

## Phylogenies, the Comparative Method, and the Conflation of Tempo and Mode

ANTIGONI KALIONTZOPOULOU<sup>1,2,\*</sup> AND DEAN C. ADAMS<sup>2,3</sup>

<sup>1</sup>CIBIO/InBio, Centro de Investigação em Biodiversidade e Recursos Genéticos, Campus Agrário de Vairão, 4485-661 Vairão, Portugal; <sup>2</sup>Department of Ecology, Evolution, and Organismal Biology; and <sup>3</sup>Department of Statistics, Iowa State University, Ames, IA 50011, USA

\*Correspondence to be sent to: CIBIO/InBio, Centro de Investigação em Biodiversidade e Recursos Genéticos, Campus Agrário de Vairão, 4485-661 Vairão, Portugal; E-mail: antigoni@cibio.up.pt.

Received 31 July 2013; reviews returned 6 October 2015; accepted 7 October 2015  
Associate Editor: Luke Harmon

**Abstract.**—The comparison of mathematical models that represent alternative hypotheses about the tempo and mode of evolutionary change is a common approach for assessing the evolutionary processes underlying phenotypic diversification. However, because model parameters are estimated simultaneously, they are inextricably linked, such that changes in tempo, the pace of evolution, and mode, the manner in which evolution occurs, may be difficult to assess separately. This may potentially complicate biological interpretation, but the extent to which this occurs has not yet been determined. In this study, we examined 160 phylogeny × trait empirical data sets, and conducted extensive numerical phylogenetic simulations, to investigate the efficacy of phylogenetic comparative methods to distinguish between models that represent different evolutionary processes in a phylogenetic context. We observed that, in some circumstances, a high uncertainty exists when attempting to distinguish between alternative evolutionary scenarios underlying phenotypic variation. When examining data sets simulated under known conditions, we found that evolutionary inference is straightforward when phenotypic patterns are generated by simple evolutionary processes that are represented by modifying a single model parameter at a time. However, inferring the exact nature of the evolutionary process that has yielded phenotypic variation when facing complex, potentially more realistic, mechanisms is more problematic. A detailed investigation of the influence of different model parameters showed that changes in evolutionary rates, marked changes in phylogenetic means, or the existence of a strong selective pull on the data, are all readily recovered by phenotypic model comparison. However, under evolutionary processes with a milder restraining pull acting on trait values, alternative models representing very different evolutionary processes may exhibit similar goodness-of-fit to the data, potentially leading to the conflation of interpretations that emphasize tempo and mode during empirical evolutionary inference. This is a mathematical and conceptual property of the considered models that, while not prohibitive for studying phenotypic evolution, should be taken into account and addressed when appropriate. [Comparative method; mode; model fit; phenotypic evolution; phylogeny; tempo.]

The phylogenetic comparative method, where species trait values are examined in light of the phylogeny of the group to infer the evolutionary processes that have shaped phenotypic diversity, is a major framework in evolutionary biology (Harvey and Pagel 1991). In recent years, remarkable advances have been made by the development of new tools for investigating macroevolutionary phenotypic patterns and testing hypotheses about the biological mechanisms that shape them. Rooted in the approaches of phylogenetic independent contrasts (Felsenstein 1985, 1988) and phylogenetic generalized least squares (PGLS; Grafen 1989; Rohlf 2001), numerous methods have been developed to investigate how phenotypes diversify over evolutionary time. Testing for diversifying selection and adaptation (Butler and King 2004) or for adaptive radiation (Harvey and Rambaut 2000; Glor 2010; Harmon et al. 2010); understanding whether morphological disparity is coupled to cladogenesis (Harmon et al. 2003; Ricklefs 2004; Rabosky and Adams 2012) or species diversification (Bokma 2002; Adams et al. 2009; Rabosky and Adams 2012); identifying phenotypic convergence and parallelism (Harmon et al. 2005; Stayton 2006; Revell et al. 2007; Adams 2010); and examining the correlation among traits through evolutionary history (Martins and Garland 1991; Pagel 1998; Revell and Collar 2009) are only some examples of how the study of phenotypic traits on phylogenies have aided biologists in understanding the processes driving diversification.

Common to all these approaches is the use of mathematical models that aim at approximating the tempo and mode of evolutionary change (Simpson 1944; Fitch and Ayala 1994). These models are rooted in similar methods first developed in paleontology to explore how phenotypes evolve. Researchers in this field have long been concerned with evolutionary tempo and mode, which they study by using data from the fossil record to infer these evolutionary parameters (Gingerich 1976; Gould and Eldredge 1977; Gould 1980; Fitch and Ayala 1994). Paleontological studies were profoundly influenced by the hallmark contribution of George Gaylord Simpson (1944) in which he used the word “tempo” to define the pace at which phenotypic evolution proceeds. Likewise, he defined “mode” as “...the study of the way, manner, or pattern of evolution, a study in which tempo is a basic factor...” (Simpson 1944). In his definitions, Simpson inextricably linked tempo and mode together: the self-contained description of how fast evolutionary changes occurs (tempo) was a basic component for describing the way in which these changes are attained (mode). Indeed, a recent investigation of the paleontological methods used to estimate and compare evolutionary rates shows that different rate metrics perform better depending on the mode of evolution (Hunt 2012). Thus, in paleontological studies, it is clear that tempo and mode are intimately related and can often not be accurately characterized independently (Hunt 2012). This observation raises an important question: is this also the case when

using phylogenetic comparative approaches to assess phenotypic evolution of extant taxa?

In modern phylogenetic comparative methods, the tempo and mode of evolution are approached through mathematical models that describe extant phenotypic variation given a phylogenetic hypothesis for the group of interest. The first breakthrough toward modeling how continuous phenotypic traits evolve on phylogenies was the introduction of a random-walk model (Brownian motion, BM; Edwards and Cavalli-Sforza 1964; Felsenstein 1973, 1985, 1988; Harvey and Pagel 1991). Under BM, phenotypic variation accumulates linearly over time and the amount of change in the value of a phenotypic trait ( $X$ ) over a small time interval ( $t$ ) can be modeled as:

$$dX(t) = \sigma dB(t) \quad (1)$$

In Equation (1),  $dB(t)$  represents independent, normally distributed, random perturbations and  $\sigma$  is the evolutionary rate or variance. The maximum-likelihood estimator of the evolutionary rate is given by:

$$\sigma = \sqrt{\frac{(\mathbf{X} - E(\mathbf{X}))^t \mathbf{C}^{-1} (\mathbf{X} - E(\mathbf{X}))}{N}} \quad (2)$$

where  $\mathbf{C}$  is the phylogenetic variance–covariance matrix,  $N$  is the number of taxa,  $\mathbf{X}$  is the vector of phenotypic trait values at the tips of the phylogeny, and  $E(\mathbf{X})$  is the expected value of  $\mathbf{X}$ , or the phylogenetic mean, corresponding to the value at the root node of the phylogeny under BM (O’Meara et al. 2006). The evolutionary rate  $\sigma$  is a central parameter of the BM model, as it captures how fast evolution proceeds. As such, it represents Simpson’s idea of evolutionary tempo.

Despite its enormous utility and wide application in evolutionary research, the BM model is sometimes too simple to represent complex evolutionary reality (Butler and King 2004; Beaulieu et al. 2012). Extensions to this model have thus been developed to allow assessing not only how fast, but also *how* evolution has generated the phenotypic patterns observed in nature. One family of these extended models aims at providing a solution for modeling the tempo of phenotypic evolution more accurately. For instance, the pace of phenotypic evolution may vary across single branches of the phylogeny (McPeck 1995; O’Meara et al. 2006; Revell 2008), between groups of taxa on a phylogeny (Garland 1992; O’Meara et al. 2006; Thomas et al. 2006, 2009; Adams 2014), across evolutionary time (Pagel 1999; Blomberg et al. 2003; Harmon et al. 2010), or among traits (Adams 2013). Such evolutionary hypotheses are tested by fitting models of evolution that encompass more than one evolutionary rate parameter across the phylogeny, and then comparing their fit to a single-rate BM.

Another family of models includes an additional term, yielding an Ornstein–Uhlenbeck (OU) process, which describes an evolutionary “pull” of trait mean value toward one or more optima through time:

$$dX(t) = \sigma dB(t) + \alpha[\beta(t) - X(t)]. \quad (3)$$

The first term of Equation (3) corresponds to the random walk component, while the second term represents the strength of selection ( $\alpha$ ) toward a phenotypic optimum ( $\beta$ ) (Butler and King 2004; Beaulieu et al. 2012). Notice that here we follow the notation of Beaulieu et al. (2012) and represent phenotypic optima as  $\beta$ , to avoid confusion with the notation  $\theta$ , sometimes used for the relative rate parameter (i.e., Thomas et al. 2006; 2009). From the above mathematical formulation, the first term of Equation (3) is dominated by the evolutionary rate  $\sigma$ . The second term represents a change in mean trait value, occurring toward an optimal state  $\beta$  under a pace proportional to  $\alpha$  (Butler and King 2004). By varying the terms  $\alpha$  and  $\beta$  of Equation (3), one can represent evolutionary changes that vary in strength and direction, correspondingly (Butler and King 2004; Beaulieu et al. 2012). For  $\alpha=0$ , Equation (3) collapses back to a BM process. Variation in the relative influence of  $\sigma$  and  $\alpha$  would then yield models that represent evolutionary processes that lie closer or further away from the simple BM model. In contrast to the first family of models, though, which focus on modifications of the speed by which evolution proceeds, these models represent a shift from a random walk (BM) to an evolutionary process that also encompasses changes in trait mean value.

