Received: 18 February 2022 | Revised: 27 May 2022 | Accepted: 28 May 2022

DOI: 10.1111/eth.13320

REVIEW

ethology | WILEY

A proposal to enhance data quality and FAIRness

Zegni Triki[1] | Redouan Bshary[2]

[1]Department of Zoology, Stockholm University, Stockholm, Sweden

[2]Institute of Biology, University of Neuchâtel, Neuchâtel, Switzerland

Correspondence
Zegni Triki, Department of Zoology, Stockholm University, Stockholm, Sweden.
Email: zegni.triki@gmail.com

Funding information
The Swiss National Science Foundation supported this study (https://www.snf.ch/en). Grant number: P400PB_199286 to ZT and 173334 to RB. The funder had no role in study design, data collection and analysis, decision to publish or manuscript preparation.

Editor: Wolfgang Goymann

Abstract

In recent years, we witnessed an increasing number of funding agencies, scientific journals and scientists agreeing that society and science benefit from open access to research data. Benefits derive mainly from increased access to knowledge for all and improved transparency and credibility in academia. However, despite the advances in open science and open data, three significant aspects still need considerable policing: data quality, the accompanying summaries with basic information of the data files (i.e. metadata) and computational codes used to generate the research outcomes. Only by having these three components together, we can achieve efficient data sharing and reuse, and hence higher transparency. Here, we present two complementary approaches that potentially can help with shared data quality: (i) data file(s) sharing should be guided step-by-step in public archives with mandatory metadata, and (ii) journals creating assistant data editor positions at editorial boards with a leading role in data quality and computational reproducibility. Forty-four editors-in-chief in the field of behaviour, ecology and evolution shared their opinion with us regarding these two approaches. Although most of the views were divided, the majority estimated that their current editorial board members do not have the necessary skills to assess the quality of shared data. Since data are the core of research studies, we should consider not only data presence but also quality as a requirement for publication.

KEYWORDS
data share, data quality, editor-in-chief, survey

With the open data movement, the practice of data sharing is expanding among biologists (Roche et al., 2022; but see Jiao et al., 2022). The issue, as it stands, is the overall insufficient quality of the archived files (Culina et al., 2020; Roche et al., 2015), in terms of Findability, Accessibility, Interoperability and Reusability, or the FAIR data principles (Wilkinson et al., 2016). It appears that one major problem is a lack of scientists' training in data management and data archiving (Roche et al., 2021; Strasser & Hampton, 2012). The question is how to ensure high-quality FAIR data archiving so that scientists can eventually access computational codes without relying on the authors whenever there is a need to access and/or reuse data and code. Here, we discuss implementing two complementary measures that we believe will strongly improve the quality of shared data and increase reuse. The first measure uses computer macroinstructions that assist authors in archiving their data following a set of step-by-step instructions. Such policy can provide a simple, automatised and standardised quality check. For example, once an author uploads the data file, the macro can create a list with the dataset's column headings while providing information on the data type in each column (e.g. numerical or categorical). Most

importantly, the macro will add a task of requesting mandatory field entries for each column heading in turn. This will ensure minimum metadata presence for every data file uploaded to public archives.

Alone, computer macroinstructions will have limitations in assessing a dataset's completeness. No computer program can determine data quality at this stage, yet a human can. At this stage comes our second proposed measure: journals could create a dedicated assistant editor position (s) to check data quality and FAIRness of submitted papers. The data assistant editor could further review the research outcomes of a given paper by reusing authors' data and code (computational reproducibility) to reproduce their findings. In our view, an expert in the journal's discipline should fulfil the position, as this will facilitate the task of the data assistant editor in evaluating the quality and FAIRness of the data and possibly the validity and relevance of the statistical approach. We think that such a task will be beyond a statistician's skills. Ideally, a biologist will appreciate and understand potential data collection constraints that prevent "perfect" datasets in our field. Subsequently, the authors will deliberately invest in data sharing to avoid publication delays.

To get feedback on to what extent these two approaches can be a promising solution to improve the quality and FAIRness of shared data and computational reproducibility, we asked 160 editors-in-chief in the field of behaviour, ecology and evolution for their opinion. We sent out the survey, to the 160 editors, on three occasions between April 28th and October 25th 2021. Forty-four editors filled in the survey anonymously (see Appendix S1). Thirty-four among the 44 were editors of scientific society journals.

