Submitted by Emeric Figuet 2017-05-18 15:28

Emeric Figuet, Marion Ballenghien, Nicolas Lartillot, Nicolas Galtier

Reconstruction of body mass evolution in the Cetartiodactyla and mammals using phylogenomic data.

https://doi.org/10.1101/139147

Keywords: ancestral characters, Bayesian inference, mammals, phylogeny, substitution rate, GC content

Abstract

Reconstructing ancestral characters on a phylogeny is an arduous task because the observed states at the tips of the tree correspond to a single realization of the underlying evolutionary process. Recently, it was proposed that ancestral traits can be indirectly estimated with the help of molecular data, based on the fact that life history traits influence substitution rates. Here we challenge these new approaches in the Cetartiodactyla, a clade of large mammals which, according to paleontology, derive from small ancestors. Analysing transcriptome data in 41 species, of which 22 were newly sequenced, we provide a dated phylogeny of the Cetartiodactyla and report a significant effect of body mass on the overall substitution rate, the synonymous vs. non-synonymous substitution rate and the dynamics of GC-content. Our molecular comparative analysis points toward relatively small Cetartiodactyla ancestors, in agreement with the fossil record, even though our data set almost exclusively consists of large species. This analysis demonstrates the potential of phylogenomic methods for ancestral trait reconstruction and gives credit to recent suggestions that the ancestor to placental mammals was a relatively large and long-lived animal.

This preprint worth a revision

by Bruce Rannala, 2017-09-30 07:35

Manuscript: https://doi.org/10.1101/139147

Decision & reviews

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 Ne 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.

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 :
Thanks for these comments. We added a section explicitly discussing uncertainty in phylogeny and the incomplete lineage sorting (ILS) problem, plus new analyses related to this issue. We also show the RaxML tree with branch lengths in the revised version, as well as additional results and information as requested by the reviewers.

* Page 9, phylogeny reconstruction: if dN/dS systematically varies across the group and the cause is a decreased Ne in larger species this might create more uncertainity of relationships among small species than among large species -- I wonder whether this could be a source of bias?

We agree that ILS is a potential problem in such analyses, as recently highlighted in a number of papers, e.g., from Matthew Hahn's group. For various reasons we do not believe that ILS is a major issue in this specific dataset. In presence of ILS, a fraction of substitutions can be incorrectly mapped, when the true gene tree differs from the reference species tree. We here argue that this fraction is probably very small.

ILS can only happen when two successive speciation events are close enough in time that ancestral polymorphism is not yet sorted when the second split occurs. Quantitatively, for ILS to be effective, the time between the two splits, Ts, must be of the order of magnitude of (or less than) the average coalescence time in the ancestral population, Tc, which in a panmictic population is twice the effective population size. If Ts>>Tc then most polymorphisms will be sorted before the second split happens. Multiplying the two numbers by the mutation rate, u, we reach the condition that ILS can only be effective if the length of the branch separating the two splits, Ts.u, is of the same order of magnitude as within-species polymorphism, Tc.u.

Published estimates of non-coding genetic polymorphism in Cetartiodactyla range between 0.025% and ~0.27%, implying coding sequence polymorphism of the order of 0.01%-0.1%. So only internal branches shorter than 0.001 have a chance to be affected by ILS in this analysis. Only two internal branches in our reference RaxML tree (Appendix B) have a length lower than 0.001 – Kobus-Pantholops ancestor and Boselaphus-Tragelaphus ancestor. These two branches represent a very small fraction of total tree length, hence our conclusion that incorrect mapping due to ILS is probably negligible in this analysis. This very argument was made by Scornavacca and Galtier (2017), who suggested that ILS is only a minor determinant of gene tree/species tree conflicts in mammals. This is also consistent with the relatively weak effect of ILS on branch length estimation reported by Mendes and Hahn (2016) from simulations.

We added a paragraph in the Discussion section discussing this issue (lines 394 to 405), citing Hahn and Nakhleh (2016), Mendes and Hahn (2016), Scornavacca and Galtier (2017) and references regarding within-species polymorphism in Cetartiodactyls. We also point out that our analysis is mainly based on the ratio on nonsynonymous to synonymous rates, and we see no strong reason why ILS should impact one of the two rates more than the other.

Have the authors considered trying a species tree inference method that accounts for incomplete lineage sorting (which would have more effect with larger Ne) 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.
Following this comment we used the ILS-aware ASTRAL method and recovered a tree that differed from our reference RaxML tree by just one internal branch, namely, the ancestral branch to Kobus (see response to previous comment above, Figure S2 and Appendix B). We performed substitution mapping and downstream analyses with this new phylogeny, and obtained results very similar to our main analysis. The results of this new analysis are shown in Table S5. We added a sentence in Materials & Methods (line 222) and modified the Results section (line 308) to account for the test of this alternate topology.

and I would like to see the **raxML tree with branch lengths included in the paper** (as suggested by the anonymous reviewer).
We added the phylogeny produced by raxML with branch lengths and bootstrap values in supplementary material Figure 2 and Appendix B (newick trees).

* 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).
Please note that we do not report p-values for the Bayesian analyses (only for the classical regression analyses, which are indeed frequentist). For the analyses conducted with coevol, what we report instead is the posterior probability that the correlation is positive. As shown in Lartillot and Poujol, 2011, rejecting the no-correlation when this pp<0.025 or >0.975 leads to a good control of the false positive rate.

