A new phylogenetic protocol: dealing with model misspecification and confirmation bias in molecular phylogenetics

Lars S. Jermiin\textsuperscript{1,2,3,4,*}, Renee A. Catullo\textsuperscript{1,2,5} and Barbara R. Holland\textsuperscript{6}

\textsuperscript{1}CSIRO Land & Water, Canberra, ACT 2601, Australia, \textsuperscript{2}Research School of Biology, Australian National University, Canberra, ACT 2601, Australia, \textsuperscript{3}School of Biology & Environment Science, University College Dublin, Belfield, Dublin 4, Ireland, \textsuperscript{4}Earth Institute, University College Dublin, Belfield, Dublin 4, Ireland, \textsuperscript{5}School of Science and Health & Hawkesbury Institute of the Environment, Western Sydney University, Penrith, NSW 2751, Australia and \textsuperscript{6}School of Natural Sciences, University of Tasmania, Hobart, TAS 7001, Australia

Received January 17, 2020; Revised May 18, 2020; Editorial Decision May 28, 2020; Accepted June 04, 2020

ABSTRACT

Molecular phylogenetics plays a key role in comparative genomics and has increasingly significant impacts on science, industry, government, public health and society. In this paper, we posit that the current phylogenetic protocol is missing two critical steps, and that their absence allows model misspecification and confirmation bias to unduly influence phylogenetic estimates. Based on the potential offered by well-established but under-used procedures, such as assessment of phylogenetic assumptions and tests of goodness of fit, we introduce a new phylogenetic protocol that will reduce confirmation bias and increase the accuracy of phylogenetic estimates.

INTRODUCTION

Molecular phylogenetics plays a pivotal role in the analysis of genomic data and has already had a significant, wide-reaching impact in science, industry, government, public health and society (Table 1). Although the science and methodology behind applied phylogenetics is increasingly well understood within parts of the scientific community (1,2), there is still a worryingly large body of research where the phylogenetic analysis was done with little attention to the consequences of a statistical misfit between the phylogenetic data and the assumptions that underpin the phylogenetic methods.

One reason for this is that phylogenetics relies extensively on statistics, mathematics and computer science, and many users of phylogenetic methods find the relevant sections of these disciplines challenging to comprehend. Another reason is that methods and software often are chosen because they already are popular and/or easy to use, rather than because they are the most appropriate for the scientific questions and phylogenetic data at hand. A third reason is that much of the phylogenetic research done to date has relied on phylogenetic protocols (3–8), which have evolved to become a standard to which it seems sensible to adhere. Although these protocols vary, they have, at their core, a common set of sensible features (i.e. procedural steps) that are linked in a seemingly logical manner (see below).

Here we posit that, although the current phylogenetic protocol has many useful features, it is missing two crucial components whereby the quality of fit between the data and models applied is assessed. This means that using the phylogenetic protocol in its current form may lead to biased conclusions. We suggest a modification to the protocol that will make it more robust and reliable, and that will identify where new methods are needed to enable ease of use and statistical accuracy in the field.

The current phylogenetic protocol

Phylogenetic analysis of alignments of nucleotides or amino acids usually follows a protocol like that in Figure 1. Initially, the phylogenetic data are chosen on the assumption that they will allow the researchers to solve a particular scientific problem. This choice of sequence data is often based on prior knowledge, developed locally or extracted from the literature. Then, a multiple sequence alignment (MSA) method is chosen, often on the basis of prior experience with a specific method. The sequences are then aligned, the aim being to obtain an MSA, wherein homologous characters (i.e. nucleotides or amino acids) are aligned. In practice, it is often necessary to insert alignment gaps between some of the characters in some of the sequences to obtain an optimal MSA— in some cases, there may be characters in different sequences that cannot be aligned reliably.

Then follows the task of selecting sites that will be used to infer the phylogenetic tree. The rationale behind doing
Table 1. Examples of phylogenetic research, divided into areas, based on impact and/or relevance

| Area           | Examples                                                                 |
|----------------|--------------------------------------------------------------------------|
| **Science**    | Provide accurate estimates of the evolution of, for example, species (9–12) |
|                | Provide accurate estimates of evolutionary processes at the molecular level (13,14) |
|                | Addressing macro-evolutionary questions pertaining to birds and mammals (15,16) |
|                | Understand biogeographic patterns and diversity (17,18)                   |
|                | Reconstruction of ancestral states (19,20)                                |
| **Industry**   | Facilitate the design and engineering of novel enzymes (21) and drugs (22–24) |
| **Government** | Reveal likely sources and dispersal routes of agricultural pests and pathogens (25–29) |
|                | Assign conservation priorities to species or biogeographic regions based on estimates of genetic diversity (30–32) |
| **Public health** | Reveal the origin and spread of human pathogens (33–36)                  |
|                | Predict the evolution of human influenza A (37)                         |
|                | Identify the natural reservoir of the virus causing COVID-19 (38–40)      |
|                | Reveal the origin and evolution of cancers (41,42)                       |
| **Society**    | Reveal the evolution of human language (43,44)                           |
|                | Map the relationship among ancient texts (45), tales (46) and music (47)  |
|                | Reveal evolution of humans since their divergence from other primates (48–52) |

![Figure 1. The current phylogenetic protocol. Solid arrows show the order of actions normally taken during a phylogenetic analysis. Dashed arrows show feedback loops often employed in phylogenetic research. For details, see the main text.](image-url)

so is to maximize the signal-to-noise ratio in the MSA. By omitting poorly aligned and highly variable sections of the MSA, which are thought to create noise due to the difficulty of establishing true homology for each site (defined as similarity due to historical relationships by descent (53)), it is hoped that the resulting sub-MSA will retain a strong historical signal (defined as the order and timing of divergence events (54)) that will allow users to obtain an accurate phylogeny. The choice of sites to retain is made by visual inspection of the MSA or by using purpose-built software (55–65). The automated ways of filtering MSAs have been questioned (66).

