



EDITORIAL

THE REPRODUCIBILITY CRISIS IN PHYLOGENETIC ANALYSES

PEDRO L. GODOY, ¹ DANIEL M. CASALI, ² JULIE MEACHEN, ³ ADRIANA LÓPEZ-ARBARELLO, ⁴
MICHAEL D. D'EMIC ^{5*}

¹Department of Zoology, Institute of Biosciences, University of São Paulo, São Paulo, Brazil;

²Department of Biology, Faculty of Philosophy, Sciences, and Letters, University of São Paulo, São Paulo, Brazil;

³Department of Anatomy, Des Moines University, West Des Moines, Iowa, U.S.A.;

⁴Fundación Miguel Lillo, Unidad Ejecutora Lillo (CONICET-FML), San Miguel de Tucumán, Argentina;

⁵Department of Biology, Adelphi University, Garden City, New York, U.S.A.

SUPPLEMENTARY FILE(S)—Supplementary files are available for the article for free at www.tandfonline.com/UJVP.

Citation for this article: Godoy, P. L., Casali, D. M., Meachen, J., López-Arbarello, A., and D'Emic, M. D. (2026) The reproducibility crisis in phylogenetic analyses. *Journal of Vertebrate Paleontology*. <https://doi.org/10.1080/02724634.2026.2638040>

The importance of reproducibility and replicability in science has long been recognized by both empiricists and philosophers (Popper, 1959; Lakatos, 1970; Bishop & Gill, 2020). Though definitions vary by discipline, most authors consider reproducibility as the ability to obtain the same or similar results or conclusions with the same datasets and analytical pipelines, whereas replicability refers to the ability to reinforce results or conclusions with different datasets (Natl. Acad. Sci. Eng. Med., 2019). Nevertheless, for over two decades, investigators in various disciplines have vocalized growing alarm about the dual crises of reproducibility and replicability in scientific research (e.g., Camerer et al., 2018; Ioannidis, 2005; Baker, 2016). Focusing specifically on reproducibility, this crisis has been acknowledged and quantified in different disciplines such as epidemiology (Ioannidis et al., 2005), psychology (Open Science Collaboration, 2015), and metrology (Plant & Hanisch, 2020).

Myriad arguments about the reality of and potential causes of a reproducibility/replicability crisis have been voiced (Bernard, 2023; Fanelli, 2018), including negligence or inaction by journal editors (Allison et al., 2016). A more objective problem concerns a subset of reproducibility, termed computational reproducibility, which is the ability to reproduce the analytical results of a study using its own data and code (Korbmacher et al., 2023). With that in mind, some journals such as *Proceedings of the Royal Society B*, *Journal of Evolutionary Biology*, and the *American Naturalist* have appointed data editors to guarantee the availability of all data and associated materials (e.g., analysis code) linked to a publication (Dunne et al., 2025), but this is not currently the norm.

In the field of evolutionary biology, the *Journal of Vertebrate Paleontology* (JVP) publishes a large number of “data-producing studies” (e.g., descriptions of vertebrate fossil material), which constitute a substantial contribution to large databases and broad-scale analyses (Smith et al. 2024). JVP receives hundreds of submissions (including resubmissions) each year (median annual submissions + resubmissions since 2004 = 353; Fig. 1), with a substantial number including a phylogenetic analysis, providing a rich dataset to explore how reproducible analyses are in the field of vertebrate paleontology. Since mid-2022, the *Journal*

of *Vertebrate Paleontology* has investigated the computational reproducibility of evolutionary studies by appointing first one and then two “Phylogenetics Editors” (PL Godoy and DM Casali) who are tasked with replicating phylogenetic analyses submitted to our journal. These editors attempt to execute and return results identical to those reported by the authors, using parameters described in the manuscript’s methods section and the authors’ submitted data files.

From mid-2022 to early 2026, 298 manuscripts containing a phylogenetic analysis have been received (135 submissions followed by 163 resubmissions), and the computational reproducibility of each has been recorded. This information was recorded for the purposes of manuscript evaluation during the normal course of peer review. All data recorded with submission/resubmission and associated code used to analyze these data are provided in Supplementary Information and deposited at the github link below. Among these, approximately 87% used only parsimony, a method commonly employed in phylogenetic analyses of morphological data. Studies employing a probabilistic method (maximum likelihood and/or Bayesian analysis) in addition to parsimony are much less common, accounting for approximately 12% of the received manuscripts. Less than 1% of the manuscripts solely employed Bayesian inference. When used, probabilistic analyses are usually applied to morphological datasets, although a few studies also employed these methods for combined molecular and morphological datasets.

