ARTICLE

Running head: ADEQUACY OF PHYLOGENETIC TRAIT MODELS

Model adequacy and the macroevolution of angiosperm functional traits

Matthew W. Pennell[1,*], Richard G. FitzJohn[2], William K. Cornwell[3] & Luke J. Harmon[1,4]

[1] Department of Biological Sciences & Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, ID 83844, U.S.A.

[*] Email for correspondence: mwpennell@gmail.com

[2] Department of Biological Sciences, Macquarie University, Sydney, NSW 2109, Australia; rich.fitzjohn@gmail.com

[3] School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW 2052, Australia; w.cornwell@unsw.edu.au

[4] lukeh@uidaho.edu

Keywords: phylogenetic comparative methods, model adequacy, independent contrasts, Angiosperm functional traits

Contents of supplementary material

Abstract

All models are wrong and sometimes even the best of a set of models is useless. Modern phylogenetic comparative methods (PCMs) are almost exclusively model–based and therefore making robust inferences from PCMs requires using a model of trait evolution that is a good explanation for the data. To date, researchers using PCMs have evaluated the explanatory power of a model only in terms of relative, not absolute, fit. Here we develop a general statistical framework for assessing the absolute fit, or adequacy, of phylogenetic models for the evolution of quantitative traits. We use our approach to test whether commonly used models are adequate descriptors of the macroevolutionary dynamics of real comparative data. We fit models of trait evolution to 337 comparative datasets covering three key Angiosperm functional traits and evaluated the absolute fit of the models to each dataset. Overall, the models we used are very inadequate for the evolution of these traits; this was true for many different groups and at many different scales. Furthermore, the relative support for a model had very little to do with its absolute adequacy. We argue that assessing model adequacy should be a key step in comparative analyses.

> There are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns — the ones we don't know we don't know.
>
> — Former U.S. Secretary of Defense, Donald Rumsfeld

Introduction

Phylogenetic trees and phylogenetic thinking are ubiquitous in modern biology; ecologists, geneticists, paleontologists and anthropologists now widely recognize that a historical perspective can provide important insights into evolutionary questions (Pennell and Harmon, 2013). This interest in phylogenetic biology has risen in lockstep with developments in phylogenetic comparative methods (PCMs; reviewed in O'Meara, 2012; Pennell and Harmon, 2013). Modern PCMs are almost exclusively based on probabilistic models, meaning inferences are conditional on both the phylogenetic tree and the chosen model. Selecting a good model is therefore essential for making robust inferences. Model choice should be guided by two considerations. First, does the model capture the processes relevant to the question (Hansen and Orzack, 2005; Maddison, 2006; Hansen and Bartoszek, 2012; Pennell et al., 2013)? Second, does the model provide a good statistical explanation for the data? Though there is interplay between these two questions (Hansen and Bartoszek, 2012), we focus on the latter in this paper. In phylogenetic comparative biology, this question has been addressed by comparing the relative fit of a number of models and selecting the best one for inference (Mooers et al., 1999; Harmon et al., 2010;

Hunt, 2012).

However, the relative support for a model is only a partial answer to the question of whether it is a good explanation for the data, as the best of a poor set of models is still a poor model. We also must consider whether the model is a good fit, or adequate, in absolute terms. Assessing model adequacy is a routine step in many statistical applications, especially in Bayesian statistics (Gelman et al., 2003). Failure to assess model adequacy can potentially result in erroneous inferences. A classic example of this in molecular systematics is the disputed phylogenetic relationships of rodents. A number of early molecular phylogenetic studies reported evidence that "strongly contradict[ed]" the traditional hypothesis that the order Rodentia is a monophyletic group (Graur et al., 1991; D'Erchia et al., 1996). Sullivan and Swofford (1997) demonstrated that these conclusions were misleading, and resulted entirely from using models of molecular evolution that did not adequately accommodate variation in substitution rates across sites (also see Brown, 2014). There are many similar case–studies throughout and outside of evolutionary biology.

Many models for describing the evolution of phenotypic traits along a phylogeny have been developed, as well as sophisticated machinery for fitting them to comparative data (O'Meara, 2012). Using these models, researchers from across biological disciplines have made discoveries that would not have been possible without a phylogenetic perspective. However, it is disconcerting that we often have no idea if the models used in PCMs are adequate — there are no general procedures that can be used to assess this. Even more troubling is that model adequacy is rarely ever considered when PCMs are applied. To borrow Rumsfeld's taxonomy, model adequacy has largely been an "unknown unknown" in comparative biol-

ogy.

In this paper we seek to help rectify this situation. We address two major outstanding problems in comparative biology. First, we develop a general framework for assessing the adequacy of phylogenetic models of trait evolution, specifically those for quantitative characters. Second, we assess the adequacy of commonly used models using a recently published phylogeny of Angiosperms (flowering plants) (Zanne et al., 2013) and data compiled from the literature on three important functional traits: specific leaf area, seed mass, and leaf nitrogen content. We did this to test whether the simple models used in comparative biology provide a reasonable description of the macroevolutionary dynamics of real trait data, across many different groups and time scales.

A general framework for assessing model adequacy

Though the number of models used in comparative biology is quite large, most of these fall into a relatively small set of classes (O'Meara, 2012). In this paper, we focus on one of these classes — models that describe the evolution of a single continuously valued trait. More specifically, our approach works for models that assume that trait values at the tips are multivariate normal; this applies to most models of quantitative trait evolution that have been developed to date (O'Meara, 2012).

If we have a phylogenetic tree consisting of n lineages and data on the trait values observed at each tip X ($X = x_1, x_2, \ldots, x_n$), we can fit a model $\mathcal{M}$ with parameters θ to describe the pattern of trait evolution along the phylogeny. Most analyses using comparative data aim to answer one of the

following questions: what values of θ best explain X given $\mathcal{M}$?; or, does $\mathcal{M}_1$ explain the data better than $\mathcal{M}_0$? Our approach is conceptually distinct in that we want to ask, how likely is it that model $\mathcal{M}$ with parameters θ would produce a dataset similar to X if we re–ran evolution?

There are two primary ways of fitting models to comparative data. The first is use maximum likelihood (ML), restricted maximum likelihood (REML), least–squares, etc. to obtain a point estimate of θ. The second is to estimate posterior probability distribution $\Pr(\theta|X, \mathcal{M})$ using Bayesian approaches. For the models used in comparative biology, estimating $\Pr(\theta|X, \mathcal{M})$ requires using Markov chain Monte Carlo (MCMC) machinery to sample values of θ.

While likelihood and Bayesian approaches to model–fitting are philosophically different from one another, in practice, our approach to assessing model adequacy is much the same for both: 1) fit the model of trait evolution; 2) calculate a set of summary statitics on the observed data $\mathcal{S}_X$; 3) simulate many new datasets $Y_1, Y_2, \ldots, Y_m$ under the model using the estimated parameters; 4) calculate summary statistics on the simulated data $\mathcal{S}_{Y,1}, \mathcal{S}_{Y,2}, \ldots, \mathcal{S}_{Y,m}$; 5) compare $\mathcal{S}_X$ to the distribution of $\mathcal{S}_Y$. If $\mathcal{S}_X$ deviates significantly from the distribution of $\mathcal{S}_Y$, we can reject the model as inadequate (see figure 1).

If we have a point estimate of the model parameters $\hat{\theta}$, we simulate $Y_1, Y_2, \ldots, Y_m$ on the phylogeny accoridng to $\hat{\theta}$ and $\mathcal{M}$. We then compare a single set of summary statistics $\mathcal{S}_X$ calculated from our observed data to the distribution of values for $\mathcal{S}_Y$ computed across all m simulated datasets. In statistical terminology, this procedure is known as parametric bootstrapping. Parametric bootstrapping is likely familiar to phylogenetic biologists

in the form of the Goldman–Cox test (Goldman, 1993) for assessing the adequacy of sequence evolution models.

If we have a posterior probability distribution $\Pr(\theta|X,\mathcal{M})$, we can assess model adequacy using posterior predictive simulation (Rubin, 1984; Gelman et al., 1996). We obtain new datasets by sampling from a second distribution, the posterior predictive distribution

$$\Pr(Y|X,\mathcal{M}) = \int \Pr(Y|\theta,\mathcal{M}) \Pr(\theta|X,\mathcal{M}) d\theta \quad (1)$$

where $\Pr(Y|X,\mathcal{M})$ is the probability of a new dataset Y given X and $\mathcal{M}$, averaged over the distribution of the parameters. In practice, sampling from $\Pr(Y|X,\mathcal{M})$ entails drawing samples from the joint posterior distribution of parameters $\Pr(\theta|X,\mathcal{M})$ and simulating data on the phylogeny according to the sampled parameter values. Therefore, the datasets $Y_1, Y_2, \ldots, Y_m$ are each generated from different values of θ. Posterior predictive simulation approaches have been previously developed for models in molecular phylogenetics (Bollback, 2002; Reid et al., 2013; Lewis et al., 2013; Brown, 2014), and recently for PCMs (Slater and Pennell, 2013), but have not been widely adopted in either field.

If the chosen summary statistics were properties of the data alone, such as the mean of observed trait values at the tips, we would have a single estimate of $\mathcal{S}_X$ for both likelihood and Bayesian approaches. However, in our approach for assessing model adequacy, the summary statistics depend on both the data and the parameter estimates, for reasons detailed below. Therefore for the likelihood case, we have a single set of observed summary statistics $\mathcal{S}_X$ based on the data and the point estimate of model parameters $\hat{\theta}$. For the Bayesian approach we have a distribution of observed summary

statistics $\mathcal{S}_{X,1}, \mathcal{S}_{X,2}, \ldots, \mathcal{S}_{X,m}$, each $\mathcal{S}_{X,i}$ calculated using the same data but a different set of parameter values sampled from $\Pr(\theta|X, \mathcal{M})$. In this case, we compare the distribution of values of $\mathcal{S}_X$ to the distribution of values of $\mathcal{S}_Y$.

Summary statistics

No simulated dataset will ever be exactly the same as our observed dataset. We therefore need to choose informative summary statistics in order to evaluate whether the model predicts datasets that are similar to our observed dataset in meaningful ways. As an example, consider the case of regular (i.e., non–phylogenetic) logistic regression. The goal is to fit a model that predicts the state (0 or 1) of the dependent variable. We can assess the adequacy of such a model by simulating many datasets under the fitted parameters. One good summary statistic for this type of model is the proportion of values that are in state 0. We can calculate this directly from the observed data and from each of the simulated datasets and compare the observed value of this summary statistic to the distribution of values from the simulated datasets. This works because in a simple logistic regression model, we assume that the different values for the responses are independent of each other. However, the states at the tips of the phylogeny are not independent — this is why we are using PCMs in the first place! — and thus calculating summary statistics on the data directly is not generally informative for models in comparative biology.

We account for the non–independence of the observed data by calculating summary statistics on the set of contrasts (i.e., "phylogenetically independent contrasts", *sensu* Felsenstein, 1985) computed at each node.

(We refer readers to Felsenstein, 1985; Rohlf, 2001; Blomberg et al., 2012, for details on how contrasts are calculated.) Under Brownian motion (BM) the contrasts will be i.i.d. $\sim \mathcal{N}(0, \sigma)$, where σ^2 is the BM rate parameter (Felsenstein, 1985). This i.i.d. condition allows us to perform standard statistical tests on the contrasts.

The choice of what summary statistics to use for assessing model adequacy is ultimately one of balancing statistical intuition and computational effort. We have chosen the following set of six summary statistics to compute on the contrasts because they capture a range of possible model violations and have well–understood statistical properties. All of these essentially evaluate whether the contrasts come from the distribution expected under BM.

M_{PIC} The mean of the squared contrasts. This is equivalent to the REML estimator of the Brownian motion rate parameter σ^2 (Garland et al., 1992; Rohlf, 2001). M_{PIC} is a metric of overall rate. Violations detected by M_{PIC} indicate whether the overall rate of trait evolution is over– or underestimated.

V_{PIC} The coefficient of variation (standard deviation/mean) for the absolute value of the contrasts. If V_{PIC} calculated from the observed contrasts is greater than that calculated from the simulated contrasts, it suggests that we are not properly accounting for rate heterogeneity across the phylogeny. If V_{PIC} from the observed is smaller, it suggests that contrasts are more even than the model assumes. We use the coefficient of variation rather than the variance because the mean and variance of contrasts can be highly correlated.

S_{VAR} The slope resulting from fitting a linear model to the absolute value

of the contrasts vs. their expected variances. Each (standardized) contrast has an expected variance proportional to the sum of the branch lengths connecting the node at which it is computed to its daughter lineages (Felsenstein, 1985). Under a model of BM, we expect no relationship between the contrasts and their variances. We use S_{VAR} to test if contrasts are larger or smaller than we expect based on their branch lengths. If, for example, more evolution occurred per unit time on short branches than long branches, we would observe a negative slope.

S_{ANC} The slope resulting from fitting a linear model to the absolute value of the contrasts vs. the inferred ancestral state at the corresponding node. We estimated the ancestral state using the least–squares method suggested by Felsenstein (1985) for the calculation of contrasts. We note that this is not technically an ancestral state reconstruction (see Felsenstein, 1985); it is more properly thought of as a weighted average value for each node. We used this statistic to evaluate whether there is variation in rates relative to the trait value; for example, do larger organisms evolve proportionally faster than smaller ones?

S_{HGT} The slope resulting from fitting a linear model between the absolute value of the contrasts and vs. height of the node at which they are inferred measured from the root. This is used to capture variation relative to time. It is alternatively known as the "node–height test" and has been used to detect early bursts of trait evolution during adaptive radiations (Freckleton and Harvey, 2006; Slater and Pennell, 2013).

