Syst. Biol. 63(3):293–308, 2014

DOI:10.1093/sysbio/syt066
Advance Access publication October 22, 2013

Robust Regression and Posterior Predictive Simulation Increase Power to Detect Early Bursts of Trait Evolution

GRAHAM J. SLATER[1,2,*] AND MATTHEW W. PENNELL[3,4]

[1]*Department of Ecology and Evolutionary Biology, University of California Los Angeles, 610 Charles E. Young Drive East, Los Angeles, CA, 90095-7239, USA;* [2]*Department of Paleobiology & Division of Mammals, National Museum of Natural History, Smithsonian Institution, MRC 121, PO Box 37012, Washington, DC., 20013-7012, USA;* [3]*Institute for Bioinformatics and Evolutionary Studies, University of Idaho, 441D Life Sciences South, PO Box 443051, Moscow, ID, 83844-3051, USA; and* [4]*National Evolutionary Synthesis Center, 2024 W. Main Street, Suite A200, Durham, NC, 27705-4667, USA*

[*]*Correspondence to be sent to: Graham J. Slater, Department of Paleobiology, National Museum of Natural History, Smithsonian Institution, MRC 121, PO Box 37012, Washington, DC., 20013-7012, USA; E-mail: SlaterG@si.edu.*

Received 26 November 2012; reviews returned 18 April 2013; accepted 4 October 2013
Associate Editor: Jeremy Brown

Abstract.—A central prediction of much theory on adaptive radiations is that traits should evolve rapidly during the early stages of a clade's history and subsequently slowdown in rate as niches become saturated—a so-called "Early Burst." Although a common pattern in the fossil record, evidence for early bursts of trait evolution in phylogenetic comparative data has been equivocal at best. We show here that this may not necessarily be due to the absence of this pattern in nature. Rather, commonly used methods to infer its presence perform poorly when when the strength of the burst—the rate at which phenotypic evolution declines—is small, and when some morphological convergence is present within the clade. We present two modifications to existing comparative methods that allow greater power to detect early bursts in simulated datasets. First, we develop posterior predictive simulation approaches and show that they outperform maximum likelihood approaches at identifying early bursts at moderate strength. Second, we use a robust regression procedure that allows for the identification and down-weighting of convergent taxa, leading to moderate increases in method performance. We demonstrate the utility and power of these approach by investigating the evolution of body size in cetaceans. Model fitting using maximum likelihood is equivocal with regards the mode of cetacean body size evolution. However, posterior predictive simulation combined with a robust node height test return low support for Brownian motion or rate shift models, but not the early burst model. While the jury is still out on whether early bursts are actually common in nature, our approach will hopefully facilitate more robust testing of this hypothesis. We advocate the adoption of similar posterior predictive approaches to improve the fit and to assess the adequacy of macroevolutionary models in general. [Adaptive Radiations, Early Burst, Posterior Predictive Simulations, Quantitative Characters]

At the end of the Cretaceous, just before an asteroid and its aftermath wiped out huge swaths of life on earth including the dinosaurs, most mammals were small and, at least compared to today, unremarkable in their diversity (Lillegraven et al. 1979; Alroy 1999). Modern mammalian lineages, many of which had been in existence since the mid-Cretaceous, only subsequently diversified into the diversity of forms we see today (Alroy 1999; Archibald and Deutschman 2001; Luo 2007; Meredith et al. 2011). Characterizing the rise of this diversity has long preoccupied evolutionary biologists, and an intriguing narrative emerges from the plethora of studies of the mammalian fossil record. While there have been successive waves of lineage diversification in mammals (Luo 2007), this appears not to be the case for many important ecomorphological characters. Rather, disparity in such traits as dentition and body size peaked relatively early in the radiation of mammals and then stabilized (Alroy 1999; Wesley-Hunt 2005). The ecomorphological space available to mammals appears to have been saturated early in the Cenozoic, with subsequent lineages representing variations on a few themes. In a broad sense, this is what George Gaylord Simpson (1944; 1953) had in mind when he wrote of adaptive radiations occurring within "adaptive zones."

The notion of adaptive radiations continues to intrigue evolutionary biologists today, and many agree with Schluter's (2000) suggestion that these have generated much of the biodiversity on earth . While there has been a great deal of discussion as to exactly what constitutes an adaptive radiation (Simpson 1953; Schluter 2000; Losos and Miles 2002; Glor 2010; Losos and Mahler 2010; Yoder et al. 2010), a general expectation from much of this verbal theory is that character evolution should show a fairly stereotypical pattern through time. Under the scenario classically envisaged (Simpson 1944, 1953; Schluter 2000), rates of evolution should be rapid early in a clade's history as character displacement and other mechanisms drive species to diverge and should subsequently slow as niches fill up and competition with incumbent species prevents further divergence. This pattern should leave a signature on the distribution of trait values at the tips of a phylogeny (Blomberg et al. 2003; Harmon et al. 2003, 2010; Freckleton and Harvey 2006) and thus in principle be detectable by phylogenetic comparative methods.

Although the early burst pattern of phenotypic evolution has received much support in paleontological studies (Foote 1993, 1994, 1995, 1997; Jernvall et al. 1996; Wesley-Hunt 2005; Lloyd et al. 2012), evidence for its pervasiveness in phylogenetic comparative datasets has been mixed at best. Early, rapid trait evolution has been documented in ratsnakes (Burbrink and Pyron 2010), Australian agamid lizards (Harmon et al. 2003), cetaceans (Slater et al. 2010), ovenbirds (Derryberry et al.

2011), and triggerfishes (Dornburg et al. 2011). Mahler et al. (2010) used a model of diversity-dependent trait evolution to show that the Antillean *Anolis* radiation showed patterns consistent with niche-filling within each island assemblage. In a broader comparative study, Harmon et al. (2010) examined 49 clades including some of the most iconic adaptive radiations. Using a maximum likelihood approach (see later in the text), they found very little support for the Early Burst (EB) model across their datasets; most were better explained by either a Brownian motion (BM) or Ornstein-Uhlenbeck (OU) model of evolution. From this analysis, they suggested that the pattern of a slowdown in evolutionary rate often associated with adaptive radiations was actually rare in comparative data. An analysis of body size evolution across mammalian phylogeny similarly failed to recover evidence for early and rapid evolution, instead finding stronger support for branch-specific variation in evolutionary rates (Venditti et al. 2011; but see Slater 2013).

There are several reason to believe a lack of power, rather than a lack of generality, may underlie our ability to detect early bursts of evolution in phylogenetic comparative datasets. First, for many small datasets, we lack sufficient power to reject simple models of evolution using information theoretic approaches (i.e., high Type II error rates; Boettiger et al. 2012). Of the 49 clades used by Harmon et al. (2010), 39 had $n \leq 40$ taxa and 29 had $n \leq 20$ taxa. It can easily be shown through simulation that early bursts of trait evolution are almost completely unidentifiable in clades of these sizes, even when evolution proceeds in exceptionally rapid bursts (Fig. 1). Second, we are often limited to testing for a time-dependent process using data from a single time-slice—the present day. Although the signature of an early burst should be retained in phylogenetic comparative datasets of extant species, power to detect this signature increases if extinct taxa are included (Slater et al. 2012a). Finally, the EB model assumes that trait evolution slowed at the same pace across the entire clade. If trait evolution in some lineages strayed from the underlying process, perhaps due to convergent selection or an escape from the adaptive zone, then these secondary processes may deteriorate the power of the likelihood formulation of the EB model to detect an early burst. For example, Slater et al. (2010) found evidence for an early burst of body size evolution in cetaceans only after removing two secondarily large dolphin species, from their dataset. They attributed the deviation in body size evolution in these lineages to convergence resulting from ecological replacement of large predatory sperm whale lineages that went extinct in the late Miocene (e.g., Bianucci and Landini 2006). Such a pattern does not invalidate an overall pattern of declining rates resulting from niche filling *sensu* Simpson (1944, 1953); no theory of adaptive radiation explicitly requires monophyly and extinction of taxa or entire clades within an adaptively radiating lineage will create ecological opportunity that closely related lineages are expected to fill (Schluter 2000; Glor 2010).

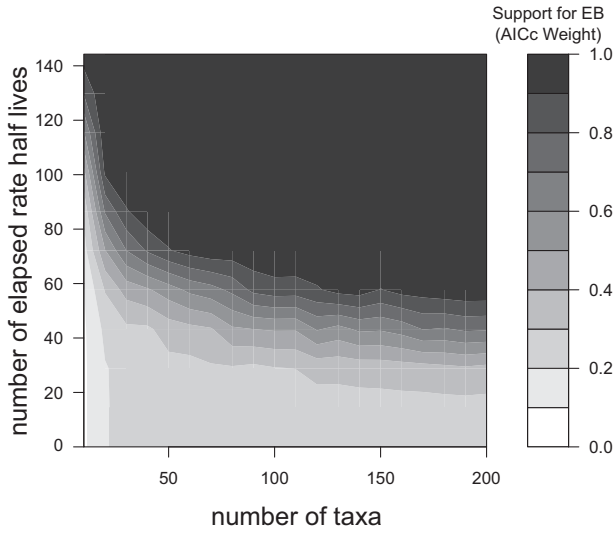

FIGURE 1. Power to detect Early Bursts (EBs) of trait evolution in comparative datasets. We simulated early bursts of trait evolution under a range of exponential decline parameters on 10,000 simulated phylogenies containing between 10 and 200 extant taxa (resulting in a 20×11 grid with an average of 446 simulations per point). We then fit Brownian Motion (BM) and EB models to the simulated data and compared model fit using Akaike Weights. Mean weights for the EB model are plotted here as a contour plot, with light colors indicating low support and dark colors indicating high support for EB. EBs can only be detected with strong support from large trees with rapidly declining rates (see text for explanation of rate half-life). Note that because BM is a special case of EB, weights are expected to be >0 when the number of elapsed rate half lives is zero.

As the statistician George Box famously remarked, "all models are wrong, but some are useful." The EB model as explicitly formulated is likely to be wrong (as are BM and OU). The question of whether it is useful for identifying early rapid evolution in comparative datasets remains to be answered.

Despite having great potential to aid in assessing model fit, predictive approaches have received little attention in the field of comparative methods in general, and for studying modes of trait evolution in particular (but see Boettiger et al. 2012). Posterior predictive approaches differ from traditional Bayesian methods in shifting the focus of model selection from choosing the model that maximizes the posterior probability of the observed data to choosing the model that best predicts the observed data via simulation (Kadane and Lazar 2004). Predictive approaches are best known to comparative biologists through the pioneering work of Nielsen (2002), whose innovative stochastic character mapping approach allowed for probabilistic inference of the locations of character state transitions on phylogenies (also see Huelsenbeck et al. 2003; Bollback 2006; Minin and Suchard 2008; Revell 2013). Simulation-based approaches have also been used to detect rate shifts (Sidlauskas 2007; Slater et al. 2012b), non-Brownian modes of evolution (Kutsukake and Innan 2013) and to assess the adequacy of models of diversification (Rabosky 2009; Rabosky et al. 2012). Related methods have also been applied to selecting models of sequence

evolution for inferring phylogenetic trees (e.g., Brown and ElDabaje 2009; Lewis et al. 2014). These results raise the possibility that posterior predictive approaches might not only provide comparative biologists with a measure of model adequacy, but also with a powerful tool for comparing the fit of different models to trait data, particularly in cases where traditional likelihood approaches are known to have low power.