Overall, there was no general agreement on who is responsible for the quality check of shared data: 32% [95% Confidence Interval (CI): lower limit, upper limit; 18.2, 45.8] of the participants suggested referees as the ones responsible for such task, while 25% [12.2, 37.8] suggested the editorial staff, 18% [6.6, 29.3] the public data repositories and 14% [3.7, 24.2] suggested the authors (Figure 1).

Regarding implementing macroinstructions in data sharing, 52% [37.2, 66.8] of the editors (strongly) agreed that macroinstructions could improve the quality and FAIRness of shared data, while 41% [26.5, 55.5] were rather neutral, with 7% [0, 14.5] disagreeing (Figure 2a). Fifty-seven per cent [42.4, 71.6] of editors identified public repositories as the key responsible for implementing such an approach, while 27% [13.9, 40.1] said it is the publishers (Figure 2b). From the editors' replies, the main potential issues of macroinstructions would be complicating the submission procedure (70%, [56.5, 83.5]), the unwillingness of data repositories to implement macroinstructions (54%, [39.3, 68.7]) and that such measure will not appeal to authors (45%, [30.3, 59.7]; multiple choices were possible; Figure 2c).

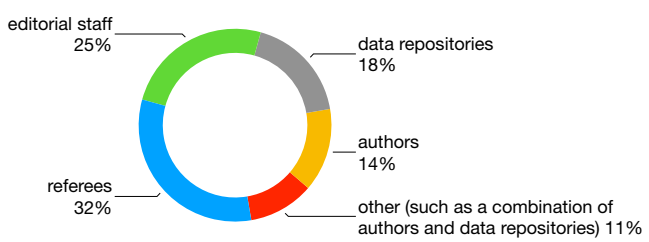


FIGURE 1 In your opinion, who is responsible for data quality check for scientific studies?

Sixty-six per cent [52.0, 79.0] of the editors viewed the addition of a dedicated data editor to their editorial board as an asset, compared to 16% [5.2, 26.8] disagreeing (Figure 3a). It became clear that such opinions stem from the trust of the editors in their editorial team skills when it comes to dealing with data checks. When asked whether their editorial board can assess the quality of shared data, 56% [42.4, 71.6] (strongly) disagreed, while 30% [15.6, 42.4] (strongly) agreed (Figure 3b). However, in a case scenario where they do have a data editor, 62% [47.7, 76.3] (strongly) agreed that this data editor should check the reuse of shared data, such as completeness of data files, accompanying metadata and codes (Figure 3c). According to the participants, the main potential issues with the implementation of a dedicated data editor would be the additional costs (93%, [85.5, 100.0]), a delay in the speed of study acceptance (61%, [46.6, 75.4]), the low attractiveness of assuming such a role (50%, [35.2, 64.8]) and the lack of appeal from the authors (34%, [20.0, 47.0]; multiple choices were possible; Figure 4a). Nevertheless, in a scenario where the editorial board does have a data editor, we asked the participants when the data editor should intervene. Again, the opinions were quite divided, with 41% [26.5, 55.5] proposing that the assessment should take place before the manuscript is sent out to review, 32% [18.2, 45.8] advocated assessment in parallel with reviews, while 9% [0.5, 17.5] chose the moment when the manuscript is ready to be accepted would be best (e.g. provisionally accepted subject to data and code checks; Figure 4b).

Despite the complexity of the current issue of shared data quality and FAIRness and the diverse opinions we received from the editors, the most frequent conclusion was that both macroinstructions and data editor would be useful (36%, [21.8, 50.2]), while 20% [8.2, 31.9] preferred coupling macroinstructions with data quality check by the reviewers. On the other hand, most importantly, only 9% [0.5, 17.5] of the editors thought the current system was satisfying (Figure 5).