D_{KS} The D–statistic obtained from Kolmolgorov–Smirnov test from comparing the distribution of contrasts to that of a normal distribution with mean 0 and standard deviation equal to the root of the mean of squared contrasts (the expected distribution of the contrasts under BM; see Felsenstein, 1985; Rohlf, 2001). We chose this to capture deviations from normality. For example, if traits evolved via a "jump–diffusion" type process (Landis et al., 2013), in which there were occasional bursts of rapid phenotypic evolution (Pennell et al., 2013), the tip data would no longer be multivariate normal owing to a few contrasts throughout the tree being much larger than the rest (i.e., the distribution of contrasts would have heavy tails).

Alternative sets of summary statisics are certainly possible. One could, for instance, calculate the median of the squared contrasts, the skew of the distribution of contrasts, etc. If the generating model was known, we could use established procedures for selecting a set of sufficient (or, approximately sufficient; Joyce and Majoram, 2008) summary statistics for that model, as is typically done when computing likelihood ratio tests. However, the aim of our approach is assess the fit of a proposed model without reference to a true model. Our summary statistics will detect many types of model misspecification but this does not mean that they will necessarily detect every type of model misspecification; researchers interested in specific questions are encouraged to explore alternate sets of summary statistics.

Beyond Brownian motion

All of the summary statistics are designed to evaluate the adequacy of a BM model of trait evolution. Our summary statistics S_{VAR}, S_{ANC}, and S_{HGT}

have been used previously in the literature with this justification (Garland et al., 1992, 1993; Díaz-Uriarte and Garland, 1996). However, if we propose a different model for the evolution of the trait, such as an Ornstein–Uhlenbeck (OU; Hansen, 1997) process, then the expected distribution of the contrasts is different. The expected distribution of contrasts under most models of trait evolution, aside from BM, is not formally characterized and even if it was, this would necessitate a specific set of summary statistics for every model proposed.

Our solution to this problem is to create what we term a "unit tree", which is a phylogenetic tree transformation that captures the dynamics of trait change under a particular evolutionary model. More formally, for a particular evolutionary model $\mathcal{M}$ (with parameter values θ), we define a unit tree as a phylogenetic tree that has the following property: the length of branch i, v_i', is equal to the amount of variance expected to accumulate over i under $\mathcal{M}, \theta$. The variance is standardized, such that the expected distribution of the trait data on the unit tree is equal to that of a Brownian Motion (BM) model with a rate σ^2 equal to 1.

One can construct a unit tree by recognizing that many models of evolution can be represented via transformations of the branch lengths of a phylogenetic tree (O'Meara, 2012; Ho and Ané, 2014), including multi–rate BM (O'Meara et al., 2006; Thomas et al., 2006; Eastman et al., 2011), mulit–optima OU (Butler and King, 2004; Beaulieu et al., 2012; Ingram and Mahler, 2013), and models in which the rate and/or process change through time (Blomberg et al., 2003; Slater, 2013).

Specifically, we create unit trees by transforming branch lengths using

the following general formula:

$$v_i' = \mathrm{E}[\mathrm{Cov}(A, B)] - \mathrm{E}[\mathrm{Cov}(A, C)] \quad (2)$$

where v_i' is the transformed length of a branch i, A and C are two lineages subtended by the node at the rootward end of i, and A and B are two lineages subtended by the tipward end of i (see figure 2). One can choose any pair of species that fits this property and then transverse the phylogeny in any direction so that each branch in the tree is transformed. For terminal branches, the formula reduces to $\mathrm{E}[\mathrm{Var}(A)] - \mathrm{E}[\mathrm{Cov}(A, B)]$. Measurement error can be incorporated by simply lengthening the terminal branches of the unit tree by the estimated standard error. As an example, we illustrate how a unit tree is constructed from the parameters of a fitted OU model (figure 2).

If the fitted model is adequate, the trait data at the tips of the unit tree will have the same distribution as data generated under a BM process with a rate of 1 and the contrasts will be $\sim \mathcal{N}(0, 1)$ — hence the name, unit tree. Creating the unit tree from the estimated model parameters prior to computing the contrasts generalizes the summary statistics to most models of quantitative trait evolution (but see Landis et al., 2013, for an exception). We also emphasize that because the contrasts are calculated on the unit tree, the summary statistics all must depend on both the data and the model — it is for this reason that the Bayesian version of our approach produces a distribution of observed summary statistics.

Once we have created the unit tree from the estimated parameters, new datasets can be simulated under the model simply using a BM process with $\sigma^2 = 1$. The distribution of summary statistics calculated on these

simulated data sets can then be compared to the summary statistics from the observed data. We summarize the entire approach in figure 1. We have implemented our method in a new R package, `arbutus` (see below for more details).

The adequacy of models for the evolution of plant functional traits

Data

We used a phylogeny of Angiosperms, containing 30,535 species, from a recent study by Zanne et al. (2013). We refer interested readers to the original publication for details on the phylogeny. For the purposes of this study, we conducted all analyses on the MLE of the phylogeny (available on DRYAD, doi:10.5061/dryad.63q27/3).

We assembled large datasets on three functionally important plant traits: specific leaf area (SLA, defined as fresh area/dry mass); seed mass; and leaf nitrogen content (% mass). Seed mass is a crucial part of species' life–history strategy (Leishman et al., 2000; Westoby et al., 2002) and SLA and leaf nitrogen content are important and widely measured components of species' carbon capture strategies (Wright et al., 2004). Understanding the macroevolutionary patterns of these three traits can provide key insights into the evolutionary processes that have shaped much of plant diversity (Cornwell et al., 2014). All data are previously published; see `https://github/richfitz/modeladequacy` for specific locations and scripts to access and process the original data. The SLA and leaf nitrogen data

comes from Wright et al. (2004) with additional SLA data from the LEDA project (Kleyer et al., 2008). Seed mass data comes the Kew database (Royal Botanical Gardens, Kew, 2014). We used an approximate grepping approach to find and correct spelling mistakes and synonymy tools from The Plant List (2014) to match the trait databases to the Zanne et al. phylogeny. The full data set includes 3293 species for SLA, of which 2200 match species in the Zanne et al. tree. For seed mass, the dataset included 22,817 species with 11,107 matched the phylogeny. For leaf nitrogen content, we have data for 1574 species with 936 included in the tree. All data was log transformed prior to analyses.

Because the vast majority of the species are only represented by a single record, it was not possible to use a species–specific estimate of trait standard error (SE) to account for either measurement error or intraspecific variation. As an alternative, we estimated a single SE for each trait by calculating the mean standard deviation for all species for which we had multiple measurements. The assumption of a constant SE across all species is unlikely to hold up to closer examination, but even a somewhat inaccurate estimate of error is better than assuming none at all (Hansen and Bartoszek, 2012).

Analysis

We first matched our trait data to the whole phylogeny and then extracted subclades from this dataset in a three ways: 1) by family; 2) by order; and 3) by cutting the tree at 50 my intervals and extracting the most inclusive clades (named or unnamed) for which the most recent common ancestor of a group was younger than the time–slice. (The crown age of Angiosperms

is estimated to be ∼243 my in the MLE tree and the tree was cut at 50, 100, 150, and 200my.) We kept only subclades for which there was at least 20 species present in both the phylogeny and trait data so that we had a reasonable ability to estimate parameters and distinguish between models (Boettiger et al., 2012; Slater and Pennell, 2013). For SLA, this left us with 72 clades, seed mass, 226 clades, and leaf nitrogen content, 39 clades (337 in total). We note that these datasets are not independent as many of the same taxa were included in family, order and multiple time–slice subtrees.

Following Harmon et al. (2010), we considered three simple models of trait evolution: 1) BM, which can be associated with genetic drift (Lande, 1976; Felsenstein, 1988; Lynch, 1990; Hansen and Martins, 1996), randomly–varying selection (Felsenstein, 1973), or the summation of many independent processes over macroevolutionary time (Hansen and Martins, 1996; Uyeda et al., 2011; Pennell et al., 2013); 2) single optimum OU, which is often assumed to represent stabilizing selection (following Lande, 1976), though we think a more meaningful interpretation is that it represents an "adaptive zone" (Hansen, 2012; Pennell and Harmon, 2013); and 3) EB (also known as ACDC), which was developed as a mathematical representation of a niche–filling process during an adaptive radiation (Blomberg et al., 2003; Harmon et al., 2010). We fit each of these models to all 337 subclades in our dataset. We then used the approach we developed to assess the adequacy of each fitted model.

All of the analyses conducted in this paper were conducted using both likelihood and Bayesian inference. We did to demonstrate the scope of our approach and because both ML and Bayesian inference are commonly used in comparative biology. We emphasize that our approach is not tied to any single statistical paradigm.

For the likelihood analyses, we fit the three models (BM, OU, and EB) using ML with the `diversitree` package (FitzJohn, 2012). We calculated the AIC score for each model. We then constructed a unit tree for each subtree, trait and model combination using the maximum likelihood estimates of the parameters. We calculated the six summary statistics described above (M_{PIC}, V_{PIC}, S_{VAR}, S_{ANC}, S_{HGT}, D_{KS}) on the contrasts of the data. We simulated 1000 datasets on each unit tree using a BM model with $\sigma^2 = 1$ and calculated the summary statistics on the contrasts of each simulated data set.

For the Bayesian analysis, we fit the same models as above using a MCMC approach, sampling parameter values using slice sampling (Neal, 2003), as implemented in `diversitree` (FitzJohn, 2012). For the BM model we set a broad uniform prior on $\sigma^2 \sim \mathcal{U}[0, 2]$, the upper bound being substantially larger than the ML estimate of σ^2 for any clade. For the OU model, we used the same prior for σ^2 and drew α values, the strength of attraction to the optimum, from a Lognormal$(\log(0.5), \log(1.5))$ distribution. A complication involved in fitting OU models is deciding what assumptions to make about the state at the root z_0. Here, we follow other authors (Butler and King, 2004; Beaulieu et al., 2012) and assume that z_0 is at the optimum. For the EB model, we again used the same prior for σ^2 and a uniform prior on a, the exponential rate of decrease in σ^2, such that $a \sim \mathcal{U}[-1, 0]$ (the minimum value is much less than we would typically expect; Slater and Pennell, 2013).

Again, for each model/trait/subtree combination, we ran a Markov chain for 10,000 generations. Preliminary investigations demonstrated that this was more than sufficient to obtain convergence and proper mixing for these simple models. After removing a burn–in of 1000 generations, we

calculated the Deviance Information Criterion (DIC, a Bayesian analog of AIC; Spiegelhalter et al., 2002) for each model. We drew 1000 samples from the joint posterior distribution (again, after removing burn–in). For each of the sampled parameter sets, we used the parameter values to construct a unit tree and calculated our six summary statistics on the contrasts. We then simulated a dataset on the same unit tree and calculated the summary statistics on the contrasts of the simulated data.

In the likelihood analyses, for each dataset, we had one set $\mathcal{S}_X$ of observed summary statistics and a 1000 sets $\mathcal{S}_{Y,1}, \mathcal{S}_{Y,2}, \ldots, \mathcal{S}_{Y,1000}$ of summary statistics calculated on data simulated on the same unit tree. In the Bayesian version, we had 1000 sets of observed summary statistics $\mathcal{S}_{X,1}, \mathcal{S}_{X,2}, \ldots, \mathcal{S}_{X,1000}$ using a different unit tree for each set and 1000 sets of simulated summary statistics $\mathcal{S}_{Y,1}, \mathcal{S}_{Y,2}, \ldots, \mathcal{S}_{Y,1000}$, each $\mathcal{S}_{Y,i}$ corresponding to the unit tree used to compute $\mathcal{S}_{X,i}$.

For both types of analyses, we report two–tailed p–values (i.e., the probability that the observed that a simulated summary statistic was more extreme than the observed). As a multivariate measure of model adequacy, we calculated the Mahalanobis distance, a scale–invariant metric, between the observed summary statistics and the mean of our simulated summary statistics, taking into account the covariance structure between the summary statistics. We took the log of the KS D–statistic, D_{KS}, as the Mahanalobis measure assumes data is multivariate normal and the D–statistic is bounded between 0 and 1. For the Bayesian analyses, we report the mean of the distribution of Mahalanobis distances. All analyses were conducted in R v3.0.2 (R Development Core Team, 2013). Scripts to reproduce all analyses are available at `https://github.com/richfitz/modeladequacy`.

A case study: seed mass evolution in the Meliaceae and Fagaceae

As an illustration of our approach, we present a case study examining seed mass evolution in two tree families, the Meliaceae, the "mahogany family", and Fagaceae, which contains oaks, chestnuts and beech trees. The trait data and phylogeny for both groups are subsets of the larger dataset used in the analysis. Superficially, these datasets are quite similar. Both are of similar size (Meliaceae: 44 species in the dataset, 550 in the clade; Fagaceae: 70 species in the dataset and 600 in the clade), age (crown age of Meliaceae: ~53my; Fagaceae: ~40my) and are ecologically comparable in terms of dispersal strategy and climatic niche.

As described above, we fit three simple models of trait evolution (BM, OU, EB) to both datasets using ML and computed AIC weights (AIC_w; Akaike, 1974; Burnham and Anderson, 2004) for the three models. For both datasets, an OU model was overwhelmingly supported ($AIC_w > 0.97$ for both groups). Therefore, looking only at relative model support, we might conclude that similar evolutionary processes are important in these two clades of trees.

