Running head: PHYLOTOCOL

Title: Phylotocol: Promoting Transparency and Overcoming Bias in Phylogenetics

Authors: Joseph F. Ryan[1,2 †*] and Melissa B. DeBiasse[1,2 †]

[1] Whitney Laboratory for Marine Bioscience, University of Florida, St. Augustine, Florida, United States of America
[2] Department of Biology, University of Florida, Gainesville, Florida, United States of America
† Both authors contributed equally to this work

** Corresponding author: E-mail: joseph.ryan@whitney.ufl.edu.*

Abstract

Research products that lack transparency and are influenced by confirmation bias lead to barriers that, when left unchecked, propagate throughout the scientific record and lead to wasted research effort. Phylogenetics is particularly vulnerable given its ever-evolving methodology and wide choice of options for conducting analyses. Great strides in transparency have been achieved in clinical research by the implementation of *a priori* protocols. Here we propose a similar approach—phylotocol—a straightforward, protocol-driven strategy tailored to the needs of phylogenetic studies. We provide a simple template and offer a flexible range of implementation frameworks, including preregistration options. Besides increasing transparency and accountability, phylotocol has the added benefits of improving study design and reproducibility, enhancing collaboration and education, and increasing the likelihood of project completion. The increased transparency afforded by wide adoption of an *a priori* system like phylotocol would have extensive benefits to science.

Keywords: phylotocol, transparency, phylogenetics, confirmation bias

The importance of reproducibility in science has been written about extensively over the past decade (Ioannidis et al. 2014; Markowetz, 2015; Nosek et al. 2015; Ihle et al. 2017), but its counterpart, transparency, has received considerably less attention. While reproducibility is important to ensure that a study's methodology is sound, without transparency, it is not sufficient to guarantee reliable scientific output (Parker et al. 2016). For example, if confirmation bias leads to the reporting of only a subset of results along with the methods required to generate those results (reporting bias), the study is technically reproducible, but lacks transparency. This is a problem across scientific disciplines, and is particularly important in phylogenetics.

Inferring relationships between genes, genomes, and species is essential for a fundamental understanding of biology. In the nearly 70 years since Hennig formalized phylogenetics (Hennig, 1950; Hennig, 1965), the field has matured immensely through the continuous development and improvement of algorithms, models, and data manipulation strategies (Whelan et al. 2001). This continuous development has led to many advances in phylogenetic methodology. However, the lack of methodological standardization has left open the door to selective reporting driven by confirmation bias.

For most phylogenetic analyses it is possible to justify a multitude of approaches from a seemingly infinite combination of algorithms, models, and data manipulation techniques. Tree reconstruction involves optimization algorithms that can employ a number of different criteria to measure the fit between tree and data including distance, parsimony, maximum likelihood, and Bayesian inference (Felsenstein, 2004). Similarly, model choice is far from standardized. There are a large number of single-matrix models (e.g., JTT) and several criteria to determine the most appropriate (e.g., AIC and BIC) (Page &

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26585v2 | CC BY 4.0 Open Access | rec: 14 May 2018, publ: 14 May 2018

Holmes, 2009). Likewise, partitioning and mixture models, which are increasingly favored over single-matrix approaches, involve a plethora of decisions for which many seemingly sound arguments can be invoked (Blair & Murphy, 2010). Data manipulation, in particular, is far from standardized; common practices include: removing unstable and quickly evolving taxa or genes, using only characters that recover 'known' clades, including only slowly evolving characters, and using conserved amino acid substitutions or indels (Salichos and Rokas, 2013). Finally, this issue is not exclusive to tree reconstruction; other phylogenetic applications (e.g., molecular clock analyses, ancestral state reconstruction, hypothesis testing, and detection of selection) each involve several competing approaches (Baum & Smith, 2012). These myriad choices in the field exacerbate susceptibility to confirmation bias.

During the course of a phylogenetic study, decisions about data analysis are often made haphazardly, motivated by assumptions or rough preliminary results, rather than an *a priori* plan realized at the conception of the project. This strategy can be particularly problematic if decisions are made in response to a result that conflicts with an expected outcome. A common assumption is that unanticipated phylogenetic outcomes are the result of an error in some aspect of the analysis. This can lead to reactionary adjustments in algorithm, model choice, or data manipulation. These types of decisions are problematic because they are greatly influenced by investigators' biases.

