Evolutionary Biology (2020) 47:220–232
https://doi.org/10.1007/s11692-020-09505-x

RESEARCH ARTICLE

Ancestral State Estimation with Phylogenetic Ridge Regression

Silvia Castiglione[1] · Carmela Serio[1] · Alessandro Mondanaro[1,2] · Marina Melchionna[1] · Francesco Carotenuto[1] · Mirko Di Febbraro[3] · Antonio Profico[4] · Davide Tamagnini[4] · Pasquale Raia[1]

Received: 8 October 2019 / Accepted: 9 June 2020 / Published online: 17 June 2020

Abstract

The inclusion of fossil phenotypes as ancestral character values at nodes in phylogenetic trees is known to increase both the power and reliability of phylogenetic comparative methods (PCMs) applications. We implemented the R function *RRphylo* as to integrate fossil phenotypic information as ancestral character values. We tested the new implementation, named *RRphylo-noder* (which is available as part of the RRphylo R package) on tree and data generated according to evolutionary processes of differing complexity and under variable sampling conditions. We compared *RRphylo-noder* performance to other available methods for ancestral state estimation, including Bayesian approaches and methods allowing rate variation between the tree branches. We additionally applied *RRphylo-noder* to two real cases studies, the evolution of body size in baleen whales and in caniform carnivores. Variable-rate methods proved to be more accurate than single-rate methods in estimating ancestral states when the pattern of phenotypic evolution changes across the tree. *RRphylo-noder* proved to be slightly more accurate and sensibly faster than Bayesian approaches, and the least sensitive to the kind of phenotypic pattern simulated. The use of fossil phenotypes as ancestral character values noticeably increases the probability to find a phenotypic trend through time when it applies to either the entire tree or just to specific clades within it. We found Cope's rule to apply to both mysticete cetaceans and caniform carnivores. The *RRphylo-noder* implementation is particularly appropriate to study phenotypic evolution in the presence of complex phenotypes generated by different processes acting in different parts the tree, and when suitable information about fossil phenotypes is at hand.

Keywords Ancestral states estimation · Fossil phylogenies · Phenotypic evolution · Phylogenetic comparative methods · Rrphylo

Introduction

Phylogenetic comparative methods (PCMs) are sought to account for species non-independence in analyses of phenotypic evolution, providing robust inference about the process and patterns of evolution (Alfaro et al. 2009; Harmon et al. 2003; Venditti et al. 2011) and the correlation between traits (Felsenstein 1985; Garland and Ives 2000; Revell 2010). The simplest PCMs refer to a constant-rate, non-directional process, namely the Brownian motion model of evolution (BM), and deviations thereof (Freckleton et al. 2002; Harvey and Pagel 1991; O'Meara 2012; Pagel 1997). Several PCMs now allow fitting more complicated evolutionary models, taking stabilizing selection and stasis into account and relaxing the assumption that a single evolutionary rate applies unambiguously across the tree (Castiglione et al. 2018; Elliot and Mooers 2014; O'Meara et al. 2006; Rabosky 2014; Smaers et al. 2016). More importantly, most of these recent PCMs

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11692-020-09505-x) contains supplementary material, which is available to authorized users.

✉ Pasquale Raia
pasquale.raia@unina.it

[1] Dipartimento Di Scienze Della Terra, Dell'Ambiente E Delle Risorse, Monte Sant'Angelo, Via Cinthia, 21 - 80126 Napoli, Italy

[2] Dipartimento Di Scienze Della Terra, Via G. La Pira, 4, 50121 Firenze, Italy

[3] Dipartimento Di Bioscienze E Territorio, University of Molise, C. da Fonte Lappone, 86090 Pesche, IS, Italy

[4] Dipartimento Di Biologia Ambientale, Sapienza Università Di Roma, Rome, Italy

allow using paleontological trees (Bapst 2013; Pennell and Harmon 2013), which is welcome since there is widespread acknowledgement that the inclusion of fossil information allow better inference about the patterns of taxonomic and phenotypic diversification (Didier et al. 2017; Finarelli and Liow 2016; Heath et al. 2014; Liow et al. 2010; Mitchell et al. 2018; Puttick et al. 2017; Schnitzler et al. 2017; Silvestro et al. 2016; Slater and Harmon 2013).

PCMs allow to estimate ancestral character states (the phenotypic values at the tree nodes) by means of maximum likelihood, treating the phenotypes at the nodes as parameters, in order to find the parameter values that maximize the probability of the tip (observed) data assuming a given evolutionary model (most commonly the BM or Ornstein–Uhlenbeck, OU, Joy et al. 2016). This procedure could be troublesome, since the accuracy of ancestral states estimation is biased if the model does not approximate reality (Chira and Thomas 2016; Cooper et al. 2016; Slater, Harmon, and Alfaro 2012). It has been shown that PCMs may actually provide ancestral estimates that compare very unfavorably to real fossil phenotypes (Webster and Purvis 2002), especially at the root (Gascuel and Steel 2014). To cope with this drawback, several approaches have been developed in order to use fossil species traits as given phenotypic values at the tree nodes to guide the estimation process (Slater and Harmon 2013). Under a Bayesian approach, this is implemented setting fossil phenotypes as node priors (Slater, Harmon, and Alfaro 2012), so that both the priors and the tips are used as observations in the parameter estimation process (Slater, Harmon, and Alfaro 2012).

Models that are not constrained to fit a single evolutionary rate across the tree are better suited to cope with phenotypic vectors whose complexity (herein defined as the result of different phenotypic patterns applying to different parts of the tree) is not captured by simpler evolutionary models (e.g. BM, OU, Early-Burst; Chira and Thomas 2016; Pennell et al. 2015). The downside of such "variable-rates" models is that they could be severely overparametrized, which might reduce their fit to the real data once topological and sampling issues are considered (Castiglione et al. 2018; Chira and Thomas 2016; Eastman et al. 2011; Elliot and Mooers 2014; Rabosky 2014; Venditti et al. 2011).

Here we present an implementation of the R function *RRphylo* (Castiglione et al. 2018) meant to estimate ancestral states taking advantage of explicit fossil information. In contrast to most other methods, under this new implementation, named *RRphylo-noder*, there is no expected distribution of trait changes during evolution (e.g. the normal distribution under BM), so that we expect *RRphylo-noder* could outperform competing methods in the presence of complex phenotypes. We tested this hypothesis by means of extensive simulations and real-case applications. We show that *RRphylo-noder* is faster and increasingly more accurate than competing methods in reconstructing ancestral states as phenotypic complexity increases.

Materials and Methods

The *RRphylo-noder* implementation is based on phylogenetic ridge regression, *RRphylo* (Castiglione et al. 2018). In *RRphylo* the phenotypic change between any two nodes (or a node and a tip) aligned along a phyletic line is described by the sum of individual contributions at each consecutive branch between the nodes, according to the equation $\Delta y = \vec{\beta}_{1} l_1 + \vec{\beta}_{2} l_2 + \ldots . \vec{\beta}_{n} l_n$. Here, n equals the number of branches intervening between the nodes, $\vec{\beta}_{1\ldots n}$ is the vector of phylogenetic ridge regression coefficients (the evolutionary rates), and $l_{1\ldots n}$ are the branch lengths. The vector of regression coefficients $\vec{\beta}$ is computed simultaneously for all the branches in the tree by applying a normalization factor λ which avoids fitting extreme β values and prevents multicollinearity (James et al. 2013).

RRphylo-noder integrates the phenotypic information at internal nodes in the estimation of evolutionary rates and ancestral character states. Given a vector ***n*** of phenotypic values known in advance to be placed at internal nodes (***fossil.states***), a vector of false tips ***ftips*** of length ***n*** is added to the tree. Each i_{th} element of ***ftips*** is phenotypically identical to the corresponding $\boldsymbol{fossil.states}_{\mathbf{i}}$ and is attached to the tree at the position of $\boldsymbol{fossil.states}_{\mathbf{i}}$ with a branch of length = 0. Then, the vector of regression coefficients $(\vec{\beta})$ is estimated by means of *RRphylo* by using the modified tree and phenotype (which include both ***ftips*** and the real tips). Since the branch lengths of ***ftips*** are equal to zero, the phenotypic rate between each $\boldsymbol{ftips}_{\mathbf{i}}$ and the corresponding node is zero, which means the ***fossil.states*** and their corresponding ***ftips*** will have the same phenotypic estimates. After β coefficients are estimated, the vector of phenotypic values at nodes $\vec{a}$ is calculated as usual as:

$$\vec{a} = L'\vec{\beta}$$

where each row of *L'* represents the path of branch lengths moving from a specific node in the tree. The final step of the algorithm consists in removing ***ftips*** from the tree, and from the rate and phenotypic vectors.

