SUPPORTING INFORMATION FOR

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]

1 *Department of Biological Sciences & Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, ID 83844, U.S.A.*
2 *Department of Biological Sciences, Macquarie University, Sydney, NSW 2109, Australia*
3 *School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, NSW 2052, Australia*

mwpennell@gmail.com

RESULTS FROM BAYESIAN ANALYSES

As with the likelihood results (described in main text), OU models were highly supported across many datasets; 177/337 clades had the highest DIC weight (DIC_w) on an OU model; 156 of them with greater than 75% of the total DIC_w (see figure S5). While a generally similar pattern of model support holds for both likelihood and Bayesian inference, the likelihood analyses are much cleaner (compare figure 3 and figure S5). This differnce 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 ~0.731. The rare clades where 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 poor across the majority of the datasets (figure S6). 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, we detected deviations from the expectations of the best supported model using at least one test statistic in 35 cases, 26 by at least two, and 19 by three or more. For the seed mass data, we detected deviations with at least one test statistic in 173 cases (by two or more in 109 datasets and by at least three in 72 cases). 24/39 leaf nitrogen datasets were found to be inadequately described by the best supported model with at least one test statistic (13 by at least two and 10 by at least three).

Also, similar to the likelihood analyses, the frequency at which deviations were found differed between the test statistics. In 171 cases, we detected model misspecification with C_{VAR} and with S_{VAR}, 141 (M_{SIG}: 24, S_{ASR}: 101, S_{HGT}: 78, D_{CDF}: 67). We did not detect deviations from the expectations of the best-supported model in 105 datasets. As with the likelihood analyses, we were more likely to detect model deviations when examining larger clades (figure S7).

SUPPLEMENTARY FIGURES

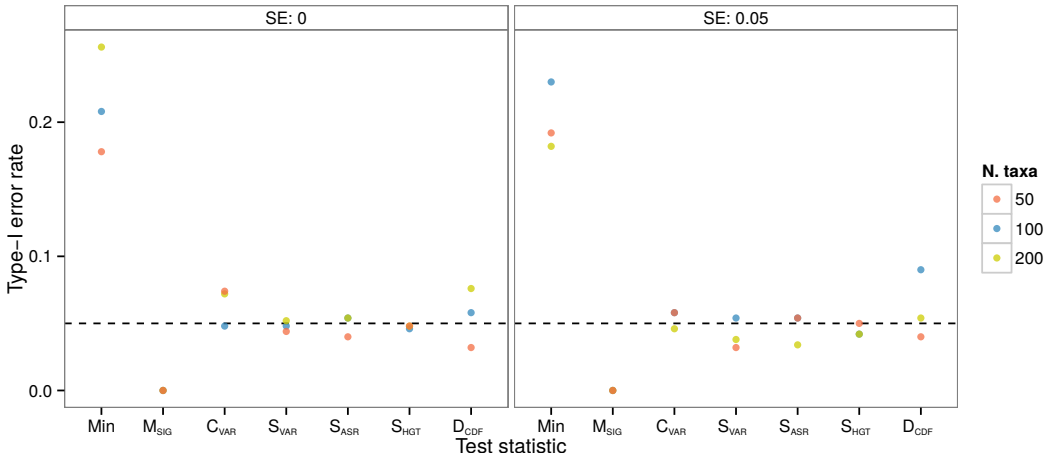

FIGURE S1: Type-1 error rates for data simulated under a Brownian motion (BM) model. We simulated 500 datasets under for 3 different tree sizes ($N = \{50, 100, 200\}$, represented by the different colors) and two known values of standard error of observed species means (0 and 0.05, left and right panel, respectively). The Type-1 error rates for each test statistic are consistently around or lower than a 0.05 threshold. However, the frequency at which *at least one* of the test statistics deviated significantly from the expectations (the variable "Min" on the left side of each plot) was substantially greater, rising to above 20% in some cases. See text for why we decided against correcting for the effect of multiple comparisons in the analysis.

