Phylodynamic model adequacy using posterior predictive simulations

Sebastian Duchene[1]*, Remco Bouckaert[2,3], David A. Duchene[4], Tanja Stadler[5,6], Alexei J. Drummond[2]

[1]Department of Biochemistry and Molecular Biology, Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Australia.

[2]Centre for Computational Evolution, University of Auckland, Auckland, New Zealand.

[3]Max Planck Institute for the Science of Human History, Jena, Germany.

[4]School of Life and Environmental Sciences, University of Sydney, Sydney, Australia.

[5]Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland.

[6]Swiss Institute of Bioinformatics, Lausanne, Switzerland.

* To whom correspondence should be addressed:

Sebastian Duchene

Email: sebastian.duchene@unimelb.edu.au

Abstract

Rapidly evolving pathogens, such as viruses and bacteria, accumulate genetic change at a similar timescale over which their epidemiological processes occur, such that it is possible to make inferences about their infectious spread using phylogenetic time-trees. To do this, it is necessary to choose a phylodynamic model. However, the resulting inferences are contingent on whether the model is an adequate description of the data. Model adequacy methods allow formal rejection of a model if it cannot generate the main features of the data. We present TreeModelAdequacy (TMA), a package for the popular BEAST2 software, that allows assessing the adequacy of phylodynamic models. We illustrate its utility using simulations and an Ebola virus data set from the start of the 2013-2016 West African outbreak. The Ebola data were adequately described by the coalescent exponential-growth, the birth-death, and the birth-death susceptible-infected-recovered models, consistent with the rapid infectious spread of the virus.

Keywords

Model adequacy, Bayesian phylogenetics, Phylodynamics, Viral evolution, Posterior predictive simulation, BEAST2

Phylogenetic trees depict the evolutionary relationships between groups of organisms. In the context of infectious diseases, pathogen genetic data can be used to infer such trees. By

making some assumption about the substitution process and including independent information about time, one can calibrate the molecular clock to obtain time-trees, where the branch lengths correspond to units of time, and internal nodes of the tree represent the timing of divergence events. For rapidly evolving viruses and bacteria it is possible to use the sampling times as time-calibrations. In these organisms, genetic change and epidemiological, or ecological, processes occur over a similar timescale. Thus, time-trees can be informative about epidemiological dynamics, a field of research known as phylodynamics (Grenfell et al., 2004; Holmes, Zhang, Simmonds, Rogers, & Brown, 1993; Kühnert, Wu, & Drummond, 2011; E. M. Volz, Koelle, & Bedford, 2013).

Phylodynamic models describe the distribution of node times, branch lengths, and sampling times. In full Bayesian phylogenetic analyses, the tree, parameters for the molecular clock, the substitution model and the phylodynamic model can be estimated simultaneously using molecular data. In this Bayesian framework, the phylodynamic model is effectively a 'tree prior' (e.g. du Plessis and Stadler 2015). The simplest phylodynamic models are the coalescent exponential-growth (CE) and the coalescent constant-size (CC). These two models have very different expectations about the shape of phylogenetic trees, with exponentially growing populations tending to produce trees with longer external branches than those evolving under constant population sizes (O'Meara, 2012; Volz et al., 2013). Alternative phylodynamic models are the birth-death models, which include a parameter to describe the sampling rate, so they have an expectation on the number of taxa and their distribution over time (Stadler, 2010; Stadler et al., 2012).

Choosing an appropriate model is important to draw reliable inferences from parameters of interest. For instance, the CE and the constant birth-death (BD) models can estimate the basic reproductive number, R_0 (Frost & Volz, 2010; Stadler et al., 2012; Volz et al., 2013),

which is the average number of secondary infections in a fully susceptible population. Failing to account for complex epidemiological dynamics can bias the estimate of this key parameter (Ratmann et al., 2016; Stadler, Kühnert, Rasmussen, & du Plessis, 2014). A Bayesian approach to selecting a phylodynamic model is to estimate marginal likelihoods for a pool of models and selecting that with the highest marginal likelihood (Baele et al. 2012; Baele et al. 2016), but it is also possible to obtain weighted averages of parameter estimates based on the support for each model (e.g. Baele, Li, Drummond, Suchard, & Lemey, 2013; R. R. Bouckaert & Drummond, 2017; Huelsenbeck, Larget, & Alfaro, 2004; Li & Drummond, 2012).

