Palaeoentomology 007 (2): 176–183
https://www.mapress.com/pe/

Comments and Perspectives

ISSN 2624-2826 (print edition)
PALAEOENTOMOLOGY
ISSN 2624-2834 (online edition)

https://doi.org/10.11646/palaeoentomology.7.2.2
http://zoobank.org/urn:lsid:zoobank.org:pub:39A6126D-BB4C-44F8-B32F-2FB957FE6F91

Resolving incongruences in insect phylogenomics: A reply to Boudinot *et al.* (2023)

CHEN-YANG CAI[1, *], ERIK TIHELKA[2, 3, *], DAVIDE PISANI[2, 4] & PHILIP C. J. DONOGHUE[2]
[1]*State Key Laboratory of Palaeobiology and Stratigraphy, Nanjing Institute of Geology and Palaeontology, and Centre for Excellence in Life and Paleoenvironment, Chinese Academy of Sciences, Nanjing 210008, China*
[2]*School of Earth Sciences University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol, BS8 1TQ, UK*
[3]*Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge, CB2 3EQ UK*
[4]*School of Life Sciences University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol, BS8 1TQ, UK*
cycai@nigpas.ac.cn; https://orcid.org/0000-0002-9283-8323
et532@cam.ac.uk; https://orcid.org/0000-0002-5048-5355
davide.pisani@bristol.ac.uk; https://orcid.org/0000-0003-0949-6682
phil.donoghue@bristol.ac.uk; https://orcid.org/0000-0003-3116-7463
**Corresponding authors*

Abstract

Over the last two decades, advances in molecular phylogenetics have established a new understanding of beetle phylogeny. However, some historically contentious relationships, particularly among early-diverging beetle clades, remain to be resolved. In a recent paper (Cai *et al*., 2022), we identified model-dependent signals in beetle phylogeny and showed how the removal of the most compositionally heterogeneous sites, in combination with the use of across-site compositionally heterogeneous models leads to results that are more congruent with the distribution of morphological characters and the beetle fossil record. In their reply, Boudinot *et al*. (2023) suggested that our analyses are affected by a range of shortcomings, encompassing almost every aspect of our study. Unfortunately, the arguments presented by Boudinot *et al*. (2023) are based on misinterpretation of the results of statistical tests, as well as misconceptions concerning substitution models, model testing and its role in phylogenomics. Here we clarify these misconceptions and show that the critiques raised by Boudinot *et al*. (2023) have no merit.

Keywords: Coleoptera, phylogenetics, compositional heterogeneity, model testing, systematic bias

Introduction

Over the past decade, advances in next-generation sequencing opened up a new era for studying the evolution of insects. The plummeting cost of sequencing reads, the increasing affordability and portability of sequencing platforms, advances in extracting genomic data from museum specimens and importantly, coordinated sequencing efforts across the taxonomic spectrum, have yielded a trove of data for all major groups of this most prolific branch of the animal tree (Trautwein *et al*., 2012; Misof *et al*., 2014; Kjer *et al*., 2016). In recent years, several studies addressed the phylogeny of beetles using truly massive datasets, including multiple gene markers, transcriptomes, and genomes (McKenna *et al*., 2015, 2019; Toussaint *et al*., 2017; Zhang *et al*., 2018). In our own recent contribution (Cai *et al*., 2022), we explored the impacts of across-site compositional heterogeneity on inferring beetle phylogeny and proposed updates to the higher classification of Coleoptera, to incorporate relationships that are supported by our results as well as morphological evidence.

We are pleased that our work has attracted widespread interest including that of Boudinot *et al.* (2023, first published online in 2022) who have undertaken a sweeping appraisal of our study and found it wanting on almost all counts. On behalf of the original authors of the study, we herein explore the key points raised by Boudinot and colleagues and show how they are based on misinterpretations of our results and misunderstandings of models, model testing, and model-based phylogenetics.

Material and methods

To examine model adequacy, posterior predictive analyses were performed using PhyloBayes MP1.7 under the GTR model. To showcase how the CAT-GTR model adapts the number of compositional categories to fit the data at hand, we simulated 105,000 sites, across-site compositionally homogeneous amino acid datasets. The datasets were