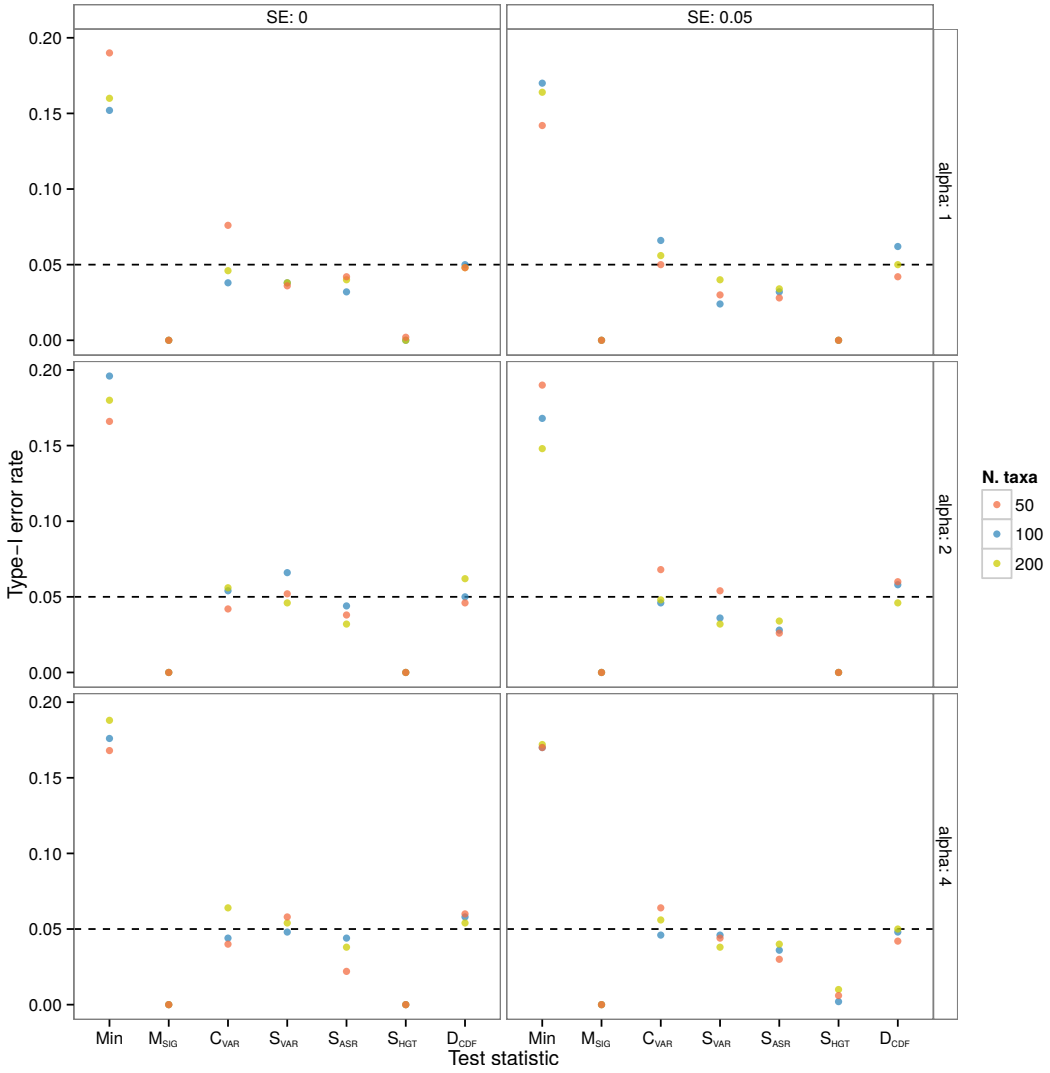

FIGURE S2: Type-1 error rates for data simulated under an Ornstein-Uhlenbeck (OU) model. We simulated 500 datasets under for 3 different tree sizes ($N = \{50, 100, 200\}$, represented by the different colors) and two known values of standard error of observed species means (0 and 0.05, left and right panel, respectively). We also simulated under three values for the α parameter ($\alpha = \{1,2,4\}$, top, middle and bottom panel), representing phylogenetic half-lives of 69%, 35%, 17% of total tree depth, respectively. The Type-1 error rates for each test statistic are consistently around or lower than a 0.05 threshold. However, the frequency at which *at least one* of the test statistics deviated significantly from the expectations (the variable "Min" on the left side of each plot) was substantially greater, approaching 20% in some cases. See text for why we decided against correcting for the effect of multiple comparisons in the analysis.

