PhyloMAd

v1.0 Manual

by David A. Duchêne, Sebastian Duchêne, and Simon Y.W. Ho

Contents

1. Citation, Availability, and Licence

The recommended citation for PhyloMAd is:

- Duchêne DA, Duchêne S, Ho SYW. In prep. PhyloMAd: Efficient assessment of phylogenomic model adequacy.

If you have used the combination of data-based statistics known to be informative under the PhyloMAd framework, please cite:

- Duchêne DA, Duchêne S, Ho SYW. In prep. The parallel between performance and adequacy of substitution models in phylogenomics.

If you have used the suggested interpretation of the χ^2 statistic for compositional heterogeneity, please cite:

- Duchêne DA, Duchêne S, Ho SYW. 2017. New statistical criteria detect phylogenetic bias caused by compositional heterogeneity. Mol Biol Evol. 34:1529–1534.

2. Support

Please report any bugs as issues in the github repository, or contact David A. Duchêne (david.duchene@sydney.edu.au) for troubleshooting or support.

3. Overview

Statistical phylogenetic methods rely on models of the evolutionary process. For example, phylogenetic analysis often depends on nucleotide substitution models, whereas molecular dating employs models of rate variation across branches (clock models). However, these models sometimes provide poor descriptions of biological reality and lead to biased estimates of the phylogeny and other parameters (Goldman 1993; Minin et al. 2003; Ripplinger and Sullivan 2008).

PhyloMAd assesses the adequacy of models in phylogenetics. This is distinct from comparing the relative fit of different candidate models (e.g., using the Akaike information criterion), because it allows the assessment of the individual merits of a model. To perform this assessment, PhyloMAd relies on using simulations to generate data sets that resemble the empirical data, and calculating descriptive statistics to identify differences between the data coming from the model (i.e., the simulations) and the empirical data. The simulations are also known as predictive simulations, or posterior predictive simulations when the assessment is made in a Bayesian framework. If the data originating from the model are dissimilar to the empirical data, then the model is rejected.

Many test statistics have been proposed for assessing model adequacy and they fall into two major groups. Data-based statistics describe features of the data set itself (here the data set is assumed to be an alignment of nucleotide or amino acid sequences; Fig. 1a). In contrast, inference-based statistics are calculated from the estimates made using the data and the candidate model (Fig. 1b; Brown 2014). Test statistics have been proposed for assessing substitution models, clock models (Fig. 1c), and a diversity of other models used in phylogenetic analysis (e.g., the tree prior or the multi-species coalescent; Goldman 1993; Reid et al. 2014; Duchêne et al. 2015). Selecting and interpreting test statistics remains an active area of research (e.g. Duchêne et al. 2017), but section 7 of this manual describes some of them and the interpretation of their results.

Assessment and interpretation of results in PhyloMAd is performed through a graphical user interface. The software can be used for fast assessment of the adequacy of the most commonly used models of substitution and among-lineage rate variation (clock models). PhyloMAd implements a fast method of assessing substitution models using maximum likelihood, which is tailored for analyses of large, multi-locus data sets. The results are reported in the form of summaries and graphical plots.