Alarming, when considering initial submissions, about 26% (35/135) of analyses would not execute, about 64% (87/135) of Methods sections lacked sufficient information about the search procedure(s) employed necessary to reproduce analyses, and for 31% (42/135) we did not obtain the same strict consensus topology that the authors presented in their manuscript (Fig. 2). These percentages improved after resubmission and feedback from a phylogenetics editor, decreasing to approximately 10%, 28%, and 25%, respectively (Fig. 2), until decreasing to zero for accepted articles after a varying number of resubmissions.

Phylogenetics Editors commonly encountered a number of additional issues that did not strictly create a situation of computational reproducibility, but nonetheless hindered reproducibility and/or did not adhere to best practices (Fig. 2). Among all problems detected across initial submission and resubmissions, the most frequent was that the authors did not supply a file with the most parsimonious trees (MPTs) as requested in the

*Corresponding author.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/ujvp.

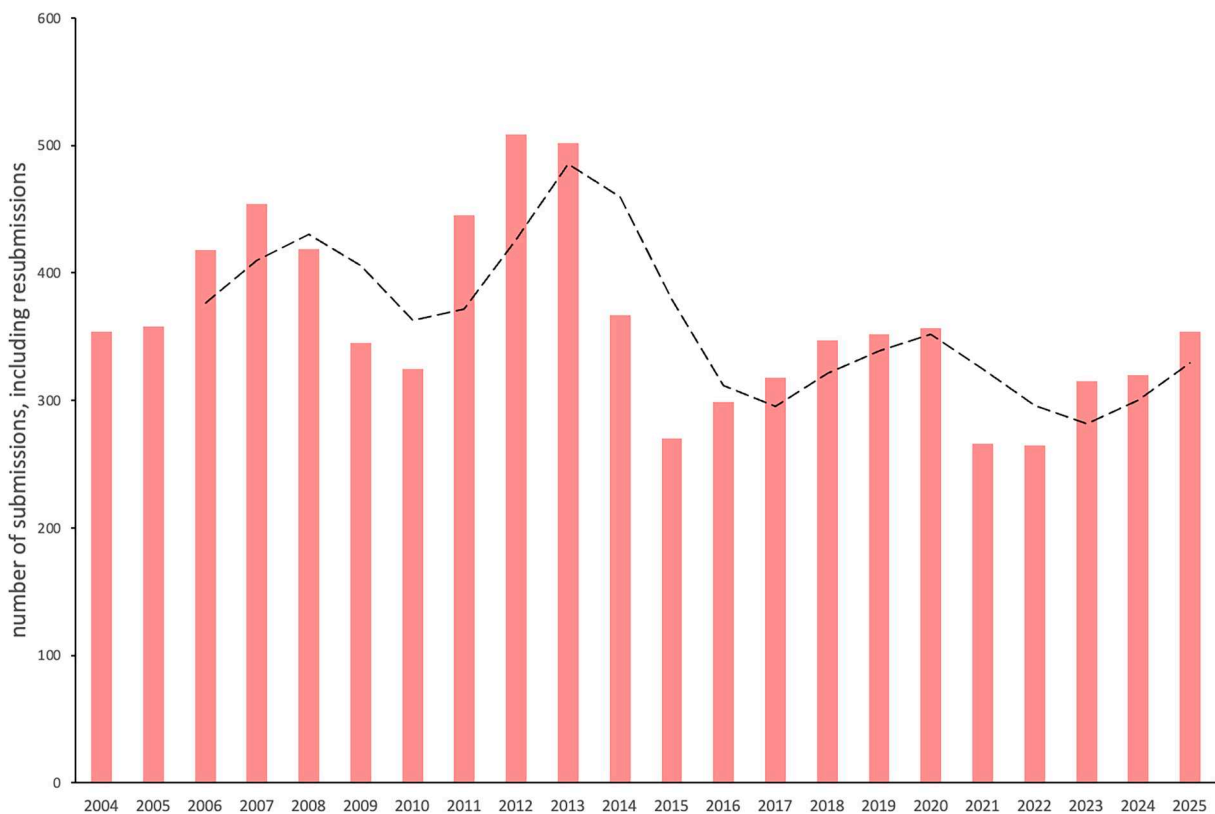


FIGURE 1. Submissions (including resubmissions) to the *Journal of Vertebrate Paleontology* through time. Dashed line shows a three-year moving average.

journal guidelines (~21%). About 11% of problems were related to mismatches between the strict consensus presented by authors and those obtained by phylogenetic reviewers.