* Tables 1 and 2: I think the legends must be reversed.
Thank you for pointing this out. Legends were reassigned to their due table.

* Figure 2: Are the points on this graph mean posterior dN/dS versus log_10(BM)? This should be stated in the legend.
We expanded the legend of Figure 2 to make it explicit that dN/dS values were obtained by substitution mapping and therefore do not correspond to posterior dN/dS from *coevol*.

Reviewed by anonymous reviewer, 2017-09-20 23:51

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 (l124), 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.

Thanks for your comments on our manuscript. We tried to be more exhaustive of a few aspects of our analyses in order to appeal more easily to non-specialists (see responses below). We also emphasise in the main text that details about the different methodologies are described extensively in supplementary material Appendix A (lines 194, 226, 255, 265).

* 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.

We modified the discussion to explicitly mention this discrepancy (lines 408 to 410) – which we honestly do not fully understand at the moment – and suggest perspectives for future attempts to clarify its origins (line 433: *'Another promising approach involves analysing the variance, not just the mean, of substitution rates across genes (Wu et al. 2017).'*). This is a large amount of work, owing to the complexity of the *coevol* method, and the fact that each run takes several weeks. We do think that empirical analyses of large data sets such as this one provide an interesting source of information for improving methods, as suggested, but we see these methodological developments as a distinct project.

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.
This is a sound remark that the use of an ultra-metric tree in the *coevol* analysis might constitute a source of discrepancy between the results of the two kinds of analyses. We modified the manuscript to empathise the use of a classical molecular phylogenetic tree in the first and third part (about GC content), and a calibrated ultra-metric tree with *coevol* (lines 225, 262). We also mentioned it as a possible source of discrepancy between the two analyses in the Discussion section (lines 410 to 412: *'Please note that a time tree is used in the Bayesian analysis, whereas dN/dS were estimated using a non-ultrametric tree with substitution mapping. This could explain in part the discrepancies between the two analyses.'*).

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.
raxML phylogeny was provided both in Figure S2 and Appendix B (newick tree).

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.
The strength of the correlations between dN/dS and LHTs are quite dependant on the molecular and taxonomic dataset used as well as the life-history trait ; and can hardly be directly compared according to the used methodology. It is true however, that in order to perform ancestral reconstruction, not only are such correlations necessary, but they should also be strong enough to allow for satisfactory confidence intervals. For example, at the placental mammal scale, a strong correlation was found between dN/dS and longevity (r=0.86) using a large genomic dataset (Romiguier et al. 2013, MBE). As for GC-content, although its relation with LHTs cannot be directly exploited for ancestral reconstruction (due to phylogenetic inertia), a strong correlation was also retrieved with longevity when using the 'time-corrected index of GC3 conservation' described in Romiguier et al. 2013, which we reused in the manuscript.

* l151: I am not sure what is meaning of this sentence. Does this refer to a better reconstruction of ancestral LHT ?

It does refer to a better reconstruction of ancestral LHT, we slightly modified the sentence to make it clearer (line 153: '*An analysis of complete mitochondrial genomes in 201 species of cetartiodactyls suggested that DNA-aided reconstruction* ***can estimate ancestral LHTs more accurately*** *than classical method*').

* l169: 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.

In *coevol*, the correlation between traits and rates is modelled jointly with the evolutionary process, as now clarified in the previous sentence (line 166: '*the* ***correlated*** *evolution of life-history traits and molecular parameters across the phylogenetic tree is modelled as a multivariate Brownian motion [...]*').

* l197: few words on the "home made scripts" would be welcome. How do they filter out mis-aligned regions ?

We added a sentence in the Material and Methods section that briefly describes how we cleaned alignments for obviously misaligned regions (line 197: '*Briefly, this was done by scanning each sequence for small regions flanked by undetermined base pairs, and removing them if they shared less than 70% similarity with any other sequence.*').

* l209: any justification for the log_10 transformation ?

The reason life-history traits are classically log-transformed in such analyses is that their observed relation with census population size (and by extension effective population size) is non linear (Damuth 1981, White et al., 2007).

* l261: any insight on why this index rather than any other ?

The rationale for this index is described in the Romiguier et al., 2013 paper ; it seeks to predict the loss of correlation between gene GC-content in a pair of species as a function of time and life-history trait in the species, with the idea that small-sized, large Ne species will experience stronger gBGC and therefore diverge faster in terms of GC patterns. This was designed specifically with the purpose of ancestral reconstruction and is therefore reused as is in this manuscript, where we aim at assessing the reliability of existing approaches for ancestral reconstruction using molecular data.

* l289: Kr/Kc ratio... As it is not so standard, can you define it ?
Definition for the Kr/Kc ratio was given in Appendix A, but was indeed lacking in the main text. We recalled its definition at the first mention of the ratio in the Results section (line 294).

* 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.
Unfortunately, the way the analyses were conducted does not allow us to extract the necessary information for drawing such distributions and reporting point estimates, unless we start these chains all over again which is computationally heavy. Please note that the posterior distribution of all parameters have been taken into account in the estimation of the correlations between molecular parameters and life-history traits and the estimation of credibility intervals.