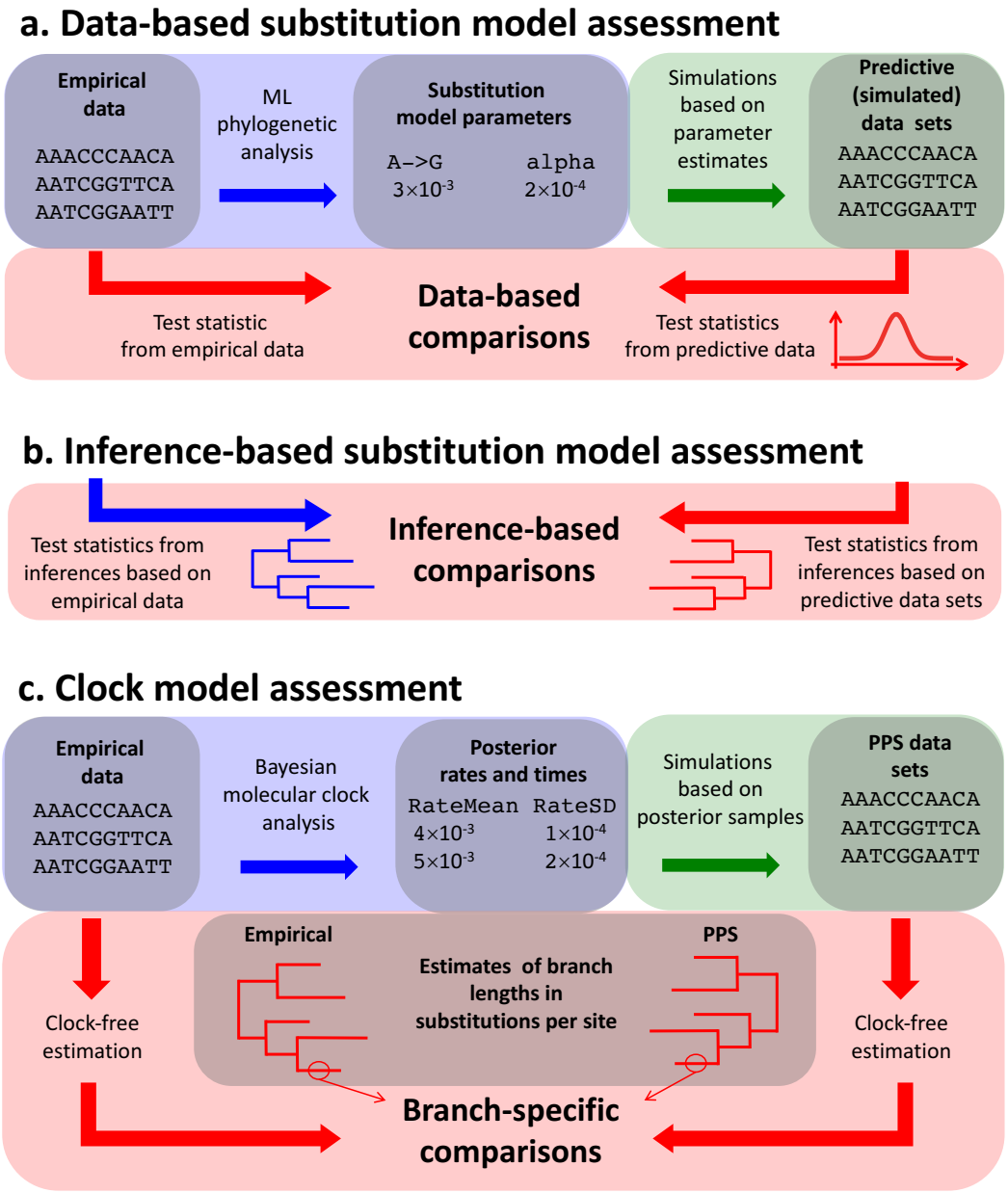

Figure 1. Three of the existing approaches to using predictive simulations to assess model adequacy in phylogenetics. (a) Data-based methods of assessment use characteristics of the data for model assessment, like the multinomial likelihood or the GC content. (b) Inference-based methods focus on characteristics of the estimates from the data, such as the tree length or the mean branch support across the inferred topologies. (c) The method for assessing Bayesian clock models uses estimates from clock-free methods. The number of substitutions per site expected along each branch under the clock model are then compared with those inferred in a clock-free analysis.

4. Download and Installation

PhyloMAd requires that the R statistical computing language (version ≥3.2) is installed. R is freely available from the R Project website (www.r-project.org).

There are three ways to obtain PhyloMAd:

1. Download and unzip the PhyloMAd repository by clicking the following link: https://github.com/duchene/phylomad/archive/master.zip

2. Go to https://github.com/duchene/phylomad and click on the *Clone or download* button.

3. Open a bash shell and type the following (assuming that git is installed):

```
git clone https://github.com/duchene/modadclocks.git
```

After it has been downloaded, PhyloMAd can be executed by opening the folder and double-clicking the icon according to the platform that you are using. For Mac machines, click on *runMac.command* and for Windows click on *runWin.vbs*. If needed, this will install all of the required R packages, so when the software is first opened on a given machine it might take several minutes.

If you have difficulty opening the program, you might want to try opening a bash shell, setting your directory to the PhyloMAd folder and executing the R script by hand.

```
cd path_to_PhyloMAd
Rscript phylomad.Rscript
```

On Mac machines, PhyloMAd will open a terminal window that will log the progress. In Windows, a log file with the screen progress will be saved in the main PhyloMAd folder. This file is not needed other than for checking progress, so it is safe to delete it when the program is closed. It is also safe to close all the windows associated with PhyloMAd when analyses are complete.