Examining model adequacy provides a different perspective. We took the MLE of the parameters from the OU models for each dataset and constructed a unit tree based on those parameters. We calculated our six summary statistics on the contrasts of the data, then simulated 1000 datasets on the unit tree and calculated the summary statistics on the contrasts of each simulated dataset (figure 3). For seed mass evolution in Meliaceae, the OU model was an adequate model; all six observed summary statistics were in the middle of the distribution of simulated sum-

mary statistics ($M_{\text{PIC}} : p = 0.420$, $V_{\text{PIC}} : p = 0.533$, $S_{\text{VAR}} : p = 0.605$, $S_{\text{ANC}} : p = 0.494$, $S_{\text{HGT}} : p = 0.122$, $D_{\text{KS}} : p = 0.677$). In contrast, for Fagaceae we found that the OU model was inadequate with M_{PIC}, the REML estimate of σ^2 was significantly lower than the expectation based on the model ($p \sim 0$), suggesting that the process of evolution that gave rise to this data was more complex that that captured by a simple OU process (we return to this in the Discussion). The rest of the observed summary statistics did not differ significantly from the simulated summary statistics ($V_{\text{PIC}} : p = 0.607, S_{\text{VAR}} : p = 0.188$, $S_{\text{ANC}} : p = 0.404$, $S_{\text{HGT}} : p = 0.883$, $D_{\text{KS}} : p = 0.957$). This example serves to illustrate the distinction between the conventional approach to model selection in PCMs and model adequacy. For both of these datasets, OU best supported model, but is only adequate for one of them. Selecting amongst a limited pool of models does not give a complete picture of whether a model is a good explanation for the data.

Results

Despite the potential for differences in the adequacy of models fit with likelihood versus Bayesian inference, we find the results to be broadly similar. For clarity and conciseness, we present only the results from the likelihood analyses here. Results from the Bayesian analysis are presented in the Supplemental Material. Full results from all analyses are available at `https://github.com/richfitz/modeladequacy`.

Across the 337 subclades, we found widespread support for OU models. For 236 of clades, OU had the highest AIC_w. OU had $\sim$100% of the AIC_w

in 27 clades and >75% of the weight in 184 clades (figure 4). Similar to the analysis of Harmon et al. (Harmon et al., 2010) we found very little support for EB models (only 6 clades supported EB with >75% AIC_w), suggesting that "early bursts" of trait evolution may indeed be rare in comparative data (but see Slater and Pennell, 2013). Larger clades were likely to have high support for a single model (of the 101 clades consisting of more than 100 taxa, 53 had >90 of the AIC weight on a single model), and that was overwhelmingly likely to be an OU model (52/53 clades).

We limit our analyses of model adequacy to only the most highly supported model in the candidate set, as supported by AIC. We did this to present a best–case scenario; if a model had very little relative support, it would be unremarkable if it also had poor adequacy (but see Ripplinger and Sullivan, 2010). Even considering only the best of the set, in general, the models had strikingly poor adequacy (figure 5). Of the 72 comparative datasets of SLA, all 72 rejected the best model by at least one summary statistic (using a cut–off of $p = 0.05$), 33 by at least two, and 17 by three or more. Results were similar in the seed mass data (of the 226 seed mass datasets, 185 were rejected by at least one summary statistic, 128 by at least two and 74 by three or more) and leaf nitrogen content (of the 39 datasets, all 39 could be rejected by at least one, 24 by at least two, and 11 by three or more summary statistic). Some summary statistics were much more likely to detect model violations than others. The best model was rejected by M_{PIC} in 250 datasets and by V_{PIC} in 174; the frequency of rejection is substantially lower by the other summary statistics (S_{VAR}: 65, S_{ANC}: 66, S_{HGT}: 49, D_{KS}: 71). Across all 337 datasets, only 41 are adequately modeled by either BM, OU or EB — all of these are seed mass datasets. This is extremely worrisome as these are the most commonly used models for quantitative

traits in comparative biology.

As the subclades are not independent (overlapping sets of taxa are present in family, order and time–slice phylogenies), conventional statistics, such as linear regression, are not straightforward to apply across datasets. Nonetheless, the trend is clear: the larger the phylogeny, the more likely OU is to be highly supported and the more likely the model is to be inadequate. There is a strong relationship between the size of a subclade and the overall distance between observed and simulated summary statistics, as measured by the Mahanalobis distance (figure 6). This is not simply an artifact of conducting the analyses using a larger number of contrasts for the summary statistics — if the model was adequate at all scales, there would be no relationship between the Mahalanobis distance and the size of the phylogeny. As stated above, larger clades also tended to support a single model, meaning that the datasets for which the best model had a very poor absolute fit also had the most substantial difference between the relative fits of the three models (figure S1). There was no relationship between clade age and model adequacy (figure S2).

Discussion

The distinction between relative and absolute fit is an important one. A model may provide the best explanation for a dataset compared to a few other models but still be a very poor explanation in terms of capturing the patterns of variation present in the data. To again draw an analogy with more conventional statistics, consider fitting a regular linear regression model to a dataset. In this case, we can often identify if the model is

a good absolute fit simply by plotting the data. A number of distributions may produce similar results (i.e., intercept, slope, p–value, etc.) but our inferences are contingent upon the relationship being linear (for a classic case study, see Anscombe, 1973). For phylogenetic models of trait evolution, such simple diagnostic measures are usually not informative for assessing model adequacy due to the complex pattern of dependency between data points — therefore, alternate statistical procedures are necessary. Here we have developed a general tool for this purpose.

In our analyses of the evolution of three key Angiosperm functional traits, the fact that a model — most often, OU — was highly supported relative to the others, had little to do with the model's absolute explanatory power. Overall, the adequacy of the three simple, but commonly used, models of trait evolution was woefully poor (figure 5). Perhaps surprisingly, this was true at across all scales, though models were increasingly inadequate for larger clades (figure 6). Our results raise serious concerns about current practices in comparative biology. Relying on a single model, such as BM, or even the best among a small subset of models, for inference has the potential to greatly mislead inferences about the processes that have driven the evolution of traits at phylogenetic scales.

Many researchers have expressed concern that the models used in comparative biology are often inappropriate, for either biological or statistical reasons (Felsenstein, 1985, 1988; Harvey and Pagel, 1991; Garland et al., 1992; Díaz-Uriarte and Garland, 1996; Hansen and Martins, 1996; Price, 1997; Garland et al., 1999; Garland and Ives, 2000; Hansen and Orzack, 2005; Hansen and Bartoszek, 2012; Felsenstein, 2012; Boettiger et al., 2012; Slater and Pennell, 2013). Indeed, many of these issues were raised in the early days of the field; we are far from the the first to point this out. In this

paper we have developed an approach to actually quantify when a model provides a poor explanation for the data and demonstrate that the concerns about model adequacy are just not theoretical but have real implications for data analysis and interpretation.

The 337 comparative datasets we analyzed varied in terms of traits, size, age and placement in the Angiosperm phylogeny. Nonetheless, several general patterns emerge. An OU model, was by and large, the most supported of the three we examined. In an analysis of 67 comparative datasets consisting of size and shape data from a variety of animal taxa, Harmon et al. (Harmon et al., 2010) also found substantial support for OU models, though for their datasets, BM was more commonly chosen by AIC (we note, however, that many of their datasets were quite small; see Slater and Pennell, 2013). Since their paper, a number of studies conducted in a diverse array of groups have also found OU models to be preferred over BM models (e.g., Burbrink et al., 2012; Quintero and Wiens, 2013; López-Fernández et al., 2013).

The tendency of OU to explain data better than BM has inspired diverse process–based explanations, including stabilizing selection, evolutionary constraints and the presence of "adaptive zones" (Hansen and Martins, 1996; Butler and King, 2004; Hansen, 2012; Pennell and Harmon, 2013). If the widespread support for OU models was indeed caused by the biological processes that have been proposed, we would expect that an OU model would also be widely adequate. However, this is not what we found. OU models are inadequate with our summary statistics, most often with M_{PIC} and V_{PIC} but frequently with others as well. OU models often failed to capture other important types of heterogeneity — variation with respect to branch lengths (S_{VAR}), trait values (S_{ANC}) and time (S_{HGT}). Additionally,

a substantial number of datasets were not well–modeled by a multivariate normal distribution (D_{KS}). These results suggest a statistical explanation for the high support for OU models. OU predicts higher variance near the tips of the phylogeny than do BM or EB models (see figure 1 in Harmon et al., 2010). Heterogeneous evolutionary processes, phylogenetic misestimation and measurement error (Houle et al., 2011; Hansen and Bartoszek, 2012) could also produce such a pattern. In light of our results from model adequacy, it seems likely that OU is widely supported because it is able to accommodate more "slop" than the other models. This is not to say that the processes captured by OU models are unimportant in macroevolution, but rather that OU models may be favored for reasons that are more statistical than biological.

The way in which the observed summary statistics deviate from the simulated values also supports the claim that the widespread support for OU is largely a statistical artifact. Model violations were most frequently detected by the global rate estimate M_{PIC}, in all cases M_{PIC} calculated from the observed contrasts was less than the mean of M_{PIC} calculated from the contrasts of the data simulated on the unit tree. If the evolutionary process (or, alternatively, phylogenetic/measurement error) is heterogeneous across the tree, the lineages in some parts of the clade will be much more divergent than in others. The only way for the model to account for the highly divergent groups is to estimate a large σ^2 (and/or a small α parameter for the OU model). The unit tree formed by these parameter estimates will have long branches across the entire tree. In the less divergent parts of the tree, the contrasts calculated on this unit tree will be small, relative to what we expect under BM. So perhaps counter–intuitvely, when heterogeneity in processes across taxa cause the estimated global rates of

divergence to be inflated, resulting in a lower value for M_{PIC}. For similar reasons, values of V_{PIC} calculated from the observed data tended to be *larger* than the simulated values, though not as consistently as M_{PIC} is smaller.

Returning to our case study, it is unclear why an OU model appears to be an adequate model of seed mass evolution in Meliaceae but not in Fagaceae (as detected by deviations in M_{PIC}; figure 3). Both are woody lineages with large, usually vertebrate dispersed seeds (Pannell and Koziol, 1987; Manos et al., 2001). One intriguing difference between the two clades is that some groups within Fagaceae have recently radiated, e.g., within the oaks (*Quercus*) (Simeone et al., 2013). The same ecological processes may be implicated in driving rapid speciation and heterogeneous patterns of trait evolution across the group (Schluter, 2000), but we do not have any evidence that this is the case. Alternatively, the poor adequacy may not be due to the features of the model per se but may indicate problems in the data. Failure to fully account for measurement error, topological and branch length errors, and "outlier lineages" (see Slater and Pennell, 2013) can all create a poor fit between the model and data, even if the model is generally capturing the pattern of trait evolution along the phylogeny.

If a model is a poor fit, there are a number of ways forward. First, we recommend that researchers take a close look at their data. Has measurement error been appropriately incorporated into the analysis?; are a few contrasts much larger than the rest?; and if so, are these associated with poorly supported nodes in the phylogeny? Such simple diagnostics will likely prove to be informative in many cases. We suspect that for many of the clades in which the models were all rejected, the inadequacy of the model could be traced to idiosyncratic problems in the underlying datasets.

Second, researchers may need to incorporate additional biological realism into their models. There has been a great deal of progress recently towards developing more complex models of trait evolution, such as those allowing multiple rates (O'Meara et al., 2006; Thomas et al., 2006; Eastman et al., 2011), multiple OU optima (Butler and King, 2004; Beaulieu et al., 2012; Ingram and Mahler, 2013) and combinations of processes through time (Slater, 2013). We chose the models we did for this analysis because they are commonly used and because we they allowed a direct and straightforward comparison between different clades in our study as well as with other similiar studies (e.g., Harmon et al., 2010). We were certainly not so naïve as to think they would be appropriate at very large scales; for example, we knew that a single rate BM model was unrealistic for the evolution of seed mass across all Angiosperms (Moles et al., 2005). Though the three models were widely rejected across our datasets, it is likely that models that include heterogeneous processes will provide good explanations for the evolution of these functional traits in many groups. The patterns of deviation between the observed and simulated summary statistics will often reflect interesting biology and can help guide researchers as to what additional processes are likely to be important in their group and need to be incorporated in the model of trait evolution. And our approach can, in turn, be applied to these more complex models to evaluate if they are indeed adequate.

To be clear, we are not advocating in this paper that the p–values obtained from simulations be used as a means of model selection; rather, we view model checking as a step that should follow model choice and fitting (see Gelman et al., 2003, ch. 6, for discussion). If a model is inadequate, we recommend that researchers use the summary statistics to think

critically about what type of additional processes need to be considered, perhaps perform some sort of model selection on a new set of models, and then evaluate the adequacy of a newly proposed model. Simply increasing the complexity of the model until it is no longer rejected by the summary statistics is not advisable. Model selection criteria are designed to balance bias and variance (the variance of a model increases with the number of free parameters). We do not provide any mechanism to penalize the addition of extra parameters. In molecular phylogenetics, it is clear that some form of model selection is crucial for making reliable inferences (Sullivan and Joyce, 2005; Ripplinger and Sullivan, 2008), but the relationship between different model selection criteria and model adequacy measures is complex (Ripplinger and Sullivan, 2010; Boettiger et al., 2012).

However, it may be possible to extend our approach with an eye towards model selection. Slater and Pennell (2013) developed their posterior predictive simulation approach (which is related to our method) to distinguish between a BM model and one where rates of evolution decreased through time. They chose summary statistics (including the node–height test Freckleton and Harvey, 2006, S_{HGT} is our study) specifically to address this question. Slater and Pennell found using posterior predictive fit as a model selection criterion to be much more powerful than comparing models using AIC or likelihood ratio tests, particularly when "outlier taxa" (lineages where the pattern of evolution deviates from the overall model) were included in the analysis. The logic of Slater and Pennell could be extended to other scenarios; to test some evolutionary hypotheses, we may care a lot about whether a model explains varation along some axes but be less concerned about others. This is a question–specific approach to model selection and has been developed in the context of molecular phylogenetics

(Bollback, 2002; Lewis et al., 2013). This is also the essence of the Decision–Theoretic approach to model selection (Robert, 2007), which has also been well–used in phylogenetics (Minin et al., 2003), but has not previously been considered in PCMs.

