1   Running head: PHYLOTOCOL
2
3   Title: Phylotocol: Promoting Transparency and Overcoming Bias in Phylogenetics
4
5   Authors: Melissa B. DeBiasse[1,2] and Joseph F. Ryan[1,2] *
6
7   [1] Whitney Laboratory for Marine Bioscience, University of Florida, St. Augustine, Florida,
8   United States of America
9   [2] Department of Biology, University of Florida, Gainesville, Florida, United States of America
10
11   *Corresponding author: E-mail: joseph.ryan@whitney.ufl.edu.*
12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30   **Abstract**

31

32   Research products that lack transparency and are influenced by confirmation bias lead to

33   barriers that, when left unchecked, propagate throughout the scientific record and lead to wasted

34   research effort. Phylogenetics is particularly vulnerable given its ever-evolving methodology and

35   wide choice of options for conducting analyses. Great strides in transparency have been achieved

36   in clinical research by the implementation of *a priori* protocols. Here we propose a similar

37   approach—phylotocol—a straightforward, protocol-driven strategy tailored to the needs of

38   phylogenetic studies. We provide a simple template and offer a flexible range of implementation

39   frameworks, including preregistration options. Besides increasing transparency and

40   accountability, phylotocol has the added benefits of improving study design and reproducibility,

41   enhancing collaboration and education, and increasing the likelihood of project completion. The

42   increased transparency afforded by wide adoption of an *a priori* system like phylotocol would

43   have extensive benefits to science.

44

45   **Keywords:** phylotocol, transparency, phylogenetics, confirmation bias

46

47

2

48   The importance of reproducibility in science has been written about extensively over the

49   past decade (Markowetz, 2015; Baker, 2016), but its counterpart, transparency, has received

50   considerably less attention (Nosek et al. 2015; Ihle et al. 2017). While reproducibility is

51   important to ensure that a study's methodology is sound, without transparency, reproducibility is

52   not sufficient to guarantee reliable scientific output (Parker et al. 2016). For example, if

53   confirmation bias leads to the reporting of only a subset of results along with the methods

54   required to generate those results (reporting bias), the study is technically reproducible, but lacks

55   transparency. This is a problem across scientific disciplines, and is particularly important in

56   phylogenetics.

57   Inferring relationships between genes, genomes, and species is essential for a

58   fundamental understanding of biology. In the nearly 70 years since Hennig formalized

59   phylogenetics (Hennig, 1950; Hennig, 1965), the field has matured immensely through the

60   continuous development and improvement of algorithms, models, and data manipulation

61   strategies (Whelan et al. 2001). This continuous development has led to many advances in

62   phylogenetic methodology. However, the lack of methodological standardization has left open

63   the door to selective reporting driven by confirmation bias.

64   The current state of phylogenetics makes it possible to justify a multitude of approaches

65   from a seemingly infinite combination of algorithms, models, and data manipulation techniques.

66   In many cases, widely used methods have been shown to either outperform or underperform

67   depending on the context without a deep understanding of why (e.g. maximum-likelihood

68   methods (Zhou et al. 2018) and multi-species coalescent based methods (Mirarab et al. 2014;

69   Chou et al. 2015)). Similarly, choosing between models of sequence evolution (e.g. general time-

70   reversible models or models with fixed versus site-heterogeneous matrices) can be equally

3

71   complex to rationalize. There are many conflicting opinions regarding decisions related to

72   inclusion/exclusion of ingroup and outgroup taxa and characters. Several common practices (e.g.

73   removal of quickly evolving genes or taxa, removal of sites containing gaps) have been shown to

74   have negative impact (Salichos and Rokas, 2013). On the other hand, recent research has shown

75   that a small number of outlier genes can have a large effect on phylogenetic inference (Shen et

76   al. 2017; Walker et al. 2018), in which case removing data might be justified.

77   The deluge of method choices is not limited to tree inference. Once a phylogeny is

78   inferred, there are many methods for reconciling conflicting topologies through hypothesis

79   testing (e.g. Goldman et al. 2000) or testing the fit of a tree to its underlying data (e.g. Reid et al

80   2013). Likewise, downstream analyses that use trees to ask evolutionary and ecological questions

81   (e.g. ancestral state reconstruction, divergence dating, branch specific tests for selection) all have

82   their own suite of methods and parameters on which researchers must decide. These complex

83   choices at every step of the process across phylogenetics exacerbate our susceptibility to

84   confirmation bias.