Once you have opened PhyloMAd, start by selecting the type of model that you wish to assess using the selection panel on the left of the screen.

5. Assessing Substitution Models

This section will take you through the options for assessing substitution models.

a. *Data selection.* The **Data tab** allows the selection of the data set for which the model will be assessed. Sequence alignments can be selected by pressing the *Browse…* button and navigating to the folder containing the data. Multiple alignments, corresponding to multiple partition subsets, can be selected simultaneously if they are in the same folder.

b. *File format.* The default format is NEXUS, but Phylip and FASTA formats can be selected using the buttons below the file input.

c. *Tree file selection.* It might be appropriate to fix the tree topology, for example, if the topology is known from independent evidence. The tree file can be selected using the *Browse* button below the space for selecting the data file. The default is to estimate the tree in every analysis, so if the tree is to be fixed then the tree format needs to be selected in this section.

d. *Model.* The model to be assessed can be selected from the **Model tab** using a drop-down menu. PhyloMAd allows the assessment of nucleotide substitution models of the general time-reversible (GTR) family, and four commonly used amino acid models, with the default option of including gamma-distributed rates across sites below the model drop-down menu.

e. *Test statistics.* The test statistics that will be calculated for model assessment can be selected from the **Test statistics tab** (Fig. 2). Details about each of the available statistics can be found in section 7 of this manual.

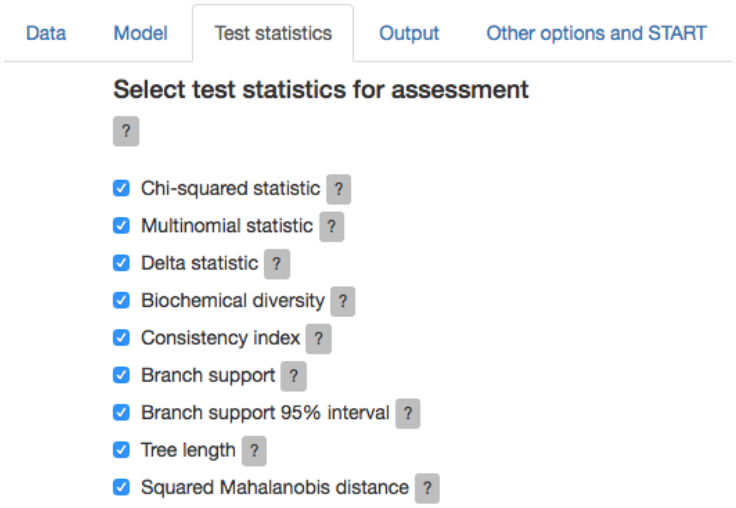

Figure 2. Screenshot of the checkboxes to select test statistics for assessment. The Squared Mahalanobis distance is a summary statistic for the other test statistics selected in this tab.

f. *Output.* The output can include several components. The default options will make, for each locus in the data set, a folder containing the metrics of adequacy and histograms showing the results.

The metrics of adequacy for individual loci include two metrics of the distance between the empirical test statistic and those of the predictive data sets: one is the tail area probability (the proportion of test statistics for simulations that are lower than that calculated for the empirical data), and the other is the number of standard deviations of the predictive distribution between the mean of the distribution and the test statistic for the empirical data. The table also includes the test statistics calculated for the empirical data.

The test plots option, which is selected by default, leads to histograms of the test statistics calculated for predictive data, with the statistic calculated for the empirical data shown by a red line. These plots are useful for qualitative assessment of the difference between simulations and the empirical data.