Model adequacy in phylogenetics

Model selection methods only allow a relative comparison of a set of models, but they cannot determine whether any of the models in question could have generated the data at hand (i.e. absolute model fit). Such information however is key to avoid unreliable inferences from a model and to improve our understanding of the biological processes that produced the data. Absolute model fit can be assessed via model adequacy methods, where a model is considered 'adequate' if it is capable of generating the main features of the empirical data. Consequently, model adequacy allows the user to reject a model, instead of ranking it with respect to other models, as is the case with model testing (e.g. Goldman 1993; Bollback 2002; Ripplinger and Sullivan 2010; Brown 2014).

Model adequacy is typically conducted by fitting a model to the empirical data, and generating synthetic data from the model in question, a procedure that is similar to a parametric bootstrap (Goldman, 1993). The adequacy of the model is determined depending on whether the synthetic data are similar to the empirical data, according to a descriptive test statistic. This requires obtaining test statistics that summarise key aspects

of the data. When the data consist of phylogenetic trees, some intuitive test statistics can include the ratio of external to internal branch lengths, the tree height, and measures of phylogenetic tree imbalance (Drummond & Suchard, 2008). Clearly, designing test statistics is not trivial, but they should attempt to explicitly test some of the assumptions of the model. For example, the CE and CC models have different expectations of the ratio of external to internal branch lengths, such that this may be a useful test statistic.

Bayesian model adequacy

The most theoretically justifiable way to conduct Bayesian model adequacy assessment consists of a posterior predictive framework (Bollback, 2002; Brown, 2014a, 2014b; Höhna, Coghill, Mount, Thomson, & Brown, 2017; Lewis, Xie, Chen, Fan, & Kuo, 2014). This consists in approximating the posterior distribution of the model in question given the empirical data, for example using Markov chain Monte Carlo (MCMC). Samples from the MCMC are drawn to simulate data sets under the model used for the empirical analysis. For example, the posterior distribution of the growth rate and population size parameters of the CE model can be sampled to simulate phylogenetic trees. Such simulations are known as posterior predictive simulations. Test statistics are then calculated for every posterior predictive simulation (i.e. for every simulated tree) to generate a distribution of values according to the model. A posterior predictive probability, similar to the frequentist *p*-value, can be calculated by determining where the value of the test statistic for the empirical data (i.e. the empirical phylogenetic tree) falls with respect to the posterior predictive distribution. We refer to the posterior predictive probability as p_{pp} to differentiate it from the frequentist *p*-value. A useful guideline to determine whether the model is adequate is to determine whether p_{pp}, falls between 0.025 and 0.975, such that it is within the 95% credible interval (Bollback, 2002; Brown, 2014a; Höhna et al., 2017). Combining multiple test

statistics leads to multiple testing, but this problem can be addressed by using a multivariate p_{pp} values (Drummond & Suchard, 2008).

"TreeModelAdequacy" package in BEAST2

We implemented a computational framework to assess the adequacy of phylodynamic models as a package for BEAST2 (Bouckaert et al., 2014). Analyses as outlined in Fig 1 are easy to set up through BEAUti, the graphical user interface for BEAST. Our package, TreeModelAdequacy (TMA), takes a tree with branch lengths proportional to time as the empirical data. The tree can be a summary tree from BEAST2, or estimated using a different method. Although our implementation assumes a single tree topology and divergence times, accounting for phylogenetic uncertainty can be done by analysing a set of credible trees.

We fit phylodynamic models available in BEAST2 by approximating the posterior distribution of the parameters of the model using MCMC. To generate the posterior predictive simulations, we draw samples from the posterior for the parameters from the MCMC after removing the burn-in phase, and we simulate phylogenetic trees using MASTER (Vaughan & Drummond, 2013). The last step consists in calculating test statistics for the empirical data and for the posterior predictive simulations, which depends on the TreeStat2 package (available at http://github.com/alexeid/TreeStat2). The user can select a large number of test statistics. At the end of the analysis, p_{pp} values and quantiles for the posterior predictive distribution are shown, but they can also be visualised in Tracer (available at: http://beast.bio.ed.ac.uk/tracer) or using an R script included in the package. At present, the range of phylodynamic models that can be assessed includes the CC and CE coalescent models, the constant BD with serial sampling (Stadler et al. 2012), and the birth-death susceptible-infected-recovered model (BDISR; Kühnert et al. 2014).