In this article, we develop two posterior predictive approaches for detecting early bursts of trait evolution based on existing comparative methods. We then use simulations to evaluate their power relative to a more commonly used maximum likelihood approach. Perhaps surprisingly, we find that the maximum likelihood approach has somewhat lower power to detect early bursts than the two predictive approaches, particularly when the decline in rate is not incredibly strong. This difference is greatly exacerbated when even a moderate number of 'outlier' taxa (lineages that do not fit the ancestral evolutionary model) are introduced into the clade. We present an additional modification to one of our predictive approaches that makes use of a robust regression procedure to reduce the influence of such outliers, and show that this improves our power to detect early bursts of trait evolution in simulated datasets. Finally, we demonstrate the utility of these approaches using a dataset of cetacean body lengths (Slater et al. 2010), and show that our robust regression procedure provides little support for two alternative models of trait evolution while favoring the early burst hypothesis. Although we cannot claim anything from this study regarding the generality of the early burst pattern in nature, we do suggest that the evidence refuting its frequency is still equivocal and requires further attention.

Materials and Methods

Methods for Detecting Early Bursts

We begin by briefly reviewing the three commonly used approaches for detecting early bursts of trait evolution using phylogenetic comparative data: disparity through time analysis (Harmon et al. 2003), the node height test (Freckleton and Harvey 2006), and maximum likelihood (Harmon et al. 2010). Disparity Through Time analysis (henceforth DTT; Harmon et al. 2003) is similar in spirit to approaches used by paleobiologists (e.g., Foote 1994) in that the method uses the average pairwise Euclidean distance between species as a measure of disparity. Disparity is first computed for the entire clade, and subsequently for each subclade in the phylogeny. Subclade disparities are then standardized by total clade disparity to produce a measure of relative subclade disparity. To compute disparity through time, we move from root to tip, computing the average relative disparity at each node as the mean relative disparity of all clades with ancestral lineages alive at that time. Low values of average subclade disparity indicate that, on average, individual clades alive at that time contain a small amount of total morphological variation, relative to the entire clade. Under the early burst scenario, average subclade disparity is expected to decline rapidly during the early part of clade history as lineages rapidly diverge into distinct adaptive zones. Later, disparity is expected to level off as divergence slows and subclades diversify within adaptive zones, leading to low disparity within deeply divergent subclades, relative to total clade disparity. Harmon et al. (2003) introduced the Morphological Disparity Index (MDI) as a means of testing whether the observed disparity through time curve differs from the expectation of a time-homogeneous Brownian motion process. Briefly, a large number of datasets are simulated under a BM model using the maximum likelihood estimate of the Brownian rate parameter and disparity through time curves computed for each simulation. MDI is then computed as the area between the average curve from the simulated data and the curve for the observed data; this is in essence a form of parametric bootstrapping. A negative MDI, meaning that the majority of the area between the curves falls below the curve from the simulated data, indicates that subclades contain lower levels of disparity than predicted under a constant rates process, and supports the notion of a slowdown in rate. Slater et al. (2010) presented a small modification of this approach wherein the area between the observed disparity curve and all simulated curves are computed and the proportion of cases in which the area between the curves ≤ 0, that is the proportion of cases in which the curve for the observed data falls below the simulated data, is taken as a probability that the MDI is significantly more negative than expected under a time-homogeneous BM model.

The node height test (Freckleton and Harvey 2006) uses the relationship between the absolute magnitude of standardized independent contrasts (Felsenstein 1985) and the height above the root of the node at which they were computed to identify early bursts of trait evolution. A significant, negative relationship between the two (i.e., larger contrasts occurring deeper in the tree) indicates that the rate of trait evolution had slowed through time. This method is similar to approaches used to evaluate the fit of a Brownian Motion model to trait data before performing a contrasts (Felsenstein 1985) or phylogenetic generalized least squares (Grafen 1989) analysis to test for an evolutionary correlation between quantitative characters. In the context of the node height test however, it is used to reject the null hypothesis of rate constancy through time, rather than to justify its use as an appropriate assumption (Freckleton and Harvey 2006). Freckleton and Harvey (2006) suggested using a randomization procedure to assess the significance of a node height test slope. To our knowledge, no simulation-based procedure has been utilized to evaluate whether the slope of a node-height test differs significantly from a null expectation.

The last approach is to use maximum likelihood (ML) to fit a model in which the rate of trait evolution is

permitted to decrease through time (an Early Burst model; hereafter EB) and to compare this to a model of a trait evolving under a time-homogeneous BM using some model selection criterion such as a Likelihood Ratio Test (LRT), Akaike Information Criteria (AIC: Akaike 1974) or Bayesian Information Criteria (BIC: Schwarz 1978). BM and EB are conceptually very similar in that they both assume that the trait values at the tips come from a multivariate normal distribution with the expected value for all tips equal to the state at the root of the tree. In fact all current univariate models of continuous trait evolution, including the OU process (Hansen 1997; Butler and King 2004; Beaulieu et al. 2012) and the commonly used "Pagel" models, (λ, κ, δ; Pagel 1997, 1999), can be generalized in the same form. Following the notation of O'Meara et al. (2006), we denote $\mathbf{X}$ to be a column-vector of tip values and $\mathbf{C}$ to be a $N \times N$ matrix (where N is the number of tips) describing the phylogeny in terms of branch length such that $C_{i,j}$ is the distance from the root node to the most recent common ancestor of tips i and j. The (log) likelihood $\mathcal{L}$ of the model can therefore be written as:

$$\log(\mathcal{L}) = \log\left[\frac{\exp\left[-\frac{1}{2}[\mathbf{X}-\mathrm{E}(\mathbf{X})]'\mathbf{V}^{-1}[\mathbf{X}-\mathrm{E}(\mathbf{X})]\right]}{\sqrt{(2\pi)^N \times \det(\mathbf{V})}}\right], \quad (1)$$

where $\mathbf{V}$ is the Variance-Covariance (VCV) matrix for the tips under a given model. For example, under BM, the elements of $\mathbf{V}$ are given by

$$V_{ij} = \sigma^2 C_{i,j}, \quad (2)$$

meaning that the rate of evolution is constant across the clade and the variance accumulates proportional to the branch lengths. Under the "Accelerating Change/Decelerating Change" (AC/DC) model (Blomberg et al. 2003), the rate of trait evolution is permitted to accelerate or decelerate exponentially as a function of time. Note that we could also formulate a "linear change" model where the rate increases or decreases linearly with time. We limit our focus here to the exponential change model as it better fits the Simpsonian view of adaptive radiation as rapid early phenotypic diversification that subsequently slows as niches within adaptive zones become saturated. We follow Harmon et al. (2010) and constrain the model such that the rate of change is only allowed to decrease towards the present and refer to this constrained form as the EB Model. If the starting rate of evolution is σ_0^2 and the parameter describing the rate at which σ^2 changes is denoted as r then:

$$V_{ij} = \int_0^{C_{i,j}} \sigma_0^2 e^{rt}\, dt = \sigma_0^2\left(\frac{e^{rC_{i,j}}-1}{r}\right) \quad (3)$$

Note that if $r = 0$, this model reduces to BM as σ^2 remains constant throughout the clade's history.

Posterior Predictive Approach

An alternative way to assess the fit of BM and EB models to comparative data is to simulate under both models and compare how well each does at predicting the observed distribution of trait values. If the observed data truly evolved under an EB-like process, then simulating under BM should do a very poor job of predicting them. Simulating under EB, on the other hand, should result in trait values that are very similarly distributed. Both DTT and the node height test can readily be accommodated in such a posterior predictive framework as both produce a single summary statistic describing how phenotypic disparity changes through time. By generating a large number of summary statistics from data simulated under each process, we can compare our observed summary statistics to these predictive distributions and generate posterior predictive P values for each. In the case of an EB-like process, data simulated under BM should result in low posterior predictive P values (< 0.05), while simulations under EB should result in P values of ≈ 0.5.

We wrote a simple Markov chain Monte Carlo (MCMC) algorithm to sample model parameters for BM and EB from their posterior distributions, and to generate simulated datasets under the sampled values (Fig. 2). Our sampler followed other recent MCMC implementations applied to comparative datasets (e.g., Revell et al. 2012; Slater et al. 2012b), so we will not give full details here. R code to perform the analyses has been deposited at http://datadryad.org (doi:10.5061/dryad.6m2q0) and is incorporated in a forthcoming release of the `geiger` package (Harmon et al. 2008; Eastman, J., Pennell, M., Slater, G., Brown, J., FitzJohn, R., Alfaro, M., Harmon, L., Unpublished data) for R (R Development Core Team 2013). Briefly, at each step in the MCMC we proposed a new value for the model parameters, the Brownian rate (σ^2) and, in the case of the EB model, the exponential change parameter (r). We used the restricted ML algorithm of Felsenstein (1973) to evaluate the likelihood, following equation 2.5 in Freckleton and Jetz (2009), and accepted or rejected proposals based on the standard Metropolis–Hastings acceptance ratio. We set a flat improper prior of $\mathcal{U}[-\infty, \infty]$ on $\log(\sigma^2)$. Large negative values for the exponential change parameter of the EB model can cause singularity issues when attempting to invert the transformed phylogenetic variance-covariance matrix. We therefore used a lower bound on the uniform prior on this parameter that was computed from the height, T, of the tree:

$$r_{min} = \frac{\log(\sigma_T^2)}{t_0}, \quad (4)$$

where σ_T^2 is the end rate of evolution expressed as a proportion of the initial rate and t_0 is the root age of the tree. We used an arbitrary value of $\sigma_T^2 = 10^{-5}$ to define our lower bound on r.

Our MCMC differed in one significant respect from other implementations for comparative data.