Evaluating the temporal trends related to these problems over the last 3.5 years, from the second half of 2022 to the end of 2025, allows us to evaluate the impact of the Phylogenetics Editors on

the community submitting manuscripts to JVP. In these data, there are only trends of improvement regarding the inclusion of character and state names in the matrix and authors providing the file with the MPTs (Fig. 3). Interestingly, these two problems were also the ones that were more prevalent in the second half of 2022, with over 50% of the submissions presenting them (Fig. 3).

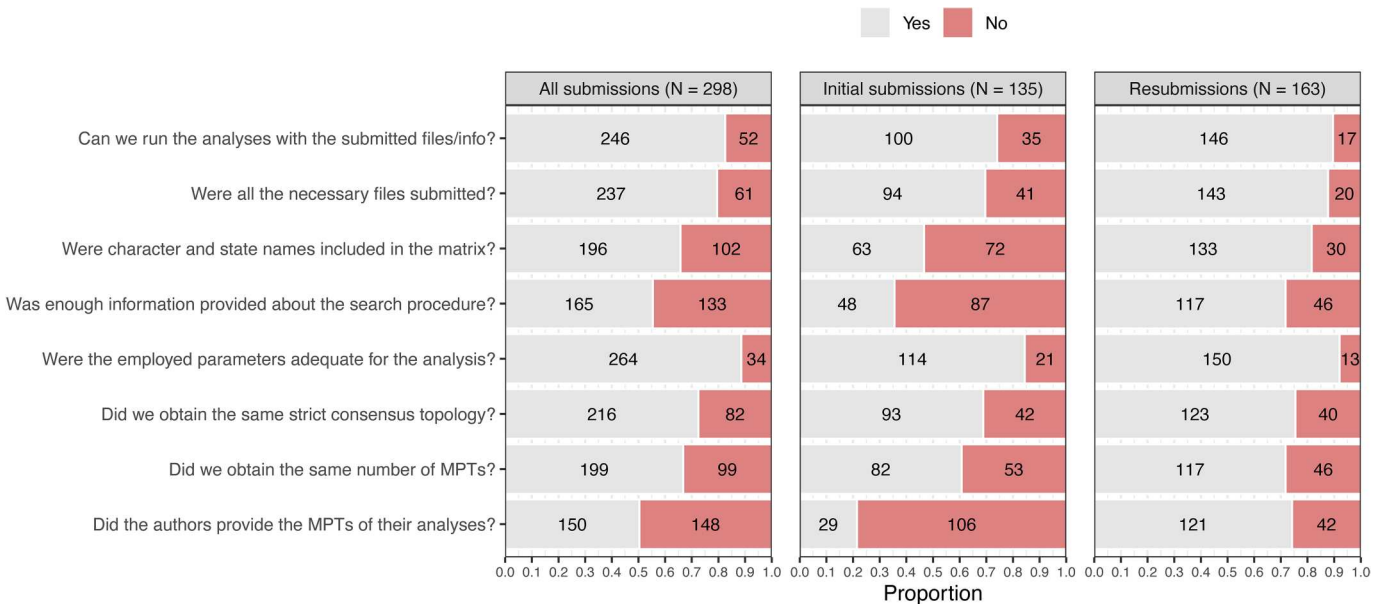


FIGURE 2. Proportion of all submissions, initial submissions, and resubmissions for which a phylogenetics editor encountered problems related to reproducing an analysis for the reasons listed. Gray boxes indicate “yes” answers and red boxes indicate “no” answers.