Our implementation allows parallelisation of the tree simulation step, which can increase computational speed when a large number of posterior predictive simulations are required. This step can also be conducted independently on a computer cluster. In this case, our standalone application TreeModelAdequacyAnalyser can compute test statistics and p_{pp} values, even for trees generated in a different program than BEAST2. TMA is open source and freely available under a LGPL licence. It can be downloaded from BEAUTi2 (part of BEAST2), and the documentation and example files are available at: http://github.com/sebastianduchene/tree_model_adequacy.

Validation using simulations

To verify our implementation we conducted a simulation experiment. We simulated 100 trees under each of the four phylodynamic models (CC, CE, BD, and BDSIR) using MASTER and assessed their adequacy according to seven test statistics (Table S1 and Table 1). The parameters for our simulations were based on analyses of 72 whole genomes sequences of Ebola virus (Gire et al., 2014). Our expectation is that p_{pp} values should be around 0.50 when the model assessed matches that used to generate the data. In contrast, when the model assessed does not match that used to generate the data, some of the p_{pp} values should correspond to the tails of the posterior predictive distribution. For example, when we assessed the CC model for our trees simulated under the same model, the mean p_{pp} value over 100 simulations for the tree height was 0.46. The remaining test statistics had p_{pp} values between 0.41 and 0.56. In contrast, when we assessed the CC model for trees generated under the BD and CE models, the p_{pp} value for the tree height was 0.00, indicating that all posterior predictive simulations generated trees with older tree heights (Table S1). Interestingly, some test statistics, including the tree height had mean p_{pp} values between 0.025 and 0.975 when the CE model was assessed for trees generated under the

183 CC model, suggesting that the CE model is an adequate description of trees generated
184 under the CC. This occurs because the CC model is equivalent to a CE when the exponential
185 growth rate is zero. Importantly, the Colless test statistic that measures phylogenetic tree
186 imbalance had mean p_{pp} values that ranged from 0.10 to 0.75 across simulations, indicating
187 that the four models generate similar degrees of phylogenetic tree imbalance (Drummond
188 & Suchard, 2008).
189
190 **Phylodynamic model adequacy in a West African Ebola virus epidemic**
191 We obtained a phylogenetic tree inferred in a previous study from 72 Ebola virus whole
192 genome samples collected during the 2013-2016 epidemic (Gire et al., 2014). The samples
193 were collected from May to July 2014 in Sierra Leone. These data have been used in
194 previous studies to estimate epidemiological parameters, with estimates of R_o ranging from
195 1.5 to 2.5, depending on the phylodynamic model (Stadler et al., 2014; Volz & Pond, 2014).
196 An important consideration about our analysis is that we assume that the tree topology and
197 divergence time estimates are sufficiently accurate.
198
199 We inferred phylodynamic parameters for the Ebola virus tree using four models; CC, CE,
200 the BD, and the BDSIR (Kühnert et al., 2014). The CE, BD, and BDSIR models can estimate
201 R_o if information about the sampling process (for the birth-death models) or present
202 number of infected individuals (for the coalescent) is available. In this case, we assumed
203 that the sampling proportion was 0.7 (Gire et al., 2014) by fixing this parameter in the BD
204 and BDSIR models. For the remaining parameters of these two models, we used the same
205 prior distributions as in a previous analysis of these data (Stadler et al., 2014). For the CE
206 model we used a Laplace distribution and a OneOnX distribution as priors for the growth
207 rate and for the effective population size, respectively. We ran an MCMC of 10^7 steps and
208 generated 1,000 posterior predictive simulations, and we computed nine test statistics

(Table 1). Importantly, two of these test statistics, the number of tips and mean sampling times, can only be computed for the BD model, which is not conditioned on the sampling times and number of tips (Boskova, Bonhoeffer, & Stadler, 2014; Stadler, 2010).

To compare the different models, we considered each test statistic individually and calculated a multinomial p_{pp} value using the Mahalanobis distance (Table 1). The latter method suggested that the CE, BD, and BDSIR models may be adequate descriptions of the Ebola virus tree, with p_{pp} values of 0.467, 0.47 and 0.88, respectively. Two of seven test statistics, the tree height and the ratio of external to internal branch lengths for the Ebola virus tree were not plausible under the CC model, with p_{pp} values of 0.035 and 0.99, respectively. Moreover, its posterior predictive distribution had little overlap with those from the other models (Fig 2). This finding is expected because these data were collected at an early phase of the epidemic, where the number of infections should grow rapidly. In contrast, the CE model appears to describe this data set more closely, with p_{pp} values that ranged between 0.89 and 0.26 for all test statistics.