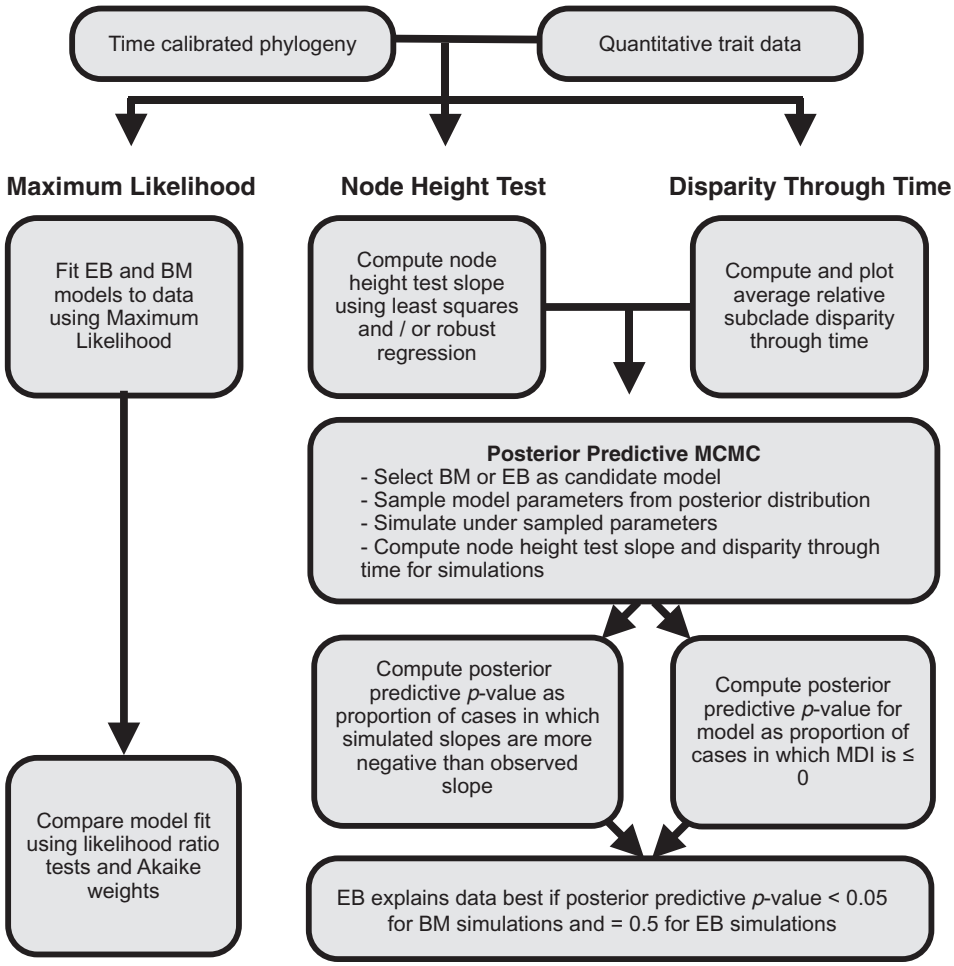

FIGURE 2. Flow chart showing the steps involved in testing for early bursts of trait evolution in our simulated datasets. See Materials and Methods for details on each analysis.

In addition to sampling parameter values from the chain at each sampling step, we also generated a simulated dataset under the fitted model using the current model parameters (Fig. 2). We then computed the slope for the node height test performed on the simulated data and MDI between the observed and simulated data, recording these values to our output file along with sampled parameters. We used the natural logarithm of the absolute value of standardized contrasts in our node height computations, rather than untransformed absolute values of standardized contrasts as in Freckleton and Harvey (2006) to ensure that we were testing for exponential declines in rate. At the end of each MCMC run, we obtained a sample of parameter values from the posterior distribution of the fitted model, as well as posterior predictive distributions for the node height test slope and MDI. We subsequently computed a posterior predictive *P* value for MDI by computing the quantile of the posterior predictive sample containing zero (that is, the quantile at which there is no difference between the observed and simulated data). We obtained the equivalent probability for the node height test by computing the proportion of cases in which the posterior predictive samples for the node height slope were less than or equal to the slope for the observed data. Computing a posterior predictive *P* value in this way does not allow one to select a model as the best fitting in the way that AIC scores might. Rather expected values lying in the tails of the predictive distributions indicate poor predictive ability of the fitted model, while values falling close to the center of the distribution indicate that the model is better at predicting the observed data.

Comparison of Method Performance

We performed a series of simulations to compare the power of our posterior predictive formulations of DTT and the node height test to that of the Maximum likelihood EB model. We did not compare our posterior predictive versions to standard formulations of DTT and the node height test as the latter provide only a measure of whether or not an observed dataset fits with predictions of the EB model, rather than allowing comparisons of model adequacy. For each simulation, we generated a phylogenetic tree under a birth–death process with a speciation rate $\lambda = 0.1$ and an extinction rate $\mu = 0.09$ using the R package `TreeSim`

(Stadler 2011). We conditioned simulations on 100 extant taxa, and subsequently rescaled the total depth of all trees to 10 time units. Preliminary simulations (data not shown) using a range of λ and μ values did not change our results and so we use the same parameters throughout this paper for the sake of consistency. We then simulated trait evolution under both BM and EB models. For BM, variation in the rate of character evolution should not affect model fitting, as rate variation has no effect on phylogenetic signal in trait data (Revell et al. 2008). Nevertheless, we confirmed this by simulating with σ^2 drawn from a range comprising (0.001, 0.1, 1, 10). For EB, choosing a meaningful range for r is difficult as the impact of variation in this parameter on the distribution of trait values and, subsequently, on our ability to detect the EB model is explicitly dependent on the age of the clade being examined; for a given value of r, as the age of a clade increases, the rate has more time to decay from its initial value and the resulting impact on the distribution of trait values among and within clades becomes more marked. For comparative purposes, it can be more useful to think of r in terms of its impact on the half-life of the rate, $t_{1/2}$ rather than in terms of the absolute magnitude of the parameter itself. For an exponentially decaying process occurring at rate r, the half-life is given by

$$t_{1/2} = \frac{log(2)}{r}. \quad (5)$$

We selected a range of r values that resulted in between 1 and 10 half-lives elapsing over the 10-million-year history of our simulated clades. Thus, at the lower limit of sampled values, an r corresponding to one half life will result in the rate decreasing to one half of its initial value over the history of the clade. Conversely, an r corresponding to 10 elapsed half-lives results in the rate decreasing by half after only one time unit, and ending at one-tenth of one percent of its initial value.

We fit both BM and EB models to each simulated dataset using our posterior predictive MCMC. After performing some initial runs to assess the trade-off between run time and sampling efficiency, and to determine the typical number of generations required to acheive convergence, we decided to run each chain for 10,000 generations, sampling every 10 generations and with the first 20 samples discarded as burn-in. We then assessed the fit of the fitted model to our simulated datasets using posterior predictive P-values for MDI and the node height test. To compare our posterior predictive approach to the more commonly used ML procedure, we also fit BM and EB models to each simulated dataset using the `fitContinuous` function in the `geiger` package (Harmon et al. 2008), and assessed model fit using likelihood ratio tests and small sample corrected Akaike weights (Burnham and Anderson 2002).

Outlier Taxa and Robust Regression

The EB model assumes that all taxa in the phylogeny evolve under the EB process. In reality this is unlikely to be the case; extinction, migration, and competition can create novel ecological opportunities that may lead one or more lineages to increase their rate of evolution or to jump to a new adaptive zone. We investigated the impact of these outlier taxa using simulations and attempted to account for their influence by using a modification to the node height test.

To test the robustness of the three approaches to "outlier" taxa we conducted an additional set of analyses in which we attempted to simulate convergent evolution under a single OU-like process. Again setting $\lambda = 0.1$ and $\mu = 0.09$, we simulated phylogenetic trees of 100 taxa under a birth–death model of diversification, and rescaled the trees to a root height of 10 time units. We then simulated an early burst of trait evolution over the whole tree under a strong burst that should be detectable by all methods by setting $\sigma_0^2 = 1$ and $t_{1/2} = 1.4$ ($r = -0.5$). We added outlier taxa in two ways: 1) such that they were randomly distributed throughout the phylogeny, and 2) such that they were phylogenetically clustered. The first scenario appoximates the case where there has been one or more independent occurrences of an uptake in evolutionary rate in the group, as might happen if lineages from multiple clades attempt to fill a recently vacated niche. The second scenario represents the case where an entire clade "escapes" from an ancestral adaptive zone and radiates into a new niche(s) (Etienne and Haegeman 2012).

For the phylogenetically random convergence case we randomly selected n terminal branches, where $n \in \{0, 1, 2, \ldots, 5, 10, 15, 20, 30, \ldots, 50\}$. We assumed for the purposes of simulation that these taxa "jumped" to a new adaptive zone with its own optimal trait value. To achieve this, we replaced the trait values for the convergent tips with values drawn from a narrow (sd $= 0.01$) normal distribution. To ensure that the new trait values were reasonable with respect to the underlying clade, we set the mean of the normal distribution from which we drew new values equal to 80th percentile value from the original data. This decision was arbitrary but motivated by biological realism—we suspect that for most cases in which outlier taxa are likely to cloud an underlying early burst, the outliers are most likely to shift to niches previously occupied by members of their clade, and thus jump to to new trait values from within the range of observed values. For the phylogenetically clustered outliers we again selected n tips to replace but drew them such they formed a monophyletic clade. We then simulated a new set of traits for this entire clade under a BM model with $\sigma^2 = 0.05$ and a mean again equal to the 80th percentile value of the original data.

In an attempt to improve on our ability to detect early bursts in the presence of outlier taxa, we made one modification to to our posterior predictive MCMC. In addition to using ordinary least squares regression to test for a relationship between contrast size and node height in the node height test, we also used a robust regression. Robust regression approaches are designed to reduce the influence of outlying data points by identifying and down-weighting them, rather than

removing them entirely. A robust regression procedure has been used for this very reason by Gingerich (1993) when estimating changes in evolutionary rate relative to sampling interval length from paleontological data. We used a ML-type (M-estimation) robust regression procedure (Huber 1973) implemented using the `rlm` function in the R package `MASS` (Venables and Ripley 2002). M-estimation procedures generate weights for each observation that are used to reduce the influence of outlier data on computation of the slope. Weights are estimated using an iteratively reweighted least squares procedure (Holland and Welsch 1977), optimizing the equation

$$\sum_{i=1}^{n} w_i(y_i - x'b)x_i' = 0,$$

where w_i is the weight and $(y_i - x'b)x_i$ is the residual of the i-th observation. We used the Huber weighting scheme (Huber 1973), which applies weights of 1 to observations that do not deviate from the model predicted value (i.e., those with residuals $\approx$ 0). Observations with larger residuals are down-weighted by applying weights < 1, with smaller weight values applied to larger residuals. We compared the performance of the node height test using both forms of regression, in both cases with null distributions for their slopes obtained via posterior predictive simulation.

Analysis of Cetacean Body Size Evolution

Extant cetaceans span a remarkable range of body sizes, from the 1.5 meter, 150 kg vaquita *Phocoena sinus* (Hohn et al. 1996) to the largest animal to have ever lived, the 24m, 190,000 kg blue whale *Balaenoptera musculus* (Nowak 1999). In a study of tempo and mode in cetacean diversification, Slater et al. (2010) found support for an early burst of body size evolution in living cetaceans using DTT and the node height test. They hypothesized that this pattern was driven by niche partitioning, primarily along dietary axes, during the radiation of modern cetaceans. However, they noted that MDI, although negative, did not quite meet significance (MDI = −0.17, P = 0.054) and the node height test only produced a significant result after removal of two young delphinid lineages (the orca *Orcinus orca*, and pilot whales *Globicephala* spp.) that appeared to have evolved large body size in response to recent shifts in diet and foraging behavior (Slater et al. 2010). In a later study looking at rates of body size evolution across all mammals Venditti et al. (2011), found no evidence for an early burst of evolution in cetaceans. Instead, they argued that the finding of Slater et al. (2010) was an artefact of rapid evolution in the lineage leading to Mysticeti, which includes the large baleen whales.