Recently, more complex models have been developed in an attempt to characterize the biological mechanisms underlying phenotypic evolution more accurately. For instance, this can be done either by allowing all  $\sigma$ ,  $\alpha$ , and  $\beta$  in Equation (3) to vary (Beaulieu et al. 2012); or by incorporating different phylogenetic means for different parts of the tree in the calculation of  $\sigma$  (O’Meara et al. 2006; Thomas et al. 2006, 2009). In each case, model parameters are simultaneously estimated, typically in concert with maximizing the corresponding likelihood equation (but see also e.g., Revell et al. 2011; Eastman et al. 2011; Revell and Reynolds 2012 for Bayesian implementations). Some of these parameters contribute to modeling trait variance across taxa through a mean value (i.e., phylogenetic means  $E(\mathbf{X})$ , optimal trait values  $\beta$ ), while others model residual variance (i.e., evolutionary rates  $\sigma$ , strength of selection  $\alpha$ ). Alternative models are then compared by evaluating their fit to the data given the underlying phylogeny through likelihood comparison methods (e.g., likelihood ratio tests, information theoretic criteria, or Monte Carlo simulations; Boettiger et al. 2012). Through this procedure, evolutionary biologists attempt to obtain a reliable model of the historical events that underlie current phenotypic variation. As models become more complex, though, inference becomes more complicated. This is because each of the mathematical parameters used to characterize phenotypic evolution in a phylogenetic context is estimated with respect to the other parameters included in the underlying model. Therefore, it is of interest to determine whether changes in model parameters can be readily assessed when using phylogenies to study phenotypic evolution.

In this article we investigate the efficacy of comparative methods to distinguish between phylogenetic comparative models that emphasize changes in different evolutionary parameters. We restrict our study to those cases where evolutionary changes are found across groups on a phylogeny. These encompass questions about how ecological, biogeographic, historical, or other life-history factors have influenced trait diversification, and they are among the most frequently examined hypotheses in phenotypic evolution. Based on empirical data, we demonstrate that it is frequently the case that models representing very different evolutionary processes are equally likely for describing the data given a phylogeny. A detailed examination of model parameters using simulations indicates that model performance is more strongly dependent on variations in mode than on variations in tempo. When complex evolutionary processes are considered and differences in means are not prominent, different models receive similar support, potentially leading to the conflation of radically different interpretations during evolutionary inference. This is a property of the considered models that, while not prohibitive for studying phenotypic evolution, should be taken into account and addressed when appropriate.

#### DATA SETS AND MODELS CONSIDERED

In an effort to determine whether different model parameters could be unambiguously approached using standard phylogenetic comparative techniques, we examined previously published data sets and conducted numerical simulations. Empirical data sets were used to investigate the degree of ambiguity encountered when using phylogenetic comparative models on real biological data. In continuation, we used simulations to build specific evolutionary scenarios, where the evolutionary process underlying phenotypic variation was considered known. This allowed us to address whether the inferred evolutionary process found by comparing the fit of alternative evolutionary models to the data matched the known process under which the phenotypes were actually generated.

For each phylogeny  $\times$  trait data set, empirical or simulated, we fit six evolutionary models, which represent different hypotheses about the process underlying evolutionary change. The simplest model examined, often considered as a null hypothesis, was a simple BM, with a single evolutionary rate across all branches (BM1). In a BM framework, we also examined two other models, which consider variation in evolutionary rates across groups of taxa. The first fits a different evolutionary rate for each group based on a single phylogenetic mean (BMS; i.e., the “non-censored” approach of O’Meara et al. 2006). The second fits a different evolutionary rate for each group, also considering different evolutionary means for each group (BMSG; i.e., the “censored” approach of O’Meara et al. 2006; see also Thomas et al. 2006, 2009). Note that the

method proposed by Thomas et al. (2006) corresponds to the censored approach developed by O’Meara et al. (2006), when at least one of the examined groups is monophyletic (see Online Appendix 3, available on Dryad at <http://dx.doi.org/10.5061/dryad.2ss46>). In an OU framework, we fit three models: the first with a single rate and a single optimum for all taxa (OU1), the second with a single rate but different optima for each group of taxa (OUM), and the third with different optima and different rates for each group of taxa (OUMV) (Butler and King 2004; Beaulieu et al. 2012). We did not consider variation in the strength of selection ( $\alpha$ ) across groups, as this parameter has only recently been allowed to vary in evolutionary models and the inference of differences in  $\alpha$  has been shown to bear a high uncertainty (Beaulieu et al. 2012; but see also below for the influence of  $\alpha$  on model inference). Models were fit using the R-packages OUwie (Beaulieu et al. 2012) and motmot (Thomas and Freckleton 2012). Before engaging in any model comparisons, we first used simulations to confirm that the two software packages provide comparable parameter estimates and likelihood scores (Online Appendix 4, available on Dryad at <http://dx.doi.org/10.5061/dryad.2ss46>). Models that did not reach convergence were excluded from model comparison procedures (see Online Appendix 1, available on Dryad at <http://dx.doi.org/10.5061/dryad.2ss46>, for details).

Once all models were computed, we compared their to the data using likelihood approaches. This is a strategy with a long history in ecology and evolutionary biology, and several different measures may be used to compare the goodness-of-fit of different models (Burnham and Anderson 2002). One of the most commonly used in recent studies of phenotypic evolution is Akaike’s information criterion (AIC). Model comparisons are typically performed by ranking the models based on their AIC. Models that lie less than four AIC units from the model with the lowest value ( $\Delta AIC < 4$ ) are usually considered as indistinguishable in terms of their goodness-of-fit to the data (Burnham and Anderson 2002). Here we followed a two-step procedure to evaluate the goodness-of-fit of alternative models. First, for each phylogeny  $\times$  trait data set we ranked all models fitted based on their  $\Delta AICc$  and retained only those with  $\Delta AICc < 4$ . We used  $AICc$ , which implements a correction to AIC for finite sample size (Burnham and Anderson 2002). Second, we calculated pairwise  $\Delta AICc$  values to evaluate similarity in fit between pairs of models lying in the  $\Delta AICc < 4$  interval. Notice, however, that this approach can be somewhat problematic, because our comparisons involved several pairs of nested models. Under such circumstances, the log likelihood of the simpler of two models can never exceed that of the more complex one (Hunt 2006). As such, if there is no difference in likelihood between the two alternative models,  $\Delta AIC$  is driven exclusively by the difference in the number of parameters and it will be, for example,  $\Delta AIC \approx 2$  if the models differ by only one in the number of estimated parameters (Burnham and Anderson 2002,

TABLE 1. Summary of tree size ( $n$ ), group distribution across the phylogeny, and number of phylogeny  $\times$  trait comparisons included ( $N_{\text{phy} \times \text{trt}}$ ) for the different empirical data sets examined

Group	Taxon	Trait	$n$	Groups	$N_{\text{phy} \times \text{trt}}$	Source
Plants	Grasses	Ecological niche	141	Clustered	10	Edwards and Still (2008)
	<i>Rafflesia</i>	Flower size	19	Clustered	1	Barkman et al. (2008)
	Sedges	Chromosome number, genome size	87	Clustered	4	Chung et al. (2012)
	Sedges	Chromosome number	53	Clustered	8	Hipp (2007)
Fish	Centrarchid fishes	Jaw morphology	27	Clustered	12	Collar and Wainwright (2006)
	Centrarchid fishes	Jaw morphology	29	Random	12	Collar et al. (2009)
	Cichlid fishes	Jaw morphology	79	Clustered	5	Hulsey et al. (2010)
	Parrotfishes	Jaw morphology	122	Clustered	18	Price et al. (2010)
	Parrotfishes	Jaw morphology	118	Random	8	Price et al. (2011)
	Pupfish	Body size/shape	48	Clustered	32	Martin and Wainwright (2011)
Anura	Fanged frogs	Body size	21	Random	4	Setiadi et al. (2011)
	Hylid frogs	Body size	220	Clustered	1	Wiens et al. (2011)
	<i>Osteopilus</i> frogs	Body size	171	Clustered	1	Moen and Wiens (2009)
Squamata	<i>Anolis</i> lizards	Body size	160	Random	12	Thomas et al. (2009)
	<i>Phelsuma</i> geckos	Body size/shape	20	Clustered	15	Harmon et al. (2008)
	Monitor lizards	Body size	37	Random	3	Collar et al. (2011)
Birds	Birds	Body size, metabolic rate, temperature	44	Clustered	3	Swanson and Garland (2009)
	Passerine birds	Body size, metabolic rate	60	Clustered	2	Swanson and Bozinovic (2012)
Mammals	All mammals	Body size	842/539	Random	4	Raia and Meiri (2011)
	Cetaceans	Body size	68	Random	3	Slater et al. (2010)
	Chiroptera	Skull morphology, trophic level	81	Clustered	2	Dumont et al. (2011)

p. 131). In such cases, one or more models may lie in the  $\Delta\text{AICc} < 4$  interval, but if the best model is nested in others it should be preferred, and there is in reality no uncertainty in terms of model selection. Taking this property of  $\Delta\text{AIC}$  into account, we examine comparisons of nested and non-nested pairs of models separately when considering the results obtained in relation to variations of evolutionary model parameters.