Simulations

We tested *RRphylo-noder* accuracy and compared it to other available methods for ancestral states estimation. The goals of the simulations were to assess (1) the accuracy of both single-rate and variable-rates methods to ancestral state estimation with complex phenotypes, (2) the effect of sampling on ancestral state estimation and (3) the impact of

using fossil phenotypes as ancestral character states known in advance. We used ape's *ace* (Paradis and Schliep 2018) to represent the simplest, most straightforward method for ancestral character state estimation under BM. The function *fastAnc* in phytools (Revell 2012) is based on BM as *ace*, but additionally allows specifying phenotypic states at nodes. Elliot and Mooers' *StableTraits* (Elliot and Mooers 2014) estimates ancestral states using a generalization of BM to the stable random walk, represented by a symmetrical, zero-centered distribution of phenotypic increments during the evolutionary time defined by the parameters α (the index of stability) and c (the scale). *StableTraits* performs the Bayesian estimation of ancestral states fitting the α and c parameters and allows the comparison with BM. As in *RRphylo-noder*, node priors can be used by grafting zero-branch length false tips to specified nodes. We implemented both *StableTraits* and *StableTraits-Brownian* (that is *StableTraits* referring BM to estimate ancestral states) and wrote a wrapper around *StableTraits* (named *StableTraitsR*) which allows using the function within the R environment and with different operating systems and to specify node priors. We tested both *RRphylo* and *RRphylo-noder* along with all of these other methods. In sum, we used two methods which do not allow to specify phenotypic values known in advance at nodes (*ace* and *RRphylo*), two single-rate methods (*ace* and *fastAnc*), and four methods which allow the evolutionary rate to change across the tree (which we collectively refer to as 'variable-rate' models: *StableTraits-Brownian, StableTraits, RRphylo* and *RRphylo-noder*). We tested such methods on forty different kinds of phenotypes to assess their performance under phenotypes differing in terms of complexity and under different sampling regimes.

Simulating Trees and Starting (basic) Phenotypes

To produce the phenotypes, we started by creating 100 random phylogenetic trees having 80 species at least (average = 164 species, range = 127–238 species) by using the function *sim.bdtree* in the R package geiger (Harmon et al. 2007) setting the birth rate at 0.5 and the death rate at 0.2. For each tree, we first generated a phenotypic vector ***y*** under BM, recording both tip (species values) and node (ancestral values) phenotypes, by using the function *fastBM* in the R package phytools (Revell 2012). We set the phenotypic mean (i.e. the value at the tree root) at 0 and σ^2 (the Brownian rate) at 1. A second phenotype ***yt*** was produced on the same tree by modelling evolution according to Brownian motion with trend (i.e. a trend in the phenotypic mean over time, hereafter referred as 'phenotypic drift') using *fastBM* and specifying the *mu* parameter (the intensity of the drift) at 0.5. For ***yt*** as well, we recorded both the species and ancestral values. The starting phenotypes ***y*** and ***yt*** were used as they are and then manipulated to derive more complex phenotypic vectors as explained below.

Simulating Complex Phenotypic Evolution with and Without Ancestral States known in Advance

To test the effect of providing known phenotypes at some nodes, we randomly selected from the starting phenotype a number of nodes ***N*** equal to 5% of the nodes in the tree and their ancestral values. The ***N*** ancestral values were used as phenotypic values at nodes known in advance (***fossil.states*** under the *RRphylo-noder* terminology or node priors under the Bayesian approach terminology). To use ancestral values which were not evolved according to BM (or BM with trend either), we produced a further phenotype shuffling the ***N*** ancestral values across the tree nodes. The advantage of shuffling ancestral values is that under BM they are phylogenetically weighted means of the tip phenotypes descending from them, meaning they are constrained within the range of values of the descendant phenotypes, whereas after the shuffling the ancestral values could be outside the phenotypic range of tips descending from them (this is true, for instance, of any clade following Cope's rule). A further manipulation consists in randomly selecting half of the tree nodes to apply a phenotypic trend to the clades descending from them. This way, the tree phenotypic vector is generated by a complex process, whereby a number of clades evolve under phenotypic drift while others do not (by using ***y*** as the starting phenotype), or the intensity of drift for individual clades differs from, and could even reverse, the drift imposed to the entire tree (by using ***yt*** as the starting phenotype). Such phenotypic drifts specifically applied to selected clades was imposed by multiplying, for each selected clade, the distances of each tip to their corresponding common ancestor by a constant, according to the equation $y' = y + time * 0.5$, where y' is the 'drifted' phenotype and *time* is the vector of tip to common ancestor distances. Such complex phenotypes are often reported in literature (Laurin 2004; Hone et al. 2005; Monroe and Bokma 2010; Raia et al. 2012; Benson et al. 2014a; Baker et al. 2015), though their existence is impossible to recognize with abstract evolutionary models (e.g. BM, early-burst) as they applying to the tree as a whole (but see Slater and Pennell 2013).

The whole set of manipulations derives from either ***y*** (BM) or ***yt*** (BM with trend), a second phenotype with shuffled ancestral values, and a number of complex phenotypes where patterns of phenotypic drift change idiosyncratically across the tree, plus any combination of them (Table 1). All these phenotypic vectors were eventually used to test the effect of sampling on ancestral state estimation as described below.

Table 1 Simulated kinds of phenotypes. The columns indicate how the phenotypes were made more complex departing from the initial (either Brownian motion BM, or BM with trend, DR) applying a number of manipulations and their combinations

Starting phenotype	Shuffling	Drift at individual clades	Random sampling		Biased sampling	
BM	shu	ds5	s1—random	s5—random	s1—bias	s5—bias
		–	s1—random	s5—random	s1—bias	s5—bias
	–	ds5	s1—random	s5—random	s1—bias	s5—bias
		–	s1—random	s5—random	s1—bias	s5—bias
DR	shu	ds5	s1—random	s5—random	s1—bias	s5—bias
		–	s1—random	s5—random	s1—bias	s5—bias
	–	ds5	s1—random	s5—random	s1—bias	s5—bias
		–	s1—random	s5—random	s1—bias	s5—bias

shu the phenotypic values of ancestral values (simulated according to BM or DR either) are shuffled before the analyses, *ds5* half of the ancestral values are selected randomly to be applied a phenotypic drift through time, *s1* 10% of species are removed either at 'random' or depending on the value of the phenotype ('bias') *s5* the same as *s1* but removing 50% of species. Starting from the left to the right, the initial phenotype is manipulated making it more complex. For instance, starting from the Brownian motion simulation BM, the phenotype is changed shuffling known ancestral states among selected nodes (*shu*), then applying a phenotypic drift to one half of the clades subtended by such selected nodes (*ds5*) and then removing either at random or with biased sampling up to one half of the species from the tree (e.g. *s5* biased)

Simulating the Effect of Sampling

After different kinds of phenotypes have been produced according to the manipulations described so far, a number of tips were removed from the tree as to test the effect of incomplete sampling. Two different kinds of subsampling strategies were applied, either 'biased or 'random'. Under the former, the chance of a species to be removed from the tree is inversely proportional to its phenotypic value. This corresponds to real case situations such as, for instance, to the higher chance to fossilize (hence to be sampled in the tree) for large versus small organisms (Behrensmeyer et al. 2000; Meloro et al. 2007). Two sampling schemes were applied. The first consists in removing 10% of the species from the tree, the second consists in removing 50% of the species, under both the 'random' and the 'biased' designs.

Overall, these procedures originate 20 different kinds of phenotypes deriving from the starting simulation (i.e. $\boldsymbol{y}$ or $\boldsymbol{yt}$, either). Each phenotypic combination was repeated 10 times originating a total of $20 \times 2 \times 10 = 400$ simulations. The schematic description of the simulation procedures is presented in Table 1.