There are a number of additional ways our approach could be extended. First, we have only considered a limited set of summary statistics. We chose them because each of these has a clear statistical expectation and observed deviations from them have intuitive biological explanations. However, they are certainly a subset of all possible summary statistics that could be applied. For example, because contrasts are i.i.d., there should be no autocorrelation between neighboring contrasts; the summary statistics could be expanded to detect non–zero autocorrelation. Second, our approach can be applied equally well to phylogenetic regression models, such as phylogenetic generalized least squares (Grafen, 1989) or phylogenetic mixed models (Lynch, 1991; Hadfield and Nakagawa, 2010), where concerns regarding model adequacy are just as pertinent (Hansen and Bartoszek, 2012). In most phylogenetic regression models, the trait model describes the pattern of covariance between the residuals rather than the data (Rohlf, 2001, 2006; Freckleton et al., 2011) (but see Hansen et al., 2008). If we form the unit tree from estimated model parameters and calculate the summary statistics on the contrasts of the residuals, we can apply our approach. While our approach can be used to assess the adequacy of the phylogenetic component of regression models "out of the box", additional steps are required to assess the adequacy of the linear component. Third, our method was designed for quantitative trait models that assume data can be modeled with a multivariate normal distribution. We need model adequacy approaches for other types of traits (e.g., binary, multistate, ordinal) and for quantita-

tive models that do not predict a multivariate normal distribution of data (Landis et al., 2013).

Arbutus

We have implemented our approach in a new R package `arbutus`. It is available on github `https://github.com/mwpennell/arbutus`. For this project, we have also adopted code from the `ape` (Paradis et al., 2004), `geiger` (Pennell et al., in press), `diversitree` (FitzJohn, 2012) and `ggplot2` (Wickham, 2009) libraries. We have written functions to parse the output of a number of different programs for fitting trait evolution models (see the `arbutus` website for an up–to–date list of supported models and packages). As this approach was developed to be general, we have written the code in such a way that users can include their own summary statistics and trait models in the analyses; we include demonstrations of how this can be done on the project website.

Concluding remarks

Attempts to assess the adequacy of phylogenetic models are almost as old as modern comparative phylogenetic biology. In the 1980s and 1990s much discussion surrounded the appropriateness of various methods and models (Felsenstein, 1985, 1988; Harvey and Pagel, 1991; Garland et al., 1992; Díaz-Uriarte and Garland, 1996; Price, 1997; Garland et al., 1999; Garland and Ives, 2000). Ironically, as PCMs have become more widely adopted, criticism of modeling assumptions has waned (but see Losos, 2011; Felsen-

stein, 2012; Hansen and Bartoszek, 2012, for recent discussions). In our analysis of Angiosperm functional traits, we demonstrate that these concerns regarding model inadequacy are not just theoretical. Commonly used models in comparative biology were extremely poor explanations for the data; this was true both across the tree and through time. The recent development of models that incorporate different types of heterogeneity is encouraging — invoking increased complexity of processes and patterns will likely be necessary for making robust evolutionary inferences from comparative data. Evaluating the fit of macroevolutionary models should be an important component of any comparative analysis. We hope that our study can help move statistical model adequacy from being an "unknown unknown" to being a "known known", or at least a "known unknown". That is to say, we may not know exactly why our model is not capturing the variation in the data, but at least we will know it is not.

Acknowledgments

We would like to thank the members of the Tempo and Mode of Trait Evolution Working Group at the National Evolutionary Synthesis Center (NESCent) for their suggestions and encouragement. We thank Josef Uyeda, Daniel Caetano and Will Pearse for their thoughtful comments on the manuscript. Paul Joyce and Graham Slater also provided valuable insights into this project. Last, we are grateful to the researchers who made their data available; this project would not have been possible without it. MWP was supported by a NESCent graduate fellowship and a NSERC post–graduate fellowship. This work was also supported by NSF grants awarded to LJH (DEB 0919499 and 1208912).

References

Akaike, H. 1974. A new look at the statistical model identification. Automatic Control, IEEE Transactions on 19:716–723.

Anscombe, F. J. 1973. Graphs in statistical analysis. The American Statistician 27:17–21.

Beaulieu, J. M., D.-C. Jhwueng, C. Boettiger, and B. C. O'Meara. 2012. Modeling stabilizing selection: Expanding the ornstein-uhlenbeck model of adaptive evolution. Evolution 66:2369–2383.

Blomberg, S. P., T. Garland, and A. R. Ives. 2003. Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. Evolution 57:717–745.

Blomberg, S. P., J. G. Lefevre, J. A. Wells, and M. Waterhouse. 2012. Independent contrasts and pgls regression estimators are equivalent. Systematic Biology 61:382–391.

Boettiger, C., G. Coop, and P. Ralph. 2012. Is your phylogeny informative? measuring the power of comparative methods. Evolution 66:2240–2251.

Bollback, J. P. 2002. Bayesian model adequacy and choice in phylogenetics. Molecular Biology and Evolution 19:1171–1180.

Brown, J. M. 2014. Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. Systematic Biology DOI:10.1093/sysbio/syu002.

Burbrink, F. T., X. Chen, E. A. Myers, M. C. Brandley, and R. A. Pyron. 2012. Evidence for determinism in species diversification and contin-

gency in phenotypic evolution during adaptive radiation. Proceedings of the Royal Society B: Biological Sciences 279:4817–4826.

Burnham, K., and D. Anderson. 2004. Model selection and multi-model inference: a practical information-theoretic approach. Springer.

Butler, M. A., and A. A. King. 2004. Phylogenetic comparative analysis: A modeling approach for adaptive evolution. The American Naturalist 164:683–695.

Cornwell, W. K., M. Westoby, D. S. Falster, R. G. FitzJohn, B. C. O'Meara, M. W. Pennell, D. J. McGlinn, J. Eastman, A. T. Moles, P. B. Reich, D. C. Tank, I. J. Wright, L. Aarssen, J. M. Beaulieu, R. M. Kooyman, M. R. Leishman, E. T. Miller, U. Niinemets, J. Oleksyn, A. Ordonez, D. L. Royer, S. A. Smith, P. F. Stevens, L. Warman, P. Wilf, and A. E. Zanne. 2014. Functional distinctiveness of major plant lineages. Journal of Ecology 102:345–356.

D'Erchia, A., C. Gissi, G. Pesole, C. Saccone, and U. Arnason. 1996. The guinea-pig is not a rodent. Nature 381:587–600.

Díaz-Uriarte, R., and T. Garland. 1996. Testing hypotheses of correlated evolution using phylogenetically independent contrasts: Sensitivity to deviations from brownian motion. Systematic Biology 45:27–47.

Eastman, J. M., M. E. Alfaro, P. Joyce, A. L. Hipp, and L. J. Harmon. 2011. A novel comparative method for identifying shifts in the rate of character evolution on trees. Evolution 65:3578–3589.

Felsenstein, J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. American Journal of Human Genetics 25:471–492.

———. 1985. Phylogenies and the comparative method. The American Naturalist 125:1–15.

———. 1988. Phylogenies and quantitative characters. Annual Review of Ecology and Systematics 19:445–471.

———. 2012. A comparative method for both discrete and continuous characters using the threshold model. The American Naturalist 179:145–156.

FitzJohn, R. G. 2012. Diversitree: comparative phylogenetic analyses of diversification in r. Methods in Ecology and Evolution 3:1084–1092.

Freckleton, R. P., N. Cooper, and W. Jetz. 2011. Comparative methods as a statistical fix: The dangers of ignoring an evolutionary model. The American Naturalist 178:E10–E17.

Freckleton, R. P., and P. H. Harvey. 2006. Detecting non-brownian trait evolution in adaptive radiations. PLoS Biol 4:e373.

Garland, T., A. W. Dickerman, C. M. Janis, and J. A. Jones. 1993. Phylogenetic analysis of covariance by computer simulation. Systematic Biology 42:265–292.

Garland, T., P. H. Harvey, and A. R. Ives. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. Systematic Biology 41:18–32.

Garland, T., and A. R. Ives. 2000. Using the past to predict the present: Confidence intervals for regression equations in phylogenetic comparative methods. The American Naturalist 155:346–364.

Garland, T., P. E. Midford, and A. R. Ives. 1999. An introduction to phylogenetically based statistical methods, with a new method for confidence intervals on ancestral values. American Zoologist 39:374–388.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2003. Bayesian Data Analysis. 2nd Edition. Chapman & Hall/CRC.

Gelman, A., X. Meng, and H. Stern. 1996. Posterior predictive assessment of model fitness via realized discrepencies (with discussion). Statistica Sinica 6:733–807.

Goldman, N. 1993. Statistical tests of models of dna substitution. Journal of Molecular Evolution 36:182–198.

Grafen, A. 1989. The phylogenetic regression. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 326:119–157.

Graur, D., W. A. Hide, and W. Li. 1991. Is the guinea-pig a rodent? Nature 351:649–652.

Hadfield, J. D., and S. Nakagawa. 2010. General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. Journal of Evolutionary Biology 23:494–508.

Hansen, T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351.

———. 2012. Adaptive landscapes and macroevolutionary dynamics. Pages 205–221 *in* E. Svensson and R. Calsbeek, eds. The Adaptive Landscape in Evolutionary Biology. Oxford University Press.

Hansen, T. F., and K. Bartoszek. 2012. Interpreting the evolutionary regression: The interplay between observational and biological errors in phylogenetic comparative studies. Systematic Biology 61:413–425.

Hansen, T. F., and E. P. Martins. 1996. Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific data. Evolution 50:1404–1417.

Hansen, T. F., and S. H. Orzack. 2005. Assessing current adaptation and phylogenetic inertia as explanations of trait evolution: the need for controlled comparisons. Evolution 59:2063–2072.

Hansen, T. F., J. Pienaar, and S. H. Orzack. 2008. A comparative method for stuyding adaptation to a randomly evolving environment. Evolution 62:1965–1977.

Harmon, L. J., J. B. Losos, T. Jonathan Davies, R. G. Gillespie, J. L. Gittleman, W. Bryan Jennings, K. H. Kozak, M. A. McPeek, F. Moreno-Roark, T. J. Near, A. Purvis, R. E. Ricklefs, D. Schluter, J. A. Schulte II, O. Seehausen, B. L. Sidlauskas, O. Torres-Carvajal, J. T. Weir, and A. Ø. Mooers. 2010. Early bursts of body size and shape evolution are rare in comparative data. Evolution 64:2385–2396.

Harvey, P. H., and M. D. Pagel. 1991. The comparative method in evolutionary biology. Oxford University Press.

Ho, L. S. T., and C. Ané. 2014. A linear-time algorithm for gaussian and non-gaussian trait evolution models. Systematic Biology DOI:10.1093/sysbio/syu005.

Houle, D., C. Pelabon, G. P. Wagner, and T. F. Hansen. 2011. Measurement and meaning in biology. The Quarterly Review of Biology 86:3–34.

Hunt, G. 2012. Measuring rates of phenotypic evolution and the inseparability of tempo and mode. Paleobiology 38:351–373.

Ingram, T., and D. L. Mahler. 2013. Surface: detecting convergent evolution from comparative data by fitting ornstein-uhlenbeck models with stepwise aic. Methods in Ecology and Evolution 4:416–425.

Joyce, P., and P. Majoram. 2008. Approximately sufficient statistics and bayesian computation. Statistical applications in genetics and molecular biology 7:1–16.

Kleyer, M., R. Bekker, I. Knevel, J. Bakker, K. Thompson, M. Sonnenschein, P. Poschlod, J. Van Groenendael, L. Klimeš, J. Klimešová, et al. 2008. The leda traitbase: a database of life-history traits of the northwest european flora. Journal of Ecology 96:1266–1274.

Lande, R. 1976. Natural selection and random genetic drift in phenotypic evolution. Evolution 30:314–334.

Landis, M. J., J. G. Schraiber, and M. Liang. 2013. Phylogenetic analysis using lévy processes: Finding jumps in the evolution of continuous traits. Systematic Biology 62:193–204.

Leishman, M. R., W. I. J., A. T. Moles, and M. Westoby. 2000. The evolutionary ecology of seed size. Pages 31–57 *in* M. Fenner, ed. Seeds: The Ecology of Regeneration in Plant Communities. CAB Int.

Lewis, P. O., W. Xie, M.-H. Chen, Y. Fan, and L. Kuo. 2013. Posterior predictive bayesian phylogenetic model selection. Systematic Biology DOI:10.1093/sysbio/syt068.

López-Fernández, H., J. H. Arbour, K. O. Winemiller, and R. L. Honeycutt. 2013. Testing for ancient adaptive radiations in neotropical cichlid fishes. Evolution 67:1321–1337.

Losos, J. B. 2011. Seeing the forest for the trees: The limitations of phylogenies in comparative biology. The American Naturalist 177:709–727.

Lynch, M. 1990. The rate of morphological evolution in mammals from the standpoint of the neutral expectation. The American Naturalist 136:727–741.

———. 1991. Methods for the analysis of comparative data in evolutionary biology. Evolution 45:1065–1080.

Maddison, W. P. 2006. Confounding asymmetries in evolutionary diversification and character change. Evolution 60:1743–1746.

Manos, P. S., Z.-K. Zhou, and C. H. Cannon. 2001. Systematics of fagaceae: Phylogenetic tests of reproductive trait evolution. International Journal of Plant Sciences 162:1361–1379.

Minin, V., Z. Abdo, P. Joyce, and J. Sullivan. 2003. Performance-based selection of likelihood models for phylogeny estimation. Systematic Biology 52:674–683.

Moles, A. T., D. D. Ackerly, C. O. Webb, J. C. Tweddle, J. B. Dickie, and M. Westoby. 2005. A brief history of seed size. Science 307:576–580.