We used our posterior predictive approach to assess three models of body size evolution in living cetaceans. We first fit a single-rate, time-homogeneous BM model and an EB model, as in our simulation tests. We also fit a model in which body size evolved at a constant rate along all branches of cetacean phylogeny but with an elevated rate permitted in the stem lineage leading to crown mysticetes (henceforth the "mysticete-shift" model; Venditti et al. 2011). In practice, this was accomplished by fitting a branch length scalar, ϕ, to the branch leading to mysticetes. We performed two runs of 1 million generations for each model, sampling from the chain every 100 steps. After visually checking that both runs converged on the target distribution using the `Tracer` software (Rambaut and Drummond 2007), we conservatively discarded the first 20% of samples as burn-in. We compared the predictive performance the three models by computing posterior predictive P values from the post burn-in predictive distributions of MDI and from least squares and robust regression node height slopes for each model. To contrast the results of posterior predictive simulations with inferences obtained from model fitting approaches, we also used ML to fit the BM, EB and mysticete-shift models, using a modified version of the `fitContinuous()` function (Harmon et al. 2008). This function, along with code and data used to perform the analyses has been deposited on Dryad.

Results

Comparison of Method Performance

Disparity through time, the node height test, and maximum likelihood all exhibited low power to detect early bursts of trait evolution when few half lives had elapsed over the history of the simulated clades (Fig. 3a). All three approaches required >6 half-lives to have passed before the EB model was preferred on average with strong support (>0.95; Fig. 3a, b). Support for EB over BM increased with the number of elapsed half lives under both likelihood approaches, but the increase was more rapid for the likelihood ratio test (Fig. 3a). Mean support for EB based on Akaike Weights was lower than 0.5 when as many as 3 half-lives had elapsed. The two posterior predictive approaches performed reasonably. For EB with $t_{1/2} = \infty$ (i.e., BM) mean posterior predictive P values for both were 0.5, as expected for a nested null model, and as the number of elapsed half-lives increased support for BM decreased (Table 1 and Fig. 3a). At faster rate declines (number of elapsed half lives ≥ 9), all approaches performed comparably, although power for the MDI approach was lower than for the node height or likelihood ratio tests. Posterior predictive P values derived using simulation from the posterior distribution of parameters for the EB model followed expectation (Fig. 3c), although weak bursts resulted in posterior predictive P values that were lower than the expected value of 0.5; this can be explained by an upper bound of 0 on our uniform prior for r. Mean P values rapidly converge on 0.5 as the number of elapsed half-lives increases.

When BM was the generating model, variation in σ^2 had no effect on our ability to select among models, as expected and model selection perfomance was appropriate. Full details of these results are provided in the Supplementary Materials.

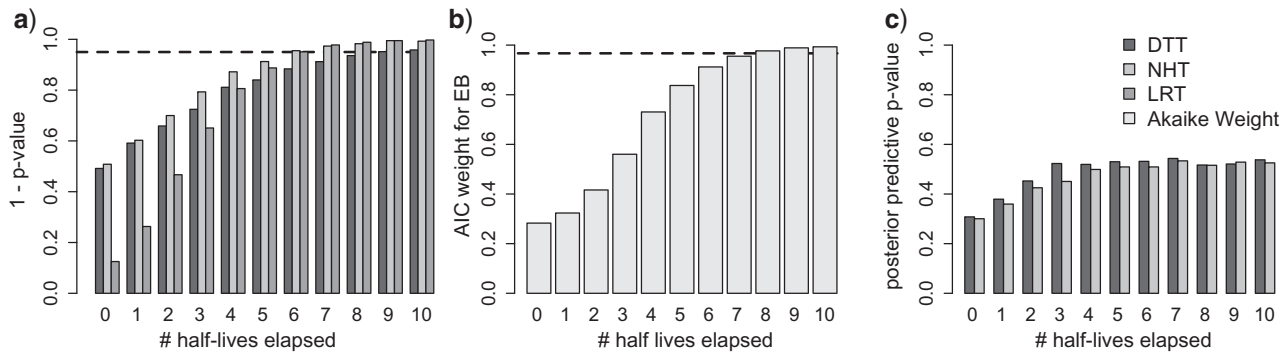

FIGURE 3. Distinguishing early bursts of trait evolution from Brownian motion in simulated datasets. Support for EB is shown for elapsed rate half-lives (see text) ranging from 0 (i.e., BM) to 10. Support for EB is derived from a) mean posterior predictive *P*-values for Disparity Through Time and the Node Height Test when fitting and simulating under a BM model, and *P*-values from likelihood ratio tests comparing the fit of EB and BM models, and; b) median Akaike weights. Quantiles of predictive distributions derived under an EB model that contain the expected summary statistic value are shown in c). Here, values close to 0.5 indicate a close correspondence between simulated and observed data. Note that a) shows 1-*P* values and thus values ≥ 0.95 indicate low support for BM in favor of EB. Values are shown in this way to facilitate visual comparison with support under Akaike weights.

TABLE 1. Power (1 – Type II error) to detect Early Bursts of trait evolution under different burst strengths for the four frequentist approaches

Test	Number of rate half-lives elapsed										
	0	1	2	3	4	5	6	7	8	9	10
Morphological disparity index	0.05	0.11	0.19	0.26	0.43	0.49	0.61	0.70	0.76	0.81	0.84
Node height test	0.06	0.11	0.18	0.36	0.56	0.70	0.82	0.90	0.93	0.98	0.98
Likelihood ratio test	0.00	0.04	0.15	0.35	0.59	0.73	0.88	0.95	0.98	0.99	1.00
Node height test (robust regression)	0.06	0.12	0.21	0.42	0.64	0.78	0.88	0.93	0.95	0.99	0.99

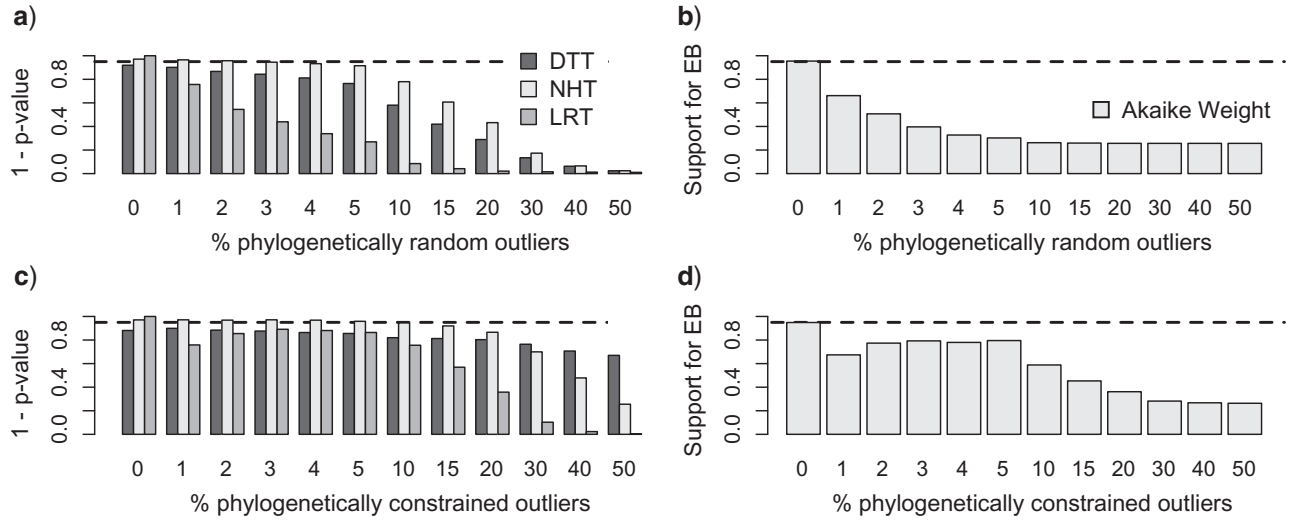

FIGURE 4. Detecting early bursts in the face of; a, b) phylogenetically random and, c, d) phylogenetically clustered convergence. a & c) show support for EB using posterior predictive approaches and likelihood ratio tests (mean *P*-values), while b & d) show support for EB derived from Akaike weights. Node height tests were performed using ordinary least squares regressions.

Outlier Taxa and Robust Regression

Adding convergent taxa had a substantial impact on our ability to detect an underlying early burst of trait evolution. When convergent taxa were added randomly through the phylogeny, Akaike weights for the EB model were low, even when only one or two convergent taxa were added (Fig. 4b), even though the underlying burst was of a strength that should be detectable. The likelihood ratio test performed slightly better, but posterior predictive *P* values for MDI and, in particular, the node height test performed much better (Fig. 4a). We were on average able to strongly discount the null hypothesis of a BM process with up to 5% of taxa exhibiting phylogenetically random convergence using the posterior predictive node height

TABLE 2. Power for the two basic posterior predictive approaches, the likelihood ratio test, and the posterior predictive node height test with robust regression in cases with convergent taxa

Type of convergence	Test	Number of convergent taxa											
		0	1	2	3	4	5	10	15	20	30	40	50
Random Convergence	Morphological disparity index	0.7	0.66	0.56	0.46	0.4	0.29	0.09	0.03	0.01	0.01	0	0
	Node height test	0.88	0.89	0.85	0.79	0.74	0.7	0.37	0.17	0.06	0.01	0	0
	Likelihood ratio test	1	0.75	0.54	0.44	0.34	0.27	0.08	0.04	0.02	0.01	0.01	0.01
	Node height test (robust regression)	0.93	0.93	0.92	0.89	0.88	0.87	0.64	0.39	0.19	0.02	0	0
Phylogenetically Clustered Convergence	Morphological disparity index	0.62	0.64	0.59	0.56	0.56	0.53	0.48	0.46	0.43	0.42	0.37	0.33
	Node height test	0.88	0.88	0.87	0.88	0.88	0.86	0.78	0.73	0.58	0.27	0.09	0
	Likelihood ratio test	1	0.76	0.85	0.89	0.88	0.86	0.75	0.57	0.35	0.1	0.02	0
	Node height test (robust regression)	0.93	0.93	0.92	0.93	0.92	0.9	0.89	0.85	0.75	0.49	0.2	0.02

test. These qualitative observations are confirmed by computation of power of the three approaches (Table 2). All three performed better for phylogenetically clustered convergence. On average, an underlying strong burst could be detected on the basis of Akaike weights with convergent clades of up to five taxa added and we could strongly discount the null BM model using MDI with clades of up to 10% convergence (Fig. 4d). Again, the posterior predictive node height test out-performed the other approaches (Table 2), on average providing poor support for the null with convergent clades containing up to 20% of the total number of taxa added to our datasets (Fig. 4c).

The use of robust regression in place of ordinary least squares regression in the node height test resulted in a slight, but notable improvement in power (Table 2) and decrease in support the null model for both types of convergence. Using robust regression, we found poor support for the null BM model with up to 10% phylogenetically random convergent taxa (Fig. 5a). For ordinary least squares regression this behavior dropped to tolerating only 5% convergent taxa. The greatest benefits again occurred for phylogenetically clustered convergence. Here, we found, on average, poor support for BM with up to 30% convergent taxa. For OLS, this figure dropped to ≈20 convergent taxa (Fig. 5b).