Testing Ancestral State Reconstruction Methods Accuracy

The ultimate goal of any ancestral state estimation method is to predict phenotypes at nodes. Under perfect prediction, regressing the simulated phenotypes at nodes against their predictions originates a regression slope = 1 and intercept = 0. For each method, we calculated the slope and the intercept of the regression between simulated and fitted values, and the root mean squared error (*rmse*) of the regression. Under *fastAnc*, ancestral values are not fitted but given as phenotypic values at nodes known in advance. This means the more ancestral values are provided the lower *fastAnc*'s *rmse* will be. Thus, rather than a measure of the goodness of fit of *fastAnc*, its corresponding *rmse* depends on how many ancestral values are provided. For this reason, beyond *rmse* we calculated the *rmse* over the fitted nodes only (*reduced rmse*) in order to compare *fastAnc* to the other methods. Since *ace* and *RRphylo* make no use of ancestral values, we used only *rmse* to compare the ancestral state estimates of these methods. Mean *rmse* between methods was compared by means of repeated-measures ANOVA using the function *lme* in the R package *nlme* (Pinheiro et al. 2014) taking the kind of phenotype tested as the random effect. The performances of the methods were compared by using the function *glht* in the package multcomp (Hothorn et al. 2016).

The effects of changing the pattern of phenotypic drift across the tree, shuffling ancestral values and applying different sampling schemes, combine into phenotypes of increasing complexity, which deviate more and more from the starting $\boldsymbol{y}$ and $\boldsymbol{yt}$ vectors as different factors are added. To assess how such phenotypic complexity affects methods' performance, for each method we averaged the 10 *rmse* (and *reduced rmse* as well) per type of phenotype, and collated *rmse*s for the different phenotypes from the smallest to the largest *rmse* as predicted by ape's function *ace*. In this way, the forty *rmse* estimates were effectively ordered from the most similar, to the most dissimilar from BM, which is the evolutionary model *ace* is based upon. Then, we used categorical regression to calculate the slope of methods' *rmse* (the response variable) against the phenotypic kinds collated from the most similar to Brownian Motion (BM) to the most dissimilar (DR.shu.ds5.s5.bias: a kind of phenotype first produced according to Brownian Motion with trend, and then modified applying shuffling to 5% of the ancestral

states, imposing a phenotypic drift to 50% of the individual clades within the tree, and eventually pruned of 50% of the tips via biased sampling, see Table 1) used as the predictor variable. This way we estimated how *rmse* grows away from BM, for each method. We compared the regression slopes and estimated marginal means per method by using the functions *emtrends* and *emmeans*, respectively, in the package emmeans (Lenth et al. 2018). Ideally, the shallower the slope of the regression the less sensitive to the phenotype type a method is. Similarly, lower estimated marginal means indicate better prediction accuracy across phenotypes. This same procedure was repeated on *reduced rmse*, using *StableTraits-Brownian* predictions to collate phenotypes from the simplest (i.e. most similar to BM) to the most complex.

Eventually, we used geiger's function *fitContinuous* to compare the penalized AIC (AICc) obtained by fitting the BM and BM with trend (which is named 'drift' in geiger) to the tree and data on the original phenotypes (***y*** and ***yt***) and after the manipulations. Given the manipulation procedures invert the sign of the phenotypic drift for some clades in the tree (when the original phenotype was BM with trend), or add a drift where there was none (when the original phenotype was BM) we expect *fitContinuous* should fail to recognize the original phenotypic trend (either no such trend under BM or a drift under BM with trend) because of the manipulations. We compared *fitContinuous* results to the corresponding figures obtained by applying *search.trend* to the whole tree.

Testing the Importance of Sampling and Phylogenetic Uncertainty

A perfect prediction of ancestral states values (i.e. corresponding to the slope = 1 and intercept = 0), could indicate that a method is particularly accurate but could also depend on a method being overfit. Overfitting is the major drawback for overparametrized methods such as *RRphylo* and *RRphylo-noder*. An overfit method may appear superior to other methods when assessed for prediction accuracy based on simulated data but could fail to capture the fundamental processes that led to the observed patterns in real data, which represent a subset of the real diversity of the clades, providing much reduced prediction accuracy.

To evaluate the potential for overfit, we applied ANOVA and post-hoc TukeyHSD test to assess whether the slope and intercept of the regression between observed and estimated ancestral states differ among sampling schemes. We similarly assessed whether the phenotypic deviations between known (simulated) and fitted ancestral states estimates change per sampling scheme. Phenotypic deviation was calculated as average percent deviation of the fitted versus simulated ancestral states.

We further measured the ability of *RRphylo-noder* to capture the processes producing the observed patterns in real data. In particular, we analysed the ability of *RRphylo-noder* to reveal the existence of phenotypic drift for the clades that were designed to be so, in spite of sampling. This could be accomplished by using the RRphylo package function *search.trend* (Castiglione et al. 2019). The *search.trend* algorithm uses the *RRphylo* phenotypic estimates at nodes and the tip values to test whether there is a phenotypic drift (a change in the mean phenotype over time) departing significantly from the Brownian motion expectation. The algorithm can be applied indifferently to the entire tree or to a selection of internal nodes into the tree (Castiglione et al. 2019). For each clade selected, we calculated the intercept of the regression between the clade phenotypes and time (i.e. the distance of each tip from the tree root) and assumed these intercepts as the ancestral values for each clade, to be passed on to *RRphylo-noder* as known ancestral values. Overfit should result in reduced power to retrieve the imposed phenotypic drift under subsampling. At the same time, this allows to test whether the use of known ancestral values increases the power of *RRphylo-noder* to retrieve the true evolutionary process simulated on the tree. In the RRphylo package, we further provide a new function, named *overfitRR*, which tests whether a given phenotypic or rate pattern (either at specific nodes or for the entire tree) is robust to sampling and to phylogenetic uncertainty, that is an important source of concern in phylogenetic comparative methods, especially working with fossil phylogenies (Bapst et al. 2016).

Real Cases

We tested the *RRphylo-noder* method on two real cases. First, we inspected the evolution of body size in mysticete cetaceans. The second real case regards the evolution of body size in caniform carnivores (Online Resource 1).

Baleen whales are among the largest species ever lived. Yet, early Mysticeti include much smaller representatives (Fitzgerald 2012; Serio et al. 2019). The sister group to baleen whales, the Oligocene Aetiocetidae, were toothed whales up to 8 m in body length. We assembled a cetacean phylogenetic tree from the backbone phylogenies in Montgomery et al. (2013) and Marx and Fordyce (2015). The composite phylogeny includes 116 species we had body size estimates for (Online Resource 2). Thirty-six species in the tree are extinct (10 archaeoceti, 23 odontoceti, 3 mysticeti). We tested whether baleen whales body size increased over time, in keeping with Cope's rule (Hone and Benton 2005; Raia et al. 2012) by using *search.trend* (Castiglione et al. 2019). Then, we used *RRphylo-noder* setting *Mystacodon selenensis* (Lambert et al. 2017) body size as the mysticete most recent common ancestor prior. *Mystacodon selenensis* was almost the size of a bottlenose dolphin (McQuate 2017).

It is considered a stem mysticete, perhaps sister to Llanocetidae (Fordyce and Marx 2018). We therefore settled *Mystacodon* body size at 150 kg, which is typical for a bottlenose dolphin. We performed *RRphylo* and *search.trend* on the phylogeny as a whole, either with *Mystacodon* as ancestor of mysticetes, or without it. In addition, we performed, for the sake of comparison, both *RRphylo* and *search.trend* on the same tree and data, but removing extinct mysticeti. Body size estimates for the mysticete most recent common ancestor were further estimated by means of the functions *ace*, *fastAnc*, *StableTraits, StableTraits-Brownian* and *RRphylo*, always the full tree (i.e. with fossil species) without using *Mystacodon* body size as the ancestral value to Mysticeti.

To assess the potential for overfit, we developed and applied the newly-implemented RRphylo function *overfitRR* to test the effect of sampling on results produced by *search.trend*. This function randomly removes a number of tips corresponding to 25% of the tree size and swaps species phylogenetic position (thereby accounting for phylogenetic uncertainty) by using the RRphylo function *swapONE*. Then, it performs *search.trend* on pruned tree and data. The procedure is repeated 100 times and the percentage of significant results returned. In this case, we specified the Mysticeti clade to be tested for temporal trends in phenotypic (body size) mean and rates.