In clinical trials, where the outcomes of a study can lead to decisions that put human lives at risk, biases have been explicitly controlled for and transparency and reproducibility ensured through the requirement of *a priori* protocols that outline objective(s), design, methodology, statistical considerations, and study organization (Laine et al. 2007; Zarin

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26585v2 | CC BY 4.0 Open Access | rec: 14 May 2018, publ: 14 May 2018

and Tse, 2013; Zarin et al. 2017). Protocols must be registered to a governmental regulatory agency, funding agency and/or an institutional review board prior to the start of a study. Any changes (amendments) to a protocol require explicit justification and an updated version of the protocol (Getz et al. 2016). Many journals require protocols to be published with clinical trial publications, providing further motivation for their implementation. Protocols greatly reduce, if not eliminate, the potential for researcher bias and in the process ensure the safety of subjects and the integrity of the trial.

Recently, preregistration of research designs has been put forth as a framework for promoting transparency in the fields of Behavioral Ecology (Ihle et al. 2017), Ecology and Evolution (Parker et al 2016), and Psychology (Hartgerink and Wicherts 2016). The proposed measures are largely comparable to protocol registration in clinical trials and provide effective means to promote transparency in each particular field. Responses to these efforts have been positive-(e.g., Blumstein, 2017; Parker and Nakawaga, 2017; Forstmeier, 2017), negative (Koenig, 2017), and mixed (Cockburn, 2017; Hatchwell, 2017). The biggest barrier to widespread adoption to preregistration is the administrative effort associated with its implementation, perceived restrictions on scientific creativity and exploratory analyses, and the perceived potential for project ideas to be scooped. A major challenge for preregistration and similar approaches moving forward is balancing the increase in transparency with the ease of implementation and flexibility.

Here we argue that the field of phylogenetics would benefit tremendously from increased transparency and propose an *a priori* protocol-driven approach—phylotocol—that can be easily incorporated into phylogenetic studies. Below we describe a phylotocol template in detail, propose a set of guidelines for its use, and discuss how it can reduce bias

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26585v2 | CC BY 4.0 Open Access | rec: 14 May 2018, publ: 14 May 2018

and improve transparency and reproducibility in phylogenetics with minimal burdens on researchers' time. We have purposely proposed a loose framework to allow best practices to shape future implementations as more researchers adopt phylotocol.

The primary objectives of phylotocol is to add transparency to phylogenetic studies by front-loading decisions and foster accountability by requiring that changes made during the course of a study are documented and justified. There are also several ancillary benefits associated with phylotocol that we describe below. The goal of this manuscript is to initiate a dialogue about transparency in phylogenetics and present a framework that may help the community to implement research transparency.