 Submitted: 1 Apr. 2024; accepted by D.-Y. Huang: 11 Apr. 2024; published: 19 Apr. 2024

simulated using the software Elynx (Schrempf, 2019) under WAG+G (4 rate categories and a = 0.6) and the 8-taxon tree in Fig. 1B. The ten datasets were analysed in PhyloBayes MPI 1.9 under CAT-GTR+G and GTR+G. For the CAT-GTR+G analyses 3 runs were completed to test for convergence. For the GTR+G dataset only one run was completed. The chains from the GTR+G analyses were used to perform posterior predictive tests (burn in = 5,000, subsampling frequency = 50) and ensure that the simulated data were, indeed, across-site compositionally homogeneous. The trace files from the CAT-GTR analyses were parsed and the number of categories used for each one of the 30,000 cycles copied to a text file. The text files were run using a custom-made Perl script that calculated the mode for the number of categories for each chain of each dataset. We tested how these values changed as the burin and the number of cycles increased (Table 1, Fig. 1B). To measure central tendency, we used the mode because it is expected that also after burn in Bayesian analyses will continue to explore the parameter space and hence, while the majority of cycles will use one site frequency category, some cycles (a minority of them) will still test the use of larger number of site frequency categories with the average expected to be slightly higher than 1. To compare the fit of analyses conducted with partitioned models and the compositionally heterogeneous LG+C60+F+G, analyses were run in IQ-TREE 1.6.12. All analysed files and software output are publicly available from the Dryad online repository: 10.5061/dryad.bzkh189h7.

Discussion

Phylogenies can be simultaneously closely comparable and significantly different

In Cai *et al*. (2022), we analysed the sequence data under both across-site composition-homogeneous and heterogeneous models using two state-of-the-art software packages implementing, respectively, Maximum Likelihood (ML) and Bayesian inference. Boudinot *et al*. (2023) criticised our study for not repeating each of our ML analyses ten times to ensure our results were consistent. Boudinot *et al*. (2023) went on to argue that our across-site composition-homogeneous and heterogeneous analyses achieved very comparable results, contrary to our 'exaggerated claims'. Indeed, the phylogenies we obtained under different models are broadly comparable, as is invariably the case because not all bipartitions in a tree are equally difficult to resolve and most will be identically resolved by all models. However, some relationships are harder to resolve and it is these 'tricky nodes' that different models and differently curated datasets usually resolve differently. Accordingly, it is clear that phylogenies can be simultaneously broadly similar, yet significantly different. Overall, the similarity of the trees from all of our analyses indicates that our original IQ-TREE and PhyloBayes analyses are reliable since they converged on the same region of tree space. Hence, there is little or no rationale to repeat individual ML analyses

TABLE 1. The mode of site frequency categories inferred by CAT-GTR for 10 simulated across-site compositionally homogeneous datasets. This table illustrates how the number of site frequency categories used in CAT-GTR analyses of simulated across-site compositionally homogeneous datasets tends to one as the analysis progress, and larger burn ins and numbers of post-burn in samples are used.

Burn in / Number of Cycles	Number of estimated site frequency categories
500 / 1,000	2.0
1,000 / 2,000	2.2
2,000 / 4,000	1.7
3,000 / 6,000	1.4
4,000 / 8,000	1.4
5,000 / 10,000	1.4
6,000 / 12,000	1.4
7,000 / 14,000	1.0
8,000 / 16,000	1.0
9,000 / 18,000	1.0
10,000 / 20,000	1.0
11,000 / 22,000	1.0
12,000 / 24,000	1.0
13,000 / 26,000	1.0
14,000 / 28,000	1.0
15,000 / 30,000	1.0