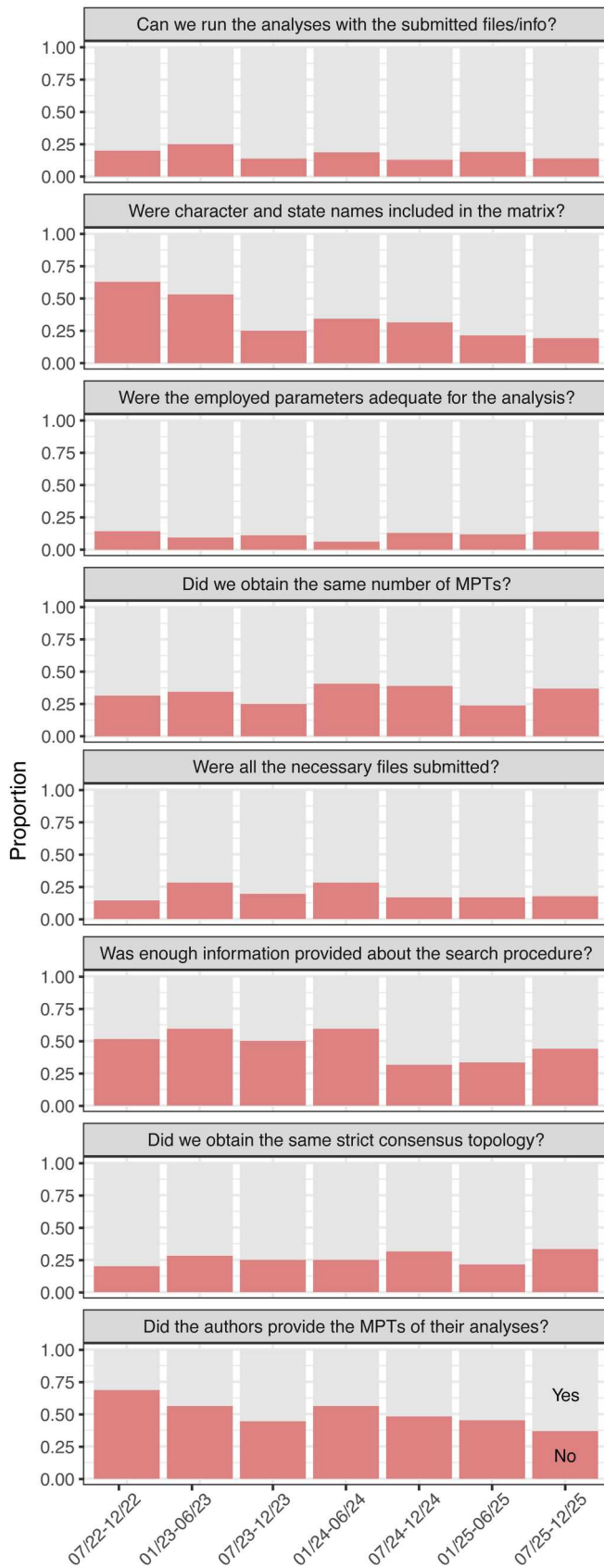


FIGURE 3. Trends in reproducibility issues and adherence to best practices over time. Gray boxes indicate “yes” answers and red boxes indicate “no” answers.

Another problem that showed a similarly high level in the second half of 2022 was the lack of sufficient information about the search procedure(s) employed, which also showed a decreasing trend, albeit a subtle one. The frequency of the other monitored issues fluctuated across these three and a half years around much lower levels (ca. 25%; Fig. 3). The persistence of these issues, which included different results from the analyses, both in terms of the number of MPTs and the strict consensus obtained by authors and reviewers, suggest these may represent hard-to-change practices, and further highlight the importance of the activities of Phylogenetics Editors.

One temporal trend might seem particularly problematic: the persistence of high rates of different consensus topologies reported by the authors in comparison to the ones found by the Phylogenetics Editors (Fig. 3). In contrast to other trends, for which we expect decreased rates over time as a result of the paleontological community becoming aware of the JVP requirements regarding reproducibility of phylogenetic analyses, this particular issue showed an independent pattern (Fig. 3). This reinforces the importance of ensuring reproducibility as well as the positive impact of the Phylogenetics Editors, which were able to flag these issues with the topologies, impeding their propagation to the published literature.