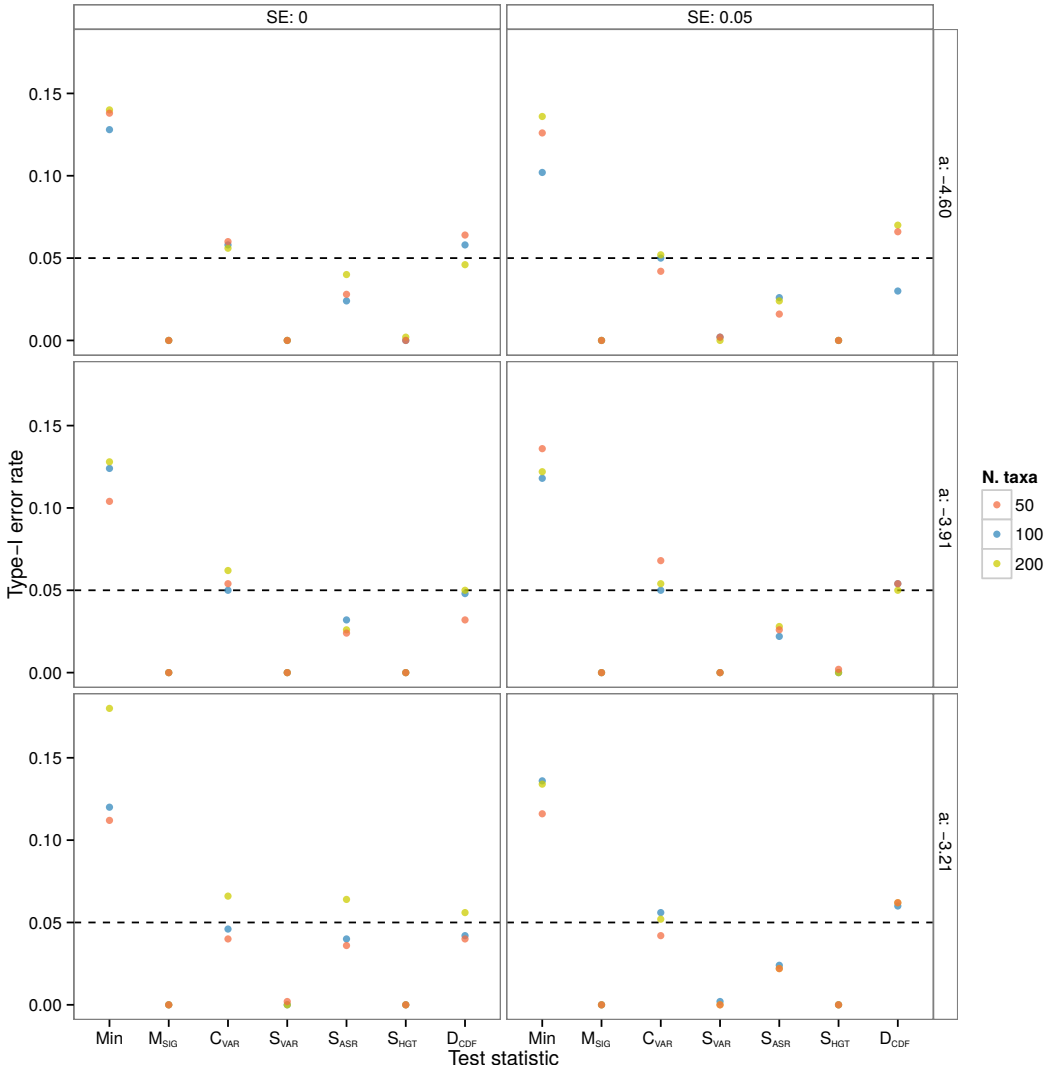

FIGURE S3: Type-1 error rates for data simulated under an Ornstein-Uhlenbeck (OU) model. We simulated 500 datasets under for 3 different tree sizes ($N = \{50, 100, 200\}$, represented by the different colors) and two known values of standard error of observed species means (0 and 0.05, left and right panel, respectively). We also simulated under three values for the exponential rate of slowdown, a ($a = \{\log(0.01), \log(0.02), \log(0.04)\}$, top, middle and bottom panel), which translate to the rate of trait evolution halfing every 0.15, 0.17, and 0.21 time units, respectively (note that the tree was scaled so the total depth was equal to unity). The Type-1 error rates for each test statistic are consistently around or lower than a 0.05 threshold. However, the frequency at which *at least one* of the test statistics deviated significantly from the expectations (the variable "Min" on the left side of each plot) was substantially greater, approaching 15% in some cases. See text for why we decided against correcting for the effect of multiple comparisons in the analysis.