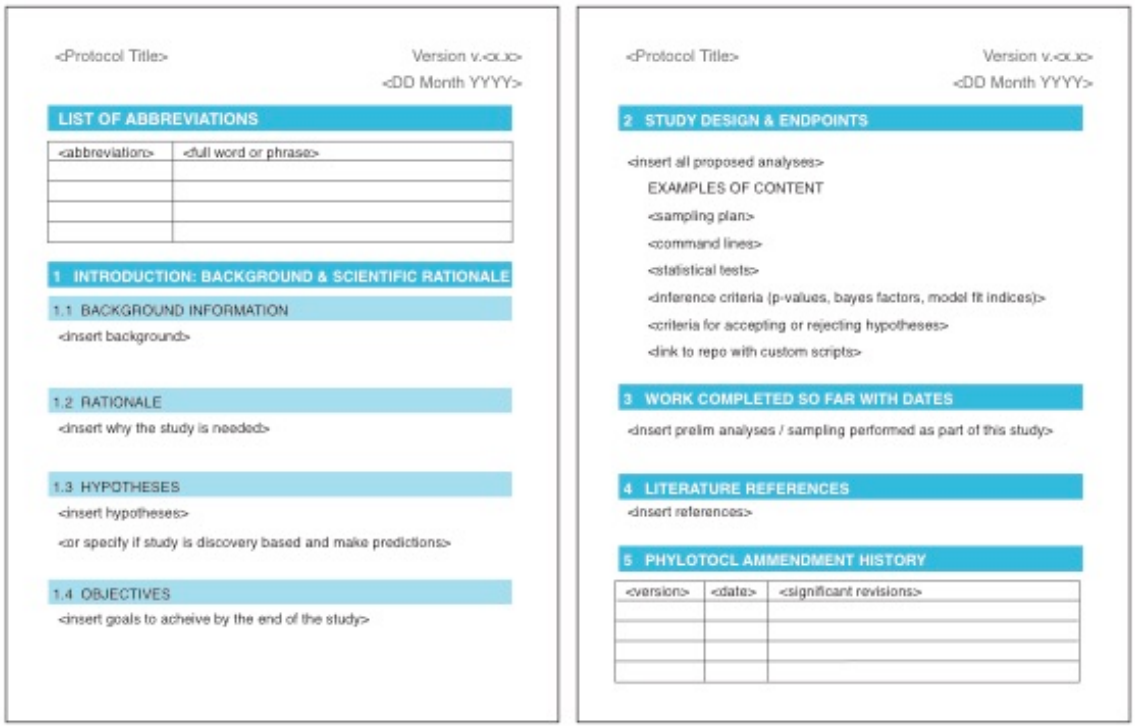

Figure 1: Phylotocol template. Based on the NIH clinical trial protocol, phylotocol layout has been tailored to match the needs of phylogenetic research. The format and the amount of information required are flexible. The figure displays the Microsoft Word version of the template, but there is also a markdown version. A phylotocol can be uploaded to any online repository or used as the basis for preregistration (see Implementation section below).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26585v2 | CC BY 4.0 Open Access | rec: 14 May 2018, publ: 14 May 2018

Anatomy of phylotocol

The template phylotocol is based off the clinical trial protocol established by the National Institutes of Health (Hudson et al. 2016; National Institutes of Health, 2017a; National Institutes of Health, 2017b) and has seven major sections: 1) Title, 2) Abbreviations, 3) Introduction, 4) Study design, 5) Steps completed, 6) References, and 7) Appendix with version history (Fig. 1). While there are certainly other sections that could be added, the minimalist approach of phylotocol aims to reduce unnecessary burden and is therefore an important motivator for potential users. The format is flexible and can be customized to the requirements, preference, and computational expertise of a particular user. We have created versions of the template in both Microsoft Word and markdown formats.

Phylotocol is not a duplication or replacement for a methods section. Instead, it outlines all decisions that could affect the final outcome of a set of studies. Some common decisions include: (1) central hypotheses, (2) which taxa will be included, (3) which methods will be applied, (4) which models will be implemented, and (5) which criteria will be used to validate or reject hypotheses. In addition, many parameter settings (e.g., number of starting trees, seeds used for programs with random processes, minimum occupancy of phylogenomic matrices) can influence the outcome of a study, and therefore should be considered for inclusion in a phylotocol. It is also important to anticipate difficult decisions; for example, when applying different algorithms, models, etc. to the same data matrix, it is imperative to provide explicit criteria for how to evaluate conflicting results.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26585v2 | CC BY 4.0 Open Access | rec: 14 May 2018, publ: 14 May 2018

Primary objectives of phylotocol

Transparency

By specifying the objectives and outlining the full methodology before any analyses are started, the phylotocol promotes transparency and reduces biases on the part of researchers. Once the project is underway, any changes to the analyses are documented in the phylotocol. This ensures that all steps in the analysis pipeline are made available, even those that failed, were replaced by other methods, or that motivated a downstream analysis but were not themselves included in the final manuscript. When decision-making processes are transparent, readers, reviewers, and editors are better able to contextualize, interpret, and evaluate the merits of a study.

Accountability

Transparency generates accountability (Mellor et al. 2018). In phylogenetics, as in other fields, it can be tempting to discount results that conflict with prior assumptions and then perform additional analyses until an expected result is realized. The implementation of phylotocol helps alleviate this temptation by holding researchers publically accountable for all decisions.

Auxiliary benefits of phylotocol

Although the primary goals of phylotocol are transparency and accountability, the process offers a number of additional benefits.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26585v2 | CC BY 4.0 Open Access | rec: 14 May 2018, publ: 14 May 2018

Designing a better study

Outlining each step of a study in a phylotocol *a priori* can bring about a more robust plan. The process of explicitly transcribing procedures and guidelines for the interpretation of results can identify important steps that may have previously been overlooked. In addition, logical flaws in experimental design can be recognized. Catching these obstacles early in the process can lead to huge savings in time and/or money.

