EvoPhylo: an R package for pre- and postprocessing of morphological data from relaxed clock Bayesian phylogenetics

Tiago R. Simões[1,*] Noah Greifer[2], Joëlle Barido-Sottani[3] and Stephanie E. Pierce[1]

[1]*Museum of Comparative Zoology & Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA.*

[2]*Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138, USA.*

[3]Institut de Biologie de l'ENS (IBENS), École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France.

*Corresponding Author: tsimoes@fas.harvard.edu

Abstract

1. Relaxed clock Bayesian evolutionary inference (BEI) enables the co-estimation of phylogenetic trees and evolutionary parameters associated with models of character and lineage evolution. Fast advances in new model developments over the past decade have boosted BEI as a major macroevolutionary analytical framework using morphological and/or molecular data across vastly different study systems. However, there is a limited availability of bioinformatic tools to pre- and post-process data from BEI, such as identifying morphological data partitions, or statistically testing and creating publication quality plots of evolutionary hypotheses using the output from BEI.

2. Here we introduce *EvoPhylo*, an R package to perform automated morphological character partitioning for phylogenetic analyses and analyze macroevolutionary parameter outputs from relaxed clock (time-calibrated) BEI.

3. We present the theoretical background behind *EvoPhylo*'s functions and analytical tools for evolutionary hypothesis testing, its potential uses, and interpretation of its results with a series of vignettes and links to a step-by-step tutorial.

4. *EvoPhylo* will facilitate utilization of Bayesian relaxed clocks as a tool for macroevolutionary inference across a wide range of users and fields of research, especially those that use morphological datasets.

Keywords: Bayesian phylogenetics, character partitioning, evolutionary rates, selection, diversification rates, morphology, R.

1. INTRODUCTION

Macroevolutionary research programs have historically relied upon the utilization of a given phylogenetic tree (or set of trees) to subsequently estimate the tempo and mode (rates and model) of morphological, ecological, and molecular traits (Morlon, 2014; Pennell & Harmon, 2013). These techniques, known collectively as phylogenetic comparative methods (Felsenstein, 1985), have revolutionized quantitative approaches to infer processes and patterns of evolution for a wide spectrum of living and fossil organisms across vastly different scales of time (Morlon, 2014; Pennell & Harmon, 2013; Slater & Harmon, 2013). In such approaches, evolutionary parameter estimates are obtained *a posteriori* from phylogenetic inference, and the structure of the phylogenetic tree (tree topology) and its branch lengths (as accumulated substitutions or as units of time) are used as input for downstream analyses and are necessarily treated as a fixed parameter. However, the true topology and branch lengths of phylogenetic trees are never known with certainty. Additionally, just as phylogenetic trees are necessary to estimate the tempo and mode of lineage and character evolution, understanding the tempo and mode of character and lineage evolution are also necessary to infer phylogenetic trees to begin with.

Bayesian evolutionary inference (BEI) using relaxed clocks circumvent such conundrums by jointly estimating tree topology and branch lengths along with evolutionary parameters, including divergence times, evolutionary rates, and rates of lineage diversification, using molecular data, morphological data, or both (Drummond et al., 2006; Gavryushkina et al., 2017; Höhna et al., 2016; Lee et al., 2014; A. Wright et al., 2020). However, until recently it was not feasible to conduct such analyses beyond relatively small datasets due to: 1) the high computational burden of estimating joint posterior probabilities of dozens of parameters; 2) the limited availability of tree and clock models concomitant with a limited understanding of the performance; and 3) limited bioinformatics tools to assess evolutionary parameters output by such analyses. Fortunately, the last decade was marked by increased academic access to high performance computing facilities, including the CIPRES Gateway (Miller et al., 2012). Additionally, there have been major advances on tree modeling, such as the fossilized birth-death (FBD) tree model and its skyline variant (SFBD), which allow speciation, extinction and fossilization parameters to vary across time bins (Gavryushkina et al., 2014; Heath et al., 2014; Stadler, 2010, 2011; Zhang et al., 2016). More recently, performance studies revealed that these models can provide accurate estimates of macroevolutionary parameters, including net diversification, turnover, and fossil sampling rates (Luo et al., 2020; Warnock et al., 2020). Relaxed clocks can also provide reliable rate estimates even with highly limited taxonomic sampling (Ho et al., 2005), and a variety of new clock models have been proposed (Bielejec et al., 2014; Fourment & Darling, 2018; Zhang, 2021). As a result of these advances, there has been a recent boost in macroevolution studies using BEI to infer evolutionary parameters for various modern and extinct lineages representing datasets of various compositions and sizes (King et al., 2017; Lee et al., 2013, 2014; Simões, Vernygora, et al., 2020; Simões & Pierce, 2021; A. Wright et al., 2020).

Bioinformatics tools to explore the rich amount of data output from BEI have also been thoroughly expanded. Such tools include software and packages to analyze the posterior trace files between multiple runs or MCMC chains, such as the standalone program Tracer (Rambaut et al., 2018) and the R package *RWTY* (Warren et al., 2017), or to visualize and plot divergence times and rates of evolution parameters on trees, including FigTree (Rambaut, 2018), DensiTree (R. R. Bouckaert, 2010), the R package *ggtree* (Yu et al., 2017). However, there are few tools currently available to extract, plot, summarize statistically, and conduct further downstream analyses from evolutionary parameters obtained from relaxed clock BEI. These include statistically testing the difference of evolution rates between clock partitions and/or evolutionary lineages, how such differences impact our understanding of the mode of selection upon those lineages, or the rate of diversification dynamics across time—but see RevGadgets (Tribble et al., 2022) for a recent implementation of the latter for outputs from the software package RevBayes (Höhna et al., 2016). Additionally, differently from molecular data—e.g., (Duchêne et al., 2014; Lanfear et al., 2016)—there are limited attempts to pre-process morphological datasets to detect data partitions that should be analyzed using independent evolutionary clock models for BEI.

Here we introduce *EvoPhylo,* an R package to perform pre- and post-processing of the input and output from BEI. It includes automated partitioning of phenotypic (i.e., morphological) character data for BEI, and statistical tools to plot and analyze macroevolutionary parameter outputs from clock (time-calibrated) BEI analyses. In this paper, we present the theoretical background behind *EvoPhylo* and describe its potential uses, overall functionality, and the interpretation of its results through demonstration with real datasets and links to online vignettes with step-by-step tutorials.