Having generated a sub-MSA, the next step in the protocol is to select a phylogenetic method. The choice of phylogenetic method implies accepting the assumptions on which the method rests. For example, irrespective of whether the data comprise nucleotides, codons, or amino acids, it is often assumed that the sequences evolved along a single bifurcating tree and that the evolutionary processes operating at the variable sites in the sequences are independent and distributed identically (in a probabilistic sense). If model-based molecular phylogenetic methods are chosen, the underlying assumption usually is that the evolutionary processes operating at the variable sites can be approximated accurately by using Markov models that are stationary, reversible and homogeneous (67–69) over time (the assumption of evolution under SRH conditions) (for more details about the most common phylogenetic assumptions, see Appendix 1). In practice, the choice is one between methods assuming that the underlying evolutionary processes can be modeled using a Markov model of nucleotide or amino-acid substitutions (i.e. distance methods (70–75), likelihood methods (70,72,74–80), Bayesian methods (81–88)) or non-parametric phylogenetic methods (i.e. parsimony methods (70,72,74,75,89–91)). In reality, many researchers analyze their data using a range of model-based phylogenetic methods, and reports that only use parsimony methods are increasingly rare. Depending on the chosen phylogenetic method, researchers may have to select a suitable model of sequence evolution (i.e. a model that combines the substitution model and the rates-across-sites model) to apply to the sub-MSA. This choice is often made by using model selection methods (14,92–103).

Having chosen a phylogenetic method and, in relevant cases, a suitable model of sequence evolution, the next step involves obtaining accurate estimates of the tree and evolutionary processes that led to the data. Phylogenetic methods are implemented in many software packages (70–86,89–91) and, depending on the methods chosen, users often also obtain the non-parametric bootstrap probability (104) or clade credibility (105) to measure support for divergence events in the phylogeny.

Having inferred the phylogeny, the final step in the protocol is to interpret the result. Under some conditions—most commonly, the inclusion of out-group sequences—the tree may be drawn and interpreted as a rooted phylogeny, in which case the order of divergence events and the lengths of the individual edges may be used to infer, for example, tempo and mode of evolution of the data. Often, the
inferred evolutionary relationships. Often, too, there are surprises, which are difficult to understand and explain. If the phylogenetic estimate is convincing and newsworthy, the discoveries may be reported, for example, through papers in peer-reviewed journals.

On the other hand, if the surprises are too numerous or unbelievable, the researchers may begin the task of finding out what may have ‘gone wrong’ during the phylogenetic analysis. This process is depicted as dashed feedback loops in Figure 1. The researchers may analyze their data differently (e.g. use other Markov models, use other phylogenetic methods, use a different sub-MSA, align the sequences differently, use a different alignment method or use another dataset), and given enough patience, they may reach a conclusion about their data and attempt to publish their results.

**Problems with the current phylogenetic protocol**

Although the current phylogenetic protocol has led to many important discoveries, it also has left many scientists with strong doubts about or, alternatively, undue confidence in the estimates. The literature is rife with examples where analyses of the same data have led to disagreements among experts about what is the ‘right’ phylogeny (cf. e.g. (106–108)). Such disagreements are confusing, especially for non-experts and the public. To understand why these disagreements might arise, it is necessary to understand the challenges that applied phylogenetic research still faces.

While it is clear that the right data are needed to answer the scientific question at hand, making that choice is not always as trivial as it might seem. In some cases, the sequences may have evolved too slowly and/or be too short, in which case there may not be enough information in the data, or they have evolved so fast that the historical signal has largely been lost (109). In rarely reported cases, the data are not what they purport to be (110).

Next, there is no consensus on what constitutes an optimal MSA. In simple terms, what is required is an accurate MSA where every site is a correct homology statement. Currently, there is no automatic procedure for assessing homology (53). Frequently, different methods return different MSAs, implying different homology statements, but they cannot all be right. Averaging MSAs has been proposed as a solution (111). Producing an MSA will often involve manual modifications following visual inspection, which introduces subjectivity and a lack of reproducibility. One way to mitigate this problem is to rely on reviews and simulation-based comparisons of MSA methods (53,112–119), but these reports seem to have had less impact than the assumptions underpinning each phylogenetic method are often poorly understood. Indeed, strong violation of phylogenetic assumptions could lead to similar, but nevertheless, wrong trees being inferred using different phylogenetic methods (123,124).

Over the last two decades, the choice of a suitable model of sequence evolution has often been made by using purpose-built model-selection methods (14,92–103). Assuming a tree, these methods step through a list of predefined models, evaluating each of them, one by one, until the list is exhausted. This approach is sensible if the true or most appropriate model is included in the list of predefined models. On the other hand, if the true model is not included in this list, then the popular model-selection methods will never be able to return the true model. They will return an optimal model, but it will be conditional on the models included in the list. Unfortunately, most model-selection methods only consider time-reversible Markov models. If the sequences have evolved along a single tree but under non-reversible Markovian conditions (i.e. under non-SRH conditions), then there is no way that a single, time-reversible Markov model is sufficient to approximate the evolutionary processes across all edges of the tree (125). Therefore, it is worrying that many researchers still ignore or dismiss the implication of compositional heterogeneity across sequences (124). This type of heterogeneity indicates that the evolutionary processes have changed across the lineages. This implication must be taken seriously when data are analyzed phylogenetically.

The choice of phylogenetic program is often driven by prior experiences and transaction costs (i.e. the time it takes to become a confident and competent user of the software) rather than by a profound understanding of the strengths, limitations and weaknesses of the available software. However, this may not substantially minimize the accuracy of the phylogenetic estimate, as long as the data do not violate the assumptions on which the phylogenetic methods are based, and the phylogenetic methods search tree space and model space thoroughly.

The bootstrap probability (104) and clade credibility (105) are often thought of as metrics of the accuracy of
the phylogenetic estimate or the confidence we might have in the inferred divergence events. Unfortunately, doing so is unwise because they measure consistency of the estimate (126)—a phylogenetic estimate may consistently point to an incorrect tree.