Trying to imagine the perspective of our colleagues, we are optimistic that a data editor will be an asset for editorial boards. Suppose there is an issue with data quality/reuse and computational reproducibility, it is tremendously advantageous for the authors and the editorial staff to find out before the study is published. We see it as an encouraging sign that most editors (70%, [56.5, 83.5]) would allow revision and resubmission in the scenario where the data editor would detect problems with the shared data. Fixing the mismatch may change the findings (e.g. from statistically significant to nonsignificant) and, eventually, the conclusions. Therefore, this should become a less important criterion for acceptance as long as the scientific questions and methods remain sound (Brembs, 2019). Nevertheless, 9% [0.5, 17.5] advised rejection of the manuscript in such case, but the remaining 21% [9.0, 33.0] suggested other measures like "revise and accept" and inviting for a thorough investigation by the authors, reviewers and/or editors.

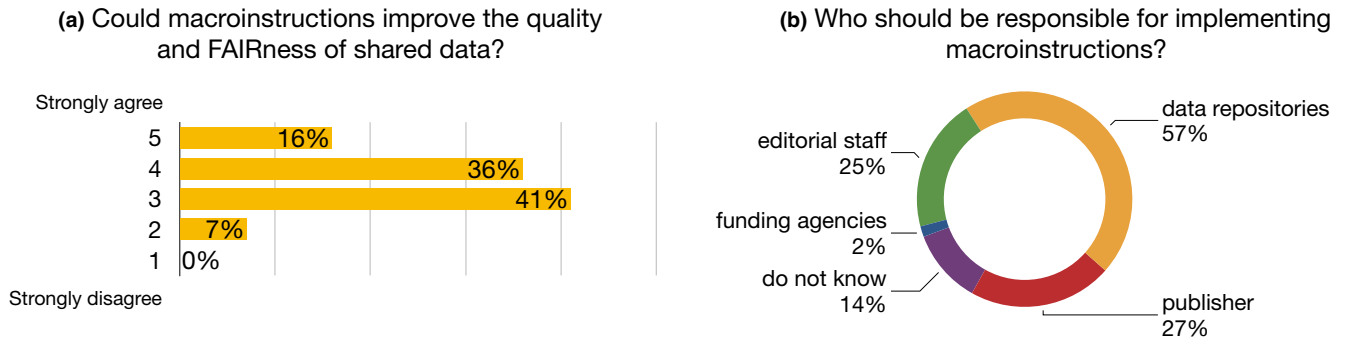
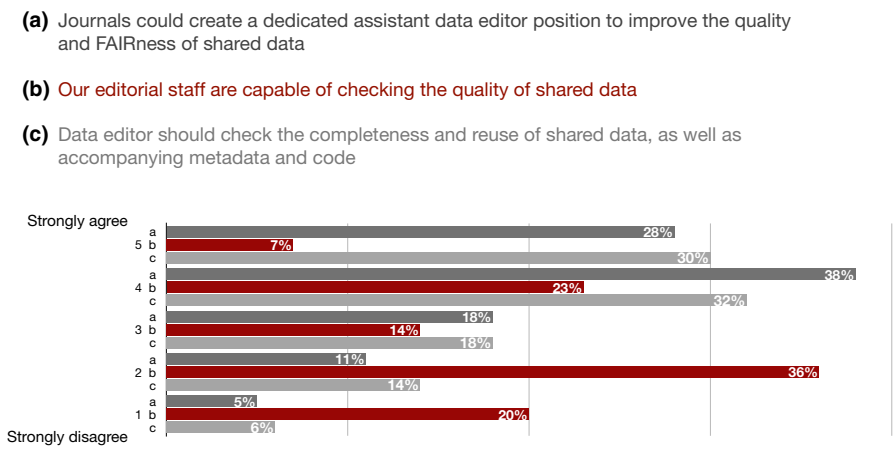


FIGURE 2 Replies linked to macroinstructions

FIGURE 3 Replies linked to data editor (part I)