2. CHARACTER PARTITIONING

2.1 Clustering method

A common approach to data partitioning (i.e., clustering) is the extraction of Euclidean distances between data points, from which a distance matrix "D" is calculated, and subsequently used to detect data partitions (clusters) using K-means, or ordination approaches such as principal coordinate analysis (PCoA). Indeed, the first attempts to automatically partition morphological characters have explored these approaches (Goswami & Polly, 2010; Lanfear et al., 2016). However, recent studies have indicated that Euclidean distances can be extremely sensitive to missing data, and alternative choices such as Gower distances (Gower, 1971) provide more suitable alternatives for the handling of missing data (Lehmann et al., 2019; Lloyd, 2016). This issue creates a subsequent problem for estimating clusters using K-means, as the latter depends on a Euclidean-based distance matrix. Further, K-means are based on measuring the distance between samples and cluster centroids (i.e., the center of mass or mean vector of the cluster). The mean vector is particularly sensitive to outliers (as any other mean estimate) (Rencher & Christensen, 2012), making its use especially problematic for small-sized clusters or clusters of drastically different sizes, which are to be expected from most standard sized morphological datasets.

EvoPhylo uses Gower distances to create the inter-character distance matrix "D" and conducts a clustering analysis of morphological data with partitioning around medoids (PAM, also known as K-medoids), which can estimate clusters (i.e., partitions) using Gower distances, following its first implementation by Simões & Pierce (2021). PAM is analogous to K-means, but the resulting clusters are centered around medoids instead of around centroids, making them less sensitive to outliers and heterogeneous cluster sizes (Budiaji & Leisch, 2019; Rencher & Christensen, 2012).

To define how many clusters the data could be partitioned into, various PAM partitioning schemes are tested and the quality of each clustering scheme is determined using the silhouette index (Si) approach (Rousseeuw, 1987), a method that estimates how well an object falls within its cluster compared to other clusters (Fig. 1). The PAM partitioning schemes to be tested should range from K=2 to a large number of partitions (user-defined, default K = 10). The best partitioning scheme from PAM+Si can be exported into a Nexus file with the *cluster_to_nexus* function, including the list of characters and their respective partitions (Fig. 2). The contents can be copied and pasted directly into a Mr. Bayes commands block for a partitioned clock Bayesian inference analysis.

```
### 1. Generate distance matrix
#Load data matrix and
#produce a Gower distance matrix
d_matrix <- get_gower_dist(
  "DataMatrix.nex",
   numeric = FALSE)

### 2. Estimate optimal number of partitions
sw <- get_sil_widths(
   dist_matrix,
   max.k = 10)
plot(sw, color = "blue", size = 1)
```

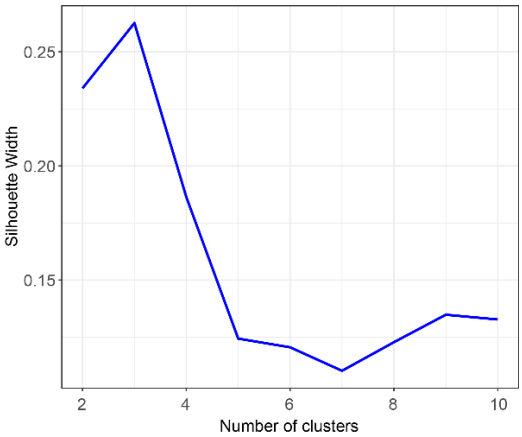


Fig. 1. Silhouette index plot indicating the higher quality of clustering when the number of partitions (k) = 3.

2.2 Selecting best candidate partitioning scheme

For further (and independent) testing of the quality of the chosen partitioning scheme, we also provide a graphic visualization approach based on a Barnes-Hut t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van Der Maaten & Hinton, 2008). More traditional ordination procedures, such as principal components analysis (PCA, for continuous data) or PCoA (for discrete data), can preserve the linear relationship between data points at a lower dimensionality. However, because those procedures try to preserve the local distances between data points, they become less efficient at characterizing the overall structure of high dimensional data. Here, it is more important to reduce the local linear distance between similar (neighboring) data points while maximizing the distance

between distant datapoints (Van Der Maaten & Hinton, 2008); for such cases, nonlinear ordination procedures are preferred for observing the overall data structure in a reduced number of dimensions. t-SNE has been demonstrated to be more efficient at preserving both local and global structures when reducing high dimensional data into only two or three dimensions compared to other nonlinear ordination procedures (Van Der Maaten, 2009), thus offering an important advantage over previously utilized graphic approaches to determine morphological clusters such as PCoA.

EvoPhylo combines PAM+Si clustering with t-SNE within the function *make_clusters*, by allowing the user to request displaying the distance between data points and in ordination space through the argument *tsne=TRUE*. Users can choose the representation of two or more dimensions and also the variable theta, which controls the speed/trade off accuracy of t-SNE calculations, through the *tsne_dim* and *tsne_theta* arguments, respectively.

EvoPhylo automatically colors individual data points in the t-SNE plots according to the partitioning scheme identified with PAM+Si, allowing users to quickly verify if both strategies converge on the number and composition of each character partition. This is the case with the example dataset used here from Simões & Pierce (2021) (Fig. 2). If there is a mismatch between the partitioning scheme from PAM+Si and that displayed in the t-SNE plots, we recommend re-plotting t-SNEs using another coloring scheme for the data points, such as one based on anatomically defined character partitions. The latter can be accomplished by directly utilizing arguments within the *Rtsne* function of the *Rtsne* package (Krijthe, 2015). If there is a closer correspondence between tSNEs and anatomical partitioning as compared to PAM+Si and tSNEs, it is reasonable to follow anatomical partitioning.

```
### 3. Calculate partitions (clusters)
#PAM+Si (plotted using t-SNE)
clusters <- make_clusters(
    dist_matrix,
    k = 3, tsne = TRUE,
    tsne_dim = 2)

plot(clusters, nrow = 1,
     max.overlaps = 12)

#Write partitions to Nexus file
cluster_to_nexus(clusters,
    file = "Clusters_Nexus.txt")
```

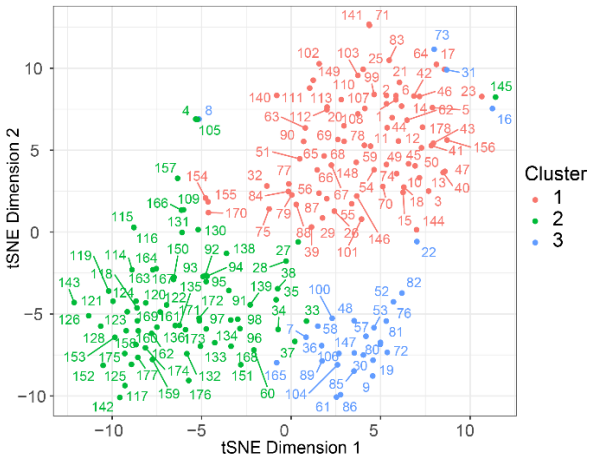


Fig. 2. Plot of identified morphological partitions using tSNE of the first two dimensions with data points colored according to the partitioning scheme determined by PAM+Si.