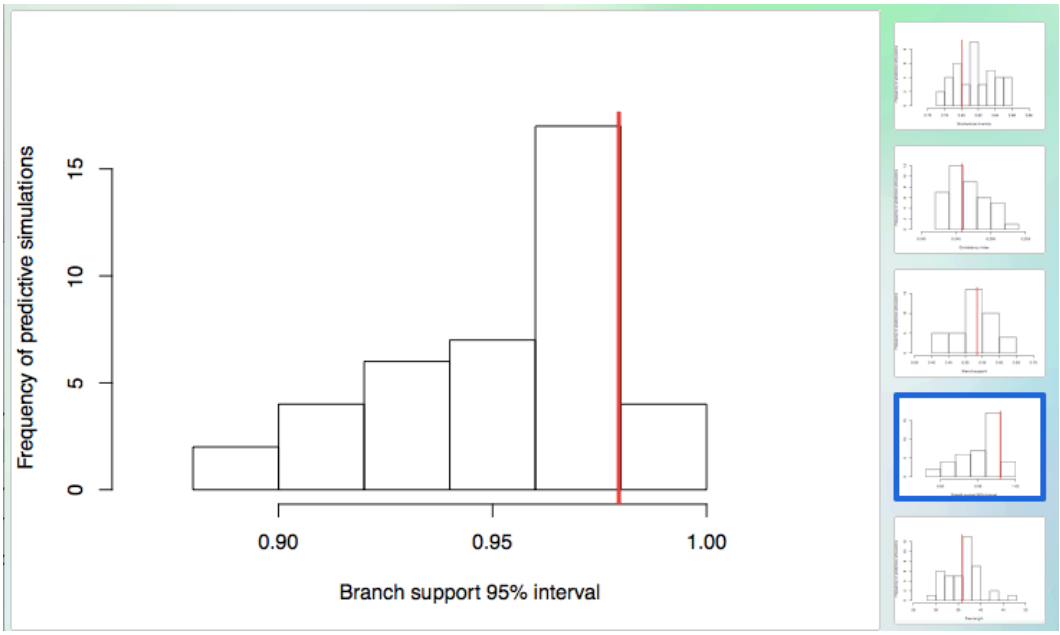

Figure 3. Example of histograms for qualitative assessment of the difference between test statistics calculated for predictive data and that calculated for the empirical data set. The histogram shows the distribution of test statistics for the simulated data sets, which can be compared with the red line, representing the value of the statistic calculated for the empirical data set.

If the χ^2 statistic has been selected, PhyloMAd will also output a text file with the interpretation of the result according to a comprehensive simulation study of the statistic (Duchêne et al. 2017).

The option of producing only the summary file overrides the other output options, and leads to a single summary output file. This file contains a table in which rows contain the details of assessment of each locus. The columns include all of the table of metrics of adequacy (tail area probabilities, the test statistic calculated for the empirical data set, and the number of standard deviations from the predictive distribution). In addition, the final column contains the interpretation of the χ^2 statistic test in terms of the risk of bias caused by compositional heterogeneity across taxa.

Other options will save the phylogenetic tree estimated from the empirical data, the predictive data alignments, and the tree estimated from each predictive data set (Fig. 3).

g. *Other options, starting, and monitoring the analysis.* Final settings can be given in the **Other options and START** tab. These settings include the number of predictive data sets to be used for assessment. At least 100 data sets should be used for assessment to be robust, but a smaller number can be used if the user merely wishes to test the program. It is also highly advisable to use multiple cores for analysis if allowed by the machine, in order to make the analysis more efficient.

To proceed with the analysis, click on the button to START ASSESSMENT and follow the next prompt. After a few minutes, data should start to be stored in the output folder, and the analyses should be logged in the PhyloMAd terminal window (Mac) or in the log file in the main PhyloMAd folder (Windows).

h. *Interpreting the results.* The aim of assessing model adequacy is to determine whether the empirical data are similar to data generated under the model, such that the model provides a reasonable description of the evolutionary process. The histograms produced by PhyloMAd provide an indication of the distance between predictive and empirical data. For quantitative metrics of the distance between the predictive data sets and the empirical data, it is perhaps most informative to report the number of standard deviations between the predictive distributions and the empirical data (Duchêne et al. 2017).

Our understanding of the performance of most test statistics remains poor, so a large distance between the predictive distribution and the empirical data might not necessarily indicate a high likelihood of bias. However, PhyloMAd provides an interpretation of the χ^2 statistic test. Other meaningful thresholds for interpreting tests will be made available in the near future. One existing approach is to consider the model inadequate if the empirical test statistic falls outside the central 95% of the predictive distribution of statistics.

6. Assessing Clock Models

This section will take you through the options for assessing Bayesian clock models.

a. *Data selection.* In order to assess the clock model, PhyloMAd requires the output of a BEAST 2 analysis (the log and tree files), in addition to the empirical data alignment (Fig. 4). These can be provided by pressing the *Browse*... buttons in the **Data tab** and navigating the folders to select the data. PhyloMAd will extract information from these files in order to generate predictive data sets and assess the clock model. Only a single data file can be selected for each section.

In this tab, it is also possible to select a percentage of the posterior samples to be discarded as burn-in.