The BD model performed better than the CC model, but one test statistic, the slope ratio of the lineages-through-time plot, had a very low p_{pp} value for the Ebola virus tree, although still within the 95% credible interval, at 0.04. Interestingly, this test statistic appears informative for inferring epidemiological parameters using Approximate Bayesian Computation (Saulnier, Alizon, & Gascuel, 2017). The BD model sometimes produced very wide posterior predictive distributions (Fig 2), which is probably because the trees generated by this model could have different numbers of tips. The number of tips and mean age of tips appeared to be adequately described by the BD model, with p_{pp} values of 0.50 and 0.81, respectively, such that they might be informative about the performance of the model in describing the sampling process.

The BDSIR model was also adequate, with a multinomial p_{pp} value of 0.88 and only three test statistics with p_{pp} values of < 0.025 or > 0.975 (Table 1). The parameters of this model include birth, death, and sampling rates, and the number of susceptible individuals, which is assumed to be infinite in the BD model. As such, the BDSIR and BD models are similar when the number of susceptible individuals in the BDSIR is very large. Overall, these results suggest that only models that allow exponential growth (CE, BD, and BDSIR) could predict some of the main features of the Ebola virus data set, which is consistent with the spread of the virus when the sequences were collected.

For the CE, BD, and BDISIR models we can estimate R_o and the infectious period, $1/\delta$. The R_o median estimates were: 1.6 (95% credible interval (CI): 1.12 – 2.2) for the BD, 1.21 (95% CI: 1.1 – 1.5), for the CE, and 1.59 (95% CI: 1.22 – 1.94) for the BDSIR. Estimates for the infectious period in calendar days were: 5.46 (95% CI: 4.14 – 7.24) for the BD, 2.90 (95% CI: 1.75 – 5.50) for the CE, and 5.21 (95% CI: 4.24 – 7.02) for the BDSIR. The estimates from these models, which appeared to be adequate, were very similar and overlapped with those from previous studies (Stadler et al., 2014).

Conclusion

Model adequacy methods are useful to understand the biological processes that generate the data, such as the evolutionary branching process. For example, our approach reveals that in July of 2014 the West African Ebola outbreak was still growing exponentially. Future research should assess more sophisticated phylodynamic models, such as those that account for population structure (Kühnert, Stadler, Vaughan, & Drummond, 2016), and their impacts on estimating parameters of interest. Our simulations corroborate previous findings that the effectiveness of model adequacy depends on the test statistics used and

the information content of the data (i.e. the accuracy of the phylogenetic tree). For this reason, studies proposing novel test statistics will benefit from investigation using simulation studies. Model adequacy software, such as TMA, will be key to address these questions.

Acknowledgements

The authors thank Timothy Vaughan for valuable discussions. SD was supported by a McKenzie fellowship and a Dyason grant from the University of Melbourne. AJD was supported by a Rutherford fellowship (http://www.royalsociety.org.nz/programmes/funds/rutherford-discovery/) from the Royal Society of New Zealand. TS was supported in part by the European Research Council under the Seventh Framework Programme of the European Commission (PhyPD: grant agreement number 335529).

Conflict of interest

The authors gave final approval to the manuscript and have no conflict of interest to declare.

Author contributions

SD, RB, and AJD wrote the computer code. SD and DD analysed the data. TS, SD, and AJD designed the experiments. SD wrote the manuscript with input from all the authors.

Figure Legends

Fig 1. Posterior predictive simulation framework implemented in the TreeModelAdequacy package. The first step consists in Bayesian analysis in BEAST under model M to estimate the posterior distribution of parameters η, shown with the arrow and posterior density in orange. Samples from the posterior are drawn to simulate phylogenetic trees, known as posterior predictive simulations, using MASTER as shown by the green arrow. The posterior predictive simulations are analysed in TreeStat to generate the posterior predictive distribution of test statistic T_a, shown by the blue arrow and probability density. Finally, T_a is also computed for the tree from the empirical data using TreeStat, shown by the red arrow, to calculate a p_{pp} value.