There is much to gain from improving code sharing, data quality and data FAIRness, such as generating more reliable research and improved transparency and credibility in science (Ihle et al., 2017; Jacobsen et al., 2020; Jeliazkova et al., 2021; Stall et al., 2019). Therefore, given their position and role in academia, public repositories and journals/publishers should invest the most and take the initiative of providing macroinstructions for data sharing and FAIRness. On the other hand, creating data editor positions is a task for scientific journals. By checking the files, data editors may detect mistakes in data files, analyses or codes that warrant corrections, which might affect the study conclusions. This in itself will help journals reduce the number of published errata and possible retractions. Furthermore, we anticipate that data editors will promote the establishment of macroinstructions at the journal level to improve their own efficiency. We agree with the editors-in-chief's opinion that the data editor's duties might become overwhelming without such improved efficiency. Nevertheless, being a data editor can be highly attractive to junior colleagues. The new generation of postdocs and newly graduated PhDs in the field of behaviour, ecology and evolution, for instance, have tremendous statistical knowledge and skills to offer. They can be an excellent asset for journals, and in return, they enrich their CVs with a demonstrated key competence. Furthermore, we do not think the financial argument against a data editor holds well. In many journals, (associate) editors, for example, are unpaid (like reviewers [Aczel et al., 2021]) or receive very little compensation. If having data editor (s) on the editorial board adds costs for journals, maybe the latter can consider having this as part of the publication fees (given that unprivileged authors and universities benefit from fee waivers). Overall, we believe that the returned benefits over time should overcome any costs, providing journals with data editors a competitive edge.

Some editors brought our attention to an already existing tool somewhat close to our suggestion for macroinstructions, DataSeer. DataSeer is a tool that verifies whether shared data match the text

