B). These results underscore Njst's robustness to data heterogeneity, contrasting sharply with ASTRAL's evident sensitivity.

Springer and Gatesy claim that our intergenic segments likely contain a mixture of orthologs and paralogs, citing the identification of CR1-LINE retrolements. However, they present no evidence to demonstrate that the presence of retrolements in intergenic segments adversely affects phylogenetic

analysis. Recent studies indicate that even in the presence of paralogs, species tree inference can be as accurate as analyses based solely on orthologs (6, 7). This was demonstrated by NJst and RAXML trees built exclusively using intergenic segments (see *SI Appendix*, Figs. S1–S5 in ref. 2), which showed close alignment with inferred intraordinal phylogenies from other data types, maintaining congruence with previous studies (8–10). These findings suggest that the purported impact of paralogs on phylogenetic analysis is actually negligible.

Despite extensive filtering and screening, genome-scale data often contains a small number of long branches and paralogs due to its intrinsic heterogeneity (5). However,

Springer and Gatesy overstate their influence on phylogenetic analysis. Their puriest approach, demanding "absolutely clean" data, is unrealistic and would drastically reduce usable data, thus hindering phylogenetic inference (11). Our analysis demonstrates that these factors have minimal impact on phylogenetic analysis and that Springer and Gatesy exaggerate their influence.

ACKNOWLEDGMENTS. This work was supported by funds from the National Natural Science Foundation of China (Grant 32020103005 to S.W. and Y.Q.; Grant 31772441 to S.W. and F.E.R.) and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

1. M. S. Springer, J. Gatesy, A new phylogeny for Aves is compromised by pervasive misalignment and homology problems. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2406494121 (2024).
2. S. Wu *et al.*, Genomes, fossils, and the concurrent rise of modern birds and flowering plants in the Late Cretaceous. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2319696121 (2024).
3. S. Mirarab, T. Warnow, ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* **31**, i44–i52 (2015).
4. P. Vachaspati, T. Warnow, ASTRID: Accurate species tree from internode distances. *BMC Genomics* **16** (suppl. 10), 53 (2015).
5. Y. Du, S. Wu, S. Edwards, L. Liu, The effect of alignment uncertainty, substitution models and priors in building and dating the mammal tree of life. *Bmc Evol. Biol.* **19**, 203 (2019).
6. M. Smith, M. Hahn, The frequency and topology of pseudoorthologs. *Syst. Biol.* **71**, 649–659 (2022).
7. Z. Yan, M. Smith, P. Du, M. Hahn, L. Nakhleh, Species tree inference methods intended to deal with incomplete lineage sorting are robust to the presence of paralogs. *Syst. Biol.* **71**, 367–381 (2022).
8. S. Hackett *et al.*, A phylogenomic study of birds reveals their evolutionary history. *Science* **320**, 1763–1768 (2008).
9. E. D. Jarvis *et al.*, Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320–1331 (2014).
10. R. Prum *et al.*, A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* **526**, 569–573 (2015).
11. M. Smith, D. Vanderpool, M. Hahn, Using all gene families vastly expands data available for phylogenomic inference. *Mol. Biol. Evol.* **39**, msac112 (2022).