Finally, once well-supported phylogenetic estimates have been inferred, prior expectations are likely to influence whether the estimates are considered reliable and newsworthy. In some cases, where information on the phylogeny is known (e.g. serially sampled viral genomes), not meeting the prior expectations may signal a problem with the phylogenetic analysis. However, if a researcher’s expectations are met by the phylogenetic results, it is more likely that a report will be written without a further assessment of what might have gone wrong during the analysis. This tendency—allowing prior expectations to influence the interpretation of phylogenetic estimates—is called confirmation bias. Confirmation bias is not discussed in phylogenetics, even though it is a recognized problem in other disciplines (e.g. psychology and social science (127)), so it is timely that the phylogenetic community takes onboard the serious implications of this.

The new phylogenetic protocol

Although the current phylogenetic protocol has many shortcomings, it also has many good attributes, including that it is easy to apply and implement as a pipeline. However, to mitigate its limitations, it will be necessary to redesign the protocol to accommodate well-established, but largely ignored, procedures as well as new feedback loops.

Figure 2 shows a picture of the new phylogenetic protocol. It shares many features found in the current protocol (e.g. the first four steps). However, the fifth step (assess phylogenetic assumptions) will be novel to many researchers. As all phylogenetic methods are based on assumptions, it is sensible to validate these assumptions at this point in the protocol. Since many phylogenetic methods assume that the data (e.g. different genes) have evolved over the same tree, and that the chosen data partitions have evolved independently under the same time-reversible Markovian conditions, it is wise to survey the sub-MSAs for evidence that the sequences actually have evolved under these conditions. If the data violate these phylogenetic assumptions, then it will be wise to avoid these phylogenetic methods and to employ other phylogenetic methods. Alternatively, it may be worth following the relevant feedback loops in Figure 2—perhaps something led to a biased sub-MSA? The relevance and benefits of this step are illustrated using a case study (Appendix 2), which focuses on determining whether a dataset is consistent with the phylogenetic assumption of evolution under time-reversible conditions. Assessments of other phylogenetic assumptions require other types of tests and surveys. Some of the relevant questions and methods are listed in Table 2.

Next follows the choice of phylogenetic method, but now this choice is made on the basis of the previous step, rather than on cultural or computational reasons. If the sequences have evolved on a single tree under time-reversible Markovian conditions, there is a large set of phylogenetic methods to choose from (70–86,89–91). On the other hand, if these data have evolved under more complex Markovian conditions, the number of suitable phylogenetic methods is rather limited (13,68,151–176), and most of these methods are aimed at finding the optimal model of sequence evolution for a given tree rather than finding the optimal set of trees. Users of phylogenetic methods therefore are sometimes confronted by a dilemma: Do they abandon their dataset because it has evolved under non-time-reversible conditions and because there are no appropriate phylogenetic methods for such data, or do they take the risk and employ the phylogenetic methods that assume evolution under time-reversible conditions? Fortunately, there may be a way around this dilemma.

Having inferred the phylogeny using model-based phylogenetic methods, it is possible to test the fit between tree, model and data (step 10 of the new protocol). A suitable test of goodness of fit was proposed in 1993 (177) (Figure 3). In brief, using the inferred optimal tree, including the edge lengths, it is possible to simulate datasets under the null model (i.e. the inferred optimal model of sequence evolution with its parameter values included). This is called a parametric bootstrap. Given the optimal tree (i.e. $T$ in Figure 3) and the optimal model of sequence evolution (i.e. $M$ in Figure 3), several sequence-generating programs (13,165,178–181) facilitate production of pseudo-data. Having generated, say, $m = 1000$ pseudo-data, the next step involves finding the difference ($\delta$) between the unconstrained log-likelihood (i.e. the log-likelihood inferred without assuming
Table 2. Important questions requiring pre-phylogenetic surveys of phylogenetic data, accompanied by publications describing methods developed to carry such surveys

Are evolutionary processes Markovian?
Two papers consider this question (128,129), and one of these presents methods in the context of recoding nucleotides (129).

Are evolutionary processes independent and identically distributed?
Four papers consider this question and present methods for assessing phylogenetic data (130–133).

Are evolutionary processes stationary, reversible and homogeneous?
Several papers consider this question and present methods for assessing phylogenetic data (54,134–143,200,201).

Are evolutionary processes heterotachous?
Several papers consider this question (144,181,196–198) and some of these present methods for inferring phylogenetic trees under such conditions (181,196–198).

Is the phylogenetic signal in phylogenetic data tree-like?
Several papers consider this question and present methods for surveying phylogenetic data (145,146,202,203).

Is the historical signal in phylogenetic data decayed?
Several papers consider this question and present methods for surveying phylogenetic data (147,204–206).

Is the compositional signal in phylogenetic data stronger than expected?
One paper considers this question and presents a method for surveying phylogenetic data (147).

How phylogenetically informative are phylogenetic data?
Three papers consider this question and present methods for surveying phylogenetic data (148–150).

---

Figure 3. Diagram showing the parametric bootstrap procedure that may be used to conduct a suitable goodness-of-fit test. The procedure includes three steps. For details, see the main text.

---

a tree and a model of sequence evolution) and constrained log-likelihood (i.e. the log-likelihood inferred assuming a tree and a model of sequence evolution)—that is, computing $\delta = \ln L(D) - \ln L(D|T,M)$, where $D$ is the data, $T$ is the optimal tree and $M$ is the optimal model of sequence evolution. If $\delta$ is greater for the real data than it is for the pseudo-data, then that result reveals a poor fit between tree, model and data (141). The approach described here works well for likelihood-based phylogenetic analysis. A similar approach that relies on posterior predictive distributions is available for Bayesian-based phylogenetic analysis (182).