Reproducibility

Although a phylotocol need not contain all the details necessary to carry out a study, it can be helpful in producing a highly reproducible set of methods. Unlike in wet-lab based experimental biology, keeping a detailed notebook is less commonplace in phylogenetics. When a phylotocol is implemented, all steps in the analyses are documented, and therefore reported more accurately. Furthermore, at the conclusion of a study, a phylotocol serves as a key reference document for constructing the methods section of a manuscript.

Collaboration

Creating and following a phylotocol can facilitate seamless collaborations among research groups. When designing a project, early drafts of a phylotocol can help with the planning phases. Getting input early from collaborators can strengthen a study while also ensuring that effort between collaborators does not overlap. Listing all steps also allows computational, personnel, budgetary, and/or other resource needs to be assessed. Using phylotocol leads to efficient planning and distribution of effort (e.g., computation allocation) across collaborators. Finalizing the steps for analyses before a project is

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26585v2 | CC BY 4.0 Open Access | rec: 14 May 2018, publ: 14 May 2018

initiated ensures that all members of a team are in agreement and on task, potentially avoiding misunderstandings and/or conflicts down the line.

Education

Phylotocol provides an excellent framework from which to train early career scientists. By making all decisions at the start of a project, it becomes natural to either collectively draft a phylotocol or for students to prepare a first draft and discuss all decisions before the work is initiated. During this process, students gain a deeper understanding of the components of the study and have a roadmap from which to work throughout the project. Likewise, mentors can be sure that effort is focused appropriately.

Previous phylotocols are also useful references for new lab members who want to quickly get up to speed on how the lab performs particular analyses. These previous phylotocols can act as a template from which to start new analyses, particularly when they include command lines for commonly used programs. Phylotocols can easily be incorporated into undergraduate and graduate courses as a tool to teach methodology, the importance of robust experimental design, and to reinforce the concepts of transparency and reproducibility in science.

Project completion

The inherent open-endedness of scientific endeavors can often be intimidating and create a barrier to project completion. Implementing a phylotocol can remove this barrier by providing explicit starting and stopping points for a project and the motivation to complete the study as planned. The phylotocol quantifies the number of objectives a

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26585v2 | CC BY 4.0 Open Access | rec: 14 May 2018, publ: 14 May 2018

project requires and helps researchers prioritize each step. Beginning and completing a manuscript for the project will also be less daunting because the background information, study justification, methods, and references will already be compiled in the phylotocol. Starting new projects hinders the ability to complete existing projects; a phylotocol serves as a gentle impediment to spontaneously starting tangential projects and therefore increases productivity.

Implementation

Scientists can use several strategies to implement phylotocol. The simplest is to create a document on one's computer at the beginning of a study and use it as a private guide for the study. This can be incredibly powerful, but does not maximize transparency since others have to trust that the analyses were planned *a priori*. Another strategy is to post a phylotocol to a public data repository (e.g., FigShare, Zenodo) or a software repository (e.g., GitHub) before beginning a study. An advantage of online data repositories is that a digital object identifier (DOI) can be issued for a phylotocol (versioning is also available for phylotocol updates). These platforms also provide free private space that can later be made public with a timestamp. A development platform like GitHub includes timestamps and makes updates to a phylotocol a more natural process (especially with the markdown version of phylotocol), but GitHub currently charges for private space and does not include the built-in ability to generate a DOI (although this can be achieved in conjunction with Zenodo or FigShare).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26585v2 | CC BY 4.0 Open Access | rec: 14 May 2018, publ: 14 May 2018

Another option is to post a phylotocol within the Open Science Framework (OSF) (Nosek et al. 2015). This platform has a specific interface for preregistration; we recommend choosing the "Open-Ended Registration" option and pasting a text version of phylotocol into the box. Preregistration includes advantages in addition to those provided in other repositories. The OSF registry has a built-in embargo system allowing a registered project to remain private for up to four years. The OSF interface allows users to connect registrations to workflow management tools (e.g., Dataverse, Dropbox, figshare, Github, and others, see: http://help.osf.io/m/addons), so that work from disparate members of a research team can be connected, persistently stored, and cited in one location. When it is ready to disseminate early findings, any file on the OSF can be shared as a preprint (https://osf.io/preprints) prior to formal publication in a journal.