Results

Ancestral State Reconstruction Methods Accuracy: Overall Results

The regression of simulated versus fitted ancestral states gives slope consistently close to 1 and intercept close to 0 for all methods when the starting phenotype was simulated according to BM (Fig. 1, Supplementary Table S1). The same applies when the starting phenotype was simulated according to BM with trend, but intercepts tend to become negative for all methods in this case (Fig. 1, Supplementary Table S2). ANOVA indicated that for 16 out of 40 different kinds of phenotypes there are significant differences among methods in terms of *rmse*. Among them, TukeyHSD indicates *StableTraits-Brownian*, *RRphylo-noder* and *StableTraits* resulted the most accurate methods overall, being selected 16, 15 and 15 times respectively among the best models. The methods *ace* (3) and *RRphylo* (1) still figure among the best candidate models. In terms of *reduced rmse*, only four times we found significant differences between methods. *StableTraits* was selected as the best method 4 times, *StableTraits-Brownian*, *RRphylo-noder* and *fastAnc* 2 times each. Details about the effect of individual factors are available as supplementary information. In general, both *StableTraits* methods and *RRphylo-noder* perform equally well, under a variety of sampling conditions and across different kinds of phenotypes (Online Resource 1). Variable rates methods consistently outperform single-rate methods. Only four times

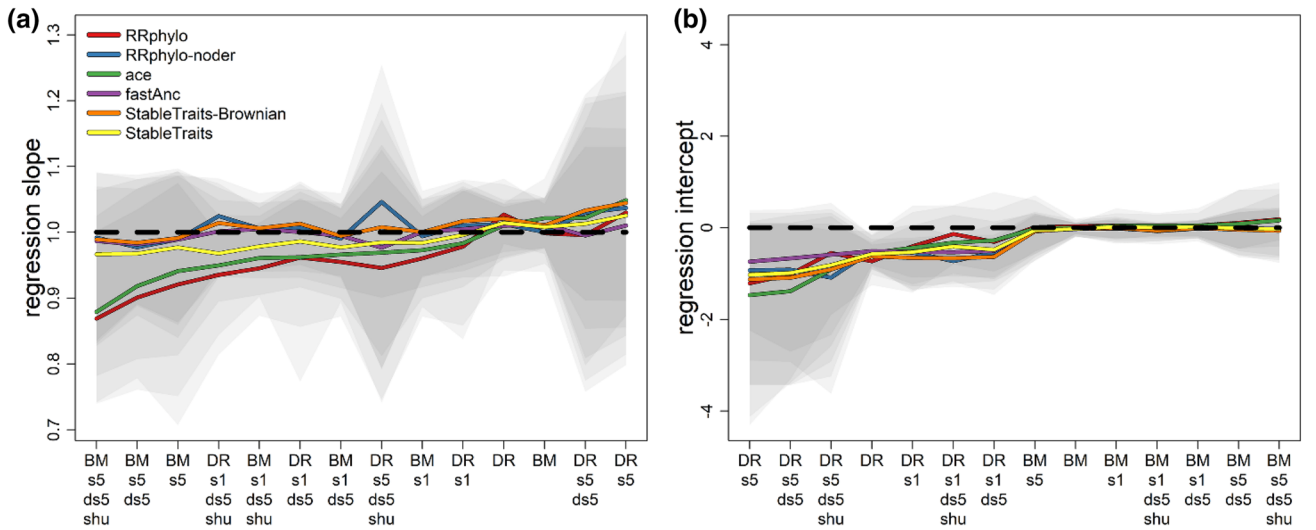

Fig. 1 Patterns of slope and intercept values derived by regressing ancestral states estimated by each method against original values, under different simulation models. A subsample of models is selected as to represent increasing levels of complexity. **a** Along x-axis, models are sorted in ascending order of slope predicted by using *ace* estimates. **b** Along x-axis, models are sorted in ascending order of intercept predicted by using *ace* estimates. Gray shaded areas represent the 95% confidence intervals for each method

methods without specified ancestral values (ape's *ace* and *RRphylo*) perform as well as methods that allow for their specification.

As the complexity of the evolutionary process generating the simulated phenotypes increases, *RRphylo-noder* and *StableTraits*, in this order, performs best (Fig. 2, Table 2). The estimated marginal means of the regression indicate that all methods with prior phenotypic knowledge about specific nodes in the tree outperform methods with no such information (Table 2) and *RRphylo-noder* is the most accurate method overall. Collating methods' *rmse* differently does not affect this result. *RRphylo-noder* remains the least sensitive to change in the type of phenotype simulated under all possible ordering.

When the original phenotype was simulated according to BM, *fitContinuous* indicated BM describes the data better than the 'drift' model 82% of the times at $\Delta AICc > 2$ (Burnham and Anderson 2004). With manipulated phenotypes (i.e. phenotypes with values at nodes estimated through a non-BM process) this percentage decreases to 79%. The corresponding figures for *search.trend* are 97.5% on the original BM-generated phenotype $\boldsymbol{y}$, 93% on the phenotype without- and 92% with specifying ancestral values derived from $\boldsymbol{y}$ via manipulations. Starting from the phenotype $\boldsymbol{yt}$, generated through the BM with trend process, *fitContinuous* indicated the 'drift' model is more appropriate than BM 100% of the cases with the original phenotype and 98.5% with the manipulated (i.e. phenotypes with values at nodes

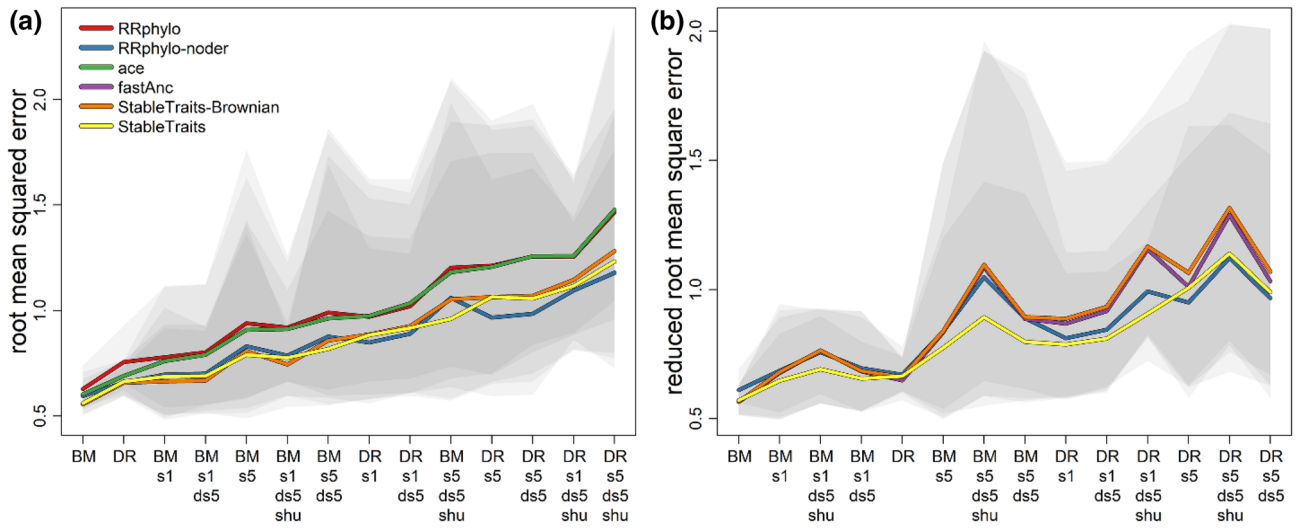

Fig. 2 Patterns of *rmse* and *reduced rmse* across different phenotypes and sampling schemes. A subsample of models is selected as to represent increasing levels of complexity. **a** Along x-axis, models are sorted in ascending order of *rmse* predicted by *ace*. **b** Along x-axis, models are sorted in ascending order of *reduced rmse* predicted by *StableTraits-Brownian*. Gray shaded areas represent the 95% confidence intervals for each method

Table 2 Methods performance and differences between methods across all types of phenotypes

Method	Slope	95% CI		Emm	95% CI	
rmse						
RRphylo-noder	0.012	0.010	0.013	0.918	0.902	0.934
StableTraits	0.014	0.013	0.015	0.937	0.921	0.953
StableTraits-Brownian	0.015	0.014	0.017	0.945	0.929	0.961
RRphylo	0.016	0.015	0.017	1.062	1.046	1.078
ace	0.017	0.016	0.018	1.070	1.054	1.086
Reduced rmse						
RRphylo-noder	0.011	0.009	0.013	0.918	0.902	0.935
fastAnc	0.014	0.012	0.015	0.910	0.894	0.927
StableTraits	0.015	0.013	0.016	0.937	0.920	0.954
StableTraits-Brownian	0.016	0.014	0.017	0.945	0.928	0.962

'slope' represents the increase in *rmse* or *reduced rmse* either as the phenotype deviates from Brownian motion. 'emm' represents the estimated marginal mean of *rmse* or *reduced rmse* either per method

estimated with a non-BM with trend process) phenotypes derived from ***yt***. The corresponding figures for *search.trend* are 79% (original), 84% (manipulated) and 82.5% (manipulated with specified ancestral values).