2.3 Data treatment and import

Categorical data (such as discrete morphological characters) should be treated as factors when imported to calculate character distances, as the symbols used to represent different character states are arbitrary (e.g., could be equally represented by letters, such as for DNA data). If continuous variables are used as phylogenetic characters, those should be read in from a separate file and treated as numeric data, since input values for each state (e.g., 0.234; 2.456; 3.567; etc.) represent true distances between data points.

Additionally, most morphological datasets have a portion of inapplicable or missing characters, which introduce problems to calculate distance matrices. Inapplicable and missing data (typically scored as "-" and "?", respectively) are interpreted as extra states relative to numerical symbols typically used for different character states ("0", "1", "2", etc.). Therefore, there are a few options users may follow for handling morphological phylogenetic datasets to account for inapplicable/missing data before importing it into *EvoPhylo*. Users may either convert inapplicable/missing to "NA" or they may choose to keep the original symbols.

As demonstrated by the example provided in the online vignette, converting inapplicable/missing conditions to "NA" introduces "NaN" scores to every pairwise comparison involving two characters with "NA" when calculating a distance matrix. Statistical tests and clustering methods cannot utilize such matrices with "NaN" as data entries, and so the removal of observations contributing to excessive NaN would have to be performed—such as done by the package *Claddis* (Lloyd, 2016) when calculating an inter-taxon distance matrices to estimate morphospace. However, removing observations with excessive inapplicable/missing data is not possible for character partitioning because each character in the dataset must be assigned to at least one partition (regardless of the amount of missing or inapplicable data). Furthermore, comparisons between any characters in which one character has an "NA" score will result in a distance of 0 between these same characters (Table 2 in the online vignette). Therefore, the implicit assumption with this strategy is that unknown characters contribute 0 distance (i.e., unknown states are assumed to be equal to the known states), which biases the distance matrix by minimizing the overall distance between characters to the lowest possible values.

Alternatively, users may keep the original inapplicable/missing data (although all must be represented by the same symbol, e.g., all as "?"), and such states will be treated as a distinct categorical variable relative to numeric symbols. As a result, pairwise comparisons with characters with unknown states avoid the introduction of 'NaN" in the distance matrix. This approach assumes that unknown states are always different from any known states, which will bias the distance matrix by increasing the overall distance between characters. Fortunately, however, Gower distances (as used here) are normalized by the number of variables in the dataset (number of taxa in this case) (Gower, 1971), which reduces this bias. For instance, in a simple comparison between two characters sampled from two taxa (A and B), e.g., character 6 (1,1) and character 7 (NA, 1) from the example in the online vignette, the raw distance between these characters is 1.0, but the Gower distance between them is 1/2 = 0.5. Therefore, we recommend this approach to

calculate inter-character distance matrices, which only requires users to convert all inapplicable/missing scores in their datasets to "?" symbols before importing into *EvoPhylo*.

We note, however, that there is no objective solution to the problem of inapplicable/missing data to estimate distance matrices, besides potentially negatively impacting the accuracy of phylogenetic analyses—e.g., (Vernygora et al., 2020; A. M. Wright & Hillis, 2014), but see Keating (2020). We thus suggest avoiding or removing such characters from morphological phylogenetic datasets whenever possible as a general good practice.

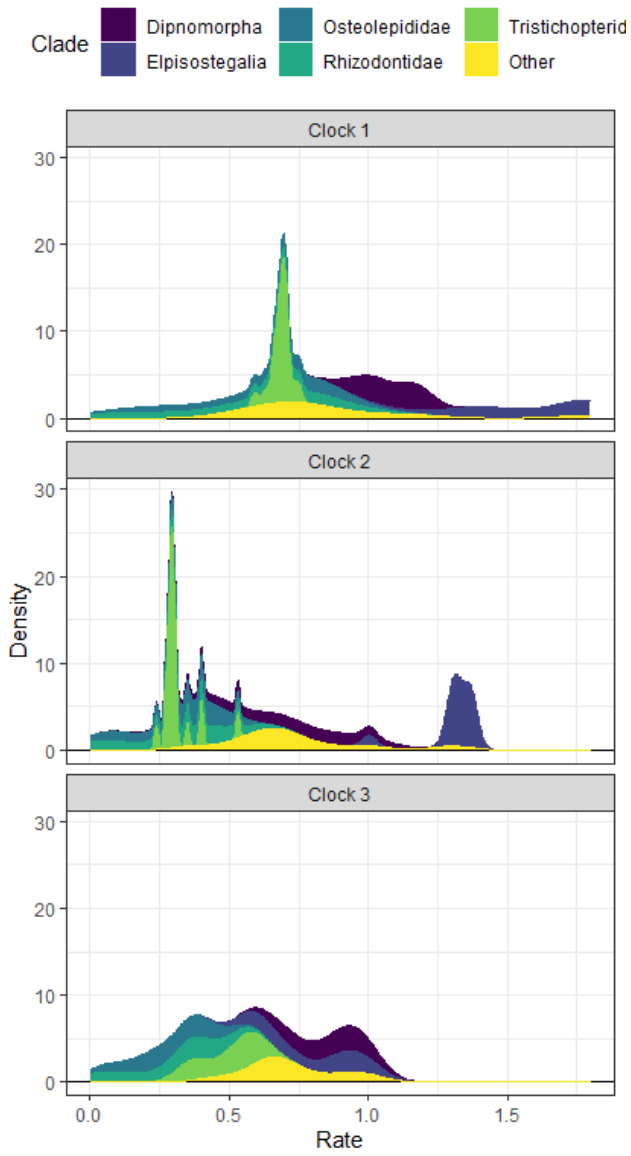
3. CLOCK RATES AND SELECTION MODE

With the assumption that morphological evolution is mostly driven by adaptive change, it is possible to infer the mode of natural selection operating upon particular regions of the phenotype (e.g., morphological or morphological partitions) and across distinct clades in a phylogeny as a function of their morphological evolutionary rates (Baker et al., 2016; Revell et al., 2012; Simões & Pierce, 2021; Venditti et al., 2011). Evolutionary rates that are significantly accelerated relative to the background rates provide support for positive or directional morphological selection in analogy with the d_N/d_S ratio in molecular evolution, whereas strongly decelerating rates indicate stabilizing selection, stasis or constraint (Baker et al., 2016; Yang, 2014). This concept was first applied to morphological traits using continuous data in phylogenetic comparative methods in the program BayesTraits (Baker et al., 2016) and later extended to discrete data and evolutionary rates estimated with Bayesian molecular or morphological clocks (Simões & Pierce, 2021), and it is the basis for inferring the strength and mode of selection in *EvoPhylo*.

The original approach in BayesTraits takes the clock rate on every tree branch (Δv), which is then compared to the background rate of evolution (Δb), forming the rate scalar ratio ($r = \Delta v / \Delta b$), as defined by Baker et al., (2016). This measure is equivalent to the interpretation of relative rates of character evolution produced by relaxed Bayesian clocks, in which estimates greater than 1 indicate rates above background rate levels (the base of the clock rate) and are therefore accelerating, whereas relative branch rate values less than 1 indicate values below background rate levels, implying a decrease in the rates of evolution in that branch (Ronquist et al., 2019).