Had the Phylogenetics Editors not undertaken their reviews, it is likely that the majority of these irreproducible and potentially erroneous results would have entered the literature. Most of the analyzed studies employed parsimony; the problem may be even more severe in cases where other approaches are used, such as maximum likelihood (Shen et al., 2020). For molecular analyses, model-building specification code has recently been developed (Drummond et al., 2023), as well as analytical protocols to detect phylogenetic relatedness on a fine scale (Vander-Waal et al., 2024) and an R package to facilitate phylogenetic reproducibility when interacting with GenBank (Román-Palacios, 2023). All of these efforts aim to facilitate and bolster reproducibility and transparency in methodological reporting, but put the burden of effort on the authors rather than editors or reviewers.

In common practice, neither verification that a submitted phylogenetic dataset will execute, nor validation of the analytical output against the results reported in a paper are prerequisite for publication. Instead, the computational reproducibility of phylogenetic analyses is evaluated by a small fraction of peer reviewers or editors on a voluntary basis. While much has been written about reproducibility in fields such as biomedicine (Cobey et al., 2024), social science (Camerer et al., 2018; Altmejd et al., 2019) and various other scientific disciplines (Baker, 2016), to our knowledge there has been little formal study of the potential reproducibility crisis in phylogenetic analysis (but see Shen et al., 2020) or its downstream effects on broader conclusions in evolutionary biology and ecology. Investigation of the computational reproducibility of molecular phylogenies in other journals may reveal similar issues.

To solve this issue, rather than putting the burden on reviewers, adding to an often over-taxed peer review system (Peng et al., 2015), we recommend that more journals onboard a Phylogenetics Editor who can enforce standards regarding the computational reproducibility of phylogenetic analyses. Although mandatory data availability policies are now commonplace in many fields, enforcement varies within and among journals, as does the availability of accompanying code necessary to reproduce results (Culina et al., 2020). By appointing a Phylogenetics Editor to enforce data availability and methodological reporting standards, phylogenetic analysis can avoid some of the pitfalls of the “reproducibility crisis” and move instead towards a “replicability upgrade” (Baker, 2016).

ACKNOWLEDGMENTS

We thank Matthew Warke, Jessica Jacobs, and Jennifer Norman of Taylor and Francis for discussion, data, and logistical support. Thanks to Keegan Melstrom (University of Central Oklahoma) for providing helpful feedback on the manuscript.

ORCID

Pedro L. Godoy  <http://orcid.org/0000-0003-4519-5094>
Daniel M. Casali  <http://orcid.org/0000-0002-0100-3735>
Julie Meachen  <http://orcid.org/0000-0002-2526-2045>
Adriana López-Arbarello  <http://orcid.org/0000-0002-2924-3319>
Michael D. D'Emic  <http://orcid.org/0000-0002-6162-7005>

DATA AVAILABILITY STATEMENT

All data and scripts necessary to reproduce our analysis are provided in Supplementary Information and are reposted at https://github.com/danielmcasali/JVP_reproducibility_plots

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author(s).