85   During the course of a phylogenetic study, decisions about data analysis are often made

86   haphazardly, motivated by assumptions or rough preliminary results, rather than an *a priori* plan

87   realized at the conception of the project. This strategy can be particularly problematic if

88   decisions are made in response to a result that conflicts with an expected outcome. A common

89   assumption is that unanticipated phylogenetic outcomes are the result of an error in some aspect

90   of the analysis, rather than a real biological pattern. This can lead to reactionary adjustments in

91   algorithm, model choice, or data manipulation. These types of decisions are problematic because

92   they are greatly influenced by investigators' biases.

4

93   In clinical trials, where the outcomes of a study can lead to decisions that put human lives

94   at risk, biases have been explicitly controlled for, and transparency and reproducibility ensured,

95   through the requirement of *a priori* protocols that outline objective(s), design, methodology,

96   statistical considerations, and study organization (Laine et al. 2007; Zarin and Tse, 2013; Zarin et

97   al. 2017). Protocols must be registered to a governmental regulatory agency, funding agency

98   and/or an institutional review board prior to the start of a study. Any changes (amendments) to a

99   protocol require explicit justification and an updated version of the protocol (Getz et al. 2016).

100   Many journals require protocols to be published with clinical trial publications, providing further

101   motivation for their implementation. Protocols greatly reduce, if not eliminate, the potential for

102   researcher bias and in the process ensure the safety of subjects and the integrity of the trial.

103   Recently, preregistration of research designs has been put forth as a framework for

104   promoting transparency in the fields of Behavioral Ecology (Ihle et al. 2017), Ecology and

105   Evolution (Parker et al 2016), and Psychology (Hartgerink and Wicherts 2016). The proposed

106   measures are largely comparable to protocol registration in clinical trials and provide effective

107   means to promote transparency in each particular field. Responses to these efforts have been

108   positive (e.g., Blumstein, 2017; Parker and Nakawaga, 2017; Forstmeier, 2017), negative

109   (Koenig, 2017), and mixed (Cockburn, 2017; Hatchwell, 2017). The biggest barrier to

110   widespread adoption to preregistration is the administrative effort associated with its

111   implementation, perceived restrictions on scientific creativity and exploratory analyses, and the

112   perceived potential for project ideas to be scooped. A major challenge for preregistration and

113   similar approaches moving forward is balancing the increase in transparency with the ease of

114   implementation and flexibility.

5

115   Here we argue that the field of phylogenetics would benefit tremendously from increased

116   transparency and propose an *a priori* protocol-driven approach—phylotocol—that can be easily

117   incorporated into phylogenetic studies. Below we describe a phylotocol template in detail,

118   propose a set of guidelines for its use, and discuss how it can reduce bias and improve

119   transparency and reproducibility in phylogenetics with minimal burdens on researchers' time.

120   We have purposely proposed a loose framework to allow best practices to shape future

121   implementations as more researchers adopt phylotocol.

122   The primary objectives of phylotocol is to add transparency to phylogenetic studies by

123   front-loading decisions and foster accountability by requiring that changes made during the

124   course of a study are documented and justified. There are also several ancillary benefits

125   associated with phylotocol that we describe below. The goal of this manuscript is to initiate a

126   dialogue about transparency in phylogenetics and present a framework that may help the

127   community to implement research transparency.

128

129

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).

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

7

145   computational expertise of a particular user. We have created versions of the template in both

146   Microsoft Word and markdown formats.

147   Phylotocol is not a duplication or replacement for a methods section. Instead, it outlines

148   all decisions that could affect the final outcome of a set of studies. Some common decisions

149   include: (1) central hypotheses, (2) which taxa will be included, (3) which methods will be

150   applied, (4) which models will be implemented, and (5) which criteria will be used to validate or

151   reject hypotheses. In addition, many parameter settings (e.g., number of starting trees, seeds used

152   for programs with random processes, minimum occupancy of phylogenomic matrices) can

153   influence the outcome of a study, and therefore should be considered for inclusion in a

154   phylotocol. It is also important to anticipate difficult decisions; for example, when applying

155   different algorithms, models, etc. to the same data matrix, it is imperative to provide explicit

156   criteria for how to evaluate conflicting results.

157

158   **Primary objectives of phylotocol**

159

160   *Transparency*

161   By specifying the objectives and outlining the full methodology before any analyses are

162   started, the phylotocol promotes transparency and reduces biases on the part of researchers. Once

163   the project is underway, any changes to the analyses are documented in the phylotocol. This

164   ensures that all steps in the analysis pipeline are made available, even those that failed, were

165   replaced by other methods, or that motivated a downstream analysis but were not themselves

166   included in the final manuscript. When decision-making processes are transparent, readers,

8

167   reviewers, and editors are better able to contextualize, interpret, and evaluate the merits of a

168   study.