To draw evolution rates from Bayesian trees and infer selection mode, users must first use the function *get_clockrate_table* to extract relative clock rate values from every branch of a relaxed clock Bayesian inference tree—i.e., median or mean rate values embedded in summary tree files produced by relaxed clock Bayesian inference. An argument *drop_dummyextant* is available to allow users to automatically remove a "dummy" extant taxon introduced for the offsetting of all tree node ages when analyzing fossil-only datasets that incorporate uncertainty in the age of every tip age—see discussions in Simões & Pierce (2021) for further details.

```
### 1. Import summary tree
tree<-treeio::read.mrbayes(
    "MultiClockTree/Tree_3p.t")

### 2. Get and export rate tables
RateTable_Medians_no_clades <-
    get_clockrate_table
    (tree, summary = "median")

write.csv(RateTable_Medians_no_clades,
    file="RateTable_Medians.csv")

### 3. Import rate table
#with custom clade memberships added
RateTable_Medians<- read.csv(
    "RateTable_Medians_Clades.csv",
    header = TRUE)

### 4. Get summary stats
#for each clade/clock partition
clockrate_summary(RateTable_Medians,
    "Sum_RateTable_Medians_Clades.csv",
    digits=2)

### 5. Plot rates by clock partition/clade
# with stacked plots (viridis)
clockrate_dens_plot(RateTable_Medians,
    stack = TRUE, nrow = 3,
    scales = "fixed")+
    ggplot2::scale_color_viridis_d() +
    ggplot2::scale_fill_viridis_d()

### 7. Get rate linear models
p1<- clockrate_reg_plot(RateTable_Medians,
    clock_x = 1, clock_y = 2)
p2<- clockrate_reg_plot(RateTable_Medians,
    clock_x = 1, clock_y = 3)
p3<- clockrate_reg_plot(RateTable_Medians,
    clock_x = 2, clock_y = 3)

p1 + p2 + p3 + plot_layout(nrow = 1)
```



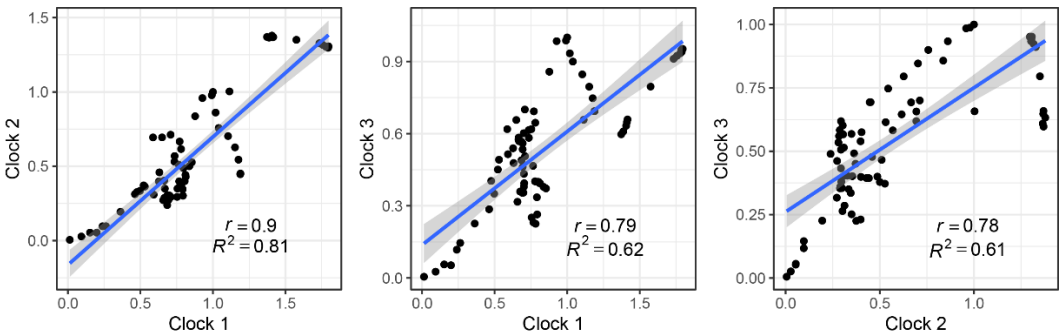


Fig. 3. Summary statistics and plots for clock (evolutionary) rates by clade and clock partitions.

At this stage, rate tables must have customizable clade names (specific for each dataset and tree topology). This can be done within R or by exporting rates tables to a CSV file (and edited in, e.g., Microsoft Excel) and manually adding a "clade" column using the tree node numbers as reference; a sample dataset of this kind is provided with *EvoPhylo* and can be called with *rate_table_clades_means*. The new rates tables with added clade names must then be used for downstream analyses. Detailed examples are provided in the online vignette.

```
### 1. Import and transform mean rate table with custom clade memberships
RateTable_Means<- read.csv("RateTable_Means_Clades.csv", header = TRUE)
RatesByClade <- clock_reshape(RateTable_Means)

### 2. Import all log (.p) files from all runs and combine them
Comb_posterior3p <- combine_log("LogFiles3p", burnin = 0.25, downsample = 2500)

### 3. Pairwise t-tests of Rate values
RateSign_tests<- get_pwt_rates(RateTable_Means, Comb_posterior3p)

### 4. Plot selection strength on the summary tree using thresholds for each clock partition
S1<-plot_treerates_sgn(tree, Comb_posterior3p,
  clock = 1,              #Show rates for clock partition 1
  summary = "mean",       #sets summary stats to get from summary tree nodes
  branch_size = 1.5, tip_size = 3,                          #sets size for tree elements
  xlim = c(-450, -260), nbreaks = 8, geo_size = list(3, 3),  #sets geoscale
  threshold = c("1 SD", "3 SD"))                            #sets selection threshold

#Repeat previous step for clock partitions 2 and 3 (objects S2 and S3)

S1|S2|S3
```

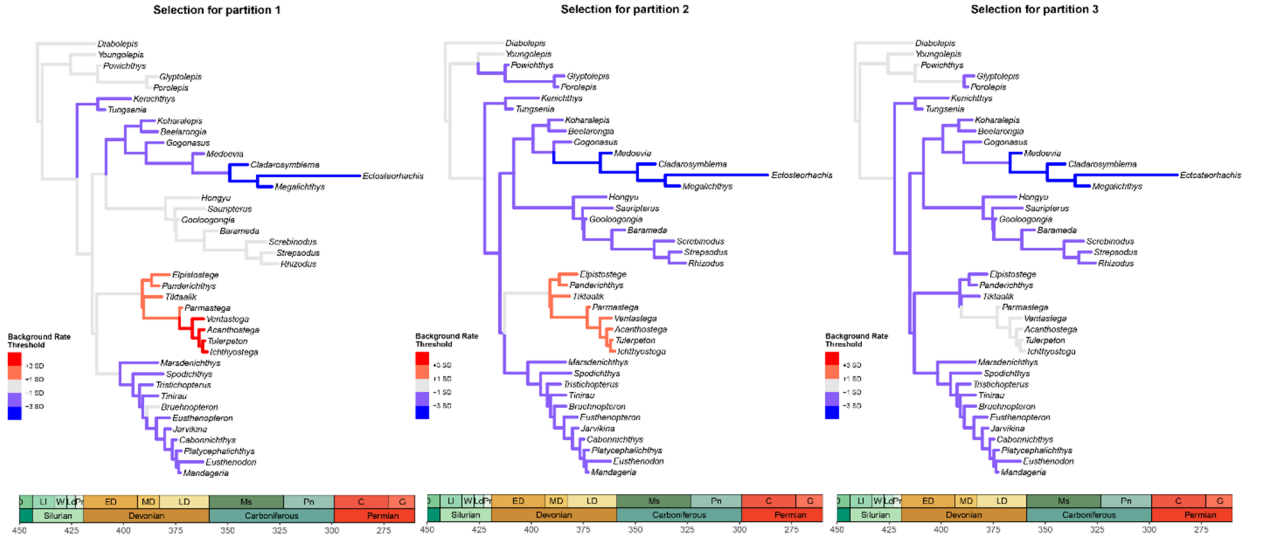


Fig. 4. Relative rates of evolution and inferred mode of selection across morphological partitions. Scale bars indicates evolutionary rate thresholds for inferring selection mode: 1 standard deviation (weak) and 3 standard deviation (very strong) evidence for positive (red spectrum) or stabilizing (blue spectrum) modes of selection at each branch for every morphological partition.