Mooers, A. Ø., S. M. Vamosi, and D. Schluter. 1999. Using phylogenies to test macroevolutionary hypotheses of trait evolution in cranes (gruinae). The American Naturalist 154:249–259.

Neal, R. M. 2003. Slice sampling. The Annals of Statistics 31:705–741.

O'Meara, B. C. 2012. Evolutionary inferences from phylogenies: A review of methods. Annual Review of Ecology, Evolution, and Systematics 43:267–285.

O'Meara, B. C., C. Ané, M. J. Sanderson, and P. C. Wainwright. 2006. Testing for different rates of continuous trait evolution using likelihood. Evolution 60:922–933.

Pannell, C. M., and M. J. Koziol. 1987. Ecological and phytochemical diversity of arillate seeds in aglaia (meliaceae): A study of vertebrate dispersal in tropical trees. Philosophical Transactions of the Royal Society of London. B, Biological Sciences 316:303–333.

Paradis, E., J. Claude, and K. Strimmer. 2004. Ape: Analyses of phylogenetics and evolution in r language. Bioinformatics 20:289–290.

Pennell, M. W., J. M. Eastman, G. J. Slater, J. W. Brown, J. C. Uyeda, R. G. FitzJohn, M. E. Alfaro, and L. J. Harmon. in press. geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. Bioinformatics .

Pennell, M. W., and L. J. Harmon. 2013. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. Annals of the New York Academy of Sciences 1289:90–105.

Pennell, M. W., L. J. Harmon, and J. C. Uyeda. 2013. Is there room for punctuated equilibrium in macroevolution? Trends in Ecology & Evolution 29:23–32.

Price, T. 1997. Correlated evolution and independent contrasts. Philo-

sophical Transactions of the Royal Society of London. Series B: Biological Sciences 352:519–529.

Quintero, I., and J. J. Wiens. 2013. Rates of projected climate change dramatically exceed past rates of climatic niche evolution among vertebrate species. Ecology Letters 16:1095–1103.

R Development Core Team. 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Reid, N. M., S. M. Hird, J. M. Brown, T. A. Pelletier, J. D. McVay, J. D. Satler, and B. C. Carstens. 2013. Poor fit to the multispecies coalescent is widely detectable in empirical data. Systematic Biology DOI:10.1093/sysbio/syt057.

Ripplinger, J., and J. Sullivan. 2008. Does choice in model selection affect maximum likelihood analysis? Systematic Biology 57:76–85.

———. 2010. Assessment of substitution model adequacy using frequentist and bayesian methods. Molecular Biology and Evolution 27:2790–2803.

Robert, C. P. 2007. The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation, 2nd Ed. Springer.

Rohlf, F. J. 2001. Comparative methods for the analysis of continuous variables: geometric interpretations. Evolution 55:2143–2160.

———. 2006. A comment on phylogenetic regression. Evolution 60:1509–1515.

Royal Botanical Gardens, Kew. 2014. Seed Information Database (SID), Version 7.1, Accessed 25 March. Http://data.kew.org/sid.

Rubin, D. 1984. Bayesian justifiable and relevant frequency calculations for the applied statistician. Annals of statistics 12:1151–1172.

Schluter, D. 2000. The Ecology of Adaptive Radiations. Oxford University Press.

Simeone, M. C., R. Piredda, A. Papini, F. Vessella, and B. Schirone. 2013. Application of plastid and nuclear markers to dna barcoding of euro-mediterranean oaks (quercus, fagaceae): problems, prospects and phylogenetic implications. Botanical Journal of the Linnean Society 172:478–499.

Slater, G. J. 2013. Phylogenetic evidence for a shift in the mode of mammalian body size evolution at the cretaceous-palaeogene boundary. Methods in Ecology and Evolution 4:734–744.

Slater, G. J., and M. W. Pennell. 2013. Robust regression and posterior predictive simulation increase power to detect early bursts of trait evolution. Systematic Biology DOI:10.1093/sysbio/syt066.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde. 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64:583–639.

Sullivan, J., and P. Joyce. 2005. Model selection in phylogenetics. Annual Review of Ecology, Evolution, and Systematics 36:445–466.

Sullivan, J., and D. Swofford. 1997. Are guinea pigs rodents? the importance of adequate models in molecular phylogenetics. Journal of Mammalian Evolution 4:77–86.

The Plant List. 2014. Version 1.1. Published on the internet. Accessed 11 March. Http://www.theplantlist.org.

Thomas, G. H., R. P. Freckleton, and T. Székely. 2006. Comparative analyses of the influence of developmental mode on phenotypic diversification rates in shorebirds. Proceedings of the Royal Society B: Biological Sciences 273:1619–1624.

Uyeda, J. C., T. F. Hansen, S. J. Arnold, and J. Pienaar. 2011. The million-year wait for macroevolutionary bursts. Proceedings of the National Academy of Sciences 108:15908–15913.

Westoby, M., D. S. Falster, A. T. Moles, P. A. Vesk, and I. J. Wright. 2002. Plant ecological strategies: Some leading dimensions of variation between species. Annual Review of Ecology and Systematics 33:125–159.

Wickham, H. 2009. ggplot2: elegant graphics for data analysis. Springer New York.

Wright, I. J., P. B. Reich, M. Westoby, D. D. Ackerly, Z. Baruch, F. Bongers, J. Cavender-Bares, T. Chapin, J. H. Cornelissen, M. Diemer, J. Flexas, E. Garnier, P. K. Groom, J. Gulias, K. Hikosaka, B. B. Lamont, T. Lee, W. Lee, C. Lusk, J. J. Midgley, M. Navas, U. Niinemets, J. Oleksyn, N. Osada, H. Poorter, P. Poot, L. Prior, V. I. Pyankov, C. Roumet, S. C. Thomas, M. G. Tjoelker, E. J. Veneklass, and R. Villar. 2004. The worldwide leaf economics spectrum. Nature 428:821–827.

Zanne, A. E., D. C. Tank, W. K. Cornwell, J. M. Eastman, S. A. Smith, D. J. McGlinn, R. G. FitzJohn, B. C. O'Meara, A. T. Moles, D. L. Royer, I. J. Wright, L. Aarssen, R. Bertin, J. Dickie, F. Hemmings, M. R. Leishman, K. Liu, J. Oleksyn, A. Ordonez, P. B. Reich, R. Sargent, D. E. Soltis, P. S.

Soltis, P. F. Stevens, N. G. Swenson, L. Warman, M. Westoby, and J. M. Beaulieu. 2013. Three keys to survival in the cold: growth habit, leaf phenology and vessel diameter mold the evolution of angiosperm freezing tolerance. Nature DOI:10.1038/nature12872.

Results from Bayesian analyses

As with the likelihood results (described in main text), OU models were highly supported across many datasets; 178/337 clades had the highest DIC weight (DIC_w) on an OU model; 157 of them with greater than 75% of the total DIC_w (see figure S3). While a generally similar pattern of model support holds for both likelihood and Bayesian inference, the likelihood analyses are much cleaner (compare figure 4 and figure S3). This difference can be explained by the fact that there is a tight statistical relationship between the AIC values for these three models. If two models have identical likelihoods, the AIC scores, defined as $-2\mathcal{L} + 2k$ [where $\mathcal{L}$ is the log-likelihood of the model and k is the number of parameters] will differ by 2. As BM is a special case of both OU and EB, in opposite directions in model space, the highest AIC_w possible for BM is $\sim$0.731). The rare clades were both OU and EB have higher support than BM likely reflect problems in optimization.) Calculating DIC values from posterior samples is inherently more stochastic; if there is little information in data, the best DIC model will depend on the values sampled by the chain.

For the model adequacy results, the results were also very similar to that of the likelihood analyses (compare to Results section in the main text). The adequacy of these simple models was woefully poor across most of the datasets (figure S4). Again, we limit our analyses of model adequacy to only the most highly supported model in the candidate set.

Of the 72 comparative datasets of SLA, the best supported model was rejected by at least one summary statistic in all 72 cases, 31 by at least two, and 17 by three or more. For the seed mass data, the model was rejected by one or mor of the summary statistics in 185 datasets (by two or more in

128 datasets and by at least three in 75 cases). All 39 leaf nitrogen datasets again rejected the best supported model with at least one summary statistic (18 by at least two and 7 by at least three).

Also, similar to the likelihood analyses, the frequency of rejection differed between the summary statistic. M_{PIC} rejected the model in 246 datasets, and V_{PIC} in 169 (S_{VAR}: 76, S_{ANC}: 70, S_{HGT}: 36, D_{KS}: 68). Again, only 41 datasets were adequately modeled by one of the three models in our candidate set. There was a strong relationship between model (in)adequacy and clade size (figure S5), but not for clade age (figure S6).

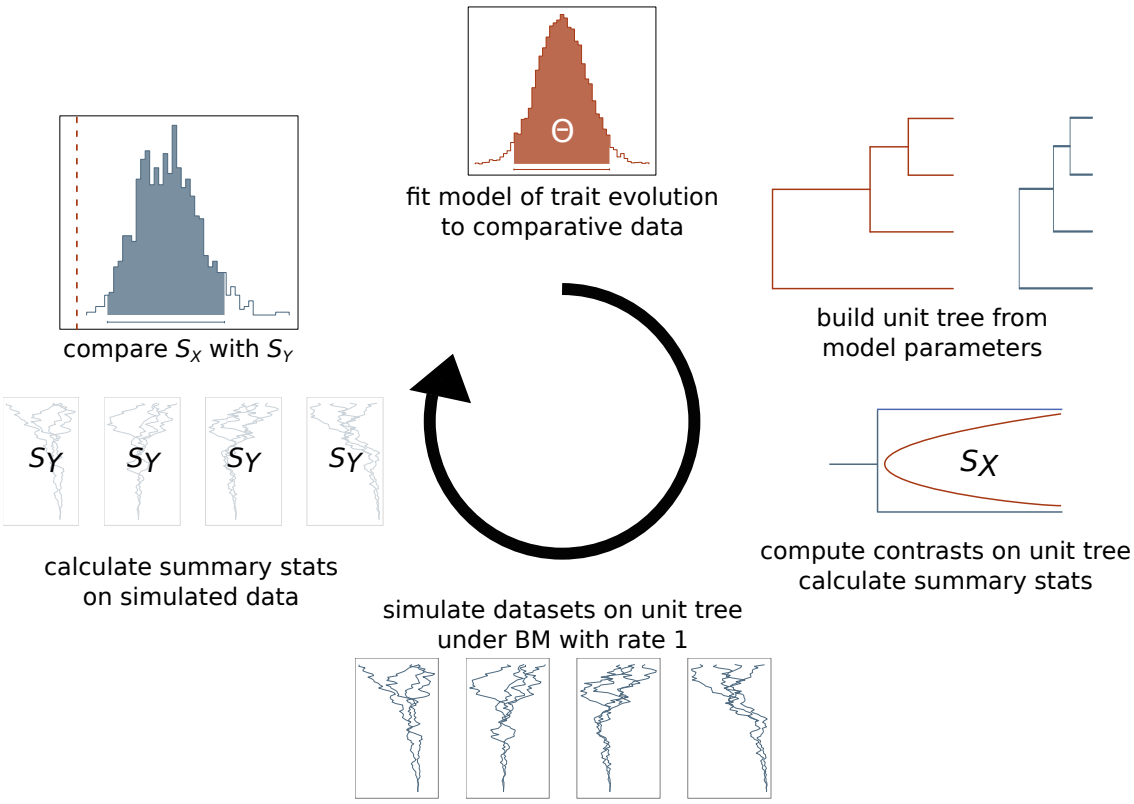

Figure 1: Schematic diagram representing our approach for assessing model adequacy. 1) Fit a model of trait evolution to the data; 2) use the estimated model parameters to build a unit tree; 3) compute the contrasts from the data on the unit tree and calculate a set of summary statistics $\mathcal{S}_X$; 4) simulate a large number of datasets on the unit tree, using a BM model with $\sigma^2 = 1$; 5) calculate the summary statistics on the contrasts of each simulated dataset $\mathcal{S}_Y$; and 6) compare the observed and simulated summary statistics. If the observed summary statistic lies in the tails of the distribution of simulated summary statistics the model can be rejected as inadequate. The rotational circle in the center of the diagram indicates that assessing model adequacy is an iterative process. If a model is rejected as inadequate, the next step is to propose a new model and repeat the procedure.