Some journals have started embracing preregistration. For example, *BMC Ecology* has started a publication format called "Registered Report." In this model, submission is a two-stage process where authors propose a study in the first submission, and if a preregistration is approved, the journal agrees to publish the results, regardless of the outcome. This accomplishes the aims of preregistration, and also ensures that results are published whether or not they confirm a hypothesis.

Discussion

Motivations for Individual Researchers

The ultimate goal of researchers is to make discoveries and formulate theories that stand up to rigorous testing, and eventually become widely accepted as truth. The

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26585v2 | CC BY 4.0 Open Access | rec: 14 May 2018, publ: 14 May 2018

suspicion of bias, especially for controversial topics where two or more research groups report conflicting results, impedes this goal. By adopting transparent practices, researchers reduce the perception of manipulation and help cultivate confidence in the reliability and robustness of their science. In this way, transparent research practices like phylotocol help maximize research impact.

Implications for Science

Two main trajectories in science are building upon prior knowledge and overturning existing perspectives. When inaccuracies in the scientific record are due to confirmation bias, the progress of science is impeded. First, research built upon published inaccuracies is typically doomed from the outset. Likewise, overturning false conclusions requires considerable work, reducing time and funds that can be applied to forward-looking efforts. It is difficult to estimate how much of the scientific record is inaccurate due to confirmation bias in phylogenetics, but it in the absence of transparency, the potential for bias is enormous, as is the potential for falsities and wasted research effort. As such, efforts like phylotocol to decrease bias in phylogenetics could potentially lead to greater productivity in our field by reducing superfluous research.

Conclusion

Phylotocol is a powerful tool to increase transparency and accountability in phylogenetics. It has great potential to improve how phylogenetic research is conducted, interpreted, communicated, and perceived. The implementation is straightforward and offers a range of auxiliary benefits, including making contributions to study design,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26585v2 | CC BY 4.0 Open Access | rec: 14 May 2018, publ: 14 May 2018

reproducibility, collaboration, and education. Phylotocol can bolster scientific productivity both at the level of the individual researcher as well as in the broader context of the scientific record. While phylotocol is a simple idea, its repercussions could be far reaching if widely implemented.

Acknowledgements

This material is based in part upon work supported by the National Science Foundation under Grant Number 1542597. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank Malika Ihle, Jessica Whelpley, Daniel Sasson, David Mellor, Tim Parker, Lyndon Coghill, Daniel Cooke, Jacob Esselstyn, Teisha King, Laura Lagomarsino, Rafael Marcondes, Genevieve Mount, Zachary Rodriguez, and Mark Swanson, for comments on earlier versions of the manuscript. The views expressed in this paper do not necessarily reflect the views of those acknowledged.

References:

Baum DA, Smith SD. Tree thinking: an introduction to phylogenetic biology. InTree thinking: an introduction to phylogenetic biology 2012.

Blair C, Murphy RW. Recent trends in molecular phylogenetic analysis: where to next? Journal of Heredity. 2010 Aug 8;102(1):130-8.

Blumstein DT. Research credibility: the devil is in the details: a comment on Ihle et al. Behavioral Ecology. 2017;28(2):355-.

Cockburn A. Long-term data as infrastructure: a comment on Ihle et al. Behavioral Ecology. 2017 Apr 1;28(2):357-.

Felsenstein J, Felenstein J. Inferring phylogenies. Sunderland, MA: Sinauer associates; 2004 Jan.

Forstmeier W. Preregister now for an upgrade to Behavioral Ecology 2.0: a comment on Ihle et al. Behavioral Ecology. 2017 Apr 1;28(2):358-9.

Getz KA, Stergiopoulos S, Short M, Surgeon L, Krauss R, Pretorius S, Desmond J, Dunn D. The impact of protocol amendments on clinical trial performance and cost. Therapeutic Innovation & Regulatory Science. 2016 Jul;50(4):436-41.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26585v2 | CC BY 4.0 Open Access | rec: 14 May 2018, publ: 14 May 2018

Hartgerink, Chris HJ, and Jelte M. Wicherts. "Research practices and assessment of research misconduct." *ScienceOpen Research* (2016).

Hatchwell BJ. Replication in behavioural ecology: a comment on Ihle et al. Behavioral Ecology. 2017 Apr 1;28(2):360-.

Hennig W. Grundzuge einer Theorie der phylogenetischen Systematik. 1950.