Summary statistics of evolutionary rates for each designated clade or clock partition can be extracted from the rates tables and summarized and plotted using the functions *clockrate_summary* and *clockrate_dens_plot*, respectively (Fig. 3). Linear regression models between clock rates are available through the *clockrate_reg_plot* function (Fig. 3), enabling the user to verify the degree of correlation between separate clock partitions. These correlations can be used as the basis to test, for instance, correlated evolution among separate morphological partitions and thus act as a test for evolutionary integration among such partitions (Simões et al., 2020). For plotting individual clock rates and their variance throughout branches in summary evolutionary trees, we suggest several functions available in the package *ggtree* (Yu et al., 2017).

In order to infer selection mode, users must obtain posterior estimates for the base of the clock rate value, which are reported in parameter log files from Bayesian inference software. Extracting this parameter from parameter files—and other parameters to be used later for FBD diversification rates (see more below)—requires importing all parameter files and combining them into a single file. This is done with the function *combine_log*, which also allows users to drop samples from generations in the beginning of each log file (i.e., discarded as burn-in) and/or downsampled to reduce the size of the output object (Fig. 4). Hence, *combine_log* is functionally analogous to LogCombiner from the BEAST2 software package (Bouckaert et al., 2019), but specifically targeted to parameter files produced by Mr. Bayes. In practice, users can also use LogCombiner to combine parameter log (.p) files from Mr. Bayes, but we chose to include a standalone function for this purpose to avoid dependency on external software and to conduct all analyses in this pipeline within the R environment.

Once rate tables (with customized clade names) and a single parameter file are available, users can deploy the *get_pwt_rates* function, which converts relative rates to absolute rate values and compares rates across every branch and every clock partition to the base of the clock rate (background rate), to measure the degree of rate deviation from background levels (Fig. 4). Thresholds must be defined to establish the degree of rate deviation from background levels that will be used to indicate whether branches and/or morphological partitions are significantly accelerating or decelerating. *EvoPhylo* allows users to utilize flexible thresholds that take into account the dispersion of the distribution of the base rates obtained from the posterior parameter files. For instance, Simões & Pierce (2021) established ±1 standard deviation (1σ) from the background mean rate as their threshold: a rate of evolution on a given branch greater than the mean background rate +1 standard deviation ($\Delta v > \mu_{\Delta v} + 1\sigma$) indicates an instance of positive selection; a rate of evolution on a branch less than the mean background rate -1 standard deviation ($\Delta v < \mu_{\Delta b} - 1\sigma$) indicates an instance of stabilizing selection or stasis; and a rate of evolution on a branch within 1 standard deviation of the mean background rate ($\mu_{\Delta b} - 1\sigma < \Delta v < \mu_{\Delta b} + 1\sigma$) indicates an evolutionary rate not significantly different from the null hypothesis of neutral evolution.

EvoPhylo allows users to compute multiple threshold levels across the tree using, e.g., one, two, three, or more standard deviations. Users can plot only one of these thresholds or all of them combined onto the evolutionary tree to assess the degree upon which clades are evolving faster or slower compared to background rates, with direct implications for interpreting the mode of

selection operating upon the morphological traits (Fig. 4). Hence, here we suggest the interpretation of the threshold values as: $\pm 1\sigma$ (p = 0.32), $\pm 2\sigma$ (p = 0.05), $\pm 3\sigma$ (p = 0.01) to indicate weak, strong, and very strong evidence for deviation from background rates, respectively. These thresholds can all be supplied to the *plot_treerates_sgn* function, which plots the summary Bayesian evolutionary tree across branches to infer selection mode.

```
### 1. Reshape combined log file from previous steps.
posterior3p_long <- FBD_reshape(Comb_posterior3p)

### 2. Summary stats for FBD parameters by time bin
t3.1 <- FBD_summary(posterior3p_long)

### 3. Test for assumptions: normality and homoscedasticity for FBD parameters
# Results = Shapiro-Wilk, Bartlett's  and Fligner-Killeen tests
t3.2 <- FBD_tests1(posterior3p_long)

### 4. Visualize deviations from normality and similarity of variances
FBD_normality_plot(posterior3p_long)

### 5. Test for significant FBD shifts between time bins for each FBD parameter
#Results = Pairwise t-tests and Mann-Whitney tests
t3.3 <- FBD_tests2(posterior3p_long)
```