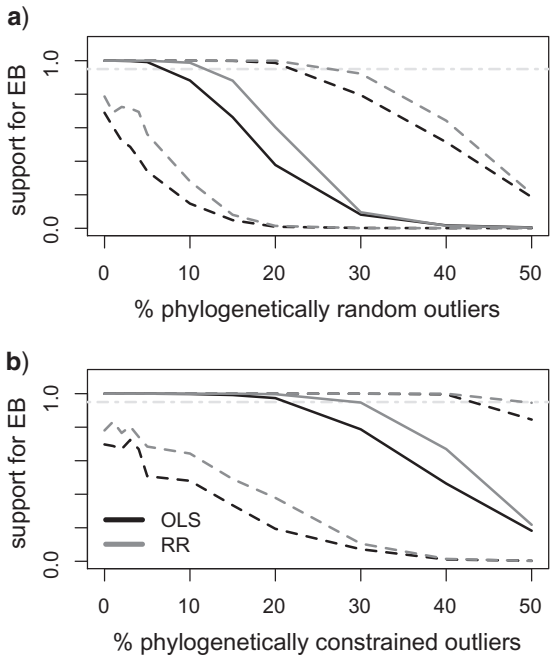

FIGURE 5. Comparison of median performance of the node height test using least squares and robust regression with a) phylogenetically random b) and phylogenetically clustered convergence. Dashed lines indicate 95% quantiles.

Analysis of Cetacean Body Size Evolution

We first used ML to fit three models to our cetacean body length data: a time-homogenous BM model, an EB model, and the mysticete-shift model that allowed for elevated rates in the stem mysticete lineage. ML parameter estimates for the early burst model (Table 3) suggest declining rates of body length evolution with an initial rate twice that estimated under BM and an exponential decline parameter of −0.023, equivalent to a rate half-life of 30.14 million years. For the mysticete-shift model, parameter estimates indicate that a rate increase of 12.5 × is required along the branch leading to crown mysticetes to explain the distribution of extant cetacean body sizes. However, Akaike weights demonstrate equivocal support for all three models (Table 3). Brownian motion is the best fitting model, but receives only 41% of the Akaike weight and, while the other two models receive slightly less weight, they cannot be ruled out. Although the EB model receives the lowest support among the candidate models, based on ML parameter estimates for this model and the crown age of cetaceans in our phylogeny, only 1.22 rate half-lives would have passed over the duration of cetacean evolution. Taken in conjunction with our simulation results (Fig. 1), this finding suggests that even if cetaceans did undergo an early burst of body size evolution, we should not expect to find support for this model using ML with a dataset comprising extant taxa only.

Posterior predictive simulations reveal a slightly clearer picture of body size evolution in cetaceans. Predictive distributions for MDI under all three candidate models are left-shifted (Fig. 6), indicating that the observed data show a greater degree of among-clade partitioning of phenotypic variation than is expected from the fitted models. This is particularly accentuated

TABLE 3. AICc scores and weights for the three candidate models fitted to cetacean body length data. The final two columns give ML estimates of model-specific parameters r (for EB) and the branch-specific rate scalar ϕ (Mysticete shift)

Model	AICc	ΔAICc	Weights	σ^2	r	ϕ
BM	50.51	0.00	0.41	0.012	—	—
Mysticete shift	50.75	0.24	0.37	0.011	—	12.544
EB	51.75	1.24	0.22	0.024	−0.023	—

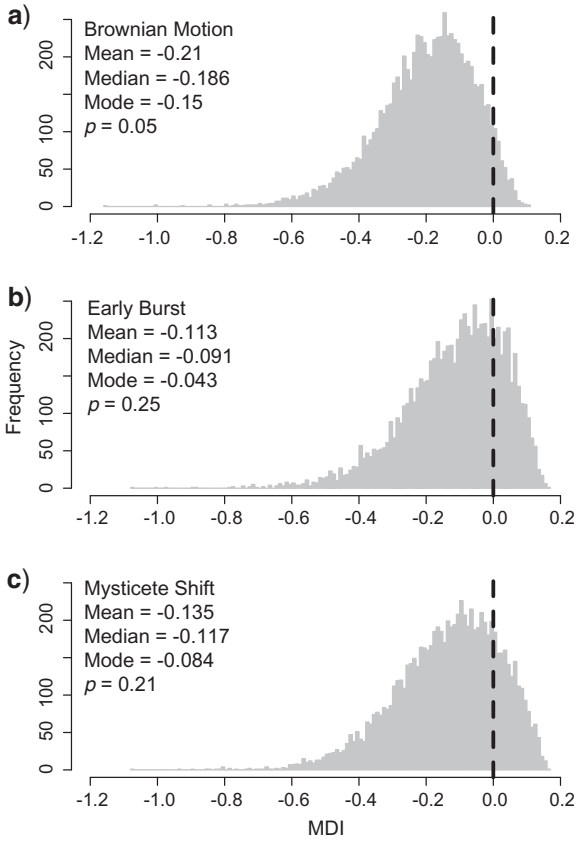

FIGURE 6. Posterior predictive distributions of MDI for cetacean body length data generated under a) BM b) EB, and c) Mysticete shift models. Dashed verticle lines indicate the expected value of 0.

for the time-homogeneous BM model (Fig. 6a), where the expected value of 0 falls at the 95th percentile of the predictive distribution. The EB and mysticete-shift simulations provide a closer fit, although the expected value of zero falls between the 75th and 80th percentiles for both. Visual inspection of the predictive distributions (Fig. 6b,c) further suggests that the mysticete-shift model and the EB model are a much closer fit to the data than the homogeneous BM process, an observation confirmed by location statistics for the three distributions (Fig. 6).

The node height test using ordinary least squares regression shows a negative relationship between the ln(absolute value of contrasts) and node heights, although this relationship is not significance at $\alpha = 0.05$ ($\beta_{ols} = -0.046, p = 0.051$). Posterior predictive distributions of node height test slopes support the results from ML and MDI; all models appear to be poor predictors of the relationship between contrast size and node height, which is more negative than expected from the candidate model pool (Fig. 7a–c and Table 4).

The use of a robust regression estimator of the slope alters this finding substantially. Huber weights for 12 contrasts were <1 (Table 5; recall that weights <1 imply larger residuals and greater influence on the OLS regression slope). Mapping these weights onto the cetacean phylogeny (Fig. 8) shows that the two outlier contrasts identified as influential by Slater et al. (2010) (those between *Orcinus orca* and *Orcaella brevirostris* node 128, and between *Feresa attenuata* and *Globicephala* spp. node 126) are indeed down-weighted in our analyses. However the most heavily downweighted constrasts are those between two pairs of recently diverged river dolphins (*Platinista* spp. node 82, and *Inia* spp. node 138) that do not differ in body length, and between *Tursiops aduncus* and a clade comprising *Stenella* and *Delphinus delphinus* (node 110). Several other contrasts, including several among species of beaked whales (Ziphiidae), are also more substantially down-weighted. Repeating the analyses with untransformed contrasts (i.e., a linear rate decline model, data not shown) confirms that the identification of additional influential contrasts here compared with here Slater et al. (2010) results from the use of log-transformed contrasts to test for exponential rate decay. Given that 9 out of 12 influential contrasts exhibit negative residuals, it is not surprising that the robust estimate of the slope is less negative than that obtained from ordinary least squares ($\beta_{robust} = -0.026$). Using our posterior predictive simulations in conjunction with the robust estimate of the node height test slope (Fig. 7d–f), we find that both BM and the mysticete-shift model exhibit low predictive power for the observed distribution of cetacean body lengths (Fig. 7d–f, $p_{BM} = 0.02, p_{shiftmodel} = 0.02$). However, the observed slope compares more favorably to the distribution of robust regression slopes for the EB model ($p_{EB} = 0.26$). Although the correspondence is not perfect (Fig. 7e), we conclude that EB appears to be the most predictive model, of those considered here, for the distribution of extant cetacean body lengths.

DISCUSSION

Evolutionary biologists have increasingly come to rely on ML or, more recently, Bayesian inference for assessing the fit of evolutionary models to quantitative trait data. The typical work-flow in a macroevolutionary study involves fitting a series of models using ML and then comparing the fit of the candidate pool to the data using likelihood ratio tests or information theoretic criteria (Mooers et al. 1999). If AIC is used, the model with the lowest AIC score or highest Akaike weight is then declared the winner, and inferences are subsequently drawn about the tempo and mode of trait evolution in the clade being investigated. These approaches have

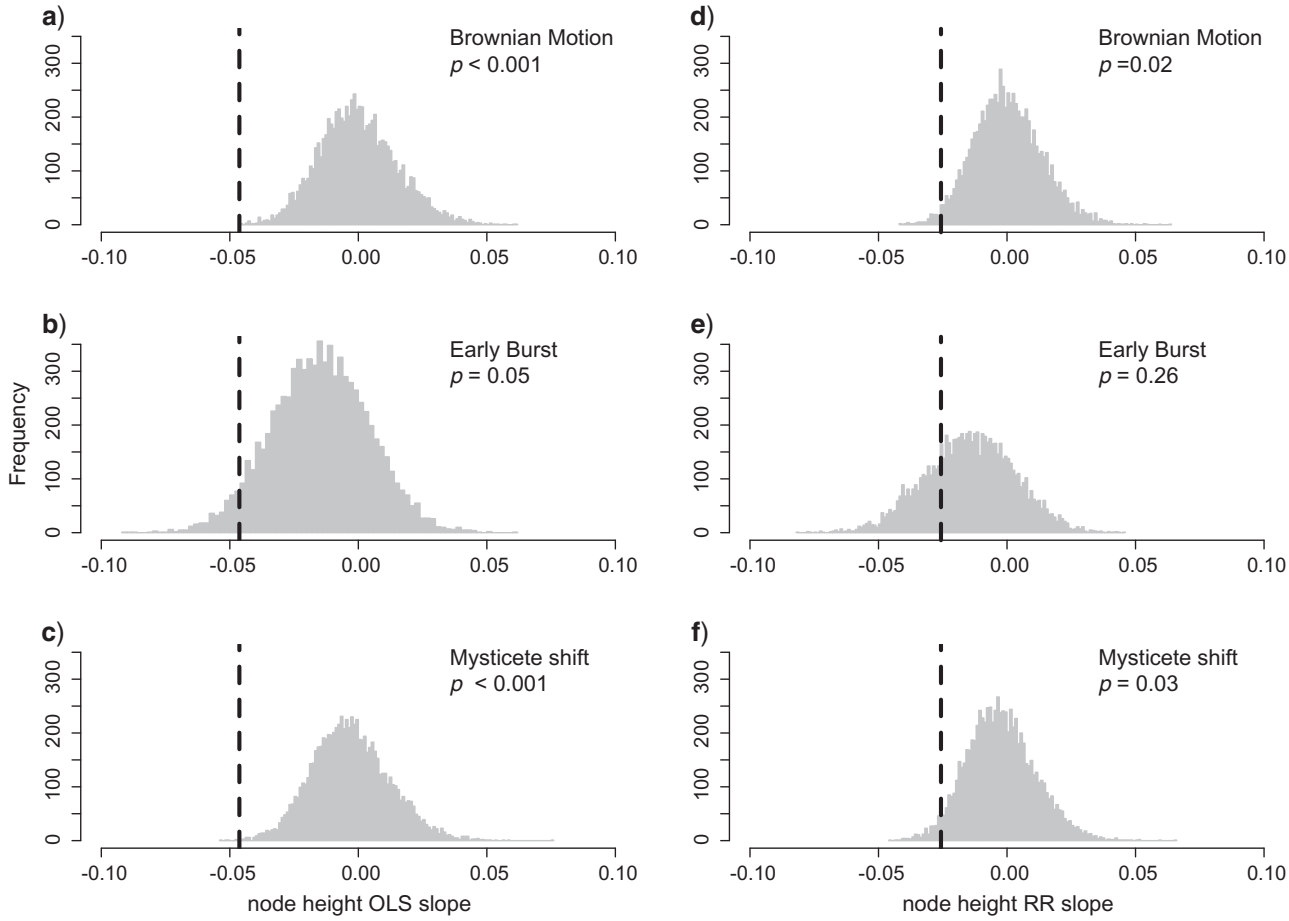

FIGURE 7. Posterior predictive distributions of node height test slopes using least squares and robust regression for cetacean body length data. Plots show a, d) BM, b, e) EB, and c, f) Mysticete-shift models. The dashed lines indicate the value of the observed slopes derived from the cetacean length data.

