



# Peer Community In Evolutionary Biology

## Predicting small ancestors using contemporary genomes of large mammals

**Bruce Rannala**  based on peer reviews by **Bruce Rannala**  and 1 anonymous reviewer

Emeric Figuet, Marion Ballenghien, Nicolas Lartillot, Nicolas Galtier (2017) Reconstruction of body mass evolution in the Cetartiodactyla and mammals using phylogenomic data. Missing preprint\_server, ver. Missing article\_version, peer-reviewed and recommended by Peer Community in Evolutionary Biology. <https://doi.org/10.1101/139147>

Submitted: 18 May 2017, Recommended: 05 December 2017

**Cite this recommendation as:**

Rannala, B. (2017) Predicting small ancestors using contemporary genomes of large mammals. *Peer Community in Evolutionary Biology*, 100042. <https://doi.org/10.24072/pci.evolbiol.100042>

Published: 05 December 2017

Copyright: This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

---

Recent methodological developments and increased genome sequencing efforts have introduced the tantalizing possibility of inferring ancestral phenotypes using DNA from contemporary species. One intriguing application of this idea is to exploit the apparent correlation between substitution rates and body size to infer ancestral species' body sizes using the inferred patterns of substitution rate variation among species lineages based on genomes of extant species [1]. The recommended paper by Figuet et al. [2] examines the utility of such approaches by analyzing the Cetartiodactyla, a clade of large mammals that have mostly well resolved phylogenetic relationships and a reasonably good fossil record. This combination of genomic data and fossils allows a direct comparison between body size predictions obtained from the genomic data and empirical evidence from the fossil record. If predictions seem good in groups such as the Cetartiodactyla, where there is independent evidence from the fossil record, this would increase the credibility of predictions made for species with less abundant fossils. Figuet et al. [2] analyze transcriptome data for 41 species and report a significant effect of body mass on overall substitution rate, synonymous vs. non-synonymous rates, and the dynamics of GC-content, thus allowing a prediction of small ancestral body size in this group despite the fact that the extant species that were analyzed are nearly all large. A comparative method based solely on morphology and phylogenetic relationships would be very unlikely to make such a prediction. There are many sources of uncertainty in the variables and parameters associated with these types of approaches: phylogenetic uncertainty (topology and branch lengths), uncertainty about inferred substitution rates, and so on. Although the authors do not account for all these sources of uncertainty the fact that their predicted body sizes appear sensible is encouraging and undoubtedly the methods will become more statistically sophisticated over time.

## References:

- [1] Romiguier J, Ranwez V, Douzery EJP and Galtier N. 2013. Genomic evidence for large, long-lived ancestors to placental mammals. *Molecular Biology and Evolution* 30: 5–13. doi: [10.1093/molbev/mss211](<https://doi.org/10.1093/molbev/mss211>)
- [2] Figuet E, Ballenghien M, Lartillot N and Galtier N. 2017. Reconstruction of body mass evolution in the Cetartiodactyla and mammals using phylogenomic data. *bioRxiv*, ver. 3 of 4th December 2017. 139147. doi: [10.1101/139147](<https://doi.org/10.1101/139147>)

## Reviews

### Evaluation round #1

DOI or URL of the preprint: [10.1101/139147](https://doi.org/10.1101/139147)

Version of the preprint: 1

### Authors' reply, 08 November 2017

[Download author's reply](#)

### Decision by [Bruce Rannala](#) , posted 30 September 2017

#### Revise

Here are my comments to the authors:

This is an important paper that just needs a few minor changes/clarifications. The authors should revise according to the recommendations of the two reviewers (myself and an anonymous reviewer). In particular, the anonymous reviewer and I both had some concerns about the uncertainty of the phylogeny. I would like to see a bit more analysis to determine whether incomplete lineage sorting may be a source of phylogenetic ambiguity for these data and I would like to see the raxML tree with branch lengths included in the paper (as suggested by the anonymous reviewer). Please also respond directly to the following comments by myself and the anonymous reviewer:

Page 12, Correlations of substitution rates/ratios and LHTs: I am not familiar with the COEVOL program but if it is producing the posterior distribution of the correlation coefficient why not provide the posterior mean and credible set rather than a p-value? (which is not a very Bayesian thing to do).

Tables 1 and 2: I think the legends must be reversed.

Figure 2: Are the points on this graph mean posterior dN/dS versus  $\log_{10}(\text{BM})$ ? This should be stated in the legend.

I88: how strong are the reported correlations? I101: same question.

I151: I am not sure what is meaning of this sentence. Does this refer to a better reconstruction of ancestral LHT?

I169: Are you referring to phylogenetic inertia for leaf nodes? Meaning that there is no need to actually compute a correlation in actual species before proceeding to the inference.