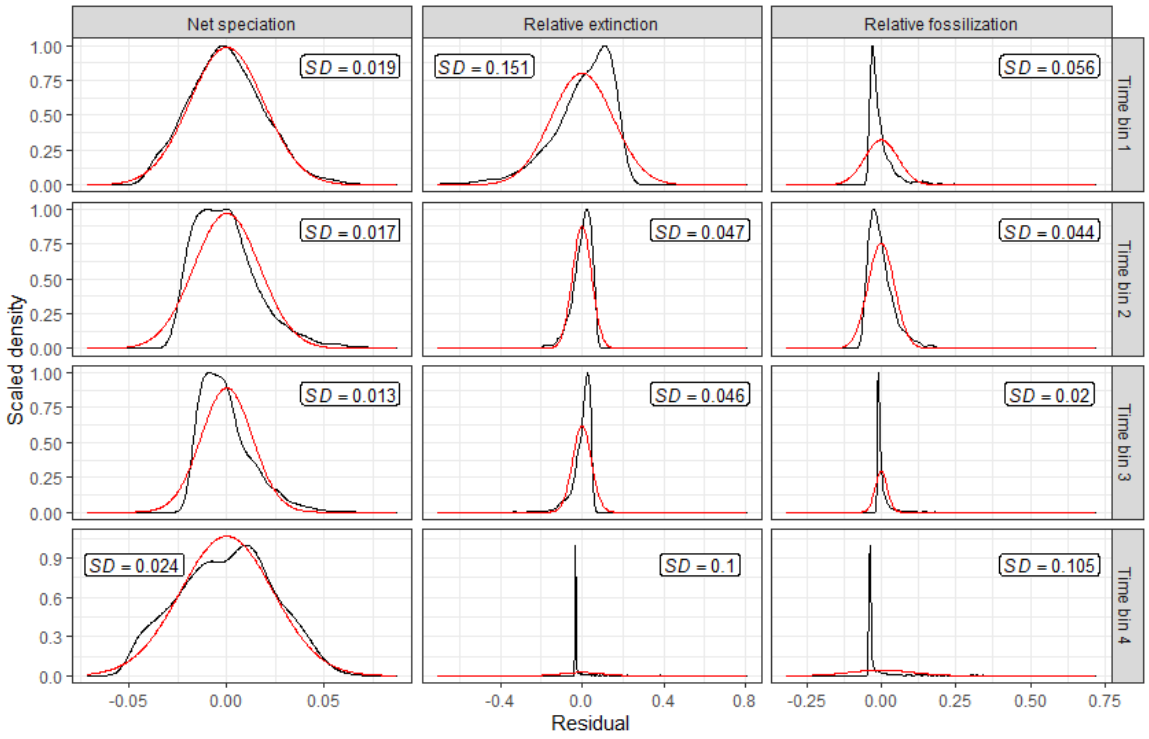


Fig. 5. Visualization of deviations from normality for each diversification parameter in the FBD model for every time bin.

4. DIVERSIFICATION RATES

The skyline variation of the fossilized birth-death tree model (SFBD) (Zhang et al., 2016) has made it possible to answer some of the most fundamental questions in macroevolution within an integrated Bayesian evolutionary inference framework and involves estimating net diversification, relative extinction (turnover), and relative fossilization across time bins. It relaxes the assumption of previous versions of the FBD model in which all diversification parameters are assumed to be constant across the tree, which is unrealistic for deep time studies. As with the birth-death skyline model (Stadler, 2011), the process starts at the root/origin (t_0 *or* t_{mtca}) and has a number (l) of rate shifting times (t_i) [t_i ($i = 1, \ldots, l$)]. The cutoff time x_{cut} represents the time after which no more fossils are sampled, and all lineages lead to extant taxa. FBD parameters must be constant within each time interval $t_l - t_{l-1}$ (or time bins), but they are allowed to vary across them. In its current implementation, the specific rate shift time points must be prespecified by the user. The output of SFBD analyses includes posterior estimates for each FBD parameter for every time bin, thus revealing fundamental aspects of shift in organismal diversity rates across time.

```
### 5. Plot the distribution of each FBD parameter by time bin with a violin plot
p1 <- FBD_dens_plot(posterior3p_long, parameter = "net_speciation",
                    type = "violin", stack = FALSE, color = "red")
p2 <- FBD_dens_plot(posterior3p_long, parameter = "relative_extinction",
                    type = "violin", stack = FALSE, color = "cyan3")
p3 <- FBD_dens_plot(posterior3p_long, parameter = "relative_fossilization",
                    type = "violin", stack = FALSE, color = "green3")
library(patchwork)
p1 + p2 + p3 + plot_layout(nrow = 1)
```

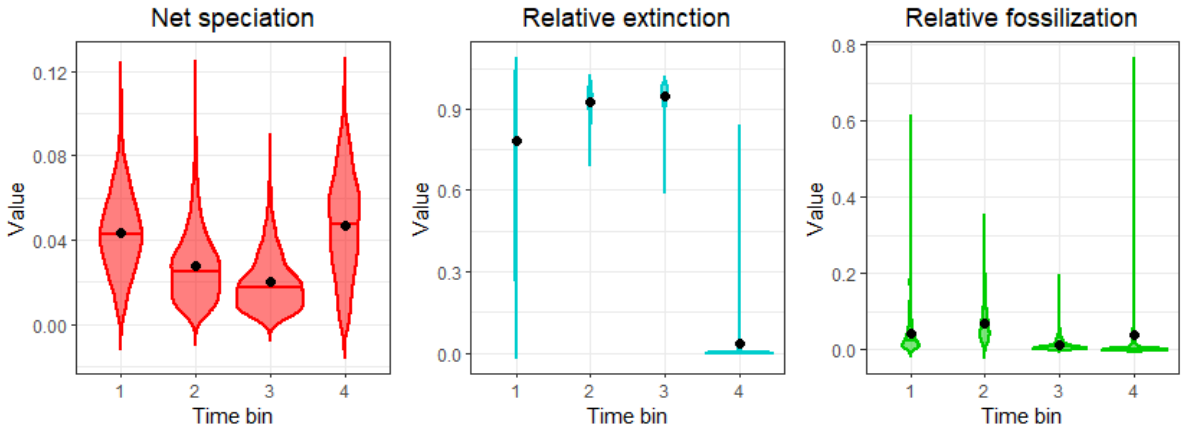


Fig. 6. Visualization of posterior estimates for each diversification parameter in the FBD model for every time bin.

EvoPhylo includes specific functions to combine and assess posterior parameters estimates from parameter log files (*combine_log*, as described above), including FBD parameters. Using the *FBD_summary* function to assess the combined parameter log file, users can produce a summary table of each specific FBD parameter for every time bin. Subsequently, users can use *FBD_tests1* to assess the normality of the distribution for each FBD parameter in each time bin using the Shapiro-Wilk normality test and visual assessment of data distribution using *FBD_normality_plot*

(Fig. 5). Additionally, *FBD_tests1* also runs a Bartlett and Fligner-Killeen tests of homogeneity of variances to assess homoscedasticity in the data. Finally, for testing between significant parameter rate shifts across time bins, *EvoPhylo* provides fast outputs of parametric (pairwise t-tests) and nonparametric (pairwise Wilcoxon rank sum, or Mann-Whitney) tests through the function *FBD_tests2*. To observe the final distribution of FBD parameters across each time bin, users can deploy *FBD_dens_plot* for each parameter of interest, using different plotting styles passed through *ggplot2* (Wickham, 2016) (Fig. 6).

5. CONCLUSIONS

Relaxed clock Bayesian evolutionary inference (BEI) is a powerful multivariate statistical approach which enables jointly estimating tree topology and macroevolutionary parameters, such as divergence times, evolutionary rates, and rates of lineage diversification. Several advances in the past decade have made BEI increasingly feasible computationally and parameter rich by the incorporation of a vast array of new trees and clock models. As a result, BEI has been increasingly adopted by evolutionary biologists working with molecular, morphological, and combined datasets to estimate time-calibrated trees and macroevolutionary dynamics across the tree of life. However, the development of bioinformatics tools to preprocess morphological data and postprocess evolutionary parameter estimates from BEI have been somewhat limited.

Here we introduce *EvoPhylo*, an R package to extract, plot, statistically summarize, and conduct further downstream analyses from evolutionary parameters obtained from relaxed clock BEI. This includes: automatically detecting partitions in morphological datasets, creating plots and summary statistics for clade and partition specific rates of morphological evolution, inferring significant shifts in evolutionary rates to infer the mode of selection across lineages and morphological partitions, and creating plots and statistically testing for shift in diversification parameters of the fossilized birth death model (net diversification, relative extinction, and relative fossilization) across time. The first version of *EvoPhylo* (v. 0.1) is designed to work with input and output data from the widely used software Mr. Bayes (Ronquist et al., 2012), but an upcoming release will expand its functionalities to also work with output data from the BEAST2 (Bouckaert et al., 2019) software package. *EvoPhylo* will thus facilitate macroevolutionary analyses using Bayesian relaxed clocks for a wide range of users and fields of research, especially those that use morphological datasets.