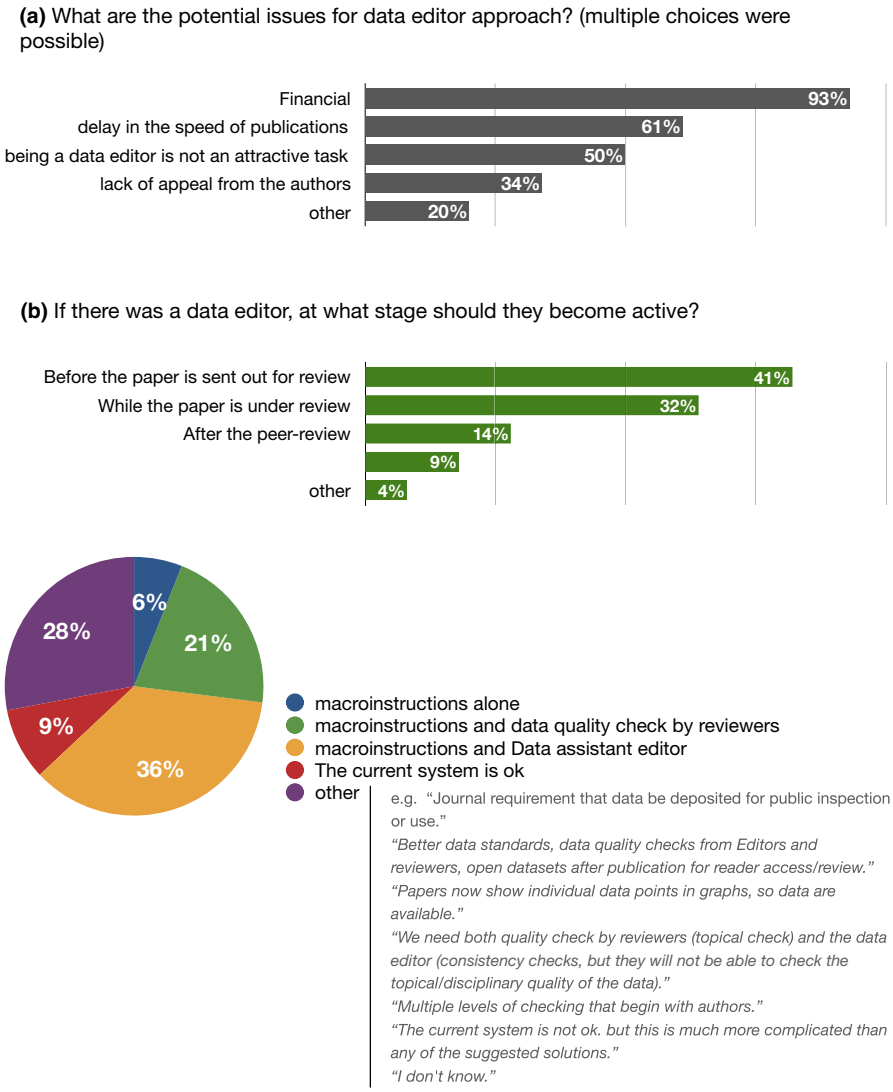


FIGURE 4 Replies linked to data editor (part II). Participants provided "other" answers in (a) such as: "Lack of expertise, lack of time"; "Too much work for one person"; "Technical skills of potential editors"; and "Subject knowledge of potential editors - especially problematic for broad/generalist journals". In (b), the responses provided in the category "other" were, for example: "Depends on the data policy of the journal"

FIGURE 5 What is the best practice for future data sharing?

describing data collection in the manuscript. It is a platform that uses artificial intelligence to help authors tailor their data sharing to the journal's requirements, following best practices (see: https://dataseer.ai). In an editorial piece published last summer (2021), Fernández-Juricic (2021) suggests that authors can acquire a certificate of code reproducibility via a third party (https://codecheck.org.uk/) before submitting their work to a journal. Additionally, there are some R packages (in R coding language [R Core Team, 2020]), such as dataMaid (Petersen & Ekstrøm, 2019) and CodeMeta (Boettiger, 2017) that can assist authors in creating metadata files for their datasets. The drawback of these programs is that they do not investigate the completeness or correctness of data and code. However, it can be a valuable tool if combined with a data editor that can check content beyond the reproducibility of the code. Apparently, a few journals in our field are already adopting the data editor approach, such as the American Naturalist journal and Ecology Letters. Recently, Daniel Bolnick (The American Naturalist, Editor-in-Chief), Tim Vines (DataSeer) and Bob Montgomerie (Data Editor at the American Naturalist journal) wrote an editorial blog post on this very topic. The authors describe how since the summer 2021 the journal has been implementing a combo approach of a small data editorial team and DataSeer to check the quality, completeness and FAIRness of shared data, and code availability and computational reproducibility. They apply such an approach only for accepted manuscripts, which potentially reduces the workload. Thus, some exemplary starting points already exist in the field to enhance scientific integrity and quality.

When asked about the best practice for future data sharing, several editors reached out and greeted the initiative as interesting and timely. Although, one editor commented: "I think you are all very naïve...," We think two important players can make our proposal work; authors interested in extra support to make their study as accessible and flawless as possible and funding agencies that are increasingly caring about open FAIR data and open code. For example, the funding agency, the Swiss National Science Foundation (SNSF), has a policy to push towards open access. It exclusively covers the publication fees for 100% open access journals, but not hybrid ones. For good or bad, this policy clearly selects against publishing in hybrid journals. One can imagine that if a similar approach applies to increase sharing high-quality data and code, it will select for better

data FAIRness. As soon as key players team up, we can increase the quality of publications by improving data FAIRness and achieving computational reproducibility. This next level of transparency will eventually also reinforce the credibility of science.

AUTHOR CONTRIBUTIONS

The authors contributed equally. A preprint version of this article is available on EcoEvoRxiv (Triki & Bshary, 2022).

ACKNOWLEDGMENTS

We kindly thank all editors-in-chief who took the time to participate in this survey and shared their opinions with us. We also thank Dominique Roche for his valuable input on an earlier version of the study.

CONFLICT OF INTEREST

Authors declare no competing interests.

DATA AVAILABILITY STATEMENT

Data is accessible on Figshare: https://doi.org/10.6084/m9.figshare.17091572 (Triki and Bshary, 2021).

ORCID

Zegni Triki https://orcid.org/0000-0001-5592-8963
Redouan Bshary https://orcid.org/0000-0001-7198-8472

REFERENCES

Aczel, B., Szaszi, B., & Holcombe, A. O. (2021). A billion-dollar donation: Estimating the cost of researchers' time spent on peer review. *Research Integrity and Peer Review*, *6*, 14. https://doi.org/10.1186/s41073-021-00118-2

Boettiger, C. (2017). Generating CodeMeta metadata for R packages. *The Journal of Open Source Software*, *2*, 454. https://doi.org/10.21105/joss.00454

Brembs, B. (2019). Reliable novelty: New should not trump true. *PLoS Biology*, *17*, e3000117. https://doi.org/10.1371/journal.pbio.3000117

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*, e3000763.