Hennig W. Phylogenetic systematics. Annual review of entomology. 1965 Jan;10(1):97-116.

Hudson KL, Lauer MS, Collins FS. Toward a new era of trust and transparency in clinical trials. Jama. 2016 Oct 4;316(13):1353-4.

Ihle, M., Winney, I.S., Krystalli, A. and Croucher, M., 2017. Striving for transparent and credible research: practical guidelines for behavioral ecologists. *Behavioral Ecology*, *28*(2), pp.348-354.

Ioannidis, J.P., 2014. How to make more published research true. *PLoS medicine*, *11*(10), p.e1001747.

Christine Laine, M.D., M.P.H.Richard Horton, F.Med.Sci.Catherine D. DeAngelis, M.D., M.P.H.Jeffrey M. Drazen, M.D.Frank A. Frizelle, M.B., Ch.B., M.Med.Sc.Fiona Godlee, M.B.,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26585v2 | CC BY 4.0 Open Access | rec: 14 May 2018, publ: 14 May 2018

B.Chir., B.Sc.Charlotte Haug, M.D., Ph.D., M.Sc.Paul C. Hébert, M.D., M.H.Sc.Sheldon Kotzin, M.L.S.Ana Marusic, M.D., Ph.D.Peush Sahni, M.S., Ph.D.Torben V. Schroeder, M.D., D.M.Sc.Harold C. Sox, M.D.Martin B. Van Der Weyden, M.D.Freek W.A. Verheugt, M.D. Clinical trial registration—looking back and moving ahead. New England Journal of Medicine. 2007 June 28; 2734-2736.

Koenig WD. Striving for science that is transparent, credible—and enjoyable: a comment on Ihle et al. Behavioral Ecology. 2017 Apr 1;28(2):358-.

Markowetz, F., 2015. Five selfish reasons to work reproducibly. *Genome biology*, *16*(1), p.274.

Mellor D, Vazire S, StephenLindsay D. Transparent science: A more credible, reproducible, and publishable way to do science.

Nosek, B.A., Alter, G., Banks, G.C., Borsboom, D., Bowman, S.D., Breckler, S.J., Buck, S., Chambers, C.D., Chin, G., Christensen, G. and Contestabile, M., 2015. Promoting an open research culture. *Science*, *348*(6242), pp.1422-1425.

National Institutes of Health. Word version of final clinical trials protocol. http://osp.od.nih.gov/wp-content/uploads/Protocol-Template-Version-1.0-040717.docx Accessed August 24, 2017a

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26585v2 | CC BY 4.0 Open Access | rec: 14 May 2018, publ: 14 May 2018

National Institutes of Health. NIH and FDA Request for Public Comment on Draft Clinical Trial Protocol Template for Phase 2 and 3 IND/IDE Studies. http://grants.nih.gov/grants/guide/notice -files/NOT-OD-16-043.html Accessed August 24, 2017b.

Page RD, Holmes EC. Molecular evolution: a phylogenetic approach. John Wiley & Sons; 2009 Jul 14.

Parker TH, Forstmeier W, Koricheva J, Fidler F, Hadfield JD, Chee YE, Kelly CD, Gurevitch J, Nakagawa S. Transparency in ecology and evolution: real problems, real solutions. Trends in ecology & evolution. 2016 Sep 30;31(9):711-9.

Parker TH, Nakagawa S. Practical models for publishing replications in behavioral ecology: a comment on Ihle et al. Behavioral Ecology. 2017 Apr 1;28(2):355-7.

Salichos L, Rokas A. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature. 2013 May 16;497(7449):327.

Whelan S, Liò P, Goldman N. Molecular phylogenetics: state-of-the-art methods for looking into the past. TRENDS in Genetics. 2001 May 1;17(5):262-72.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26585v2 | CC BY 4.0 Open Access | rec: 14 May 2018, publ: 14 May 2018

400 Zarin DA, Tse T. Trust but verify: trial registration and determining fidelity to the protocol.
401 Annals of internal medicine. 2013 Jul 2;159(1):65-7.
402
403 Zarin DA, Tse T, Williams RJ, Rajakannan T. Update on trial registration 11 years after the
404 ICMJE policy was established. New England Journal of Medicine. 2017 Jan
405 26;376(4):383-91.
406

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26585v2 | CC BY 4.0 Open Access | rec: 14 May 2018, publ: 14 May 2018