Parametric bootstrapping is computationally expensive and may be time consuming, so we recommend it be used only if the data appear to meet the assumptions of the phylogenetic method used. The advantages of using such a goodness-of-fit test are that it allows users to assess whether
a poor fit is poor enough to not be due to chance. It does not say anything about why a fit might be poor or whether the lack of fit matters (in a biological sense). If the fit is poor, then the relevant feedback loops should be followed (Figure 2)—perhaps a biasing factor was overlooked? If the phylogenetic tree and model of sequence evolution are found to fit the data, then that implies that the estimates represent a plausible explanation of the data. It is these estimates that should be reported, but only as one plausible explanation, not as the only possible explanation. This is because there may be other plausible explanations of the data that never were considered during the analysis.

The future: areas in most need of methodological research

Adherence to the new phylogenetic protocol will undoubtedly lead to improved accuracy of phylogenetic estimates and a reduction of confirmation bias. The advantage of the fifth step in the new phylogenetic protocol (i.e. assess phylogenetic assumptions) is that users are able to decide how to do the most computationally intensive parts of the phylogenetic study without wasting precious time on, for example, a high-performance computer center. Model selection, phylogenetic analysis and parametric bootstrapping are computationally intensive and time consuming, and there is a need for new, computationally efficient strategies that can be used to analyze sequences that have evolved under complex phylogenetic conditions.

The advantage of the 10th step in the new phylogenetic protocol (i.e. test goodness of fit) is its ability to answer whether an inferred phylogeny explains the data well, or not. In so doing, the step tackles the issue of confirmation bias front on. Clearly, without information gleaned from the fifth step, the parametric bootstrap might return an unwanted answer (i.e. the inferred tree and model of sequence evolution does not fit the data well), so to avoid this disappointment it is better to embrace the new phylogenetic protocol in full.

Results emerging from studies that rely on the new phylogenetic protocol might well call into question published phylogenetic research, but there is also a chance that research might gain stronger support. This is good for everyone concerned, especially since it will become easier to defend the notion that the research was done without prejudice or preference for a particular result. Objectivity should be restored in phylogenetics—it is no longer reasonable to defend phylogenetic results on the basis that they were obtained using the best available tools; if these tools do not model the evolutionary processes accurately, then that should be reported rather than be hidden away. This is critical because it increases the transparency of the research and aids other scientists to understand the intricate nature of the challenges encountered.

Notwithstanding the likely benefits offered by the new protocol and the methods supporting it, it would be unwise to assume that further development of phylogenetic methods will no longer be needed. On the contrary, there is a lot of evidence that method development will be needed in different areas:

**MSA methods:** There is a dire need for MSA methods that yield accurate homology statements. Likewise, there is a need for methods that allow users to (i) determine how accurate different MSA methods are and (ii) select MSA methods that are most suitable for the data at hand. Moreover, there is a dire need for better transparency in the way MSAs are reported in the literature (65).

**Methods for masking MSAs:** Assuming an accurate MSA has been inferred, there is a need for strategies that can be applied to identify and distinguish between poorly aligned and highly variable regions of MSA. Well-aligned but highly variable regions of MSAs may be more informative than poorly aligned regions of such MSAs, so to delete them may be unwise.

**Model-selection methods:** Model selection is a pivotal prerequisite if parametric phylogenetic methods are used. However, the Model selection methods currently employed may not be accurate (183), especially for sequences that have evolved under complex conditions (e.g. heterochalous, covarion or non-time-reversible conditions). For example, the evolutionary process may have to be considered an evolving entity in its own right. Further, a better understanding of the information criteria used is necessary (184).

**Phylogenetic methods:** While there are many accurate phylogenetic methods for analysis of data that have evolved under time-homogeneous, reversible Markovian conditions, there is a dearth of accurate phylogenetic methods suitable for analysis of data that have evolved under more complex conditions. Added to this challenge are methods that accurately consider incomplete lineage sorting of genetic markers and the special conditions associated with the analysis of single nucleotide polymorphism (SNP) data.

**Goodness-of-fit tests:** Suitable goodness-of-fit tests are available, but there is not only a need for a wider understanding of the merits of these tests, but also of how they can be tailored to suit different requirements. In particular, there is a need for programs that can generate simulated data under extremely complex evolutionary conditions. Some programs are available (13,165,180,185), but they only cater for a very limited set of conditions.

**Analysis of residuals:** Although goodness-of-fit tests can tell you whether or not the lack of fit observed is potentially due to chance, they do not answer the more useful question of whether or not that lack of fit matters or how the lack of fit arises (186,187). For this reason, residual diagnostic tools that can inform the user about the way in which their model fails to fit the data would be very useful.

In summary, while calls for better phylogenetic methods and more careful considerations of the data have occurred (125), we posit there is a need for a comprehensive overhaul of the current phylogenetic protocol. The proposed new phylogenetic protocol is unlikely to be a final product; rather, it is probably a first, but important step toward a more scientifically sound phylogenetic protocol, which should result in not only more accurate phylogenetic estimates, but also a reduction in the likelihood of confirmation bias.

**CONCLUSION**

The Holy Grail in molecular phylogenetics is being able to obtain accurate, reproducible, transparent and trustworthy phylogenetic estimates from the phylogenetic data. We are
not there yet, but encouraging progress is being made not only in the design of the phylogenetic protocol, but also in phylogenetic methodology based on the likelihood and Bayesian optimality criteria.

Notwithstanding this progress, a quantum shift in attitudes and habits is needed in the phylogenetic community—it is no longer enough to infer an optimal phylogenetic estimate. The fit between trees, models and data must be evaluated before phylogenetic estimates can be considered newworthy. We owe it to the scientific community and wider public to be as rigorous as we can—the attitude ‘She’ll be alright, mate’ is no longer appropriate in this discipline.

ACKNOWLEDGEMENTS

L.S.J. thanks the University College Dublin for its generous hospitality. We thank D. Higgins, A. Locatelli and K.H. Wolfe for their constructive feedback.