#### EVOLUTIONARY INFERENCE ON EMPIRICAL DATA SETS

We compiled a total of 160 phylogeny  $\times$  trait empirical data sets from 21 published studies that tested for differences between groups in evolutionary tempo and mode across phylogenies. These encompass a wide array of continuous phenotypic traits for several plant and animal taxa, examined on phylogenies of different sizes (Table 1). In all these studies, the authors examined the effect of some biological factor of interest on phenotypic evolution, by comparing tempo and mode attributes of the evolutionary process across biologically identified groups of interest. We used the biological hypotheses examined by the authors in the original studies to allocate taxa into groups for our comparisons. However, to delimit our analyses, we only conducted comparisons of two groups, as an increase in the number of groups can only make inference more complex. In this context, when original studies included evolutionary models with multiple rates, evolutionary means, or adaptive optima, we broke down these hypotheses into pairwise group comparisons and compared one group with the pool of all other groups considered. Furthermore, for all empirical data sets examined, we closely followed preliminary data processing operations

(e.g., logarithmic transformations and size-correction of phenotypic traits, tree-length standardizations) as described by the authors.

#### Results from Empirical Data Sets

Pairwise comparison between models within 4  $\Delta\text{AICc}$  units of the first-ranking model indicated that mean pairwise  $\Delta\text{AICc}$  values are often below the established threshold of 4 units (Fig. 1; Online Appendix 1, available on Dryad at <http://dx.doi.org/10.5061/dryad.2ss46>). This means that pairs of models representing different evolutionary hypotheses are frequently very similar in terms of their fit to the data, and thus the best estimate of the evolutionary process may not be easily identifiable. This lack of statistical distinction between models frequently occurred within the BM and within the OU families of models (Fig. 1). In these cases, the ambiguity in identifying the best model was among models encompassing variations between groups in the rate parameter or among models encompassing variations in the number of hypothesized optima. These sets, however, include models that are nested and therefore this lack of statistical distinctiveness does not hinder evolutionary inference.

However,  $\Delta\text{AICc}$  was also frequently below 4 units when comparing BMS to OU1 or OUM, and BMSG to OUM or OUMV models (Fig. 1). In these cases, statistical lack of distinctiveness in model fit also translates into an evolutionary uncertainty. Indeed, these pairs of models are not nested and they differ in the model parameters they include, representing radically different evolutionary processes. For instance, a model with two rates under BM (BMS) is frequently

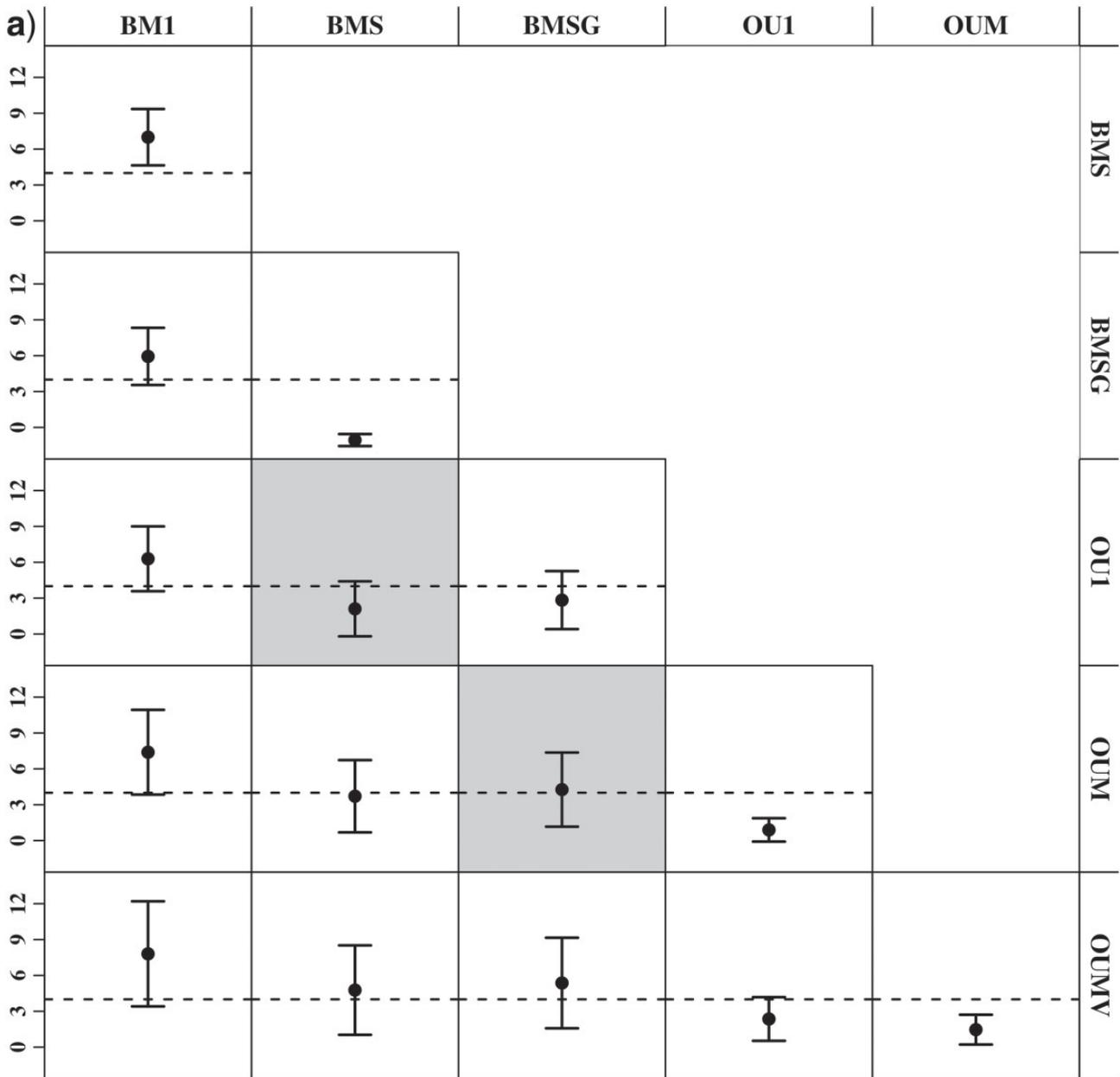


FIGURE 1. Means and corresponding 95% confidence intervals of pairwise  $\Delta AICc$  for all models fitted to empirical data sets ( $n = 160$ ). Dashed lines represent the frequently used threshold of  $\Delta AICc = 4$ . Pairs of models with the same number of parameters are shaded in gray.

indistinguishable in terms of fit to the data from an OU1 model where all species evolve toward a single optimum under a single rate. These results indicate that inferring the correct evolutionary model underlying phenotypic diversification between groups may hold a higher uncertainty than is generally appreciated.

SIMULATIONS UNDER DIFFERENT EVOLUTIONARY SCENARIOS

The empirical results above demonstrate that biologists interested in understanding phenotypic

evolution may frequently face difficulties when testing hypotheses that encompass the simultaneous modification of several model parameters. To examine the generality of this observation, we used simulations where a continuous trait was simulated on a phylogeny using a known evolutionary process, and the resulting phenotypic patterns were subsequently evaluated using several evolutionary models that differ in their included parameters. In this way, we could determine the extent to which the inferred evolutionary process found by comparing the fit of alternative evolutionary models to the data and the phylogeny matched the

known process under which the phenotypes actually evolved.

Briefly, our simulation protocol was as follows. First we simulated a continuous phenotypic trait evolving on 64-taxa, random phylogenies. We began by simulating a single set of 1000 pure-birth phylogenies ( $\lambda=1$ ,  $\mu=0$ ) using the `pmtree` function of `phytools` R-package (Revell 2012). All phylogenetic trees were scaled to unit total length, and used as the underlying phylogenetic hypotheses for all simulations. We used `OUwie` (Beaulieu et al. 2012) and `motmot` (Thomas and Freckleton 2012) R-packages to simulate trait values for two groups under the six evolutionary models described above. Group membership, represented as a binary trait evolving on the phylogeny, was randomized by sampling a single shift in the binary trait in a node relatively deep in the tree, thus yielding at least one monophyletic group. We deliberately avoided examining groups randomly distributed across the phylogeny, because rate estimates in such circumstances are known to be inflated when both means and rates vary across groups, and rate comparison methods may present a high type I error (Thomas et al. 2009). For group distribution simulations, we used `ape` (Paradis et al. 2004), `geiger` (Harmon et al. 2009), and `phangorn` (Schliep 2011) packages for R (R Development Core Team 2012).

For the continuous trait, we simulated data sets with or without differences in evolutionary rates (setting  $\sigma_1^2=1$  and the relative rate parameter  $\sigma_2^2/\sigma_1^2$  to either 1 or 6, respectively), under both BM and OU; with or without differences in phylogenetic means under BM (setting the difference in means to either 0 or 3 standard deviations of the mean phenotype at the tips; see Thomas et al. (2006, 2009)); and with or without differences in optima ( $\beta$ ) under OU (setting  $\beta_1=1$  and the relative optima parameter  $\beta_2/\beta_1$  to either 1 or 5, which correspond to an absolute difference in optima of 0 or 4, respectively). These parameter settings were chosen to closely match those observed in empirical studies, while maximizing the potential for discriminating between different models (but see below for variation in simulated parameter values). For instance, the tree size used for simulations (64 taxa) approaches that of empirical data sets (mean tree size = 80), while facilitating computation for simulations. Similarly, the mean value of relative rate estimated by the BMS and BMSG models for empirical data sets was approximately  $\sigma_2^2/\sigma_1^2=5$  (depending on the model fit) and we set this parameter to  $\sigma_2^2/\sigma_1^2=6$ . In OU-based simulations, the “rubber band” parameter (Butler and King 2004) was set to  $\alpha=1$ , which translates to a moderate phylogenetic half-life of  $t_{1/2}=\ln(2)/\alpha\approx 0.69$ . This parameter represents the time it takes for adaptation to a new optimum to become more influential than constraints from the ancestral state and, as such, it substantially influences the dynamics of OU models (Hansen 1997; see also further on for the effect of variation in  $\alpha$ ). Combinations of the above parameter settings yield the six models examined for the empirical data sets (e.g., BM1, BMS, BMSG, OU1, OUM,

and OUMV). We simulated a total of 1000 data sets under each model.