169

170   *Accountability*

171   Transparency generates accountability (Mellor et al. 2018). In phylogenetics, as in other

172   fields, it can be tempting to discount results that conflict with prior assumptions and then

173   perform additional analyses until an expected result is realized. The implementation of

174   phylotocol helps alleviate this temptation by holding researchers publically accountable for all

175   decisions.

176

177   **Auxiliary benefits of phylotocol**

178

179   Although the primary goals of phylotocol are transparency and accountability, the

180   process offers a number of additional benefits.

181

182   *Designing a better study*

183   Outlining each step of a study in a phylotocol *a priori* can bring about a more robust plan.

184   The process of explicitly transcribing procedures and guidelines for the interpretation of results

185   can identify important steps that may have previously been overlooked. In addition, logical flaws

186   in experimental design can be recognized. Catching these obstacles early in the process can lead

187   to huge savings in time and/or money.

188

189   *Reproducibility*

9

190   Although a phylotocol need not contain all the details necessary to carry out a study, it

191   can be helpful in producing a highly reproducible set of methods. Unlike in wet-lab based

192   experimental biology, keeping a detailed notebook is less commonplace in phylogenetics. When

193   a phylotocol is implemented, all steps in the analyses are documented, and therefore reported

194   more accurately. Furthermore, at the conclusion of a study, a phylotocol serves as a key

195   reference document for constructing the methods section of a manuscript.

196

197   *Collaboration*

198   Creating and following a phylotocol can facilitate seamless collaborations among

199   research groups. When designing a project, early drafts of a phylotocol can help with the

200   planning phases. Getting input early from collaborators can strengthen a study while also

201   ensuring that effort between collaborators does not overlap. Listing all steps also allows

202   computational, personnel, budgetary, and/or other resource needs to be assessed. Using

203   phylotocol leads to efficient planning and distribution of effort (e.g., computation allocation)

204   across collaborators. Finalizing the steps for analyses before a project is initiated ensures that all

205   members of a team are in agreement and on task, potentially avoiding misunderstandings and/or

206   conflicts down the line.

207

208   *Education*

209   Phylotocol provides an excellent framework from which to train early career scientists.

210   By making all decisions at the start of a project, it becomes natural to either collectively draft a

211   phylotocol or for students to prepare a first draft and discuss with their mentor all decisions

212   before the work is initiated. During this process, students gain a deeper understanding of the

10

213 components of the study and have a roadmap from which to work throughout the project.

214 Likewise, mentors can be sure that effort is focused appropriately.

215 Previous phylotocols are also useful references for new lab members who want to quickly

216 get up to speed on how the lab performs particular analyses. These previous phylotocols can act

217 as a template from which to start new analyses, particularly when they include command lines

218 for commonly used programs. Phylotocols can easily be incorporated into undergraduate and

219 graduate courses as a tool to teach methodology, the importance of robust experimental design,

220 and to reinforce the concepts of transparency and reproducibility in science.

221

222 *Project completion*

223 The inherent open-endedness of scientific endeavors can often be intimidating and create

224 a barrier to project completion. Implementing a phylotocol can remove this barrier by providing

225 explicit starting and stopping points for a project and the motivation to complete the study as

226 planned. The phylotocol quantifies the number of objectives a project requires and helps

227 researchers prioritize each step. Beginning and completing a manuscript for the project will also

228 be less daunting because the background information, study justification, methods, and

229 references will already be compiled in the phylotocol. Starting new projects hinders the ability to

230 complete existing projects; a phylotocol serves as a gentle impediment to spontaneously starting

231 tangential projects and therefore increases productivity.

232

233 **Implementation**

234

11

235 Scientists can use several strategies to implement phylotocol. The simplest is to create a

236 document on one's computer at the beginning of a study and use it as a private guide for the

237 study. This can be incredibly powerful, but does not maximize transparency since others have to

238 trust that the analyses were planned *a priori*. Another strategy is to post a phylotocol to a public

239 data repository or a software repository before beginning a study. An advantage of online data

240 repositories is that a digital object identifier (DOI) can be issued for a phylotocol (versioning is

241 also available for phylotocol updates). Many platforms also provide free private space that can

242 later be made public with a timestamp. Repositories that incorporate a version control system

243 will generate timestamps and make updates to a phylotocol a more natural process (especially

244 with the markdown version of phylotocol).