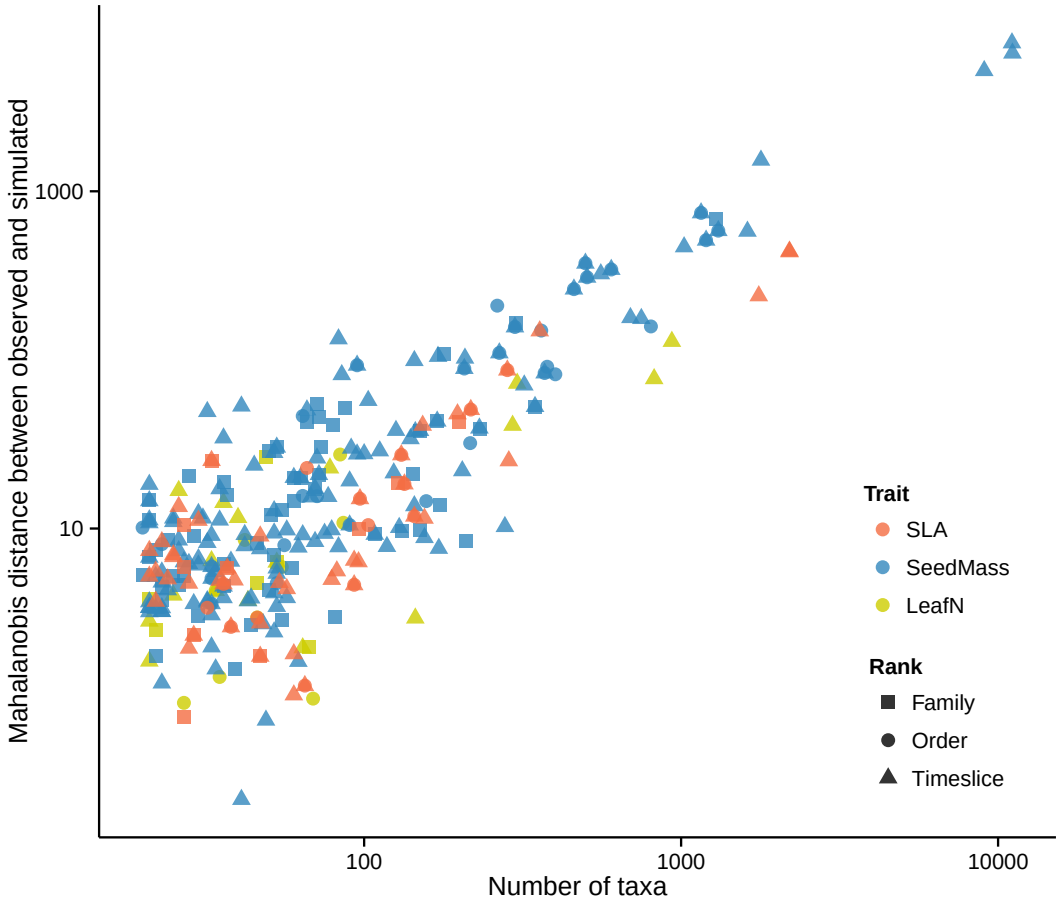

FIGURE S4: The relationship between clade size and a multivariate measure of model adequacy. The Mahalanobis distance is a scale-invariant metric that measures the distance between the observed and simulated test statistics, taking into account the covariance between test statistics. The greater the 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 clear relationship between the two–the larger the dataset, the stronger the evidence that the model does not capture variation in the data.