We then fit the same set of six models to all simulated data sets and followed the same procedure as above to filter out models that did not reach convergence. In brief, we first excluded all models that failed to converge (giving package errors and failing to provide a solution), as well as those models in which the estimated alpha parameter equaled the upper bound of the optimizing algorithm. Then, for each type of model fitted, we examined the distribution of estimates for each of the model parameters across the 1000 data sets simulated for each evolutionary process and filtered out models with relative sigma, difference in means (for the BMSG model) or relative optima (for the OUM and OUMV models) parameter values that were outside the 99% quantiles of the parameter distribution. Finally, we conducted comparisons using  $\Delta\text{AICc}$  to gain insight into potential convergence between models in terms of fit to the data.

### Simulation Results

In accordance with the pattern observed in empirical data sets, model evaluation of data simulated under known evolutionary processes indicated that, for the simulation conditions used, several of the candidate models could not be efficiently distinguished with respect to their statistical fit to the data. Importantly, while the model used to simulate the data was very often the one with the lowest mean  $\text{AICc}$  score for that simulation condition, several other evolutionary models were within 4  $\Delta\text{AICc}$  units from it or exhibited lower  $\Delta\text{AICc}$  values than it (Fig. 2). Since using simulations enables us to access the “true” underlying process, this indicates that alternative evolutionary models might be equally plausible explanations for phenotypic patterns generated under a specific evolutionary model. For the most simple model available (BM1; Fig. 2a), several other models exhibited  $\Delta\text{AICc} < 4$ , but BM1 was globally the best-fit model. However, when model parameters varied between groups inferential ambiguity increased. For instance, when all taxa evolved toward a single selective optimum (OU1; Fig. 2d), several other models were frequently in the  $\Delta\text{AICc} < 4$  interval.

Interestingly, we found that variation in evolutionary rates between groups of taxa was generally easy to detect. When data were simulated with  $\theta=\sigma_2^2/\sigma_1^2\neq 1$  (e.g., under the BMS, BMSG, and OUMV models), models that represented a process with a single rate across the phylogeny were visibly worse in terms of likelihood, and exhibited very high  $\Delta\text{AICc}$  scores (Fig. 2b,c,f). Thus, current models used in phylogenetic comparative methods are capable of diagnosing the presence of multiple evolutionary rates on a phylogeny, when evolutionary rate is known to vary across groups of taxa.

While differences in fit to the data may be reliably identified between models that contain distinct

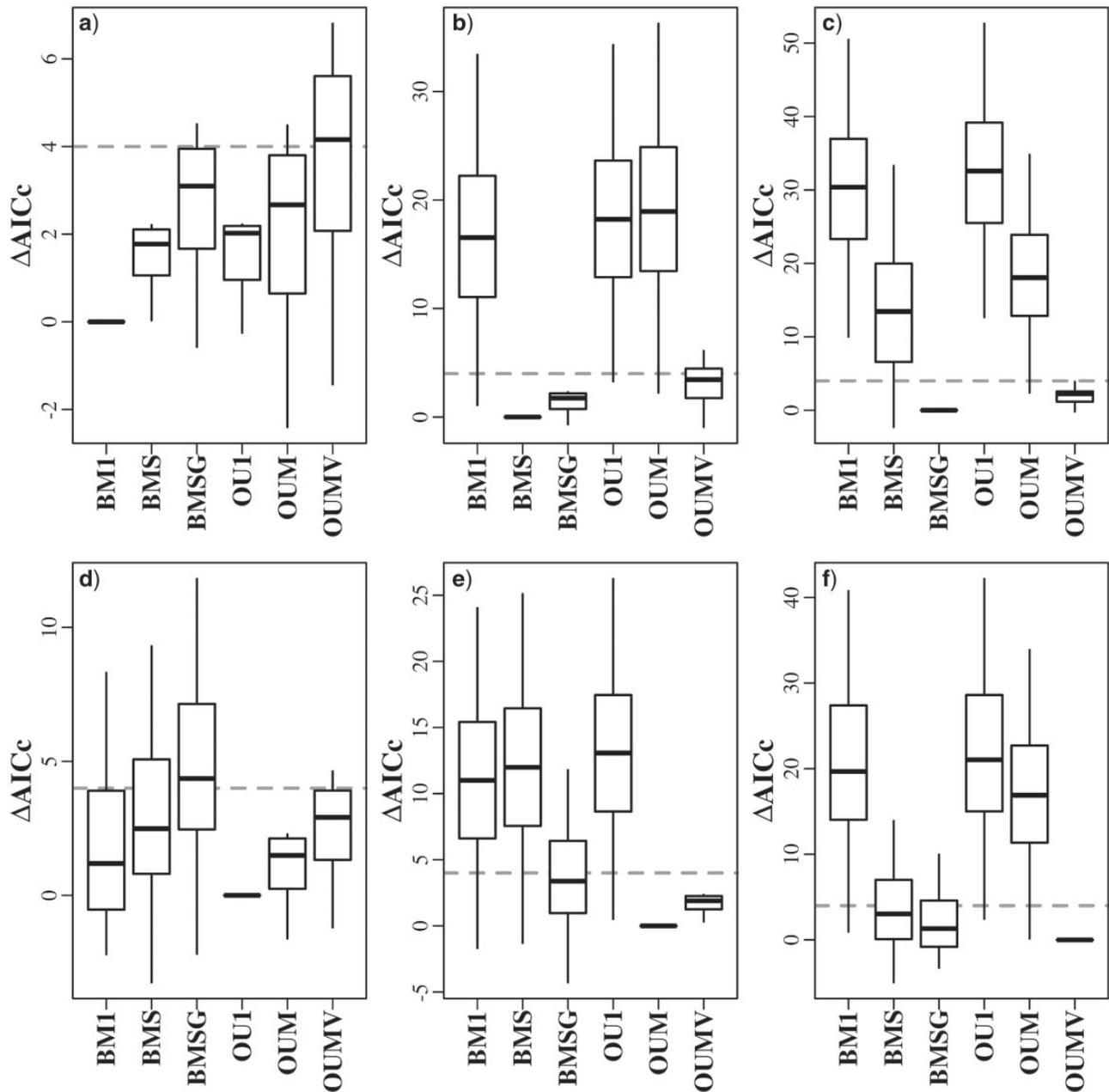


FIGURE 2. Quantile boxplots of  $\Delta\text{AICc}$  scores obtained by simulating 1000 data sets under six different models a) BM1, b) BMS, c) BMSG, d) OU1, e) OUM, f) OUMV and then fitting the same six models to each of them.  $\Delta\text{AICc}$  have been standardized relative to the simulating model ("real" model underlying the data), which therefore always has  $\Delta\text{AICc} = 0$ . The dashed horizontal line represents the frequently used threshold of  $\Delta\text{AICc} = 4$ .

evolutionary rates, variation in parameters that describe differences in evolutionary means (i.e.,  $E(\mathbf{X})$ ,  $\alpha$ , and  $\beta$  in Equations (2) and (3)) was not nearly as straightforward to identify. Specifically, for data sets simulated with multiple rates (Fig. 2b,c,f), both BM and OU models that allow for variation in rates (e.g., BMS, BMSG, and OUMV) showed similar fits to the data, irrespective of the model used to produce phenotypic variation. A similar pattern was observed for data simulated under an OU process with a single evolutionary

rate and two optima (OUM; Fig. 2e). In this case, the differentiation in phylogenetic means was easily detected, as models not encompassing this parameter (e.g., BM1, BMS, and OU1) generally exhibited high  $\Delta\text{AICc}$  scores. Thus, these results indicate that while model comparison promptly detects that multiple rates or multiple means are necessary for explaining the phenotypic data, it is not conclusive on the mode of evolution that has acted under the simulation conditions used here.

#### VARIATION IN MODEL PARAMETERS AND GOODNESS OF FIT

The examination of both empirical and simulated data sets suggests that inferring the evolutionary process underlying phenotypic diversity may be challenging. Most importantly, the results obtained by fitting evolutionary models to data simulated under known evolutionary processes suggest that explanations corresponding to models that modify the way trait variance and trait mean value are modeled are often confounded. This may suggest that similar phenotypic patterns may emerge by varying evolutionary tempo, mode, or both. Notice, however, that evolutionary inferences are made by comparing models that modify different parameters of Equation (3), where the two parts of this equation are added together to model phenotypic change. In other words, when modeling phenotypic evolution based on present trait values, one considers a first component related to the mean structure of the data (represented by  $E(\mathbf{X})$  and  $\beta$  in Equation (3)), and a second component related to the residual variance (determined by  $\sigma$  and  $\alpha$  in Equation (3); see also above). As such, variation in the relative weight of these two pieces may be responsible for the convergence in model fit observed above. To investigate this hypothesis and to pinpoint the circumstances under which sets of models may be confounded, we ran additional simulations where we varied the relative magnitude of model parameters. For this, we focused on those simulating conditions and models fit that pointed to a potential conflation of evolutionary model parameters. This way we could examine how different models perform when the relative influence of different parameters on phenotypic patterns are known.