Author contributions: L.S.J. conceived the new phylogenetic protocol. R.A.C., B.R.H. and L.S.J. wrote the paper. Dedication: To the memory of Rossiter H. Crozier (1943–2009), an evolutionary biologist, who, with his generosity and wide-reaching inquisitiveness, inspired students and scientists in Australia, and abroad.

FUNDING

The Australian Government’s Commonwealth Scientific Industrial Research Organisation (CSIRO).

Conflict of interest statement. None declared.

REFERENCES

1. Yang, Z.H. and Rannala, B. (2012) Molecular phylogenetics: principles and practice. Nat. Rev. Genet., 13, 303–314.
2. Liò, P. and Goldman, N. (1998) Models of molecular evolution and phylogeny. Genome Res., 8, 1233–1244.
3. Harrison, C.J. and Langdale, J.A. (2006) A step by step guide to phylogeny reconstruction. Plant J., 45, 561–572.
4. Hunt, T. and Vogler, A.P. (2008) A protocol for large-scale rRNA sequence analysis: towards a detailed phylogeny of Coleoptera. Mol. Phylogenet. Evol., 47, 289–301.
5. Hall, B.G. (2013) Building phylogenetic trees from molecular data with MEGA. Mol. Biol. Evol., 30, 1229–1235.
6. Lemmon, E.M. and Lemmon, A.R. (2013) High-throughput genomic data in systematics and phylogenetics. Annu. Rev. Ecol. Evol. Syst., 44, 99–121.
7. O’Halloran, D. (2014) A practical guide to phylogenetics for nonexperts. J. Vis. Exp., 84, e50975.
8. Wilding, M., Nachtshatt, M., Speight, R. and Scott, C. (2017) An improved and general streamlined phylogenetic protocol applied to the fatty acid desaturase family. Mol. Phylogenet. Evol., 115, 50–57.
9. dos Reis, M., Inoue, J., Hasegawa, M., Asher, R.J., Donoghue, P.C.J. and Yang, Z.H. (2012) Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. Proc. R. Soc. B, 279, 3491–3500.
10. Rahbel, B.R., Gitzendanner, M.A., Solits, P.S., Solits, D.E. and Burleigh, J.G. (2014) From algae to angiosperms—inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. BMC Evol. Biol., 14, 26.
11. Misof, B., Liu, S.L., Meusmann, K., Peters, R.S., Donatha, M., Mayer, C., Fransen, P.B., Ware, J., Flouri, T., Beutel, R.G. et al. (2014) Phylogenomics resolves the timing and pattern of insect evolution. Science, 346, 763–767.
12. Prum, R.O., Berv, J.S., Dornburg, A., Field, D.J., Townsend, I.P., Lemmon, E.M. and Lemmon, A.R. (2015) A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature, 526, 569–573.
13. Jayaswal, V., Wong, T.K.F., Robinson, J., Poladian, L. and Jermain, L.S. (2014) Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. Syst. Biol., 63, 726–742.
14. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., Von Haeseler, A. and Jermain, L.S. (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods, 14, 587–589.
15. Perry, D. and Phillips, M.J. (2004) The rise of birds and mammals: are microevolutionary processes sufficient for macroevolution. Trends Ecol. Evol., 19, 516–522.
16. Meridith, R.W., Janeczka, J.E., Gatesy, J., Ryder, O.A., Fisher, C.A., Teeling, E.C., Goodbla, A., Eizirik, E., Simão, T.L.L., Stadler, T. et al. (2011) Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. Science, 334, 521–524.
17. Knapp, M., Stockler, K., Havell, D., Delsuc, F., Sebastiani, F. and Lockhart, P.J. (2005) Relaxed molecular clock provides evidence for long-distance dispersal of Nothofagus (southern beech). PLoS Biol., 3, 38–43.
18. Jetz, W., Thomas, G.H., Joy, J.B., Hartmann, K. and Mooers, A.O. (2012) The global diversity of birds in space and time. Nature, 491, 444–448.
19. Marazzi, B., Ane, C., Simon, M.F., Delgado-Salinas, A., Luckow, M. and Sanderson, M.J. (2012) Locating evolutionary precursors on a phylogenetic tree. Evolution, 66, 3918–3930.
20. Pagel, M., Meade, A. and Barker, D. (2004) Bayesian estimation of ancestral character states on phylogenies. Syst. Biol., 53, 673–684.
21. Wilding, M., Peat, T.S., Kalyaanamoorthy, S., Newman, J., Scott, C. and Jermain, L.S. (2017) Reverse engineering: transaminase biocatalyst development using ancestral sequence reconstruction. Green Chem., 19, 5375–5380.
22. Searls, D.B. (2003) Pharmacophylogenomics: genes, evolution and drug targets. Nat. Rev. Drug Discov., 2, 613–623.
23. Goodfellow, M. and Fiedler, H.P. (2010) A guide to successful bioprospecting: informed by actinobacterial systematics. Antonie Van Leeuwenhoek, 98, 119–142.
24. Wright, G.D. and Poinar, H. (2012) Antibiotic resistance is ancient: implications for drug discovery. Trends Microbiol., 20, 157–159.
25. Boykin, L.M., Armstrong, K.F., Kubatko, L.S. and De Barro, P.J. (2012) Species delimitation and global biosecurity. Evol. Bioinform., 8, 1–37.
26. Hosokawa, T., Nikoh, N. and Fukatsu, T. (2014) Fine-Scale geographical origin of an insect pest invading North America. PLoS One, 9, e89107.
27. Yasaka, R., Ohba, K., Schwinghamer, M.W., Fletcher, J., Ochoa-Corona, F.M., Thomas, J.E., Ho, S.Y.W., Gibbs, A.J. and Ohshima, K. (2015) Phylodynamic evidence of the migration of turnip mosaic potyvirus from Europe to Australia and New Zealand. J. Gen. Virol., 96, 701–713.
28. Tay, W.T., Walsh, T.K.F., Downes, S., Anderson, C., Jermain, L.S., Wong, T.K.F., Piper, M.C., Chang, E.S., Macedo, I.B., Czepak, C. et al. (2017) Mitochondrial DNA and trade data support multiple origins of Helicoverpa armigera (Lepidoptera, Noctuidae) in Brazil. Sci. Rep., 7, 45302.
29. Anderson, C.J., Oakeshott, J.G., Tay, W.T., Gordon, K.H.J., Zwick, A. and Walsh, T.K. (2018) Hybridization and gene flow in the mega-pest lineage of moth, Helicoverpa, Proc. Natl. Acad. Sci. U.S.A., 115, 5034–5039.
30. Gonzalez-Orozco, C.E., Pollock, L.J., Thornhill, A.H., Mishler, B.D., Knerr, N., Laflan, S., Miller, J.T., Rosauer, D.F., Faith, D.P., Nipperess, D.A. et al. (2016) Phylogenetic approaches reveal biodiversity threats under climate change. Nat. Clim. Chang., 6, 1110–1114.
31. Rosauer, D.F., Blom, M.P.K., Bourke, G., Catalano, S., Donnellan, S., Gillespie, G., Mulder, E., Oliver, P.M., Potter, S., Pratt, R. et al. (2016) Phylogeography, hotspots and conservation priorities: an example from the top end of Australia. Biol. Conserv., 204, 83–93.
32. Tucker, C.M., Cadotte, M.W., Carvalho, S.B., Davies, T.J., Ferrier, S., Fritz, S.A., Greyer, R., Helmus, M.R., Jin, L.S., Mooers, A.O. et al. (2017) A guide to phylogenetic metrics for conservation, community ecology and macroecology. Biol. Rev., 92, 698–715.
33. Andersen, K.G., Shapiro, B.J., Matzuga, C.B., Sealoff, R., Lin, A.E., Moses, L.M., Folari, O.A., Goba, A., Odia, I., Ethane, P.E. et al.
176. Holland, B.R., Jarvis, P.D. and Sumner, I.G. (2013) Low-parameter phylogenetic inference under the general Markov model. *Syst. Biol.*, 62, 78–92.