Assessing the Impact of Sampling

When the tree is subsampled, the slope and intercepts for variable-rates methods remain close to 0 and 1 respectively, when the starting phenotype was ***y***, and slightly less than 0 and 1, respectively, when the starting phenotype was ***yt*** (Fig. 1, Supplementary Tables S7-S8).

We tested whether sampling affects these parameters as well as the percent phenotypic deviation from the simulated parameters by means of ANOVA and post-hoc testing, using the sampling intensity (either 90% of the original tree or 50%) and sampling type (either 'biased' or 'random') as factors. The results indicate that slopes, intercepts and percent deviations from the original phenotypes are never statistically different with the random sampling, except for *StableTraits* when sampling intensity is 50% (Table 3). With intense sampling (i.e. reducing tree size to 50% of the original tree) and the biased sampling design the slope and intercepts differ significantly from the unsampled tree and data for all methods. Importantly, all single-rate models perform worse, both in terms of slope and intercepts change across sampling levels (Supplementary Table S11).

We used the *search.trend* function in RRphylo package to test whether sampling affects the probability to retrieve the correct structure in the data. ANOVA and post-hoc tests indicate there is no significant difference per sampling scheme and intensity. However, the use of ancestral values sensibly increases the possibility to find a phenotypic pattern at specific clades when it is real. Under different sampling conditions this increase in power is as high as 82.2% on average (Online Resource 1).

Cetacean Body Mass

By applying *search.trend*, we found Cope's rule to apply to mysticetes, regardless of whether ancestral states are indicated as node priors or ignored. Yet, the regression slope increases adding *Mystacodon* body mass as the ancestor of Mysticeti (Fig. 3, Table 4). The results are robust to the effect of sampling (97% and 74% instances of significant phenotypic trends are found with and without *Mystacodon* body mass as the ancestral value to all Mysticeti, respectively, by removing 25% of the tips randomly with *overfitRR*).

By applying *RRphylo-noder* the cetacean phylogeny produced an estimate of 150.04 kg for the most recent common ancestor of Mysticeti, which is coincident with the ***fossil. state*** provided (i.e. the size of *Mystacodon selenensis*). The same estimate as calculated by *RRphylo* without ***fossil. state*** is 385.36 kg. We derived for comparison the corresponding values as estimated by *ape* and *fastAnc* (which is 430.70 kg and 457.85 kg, respectively), and by *StableTraits*. At 150.02 kg.

Discussion

Variable-rates methods, that is *RRphylo-noder* and the two *StableTraits* models, consistently outperform all other methods in terms of ancestral states prediction accuracy (the accuracy of *fastAnc* rapidly decreases with the number of ancestral values used or by shuffling them, Fig. 2).

Table 3 Effect of sampling on prediction performance of variable-rates methods for ancestral states estimation

	Intercept			Slope			Percent phenotypic deviation		
	100%	90%	50%	100%	90%	50%	100%	90%	50%
StableTraits-Brownian									
100%	–	0.957	< 0.01	–	0.949	< 0.01	–	0.623	0.984
90%	0.996	–	< 0.01	1	–	0.001	0.999	–	0.52
50%	0.785	0.736	–	0.096	0.1	–	0.668	0.67	–
StableTraits									
100%	–	0.979	< 0.01	–	0.979	< 0.01	–	0.616	0.983
90%	0.940	–	< 0.01	0.867	–	< 0.01	0.999	–	0.511
50%	0.905	0.727	–	0.024	0.087	–	0.701	0.698	–
RRphylo-noder									
100%	–	0.999	< 0.01	–	0.988	< 0.01	–	0.611	0.988
90%	0.840	–	< 0.01	0.955	–	< 0.01	0.616	–	0.524
50%	0.979	0.929	–	0.121	0.062	–	0.983	0.511	–

For each sampling level (100, 90 or 50% of the original tree) the results represent p-values of the post hoc Tukey HSD test, either for biased (upper triangle) and random (lower triangle) sampling designs

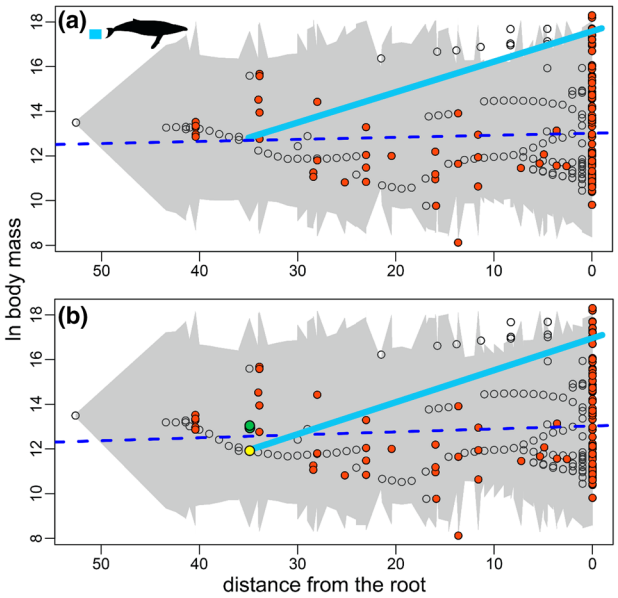

Fig. 3 Cetacean body size versus time plots. White dots represent ancestral estimates at internal nodes, orange dots represent species phenotypes. The regression of phenotypes through time for the entire phylogeny is indicated by a blue dashed line. The phenotypic trend through time for Mysticeti is represented by the solid pale blue line. Upper row: cetaceans body size evolution according to *search.trend* as produced by considering fossil mysticetes, ancestral estimates are derived by the *RRphylo* method. Lower row: cetaceans body size evolution according to *search.trend* as produced by considering fossil mysticetes, ancestral estimates are derived by the *RRphylo-noder* method. The yellow dot represents the ancestral character for *Mystacodon* as estimated by *RRphylo-noder* and *StableTraits* (fitted by using the StableTraits software), green dots represent *ace* (fitted by using ape function ace), and *fastAnc* (fitted by using phytools function fastAnc). The *ace* estimate is the lowest. The y-axis is in ln-grams, time (x-axis) represents the distance from the cetacean tree root

Table 4 Estimates for the body size at the mrca to Mysticeti, as derived by different methods

	Fitted ancestral values (kilograms)	Slope	p-value
RRphylo-noder—*Mystacodon* as ancestral value	150.04	0.020	**< 0.01**
StableTraits—*Mystacodon* as ancestral value	150.02	–	–
RRphylo	385.36	0.019	**0.01**
Ace	430.70	–	–
Fastanc	457.85	–	–

Slopes and p-values for Mysticete body size versus time regression performed by *search.trend*. Significant regression slopes are indicated in bold

However, *RRphylo-noder* is the least sensitive to changes in the complexity of phenotypes and provides the smallest root mean squared error overall (Fig. 2, Table 2). The inclusion of known phenotypic values at internal nodes substantially increases ancestral states estimation accuracy for all methods and regardless of the sampling scheme. With *RRphylo-noder*, this further translates into a nearly twofold increase in the power to detect the phenotypic drift imposed to specific clades within the tree in the simulations. The process generating the phenotypes ***y*** and ***yt*** is recognized more precisely by *search.trend* with BM and by *fitContinuous* with BM with trend.