Because statistical ambiguity was mainly encountered in models that involved variations in evolutionary rates and phylogenetic means or selective optima (i.e., BMS, BMSG, OUM, and OUMV), we first conducted simulations in which we varied these model parameters. Specifically, we simulated 1000 data sets under each of these models and setting relative rate  $\theta$  to either 3, 4, 5, or 6; difference in means under BMSG to either 1, 2, 4, or 6 standard deviations; and relative optima  $\beta_2/\beta_1$  to either 2, 5, 10, or 20. We then fit the subset of models that exhibited a mean  $\Delta\text{AICc} < 4$  during full simulation runs (see Fig. 2) to each of the simulated data sets and used  $\Delta\text{AICc}$  among the reduced model set to examine model fit to the data.

Because neither variation in rates, nor differences in phylogenetic means or selective optima, could account for the patterns observed (Fig. 3; see below for details), we focused on the strength of selection in relation to tree length, as expressed by the parameter  $\alpha$  of Equation (3) and the corresponding phylogenetic half-life ( $t_{1/2}$ ). Together with the evolutionary rate  $\sigma$ , the strength of the restraining force  $\alpha$  determines the expected covariance structure of the data at the tips of the phylogeny and as such it has a profound influence on our capacity of statistically distinguishing between different evolutionary models. To address this possibility, we

simulated 1000 data sets under each of the OU models examined above, but in this case setting the phylogenetic half-life parameter  $t_{1/2}$  to 1, 0.67, 0.5, 0.4, or 0.33. Lower values of  $t_{1/2}$  correspond to higher values of  $\alpha$ , a stronger restraining pull of the data toward one or more optimal trait values, and a more marked distinction in the expected covariance structure of the data from a BM model. All other simulating parameters were kept as described above. We then fit the six examined models to each of the simulated data sets and used  $\Delta\text{AICc}$  to examine model fit to the data.

#### *Variation in Evolutionary Rates, Means, and Optima*

From these simulations we found that for data simulated under a BM process, with increasing differences in either rates (BMS) or both rates and group means (BMSG), model parameters have little influence on relative model fit, at least within the parameter range examined here (Online Appendix 2, available on Dryad at <http://dx.doi.org/10.5061/dryad.2ss46>). This was not the case for the OU-based simulations. For data simulated with a single rate and increasing relative difference in optima (OUM), the fit of a BMSG and an OUM model was increasingly similar (Fig. 3a). That is, the higher the difference in optima between groups, the more difficult it is to distinguish between a model of random walk with different rates and means for each group (BMSG) and a model with a single evolutionary rate and different optima for each group (OUM). This suggests that when emphasis is put toward the mean structure of the data (in this case by increasing differentiation in simulated optima), the residual structures provided by the three models become increasingly similar. Indeed, as the difference in optima ( $\beta_2/\beta_1$ ) increases the rate parameters ( $\sigma$ ) estimated under each of the three models become more similar, such that the instances at which the BMSG and OUM models are indistinguishable are more and more frequent (Online Appendix 5, available on Dryad at <http://dx.doi.org/10.5061/dryad.2ss46>).

For data simulated under OU with two rates and two optima (OUMV), increasing the relative optima parameter ( $\beta_2/\beta_1$ ) eventually caused the BMS model to be visibly inappropriate for explaining the data (Fig. 3b). That is, when differences in selective optima between groups are very marked, a model that does not include any kind of mean differentiation is not sufficient for explaining phenotypic variation. This suggests that the relative fit of different models highly depends on how these partition phenotypic covariation among taxa into the mean and residual components. Interestingly, an increase in relative optima under OUMV generally resulted in a slightly higher performance for the BMSG model, when compared with the OUMV model (horizontal direction in Fig. 3b). By contrast, an increase in relative evolutionary rate ( $\theta$ ) had the opposite effect, enhancing the fit of OUMV when compared with BMSG (perpendicular direction in Fig. 3b). Putting both

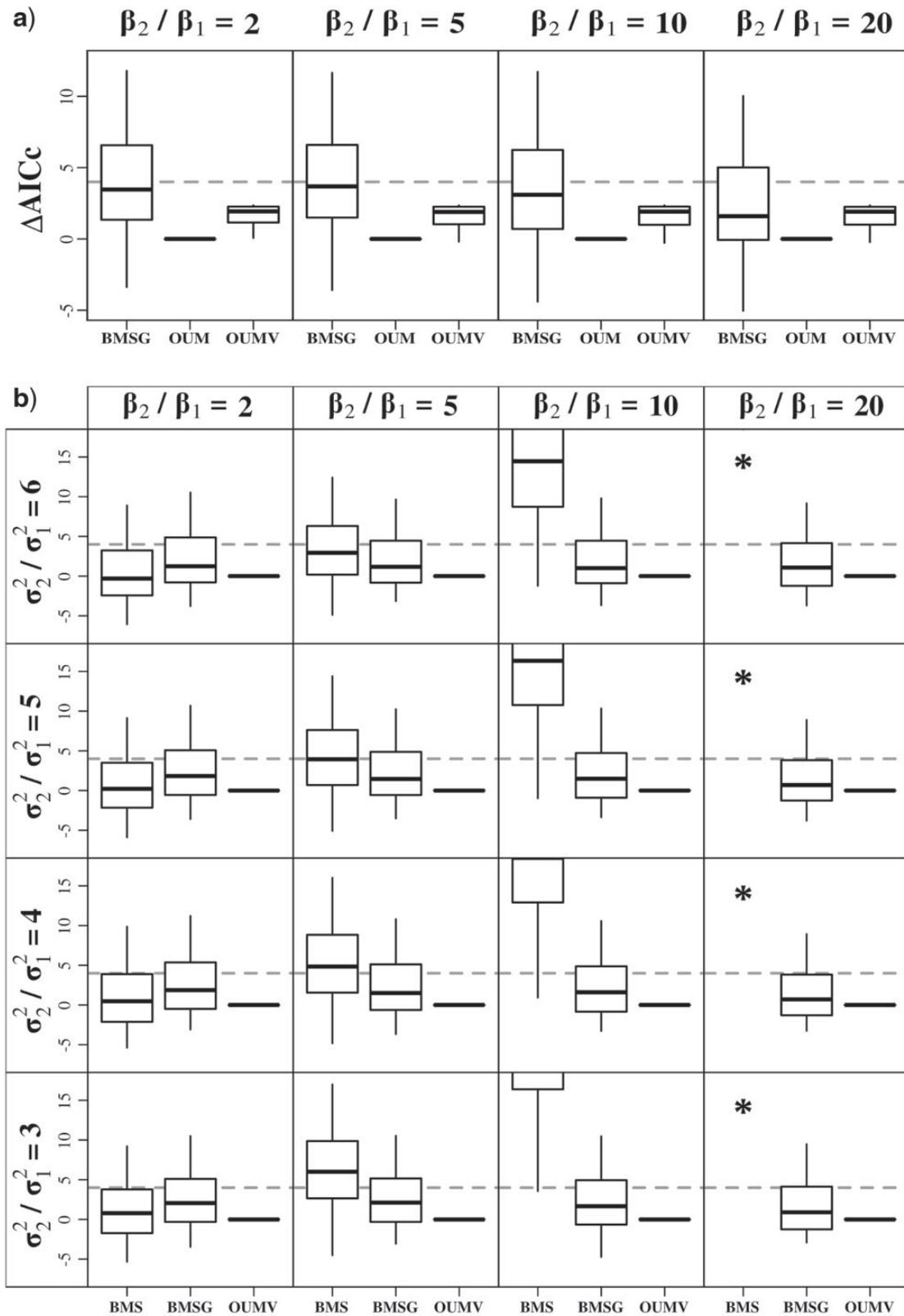


FIGURE 3. Quantile boxplots of  $\Delta AICc$  scores obtained by fitting the models that showed ambiguous  $\Delta AICc$  patterns in previous simulations (Fig. 2). In this case we simulated under OUM (a) and OUMV (b) models with varying relative rates ( $\theta$ ) and relative optima ( $\beta_2/\beta_1$ ).  $\Delta AICc$  have been standardized relative to the simulating model (“real” model underlying the data), which therefore always has  $\Delta AICc = 0$ . The dashed horizontal line represents the frequently used threshold of  $\Delta AICc = 4$ . BMS models marked with a star exhibited mean  $\Delta AICc$  scores above 4 and are off-scale in the presented graphs, to maintain the same scale in all graphs.

sources of variation together, and depending on the relative magnitude of the relative rate ( $\theta$ ) and relative optima ( $\beta_2/\beta_1$ ) parameters, simulations yield patterns of phenotypic variation where either BMSG or OUMV may be a better fit, regardless of the fact that the “real” model underlying the data is in fact OUMV. Given that both models model phenotypic variance allowing for a mean structure that includes different means for each group of taxa, this signifies that both models in fact provide similar estimates for the residual components of Equation (3) and converge toward similar covariance structures for the simulation conditions examined here. This suggests that the simulated selective pull ( $\alpha$  in Equation (3)) may not be strong enough for the simulated data to exhibit a covariance structure that is clearly different from what would occur under a BM evolutionary process.

#### *Variation in Phylogenetic Half-Life*

Simulations under varying values of  $\alpha$ , which represents the strength of selection, confirmed that this model parameter has a strong influence on the dynamics of OU models. This influence is directly reflected on our capacity for statistically distinguishing between models that represent changes in trait mean structure. Generally, lower values of alpha, which translate into increased phylogenetic half-life values, make distinguishing the evolutionary models examined here more challenging (Fig. 4). By contrast, as the restraining force represented by  $\alpha$  increases, BM models generally become less and less likely for explaining the data. Focusing on pairs of models that exhibited similar  $\Delta AICc$  scores in previous simulations, and which are not nested (i.e., OU1 vs. BMS; OUM vs. BMS; OUM vs. BMSG; and OUMV vs. BMSG), the results obtained here suggest that, in most cases, the real model underlying the data can be identified if the distinction between a BM and an OU process is sufficiently marked through a relatively strong selective influence.