245 Another option is to post a phylotocol within the Open Science Framework (OSF) (Nosek

246 et al. 2015). This platform has a specific interface for preregistration; we recommend choosing

247 the "Open-Ended Registration" option and pasting a text version of phylotocol into the box.

248 Preregistration includes advantages in addition to those provided in other repositories. The OSF

249 registry has a built-in embargo system allowing a registered project to remain private for up to

250 four years. The OSF interface allows users to connect registrations to workflow management

251 tools (see: http://help.osf.io/m/addons), so that work from disparate members of a research team

252 can be connected, persistently stored, and cited in one location. When it is ready to disseminate

253 early findings, any file on the OSF can be shared as a preprint (https://osf.io/preprints) prior to

254 formal publication in a journal.

255 Some journals have started embracing preregistration. For example, *BMC Ecology* has

256 started a publication format called "Registered Report." In this model, submission is a two-stage

257 process where authors propose a study in the first submission, and if a preregistration is

12

258 approved, the journal agrees to publish the results, regardless of the outcome. This accomplishes

259 the aims of preregistration, and also ensures that results are published whether or not they

260 confirm a hypothesis.

261

262 **Discussion**

263

264 *Motivations for Individual Researchers*

265   The ultimate goal of researchers is to make discoveries and formulate theories that stand

266 up to rigorous testing, and eventually become widely accepted as truth. The suspicion of bias,

267 especially for controversial topics where two or more research groups report conflicting results,

268 impedes this goal. By adopting transparent practices, researchers reduce the perception of

269 manipulation and help cultivate confidence in the reliability and robustness of their science. In

270 this way, transparent research practices like phylotocol help maximize research impact.

271

272 *Implications for Science*

273   Two main trajectories in science are building upon prior knowledge and overturning

274 existing perspectives. When inaccuracies in the scientific record are due to confirmation bias, the

275 progress of science is impeded. First, research built upon published inaccuracies is typically

276 doomed from the outset. Likewise, overturning false conclusions requires considerable work,

277 reducing time and funds that can be applied to forward-looking efforts. It is difficult to estimate

278 how much of the scientific record is inaccurate due to confirmation bias in phylogenetics, but it

279 in the absence of transparency, the potential for bias is enormous, as is the potential for falsities

280 and wasted research effort (Ioannidis, 2014). As such, efforts like phylotocol to decrease bias in

13

281 phylogenetics could potentially lead to greater productivity in our field by reducing superfluous

282 research.

283

284 *Conclusion*

285   Phylotocol is a powerful tool to increase transparency and accountability in

286 phylogenetics. It has great potential to improve how phylogenetic research is conducted,

287 interpreted, communicated, and perceived. The implementation is straightforward and offers a

288 range of auxiliary benefits, including making contributions to study design, reproducibility,

289 collaboration, and education. Phylotocol can bolster scientific productivity both at the level of

290 the individual researcher as well as in the broader context of the scientific record. While

291 phylotocol is a simple idea, its repercussions could be far reaching if widely implemented.

292

293 **Acknowledgements**

14

304 **References:**

305

306 Baker M. 1,500 scientists lift the lid on reproducibility. Nature News. 2016 May

307   26;533(7604):452.

308

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

310   introduction to phylogenetic biology 2012.

311

312 Blair C, Murphy RW. Recent trends in molecular phylogenetic analysis: where to next? Journal

313   of Heredity. 2010 Aug 8;102(1):130-8.

314

315 Blumstein DT. Research credibility: the devil is in the details: a comment on Ihle et al.

316   Behavioral Ecology. 2017;28(2):355-.

317

318 Cockburn A. Long-term data as infrastructure: a comment on Ihle et al. Behavioral Ecology.

319   2017 Apr 1;28(2):357-.

320

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

322

323 Forstmeier W. Preregister now for an upgrade to Behavioral Ecology 2.0: a comment on Ihle et

324   al. Behavioral Ecology. 2017 Apr 1;28(2):358-9.

325

15

326   Getz KA, Stergiopoulos S, Short M, Surgeon L, Krauss R, Pretorius S, Desmond J, Dunn D. The

327   impact of protocol amendments on clinical trial performance and cost. Therapeutic

328   Innovation & Regulatory Science. 2016 Jul;50(4):436-41.

329

330   Goldman N, Anderson JP, Rodrigo AG. Likelihood-based tests of topologies in phylogenetics.

331   Systematic biology. 2000 Dec 1;49(4):652-70.

332

333   Hartgerink, Chris HJ, and Jelte M. Wicherts. "Research practices and assessment of research

334   misconduct." *ScienceOpen Research* (2016).