5.1 Dependencies

Evophylo depends on several R packages, in particular, *ape* (Paradis & Schliep, 2019), *cluster* (Maechler et al., 2012), *deeptime* (Gearty, 2021), *ggplot2* (Wickham, 2016), *ggrepel* (Slowikowski et al., 2018), *ggtree* (Yu et al., 2017), *patchwork* (Pedersen, 2019), *treeio* (Wang et al., 2020), *Rtsne* (Krijthe, 2015), and *unglue* (Fabri, 2020).

ACKNOWLEDGMENTS

We are thankful to the input provided by the members of the Pierce lab during the conduction of this project, as well as discussions with C. Zhang. We thank the provision of a National Sciences and Engineering Council of Canada (NSERC) Postdoctoral Fellowship to TRS. We thank the provision of an European Union's Horizon 2020 Research and Innovation Programme under the Marie Sklodowska-Curie grant agreement No. 101022928 to J.B.S. Additional funds were provided by Harvard University to S.E.P.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

AUTHOR'S CONTRIBUTIONS

T.R.S. and S.E.P. conceptualize the project. T.R.S., N.G., and J.B-S contributed with code and examples. T.R.S. drafted the manuscript. All authors contributed with discussions, editing, and approved the final version of the manuscript.

PEER REVIEW

The peer review history for this article is available at XXX.

DATA AVAILABILITY STATEMENT

EvoPhylo is hosted on CRAN (https://cran.r-project.org/package=evophylo) and available on GitHub (https://github.com/tiago-simoes/evophylo). All example datasets are freely available and come bundled with the R package.

ORCID

Tiago R. Simões: https://orcid.org/0000-0003-4716-649X

Noah Greifer: https://orcid.org/0000-0003-3067-7154

Joëlle Barido-Sottani: https://orcid.org/0000-0002-5220-5468

Stephanie E. Pierce: https://orcid.org/0000-0003-0717-1841

REFERENCES

Baker, J., Meade, A., Pagel, M., & Venditti, C. (2016). Positive phenotypic selection inferred from phylogenies. *Biological Journal of the Linnean Society. Linnean Society of London*, *118*(1), 95–115.

Bielejec, F., Lemey, P., Baele, G., Rambaut, A., & Suchard, M. A. (2014). Inferring heterogeneous evolutionary processes through time: from sequence substitution to phylogeography. *Systematic Biology*, *63*(4), 493–504.

Bouckaert, R. R. (2010). DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics* , *26*(10), 1372–1373.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., Matschiner, M., Mendes, F. K., Müller, N. F., Ogilvie, H. A., du Plessis, L., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., … Drummond, A. J. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, *15*(4), e1006650.

Budiaji, W., & Leisch, F. (2019). Simple K-Medoids Partitioning Algorithm for Mixed Variable Data. *Algorithms*, *12*(9), 177.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., & Rambaut, A. (2006). Relaxed Phylogenetics and Dating with Confidence. *PLoS Biology*, *4*(5), e88.

Duchêne, S., Molak, M., & Ho, S. Y. W. (2014). ClockstaR: choosing the number of relaxed-clock models in molecular phylogenetic analysis. *Bioinformatics* , *30*(7), 1017–1019.

Fabri, A. (2020). unglue: Extract Matched Substrings Using a Pattern. *R Package Version 0.1.0*. https://CRAN.R-project.org/package=unglue

Felsenstein, J. (1985). Phylogenies and the Comparative Method. *The American Naturalist*, *125*(1), 1–15.

Fourment, M., & Darling, A. E. (2018). Local and relaxed clocks: the best of both worlds. *PeerJ*, *6*, e5140.

Gavryushkina, A., Heath, T. A., Ksepka, D. T., Stadler, T., Welch, D., & Drummond, A. J. (2017). Bayesian Total-Evidence Dating Reveals the Recent Crown Radiation of Penguins. *Systematic Biology*, *66*(1), 57–73.

Gavryushkina, A., Welch, D., Stadler, T., & Drummond, A. J. (2014). Bayesian Inference of Sampled Ancestor Trees for Epidemiology and Fossil Calibration. *PLoS Computational Biology*, *10*(12), e1003919.

Gearty, W. (2021). deeptime: Plotting Tools for Anyone Working in Deep Time. *R Package Version 0. 0.5*. https://CRAN.R-project.org/package=deeptime

Goswami, A., & Polly, P. D. (2010). The influence of character correlations on phylogenetic analyses: a case study of the carnivoran cranium. In A. Goswami & A. Friscia (Eds.), *Carnivoran Evolution: New Views on Phylogeny, Form and Function* (pp. 141–164). Cambridge University Press.

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, *27*(4), 857–871.

Heath, T. A., Huelsenbeck, J. P., & Stadler, T. (2014). The fossilized birth–death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences*, *111*(29), 2957–2966.

Ho, S. Y. W., Phillips, M. J., Drummond, A. J., & Cooper, A. (2005). Accuracy of rate estimation using relaxed-clock models with a critical focus on the early metazoan radiation. *Molecular Biology and Evolution*, *22*(5), 1355–1363.

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P., & Ronquist, F. (2016). RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language. *Systematic Biology*, *65*(4), 726–736.

Keating, J. N., Sansom, R. S., Sutton, M. D., Knight, C. G., & Garwood, R. J. (2020). Morphological Phylogenetics Evaluated Using Novel Evolutionary Simulations. *Systematic Biology*, *69*(5), 897–912.

King, B., Qiao, T., Lee, M. S. Y., Zhu, M., & Long, J. A. (2017). Bayesian Morphological Clock Methods Resurrect Placoderm Monophyly and Reveal Rapid Early Evolution in Jawed Vertebrates. *Systematic Biology*, *66*(4), 499–516.

Krijthe, J. H. (2015). Rtsne: T-distributed stochastic neighbor embedding using Barnes-Hut implementation. *R Package Version 0. 13*. https://github.com/jkrijthe/Rtsne

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., & Calcott, B. (2016). PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses. *Molecular Biology and Evolution*, *34*(3), 772–773.

Lee, M. S. Y., Cau, A., Naish, D., & Dyke, G. J. (2014). Sustained miniaturization and anatomical innovation in the dinosaurian ancestors of birds. *Science*, *345*(6196), 562–566.

Lee, M. S. Y., Soubrier, J., & Edgecombe, G. D. (2013). Rates of phenotypic and genomic evolution during the Cambrian explosion. *Current Biology: CB*, *23*(19), 1889–1895.