#### DISCUSSION

Phylogenetic comparative models are a major tool used to investigate interspecific phenotypic patterns and enhance our understanding of the historical processes that have shaped patterns of phenotypic diversity. Our examination of empirical data indicates that in many circumstances uncertainties may emerge when attempting to distinguish between alternative evolutionary processes underlying phenotypic variation. When using simulations to examine models under known conditions, we found that one can accurately identify some basic characteristics of the evolutionary process (e.g., variation in rates). However, inferring the exact nature of the evolutionary process that has yielded phenotypic variation when using phylogenetic comparative modeling to access complex, potentially more realistic, mechanisms may be more

problematic. In these cases, model inference bears a higher ambiguity, where non-nested candidate models may appear equally plausible for explaining phenotypic variation when their goodness of fit to the data is considered. This has important practical and theoretical implications.

For empirical biologists seeking to explain the patterns of phenotypic variation observed in nature, the results obtained here should serve as a cautionary tale when contrasting the fit of mathematical models on phylogenies. Indeed, by conducting comparisons of alternative models that represent hypotheses about the causes underlying phenotypic evolution across 160 empirical data sets, we found that several pairs of models may frequently receive similar support (Fig. 1). Through simulation experiments we showed that this is not due to some distinctive property of the empirical data sets examined here. Instead, the same tendency was observed when using simulations to produce patterns of phenotypic variation under known evolutionary processes. Thus, a first caution to be taken from these findings is related to the set of candidate models chosen. Because, as shown here, model comparisons can frequently yield ambiguous results, researchers will ensure stronger evolutionary inference by *a priori* limiting the set of candidate models based on previous knowledge on the biological system under examination (Burnham and Anderson 2002). Another frequent recommendation has been that, when in doubt (in terms of model selection criteria), simpler evolutionary models should be preferred, as a reduced number of model parameters can be estimated more accurately (Butler and King 2004; Beaulieu et al. 2012; Ho and Ané 2014). Our findings support this recommendation for the case of examining pairs of nested models. In such circumstances, lack of statistical distinctiveness does not translate into an evolutionary ambiguity. Specifically, when nested pairs of models are contrasted, similarity in AIC scores, or a small difference between them, may actually be the result of very similar model parameters (e.g., Hunt 2006; see also Online Appendix 5, available on Dryad at <http://dx.doi.org/10.5061/dryad.2ss46>) and the evolutionary interpretation of the data is straightforward.

In other circumstances however, this recommendation does not hold. Among the models we compared, those encompassing different evolutionary rates under BM (BMS and BMSG) and evolution toward different optima with a single (OUM) or multiple rates (OUMV) were particularly problematic. Indeed, pairs of models of these four types were frequently indistinguishable in terms of fit to the data, yielding the inference of evolutionary tempo and mode ambiguous (Figs. 1 and 2). These pairs of models are not nested, and therefore the lack of statistical distinction between them is problematic in terms of evolutionary interpretation. Technically, this is particularly relevant for pairs of non-nested models that also have the same number of parameters. This is the case, for instance, with the BMS–OU1 and BMSG–OUM pairs of models, with three and four parameters,

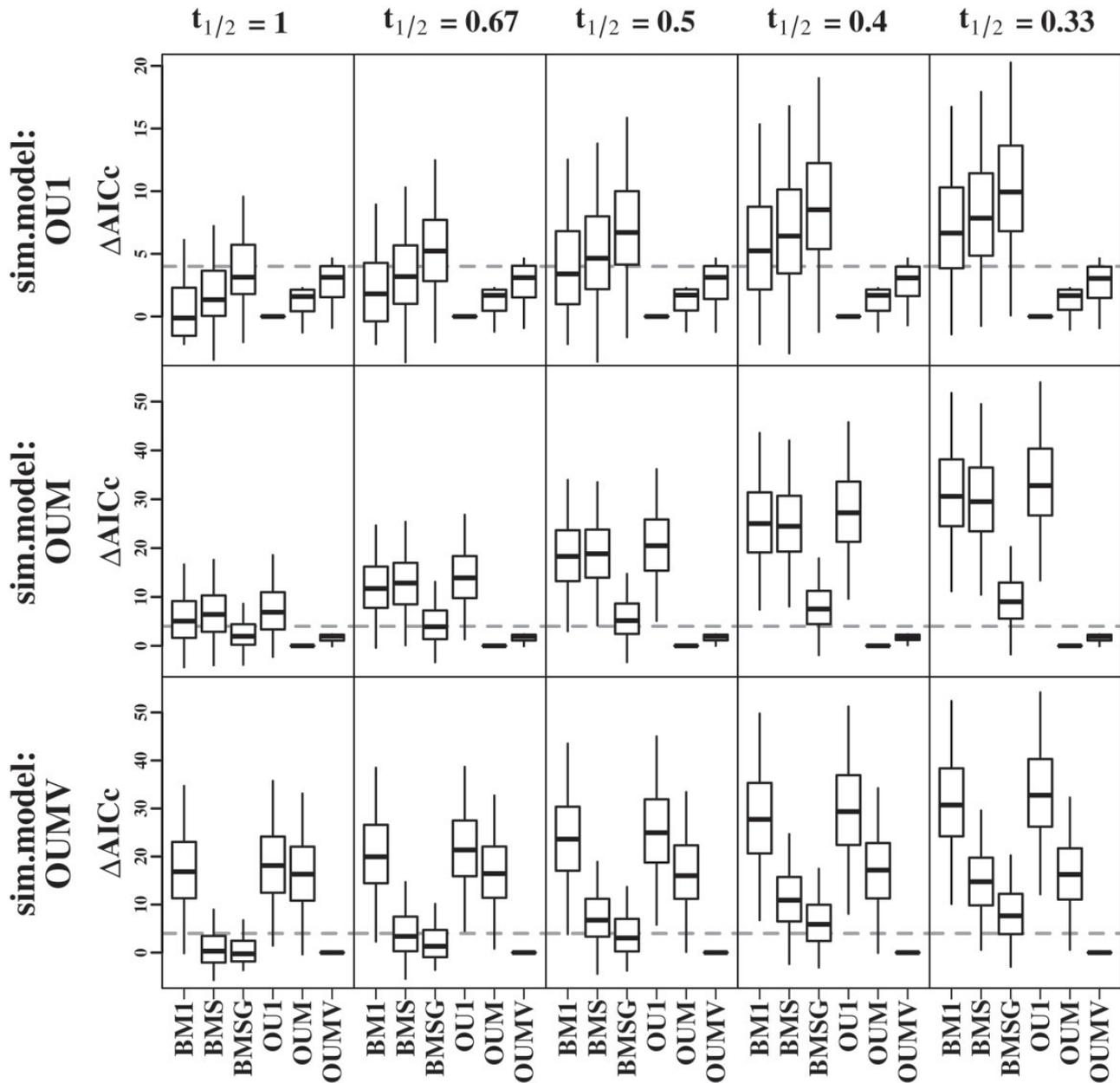


FIGURE 4. Quantile boxplots of  $\Delta\text{AICc}$  scores obtained by simulating 1000 data sets under the three examined OU models (sim.model), while varying the strength of selection  $\alpha$  and consequently the value of phylogenetic half-life.  $\Delta\text{AICc}$  have been standardized relative to the simulating model ("real" model underlying the data), which therefore always has  $\Delta\text{AICc}=0$ . The dashed horizontal line represents the frequently used threshold of  $\Delta\text{AICc} = 4$ .

respectively (Figs. 3 and 4d,e). This occurs because these models modify different pieces of Equation (3), containing either an additional rate parameter or an additional optimum parameter, resulting in the same total number of estimated parameters. In these circumstances the recommendation of choosing the simpler model is not applicable.

Even with models that do differ in the number of parameters, however, choosing the simpler model is not always straightforward, as the models compared here represent radically different evolutionary processes and would lead to very different biological interpretations.

As such, it is important to understand why they may exhibit similar fits to the data. Simulations under varying model parameters provide some insights to that respect. When phenotypic data were simulated under a diversifying evolutionary process with a single rate (OUM), an increase in simulated relative optima augmented the overlap, in terms of goodness of fit, between the simulating model and a Brownian model with two rates and different phylogenetic means (BMSG; Fig. 3a). This suggests that both models converge by allowing for different mean structures in each group and reach similar parameter estimates for the residual

structure (Online Appendix 5, available on Dryad at <http://dx.doi.org/10.5061/dryad.2ss46>). Similarly, for data simulated under a diversifying evolutionary process with different rates for the two groups (OUMV) a Brownian model with different phylogenetic means (BMSG), and a model of evolution toward two optima with two evolutionary rates (OUMV) exhibit similar fits to the data (Fig. 3b). These results, together with the similarity of the model parameters estimated under different models, suggest that the residual structure of OU-simulated data is not sufficiently different from a Brownian process, causing the conflation observed.