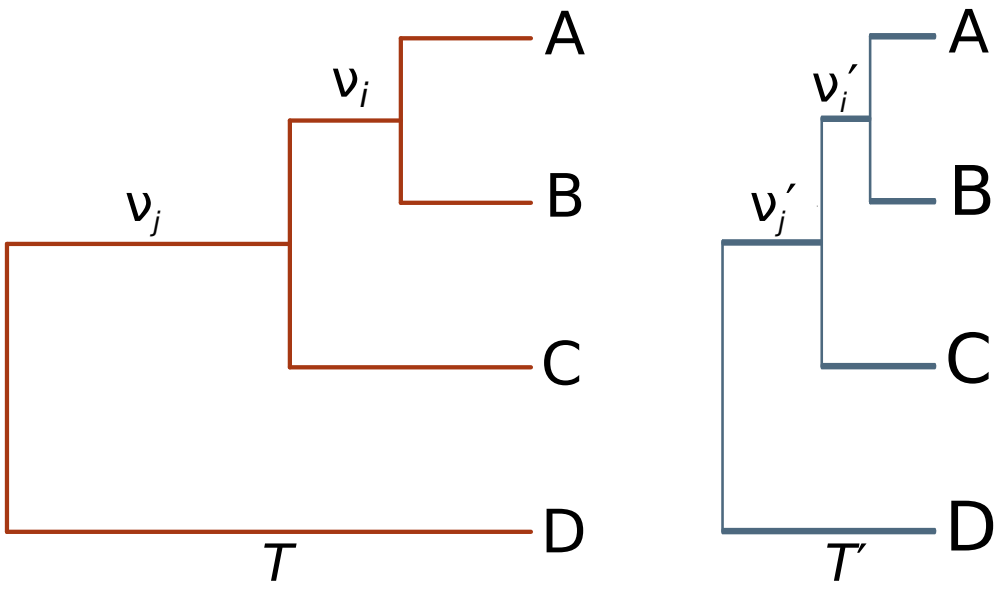

Figure 2: Illustration of how a unit tree is constructed from model parameters. The original phylogenetic tree (in red) can be represented by a variance–covariance (vcv) matrix $\mathbf{C}$. The elements $C_{m,n}$ are the shared path–length from the root to the most recent common ancestor of m and n. The diagonal elements ($m = n$) are simply the total distance from the root to the tips. Given a model $\mathcal{M}$ and parameters θ, the expected vcv of the trait values is described by a second matrix $\mathbf{\Sigma}$. In the case of an OU model, the elements of $\mathbf{\Sigma}$ can be calculated as $\Sigma_{m,n} = \frac{\sigma^2}{2\alpha} e^{-2\alpha(T_{max} - C_{m,n})}(1 - e^{-2\alpha C_{m,n}})$, where T_{max} is the depth of the tree (Hansen, 1997). We can then use the fitted parameter values (in this example, $\sigma^2 = 0.5$ and α, the strength of attraction towards the optimum, is 1), and equation 2 to construct the unit tree (blue). Focusing on a single branch i, the transformed branch length $\nu_i' = \Sigma_{A,B} - \Sigma_{A,C}$. Note that not only have the branch lengths changed relative to one another but the total tree depth T has decreased as well. While the original tree has branch lengths in units of time, the unit tree has branch lengths in units of expected (standardized) variance.

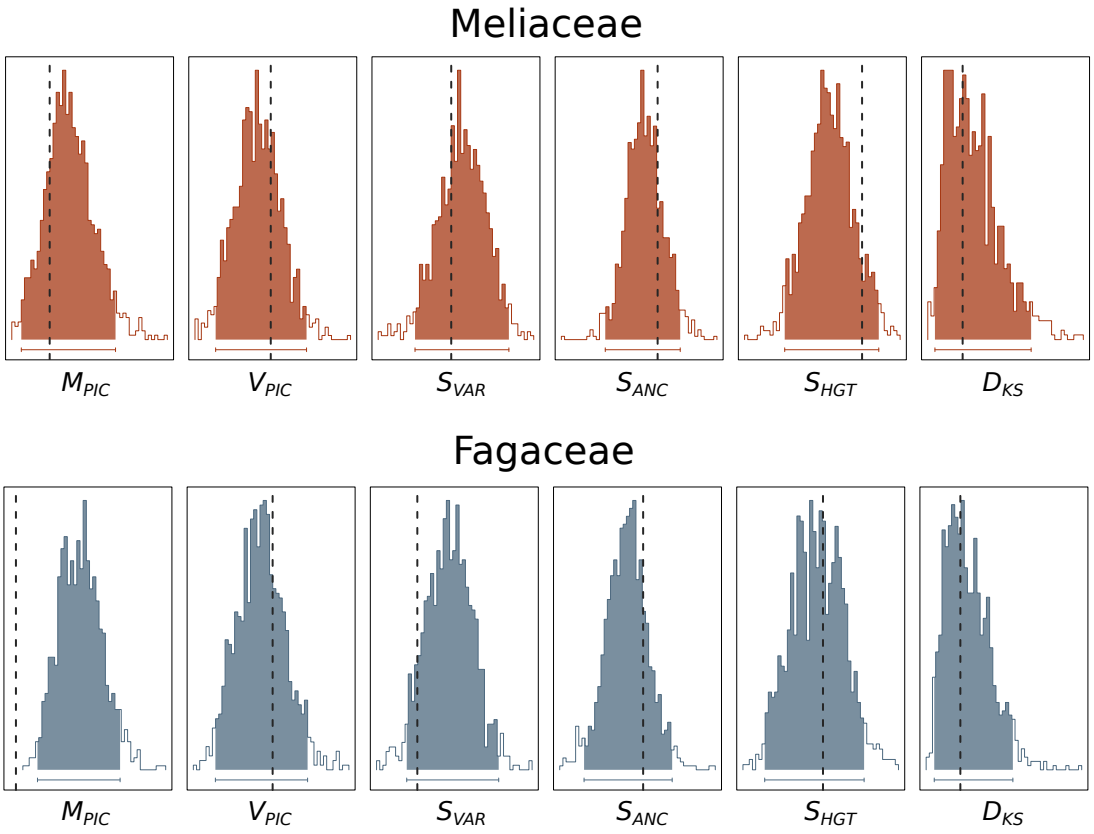

Figure 3: Illustration of our approach to model adequacy. We fit three models (BM, OU, and EB) to seed mass data from two different tree families, the Meliaceae and the Fagaceae. In both cases, an OU model (analyzed here) was strongly supported when fit with ML. The plotted distributions are the summary statistics (M_{PIC}, V_{PIC}, S_{VAR}, S_{ANC}, S_{HGT}, D_{KS}) calculated from the contrasts of the simulated data; the bars underneath the plots represent 95% of the density. The dashed vertical lines are the values of the summary statistics calculated on the contrasts of the observed data. Using our summary statistics, an OU model appears to be an adequate model for the evolution of seed mass in the Meliaceae; for all of the summary statistics, the observed summary statistic lies in the middle of the distribution of simulated summary statistics. For the Fagaceae, the rate estimate M_{PIC} from the observed data is much lower that the rate estimate calculated on the simulated datasets. We can therefore reject an OU model as inadequate for this group (see text for details).

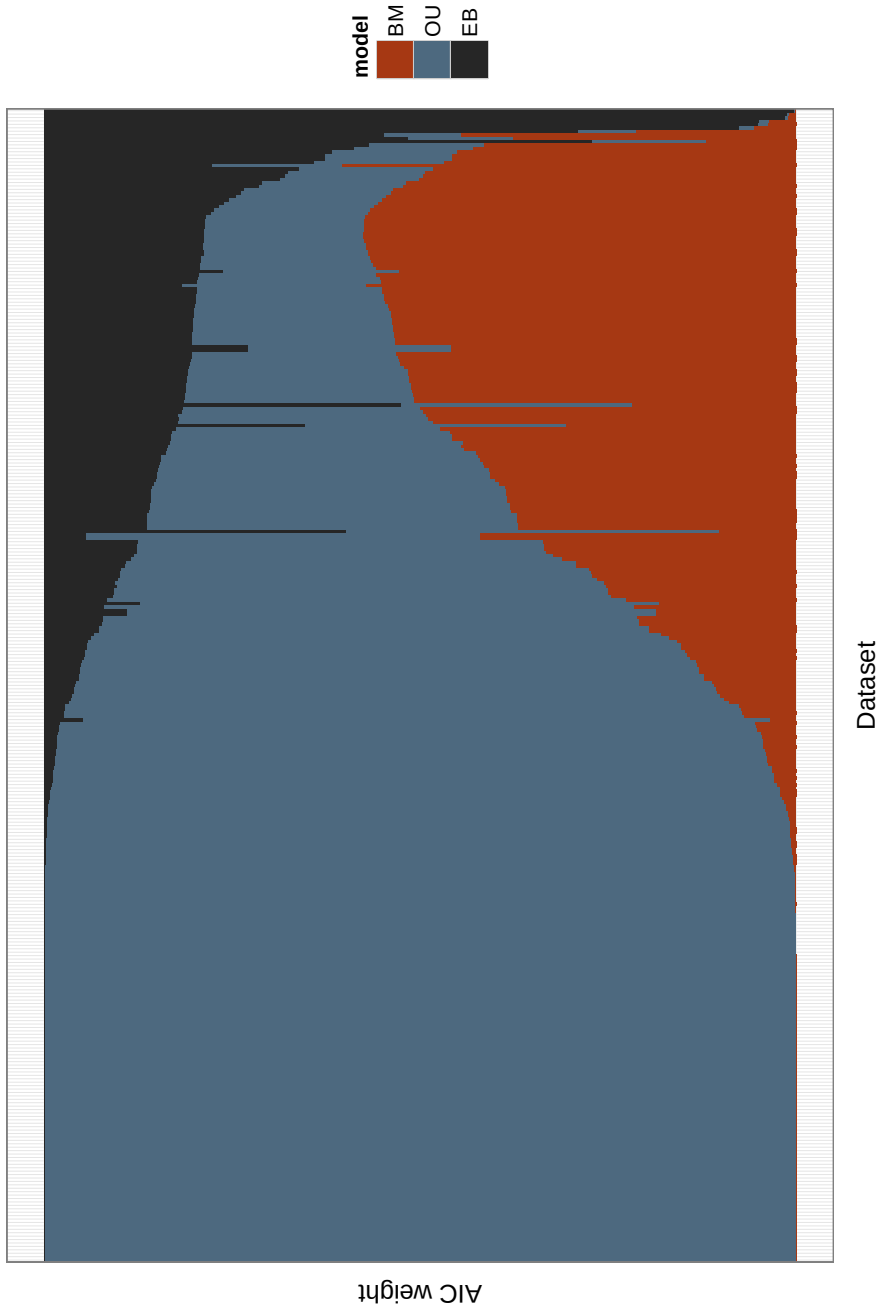

Figure 4: The relative support, as measured by AIC weight, for the three models used in our study (BM, OU, and EB) across all 337 datasets. An OU model is highly supported for a majority of the datasets.

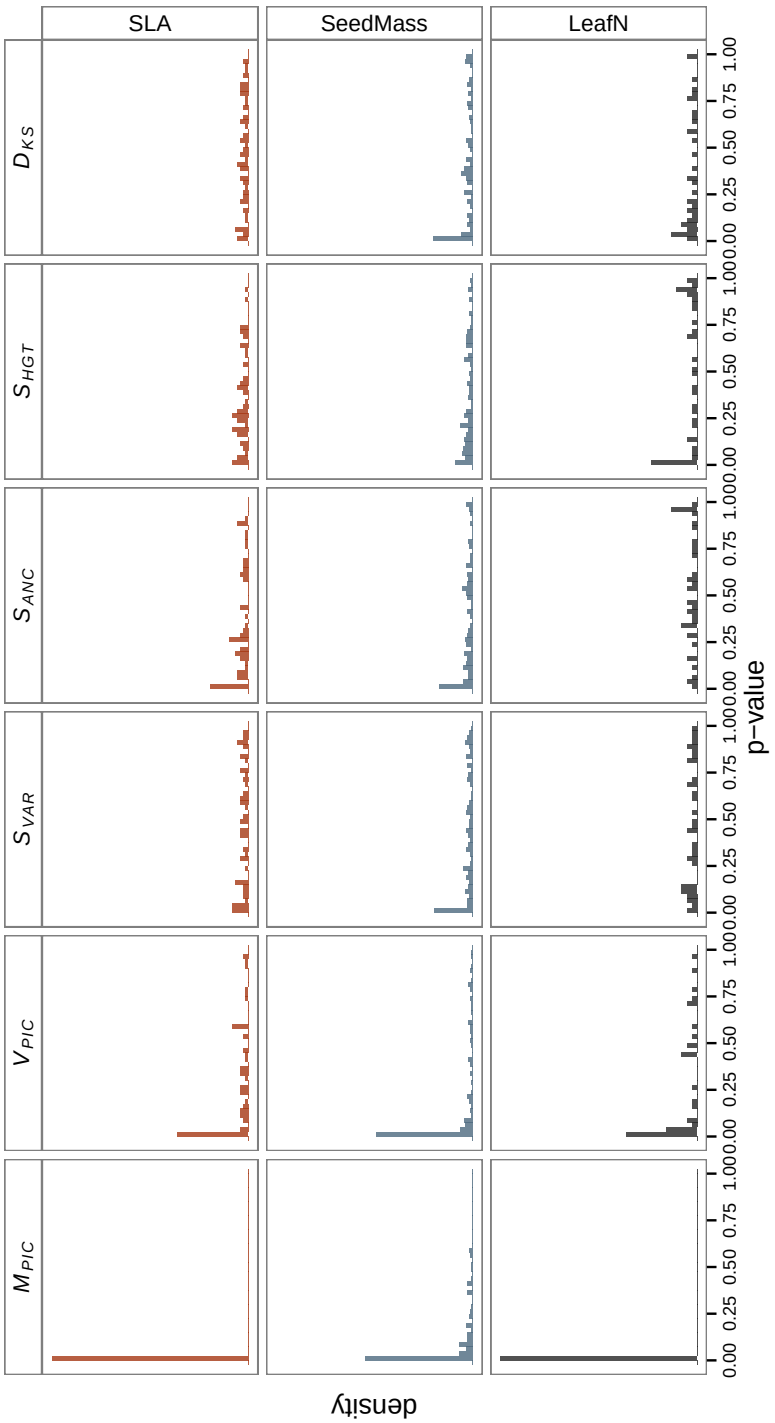

Figure 5: The distribution of p–values for our six summary statistics over all 337 datasets in our study after fitting the models using ML. The p–values are from applying our model adequacy approach to the best supported of the three models (as evaluated with AIC). For both the rate estimate M_{PIC} and the coefficient of variation V_{PIC}, the vast majority of datasets would reject the best of the three models (at $p < 0.05$).