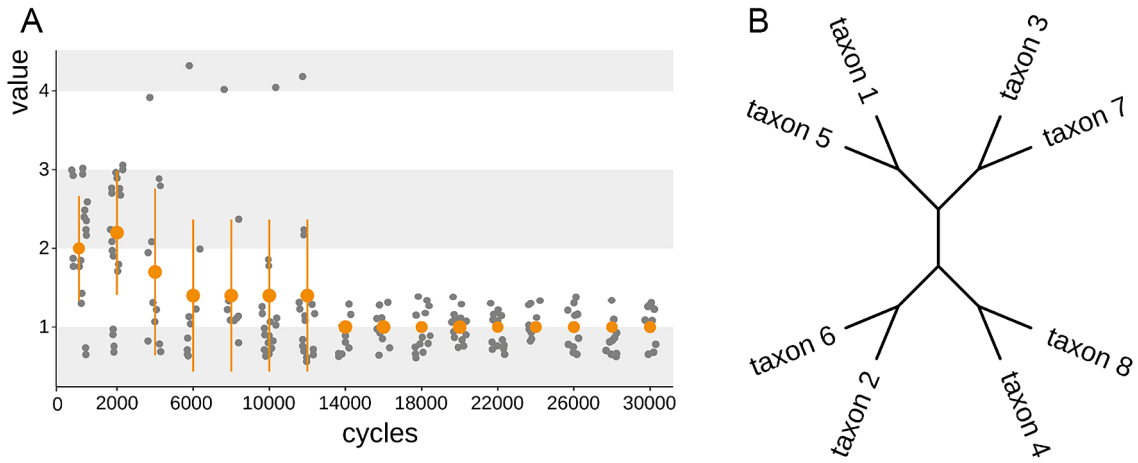

FIGURE 1. Despite being an infinite mixture model, CAT-GTR does not necessarily use an arbitrarily high number of categories. On the contrary, GTR is nested within CAT-GTR; a CAT-GTR model with one frequency category is a different way to refer to GTR. It can be expected that in CAT-GTR analyses of across-site compositionally homogeneous datasets, CAT-GTR would autotune, selecting one frequency category as optimal thus simplifying itself to a GTR model. Our simulations show that this is clearly the case demonstrating that CAT-GTR cannot even overfit across-site compositionally homogeneous datasets. **A**, With simulated compositionally homogeneous amino acid alignments, CAT-GTR correctly autotuned with the number of inferred categories decreasing as better convergence is achieved. From approximately 14,000 cycles onward the mode of the distribution being equal to one. Which indicates that from that point onward the CAT-GTR analyses successfully simplified to GTR analyses (as expected) in a Bayesian framework. **B**, The analysis correctly reconstructs the target tree.

10 times to ensure convergence, as Boudinot *et al*. (2023) advocated. Indeed, in an era of green computing, who does that?

Irrespective of which beetle phylogeny is correct, we contend that the inference of incongruent trees when a dataset is curated or analysed in different ways is the most interesting aspect of phylogenomic research, and how to make sense of these differences should be the main focus of phylogenomic studies, this is what our study did when presenting trees that are similar (but also incongruent) to those inferred from the same data by previous studies.

Testing for compositional heterogeneity

Boudinot *et al*. (2023) focused much of their criticism on Cai *et al*. (2022) and claimed that the analysed dataset is compositionally heterogeneous across sites and on their model selection protocol. Cai *et al*. (2022) used across-site compositional heterogeneous models to analyse the data. However, Boudinot *et al*. (2023) suggested that there is no evidence that the analysed dataset is compositionally heterogeneous and that as a consequence it should have been analysed using site-homogeneous models. However, Cai *et al*. (2022) did not assume that the data were compositionally heterogeneous, as Boudinot *et al*. (2023) claimed. Rather, Cai *et al*. (2022) performed standard model selection showing the data best fit across-site compositionally heterogeneous models, even after using BMGE (Criscuolo & Gribaldo, 2010), and despite the application of BaCoCa (Kück & Struck, 2014) on the filtered dataset could not reject the hypothesis that the remaining data were compositionally homogeneous. Boudinot *et al*. (2023) interpreted the inability of BaCoCa to reject the hypothesis that the data might be compositionally homogeneous as a rejection of the alternative hypothesis, concluding that there was no need to run model selection tests and that the data should just be analysed using an across-site compositional homogeneous model. This is not how the results of statistical tests should be interpreted. It is rather the case that BaCoCa confirms that the sites retained by BMGE are the least heterogeneous, to the point that the hypothesis that the data might be across-site compositional homogeneous could not be rejected. However, failure to reject the null hypothesis does not imply rejection of the alternative hypothesis, and it does not automatically follow that the retained data are therefore across-site compositionally homogeneous, as Boudinot *et al*. (2023) suggested. A failure to reject the null hypothesis simply means that the observations are potentially consistent with the null hypothesis, but that conclusion is entirely contingent on the power of the test. The use of a different, more powerful, test could still reject the null hypothesis. The approach of Cai *et al*. (2022), where a model selection step was performed on the BMGE-filtered data, is conservative and preferrable. This is because, if Boudinot *et al*. (2023) were correct, model selection would have

identified an across-site compositionally homogeneous model as best fit (also see below). That is, the approach of Cai *et al*. (2022) did not preclude the use of across-site compositionally homogeneous models, differently from the approach of Boudinot *et al*. (2023), the latter of which would unnecessarily preclude the use of across-site compositionally homogeneous models.

Cai *et al*. (2022) performed model-testing under both Bayesian and ML inference, using 10-fold cross validation in PhyloBayes and BIC in IQ-TREE. Despite the fact that the across-site compositionally heterogeneous models implemented in PhyloBayes and IQ-TREE are different, both analyses found that the data best fit across-site compositionally heterogeneous models (CAT-GTR tested in PhyloBayes and C10 tested in IQ-TREE—see Cai *et al*., 2022). This result implies that, even after the exclusion of the compositionally most heterogeneous sites, the use of across-site compositionally heterogeneous models in Cai *et al*. (2022) were fully justified.