Indeed, simulations under varying values of phylogenetic half-life indicate that, when an OU process underlies the data, the strength of selection is critical for accurately assessing alternative models (Fig. 4). For progressively lower values of  $\alpha$  relative to tree length, which translate into progressively higher phylogenetic half-life parameters, OU dynamics are not sufficiently different from a BM process, and the parameters estimated under both types of model are essentially the same (Online Appendix 5, available on Dryad at <http://dx.doi.org/10.5061/dryad.2ss46>), resulting in similar expected covariance structures. Importantly, for relatively low values of  $\alpha$  (Fig. 4), this may be interpreted as a conflation of tempo and mode, in the sense that a BM model with rate differentiation between groups (BMSG) and an OU model with moderate rate differentiation and a rubber-band component acting on trait mean value (OUMV) could be equally plausible for explaining phenotypic patterns. In these circumstances, however, the selective pull driving phenotypic evolution is not very strong, and both models converge toward similar estimates, essentially representing the same evolutionary process. Two observations are of relevance here: first, empiricists should be able to judge, based on a good knowledge of their model system, how strong the estimated pull of selection on the examined phenotypes is. Phylogenetic half-life is known to vary extensively across different traits and study organisms (Hansen 2012), such that a close examination of data set-specific estimated model parameters is necessary for understanding whether different candidate models would actually lead to different evolutionary interpretations. Second, it is important to note that the exact biological significance of the BMSG model remains somewhat obscure. Both O'Meara et al. (2006) and Thomas et al. (2006) have suggested, although in different ways, that this model encompasses a quick shift in mean trait value at the base of each subtree evolving under different rates, yielding the different phylogenetic means from which phenotypic variation is modeled under this type of BM model. This kind of quick shift could be further explored and confirmed by modeling a shift in evolutionary rates *specifically* on the branch leading to these differentiated phylogenetic means of each rate group (Revell 2008). In terms of covariance structure, however, the dynamics of this model are probably not those typically represented in BM models, which may be contributing to the

frequent resemblance of BMSG and OU models. This effect seems to be quite important for the simulation experiments conducted here under trees with 64 taxa, but it may be alleviated in larger data sets, which would provide more power for distinguishing the fit of different models.

On the other hand, the comparison of models under different evolutionary scenarios conducted here also allows us to track the circumstances under which model selection is a powerful tool for inferring how phenotypic evolution has proceeded. Specifically, when phenotypic data are generated by an evolutionary process that encompasses variation across groups only in evolutionary rates or means, model selection promptly detects such differentiation. Simulations under a BM process with different rates for each group resulted in a markedly lower performance of models not encompassing rate differences (Fig. 2b). This pattern may be associated to the observation that evolutionary rate parameters are generally easier to estimate accurately (Boettiger et al. 2012; Ho and Ané 2014). Similarly, simulations under an OU process where the two groups were driven toward different optima showed that models that do not consider any type of mean differentiation are a poor fit to the data (Fig. 2e). These results indicate that, when testing hypotheses encompassing changes in single model parameters, the comparison of phylogenetic models is a useful tool for tracing phenotypic evolution. The same is the case when faced with evolutionary processes dominated by a strong selective influence, where phenotypes are promptly driven toward one or more optima. In such circumstances, evolutionary dynamics are determined by a visible dominance of selection ( $\alpha$ ) over drift ( $\sigma$ ) and model comparison efficiently identifies the evolutionary tempo and mode underlying the data.

Another related aspect that has been recently emphasized by other authors (e.g., Beaulieu et al. 2012) and which is further reinforced by the results presented here is that model selection alone is rarely conclusive for answering a scientific question. Models can only provide the best hypothesis for explaining patterns in data, and as such they are only the first step of exploring what biological factors have acted and how (Losos 2011). In the context of evolutionary inference examining phenotypic traits on phylogenies, a powerful toolkit exists today for testing many different hypotheses which should be used to obtain multiple lines of evidence for the same question. The evaluation of OU models without *a priori* defining selective regimes (Ingram and Mahler 2013; Uyeda and Harmon 2014), the delimitation of the parameter space through the definition of biologically informed priors (Uyeda and Harmon 2014), or the quantification of statistical model adequacy for explaining phenotypic variation (Pennell et al. 2015) are definitely promising steps in this direction. This also brings into focus the need for models that consider potential variation in mode across time, or in single branches of the phylogeny. While temporal variation in evolutionary tempo has been extensively

considered (e.g., Pagel 1998; Blomberg et al. 2003; Harmon et al. 2010), implementations of evolutionary models that allow for variations in mode across time have only been introduced very recently (Slater 2013). Such variation would also be potentially relevant for the study of evolutionary tempo. Indeed, variation of the mode of evolution across phylogenetic time is known to entangle the estimation of evolutionary rate parameters in a paleontological framework (Hunt 2012). Given the conceptual similarities between the observations of Hunt (2012) regarding paleontological inference and the conclusions drawn here with respect to using phylogenies and the comparative method to explore evolutionary tempo and mode, we expect the same to occur in phylogenetic comparative models.

Finally, it is important to remark that the conflation of evolutionary tempo and mode observed under certain conditions is, in our view, not only a methodological issue associated to model selection. It is mainly a central element of how we perceive evolutionary processes through both paleontological and modern phylogenetic comparative methods. In the phylogenetic framework, model parameters such as the evolutionary rate  $\sigma^2$  and the phylogenetic half-life  $t_{1/2}$  are estimated and compared. Some of them are historically more associated to the words *tempo*, in the case of evolutionary rates, or *mode*, in the case of the  $\alpha$  parameter of OU models. However, other models—not examined here—consider variations in evolutionary rates to actually test hypotheses about tempo. The Early Burst model of fast phenotypic evolution early during diversification followed by slower evolutionary rates used to test for adaptive radiation (Harmon et al. 2010), or tests of gradualism versus punctuated equilibria (Pagel 1999) nicely illustrate how evolutionary tempo and mode are conceptually inseparable. The links between these important evolutionary concepts and the mathematical models used to approach them are not, however, always straightforward. The association of model parameters to the terms “tempo” and “mode” bears some ambiguity, and more work is needed to clarify how phylogenetic comparative models can be adequately used to describe the evolutionary process.

#### SUPPLEMENTARY MATERIAL

Data available from the Dryad Digital Repository: <http://dx.doi.org/10.5061/dryad.2ss46>.

#### FUNDING

A.K. was supported by a post-doctoral grant (SFRH/BPD/68493/2010) and an IF investigator position by Fundação para a Ciência e Tecnologia (FCT, Portugal). This work was supported in part by the NSF grant DEB-1257287 [to D.C.A.] and by Project “Biodiversity, Ecology and Global Change” co-financed by North Portugal Regional Operational Programme 2007/2013 (ON.2—O Novo Norte), under the National

Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF).

#### ACKNOWLEDGEMENTS

The authors thank L. Harmon, G. Hunt, F. Anderson, B. O’Meara, G. Thomas, and three anonymous reviewers for useful comments on previous versions of the manuscript. G. Thomas kindly provided an updated version of the *transformPhylo.sim* function of motmot R-package. C. Stack provided valuable help with the R-package RBrownie. The following people kindly provided data files from empirical studies: T. Barkman, D. Collar, A. Corl, W. Cooper, L. M. Dávalos, E. R. Dumont, E. Edwards, L. Harmon, D. Hulsey, C. Martin, D. Moen, S. Price, P. Raia, G. Slater, D. Swanson, and J. Wiens.

#### REFERENCES

- Adams D.C. 2010. Parallel evolution of character displacement driven by competitive selection in terrestrial salamanders. *BMC Evol. Biol.* 10(72):1–10.
- Adams D.C. 2013. Comparing evolutionary rates for different phenotypic traits on a phylogeny using likelihood. *Syst. Biol.* 62:181–192.
- Adams D.C. 2014. Quantifying and comparing phylogenetic evolutionary rates for shape and other high-dimensional phenotypic data. *Syst. Biol.* 63:166–177.
- Adams D.C., Berns C.M., Kozak K.H., Wiens J.J. 2009. Are rates of species diversification correlated with rates of morphological evolution? *Proc. R. Soc. B Biol. Sci.* 276:2729–2738.
- Barkman T.J., Bendiksyby M., Lim S.H., Salleh K.M., Nais J., Madulid D., Schumacher T. 2008. Accelerated rates of floral evolution at the upper size limit for flowers. *Cur. Biol.* 18:1508–1513.
- 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.
- Blomberg S.P., Garland T. Jr., Ives I.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.
- Bokma F. 2002. Detection of punctuated equilibrium from molecular phylogenies. *J. Evol. Biol.* 15:1048–1056.
- Burnham K.P., Anderson D.R. 2002. Model selection and multi-model inference: A practical information-theoretic approach. New York: Springer-Verlag.
- Butler M.A., King A.A. 2004. Phylogenetic comparative analysis?: A modeling approach for adaptive evolution. *Am. Nat.* 164:683–695.
- Chung K., Hipp A.L., Roalson E.H. 2012. Chromosome number evolves independently of genome size in a clade with nonlocalized centromeres (*Carex*: Cyperaceae). *Evolution* 66:2708–2722.
- Collar D.C., Wainwright P.C. 2006. Discordance between morphological and mechanical diversity in the feeding mechanism of centrarchid fishes. *Evolution* 60:2575–2584.
- Collar D.C., O’Meara B.C., Wainwright P.C., Near T.J. 2009. Piscivory limits diversification of feeding morphology in centrarchid fishes. *Evolution* 63:1557–1573.
- Collar D.C., Schulte J.A., Losos J.B. 2011. Evolution of extreme body size disparity in monitor lizards (*Varanus*). *Evolution* 65:2664–2680.
- Dumont E.R., Dávalos L.M., Goldberg A., Santana S.E., Rex K., Voigt C.C. 2011. Morphological innovation, diversification and invasion of a new adaptive zone. *Proc. R. Soc. B Biol. Sci.* 279:1797–1805.
- Eastman J.M., Alfaro M.E., Joyce P., Hipp A.L. and Harmon L.J. 2011. A novel comparative method for identifying shifts in the rate of character evolution on trees. *Evolution* 65:3578–3589.