Fig 2. Pairwise plots for the posterior predictive distributions of two pairs of test statistics: Colless tree imbalance and the ratio of external to internal branch lengths (A), and tree height and the slope ratio of the lineages-through-time plot (B). Each point corresponds to a posterior predictive simulated tree, such that the clouds of points represent the expectation of test statistics under each model; birth-death susceptible-infected-recovered (BDSIR; in red), birth-death (BD; in orange), coalescent exponential-growth (CE; in blue), and coalescent constant-size (CC; in green). The X symbol corresponds to the value observed for the Ebola virus tree.

Table 1. Posterior predictive probabilities (p_{pp}) for nine test statistics, 95% percentile ranges for the posterior predictive simulations (PR-PPS) and estimated values for the Ebola virus tree. Values marked with an asterisk (*) are those for which $0.025 < p_{pp} > 0.975$, indicating that the model did not adequately describe this feature.

		Coalescent Constant-size		Coalescent Exponential-growth		Birth-Death		Birth-Death SIR	
Test statistic	**Observed value for Ebola virus tree**	p_{pp}	**95%PR-PPS**	p_{pp}	**95% PR-PPS**	p_{pp}	**95% PR-PPS**	p_{pp}	**95% PR-PPS**
Longest branch length	0.047	0.06	0.04 – 0.35	0.78	0.024 – 0.075	0.90	0.021- 0.053	0.98*	0.017 – 0.044
Colless index	0.23	0.97	0.11 – 0.24	0.94	0.11 – 0.25	0.52	0.10 – 0.80	0.67	0.14 – 0.34
Slope ratio of LTT plot[1]	0.087	0.96	0.019 – 0.091	0.26	0.064 – 0.16	0.04	0.083 – 3.12	0.02*	0.15 – 0.57
Time to maximum number of lineages[1]	1.19×10^{-4}	0.83	1.1×10^{-4}- 1.2×10^{-4}	0.59	1.14×10^{-4} – 1.59×10^{-4}	0.14	8.11×10^{-5} – 3.58×10^{-3}	0.48	1.1×10^{-4}– 1.6×10^{-4}
Tree height	0.11	0.035*	0.10 – 0.45	0.72	0.075 – 0.14	0.61	0.077 – 0.13	0.99*	0.07 – 0.10
Tree length	0.93	0.050	0.90 – 1.85	0.50	0.77 – 1.1	0.49	0.19 – 1.91	0.1	0.56 – 0.97
Ratio of external to internal branch lengths	1.19	0.99*	0.34 – 1.17	0.89	0.66 – 1.38	0.54	0.75 – 1.64	0.42	0.87 – 1.73
Number of tips	72	--	--	--	--	0.50	12 – 146	--	--
Mean age of tips	0.80	--	--	--	--	0.81	2.0×10^{-4}- 9.71×10^{-1}	--	--
Multinomial P value	--	0.98*	--	0.67	--	0.47	--	0.88	--

[1] The time to the maximum number of lineages is obtained by producing a lineages-through-time (LTT) plot and calculating the time when the maximum number of lineages is reached. The slope ratio of the LTT plot is calculated by fitting two regression lines along the this plot, from the origin to the maximum number of lineages, and from the maximum number of lineages to

the youngest tip of the tree, as proposed by Saulnier et al. (2017). This test statistic displays positive association with the size of the tree, such that we standardise it by the number of tips. The -- symbol corresponds to values that were not applicable.

314 **References**

315 Baele, G., Lemey, P., Bedford, T., Rambaut, A., Suchard, M. A., & Alekseyenko, A. V. (2012).
316 Improving the accuracy of demographic and molecular clock model comparison while
317 accommodating phylogenetic uncertainty. *Molecular Biology and Evolution*, *29*(9),
318 2157–2167.

319 Baele, G., Lemey, P., & Suchard, M. A. (2016). Genealogical working distributions for
320 Bayesian model testing with phylogenetic uncertainty. *Systematic Biology*, *65*(2), 250–
321 264.

322 Baele, G., Li, W. L. S., Drummond, A. J., Suchard, M. A., & Lemey, P. (2013). Accurate model
323 selection of relaxed molecular clocks in bayesian phylogenetics. *Molecular Biology and*
324 *Evolution*, *30*(2), 239–243.

325 Bollback, J. P. (2002). Bayesian model adequacy and choice in phylogenetics. *Molecular*
326 *Biology and Evolution*, *19*(7), 1171–1180.

327 Boskova, V., Bonhoeffer, S., & Stadler, T. (2014). Inference of epidemiological dynamics
328 based on simulated phylogenies using birth-death and coalescent models. *PLOS*
329 *Comput Biol*, *10*(11), e1003913.