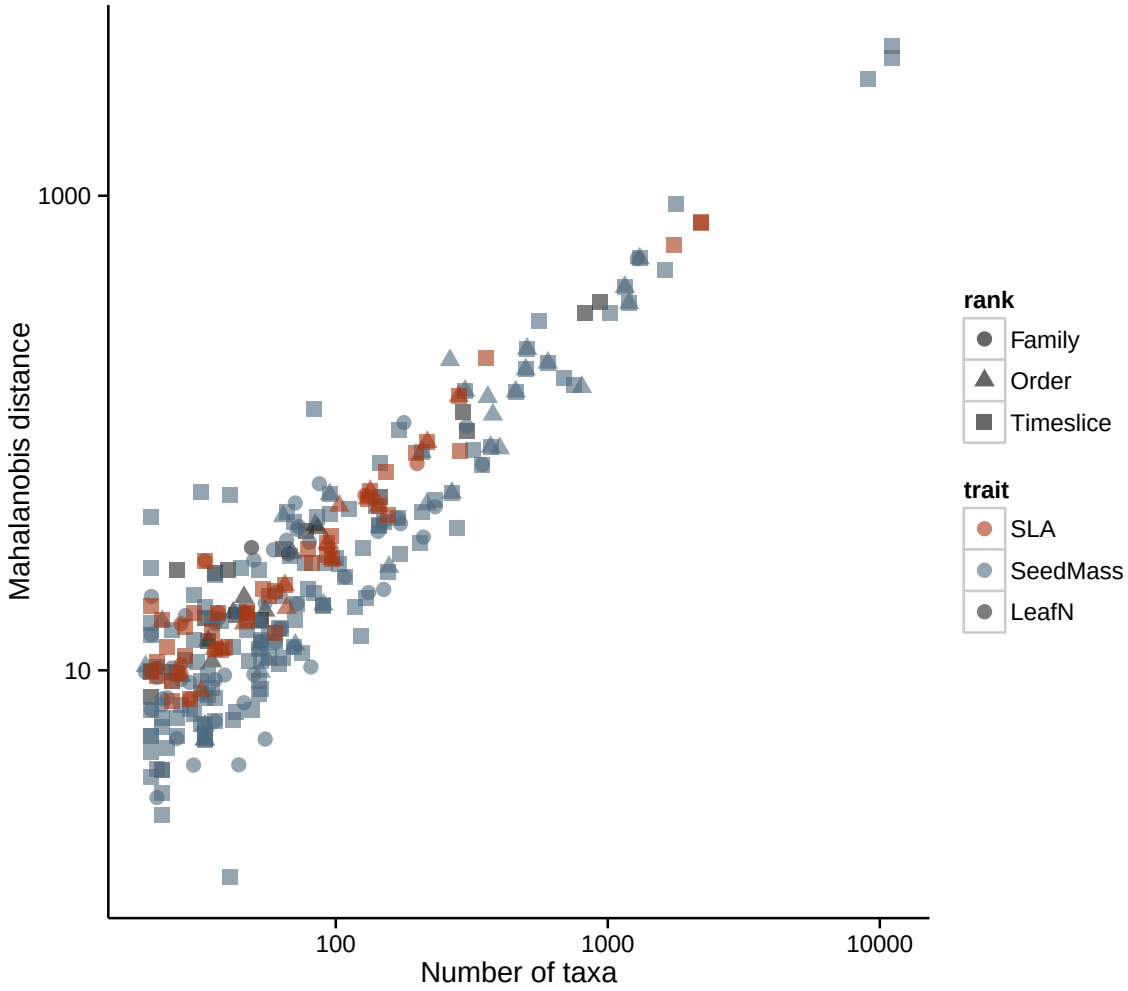

Figure 6: The relationship between clade size and a multivariate measure of model adequacy. The Mahalanobis distance is a scale–invariant metric that measures the distance between the observed and simulated summary statistics, taking into account the covariance between summary statistics. The greater the Mahalanobis distance, the worse the model captures variation in the data. Considering only the best supported model for each clade (as chosen by AIC), there is a striking relationship between the two — the larger the dataset, the worse the models performed (note the logarithmic scale). If the models were equally likely to be adequate at all scales, we would expect no relationship.

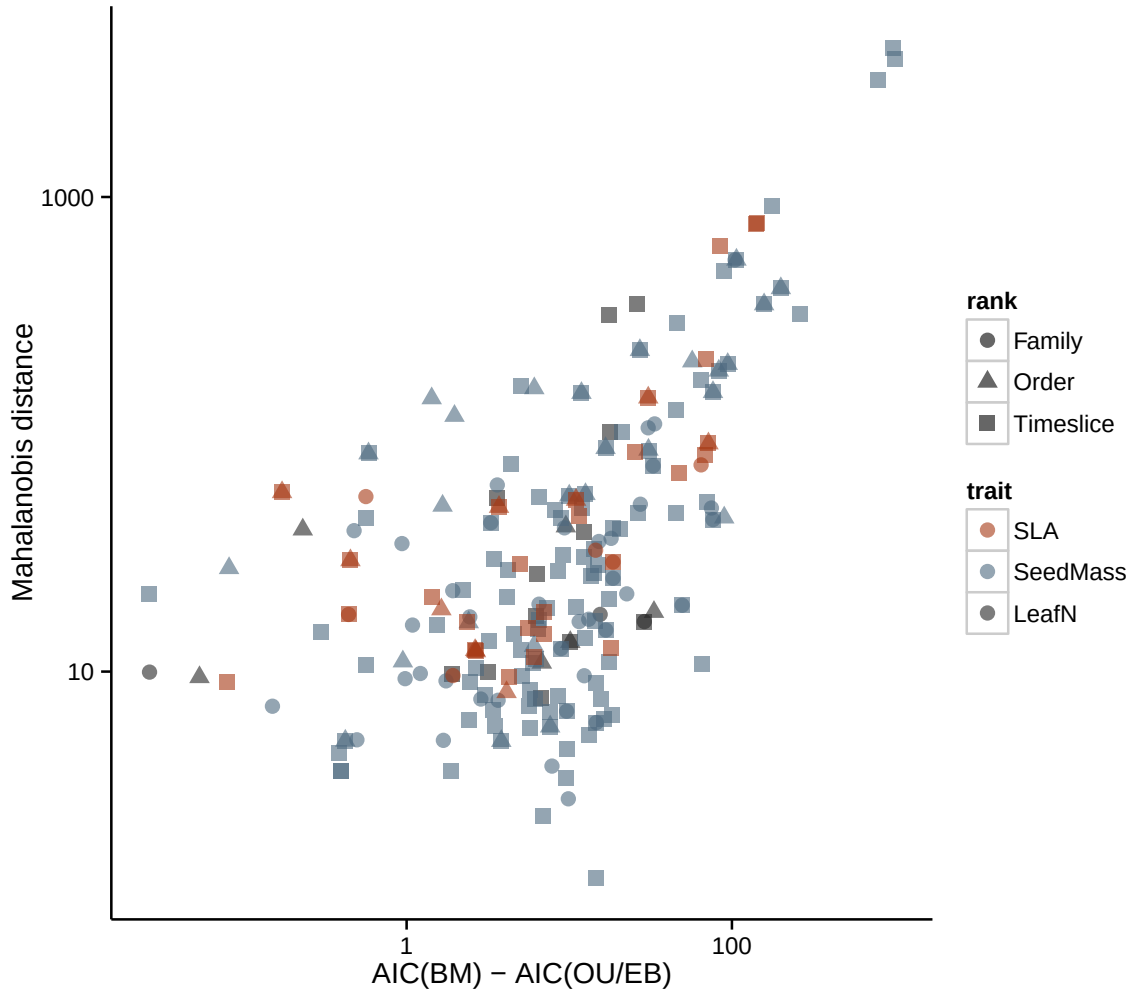

Figure S1: The relationship between relative and absolute fit. For every clade for which a more complex model (OU, EB) was favored over BM using AIC, the Mahalanobis distance between the observed summary and simulated summary statistics is plotted against the improvement in AIC for the more complex model compared to BM. (Note that as all AIC values were negative, larger differences mean greater relative support). The greater the relative fit of a more complex model, the more likely the model was to be inadequate. This result in primarily driven by clade size but serves to emphasize the distinction between relative and absolute fit.

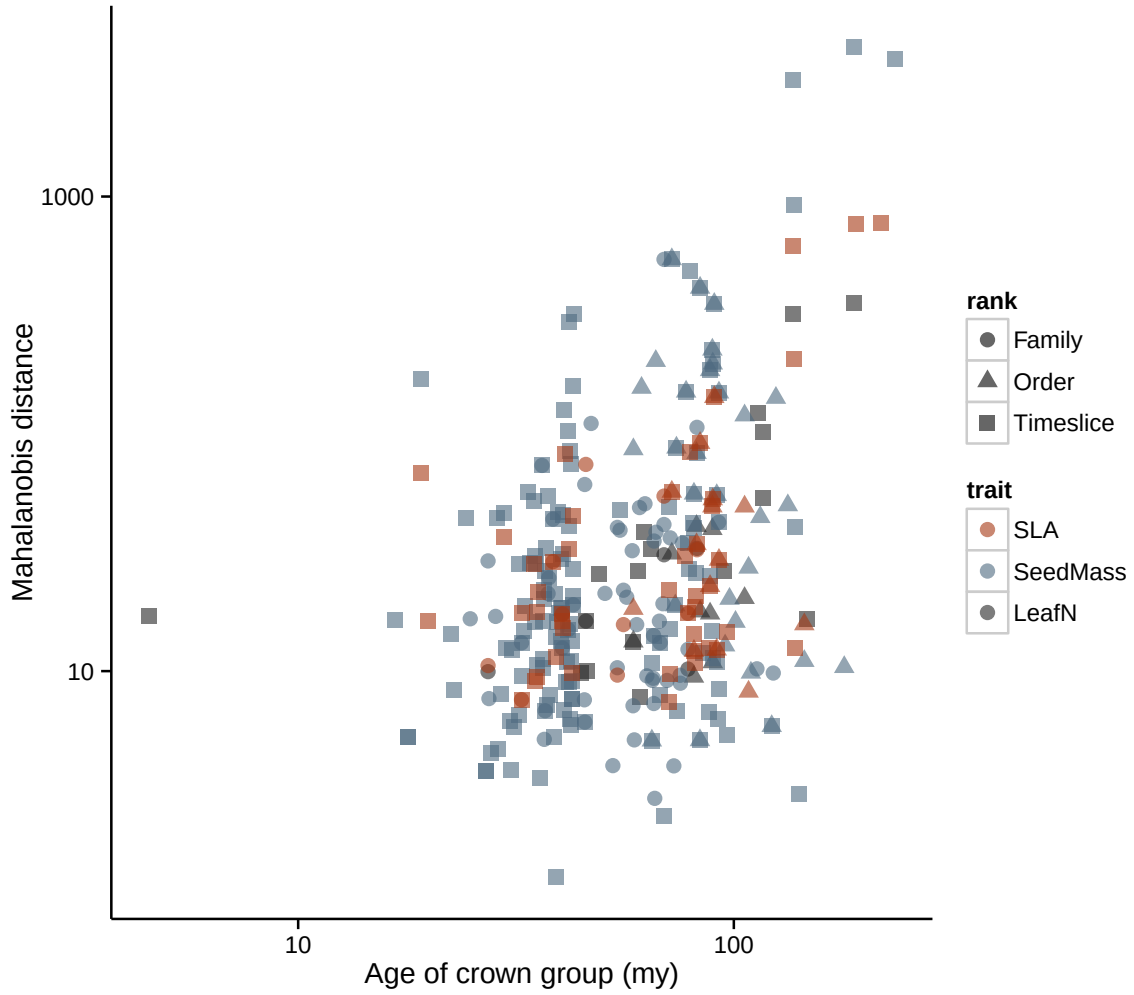

Figure S2: The relationship between clade age and a multivariate measure of model adequacy. Considering only the best supported of the three models (as selected by AIC, after fitting the models using ML), there is no apparent relationship between the age of clade and the distance of the observed and simulated summary statistics, as measured by the Mahalanobis distance. Contrast this figure with figure 6, which demonstrates a very tight relationship between clade size and model inadequacy.

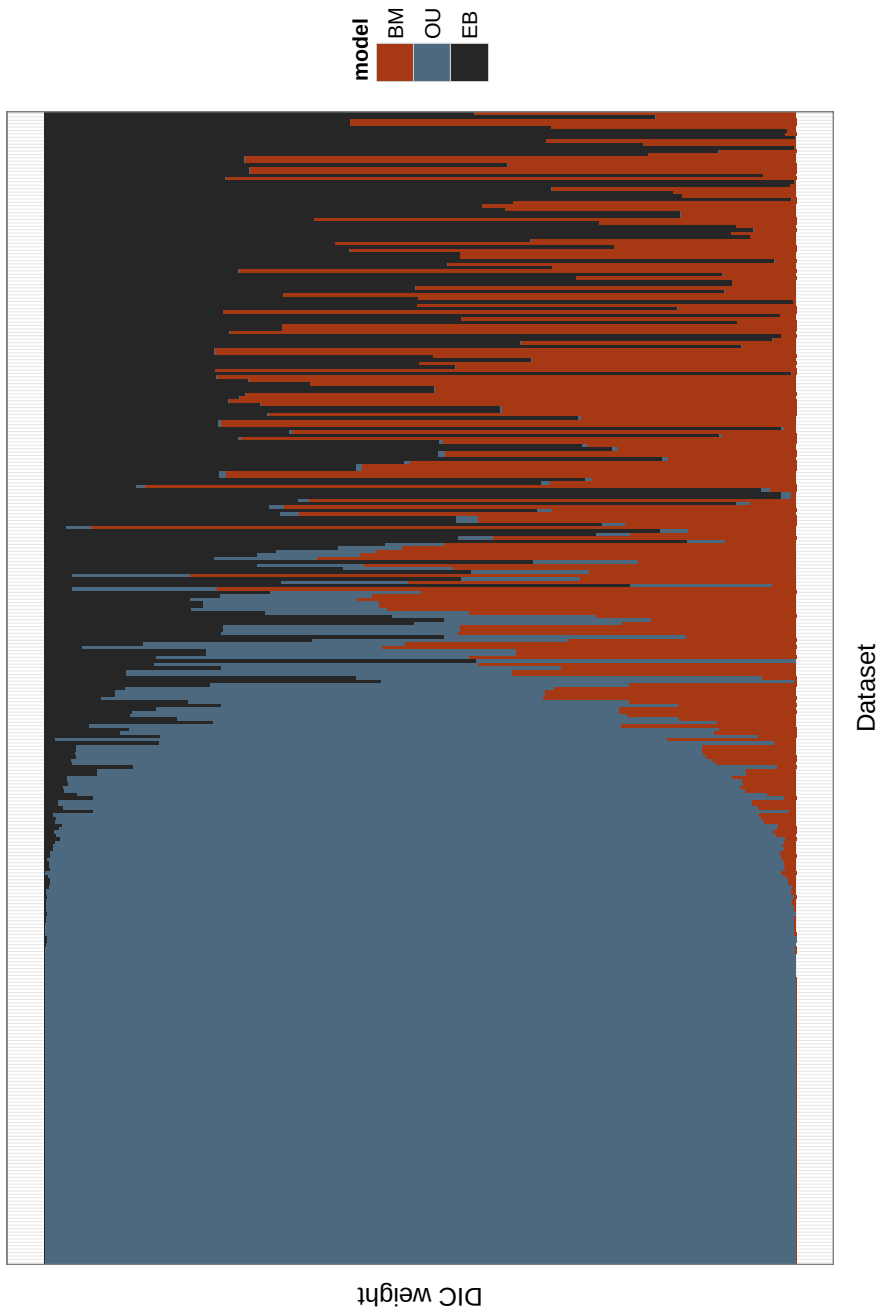

Figure S3: The relative support, as measured by DIC weight, for the three models used in our study (BM, OU, and EB) across all 337 datasets. All models were fit with MCMC. Like the model comparisons done with AIC, an OU model is highly supported for a majority of the datasets.