Boudinot *et al*. (2023) further criticised Cai *et al*. (2022) for not using Posterior Predictive Analyses (PPA) to further test the adequacy of the models used and test whether the data were across-site compositionally heterogeneous. This is a valid point and following their suggestion, we performed PPA analyses of the 16,206-site datasets under GTR. The PPA results, are in complete agreement with the model selection results showing that the BMGE-filtered dataset remains strongly compositionally heterogeneous across sites (*Z*-score PPA-DIV$_{GTR}$ = 96.11). Note that for a model to adequately describe the data the *Z*-score should be between +2 and –2; see Giacomelli *et al*., 2022), and even taking a relaxed approach to the interpretation of *Z*-scores (as Giacomelli *et al*., 2022 did), *Z*-scores higher than 10 identify an extremely poor fit of the model to the data. This result is not unexpected since PPA tests are more powerful than standard (Chi-square based) statistical tests like those implemented in BaCoCa and BMGE, and corroborate our conclusion that the 16,206-site data set is across-site compositional heterogeneous, and should be analysed using models that can account for across-site compositional heterogeneity, controlling for the concerns raised by Boudinot *et al*. (2023). We can only conclude that the focal criticism of Boudinot *et al*. (2023) was based on a misinterpretation of the results of statistical analyses and the application of model selection strategies and is without merit.

Model abuse

Evidently, the claim by Boudinot *et al*. (2023) that Cai *et al*. (2022) 'abused' across-site compositionally heterogeneous models, is unfounded; the data are compositionally heterogeneous. Nevertheless, the concern over model complexity is a common misconception and so we take this opportunity to clarify why CAT-GTR cannot be "abused". When a CAT-GTR analysis is performed, the across-site compositional heterogeneity in the data is estimated and the number of site-frequency categories needed to adequately describe it are consequently inferred. Because of the hierarchical nature of substitution models, the GTR model is nested within a CAT-GTR model, as GTR can be represented as the special case of CAT-GTR where the number of site-frequency categories (nCAT) is equal to 1. If a dataset is across-site compositionally homogeneous, it is expected that, as convergence improves, the sites will tend to be clustered in a single site-frequency category more and more frequently. As the parameter space is explored, in an MCMC analysis, we do not expect that for all cycles nCAT = 1, yet the mode of the sampled points will more closely approximate 1 as the analysis progresses, and more cycles are analysed using a nCAT1-GTR (*i.e*., a GTR) model, simply because for across-site compositionally homogeneous data, this model is expected to fit best. To express this another way, as the optimal number of categories is tuned, CAT-GTR would simplify into an across-site compositionally homogeneous GTR model. If CAT-GTR is used to analyse an across-site compositional homogeneous dataset, PhyloBayes would effectively analyse the data under a GTR, rather than using an over-parameterized model with hundreds of site-frequency categories.

We demonstrate this here with a simple simulation study where ten simulated across-site compositionally homogeneous datasets (see methods for details) are analysed under CAT-GTR in PhyloBayes. First, for each simulated dataset we performed a posterior predictive analysis under GTR+G to make sure that the simulated data were indeed, across-site compositionally homogeneous. The average Z-score for the posterior predictive analyses (PPA$_{GTR}$ = –0.58; SD=0.3) indicates that the simulated data can be adequately modelled by a single GTR matrix without the need of CAT. That is, the simulated data are across-site compositionally homogeneous. We then performed CAT-GTR analyses (three chains) of these datasets and parsed the trace files to extract the number of site-frequency categories used in each cycle. For each dataset we calculated the mode for the site frequency category parameter (Nmode in the tracefile), across the three chains. We started using a burin of 500 and considering the first 1,000 cycles only. After that, we incremented both the burn in and the number of cycles considered (see Table 1), to show how Nmode changed as convergence progressed. After 8,000 cycles, the mode for the site frequency category parameter (Nmode) stabilised to 1 and no longer changed (Fig. 1A). These results are as expected and illustrate that if the analyses of across-site compositionally homogeneous datasets are run for a sufficient number of cycles, CAT-GTR simplified itself into a GTR model avoiding to overfit the data. Despite

setting PhyloBayes to use CAT-GTR, analyses of the simulated across-site compositionally homogeneous data are performed under the GTR model.

This simple exercise proves that CAT-GTR cannot be abused and it is safe to use even when analysing across-site compositionally homogeneous data. (Note that the conclusions from this example do not necessarily extend to models using fixed number of categories, such as the C10-C60 and to CAT-Poisson since, in contrast to CAT-GTR, these models do not simplify into a GTR model). But discussing the details of how different CAT-based models behave with across-site compositionally homogeneous datasets is outside the scope of this reply. It should however be noted that, while using CAT-GTR with across-site compositionally homogeneous data is not a problem since GTR is nested within CAT-GTR, the opposite is not true. Using GTR (or LG) when the data are across-site compositionally heterogeneous is problematic because these models do not have the ability to partition the data into multiple site-frequency categories when necessary to model across-site compositionally heterogeneous data, casting doubts on results obtained using such models when they disagree with results obtained using CAT-GTR.