TABLE 4. Posterior predictive *P* values for the three candidate models fitted to cetacean body length data. *P* values are computed for the morphological disparity index, the node height test with ordinary least-squares regression and the node height test with robust regression

Model	MDI	Node height OLS	Node height RR
BM	0.05	0	0.02
EB	0.25	0.05	0.26
Mysticete shift	0.21	0	0.03

Note: Values ~0.5 indicate a good correspondence between the candidate model and the observed data, while tail-end values (<0.05, >0.95) indicate poor correspondence.

TABLE 5. Residuals and Huber weights for ln(cetacean body length contrasts) on node heights under robust regression

Node	Residuals	Weight
138	−8.473	0.161
82	−8.404	0.163
110	−3.833	0.357
93	−2.481	0.551
83	−2.194	0.624
86	−2.016	0.679
128	1.965	0.696
126	1.897	0.721
145	1.549	0.883
111	−1.510	0.906
95	−1.396	0.980
88	−1.380	0.991

many desirable properties. In particular, model selection using information theoretic approaches allows users to simultaneously compare a pool of candidate models rather than performing a series of pairwise comparisons. Furthermore, because informatic theoretic approaches do not require a model describing the null expectation, selection of a "best" model does not imply that model is correct but rather than it comes closest to describing the underlying data. However, these approaches, and indeed their advantages, also lead to limitations. For example, a given model might receive the highest weight of the candidate pool but simulating from it might fail to produce anything like the observed data. In this case, the winning model does not adequately describe the underlying evolutionary process in the clade being studied, but how are we to know this? Alternatively, we

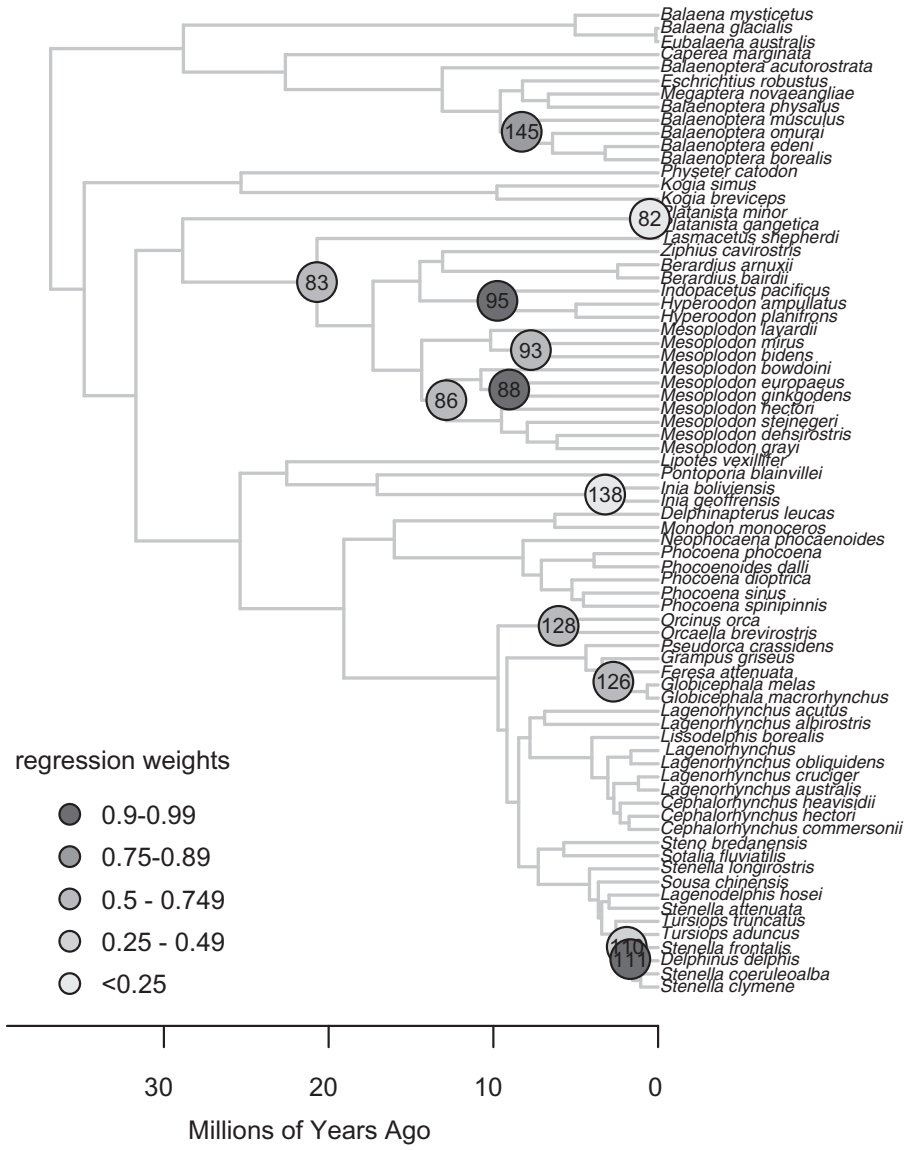

FIGURE 8. Cetacean phylogeny used in our analyses. Shaded node labels indicate nodes that were down-weighted in computation of the robust regression slope. Lighter shades indicate nodes that were more heavily down-weighted.

might find that we cannot differentiate among candidate models using information theoretic criteria. It is possible in this case however that our inability to select a winning model is not driven by the lack of a good model in our candidate pool, but rather by one or two data points that deviate from the ancestral pattern and cloud our ability to select the true model. Model fitting alone does not tell us however if this is the case.

In this article, we have demonstrated that posterior predictive approaches have the potential to greatly improve our ability to distinguish among modes of trait evolution in comparative data. We found that by using posterior predictive *P* values derived from the morphological disparity index (Harmon et al. 2003) and, in particular, the node height test slope (Freckleton and Harvey 2006), we recovered strong support for the EB model from datasets simulated under moderately strong early bursts that, in some cases, could not be detected using information theoretic approaches. This unexpected result is appealing for two main reasons. First, the "Early Burst" model is simply difficult to detect with phylogenetic comparative data (Harmon et al. 2010), and any method that increase our power to do so is helpful. Interestingly, the patterns we have found regarding the power to detect slowdowns in trait evolution are, in some ways, exactly the opposite to the findings of studies that have investigated slowdowns in lineage diversification rate (Liow et al. 2010; Pennell et al. 2012). For example, Liow et al. (2010) simulated phylogenies under time-varying diversification rates and found that the power to detect early bursts was greatest early on and that the signal was subsequently erased by extinction later in the clade's history (see also, Quental and Marshall 2010; Pennell et al. 2012). Here,

we demonstrated that under a model of declining trait evolution, we expect the signal to actually get stronger through time. This finding supports the notion that, if adaptive radiations are characterized by early bursts of both lineage and trait diversificaiton, the signal for the early burst may be better retained in trait data (Slater et al. 2010). Although Harmon et al. (2010) found little support for EB-like processes in their large number of comparative datasets, several authors have found support for decelerating rates of evolution in individual clades (Harmon et al. 2003; Burbrink and Pyron 2010; Mahler et al. 2010; Slater et al. 2010; Derryberry et al. 2011; Dornburg et al. 2011). It is notable that none of these studies used the ML formulation of the EB model to do so; in each case their findings were based on DTT, the node height test, or both. (Mahler et al. (2010) did use ML but not the standard EB model; we refer readers to their paper for details). It is possible that these methods are simply more forgiving of noisy data than the analytical approach of Harmon et al. (2010), although more work is clearly required to fully understand this phenomenon.

The second appealing aspect of posterior predictive approaches is that they provide a built-in check of model adequacy. By simultaneously fitting models and simulating from their parameter posterior distributions, we are not only able to compare model fit, but also to ask how close each comes to predicting the distribution of data observed in the focal clade. A word of caution is warranted here: The summary statistics that we used to compare EB and BM are useful for comparing time-homogeneous rate processes to those in which rates can vary through time because they use estimates of the rates along branches or the expected disparity resulting from the evolutionary process. The metrics are less suited to processes in which the expected value of a trait changes, such as an evolutionary trend (Alroy 1998; Hunt 2006, 2007) or a multi-peak OU process (Butler and King 2004; Beaulieu et al. 2012). Although this means that MDI and the node height test slope are unlikely to be useful for assessing the fit of these kinds of models, alternative summary statistics could easily be derived to do this.

To the best of our knowledge, ours is the first study investigating the influence of "outlier" taxa on model selection and fit. The implicit assumption in fitting and comparing models of trait evolution, that patterns should be homogeneous across the clade, is clearly unrealistic and it is alarming that even a small number of outlier taxa had such a strong effect on our ability to infer the true model. This is especially relevant for tests of early bursts in the context of adaptive radiations as there is no reason why we would expect *a priori* that an entire clade should show an early burst pattern, even under the simple scenario described by Simpson (1944). Some lineages may escape the ancestral adaptive zone, perhaps by moving to a new geographic region or evolving a novel key innovation (Etienne and Haegeman 2012). A number of approaches have been developed to investigate rate heterogeneity in rates of trait evolution across the tree (O'Meara et al. 2006; Thomas et al. 2006; Eastman et al. 2011; Revell et al. 2012; Slater et al. 2012b) but these have so far, been restricted to the case of multiple rate BM. We took a slightly different approach here. Rather than attempting to identify exceptional lineages, we sought to look for general patterns—that is to pull a broad signal out of the noise. As the node-height test involves fitting a linear-regression model to the data, it is natural to turn to established methods from the statistics literature, such as robust regression (Huber 1973), to down-weight the contribution of "outlier" taxa in the test. In applying our posterior predictive robust regression approach to the cetacean dataset, we were able to show that a constant rates process and a process allowing rapid evolution in the stem mysticete lineage were poor predictors of extant cetacean body lengths. The early burst model is a better, although not perfect fit to these data. Slater et al. (2010) recovered a similar result but were forced to arbitrarily remove two contrasts that they had visually identified as outliers in their node height test regression. Our use of a robust regression identified these outliers, as well as several additional influential contrasts, and allowed us to appropriately weight them when computation of the node height test slope. In the process, we removed the arbitrariness of visually identifying outliers and avoided removing them altogether.