LITERATURE CITED

- Allison, D. B., Brown, A. W., George, B. J., & Kaiser, K. A. (2016). Reproducibility: A tragedy of errors. *Nature*, 530(7588), 27–29. doi:10.1038/530027a
- Altmejd, A., Dreber, E., Forsell, J. Huber, T. Imai, M. Johannesson, M. Kirchler, G. Nave, C.F. Camerer, (2019) Predicting the replicability of social science lab experiments. *PLoS ONE* 14(12), e0225826. doi:10.1371/journal.pone.0225826
- Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604), 452–454. doi:10.1038/533452a
- Bernard, C. (2023). Stop reproducing the reproducibility crisis. *Eneuro*, 10(2), ENEURO.0032-23.2023–3. doi:10.1523/ENEURO.0032-23.2023
- Bishop, D., & Gill, E. (2020). Robert Boyle on the importance of reporting and replicating experiments. *Journal of the Royal Society of Medicine*, 113(2), 79–83. doi:10.1177/0141076820902625
- Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., ... Wu, H. (2018). Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour*, 2(9), 637–644. (2018). doi:10.1038/s41562-018-0399-z
- Cobey, K. D., Ebrahimzadeh, S., Page, M. J., Thibault, R. T., Nguyen, P.-Y., Abu-Dalfa, F., et al. (2024). Biomedical researchers' perspectives on the reproducibility of research. *PLoS Biology*, 22(11), e3002870. <https://doi.org/10.1371/journal.Pbio.3002870>
- Culina, A., van den Berg, I., Evans, S., & Sánchez-Tójar, A. (2020). Low availability of code in ecology: A call for urgent action. *PLoS Biology*, 18(7), e3000763. <https://doi.org/10.1371/journal.Pbio.3000763>
- Drummond, A. J., Chen, K., Mendes, F. K., & Xie, D. (2023). LinguaPhylo: A probabilistic model specification language for reproducible phylogenetic analyses. *PLoS Computational Biology*, 19(7), e1011226. <https://doi.org/10.1371/journal.pcbi.1011226>
- Dunne, E. M., Chattopadhyay, D., Dean, C. D., Dillon, E. M., Dowding, E. M., Godoy, P. L., ... & Raja, N. B. (2025). Data equity in paleobiology: progress, challenges, and future outlook. *Paleobiology*, 51(1), 237–249. doi:10.1017/pab.2024.61
- Fanelli, D. (2018). Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences*, 115(11), 2628–2631. doi:10.1073/pnas.1708272114
- Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, 2(8), e124. doi:10.1371/journal.pmed.0020124
- Ioannidis, J.P.A., Bernstein, J., Boffetta, P., Danesh, J., Dolan, S., Hartge, P., ... & Khoury, M. J. (2005). A network of investigator networks in human genome epidemiology. *American Journal of Epidemiology*, 162(4), 302–304. doi:10.1093/aje/kwi201
- Korbmacher, M., Azevedo, F., Pennington, C., Hartmann, H., Pownall, M., Schmidt, K., Elsherif, M., Breznau, N., Robertson, O., Kalandadze, T., Yu, S., Baker, B. J., O'Mahony, A., Olsnes, JØ-S, Shaw, J. J., Gjoneska, B., Yamada, Y., Röer, J. P., Murphy, J., ... Evans, T. (2023). The replication crisis has led to positive structural, procedural, and community changes. *Communications Psychology*, 1(1), 1–13. doi:10.1038/s44271-023-00003-2
- Lakatos, I., & Musgrave, A. (Eds.). (1970). *Criticism and the growth of knowledge: Volume 4: Proceedings of the International Colloquium in the Philosophy of Science, London, 1965*. Cambridge University Press.
- Natl. Acad. Sci. Eng. Med. (2019). *Reproducibility and Replicability in Science*. Natl. Acad. Press.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. doi:10.1126/science.aac4716
- Peng, R. (2015). The reproducibility crisis in science: A statistical counterattack. *Significance*, 12(3), 30–32. doi:10.1111/j.1740-9713.2015.00827.x
- Román-Palacios, C. (2023). The phruta r package: Increasing access, reproducibility and transparency in phylogenetic analyses. *Methods in Ecology and Evolution*, 14(9), 2284–2299. <https://doi.org/10.1111/2041-210X.14147>
- Plant, A., & Hanisch, R. (2020). Issue 2.4, Fall 2020. *Harvard Data Science Review*, 2(4). doi:10.1162/99608f92.eb6ddee4
- Popper, K. R. (1959). *The Logic of Scientific Discovery*. Routledge.
- Shen, X. X., Li, Y., Hittinger, C. T., et al. (2020). An investigation of irreproducibility in maximum likelihood phylogenetic inference. *Nature Communications*, 11(1), 6096. <https://doi.org/10.1038/s41467-020-20005-6>
- Smith, J. A., Raja, N. B., Clements, T., Dimitrijević, D., Dowding, E. M., Dunne, E. M., ... & Kocsis, Á. T. (2024). Increasing the equitability of data citation in paleontology: capacity building for the big data future. *Paleobiology*, 50(2), 165–176. doi:10.1017/pab.2023.33
- VanderWaal, K., Pamornchainavakul, N., Kikuti, M., Linhares, D. C. L., Trevisan, G., Zhang, J., Anderson, T. K., Zeller, M., Rossow, S., Holtkamp, D. J., Makau, D. N., Corzo, C. A., & Paploski, I. A. D. (2024). Phylogenetic-based methods for fine-scale classification of PRRSV-2 ORF5 sequences: a comparison of their robustness and reproducibility. *Frontiers in Virology*, 4, 1433931. doi: 10.3389/fviro.2024.1433931