The CAT model, overparameterization, and model test fairness

Boudinot *et al*. (2023) presented a discussion of the flaws of the CAT model. They started by explaining that CAT is an "infinitely complex" [*sic*] model and claim that it may overfit the data. While there is abundant evidence that under-parametrisation (underfitting) is a problem in phylogenetics, at the least in a Bayesian setting, there is no evidence that over-parameterisation (*i.e*., overfitting) can lead to tree reconstruction artefacts (Lemmon & Moriarty, 2004; Baños *et al*., 2023). Studies to date have concluded that, at the least with Bayesian analyses, at worst, over-parameterisation depresses support values, a problem that seems to disappear with increasing dataset size (Huelsenbeck & Rannala, 2004; Lemmon & Moriarty, 2004; Baños *et al*., 2023; Fabreti & Höhna, 2022, Giacomelli *et al*., 2022). Furthermore, while CAT is an infinite mixture model, when data are analysed, the actualised CAT-based model will use a finite number of parameters, which can be small (down to one category only), as the model dataset analysis of Fig. 1 shows.

Boudinot *et al*. (2023) claimed that it is unfair to compare models such as CAT-GTR against GTR because CAT-GTR has more parameters and will therefore achieve better likelihood scores. This view ignores the fact that model test statistics account for extra parameters (using penalisation scores), and models do not achieve best-fit status simply because they have more parameters. Cai *et al*. (2022) performed model testing under both Bayesian and ML inference. Testing CAT-GTR *vs*. GTR in a Bayesian setting (with 10-fold cross-validation) they concluded that CAT-GTR fit better than GTR in Bayesian analyses. Under ML using BIC, Cai *et al*. (2022) tested C10 (the across-site compositionally heterogeneous models with the least number of site-frequency categories—*i.e*. the least number of parameters) against the GTR, WAG and LG models, concluding that C10 fit better than the other tested models. This result demonstrates that across site compositionally heterogeneous models do not fit better than across-site compositionally homogeneous models because they have more parameters. The C10+G model used in IQ-TREE has 274 df (including branch lengths that need estimation when performing model tests). This is significantly less than the more poorly-fitting GTR+G model which has 482 df, 189 of which represent AA transition rates. Thus, it is not the highest number of parameters that causes across-site compositionally heterogeneous models to best-fit the data, it is instead whether or not a model can account (at least to some extent) for the across-site compositional heterogeneity in the data, which as shown by our PPA analysis (and despite results from BaCoCa), was substantial.

Model adequacy and the relative utility of CAT and partitioned models

Boudinot *et al*. (2023) suggested that we should have compared CAT-based models against partitioned models (compositionally homogeneous models applied to partitions of the data that can correspond to one or more genes), as these fit the data much better than single GTR matrices. It is unclear whether this assertion is universally true. Feuda *et al*. (2017) showed that, with two empirical datasets, partitioned models do not achieve greater model adequacy than single GTR models when the data are compositionally heterogeneous across-sites. Furthermore, it has been shown that when comparing mixture models with partitioned ones, results of model tests can be expected to have a bias favouring partitioned models (Crotty & Holland, 2022). Nonetheless, we agree with Boudinot *et al*. (2023) that testing the relative fit of partitioned models and across-site compositionally heterogeneous models would be interesting and we do so here, using the output from IQ-TREE to compare the fit of the best partitioned model against the best across-site compositionally heterogeneous model, and GTR+F+G (a representative across-site compositionally homogeneous model). We find that LG+C60+F+G (the overall best fit across-site compositionally heterogeneous model among those we now tested in IQ-TREE) achieves a BIC of 2,624,657.923 (df = 829). The best partitioned scheme (splitting the genes in 19 independently modelled partitions), has a BIC of 2,725,723.787 (df = 1031), while GTR+F+G has a BIC of 2,752,617.845 (df = 958). Despite the partitioned model

having the highest number of parameters and GTR+F+G having the second highest number of parameters, the partitioned model is the worst-fit, with the across-site compositionally heterogeneous LG+C60+F+G being best-fit. Evidently, Boudinot *et al*. (2023) are incorrect. Data partitioning does not achieve a better fit in this instance (not even with reference to GTR+G) with the across-site compositionally heterogeneous LG+C60+F+G achieving best fit despite having fewer parameters. Our result and those of Feuda *et al*. (2017) agree in illustrating that, contrary to previous suggestions (Whelan & Halanych, 2017), partitioned models cannot be used in place of across-site compositionally heterogeneous models, simply because they model a different form of heterogeneity.