330 Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., … Drummond, A. J.
331 (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLOS*
332 *Computational Biology*, *10*(4), e1003537.

333 Bouckaert, R. R., & Drummond, A. J. (2017). bModelTest: Bayesian phylogenetic site model
334 averaging and model comparison. *BMC Evolutionary Biology*, *17*(1), 42.

335 Brown, J. M. (2014a). Detection of implausible phylogenetic inferences using posterior
336 predictive assessment of model fit. *Systematic Biology*, *63*(3), 334–348.

337 Brown, J. M. (2014b). Predictive approaches to assessing the fit of evolutionary models.
338 *Systematic Biology*, *63*(3), 289–292.

339 Drummond, A. J., & Suchard, M. A. (2008). Fully Bayesian tests of neutrality using
340 genealogical summary statistics. *BMC Genetics*, *9*(1), 68.

341 du Plessis, L., & Stadler, T. (2015). Getting to the root of epidemic spread with
342 phylodynamic analysis of genomic data. *Trends in Microbiology*, *23*(7), 383–386.

343 Frost, S. D. W., & Volz, E. M. (2010). Viral phylodynamics and the search for an "effective
344 number of infections." *Philosophical Transactions of the Royal Society of London B:*
345 *Biological Sciences*, *365*(1548), 1879–1890.

346 Gire, S. K., Goba, A., Andersen, K. G., Sealfon, R. S. G., Park, D. J., Kanneh, L., … Dudas, G.
347 (2014). Genomic surveillance elucidates Ebola virus origin and transmission during the

2014 outbreak. *Science*, *345*(6202), 1369–1372.

Goldman, N. (1993). Simple diagnostic statistical tests of models for DNA substitution. *Journal of Molecular Evolution*, *37*(6), 650–661.

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L. N., Daly, J. M., Mumford, J. A., & Holmes, E. C. (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, *303*(5656), 327–332.

Höhna, S., Coghill, L. M., Mount, G. G., Thomson, R. C., & Brown, J. M. (2017). P3: Phylogenetic Posterior Prediction in RevBayes. *Molecular Biology and Evolution*. In press.

Holmes, E. C., Zhang, L. Q., Simmonds, P., Rogers, A. S., & Brown, A. J. L. (1993). Molecular investigation of human immunodeficiency virus (HIV) infection in a patient of an HIV-infected surgeon. *Journal of Infectious Diseases*, *167*(6), 1411–1414.

Huelsenbeck, J. P., Larget, B., & Alfaro, M. E. (2004). Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Molecular Biology and Evolution*, *21*(6), 1123–1133.

Kühnert, D., Stadler, T., Vaughan, T. G., & Drummond, A. J. (2014). Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth–death SIR model. *Journal of the Royal Society Interface*, *11*(94), 20131106.

Kühnert, D., Stadler, T., Vaughan, T. G., & Drummond, A. J. (2016). Phylodynamics with migration: A computational framework to quantify population structure from genomic data. *Molecular Biology and Evolution*, *33*(8), 2102–2116.

Kühnert, D., Wu, C. H., & Drummond, A. J. (2011). Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. *Infection, Genetics and Evolution*, *11*(8), 1825–1841.

Lewis, P. O., Xie, W., Chen, M.-H., Fan, Y., & Kuo, L. (2014). Posterior predictive Bayesian phylogenetic model selection. *Systematic Biology*, *63*(3), 309–321.

Li, W. L. S., & Drummond, A. J. (2012). Model averaging and Bayes factor calculation of relaxed molecular clocks in Bayesian phylogenetics. *Molecular Biology and Evolution*, *29*(2), 751–761.

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

Ratmann, O., Hodcroft, E. B., Pickles, M., Cori, A., Hall, M., Lycett, S., … Frost, S. (2016). Phylogenetic Tools for Generalized HIV-1 Epidemics: Findings from the PANGEA-HIV

Methods Comparison. *Molecular Biology and Evolution*, *34*(1), 185–203.

Ripplinger, J., & Sullivan, J. (2010). Assessment of substitution model adequacy using frequentist and Bayesian methods. *Molecular Biology and Evolution*, *27*(12), 2790–2803.

Saulnier, E., Alizon, S., & Gascuel, O. (2017). Assessing the accuracy of Approximate Bayesian Computation approaches to infer epidemiological parameters from phylogenies. *PLOS Computational Biology*, *13*(3), e1005416.