177. Goldman, N. (1993) Statistical tests of models of DNA substitution. *J. Mol. Evol.*, 36, 182–198.

178. Ramhout, A. and Grasso, N.C. (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosyst.*, 13, 235–238.

179. Fletcher, W. and Yang, Z. (2009) INDELible: a flexible simulator of biological sequence evolution. *Mol. Biol. Evol.*, 26, 1879–1888.

180. Mallo, D., de Oliveira Martins, L. and Posada, D. (2016) SimPhy: phylogenetic simulation of gene, locus, and species trees. *Syst. Biol.*, 65, 334–344.

181. Wang, H.C., Spencer, M., Sussko, E. and Roger, A.J. (2007) Testing for covarion-like evolution in protein sequences. *Mol. Biol. Evol.*, 24, 294–305.

182. Bullback, J.P. (2002) Bayesian model adequacy and choice in phylogenetics. *Mol. Biol. Evol.*, 19, 1171–1180.

183. Susko, E. and Roger, A.J. (2020) On the use of information criteria for model selection in phylogenetics. *Mol. Biol. Evol.*, 37, 549–562.

184. Dzjak, J.J., Coffman, D.L., Lanza, S.T., Li, R. and Jermin, L.S. (2020) Sensitivity and specificity of information criteria. *Brief. Bioinform.*, 21, 533–565.

185. Duchêne, D.A., Duchêne, S. and Ho, S.Y.W. (2018) PhyloMad: efficient assessment of phylogenetic model adequacy. *Bioinformatics*, 34, 2300–2301.

186. Kumar, S., Filipski, A.J., Battistuzi, F.U., Pond, S.L.K. and Tamura, K. (2012) Statistics and truth in phylogenomics. *Mol. Biol. Evol.*, 29, 457–472.

187. Holland, B.R. (2013) The rise of statistical phylogenetics. *Aust. N. Z. J. Stat.*, 55, 205–220.

188. Yang, Z. (1997) Among-site rate variation and its impact on phylogenetic analysis. *Trends Ecol. Evol.*, 11, 367–372.

189. Yang, Z. (1995) A space-time process model for the evolution of DNA sequences. *Genetics*, 139, 993–1005.

190. Kubatko, L.S., Carstens, B.C. and Knowles, L.L. (2009) STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*, 25, 971–973.

191. Heled, J. and Drummond, A.J. (2010) Bayesian inference of species tree estimation using maximum likelihood for gene trees. *Mol. Biol. Evol.*, 19, 98–107.

192. Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N.A. and RoyChoudhury, A. (2012) Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.*, 29, 1917–1932.

193. Weinreich, D.M., Delaney, N.F., DePristo, M.A. and Hartl, D.L. (2006) Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*, 312, 111–114.

194. Bridgesham, J.T., Orltund, E.A. and Thornton, J.W. (2009) An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature*, 461, 515–519.

195. Noort, S., Taylor, M.C., Russell, R.J., Jermin, L.S., Jackson, C.J., Oakeshott, J.G. and Scott, C. (2012) Intramolecular epistasis and the evolution of a new enzymatic function. *PLoS One*, 7, e39822.

196. Galteri, N. (2001) Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol. Biol. Evol.*, 18, 866–873.

197. Huelsenbeck, J.P. (2002) Testing a covariotide model of DNA substitution. *Mol. Biol. Evol.*, 19, 841–847.

198. Cotty, S.M., Minh, B.Q., Bean, N.G., Holland, B.R., Tuke, J., Jermin, L.S. and van Haeseler, A. (2020) GHOST: recovering historical signal from heterotachously evolved sequence alignments. *Syst. Biol.*, 69, 249–264.

199. Jermin, L.S., Ho, J.W.K., Lau, K.W. and Jayaswal, V. (2009) SeqVis: A tool for detecting compositional heterogeneity among aligned nucleotide sequences. In: Posada, D. (ed). *Bioinformatics for DNA Sequence Analysis*. Humana Press, Totowa, pp. 65–91.