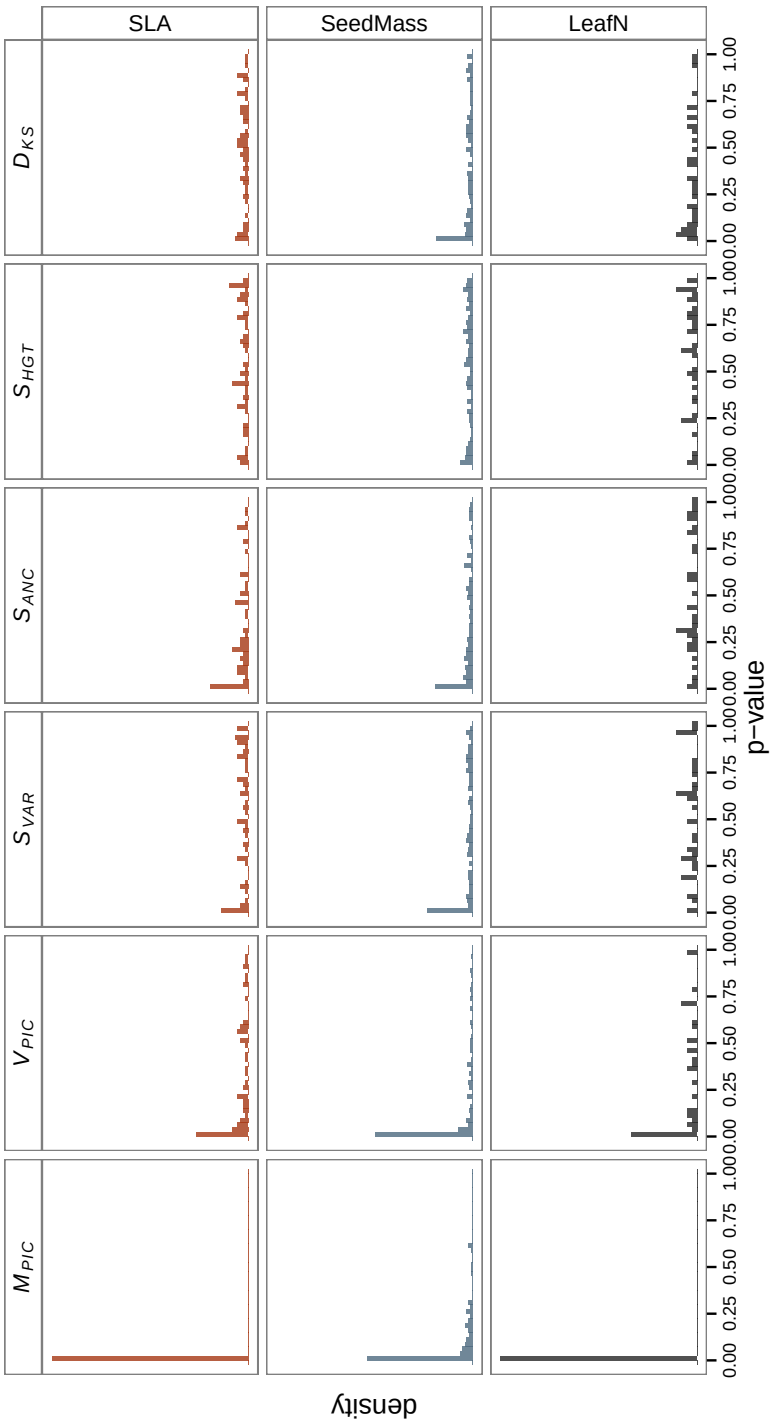

Figure S4: The distribution of p–values for our six summary statistics over all 337 datasets in our study after fitting the models using MCMC. The p–values are from applying our model adequacy approach to the best supported of the three models (as evaluated with DIC). For both the rate estimate M_{PIC} and the coefficient of variation V_{PIC}, the vast majority of datasets would reject the best of the three models (at $p < 0.05$).

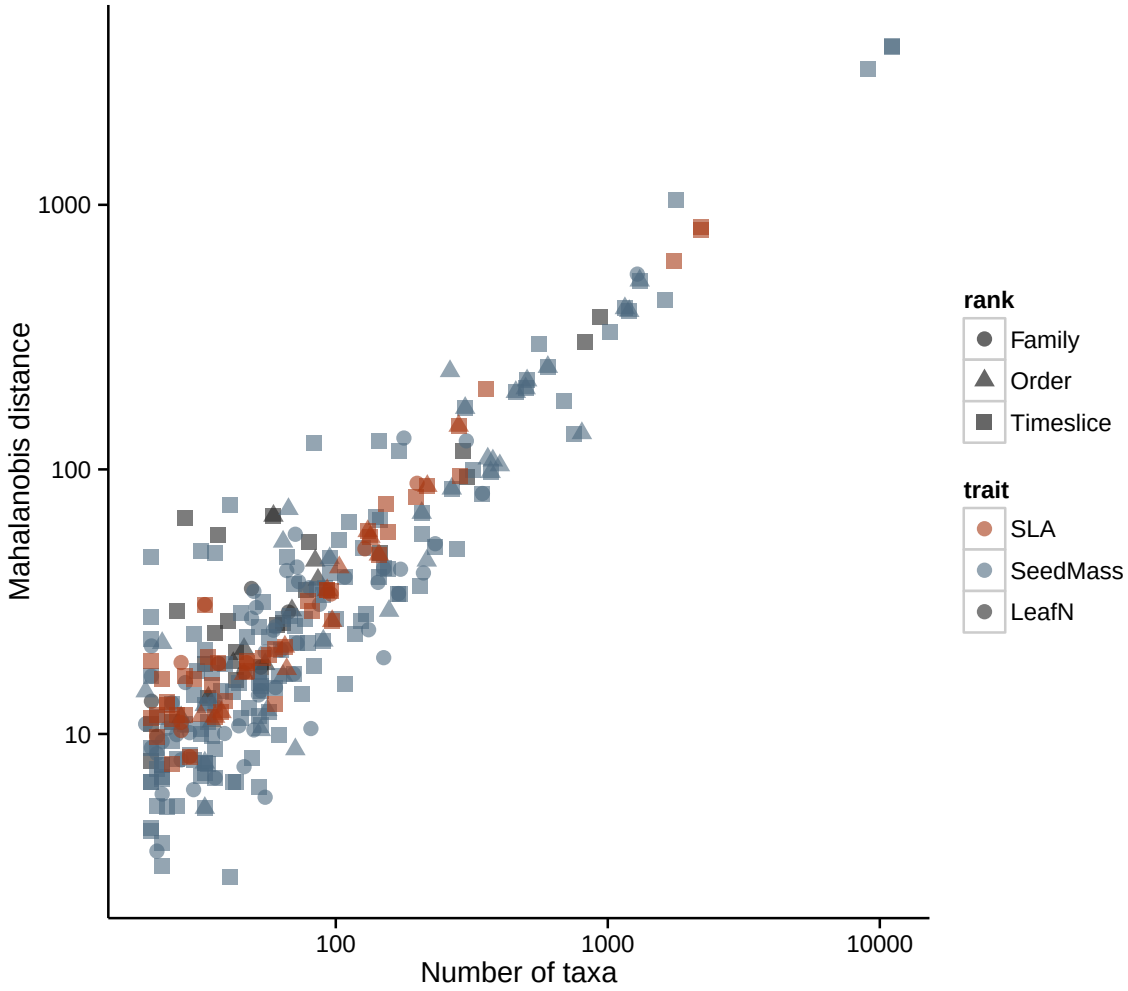

Figure S5: The relationship between clade size and a multivariate measure of model adequacy from the Bayesian analysis. The Mahalanobis distance is a scale–invariant metric that measures the distance between the observed and simulated summary statistics, taking into account the covariance between summary statistics. The greater the Mahalanobis distance, the worse the model captures variation in the data. Considering only the best supported model for each clade (as chosen by DIC), there is a striking relationship between the two — the larger the dataset, the worse the models performed (note the logarithmic scale). If the models were equally likely to be adequate at all scales, we would expect no relationship.

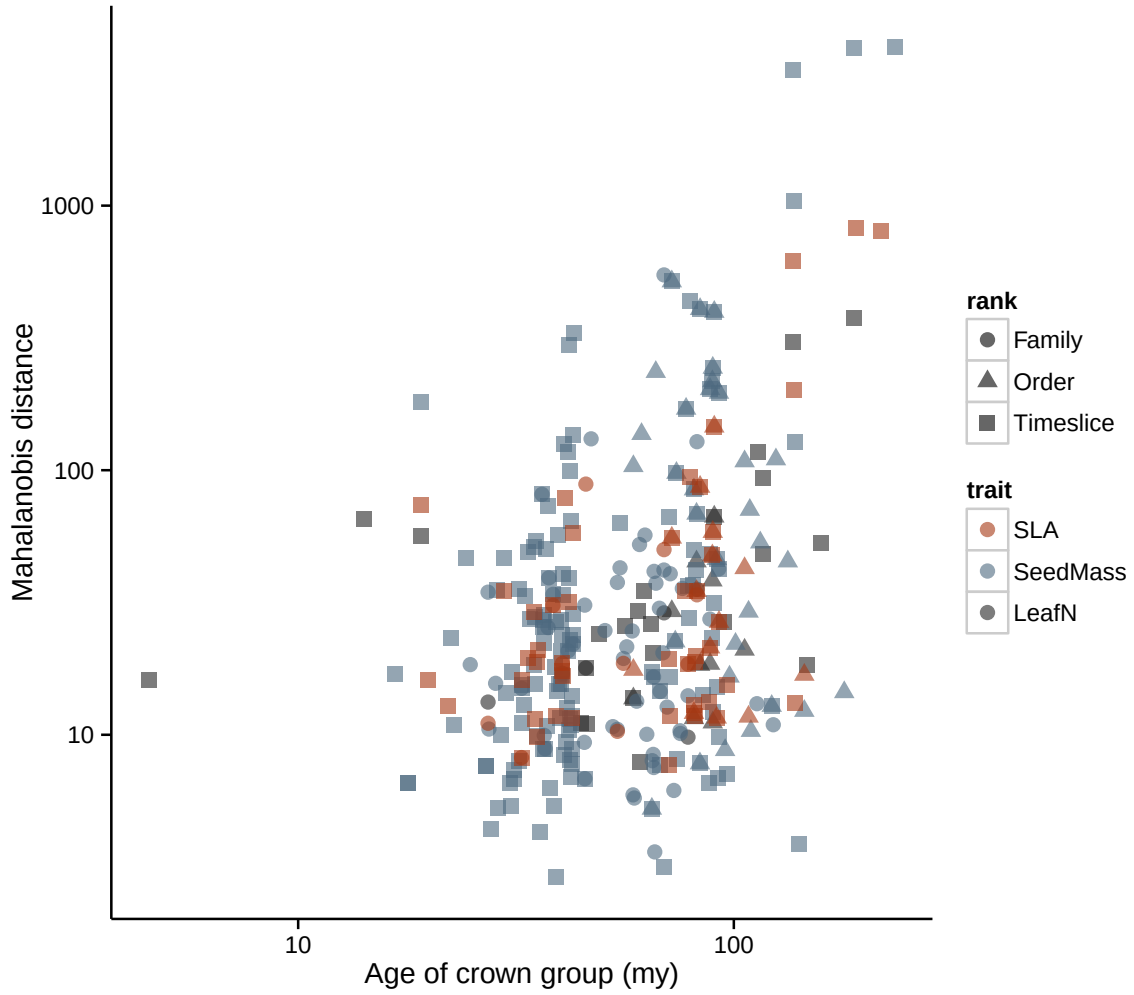

Figure S6: The relationship between clade age and a multivariate measure of model adequacy. Considering only the best supported of the three models (as selected by AIC, after fitting the models using MCMC), there is no apparent relationship between the age of clade and the distance of the observed and simulated summary statistics, as measured by the Mahalanobis distance. Contrast this figure with figure S5, which demonstrates a very tight relationship between clade size and model inadequacy.