We found that all methods provide unbiased estimation of the ancestral states, meaning the estimation error is not correlated with the phenotypic values. However, the intercepts of simulated versus estimated ancestral states tend to be less than zero when the starting phenotype was simulated under a model of phenotypic drift (Fig. 1). Since we designed BM with trend as to exhibit a positive drift, this means that all methods tend to preserve the actual phenotypic pattern at the expense of prediction accuracy. These results comply with our expectations that methods assuming an a priori evolutionary model are more prone to estimation error for all nodes other than the ancestral values known in advance, as compared to *RRphylo-noder*. Still, our results demonstrate that variable rates methods are best suited to cope with tree and data generated under complex phenotypic processes that cannot be captured by abstract evolutionary models (Chira and Thomas 2016; Slater, Harmon, and Alfaro 2012). In *RRphylo* the phenotypic difference between any parent to descendant pair in the tree is fitted as a linear transformation proportional to the time intervening between the two according to a given slope (i.e. the elements of the evolutionary rates vector $\vec{\beta}$) while minimizing rate variation within clades (Castiglione et al. 2018). In contrast, other variable-rate methods refer to a single evolutionary model describing the distribution of phenotypic changes for the whole tree (e.g. the normal distribution in BM, the stable distribution in *StableTraits*). This makes *RRphylo* the least sensible to the actual shape of the distribution of phenotypic change across the tree.

Bayesian estimation of ancestral states, as currently implemented in (at least) phytools' *anc.Bayes* and in geiger's *fitContinuousMCMC* (but see Bokma et al. 2015a, b), or *StableTraits* (Elliot and Mooers 2014), accommodates for ancestral state estimation uncertainty and provides credible estimates when node priors are used. Herein, by using both simulations and application to real cases (mysticeti and caniform carnivores, see Online Resource 1 for the latter case study) we demonstrated *RRphylo-noder* performs at least as well as such Bayesian estimation approaches, and bears the advantage of being much faster (nearly twenty times faster according to our simulations)

and less dependent to a specific distribution of phenotypic changes.

One potential problem with method fitting many parameters at once is overfit. Overfit methods tend to perform very well with given data, but they often bear the potential to provide biased and much less precise estimation when data are subsampled. We applied either mild or strong subsampling to our tree and data in the simulations and to one real case (Mysticeti). We found that both *RRphylo-noder* and *StableTraits* are robust to sampling effects but for the most severe sampling design (i.e. removing half of the tree species proportionally to the species phenotypic values). In phylogenetic ridge regression (Kratsch and McHardy 2014) a normalization factor λ is applied to avoid abnormally large phenotypic rate estimates, at the expense of prediction accuracy of tip (species) values. In *RRphylo*, λ maximum likelihood estimation is performed as to minimize the variance of rates within clades, so that phenotypes (and rates) tend to show phylogenetic signal (Castiglione et al. 2018). The structure of rate variation is constrained to maintain patterns of phenotypic evolution within clades and rates are treated as phylogenetically non-independent (Sakamoto and Venditti 2018). A similar approach to reduce overfit, that is the "inheritance" of evolutionary rates over time, is implemented in AUTEUR (Eastman et al. 2011). We guess this is the reason why *RRphylo-noder* is robust to even strong sampling effects. In *StableTraits* the danger of overfit is minimized by penalizing the effective number of parameters in the model by application of a Bayesian Predictive Information Criterion (Elliot and Mooers 2014).

We found Cope's rule to apply to mysticetes. However, the phenotypic drift becomes much more evident when the bottlenose dolphin-sized *M. selenensis* is placed at the root of the Mysticeti clade (Fig. 3). We similarly found positive evidence for Cope's rule in canids (Supplementary Fig. S13). Still in this case, this depends on the inclusion of fossil phenotypes known in advance at specific nodes (see Online Resource 1 for full details). The application of *RRphylo-noder* to these case studies demonstrates the importance of using the fossil record to guide the recognition of phenotypic patterns under a PCM context, which has been pointed out several times in other studies (Bokma et al. 2015a, b; Benson et al. 2014; Finarelli and Flynn 2006; Hunt and Slater 2016; Puttick 2016; Puttick and Thomas 2015; Webster and Purvis 2002) and remains evident here also with methods other than *RRphylo-noder* (see Fig. 3).

Although ancestral states estimation is usually difficult (Cooper et al. 2016; Joy et al. 2016) and generally constrained within the limits of actual phenotypes at the tree tips (Gascuel and Steel 2014), new approaches are being developed to provide more sensible estimates. Herein, we demonstrated that *RRphylo-noder* is virtually as powerful in fitting ancestral states as other available methods, and slightly more accurate in terms of fitting the true ancestral states when the description of phenotypic evolution is not reducible to a simple evolutionary model. This better accuracy probably depends on the fact that *RRphylo-noder* does not need to comply to the predictions of any abstract hypothesis about the tempo and mode of evolutionary change. As such, we believe *RRphylo-noder* is the most appropriate in cases of complex phenotypic distributions. This is especially true by considering that competing methods using Monte Carlo Markov Chain (MCMC) approaches require much longer computational times.

In our simulations we deliberately produced complex phenotypic patterns which change significantly across different branches of the tree. Although simple approaches such as *ace* and *fastAnc* may work well, even better than *RRphylo-noder* when the process behind phenotypic evolution is the BM, we suspect the existence of idiosyncratic evolutionary processes applying to different sections of the tree is common. For instance, body size increase over time, the well-known Cope's rule, does not apply to all mammalian clades (Monroe and Bokma 2010; Raia et al. 2012), all dinosaurs (Benson et al. 2014; Hone et al. 2005), or all early amniote clades (Laurin 2004), meaning that when these clades are tested for Cope's rule neither BM or BM with drift may serve as realistic evolutionary models. Benson et al. (2014a) presented an interesting case for change in the intensity of Cope's rule in pterosaurs because of birds' diversification starting at the beginning of the Cretaceous. The dinosaur clade including birds (maniraptoran dinosaurs) decreased, rather than increasing in body size throughout the Cretaceous (Lee et al. 2014) while the rest of the dinosaurs were still growing large. Baker et al. (2015) successfully applied a variable-rates model of phenotypic evolution to show that Cope's rule applies with different intensities to 10 out of 11 mammalian orders. There is evidence that competition and niche incumbency influenced the timing and direction of phenotypic change in turtles (Rosenzweig and McCord 1991), dinosaurs (McNab 2009; Sookias et al. 2012), and insects (Waller and Svensson 2017). We believe that even this short account is enough to suggest that referring to a simple array of evolutionary models could be risky, especially when fossil phenotypes are included. At one time the inclusion of such fossil phenotypes as the ancestral conditions for some clades in the phylogeny does increase the power and reliability of PCMs and makes it less likely that the actual pattern of phenotypic evolution complies to the prediction of any abstract evolutionary model when this is wrong. We believe *RRphylo-noder* constitutes a worthy addition to the PCM toolbox, especially in terms of providing sensible ancestral state estimates with complex phenotypes, and when the recognition of temporal trends in trait evolution is the goal.

Acknowledgements We are grateful to Gianmarco Tesone and Martina Piccolo for sharing the cetacean body size data with us. The authors declare no conflict of interest.

Data Availability The R package RRphylo is available at https://github.com/pasraia/RRphylo. Raw data and phylogenetic tree for caniform carnivores are available as example dataset in the R package geiger. Raw data and phylogenetic tree for cetaceans are available as example dataset in the R package RRphylo. The R code to reproduce all the simulations performed in this study, together with raw data for cetaceans are available as Supplementary Information.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Alfaro, M. E., Santini, F., Brock, C., Alamillo, H., Dornburg, A., Rabosky, D. L., et al. (2009). Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proceedings of the National Academy of Sciences of the United States of America, 106*(32), 13410–13414. https://doi.org/10.1073/pnas.0811087106.

Bapst, D. W. (2013). A stochastic rate-calibrated method for time-scaling phylogenies of fossil taxa. *Methods in Ecology and Evolution, 4*(8), 724–733. https://doi.org/10.1111/2041-210X.12081.