It makes sense to ask whether the contrasts identified and downweighted by the robust regression procedure make biological sense. The two nodes identified in Slater et al. (2010) were recovered here as possessing large, positive residuals consistent with interpretation in that paper. The contrast between the blue whale *Balaenoptera musculus* and a clade of other roquals also generated a positive residual that was downweighted here, perhaps suggested accelerated evolution in the lineage leading to the largest animal to have ever lived. All other downweighted contrasts possessed negative residuals. The two most heavily downweighted nodes were for pairs of recently diverged sister species that do not differ in body length, while the other node to receive a weight of < 0.5 was a contrast between a clade of similarly sized oceanic dolphins. A large proportion of down-weighted contrasts (5 of 12) occured within the beaked whales. All were due to negative residuals, indicating that rates of morphological evolution may have been extremely low in this clade of deep-diving cetaceans. However, a series of diagnostic tests including qq- and residual versus fitted value plots (data not shown) suggests that only the three contrasts with the smallest weights deviate substantially from the fitted model. We strongly advocate the use of such diagnostics when performing robust regression versions of the node height test to ensure that down-weighted contrasts are appropriately identified and not overinterpreted. It is unclear to us why robust slope estimators are not more widely used in comparative biology (although for a discussion of implementation issues, see Stromberg 2004). The use of a robust regression procedure when computing node height test slopes (Freckleton and Harvey 2006) or evolutionary correlations from phylogenetically

independent contrasts (Felsenstein 1985) would seem equally sensible after performing appropriate diagnostic checks.

Although we believe posterior predictive approaches have great potential to improve the rigor with which model selection can be achieved, there are some limitations to the approaches that we describe in this article. First, although we use robust regression to reduce the influence of outlier taxa when computing a node height test slope, these outliers still influence the sampling of parameter values on which we base our posterior predictive simulation. We would like to be able to fit models using likelihood that can identify and down weight outliers, or mixed models that allow for rate shifts within broader early bursts. Addressing these problems deserves increased attention and it seems likely that the flexibility of reversible jump Markov chain Monte Carlo approaches (e.g., Eastman et al. 2011; Venditti et al. 2011; Revell et al. 2012) will be a promising avenue for exploration here. In the meanwhile, we argue that our robust regression approach is a viable alternative. It is tractable and allows us to capture what we are most interested in: broad-scale macroevolutionary patterns. In addition, robust regression weights provide information regarding evolutionary processes in specific regions of the phylogeny that can generate promising avenues for further investigation.

The use of posterior predictive P values also leads us to a model comparison approach that has more in common with frequentist tests than the more intuitive model selection provided by information theory. In our cetacean analysis, for example, we found low support for time-homogeneous BM and BM with a rate shift along the stem myticete lineage as models for the evolution of cetacean body size, based on posterior predictive P values. The EB model clearly provided a better fit to our cetacean body length data, although the predictive P value was far from indicating strong support for this model, and it is difficult to understand the relative support we should assign to each of the three candidate models. An additional point to consider here is that our posterior predictive P value criterion provided no penalties for model complexity, leading to the very real potential for over-fitting (Burnham and Anderson 2002). As predictive approaches become more common in macroevolutionary research, the use of alternative approaches for performing model selection from predictive distributions will become more necessary (Martini and Spezzaferri 1984; Gelfand and Ghosh 1998; Kadane and Lazar 2004; Lewis et al. 2014).

In the framing of this article, we follow the lead of many researchers before us (e.g., Harmon et al. 2003, 2010) in postulating that early bursts of trait evolution should be symptomatic of adaptive radiations and that the failure to find early bursts may therefore suggest the absense of "classical" adaptive radiations. We have argued throughout this paper that the failure to observe early bursts in comparative data sets may be due to limitations of current methods and have suggested ways in which these can be circumvented. However, it is also worth considering whether our conceptual models of adaptive radiations are perhaps overly naïve and whether the predictions they make are actually realistic. Morphological divergence may continue at a relatively constant rate throughout radiations or may not be tightly linked with lineage diversification (Pennell et al. 2013); there is some evidence from the fossil record to support this view (Foote 1993; Bapst et al. 2012). Furthermore, many of the reasons we expect to find early bursts are based on concepts in ecology (such as similarity limiting co-existence) that have long been questioned by ecologists (e.g., Abrams 1975, 1983; Chesson 2000; Siepielski and McPeek 2010; Mayfield and Levine 2010; HilleRisLambers et al. 2012). More rigorous (i.e., mathematical) and ecologically explicit models that predict patterns of trait evolution during adaptive radiations are definitely warranted (Pennell and Harmon 2013). Our intuition regarding evolutionary and ecological dynamics has often been shown to be incorrect and this may indeed be the case here. For example, Ingram et al. (2012) constructed a model where a food web evolved along a phylogeny and found that detection of an early burst pattern in body size evolution was negatively associated with the degree of omnivory—something that might not have been predicted from first principles. While it remains an open question as to whether early bursts are indeed more common than Harmon et al. (2010) suggest, an equally open question is whether we should expect early bursts at all. And if so, when do we expect to observe them and in what types of traits? We hope that theoretical developments addressing the latter questions will complement further analysis of empirical data.

Supplementary Material

Supplementary material, including data files, scripts, and online-only appendices, can be found at http://datadryad.org in the Dryad data repository at (doi:10.5061/dryad.6m2q0).

Funding

G.J.S. was supported in part by National Science Foundation grant DEB 0918748 to Michael Alfaro and Luke Harmon, and in part by a Peter Buck Smithsonian Institution post-doctoral fellowship. M.W.P. was funded by a National Evolutionary Synthesis Center Graduate Fellowship, a Bioinformatics and Computational Biology Graduate Fellowship from the University of Idaho, and a National Science Foundation grant DEB 1208912 to Luke Harmon.

Acknowledgments

We thank Jeremy Brown for organizing the symposium on predictive approaches in evolutionary biology at the 2012 Evolution meeting and for inviting

our contribution. We are grateful to Michael Alfaro, Luke Harmon, Arne Mooers, and Samantha Price for much discussion of ideas about tempo and mode of trait evolution, and to Gene Hunt, Jeremy Brown and two anonymous reviewers for comments and suggestions on a previous version of the manuscript. We especially thank Rich Fitzjohn for suggesting that robust regression might be useful in accounting for outlier taxa.

References

Abrams P. 1975. Limiting similarity and the form of the competition coefficient. Theor. Population Biol. 8:356–375.
Abrams P. 1983. The theory of limiting similarity. Annual Review of Ecology and Systematics 14:359–376.
Akaike H. 1974. A new look at the statistical model identification. IEEE T Automat. Contr. 19:716–723.
Alroy J. 1998. Cope's rule and the dynamics of body mass evolution in North American fossil mammals. Science 280:731–734.
Alroy J. 1999. The fossil record of North American mammals: Evidence for a Paleocene evolutionary radiation. Syst. Biol. 48:107–118.
Archibald J.D., Deutschman D.H. 2001. Quantitative analysis of the timing of the origin and diversification of extant placental orders. J. Mammal. Evol. 8:107–124.
Bapst D.W., Bullock P.C., Melchin M.J., Sheets H.D., Mitchell C.E. 2012. Graptoloid diversity and disparity became decoupled during the Ordovician mass extinction. Proc. Natl. Acad. Sci. 109:3428–3433.
Beaulieu J.M., Jhwueng D.-C., Boettiger C., O'Meara B.C. 2012. Modeling stabilizing selection: Expanding the Ornstein-Uhlenbeck model of adaptive evolution. Evolution 66:2369–2383.
Bianucci G., Landini W. 2006. Killer sperm whale: a new basal physeteroid (Mammalia, Cetacea) from the Late Miocene of Italy. Zool. J. Linnean Soc. 148:103–131.
Blomberg S.P., Garland T., Ives A.R. 2003. Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. Evolution 57:717–745.
Boettiger C., Coop G., Ralph P. 2012. Is your phylogeny informative? measuring the power of comparative methods. Evolution 66:2240–2251.
Bollback J.P. 2006. Simmap: Stochastic character mapping of discrete traits on phylogenies. BMC Bioinform. 7:88.
Brown J.M., ElDabaje R. 2009. Puma: Bayesian analysis of partitioned (and unpartitioned) model adequacy. Bioinformatics 25:537–538.
Burbrink F.T., Pyron R.A. 2010. How does ecological opportunity influence rates of speciation, extinction, and morphological diversification in New World ratsnakes (tribe Lampropeltini)? Evolution 64:934–943.
Burnham K.P., Anderson D.R. 2002. Model selection and multimodel inference. New York, USA: Springer.
Butler M.A., King A.A. 2004. Phylogenetic comparative analysis: A modeling approach for adaptive evolution. Amer. Nat. 164:683–695.
Chesson P. 2000. Mechanisms of maintenance of species diversity. Ann. Rev. Ecol. Syst. 31:343–366.
Derryberry E.P., Claramunt S., Derryberry G., Chesser R.T., Cracraft J., Aleixo A., Pérez-Emán J., Remsen Jr. J.V., Brumfield R.T. 2011. Lineage diversification and morphological evolution in a large-scale continental radiation: The neotropical ovenbirds and woodcreepers (Aves: Furnariidae). Evolution 65:2973–2986.
Dornburg A., Sidlauskas B., Santini F., Sorenson L., Near T.J., Alfaro M.E. 2011. The influence of an innovative locomotor strategy on the phenotypic diversification of triggerfish (family: Balistidae). Evolution 65:1912–1926.
Eastman J., Pennell M., Slater G., Brown J., FitzJohn R., Alfaro M., Harmon L. in prep. Geiger 2: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. Available from http://cran.r-project.org/web/packages/geiger/index.html.
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 65:3578–3589.
Etienne R.S., Haegeman, B. 2012. A conceptual and statistical framework for adaptive radiations with a key role for diversity dependence. Amer. Nat. 180:E75–E89.
Felsenstein J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. Amer. J. Hum. Genet. 25:471–492.
Felsenstein J. 1985. Phylogenies and the comparative method. Amer. Nat. 125:1–15.
Foote M. 1993. Discordance and concordance between morphological and taxonomic diversity. Paleobiology 19:185–204.
Foote M. 1994. Morphological disparity in Ordovician-Devonian crinoids and the early saturation of morphological space. Paleobiology 20:320–344.
Foote M. 1995. Morphological diversification of Paleozoic crinoids. Paleobiology 21:273–299.
Foote, M. 1997. The evolution of morphological diversity. Ann. Rev. Ecol. Syst. 28:129–152.
Freckleton R.P., Harvey P.H. 2006. Detecting non-Brownian trait evolution in adaptive radiations. PLoS Biol. 4:e373.
Freckleton R.P., Jetz W. 2009. Space versus phylogeny: disentangling phylogenetic and spatial signals in comparative data. Proc. Royal Soc. B: Biol. Sci. 276:21–30.
Gelfand A.E., Ghosh S.K. 1998. Model choice: A minimum posterior predictive loss approach. Biometrika 85:1–11.
Gingerich P.D. 1993. Quantification and comparison of evolutionary rates. Amer. J. Sci. 293A:453–478.
Glor R.E. 2010. Phylogenetic insights on adaptive radiation. Ann. Rev. Ecol., Evol. Syst. 41:251–270.
Grafen A. 1989. The phylogenetic regression. Philos. Trans. R. Soci. London B, Biol. Sci. 326:119–157.
Hansen T.F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351.
Harmon L.J., Losos J.B., Jonathan Davies T., Gillespie R.G., Gittleman J.L., Bryan Jennings W., Kozak K.H., McPeek M.A., Moreno-Roark F., Near T.J., Purvis A., Ricklefs R.E., Schluter D., Schulte II, J.A., Seehausen O., Sidlauskas B.L., Torres-Carvajal O., Weir J.T., Mooers A.Ø. 2010. Early bursts of body size and shape evolution are rare in comparative data. Evolution 64:2385–2396.
Harmon L.J., Schulte J.A., Larson A., Losos J.B. 2003. Tempo and mode of evolutionary radiation in iguanian lizards. Science 301: 961–964.
Harmon L.J., Weir J.T., Brock C.D., Glor R.E., Challenger W. 2008. Geiger: investigating evolutionary radiations. Bioinformatics 24:129–131.
HilleRisLambers J., Adler P., Harpole W., Levine J., Mayfield M. 2012. Rethinking community assembly through the lens of coexistence theory. Ann. Rev. Ecol. Evol. Syst. 43:227–248.
Hohn A.A., Read A.J., Fernandez S. Vidal O., Findley L.T. 1996. Life history of the vaquita, *Phocoena sinus* (Phocoenidae, Cetacea). J. Zool. 239:235–251.
Holland P., Welsch R. 1977. Robust regression using iteratively reweighted least-squares. Commun. Statist. Theor. Meth. 6:813–827.
Huber P.J. 1973. Robust regression: Asymptotics, conjectures and monte carlo. Ann. Stat. 1:799–821.
Huelsenbeck J.P., Nielsen R., Bollback, J.P. 2003. Stochastic mapping of morphological characters. Syst. Biol. 52:131–158.
Hunt G. 2006. Fitting and comparing models of phyletic evolution: random walks and beyond. Paleobiology 32:578–601.
Hunt G. 2007. The relative importance of directional change, random walks, and stasis in the evolution of fossil lineages. Proc. Natl Acad. Sci. 104:18404–18408.
Ingram T., Harmon L.J., Shurin J.B. 2012. When should we expect early bursts of trait evolution in comparative data? Predictions from an evolutionary food web model. J. Evol. Biol. 25:1902–1910.
Jernvall J., Hunter J., Fortelius M. 1996. Molar tooth diversity, disparity, and ecology in Cenozoic ungulate radiations. Science 274:1489–1492.
Kadane J.B., Lazar N.A. 2004. Methods and criteria for model selection. J. Amer. Stat. Assoc. 99:279–290.
Kutsukake N., Innan H. 2013. Simulation-based likelihood approach for evolutionary models of phenotypic traits on phylogeny. Evolution 67:355–367.
Lewis P.O., Xie W., Chen M.-J., Fan Y., Kuo L. 2014. Posterior predictive Bayesian phylogenetic model selection. Syst. Biol. 63:309–321.