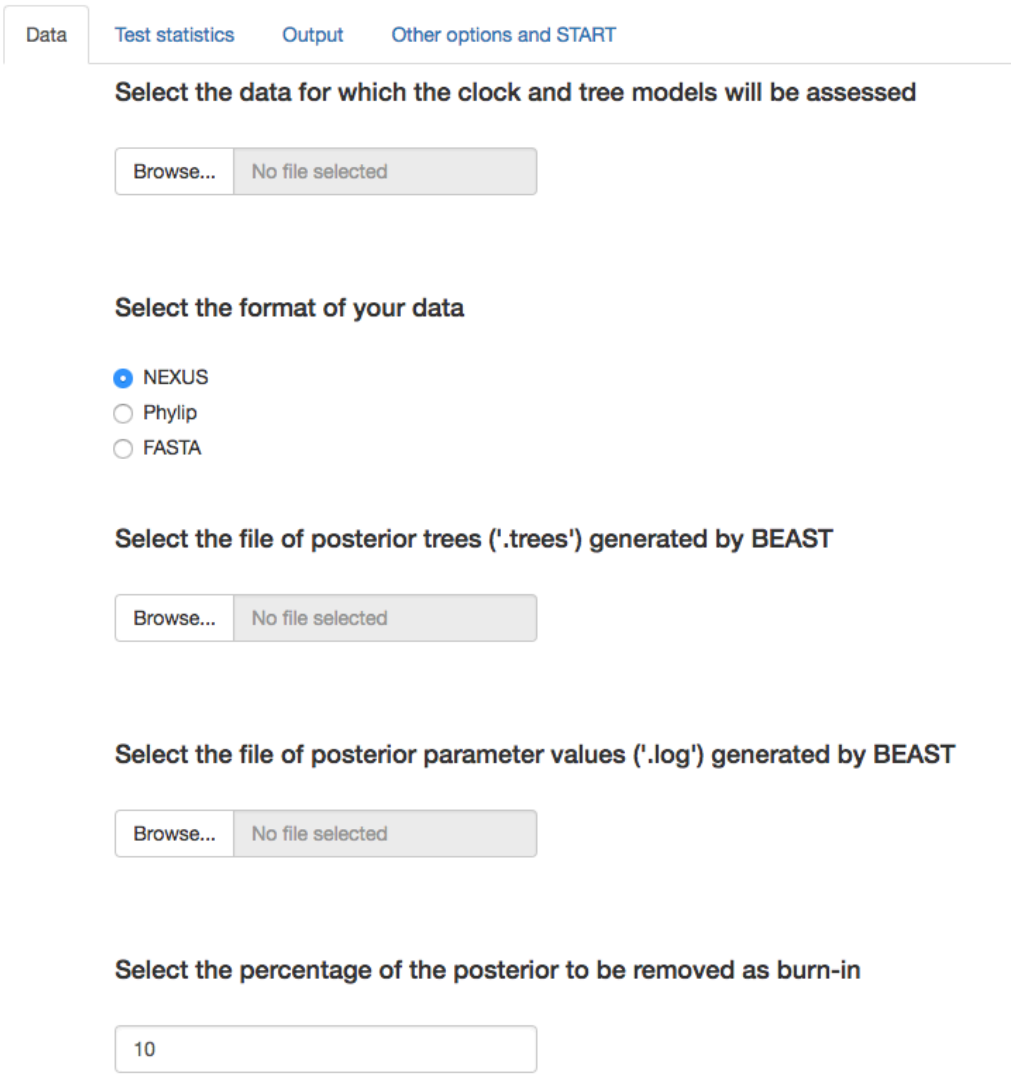

Figure 4. Screenshot of the Data tab when a Bayesian clock model is to be assessed. Clock models describe the pattern of rate variation across branches.

b. *File formats.* The data alignment can be provided in one of three formats (NEXUS, Phylip, and FASTA) using the buttons in the **Data tab**.

c. *Test statistics.* The test statistics that will be calculated for model assessment can be selected from the **Test statistics tab**. Details about each of the available statistics can be found in section 7 of this manual.

d. *Output.* Most options in the **Output tab** behave in the same way when assessing substitution models and clock models. See section 5f for detailed information about the output options that overlap across types of models being assessed.