Fossil calibrations

Boudinot *et al*. (2023) critiqued some of the 57 fossil calibrations used in our molecular clock analyses. This was made possible because our study was the first to follow best practices in justifying fossil calibrations (Parham *et al*., 2011). We do not agree with their interpretations of the fossil record, as it pertains to clock calibration, but at least Boudinot *et al*. (2023) can establish their differences of opinion and, if they wish, reproduce the divergence time analyses presented in Cai *et al*. (2022) with their preferred suite of calibrations. It is not generally possible to discriminate statistically among competing parameters, such as clock model, partition strategy, and calibration density in molecular dating, and it is our common practice to explore this parameter space (*e.g*., dos Reis *et al*., 2015; Morris *et al*., 2018; Betts *et al*., 2018), as we did in Cai *et al*. (2022), combining the results of parallel analyses that effectively integrate over the choice of uninformed parameter selection. Boudinot *et al*. (2023) instead advocate the unfounded assertion of a single, preferred, divergence time analysis, but this hardly represents the uncertainty in the evolutionary timescale of beetles. We do not anticipate that our divergence time analysis will be the last attempt to establish the evolutionary timescale of beetle diversification. Nevertheless, we remain confident that we have conducted the most extensive molecular clock analysis of beetles to date and that our calibrations will serve as a foundation for future research.

Classification

Lastly, Boudinot *et al*. (2023) stated that our revised classification of Coleoptera is "based on" CAT-GTR analyses. We have shown that there is nothing wrong with CAT-GTR and that Boudinot's *et al.* (2023) assertions are unsubstantiated by further analyses. Moreover, we point out that an update of higher coleopteran systematics has been long overdue and does not rely on the results of a single analysis. On the contrary, congruent results concerning the higher relationships of coleopteran clades have been consistently recovered in other recent phylogenomic studies (Zhang *et al*., 2018; McKenna *et al*., 2019), and many of the proposed changes were suspected for decades. Furthermore, all proposed taxonomic changes are supported by morphological justifications and revised diagnoses, discussed at length in the supplementary material of Cai *et al*. (2022). We see an update of beetle classification as necessary for maintaining clarity and reflecting our current best understanding of beetle evolution (Bouchard *et al*., 2024).

Conclusions

We agree with Boudinot *et al*. (2023) that there is no doubt that we still have a long way to go before all the major relationships in coleopteran phylogeny are resolved to everyone's' satisfaction. This has been true of all past analyses and will persist for those yet to come. Nonetheless, it is indeed promising that phylogenomic analyses of independent datasets conducted over recent years are converging on congruent topologies for many major beetle groups (Zhang *et al*., 2018; McKenna *et al*., 2019; Cai *et al*., 2022). This naturally leaves the question of how to handle the remaining incongruencies—tricky relationships that are challenging to address with currently available datasets and methods. We cannot expect to resolve difficult phylogenetic problems by simply counting how many analyses support a particular topology in a phylogenetic popularity contest; science does not proceed democratically, through majority voting. Evidence needs to be weighed and, in our opinion, the only objective measure we have is model fit. Consensus across phylogenetic analyses using models of variable fit is not a useful guide to progress but, rather, leads to self-reinforcing stagnation of scientific enquiry, as this approach inherently fail to allow rejecting conclusions inferred using outmoded models and approaches. If progress is to be made towards the ideal of a coleopteran phylogeny that everyone can agree upon, we advocate a hypothesis-driven approach, both to testing the models we used in building phylogenies, and in testing between those phylogenies and the evolutionary timescales based upon them. With reference to phylogenomic analyses, given the undeniable ability of CAT-GTR to autotune and accommodate dramatic differences in compositional heterogeneity, from across-site compositionally homogeneous dataset (Fig. 1) to highly heterogeneous datasets (compare for example measurements of fit of CAT-GTR and other models in Feuda *et al*., 2017 and Giacomelli *et al*., 2022), we advocate the use of CAT-GTR over less flexible models such as LG, LG+C60, when testing hypotheses of phylogenetic relationships

for Coleoptera, as CAT-GTR can be expected to fit most datasets better than other currently available models (*e.g*., Cai *et al*., 2023; Cai, 2024).

Acknowledgements

We thank all 13 co-authors of Cai *et al*. (2022) for their suggestions and encouragement, Professor Gergely Szollosi for helpful comments, and two anonymous reviewers for their valuable suggestions. This work has been supported by the National Natural Science Foundation of China (41925008, 42222201, 42288201). D.P. was funded by a University of Bristol, University Research Fellowship.