Fernández-Juricic, E. (2021). Why sharing data and code during peer review can enhance behavioral ecology research. *Behavioral Ecology and Sociobiology*, *75*, 103. https://doi.org/10.1007/s00265-021-03036-x

Ihle, M., Winney, I. S., Krystalli, A., & Croucher, M. (2017). Striving for transparent and credible research: Practical guidelines for behavioral ecologists. *Behavioral Ecology*, *28*, 348–354. https://doi.org/10.1093/beheco/arx003

Jacobsen, A., Kaliyaperumal, R., da Silva Santos, L. O. B., Mons, B., Schultes, E., Roos, M., & Thompson, M. (2020). A generic workflow for the data FAIRification process. *Data Intelligence*, *2*, 56–65. https://doi.org/10.1162/dint_a_00028

Jeliazkova, N., Apostolova, M. D., Andreoli, C., Barone, F., Barrick, A., Battistelli, C., Bossa, C., Botea-Petcu, A., Châtel, A., de Angelis, I., Dusinska, M., el Yamani, N., Gheorghe, D., Giusti, A., Gómez-Fernández, P., Grafström, R., Gromelski, M., Jacobsen, N. R., Jeliazkov, V., … Nymark, P. (2021). Towards FAIR nanosafety data. *Nature Nanotechnology*, *16*, 644–654. https://doi.org/10.1038/s41565-021-00911-6

Jiao, C., Li, K., Fang, Z. (2022). *Data sharing practices across knowledge domains: A dynamic examination of data availability statements in PLOS ONE publications*. ArXiv220310586 Cs.

Petersen, A. H., & Ekstrøm, C. T. (2019). dataMaid: Your assistant for documenting supervised data quality screening in R. *Journal of Statistical Software*, *90*, 1–38. https://doi.org/10.18637/jss.v090.i06

R Core Team. (2020). *A language and environment for statistical computing. Version 3.6.3*. R Foundation for Statistical Computing. https://www.R-project.org/

Roche, D. G., Berberi, I., Dhane, F., Lauzon, F., Soeharjono, S., Dakin, R., & Binning, S. A. (2022). Slow improvement to the archiving quality of open datasets shared by researchers in ecology and evolution *Proceedings of the Royal Society B*, *289*(1975), 20212780

Roche, D. G., Kruuk, L. E. B., Lanfear, R., & Binning, S. A. (2015). Public data archiving in ecology and evolution: How well are we doing? *PLoS Biology*, *13*, e1002295. https://doi.org/10.1371/journal.pbio.1002295

Roche, D. G., Raby, G. D., Norin, T., Ern, R., Scheuffele, H., Skeeles, M., Morgan, R., Andreassen, A. H., Clements, J. C., Louissaint, S., Jutfelt, F., Clark, T. D., & Binning, S. A. (2022). Paths towards greater consensus building in experimental biology. *The Journal of Experimental Biology*, *225*, jeb243559. https://doi.org/10.1242/jeb.243559

Stall, S., Yarmey, L., Cutcher-Gershenfeld, J., Hanson, B., Lehnert, K., Nosek, B., Parsons, M., Robinson, E., & Wyborn, L. (2019). Make scientific data FAIR. *Nature*, *570*, 27–29. https://doi.org/10.1038/d41586-019-01720-7

Strasser, C. A., & Hampton, S. E. (2012). The fractured lab notebook: Undergraduates and ecological data management training in the United States. *Ecosphere*, *3*, 1–18.

Triki, Z., Bshary, R. (2021). Data from: A proposal to enhance data quality and FAIRness. figshare. Dataset. https://doi.org/10.6084/m9.figshare.17091572

Triki, Z., & Bshary, R. (2022). How to enhance data FAIRness. *EcoEvoRxiv*. https://doi.org/10.32942/osf.io/tf7yv

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Triki, Z., & Bshary, R. (2022). A proposal to enhance data quality and FAIRness. *Ethology*, *00*, 1–5. https://doi.org/10.1111/eth.13320