Bapst, D. W., Wright, A. M., Matzke, N. J., & Lloyd, G. T. (2016). Topology, divergence dates, and macroevolutionary inferences vary between different tip-dating approaches applied to fossil theropods (Dinosauria). *Biology Letters, 12*, 20160237.

Baker, J., Meade, A., Pagel, M., & Venditti, C. (2015). Adaptive evolution toward larger size in mammals. *Proceedings of the National academy of Sciences of the United States of America, 112*, 5093–5098.

Bokma, F., Godinot, M., Maridet, O., Ladevèze, S., Costeur, L., Solé, F., et al. (2015a). Testing for Depéret's Rule (Body Size Increase) in Mammals using Combined Extinct and Extant Data. *Systematic Biology, 65*, 98–108.

Behrensmeyer, A. K., Kidwell, S. M., & Gastaldo, R. A. (2000). Taphonomy and paleobiology. *Paleobiology, 26*, 103–147. https://doi.org/10.1666/0094-8373(2000)26[103:TAP]2.0.CO;2.

Benson, R. B. J., Campione, N. E., Carrano, M. T., Mannion, P. D., Sullivan, C., Upchurch, P., et al. (2014). Rates of Dinosaur Body Mass Evolution Indicate 170 Million Years of Sustained Ecological Innovation on the Avian Stem Lineage. *PLoS Biology, 12*, e1001853. https://doi.org/10.1371/journal.pbio.1001853.

Bokma, F., Godinot, M., Maridet, O., Ladevèze, S., Costeur, L., Solé, F., et al. (2015b). Testing for Depéret's Rule (Body Size Increase) in Mammals using Combined Extinct and Extant Data. *Systematic Biology*. https://doi.org/10.1093/sysbio/syv075.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociolological Methods Research, 33*, 261–304.

Castiglione, S., Serio, C., Mondanaro, A., Di Febbraro, M., Profico, A., Girardi, G., et al. (2019). Simultaneous detection of macroevolutionary patterns in phenotypic means and rate of change with and within phylogenetic trees including extinct species. *PLoS ONE, 14*(1), e0210101–e210113. https://doi.org/10.1371/journal.pone.0210101.

Castiglione, S., Tesone, G., Piccolo, M., Melchionna, M., Mondanaro, A., Serio, C., et al. (2018). A new method for testing evolutionary rate variation and shifts in phenotypic evolution. *Methods in Ecology and Evolution, 9*, 974–983. https://doi.org/10.1111/2041-210X.12954.

Chira, A. M., & Thomas, G. H. (2016). The impact of rate heterogeneity on inference of phylogenetic models of trait evolution. *Journal of Evolutionary Biology, 29*(12), 2502–2518. https://doi.org/10.1111/jeb.12979.

Cooper, N., Thomas, G. H., Venditti, C., Meade, A., & Freckleton, R. P. (2016). A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies. *Biological Journal of the Linnean Society, 118*(1), 64–77. https://doi.org/10.1111/bij.12701.

Didier, G., Fau, M., & Laurin, M. (2017). Likelihood of Tree Topologies with Fossils and Diversification Rate Estimation. *Systematic Biology, 66*(6), 964–987. https://doi.org/10.1093/sysbio/syx045.

Eastman, J. M., Alfaro, M. E., Joyce, P., Hipp, A. L., & Harmon, L. J. (2011). A Novel Comparative Method For Identifying Shifts In The Rate Of Character Evolution On Trees. *Evolution, 65*(12), 3578–3589. https://doi.org/10.1111/j.1558-5646.2011.01401.x.

Elliot, M. G., & Mooers, A. O. (2014). Inferring ancestral states without assuming neutrality or gradualism using a stable model of continuous character evolution. *BMC evolutionary biology, 14*(1), 226. https://doi.org/10.1186/s12862-014-0226-8.

Felsenstein, J. (1985). Phylogenies and the comparative method. *American Naturalist, 125*, 1–15.

Finarelli, J. A., & Flynn, J. J. (2006). Ancestral state reconstruction of body size in the Caniformia (Carnivora, mammalia): The effects of incorporating data from the fossil record. *Systematic Biology, 55*(2), 301–313. https://doi.org/10.1080/10635150500541698.

Finarelli, J. A., & Liow, L. H. (2016). Diversification histories for North American and Eurasian carnivorans. *Biological Journal of the Linnean Society, 118*(1), 26–38. https://doi.org/10.1111/bij.12777.

Fitzgerald, E. M. G. (2012). Archaeocete-like jaws in a baleen whale. *Biology Letters, 8*(1), 94–96. https://doi.org/10.1098/rsbl.2011.0690.

Fordyce, R. E., & Marx, F. G. (2018). Gigantism Precedes Filter Feeding in Baleen Whale Evolution. *Current biology : CB, 28*(10), 1670–1676.e2. https://doi.org/10.1016/j.cub.2018.04.027.

Freckleton, R. P., Harvey, P. H., & Pagel, M. (2002). Phylogenetic Analysis and Comparative Data: A Test and Review of Evidence. *The American Naturalist, 160*(6), 712–726. https://doi.org/10.1086/343873.

Garland, T., Jr., & Ives, A. R. (2000). Using the Past to Predict the Present: Confidence Intervals for Regression Equations in Phylogenetic Comparative Methods. *The American Naturalist, 155*(3), 346–364. https://doi.org/10.1086/303327.

Gascuel, O., & Steel, M. (2014). Predicting the ancestral character changes in a tree is typically easier than predicting the root state. *Systematic Biology, 63*(3), 421–435. https://doi.org/10.1093/sysbio/syu010.

Harmon, L. J., Schulte, J. A., Larson, A., & Losos, J. B. (2003). Tempo and Mode of Evolutionary Radiation in Iguanian Lizards. *Science, 301*(5635), 961–964. https://doi.org/10.1126/science.1084786.

Harmon, L. J., Weir, J. T., Brock, C. D., & Glor, R. E. (2007). (2007). GEIGER: investigating evolutionary radiations. *Bioinformatics, 24*, 129–131.

Harvey, P. H., & Pagel, M. D. (1991). *The comparative method in evolutionary biology*. Oxford: Oxford University Press.

Heath, T. A., Huelsenbeck, J. P., & Stadler, T. (2014). The fossilized birth-death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences of the United States of America, 111*(29), E2957–E2966. https://doi.org/10.1073/pnas.1319091111.

Hone, D. W. E., Keesey, T. M., Pisani, D., & Purvis, A. (2005). Macroevolutionary trends in the Dinosauria: Cope's rule. *Journal*

of evolutionary biology, 18(3), 587–595. https://doi.org/10.1111/j.1420-9101.2004.00870.x.

Hone, D., & Benton, M. (2005). The evolution of large size: how does Cope's Rule work? *Trends in ecology & evolution, 20*(1), 4–6. https://doi.org/10.1016/j.tree.2004.10.012.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal, 50(3),* 346–363.

Hunt, G., & Slater, G. (2016). Integrating Paleontological and Phylogenetic Approaches to Macroevolution. *Annual Review of Ecology, Evolution, and Systematics, 47*(1), 189–213. https://doi.org/10.1146/annurev-ecolsys-112414-054207.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). New York, NY: Springer Science & Business Media.

Joy, J. B., Liang, R. H., McCloskey, R. M., Nguyen, T., & Poon, A. F. Y. (2016). Ancestral Reconstruction. *PLoS Computational Biology, 12*(7), e1004763–e1004820. https://doi.org/10.1371/journal.pcbi.1004763.

Kratsch, C., & McHardy, A. C. (2014). RidgeRace: ridge regression for continuous ancestral character estimation on phylogenetic trees. *Bioinformatics, 30*(17), i527–i533. https://doi.org/10.1093/bioinformatics/btu477.

Lambert, O., Martínez-Cáceres, M., Bianucci, G., Di Celma, C., Salas-Gismondi, R., Steurbaut, E., et al. (2017). Earliest Mysticete from the Late Eocene of Peru Sheds New Light on the Origin of Baleen Whales. *Current biology : CB, 27*(10), 1535–1541.e2. https://doi.org/10.1016/j.cub.2017.04.026.