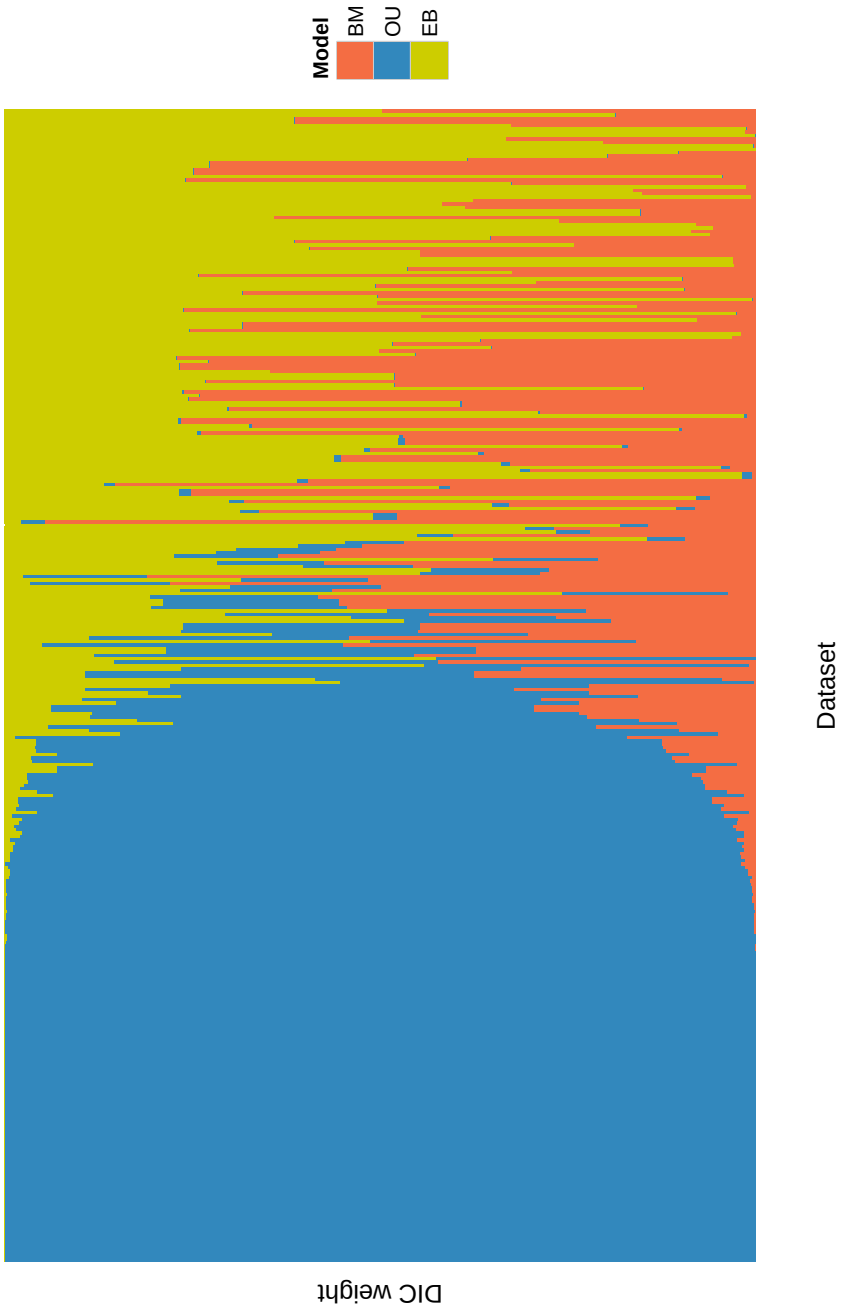

FIGURE S5: The relative support, as measured by DIC weight, for the three models used in our study (BM, OU, and EB) across all 337 datasets. All models were fit with MCMC. Like the model comparisons done with AIC, an OU model is highly supported for a majority of the datasets.

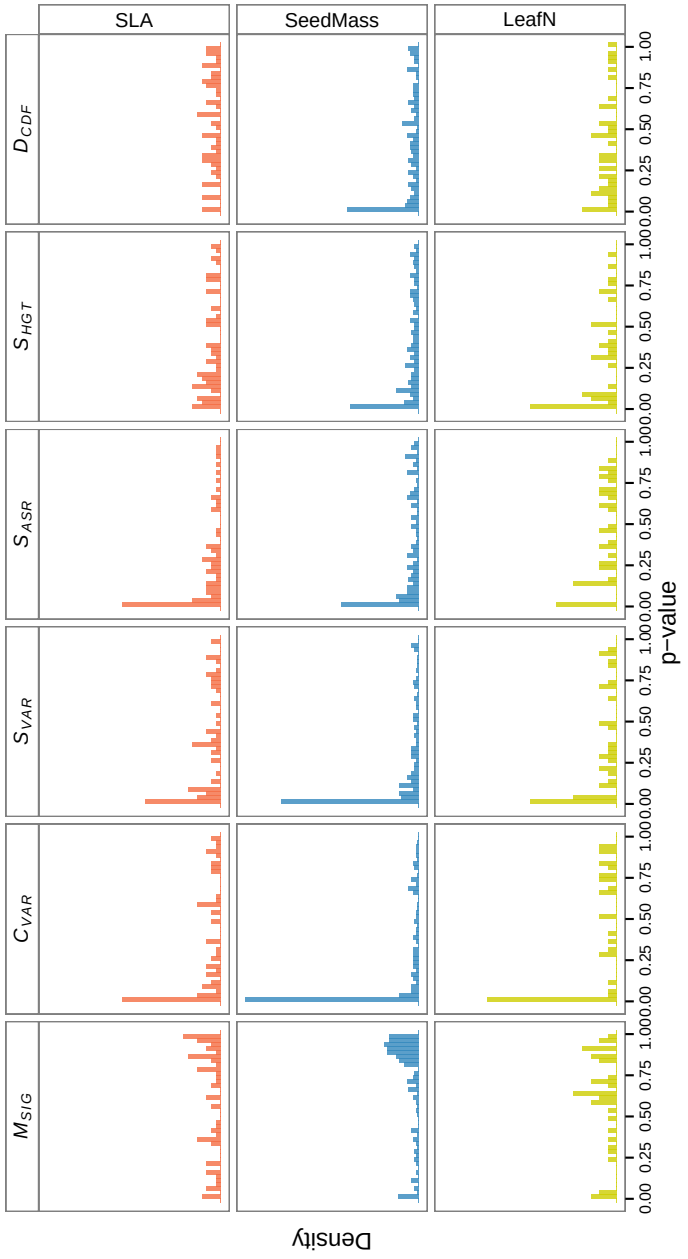

FIGURE S6: The distribution of p-values for our six test statistics over all 337 datasets in our study after fitting the models using MCMC. The p-values are from applying our model adequacy approach to the best supported of the three models (as evaluated with DIC). Many of the datasets deviate from the expectations under the best model along a variety of axes of variation. Deviations are particularly common for the coefficient of variation C_{VAR} and the slope of the contrasts against their expected variances S_{VAR}.