200. Kelly, C. (1994) A test of the Markovian model of DNA evolution. *Biometrics*, 50, 653–664.

201. Squartini, F. and Arndt, P.F. (2008) Quantifying the stationarity and time reversibility of the nucleotide substitution process. *Mol. Biol. Evol.*, 25, 2525–2535.

202. Huson, D.H. and Bryant, D. (2006) Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.*, 23, 254–267.

203. Baptiste, E., van Iersel, L., Janke, A., Kelchner, S., Kelk, S., McInerney, J.O., Morrison, D.A., Naklich, L., Steel, M., Stougie, L. et al. (2013) Networks: expanding evolutionary thinking. *Trends Genet.*, 29, 439–441.

204. Xia, X.F., Xie, Z., Salem, M., Chen, L. and Wang, Y. (2003) An index of substitution saturation and its application. *Mol. Phylogenet. Evol.*, 26, 1–7.

205. Fischer, M. and Steel, M. (2009) Sequence length bounds for resolving a deep phylogenetic divergence. *J Theor. Biol.*, 256, 247–252.

206. Mossel, E., Steel, M. and Gascuel, O. (2005) How much can evolved characters tell us about the tree that generated them? In: Gasuel, O. (ed). *Mathematics of Evolution and Phylogeny*. Oxford University Press, NY, pp. 384–422.

**APPENDIX 1. COMMON PHYLOGENETIC ASSUMPTIONS**

Phylogenetic methods developed to analyze alignments of nucleotides, codons and amino acids typically make some of the following modeling assumptions:

i. The sequences evolved along the edges of a single bifurcating tree.

ii. Each site has a fixed rate of change that does not change across the tree; however, note that this rate can be zero (invariable sites).

iii. The evolutionary processes operating at the variable sites follow independent and identically distributed Markov processes.

iv. The evolutionary processes operating at the variable sites were stationary, reversible, and homogeneous over time.

Other modeling assumptions are often made (e.g. that rate heterogeneity across the sites can be modeled accurately using a discrete $G$ distribution (188) or non-parametric, probabilistic model of rate heterogeneity across sites (189); the latter strategy is more flexible, allowing multimodal distributions of rate heterogeneity across sites to be accommodated (14)), but they are not considered here. In addition to these assumptions, it is typically assumed that the data are phylogenetically informative—that is, the historical signal has not decayed enough to make it pointless using the data in a phylogenetic study. We now describe the modeling assumptions with a minimum of mathematical and statistical terminology.

*The assumption of tree-like evolution*: Building on graph theory, this assumption—the tree-likeness assumption—states that every site in an alignment has evolved on the same rooted phylogenetic tree. The assumption applies equally to single-gene and multi-gene datasets. Some relatively recent methods (190–192) relax this assumption by assuming a single species tree but allowing the gene trees to vary from the species tree due to incomplete lineage sorting. However, if, for example, some of the sequences in the data were involved in recombination events or introgression, then the phylogenies of the sites on either side of the break points will be different and the assumption of a single underlying tree will be violated.

*The assumption of each site having a fixed rate of change that does not change across the tree*: Building on knowledge of biochemistry and natural selection, the sites in a nucleotide sequence can be divided into those that are variable and those that are invariable. Here the variable sites are those...
that are free to change while the invariable sites are those that cannot change, usually due to all changes being deleterious.

A key point here is that it is not possible, in general, to distinguish between sites that were able to change but have not done so, and sites that were unable to change due to selection. For this reason, we are limited in our ability to conclude whether sites are truly invariable. Therefore, many phylogenetic methods are designed to estimate the proportion of invariable sites. In fact, it is highly recommended to assume that a proportion of sites are invariable.

The assumption that sites are either variable or invariable may be violated by the data if the selective constraints at the sites change over time. Under this scenario, a variable site might become invariable within, for example, one lineage but remain variable in all other lineages. Our understanding of the prevalence of evolving selective constraints is not good, but intra- and inter-molecular epistasis might be contributing factors. Intra- and inter-molecular epistasis is known to constrain the order and reversibility of amino-acid substitution (193–195) but it has not yet been incorporated in phylogenetic methods (for exceptions, see 181,196–198).

The assumption of independent and identically distributed Markovian processes: Based on probability theory, this assumption is required because it gives us the statistical power that is needed to compare evolutionary hypotheses within a statistical framework. A process is Markovian if the conditional probability of change at a site in a sequence depends only on the current state and is independent of its earlier states. In other words, a Markov process has no memory. The ‘independency of the Markovian processes’ refers to the processes at the variable sites being independent of one another, and the ‘distributions of the Markovian processes being identical’ refers to the conditional probability of a change from state $i$ to state $j$ (e.g. from A to G) being the same irrespective of what variable site is considered. In other words, the assumption states that the variable sites evolve independently of one another and that the evolutionary processes are not only Markovian but also identical across the variable sites.

The assumption of stationary processes: A Markov process is stationary when the marginal probabilities of the states that the process might take remain the same irrespective of time. In mathematical terms, a Markov process is stationary when the following condition holds:

$$\text{Pr}(X(t) = j) = \pi_j = f_j,$$

where $X(t)$ denotes the state of the Markov process at time $t$, $j = A, C, G, T$ (for DNA), $\pi_j$ denotes the marginal probability of state $j$ in the process and $f_j$ represents the relative frequency of state $j$ at the variable sites of the ancestral sequence. Usually, $\pi$ is called the stationary distribution of the process (or the Markov model approximating this process). In mathematical terms, it is a vector.

The assumption of stationary processes is made as it reduces the number of parameters that needs to be optimized and, therefore, reduces the amount of time required to optimize them. If the sequences in an alignment have evolved under stationary conditions, then they are unlikely to have acquired different compositions of the states (e.g. nucleotides). On the other hand, if the sequences are compositionally heterogeneous, then they are unlikely to have evolved under stationary conditions. Evolution under non-stationary conditions now appears to be a norm rather than an exception (e.g. (142,199)).