References

Baños, H., Susko, E. & Roger, A.J. (2023) Is over-parameterization a problem for Profile Mixture Models? *Systematic Biology*, syad063.
https://doi.org/10.1093/sysbio/syad063

Bouchard, P., Bousquet, Y. Davies, A. E. & Cai, C.Y. (2024) On the nomenclatural status of type genera in Coleoptera (Insecta). *ZooKeys*, 1194, 1–981.
https://doi.org/10.3897/zookeys.1194.106440

Boudinot, B.E., Fikáček, M., Lieberman, Z.E., Kusy, D., Bocak, L., Mckenna, D.D. & Beutel, R.G. (2023) Systematic bias and the phylogeny of Coleoptera—A response to Cai *et al*. (2022) following the responses to Cai *et al*. (2020). *Systematic Entomology*, 42, 223–232.
https://doi.org/10.1111/syen.12570

Cai, C.Y. (2024) Ant backbone phylogeny resolved by modelling compositional heterogeneity among sites in genomic data. *Communications Biology*, 7 (1), 106.
https://doi.org/10.1038/s42003-024-05793-7

Cai, C.Y., Tihelka, E., Giacomelli, M., Lawrence, J.F., Ślipiński, A., Kundrata, R., Yamamoto, S., Thayer, M.K., Newton, A.F., Leschen, R.A.B., Gimmel, M.L., Lü, L., Engel, M.S., Bouchard, P., Huang, D., Pisani, D. & Donoghue, P.C.J. (2022) Integrated phylogenomics and fossil data illuminate the evolution of beetles. *Royal Society Open Science*, 9, 211771.
https://doi.org/10.1098/rsos.211771

Cai, C.Y., Tihelka, E., Liu, X. Y. & Engel, M. S. (2023) Improved modelling of compositional heterogeneity reconciles phylogenomic conflicts among lacewings. *Palaeoentomology*, 6 (1), 49–57.
https://doi.org/10.11646/palaeoentomology.6.1.8

Criscuolo, A. & Gribaldo, S. (2010) BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evolutionary Biology*, 10, 210.
https://doi.org/10.1186/1471-2148-10-210

Crotty, S.M. & Holland, B.R. (2022) Comparing partitioned models to mixture models: Do information criteria apply? *Systematic Biology*, 71, 1541–1548.
https://doi.org/10.1093/sysbio/syac003

Fabreti, L.G. & Höhna, S. (2022) Bayesian inference of phylogeny is robust to substitution model over-parameterization. *bioRxiv*.
https://doi.org/10.1101/2022.02.17.480861

Feuda, R., Dohrmann, M., Pett, W., Philippe, H., Rota-Stabelli, O., Lartillot, N., Wörheide, G. & Pisani, D. (2017) Improved modeling of compositional heterogeneity supports sponges as sister to all other animals. *Current Biology*, 27, 3864–3870. e4.
https://doi.org/10.1016/j.cub.2017.11.008

Giacomelli, M., Rossi, M.E., Lozano-Fernandez, J., Feuda, R. & Pisani, D. (2022) Resolving tricky nodes in the tree of life through amino acid recoding. *iScience*, 25, 105594.
https://doi.org/10.1016/j.isci.2022.105594

Huelsenbeck, J.P. & Rannala, B. (2004) Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Systematic Biology*, 53, 904–913.
https://doi.org/10.1080/10635150490522629

Kjer, K.M., Simon, C., Yavorskaya, M. & Beutel, R.G. (2016) Progress, pitfalls and parallel universes: a history of insect phylogenetics. *Journal of The Royal Society Interface*, 13, 20160363.
https://doi.org/10.1098/rsif.2016.0363

Kück, P. & Struck, T.H. (2014) BaCoCa—A heuristic software tool for the parallel assessment of sequence biases in hundreds of gene and taxon partitions. *Molecular Phylogenetics and Evolution*, 70, 94–98.
https://doi.org/10.1016/j.ympev.2013.09.011

Lemmon, A.R. & Moriarty, E.C. (2004) The importance of proper model assumption in Bayesian phylogenetics. *Systematic Biology*, 53, 265–277.
https://doi.org/10.1080/10635150490423520

McKenna, D.D., Shin, S., Ahrens, D., Balke, M., Beza-Beza, C., Clarke, D.J., Donath, A., Escalona, H.E., Friedrich, F., Letsch, H., Liu, S., Maddison, D., Mayer, C., Misof, B., Murin, P.J., Niehuis, O., Peters, R.S., Podsiadlowski, L., Pohl, H., Scully, E.D., Yan, E.V., Zhou, X., Ślipiński, A. & Beutel, R.G. (2019) The evolution and genomic basis of beetle diversity. *Proceedings of the National Academy of Sciences*, 116, 24729–24737.
https://doi.org/10.1073/pnas.1909655116