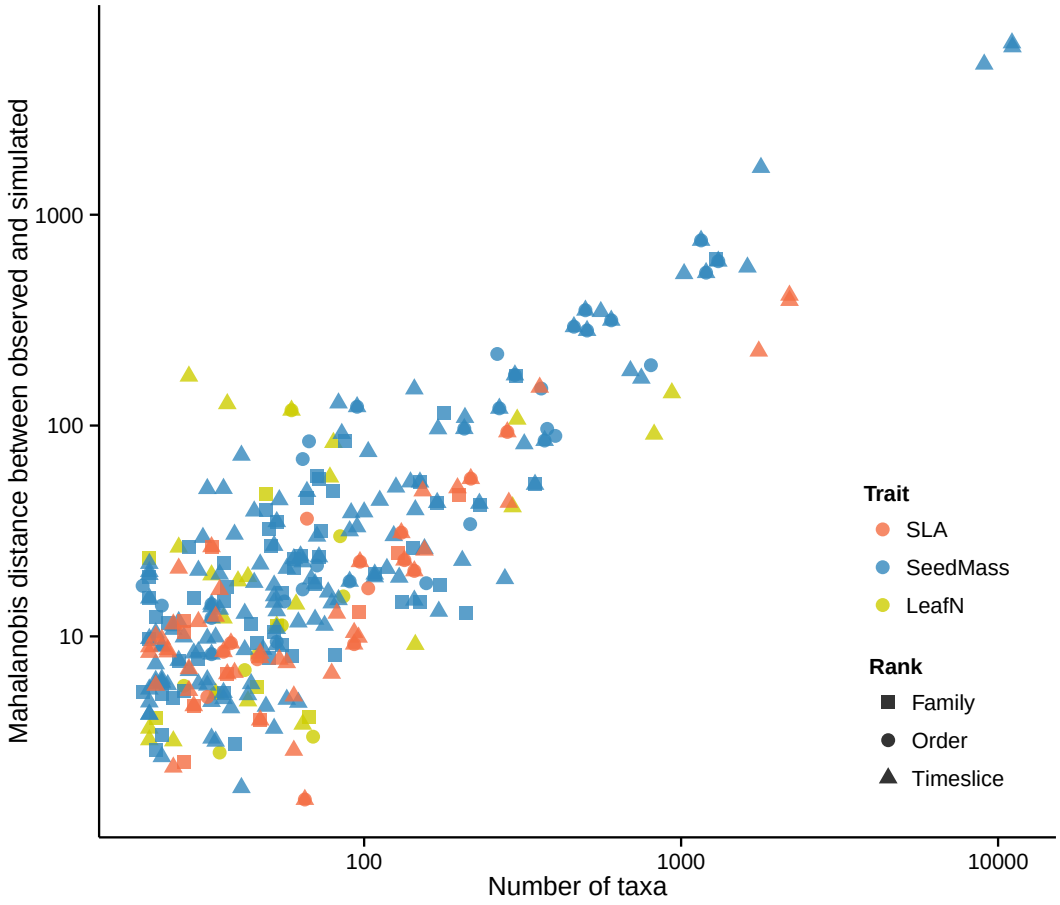

FIGURE S7: The relationship between clade size and a multivariate measure of model adequacy from the Bayesian analysis. The Mahalanobis distance is a scale-invariant metric that measures the distance between the observed and simulated test statistics, taking into account the covariance between test statistics. 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 clear relationship between the two–the larger the dataset, the stronger the evidence that the model does not capture variation in the data.