One output that is unique to assessment of clock models will be produced when the A index is selected in the **Test statistics tab**. The A index produces a predictive distribution for each branch in the tree, since each branch length has been estimated from each of the predictive data sets. One of the additional files includes plots of the tree, where branches are labelled with their corresponding *P*-value and SDPD (standard deviations from the predictive distribution). Note that the trees plotted are cladograms, such that the lengths of branches have no meaning. The other additional file contains the same information but in the form of a table where the rows are the two assessment metrics (*P*-value and SDPD) and columns are each of the branches. These data, together with the empirical tree, can be used to make further plots or analyses, and are most easily usable when loaded into R using the ape package (Popescu et al. 2012).

e. *Other options, starting, and monitoring the analysis.* Final settings can be given in the **Other options and START** tab, and are simpler when assessing clock models than when assessing substitution models. The settings include the number of predictive data sets to be used for assessment and the number of computer cores that will be used to run the analysis.

f. *Interpreting the results.* Interpreting the results is done in a similar way to that when assessing the substitution model. The only exception is the output of assessment using the *A* index, which includes results for each branch. There are two ways to interpret the branch-wise results. One is by assuming that the model is inadequately describing the evolution along branches with length that falls outside the central 95% of the lengths of posterior predictive branches ($0.05 >$ branch *P*-value > 0.95). The other way to interpret this is by using the branch-wise SDPD (standard deviations from the predictive distribution), and reject branches with SDPD values that are above a certain threshold. A threshold of 2 will be similar to that recommended for interpreting the *P*-value. Thresholds that are more strict (e.g., 1) or more lenient (e.g., 10) might be reasonable, but this will depend on the study system and they must be reported.

7. Test Statistics

Test statistics for assessing substitution models

χ^2 *statistic.*—This statistic can be calculated for a table (following Sokal and Rohlf 1981) that contains the frequencies of each of the bases, for each of the taxa in the alignment. The statistic can be used to assess whether the analysis is in risk of joining taxa due to convergence in base composition. This test is different from a χ^2 homogeneity test of the empirical data in that it does not use the χ^2 distribution to assess significance.

Multinomial likelihood.—This statistic is the likelihood of the data under a model that only describes the most general of the assumptions in the majority of substitution models: that substitution events are independent and identically distributed (i.i.d.; Reeves 1992; Goldman 1993). A model will be rejected when using this statistic if it predicts frequencies of unique site patterns that are different from those based on the process that generated the data. This statistic is intended to provide an overall assessment of model fit. Importantly, only sites with compete data are included when estimating the multinomial likelihood (i.e. sites with gaps or indels are removed).

δ *statistic*.— This statistic is the difference between the multinomial likelihood and the maximum likelihood under the model being assessed, such as those of the GTR family. The δ statistic provides insight into the "cost" of using parameters that describe the evolutionary process (Goldman 1993). As with the multinomial likelihood, this statistic is intended to provide an overall assessment of model fit.

Biochemical diversity.—This statistic is the mean number of distinct bases that occur at each site in the alignment (Lartillot et al. 2007). The motivation for this statistic is that it should reject the model if it does not account for the base composition across sites in the data. In particular for amino acid data sets, this statistic can identify cases with excessive saturation of substitutions, or if the model is unable to describe evolutionary processes across sites, such as variable rates of substitution (Lartillot et al. 2007).

Consistency index.— The consistency index provides a measure of homoplasy in the alignment given an estimate of the tree topology (Kluge and Farris 1969). It is calculated as the minimum possible number of substitutions in the data, as a proportion of the minimum number of substitutions required to describe the given tree topology. The motivation behind quantifying homoplasy is that it can assess whether the model can recover the number of substitution events and amount of phylogenetic information in the data (Lartillot et al. 2007).

Mean branch support.—The mean node support across the inferred tree. It was initially proposed as a test statistic for assessing substitution models in a Bayesian framework (Brown 2014). This statistic has been suggested to assess whether the amount of statistical support for a given tree is plausible in data generated by the model and inferred parameters. PhyloMAd

calculates the mean branch support using approximate likelihood-ratio tests in PhyML (Anisimova and Gascuel 2006).

Range in branch support.—This statistic is based on the range of node support values across the tree (Brown 2014), and can assess whether phylogenetic information is uniformly distributed across the tree. The model can be rejected if some taxa in the alignment have amounts of phylogenetic information that are implausible under the model.

Total tree length.—The sum of estimated branch lengths, or total tree length (Brown 2014). This test statistic was proposed for assessing whether the total amount of molecular evolution inferred using the empirical data is plausible under the model.