Laurin, M. (2004). The Evolution of Body Size, Cope's Rule and the Origin of Amniotes. *Systematic Biology, 53*(4), 594–622. https://doi.org/10.1080/10635150490445706.

Lee, M. S. Y., Cau, A., Naish, D., & Dyke, G. J. (2014). Dinosaur evolution Sustained miniaturization and anatomical innovation in the dinosaurian ancestors of birds. *Science, 345*(6196), 562–566. https://doi.org/10.1126/science.1252243.

Lenth, R. (2019). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.4.3.01. https://CRAN.R-project.org/package=emmeans.

Liow, L. H., Quental, T. B., & Marshall, C. R. (2010). When can decreasing diversification rates be detected with molecular phylogenies and the fossil record? *Systematic Biology, 59*(6), 646–659. https://doi.org/10.1093/sysbio/syq052.

Marx, F. G., & Fordyce, R. E. (2015). Baleen boom and bust: a synthesis of mysticete phylogeny, diversity and disparity. *Open Science, 2*(4), 140434–140434. https://doi.org/10.1098/rsos.140434.

McNab, B. K. (2009). Resources and energetics determined dinosaur maximal size. *Proceedings of the National Academy of Sciences of the United States of America, 106*(29), 12184–12188. https://doi.org/10.1073/pnas.0904000106.

McQuate, S. (2017). Fossil of oldest known baleen-whale relative unearthed in Peru. *Nature Publishing Group*. https://doi.org/10.1038/nature.2017.21966.

Meloro, C., Raia, P., & Barbera, C. (2007). Effect of predation on prey abundance and survival in Plio-Pleistocene mammalian communities. *Evolutionary Ecology Research, 9*(3), 505–525.

Mitchell, J. S., Etienne, R. S., & Rabosky, D. L. (2018). Inferring Diversification Rate Variation From Phylogenies With Fossils. *Systematic Biology, 68*(1), 1–18. https://doi.org/10.1093/sysbio/syy035.

Monroe, M. J., & Bokma, F. (2010). Little evidence for Cope's rule from Bayesian phylogenetic analysis of extant mammals. *Journal of Evolutionary Biology, 23*(9), 2017–2021. https://doi.org/10.1111/j.1420-9101.2010.02051.x.

Montgomery, S. H., Geisler, J. H., McGowen, M. R., Fox, C., Marino, L., & Gatesy, J. (2013). The Evolutionary History Of Cetacean Brain And Body Size. *Evolution, 67*(11), 3339–3353. https://doi.org/10.1111/evo.12197.

O'Meara, B. C. (2012). Evolutionary Inferences from Phylogenies: A Review of Methods. *Annual Review of Ecology, Evolution, and Systematics, 43*(1), 267–285. https://doi.org/10.1146/annurev-ecolsys-110411-160331.

O'Meara, B. C., Anè, C., Sanderson, M. J., & Wainwright, P. C. (2006). Testing for different rates of continuous trait evolution using likelihood. *Evolution, 60*, 922–933.

Pagel, M. (1997). Inferring evolutionary processes from phylogenies. *Zoologica Scripta, 26*(4), 331–348. https://doi.org/10.1111/j.1463-6409.1997.tb00423.x.

Paradis, E., Schliep, K. (2018). ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*, 526–528.

Pennell, M. W., & Harmon, L. J. (2013). An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. *Annals of the New York Academy of Sciences, 1289*(1), 90–105. https://doi.org/10.1111/nyas.12157.

Pennell, M. W., FitzJohn, R. G., Cornwell, W. K., & Harmon, L. J. (2015). Model Adequacy and the Macroevolution of Angiosperm Functional Traits. *The American Naturalist, 186*(2), E33–50. https://doi.org/10.1086/682022.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2014). nlme: Linear and nonlinear mixed effects models. R package version 3.1–117.

Puttick, M. N. (2016). Partially incorrect fossil data augment analyses of discrete trait evolution in living species. *Biology Letters, 12*(8), 20160392. https://doi.org/10.1098/rsbl.2016.0392.

Puttick, M. N., & Thomas, G. H. (2015). Fossils and living taxa agree on patterns of body mass evolution: a case study with Afrotheria. *Proceedings of the Royal Society B: Biological Sciences, 282*(1821), 20152023–20152029. https://doi.org/10.1098/rspb.2015.2023.

Puttick, M. N., O'Reilly, J. E., Tanner, A. R., Fleming, J. F., Clark, J., Holloway, L., et al. (2017). Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data. *Proceedings of the Royal Society B: Biological Sciences, 284*(1846), 20162290. https://doi.org/10.1098/rspb.2016.2290.

Rabosky, D. L. (2014). Automatic Detection of Key Innovations, Rate Shifts, and Diversity-Dependence on Phylogenetic Trees. *PLoS ONE, 9*(2), e89543–e89615. https://doi.org/10.1371/journal.pone.0089543.

Raia, P., Carotenuto, F., Passaro, F., Fulgione, D., & Fortelius, M. (2012). Ecological Specialization in Fossil Mammals Explains Cope's Rule. *The American Naturalist, 179*(3), 328–337. https://doi.org/10.1086/664081.

Revell, L. J. (2010). Phylogenetic signal and linear regression on species data. *Methods in Ecology and Evolution, 1*(4), 319–329. https://doi.org/10.1111/j.2041-210X.2010.00044.x.

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution, 3*(2), 217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x.

Rosenzweig, M. L., & McCord, R. D. (1991). Incumbent replacement: Evidence for long-term evolutionary progress. *Paleobiology, 17*, 202–213.

Sakamoto, M., & Venditti, C. (2018). Phylogenetic non-independence in rates of trait evolution. *Biology Letters, 14*(10), 20180502. https://doi.org/10.1098/rsbl.2018.0502.

Schnitzler, J., Theis, C., Polly, P. D., & Eronen, J. T. (2017). Fossils matter–understanding modes and rates of trait evolution in Musteloidea (Carnivora). *Evolutionary Ecology Research, 18*, 187–200.

Serio, C., Castiglione, S., Tesone, G., Piccolo, M., Melchionna, M., Mondanaro, A., et al. (2019). Macroevolution of Toothed Whales Exceptional Relative Brain Size. *Evolutionary Biology, 46*(4), 332–342.

Silvestro, D., Zizka, A., Bacon, C. D., Cascales-Miñana, B., Salamin, N., & Antonelli, A. (2016). Fossil biogeography: a new model to infer dispersal, extinction and sampling from palaeontological data. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371*(1691), 20150225–20150313. https://doi.org/10.1098/rstb.2015.0225.

Slater, G. J., & Harmon, L. J. (2013). Unifying fossils and phylogenies for comparative analyses of diversification and trait evolution. *Methods in Ecology and Evolution, 4*(8), 699–702. https://doi.org/10.1111/2041-210X.12091.

Slater, G. J., Harmon, L. J., & Alfaro, M. E. (2012). Integrating fossils with molecular phylogenies improves inference of trait evolution. *Evolution, 66*(12), 3931–3944. https://doi.org/10.1111/j.1558-5646.2012.01723.x.

Smaers, J. B., Mongle, C. S., & Kandler, A. (2016). A multiple variance Brownian motion framework for estimating variable rates and inferring ancestral states. *Biological Journal of the Linnean Society, 118*(1), 78–94. https://doi.org/10.1111/bij.12765.

Sookias, R. B., Butler, R. J., & Benson, R. B. J. (2012). Rise of dinosaurs reveals major body-size transitions are driven by passive processes of trait evolution. *Proceedings of the Royal Society B: Biological Sciences, 279*(1736), 2180–2187. https://doi.org/10.1098/rspb.2011.2441.

Venditti, C., Meade, A., & Pagel, M. (2011). Multiple routes to mammalian diversity. *Nature, 479*(7373), 393–396. https://doi.org/10.1038/nature10516.

Waller, J. T., & Svensson, E. I. (2017). Body size evolution in an old insect order: No evidence for Cope's Rule in spite of fitness benefits of large size. *Evolution, 71*(9), 2178–2193. https://doi.org/10.1111/evo.13302.

Webster, A. J., & Purvis, A. (2002). Testing the accuracy of methods for reconstructing ancestral states of continuous characters. *Proceedings of the Royal Society B: Biological Sciences, 269*(1487), 143–149. https://doi.org/10.1098/rspb.2001.1873.