McKenna, D.D., Wild, A.L., Kanda, K., Bellamy, C.L., Beutel, R.G., Caterino, M.S., Farnum, C.W., Hawks, D.C., Ivie, M.A., Jameson, M.L., Leschen, R.A.B., Marvaldi, A.E., Mchugh, J.V., Newton, A.F., Robertson, J.A., Thayer, M.K., Whiting, M.F., Lawrence, J.F., Ślipiński, A., Maddison, D.R. & Farrell, B.D. (2015) The beetle tree of life reveals that Coleoptera survived end-Permian mass extinction to diversify during the

Cretaceous terrestrial revolution. *Systematic Entomology*, 40, 835–880.
https://doi.org/10.1111/syen.12132

Misof, B., Liu, S., Meusemann, K., Peters, R.S., Donath, A., Mayer, C., Frandsen, P.B., Ware, J., Flouri, T., Beutel, R.G., Niehuis, O., Petersen, M., Izquierdo-Carrasco, F., Wappler, T., Rust, J., Aberer, A.J., Aspöck, U., Aspöck, H., Bartel, D., Blanke, A., Berger, S., Böhm, A., Buckley, T.R., Calcott, B., Chen, J., Friedrich, F., Fukui, M., Fujita, M., Greve, C., Grobe, P., Gu, S., Huang, Y., Jermiin, L.S., Kawahara, A.Y., Krogmann, L., Kubiak, M., Lanfear, R., Letsch, H., Li, Yiyuan, Li, Z., Li, J., Lu, H., Machida, R., Mashimo, Y., Kapli, P., McKenna, D.D., Meng, G., Nakagaki, Y., Navarrete-Heredia, J.L., Ott, M., Ou, Y., Pass, G., Podsiadlowski, L., Pohl, H., Reumont, B.M. von, Schütte, K., Sekiya, K., Shimizu, S., Ślipiński, A., Stamatakis, A., Song, W., Su, X., Szucsich, N.U., Tan, M., Tan, X., Tang, M., Tang, J., Timelthaler, G., Tomizuka, S., Trautwein, M., Tong, X., Uchifune, T., Walzl, M.G., Wiegmann, B.M., Wilbrandt, J., Wipfler, B., Wong, T.K.F., Wu, Q., Wu, G., Xie, Y., Yang, S., Yang, Q., Yeates, D.K., Yoshizawa, K., Zhang, Q., Zhang, R., Zhang, W., Zhang, Y., Zhao, J., Zhou, C., Zhou, L., Ziesmann, T., Zou, S., Li, Y., Xu, X., Zhang, Y., Yang, H., Wang, J., Wang, J., Kjer, K.M. & Zhou, X. (2014) Phylogenomics resolves the timing and pattern of insect evolution. *Science*, 346, 763–767.
https://doi.org/10.1126/science.1257570

Parham, J.F., Donoghue, P.C., Bell, C.J., Calway, T.D., Head, J.J., Holroyd, P.A., Inoue, J.G., Irmis, R.B., Joyce, W.G. & Ksepka, D.T. (2011) Best practices for justifying fossil calibrations. *Systematic Biology*, 61, 346–359.
https://doi.org/10.1093/sysbio/syr107

Schrempf, D. (2019) The ELynx Suite. https://github.com/dschrempf/elynx (Accessed April 8, 2024).

Toussaint, E.F.A., Seidel, M., Arriaga-Varela, E., Hájek, J., Král, D., Sekerka, L., Short, A.E.Z. & Fikáček, M. (2017) The peril of dating beetles. *Systematic Entomology*, 42, 1–10.
https://doi.org/10.1111/syen.12198

Trautwein, M.D., Wiegmann, B.M., Beutel, R., Kjer, K.M. & Yeates, D.K. (2012) Advances in insect phylogeny at the dawn of the postgenomic era. *Annual Review of Entomology*, 57, 449–468.
https://doi.org/10.1146/annurev-ento-120710-100538

Whelan, N. V. & Halanych, K. M. (2017) Who let the CAT out of the bag? Accurately dealing with substitutional heterogeneity in phylogenomic analyses. *Systematic Biology*, 66 (2), 232–255.
https://doi.org/10.1093/sysbio/syw084

Zhang, S.Q., Che, L.H., Li, Y., Liang, D., Pang, H., Ślipiński, A. & Zhang, P. (2018) Evolutionary history of Coleoptera revealed by extensive sampling of genes and species. *Nature Communications*, 9, 205.
https://doi.org/10.1038/s41467-017-02644-4