335

336   Hatchwell BJ. Replication in behavioural ecology: a comment on Ihle et al. Behavioral Ecology.

337   2017 Apr 1;28(2):360-.

338

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

340

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

342

343   Hudson KL, Lauer MS, Collins FS. Toward a new era of trust and transparency in clinical trials.

344   Jama. 2016 Oct 4;316(13):1353-4.

345

346   Ihle, M., Winney, I.S., Krystalli, A. and Croucher, M., 2017. Striving for transparent and

347   credible research: practical guidelines for behavioral ecologists. *Behavioral Ecology*, *28*(2),

348   pp.348-354.

16

349

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

351   p.e1001747.

352

353 Christine Laine, M.D., M.P.H.Richard Horton, F.Med.Sci.Catherine D. DeAngelis, M.D.,

354   M.P.H.Jeffrey M. Drazen, M.D.Frank A. Frizelle, M.B., Ch.B., M.Med.Sc.Fiona Godlee,

355   M.B., B.Chir., B.Sc.Charlotte Haug, M.D., Ph.D., M.Sc.Paul C. Hébert, M.D.,

356   M.H.Sc.Sheldon Kotzin, M.L.S.Ana Marusic, M.D., Ph.D.Peush Sahni, M.S., Ph.D.Torben

357   V. Schroeder, M.D., D.M.Sc.Harold C. Sox, M.D.Martin B. Van Der Weyden, M.D.Freek

358   W.A. Verheugt, M.D. Clinical trial registration—looking back and moving ahead. New

359   England Journal of Medicine. 2007 June 28; 2734-2736.

360

361 Koenig WD. Striving for science that is transparent, credible—and enjoyable: a comment on Ihle

362   et al. Behavioral Ecology. 2017 Apr 1;28(2):358-.

363

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

365

366 Mellor D, Vazire S, StephenLindsay D. Transparent science: A more credible, reproducible, and

367   publishable way to do science.

368

369 Nosek, B.A., Alter, G., Banks, G.C., Borsboom, D., Bowman, S.D., Breckler, S.J., Buck, S.,

370   Chambers, C.D., Chin, G., Christensen, G. and Contestabile, M., 2015. Promoting an open

371   research culture. *Science*, *348*(6242), pp.1422-1425.

17

372

373   National Institutes of Health. Word version of final clinical trials protocol.

374   http://osp.od.nih.gov/wp-content/uploads/Protocol-Template-Version-1.0-040717.docx

375   Accessed August 24, 2017a

376

377   National Institutes of Health. NIH and FDA Request for Public Comment on Draft Clinical Trial

378   Protocol Template for Phase 2 and 3 IND/IDE Studies.

379   http://grants.nih.gov/grants/guide/notice -files/NOT-OD-16-043.html

380   Accessed August 24, 2017b.

381

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

383   Jul 14.

384

385   Parker TH, Forstmeier W, Koricheva J, Fidler F, Hadfield JD, Chee YE, Kelly CD, Gurevitch J,

386   Nakagawa S. Transparency in ecology and evolution: real problems, real solutions. Trends in

387   ecology & evolution. 2016 Sep 30;31(9):711-9.

388

389   Parker TH, Nakagawa S. Practical models for publishing replications in behavioral ecology: a

390   comment on Ihle et al. Behavioral Ecology. 2017 Apr 1;28(2):355-7.

391

392   Reid NM, Hird SM, Brown JM, Pelletier TA, McVay JD, Satler JD, Carstens BC. Poor fit to the

393   multispecies coalescent is widely detectable in empirical data. Systematic biology. 2013 Sep

394   24;63(3):322-33.

18

395

396 Salichos L, Rokas A. Inferring ancient divergences requires genes with strong phylogenetic

397   signals. Nature. 2013 May 16;497(7449):327.

398

399 Whelan S, Liò P, Goldman N. Molecular phylogenetics: state-of-the-art methods for looking into

400   the past. TRENDS in Genetics. 2001 May 1;17(5):262-72.

401

402 Zarin DA, Tse T. Trust but verify: trial registration and determining fidelity to the protocol.

403   Annals of internal medicine. 2013 Jul 2;159(1):65-7.

404

405 Zarin DA, Tse T, Williams RJ, Rajakannan T. Update on trial registration 11 years after the

406   ICMJE policy was established. New England Journal of Medicine. 2017 Jan 26;376(4):383-

407   91.

408