- Edwards A.W.F., Cavalli-Sforza L.L. 1964. Reconstruction of evolutionary trees. In: Heywood V.H., McNeill J., editors. Phenetic and phylogenetic classification. London: Systematics Association. p. 67–76.
- Edwards E.J., Still C.J. 2008. Climate, phylogeny and the ecological distribution of C4 grasses. *Ecol. Lett.* 11:266–276.
- Felsenstein J. 1973. Maximum likelihood estimation of evolutionary trees from continuous characters. *Am. J. Hum. Gen.* 25:471–492.
- Felsenstein J. 1985. Phylogenies and the comparative method. *Am. Nat.* 125:1–15.
- Felsenstein J. 1988. Phylogenies and quantitative characters. *Ann. Rev. Ecol. Syst.* 19:445–471.
- Fitch W.M., Ayala F.J. 1994. Tempo and mode in evolution. *Proc. Natl. Acad. Sci. U. S. A.* 91:6717–6720.
- Garland T. Jr. 1992. Rate tests for phenotypic evolution using phylogenetically independent contrasts. *Am. Nat.* 140:509–519.
- Gingerich P.D. 1976. Paleontology and phylogeny: Patterns of evolution at the species level in early Tertiary mammals. *Am. J. Sci.* 276:1–28.
- Glor R.E. 2010. Phylogenetic insights on adaptive radiation. *Ann. Rev. Ecol. Evol. Syst.* 41:251–270.
- Gould S.J. 1980. G. G. Simpson, Paleontology and the modern synthesis. In: Mayr E., Provine W.B., editors. The evolutionary synthesis. Cambridge (MA): Harvard University Press. p. 153–172.
- Gould S.J., Eldredge N. 1977. Punctuated equilibria: The tempo and mode of evolution reconsidered. *Paleobiology* 3:115–151.
- Grafen A. 1989. The phylogenetic regression. *Philos. Trans. R. Soc. B* 326:119–157.
- Hansen T.F. 1997. Stabilizing selection and the comparative analysis of adaptation. *Evolution* 51:1341–1351.
- Hansen T.F. 2012. Adaptive landscapes and macroevolutionary dynamics. In: Svensson E., Calsbeek R., editors. The adaptive landscape in evolutionary biology. Oxford: Oxford University Press. p. 205–226.
- Harmon L.J., Kolbe J.J., Cheverud J.M., Losos J.B. 2005. Convergence and the multidimensional niche. *Evolution* 59:409–421.
- Harmon L.J., Melville J., Larson A., Losos J.B. 2008. The role of geography and ecological opportunity in the diversification of day geckos (*Phelsuma*). *Syst. Biol.* 57:562–573.
- 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., Brock C., Glor R., Challenger W., Hunt G. 2009. geiger: Analysis of evolutionary diversification. R package version 1.3-1. Available from: URL <http://CRAN.R-project.org/package=geiger>.
- Harmon L.J., Losos J.B., Davies T.J., Gillespie R.G., Gittleman J.L., Jennings W.B., Kozak K.H., McPeck M.A., Moreno-Roark F., Near T.J., Purvis A., Ricklefs R.E., Schluter D., Schulte J.A. II, 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.
- Harvey P.H., Pagel M.D. 1991. The comparative method in evolutionary biology. Oxford: Oxford University Press.
- Harvey P.H., Rambaut A. 2000. Comparative analyses for adaptive radiations. *Philos. Trans. R. Soc. B* 355:1599–1605.
- Hipp A.L. 2007. Nonuniform processes of chromosome evolution in sedges (*Carex*: Cyperaceae). *Evolution* 61:2175–2194.
- Ho L.S.T., Ané C. 2014. Intrinsic inference difficulties for trait evolution with Ornstein–Uhlenbeck models. *Methods Ecol. Evol.* 5: 1133–1146.
- Hulsey C.D., Mims M.C., Parnell N.F., Strelman J.T. 2010. Comparative rates of lower jaw diversification in cichlid adaptive radiations. *J. Evol. Biol.* 23:1456–1467.
- Hunt G. 2006. Fitting and comparing models of phyletic evolution: Random walks and beyond. *Paleobiology* 32:578–601.
- Hunt G. 2012. Measuring rates of phenotypic evolution and the inseparability of tempo and mode. *Paleobiology* 38:351–373.
- Ingram T., Mahler D.L. 2013. SURFACE: Detecting convergent evolution from comparative data by fitting Ornstein–Uhlenbeck models with stepwise Akaike Information Criterion. *Methods Ecol. Evol.* 4:416–425.
- Losos J.B. 2011. Seeing the forest for the trees: the limitations of phylogenies in comparative biology. *Am. Nat.* 177:709–727.
- Martin C.H., Wainwright P.C. 2011. Trophic novelty is linked to exceptional rates of morphological diversification in two adaptive radiations of *Cyprinodon* pupfish. *Evolution* 65:2197–2212.
- Martins E.P., Garland T. Jr. 1991. Phylogenetic analyses of the correlated evolution of continuous characters: A simulation study. *Evolution* 45:534–557.
- McPeck M.A. 1995. Testing hypotheses about evolutionary change on single branches of a phylogeny using evolutionary contrasts. *Am. Nat.* 145:686–703.
- Moen D.S., Wiens J.J. 2009. Phylogenetic evidence for competitively driven divergence: Body-size evolution in Caribbean treefrogs (Hylidae: *Osteopilus*). *Evolution* 63:195–214.
- 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. 1998. Inferring evolutionary processes from phylogenies. *Zool. Scr.* 26:331–348.
- Pagel M. 1999. Inferring the historical patterns of biological evolution. *Nature* 401:877–884.
- Paradis E., Claude J., Strimmer K. 2004. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* 20: 289–290.
- Pennell M.W., FitzJohn R.G., Cornwell W.K., Harmon L.J. 2015. Model adequacy and the macroevolution of angiosperm functional traits. *Am. Nat.* 186.
- Price S.A., Holzman R., Near T.J., Wainwright P.C. 2011. Coral reefs promote the evolution of morphological diversity and ecological novelty in labrid fishes. *Ecol. Lett.* 14:462–469.
- Price S.A., Wainwright P.C., Bellwood D.R., Kazancioglu E., Collar D.C., Near T.J. 2010. Functional innovations and morphological diversification in parrotfishes. *Evolution* 64:3057–3068.
- R Development Core Team. 2012. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available from: URL <http://www.R-project.org>.
- Rabosky D.L., Adams D.C. 2012. Rates of morphological evolution are correlated with species richness in salamanders. *Evolution* 66: 1807–1818.
- Raia P., Meiri S. 2011. The tempo and mode of evolution: Body sizes of island mammals. *Evolution* 65:1927–1934.
- Revell L.J. 2008. On the analysis of evolutionary change along single branches in a phylogeny. *Am. Nat.* 172:140–147.
- Revell L.J. 2012. phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* 3:217–223.
- Revell L.J., Collar D.C. 2009. Phylogenetic analysis of the evolutionary correlation using likelihood. *Evolution* 63:1090–1100.
- Revell L.J., Reynolds R.G. 2012. A new Bayesian method for fitting evolutionary models to comparative data with intraspecific variation. *Evolution* 66:2697–2707.
- Revell L.J., Johnson M.A., Schulte J.A. II, Kolbe J.J., Losos J.B. 2007. A phylogenetic test for adaptive convergence in rock-dwelling lizards. *Evolution* 61:2898–2972.
- Revell L.J., Mahler D.L., Peres-Neto P.R., Redelings B.D. 2011. A new phylogenetic method for identifying exceptional phenotypic diversification. *Evolution* 66:135–146.
- Ricklefs R.E. 2004. Cladogenesis and morphological diversification in passerine birds. *Nature* 430:338–341.
- Rohlf F.J. 2001. Comparative methods for the analysis of continuous variables: Geometric interpretations. *Evolution* 55: 2143–2160.
- Schliep K.P. 2011. phangorn: Phylogenetic analysis in R. *Bioinformatics* 27:592–593.
- Setiadi M.I., McGuire J.A., Brown R.M., Zubairi M., Iskandar D.T., Andayani N., Supriatna J., Evans B.J. 2011. Adaptive radiation and ecological opportunity in Sulawesi and Philippine fanged frog (*Limnonectes*) communities. *Am. Nat.* 178:221–240.
- Simpson G.G. 1944. Tempo and mode in evolution. New York: 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. *Methods Ecol. Evol.* 4:734–744.
- 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.

- Stayton C.T. 2006. Testing hypotheses of convergence with multivariate data: Morphological and functional convergence among herbivorous lizards. *Evolution* 60:824–841.
- Swanson D.L., Bozinovic F. 2012. Metabolic capacity and the evolution of biogeographic patterns in oscine and suboscine passerine birds. *Physiol. Biochem. Zool.* 84:185–194.
- Swanson D.L., Garland T. Jr. 2009. The evolution of high summit metabolism and cold tolerance in birds and its impact on present-day distributions. *Evolution* 63:184–194.
- Thomas G.A., Freckleton R.P. 2012. MOTMOT: Models of trait macroevolution on trees. *Methods Ecol. Evol.* 3: 145–151.
- 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. Soc. B Biol. Sci.* 273:1619–1624.
- Thomas G.H., Meiri S., Phillimore A.B. 2009. Body size diversification in *Anolis*: Novel environment and island effects. *Evolution* 63: 2017–2030.
- Uyeda J.C., Harmon L.J. 2014. A novel Bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data. *Syst. Biol.* 63:902–918.
- Wiens J.J., Pyron R.A., Moen D.S. 2011. Phylogenetic origins of local-scale diversity patterns and the causes of Amazonian megadiversity. *Ecol. Lett.* 14:643–652.