Lillegraven J.A., Kielan-Jaworowska Z., Clemens W.A. eds. 1979. Mesozoic mammals: the first two-thirds of mammalian history. Berkeley: University of California Press.
Liow L.H., Quental T.B., Marshall C.R. 2010. When can decreasing diversification rates be detected with molecular phylogenies and the fossil record? Syst. Biol. 59:646–659.
Lloyd G.T., Wang S.C., Brusatte S.L. 2012. Identifying heterogeneity in rates of morphological evolution: Discrete character change in the evolution of lungfish (Sarcopterygii; Dipnoi). Evolution 66: 330–348.
Losos J.B., Mahler D.L. 2010. Evolution since Darwin: the first 150 years chap. Adaptive radiation: the interaction of ecological opportunity, adaptation, and speciation. Sunderland (MA): Sinauer.
Losos J.B., Miles, D.B. 2002. Testing the hypothesis that a clade has adaptively radiated: Iguanid lizard clades as a case study. Amer. Nat. 160:147–157.
Luo Z.-X. 2007. Transformation and diversification in early mammal evolution. Nature 450:1011–1019.
Mahler D.L., Revell L.J., Glor R.E., Losos J.B. 2010. Ecological opportunity and the rate of morphological evolution in the diversification of greater antillean anoles. Evolution 64: 2731–2745.
Martini A.S., Spezzaferri, F. 1984. A predictive model selection criterion. J. Royal Stat. Soc. Series B 46:296–303.
Mayfield M.M., Levine, J.M. 2010. Opposing effects of competitive exclusion on the phylogenetic structure of communities. Ecol. Lett. 13:1085–1093.
Meredith R.W., Janečka J.E., Gatesy J., Ryder O.A., Fisher C.A., Teeling E.C., Goodbla A., Eizirik E., Simão T.L.L., Stadler T., Rabosky D.L., Honeycutt R.L., Flynn J.J., Ingram C.M., Steiner C., Williams T.L., Robinson T.J., Burk-Herrick A., Westerman M., Ayoub N.A., Springer M.S., Murphy W.J. 2011. Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. Science 334:521–524.
Minin V.N., Suchard M.A. 2008. Fast, accurate and simulation-free stochastic mapping. Philo. Trans. Royal Soci. B: Biol. Sci. 363:3985–3995.
Mooers A.Ø., Vamosi S.M., Schluter, D. 1999. Using phylogenies to test macroevolutionary hypotheses of trait evolution in cranes (Gruinae). Amer. Nat. 154:249–259.
Nielsen R. 2002. Mapping mutations on phylogenies. Syst. Bio. 51:729–739.
Nowak R.M. 1999. Walker's Mammals of the World. 6th ed. Baltimore (MD): The Johns Hopkins University Press.
O'Meara B.C., Ané 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. Zool. Scripta 26:331–348.
Pagel M. 1999. Inferring the historical patterns of biological evolution. Nature 401:877–884.
Pennell M.W., Harmon L.J. 2013. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. Ann. NY Acad. Sci. 1289:90–105.
Pennell M.W., Harmon L.J., Uyeda J.C. 2013. Is there room for punctuated equilibrium in macroevolution? Trends Ecol. Evol. Http://dx.doi.org/10.1016/j.tree.2013.07.004.
Pennell M.W., Sarver B.A.J., Harmon L.J. 2012. Trees of unusual size: Biased inference of early bursts from large molecular phylogenies. PLoS ONE 7:e43348.
Quental T.B., Marshall C.R. 2010. Diversity dynamics: molecular phylogenies need the fossil record. Trends Ecol. Evol. 25:434–441.
R Development Core Team. 2013. R: A Language and Environment for Statistical Computing. Available from: http://www.R-project.org. Vienna (Austria): [Internet] R Foundation for Statistical Computing.
Rabosky D.L. 2009. Ecological limits and diversification rate: alternative paradigms to explain the variation in species richness among clades and regions. Ecol. Lett. 12:735–743.
Rabosky D.L., Slater G.J., Alfaro, M.E. 2012. Clade age and species richness are decoupled across the eukaryotic tree of life. PLoS Biol. 10:e1001381.
Rambaut A., Drummond A. 2007. Tracer v1.4. Available from http://beast.bio.ed.ac.uk/Tracer.
Revell L.J. 2013. Two new graphical methods for mapping trait evolution on phylogenies. Method. Ecol. Evol. 4:754–759.
Revell L.J., Harmon L.J., Collar D.C. 2008. Phylogenetic signal, evolutionary process, and rate. Syst. Biol. 57:591–601.
Revell L.J., Mahler D.L., Peres-Neto P.R., Redelings B.D. 2012. A new phylogenetic method for identifying exceptional phenotypic diversification. Evolution 66:135–146.
Schluter D. 2000. The ecology of adaptive radiation. Oxford UK: Oxford University Press.
Schwarz G. 1978. Estimating the dimension of a model. Ann. Stat. 6:461–464.
Sidlauskas B. 2007. Testing for unequal rates of morphological diversification in the absence of a detailed phylogeny: A case study from characiform fishes. Evolution 61:299–316.
Siepielski A.M., McPeek M.A. 2010. On the evidence for species coexistence: a critique of the coexistence program. Ecology 91:3153–3164.
Simpson G.G. 1944. Tempo and mode of evolution. New York USA: Columbia University Press.
Simpson G.G. 1953. Major features of evolution. New York USA: Columbia University Press.
Slater G.J. 2013. Phylogenetic evidence for a shift in the mode of mammalian body size evolution at the Cretaceous-Palaeogene boundary. Method. Ecol. Evol. 4:734–744.
Slater G.J., Harmon L.J., Alfaro M.E. 2012a. Integrating fossils with molecular phylogenies improves inference of trait evolution. Evolution 66:3931–3944.
Slater G.J., Harmon L.J., Wegmann D., Joyce P., Revell L.J., Alfaro M.E. 2012b. Fitting models of continuous trait evolution to incompletely sampled comparative data using approximate Bayesian computation. Evolution 66:752–762.
Slater G.J., Price S.A., Santini F., Alfaro M.E. 2010. Diversity versus disparity and the radiation of modern cetaceans. Proc. R Soc. B-Biol. Sci. 277:3097–3104.
Stadler T. 2011. Simulating trees with a fixed number of extant species. Syst. Biol. 60:676–684.
Stromberg A. 2004. Why write statistical software? The case of robust statistical methods. J. Stat. Software 10:1–8.
Thomas G.H., Freckleton R.P. Székely T. 2006. Comparative analyses of the influence of developmental mode on phenotypic diversification rates in shorebirds. Proc. R Soci. B: Biol. Sci. 273:1619–1624.
Venables W.N., Ripley B.D. 2002. Modern applied statistics with S. 4th ed. New York: Springer.
Venditti C., Meade A., Pagel M. 2011. Multiple routes to mammalian diversity. Nature 479:393–396.
Wesley-Hunt G.D. 2005. The morphological diversification of carnivores in North America. Paleobiology 31:35–55.
Yoder J.B., Clancey E., Des Roches S., Eastman J.M., Gentry L., Godsoe W., Hagey T.J., Jochimsen D., Oswald B.P., Robertson J., Sarver B.A.J., Schenk J.J., Spear S.F., Harmon L.J. 2010. Ecological opportunity and the origin of adaptive radiations. J. Evol. Biol. 23:1581–1596.