The assumption of reversible processes—A Markovian process is reversible if the probability of sampling state $i$ from the stationary distribution (i.e. $\pi$) and going to state $j$ is the same as the probability of sampling state $j$ from the stationary distribution and going to state $i$. In mathematical terms, a Markov process is reversible when the following condition holds:

$$\pi_i R_{ij} = \pi_j R_{ji},$$

where $R_{ij}$ denotes the instantaneous rate of change from state $i$ to state $j$, and $i, j = A, C, G, T$ (for DNA). In mathematical terms, $R$ is a matrix. If the evolutionary process is consistent with this ‘balancing equation’, further reductions in the number of parameters can be made. An important insight into the Markovian process is that the likelihood of a tree for a given phylogenetic dataset can be computed without knowing the ancestral sequence for these data.

While the assumption of reversible processes is critical for most model-based phylogenetic methods, it is more difficult to ascertain whether sequences have evolved under reversible conditions. It is, however, clear that evolutionary processes cannot be reversible if they are non-stationary, so evolution under non-reversible conditions now also appears to be a norm, rather than an exception. A biochemical study of the evolutionary trajectories between two enzymes, atrazine chlorohydrolase (AtzA) and melamine deaminase (TriA), corroborates this picture: the optimal order of substitutions from AtzA to TriA and vice versa is the optimal order of substitution from TriA to AtzA (195).

The assumption of homogeneous processes: A Markovian process is homogeneous in time if the conditional probabilities of substitutions are constant over time. When a Markov process is consistent with this assumption, it is easy to estimate the conditional probability of change ($P$) over a given time ($\Delta t$). In mathematical terms, the following expression is used:

$$P(\Delta t) = e^{R \times \Delta t}.$$

Here, $P$ is a matrix representing the joint probability of state $i$ at the beginning of an edge in a tree and state $j$ at the end of this edge. By combining sets of $P$ matrices, one per edge, in a manner that reflects the topology of the tree (i.e. $T$), it is possible to calculate the likelihood of the tree and the model (i.e. $R$), given the phylogenetic data.

Although convenient, the assumption of homogeneous processes over time is unlikely to be realistic. This is because accumulation of substitutions over time is likely to be gradual and occurring along the edges, rather than being punctuated and occurring only at the nodes in the tree. Sometimes, the term ‘homogeneous processes’ is used to describe evolutionary processes operating along diverging lineages in a phylogenetic tree. To avoid confusion, it is sensible to write ‘time-homogeneous processes’ when this is the intended meaning and to write ‘homogeneous processes across lineages’ when this is the intended meaning.
**Figure A2.** Illustrating the merits of the fifth step in the new phylogenetic protocol. Panels A–C show the average nucleotide composition at first, second and third codon positions of the sequences (144 sequences, 413 459 codons) originally examined by Misof et al. (11). Each dot represents the nucleotide composition of a single sequence. The spread of dots in panels (A–C) reveals compositional heterogeneity at first, second and third codon positions, indicating these data violate important assumptions underlying most phylogenetic methods. The tetrahedral plots were generated using SeqVis (139). Panel D shows a two-tipped tree with an ancestral sequence and two diverged sequences. Panels E and F show the divergence matrix for these sequences at time 0 and time t. Each number in a cell of a divergence matrix corresponds to the number of sites with nucleotide i in one sequence, and nucleotide j in the other. Panels G–I show the PP plots for the data already analyzed in panels A–C. A total of 10 296 tests were done for each of the three codon positions using Homo 1.3 (http://www.csiro.au/homo/—a new version of Homo is now available (54)).

**APPENDIX 2. CASE STUDY**

To illustrate the relevance and benefits of the fifth step in the new phylogenetic protocol (i.e. assess phylogenetic assumptions), we surveyed the phylogenetic data used to infer the evolution of insects (11).

The tetrahedral plots in panels A–C of Figure A2 reveal that the nucleotide composition at the three codon positions is heterogeneous (most clear in panel C and least clear in panel B), implying that the evolutionary processes that operated at these positions are unlikely to have been time-reversible. However, the plots are deceptive because the presence of constant sites (i.e. sites with the same nucleotide or amino acid) in the data can mask how compositionally dissimilar the sequences actually are.

To learn how to resolve this issue, it is necessary to focus on the evolution of two sequences on a tree (panel D of Figure A2) and the corresponding divergence matrix, $N$, at time 0 (panel E of Figure A2) and at time $t$ (panel F of Figure A2). At time 0, the two sequences are beginning to diverge from one another, so the off-diagonal elements of $N$ are all zero. Later, $N$ may look like that in panel F of Figure A2. The off-diagonal elements of $N$ are now greater than zero, and matching off-diagonal elements of $N$ might differ (i.e. $n_{ij} \neq n_{ji}$). The degree of divergence between the two sequences can be inferred by comparing the off-diagonal elements to the diagonal elements, while the degree of difference between the two evolutionary processes can be inferred by comparing the above-diagonal elements of $N$ to the below-diagonal elements of $N$. If the two evolutionary processes were the same, the matching off-diagonal elements in panel F of Figure A2 would be similar. A lack of symmetry (i.e. $n_{ij} \neq n_{ji}$) implies that the evolutionary processes along the descendant lineages may be different.

A matched-pairs test of symmetry (138) can be used to determine whether this observed deviation from symmetry...
is statistically significant. Panels G–I of Figure A2 show the distributions of the observed and expected $P$ values from these tests for the data assessed in panels A–C of Figure A2. If the evolutionary processes operating along the descendant lineages are identical, the dots would be distributed along the diagonal line of a PP plot. However, as the dots in each plot fall far from this line, there is little evidence supporting the assumption of evolution under homogeneous conditions. The same is the case for the corresponding amino acid alignment (not shown), so it would be unwise to assume that the data evolved under time-reversible conditions. A more complex evolutionary process might explain these data. However, such methods were not available at the time when Misof et al. (11) analyzed these data, and that is still the case.