Mahalanobis distance – This statistic summarizes multiple test statistics (Mahalanobis 1936; Drummond and Suchard 2008). It is the distance between the empirical test statistics and the multivariate predictive distribution of the selected statistics. This statistic provides a summary of the tests from a group of other test statistics. For this reason, it should only be selected if more than one other statistic has also been selected.

Test statistics for assessing Bayesian clock models

Stemminess.— The proportion of the inferred tree length represented by internal branches, such that stemminess is low when terminal branches are relatively long (Fiala and Sokal 1985). This test statistic assesses whether the priors on rates or times place excessive constraints on the relationship between terminal and internal branches, such that the number of substitutions cannot be estimated accurately.

D_F.— Normalized difference between summed terminal branch lengths and summed branch lengths (tree length) in a given tree estimate. This statistic is traditionally used for detecting neutrality by comparing the total number of singleton polymorphisms in a data set with the total number of segregating sites (Fu and Li 1993). The statistic calculated in PhyloMAd is the genealogical D_F (Drummond and Suchard 2008). This test statistic is similar to stemminess, and provides an indication of whether the clock and tree priors lead to implausible ratios of summed terminal branches to tree length.

Tree length.— The summed branch lengths for a given tree estimate. This statistic can assess whether the priors on rates or times are favouring implausibly large or small amounts of molecular evolution (e.g., an exponential prior on the rates when the originating process was one in which rates have a similar and high value).

Phylogenetic imbalance.— This is also known as "comblikeness" and is the extent to which only a single root-to-tip path has undergone diversification events. In an imbalanced tree, diversification has occurred asymmetrically, such that the tree looks like a comb. The opposite is a balanced tree, and for every node in the tree the two descendent branches will

lead to the same number of tips. In PhyloMAd, imbalance is calculated using the Colless index (Colless 1982). This statistic can assess whether the priors on node times place excessive weight on particular tree shapes that are distinct from the signal in the data.

A index.— To compute this index, branch lengths are estimated from the posterior predictive data alignments under a clock-free method (estimated using PhyML), such that there is a distribution of length estimates for each branch. A posterior predictive *P*-value is then calculated for each branch using the corresponding distribution obtained with the posterior predictive alignments. The *A* index for overall assessment is the proportion of branches in the phylogram from the empirical data that have lengths falling outside the 95% quantile range of those estimated from the posterior predictive data sets.

A low *A* index indicates that a large proportion of branch rates and/or times are inconsistent with the expected number of substitutions along the branches. Under ideal conditions, an *A* index of 0.95 or higher means that the clock model accurately describes the true pattern of rate variation.

8. Literature cited

Anisimova M, Gascuel O. 2006. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst. Biol. 55:539–552.

Brown JM. 2014. Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. Syst. Biol. 63:334–348.

Colless DH. 1982. Review of phylogenetics: the theory and practice of phylogenetic systematics. Syst. Zool. 31:100–104.

Drummond AJ, Suchard MA. 2008. Fully Bayesian tests of neutrality using genealogical summary statistics. BMC Genet. 9:68.

Duchêne DA, Duchêne S, Ho SYW. 2017. New statistical criteria detect phylogenetic bias caused by compositional heterogeneity. Mol. Biol. Evol. 34:1529–1534.

Duchêne DA, Duchêne S, Holmes EC, Ho SYW. 2015. Evaluating the adequacy of molecular clock models using posterior predictive simulations. Mol. Biol. Evol. 32:2986–2995.

Fiala KL, Sokal RR. 1985. Factors determining the accuracy of cladogram estimation: evaluation using computer simulation. Evolution (N. Y). 39:609.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. Genetics 133:693–709.

Goldman N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36:182–198.

Kluge AG, Farris JS. 1969. Quantitative phyletics and the evolution of anurans. Syst. Biol. 18:1–32.

Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol. Biol. 7:4.

Mahalanobis PC. 1936. On the generalised distance in statistics. Proc. Natl. Inst. Sci. India 12:49–55.

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

Popescu A-A, Huber KT, Paradis E. 2012. ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. Bioinformatics 28:1536–1537.

Reeves JH. 1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. J. Mol. Evol. 35:17–31.

Reid NM, Hird SM, Brown JM, Pelletier TA, McVay JD, Satler JD, Carstens BC. 2014. Poor fit to the multispecies coalescent is widely detectable in empirical data. Syst. Biol. 63:322–333.

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

Sokal RR, Rohlf FJ. 1981. Biometry. San Francisco: W. H.Freeman