I197: few words on the "home made scripts" would be welcome. How do they filter out mis-aligned regions?

I209: any justification for the  $\log_{10}$  transformation?

I261: any insight on why this index rather than any other?

I289: Kr/Kc ratio... As it is not so standard, can you define it?

Table 1: what about reporting the median/mean/mode? Plots of the posterior densities would also be very informative regarding the strength and robustness of the estimations.

## Reviewed by Bruce Rannala , 20 July 2017

This paper evaluates the statistical behavior of new methods for analyzing associations between life-history traits (LHTs) and rates of molecular evolution ( $dS$  and  $dN/dS$ ). The basic idea is to study a group (Cetartiodactyla) with a fairly well resolved phylogeny and multiple fossil calibrations to evaluate whether the results seem sensible in this case. If so, that would provide some evidence that the results obtained in groups with poor fossil records might also be reasonable. The paper is well-written and the introduction does a very nice job of summarizing the LHT methods and the motivation for the study. The results (positive correlations between body mass, age at maturity and  $dN/dS$ ) fit the predictions of the reduced  $N_e$  theory as does the negative correlation with GC3. It seems the method is producing reasonable results. I have a few concerns, some minor, some less so:

Page 9, phylogeny reconstruction: if  $dN/dS$  systematically varies across the group and the cause is a decreased  $N_e$  in larger species this might create more uncertainty of relationships among small species than among large species – I wonder whether this could be a source of bias? Have the authors considered trying a species tree inference method that accounts for incomplete lineage sorting (which would have more effect with larger  $N_e$ ) to see whether the results are consistent with the tree from concatenated sequences? Later in the paper it is noted that some alternative topologies produce similar results for correlations between rates and LHTs but I am still curious.

Page 12, Correlations of substitution rates/ratios and LHTs: I am not familiar with the COEVOL program but if it is producing the posterior distribution of the correlation coefficient why not provide the posterior mean and credible set rather than a p-value? (which is not a very Bayesian thing to do).

Tables 1 and 2: I think the legends must be reversed.

Figure 2: Are the points on this graph mean posterior  $dN/dS$  versus  $\log_{10}(BM)$ ? This should be stated in the legend.

## Reviewed by anonymous reviewer 1, 20 September 2017

The ms by Figuet et al. is a case study on the inference of ancestral Life History Trait (LHT) using molecular markers, specifically  $dS$ ,  $dN/dS$  and GC3. I found the ms scientifically sound, easy to follow and quite appealing. I have only few minor points that could potentially help broadening the readership.

Since the authors aim at convincing paleontologists (1124), a special effort to make all the analyzes crystal clear for non-specialists could be a good choice. As it is now, I am not sure paleontologists will be able to follow.

My main scientific question concerns the lack of coherence between the approaches:  $dN/dS$  suggests small body-sizes (part 1 –substitution mapping–) but at the same time not (part 2 –coevol–). Conversely,  $dS$  is the main driver in the coevol approach but the length of internal branches from the raxML tree are not shown. Finally, why the GC3 signal has not been included in the coevol approach to check its consistency to the part 3. Although the three parts all point to the same direction, it would be nice to dedicate some discussion on why the metrics ( $dS$ ,  $dN/dS$  and GC3) differ in their predictions when using different approaches. The lack of coherent signal using  $dN/dS$  in the coevol framework is especially puzzling.

It would not hurt to emphasize that part 1 is done on a classical molecular phylogenetic tree (i.e. not ultra-metric) whilst the second is performed on a calibrated ultra-metric tree. I am not sure about the third part. Calibration has its own issues that could be discussed in line with my previous comment.

Can the authors show the raxML phylogeny? As it is used for the first part of the analysis, it would be nice to have a look at it.

I also have a list of minor points/interrogations that will be easily addressed. They often all are of the same nature: the text is sometimes not self-sufficient; thus providing extra-information on methods or choices may not hurt. Although interested specialists will likely know or read the cited literature, casual readers would benefit from extra pieces of information within this ms.

l88: how strong are the reported correlations? l101: same question

I151: I am not sure what is meaning of this sentence. Does this refer to a better reconstruction of ancestral LHT ?

I169: Are you referring to phylogenetic inertia for leaf nodes ? Meaning that there is no need to actually compute a correlation in actual species before proceeding to the inference.

I197: few words on the "home made scripts" would be welcome. How do they filter out mis-aligned regions ?

I209: any justification for the log<sub>10</sub> transformation ?

I261: any insight on why this index rather than any other ?

I289: Kr/Kc ratio... As it is not so standard, can you define it ?

Table 1: what about reporting the median/mean/mode ? Plots of the posterior densities would also be very informative regarding the strength and robustness of the estimations.