Lehmann, O. E. R., Ezcurra, M. D., Butler, R. J., & Lloyd, G. T. (2019). Biases with the Generalized Euclidean Distance measure in disparity analyses with high levels of missing data. *Palaeontology*, *62*(5), 837-849.

Lloyd, G. T. (2016). Estimating morphological diversity and tempo with discrete character-taxon matrices: implementation, challenges, progress, and future directions. *Biological Journal of the Linnean Society. Linnean Society of London*, *118*(1), 131–151.

Luo, A., Duchêne, D. A., Zhang, C., Zhu, C.-D., & Ho, S. Y. W. (2020). A Simulation-Based Evaluation of Tip-Dating Under the Fossilized Birth–Death Process. *Systematic Biology*, *69*(2), 325–344.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., & Others. (2012). Cluster: cluster analysis basics and extensions. *R Package Version 2.1.3*. https://CRAN.R-project.org/package=cluster

Miller, M. A., Pfeiffer, W., & Schwartz, T. (2012). The CIPRES science gateway: enabling high-impact science for phylogenetics researchers with limited resources. *Proceedings of the 1st*

Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the EXtreme to the Campus and Beyond, 1–8.

Morlon, H. (2014). Phylogenetic approaches for studying diversification. *Ecology Letters*, *17*(4), 508–525.

Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* , *35*(3), 526–528.

Pedersen, T. L. (2019). patchwork: The Composer of Plots. R package version 1.0. 0. *R Package Version 1.1.1*. https://CRAN.R-project.org/package=patchwork

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

Rambaut, A. (2018). *FigTree v1.4*. http://tree.bio.ed.ac.uk/software/figtree/

Rambaut, A., Suchard, M. A., Xie, D., & Drummond, A. J. (2018). *Tracer v1.7*. http://beast.bio.ed.ac.uk/Tracer

Rencher, A. C., & Christensen, W. F. (2012). *Methods of Multivariate Analysis* (Vol. 758). John Wiley & Sons.

Revell, L. J., Mahler, D. L., Peres-Neto, P. R., & Redelings, B. D. (2012). A new phylogenetic method for identifying exceptional phenotypic diversification. *Evolution; International Journal of Organic Evolution*, *66*(1), 135–146.

Ronquist, F., Huelsenbeck, J., Teslenko, M., & Nylander, J. A. A. (2019). *MrBayes version 3.2 manual: tutorials and model summaries*. https://nbisweden.github.io/MrBayes/manual.html

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., & Huelsenbeck, J. P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, *61*(3), 539–542.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Simões, T. R., Caldwell, M. W., & Pierce, S. E. (2020). Sphenodontian phylogeny and the impact of model choice in Bayesian morphological clock estimates of divergence times and evolutionary rates. *BMC Biology*, *18*, 191.

Simões, T. R., & Pierce, S. E. (2021). Sustained High Rates of Morphological Evolution During the Rise of Tetrapods. *Nature Ecology & Evolution*, *5*, 1403–1414.

Simões, T. R., Vernygora, O. V., Caldwell, M. W., & Pierce, S. E. (2020). Megaevolutionary dynamics and the timing of evolutionary innovation in reptiles. *Nature Communications*, *11*, 3322.

Slater, G. J., & Harmon, L. J. (2013). Unifying fossils and phylogenies for comparative analyses of diversification and trait evolution. *Methods in Ecology and Evolution / British Ecological Society*, *4*(8), 699–702.

Slowikowski, K., Schep, A., Hughes, S., Lukauskas, S., Irisson, J.-O., Kamvar, Z. N., Ryan, T., Christophe, D., Hiroaki, Y., Gramme, P., & Others. (2018). ggrepel: Automatically Position Non-Overlapping Text Labels with "ggplot2." *R Package Version 0.9.1*. https://CRAN.R-project.org/package=ggrepel

Stadler, T. (2010). Sampling-through-time in birth–death trees. *Journal of Theoretical Biology*, *267*(3), 396–404.

Stadler, T. (2011). Mammalian phylogeny reveals recent diversification rate shifts. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(15), 6187–6192.

Tribble, C. M., Freyman, W. A., Landis, M. J., Lim, J. Y., Barido-Sottani, J., Kopperud, B. T., Höhna, S., & May, M. R. (2022). RevGadgets: An R package for visualizing Bayesian phylogenetic analyses from RevBayes. *Methods in Ecology and Evolution / British Ecological Society*, *13*(2), 314–323.

Van Der Maaten, L. (2009). *Learning a parametric embedding by preserving local structure*.

Van Der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research: JMLR*, *9*(Nov), 2579–2605.

Venditti, C., Meade, A., & Pagel, M. (2011). Multiple routes to mammalian diversity. *Nature*, *479*(7373), 393–396.

Vernygora, O. V., Simões, T. R., & Campbell, E. O. (2020). Evaluating the Performance of Probabilistic Algorithms for Phylogenetic Analysis of Big Morphological Datasets: A Simulation Study. *Systematic Biology*, *69*(6), 1088–1105.

Wang, L.-G., Lam, T. T.-Y., Xu, S., Dai, Z., Zhou, L., Feng, T., Guo, P., Dunn, C. W., Jones, B. R., Bradley, T., Zhu, H., Guan, Y., Jiang, Y., & Yu, G. (2020). Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. *Molecular Biology and Evolution*, *37*(2), 599–603.

Warnock, R. C. M., Heath, T. A., & Stadler, T. (2020). Assessing the impact of incomplete species sampling on estimates of speciation and extinction rates. *Paleobiology*, *46*(2), 137–157.

Warren, D. L., Geneva, A. J., & Lanfear, R. (2017). RWTY (R We There Yet): An R Package for Examining Convergence of Bayesian Phylogenetic Analyses. *Molecular Biology and Evolution*, *34*(4), 1016–1020.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag.

Wright, A. M., & Hillis, D. M. (2014). Bayesian Analysis Using a Simple Likelihood Model Outperforms Parsimony for Estimation of Phylogeny from Discrete Morphological Data. *PloS One*, *9*(10), e109210.

Wright, A., Wagner, P., & Wright, D. (2020). *Testing character-evolution models in phylogenetic paleobiology: a case study with Cambrian echinoderms*. https://doi.org/10.32942/osf.io/ykzg5

Yang, Z. (2014). *Molecular evolution: a statistical approach*. Oxford University Press.

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2017). Ggtree : An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution / British Ecological Society*, *8*(1), 28–36.

581 Zhang, C. (2021). Selecting and averaging relaxed clock models in Bayesian tip dating of
582 Mesozoic birds. *Paleobiology*, 1–13.
583 Zhang, C., Stadler, T., Klopfstein, S., Heath, T. A., & Ronquist, F. (2016). Total-Evidence Dating
584 under the Fossilized Birth–Death Process. *Systematic Biology*, *65*(2), 228–249.
585