Stadler, T. (2010). Sampling-through-time in birth-death trees. *Journal of Theoretical Biology*, *167*(3), 696–404.

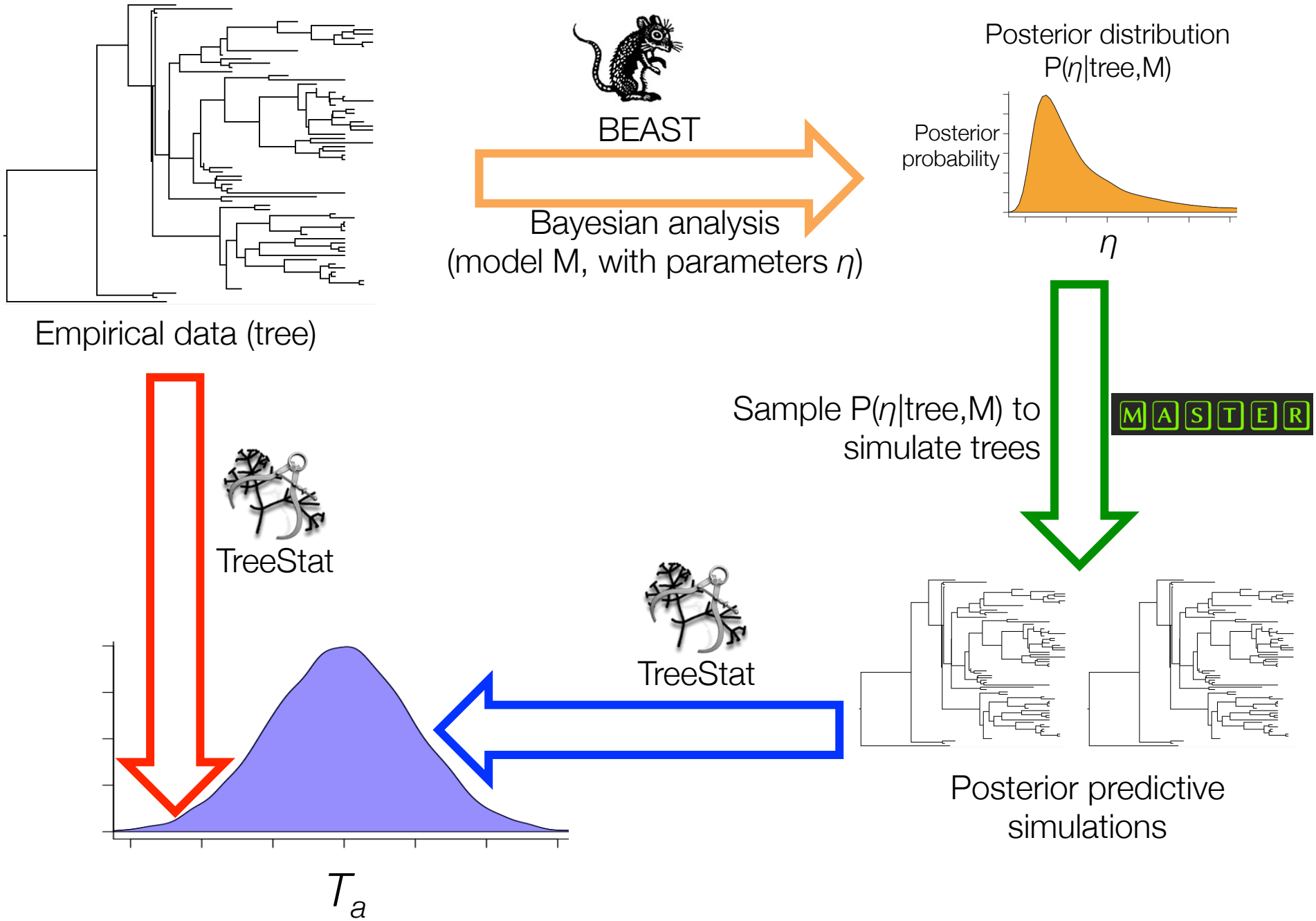
Stadler, T., Kouyos, R., von Wyl, V., Yerly, S., Böni, J., Bürgisser, P., … Xie, D. (2012). Estimating the basic reproductive number from viral sequence data. *Molecular Biology and Evolution*, *29*(1), 347–357.

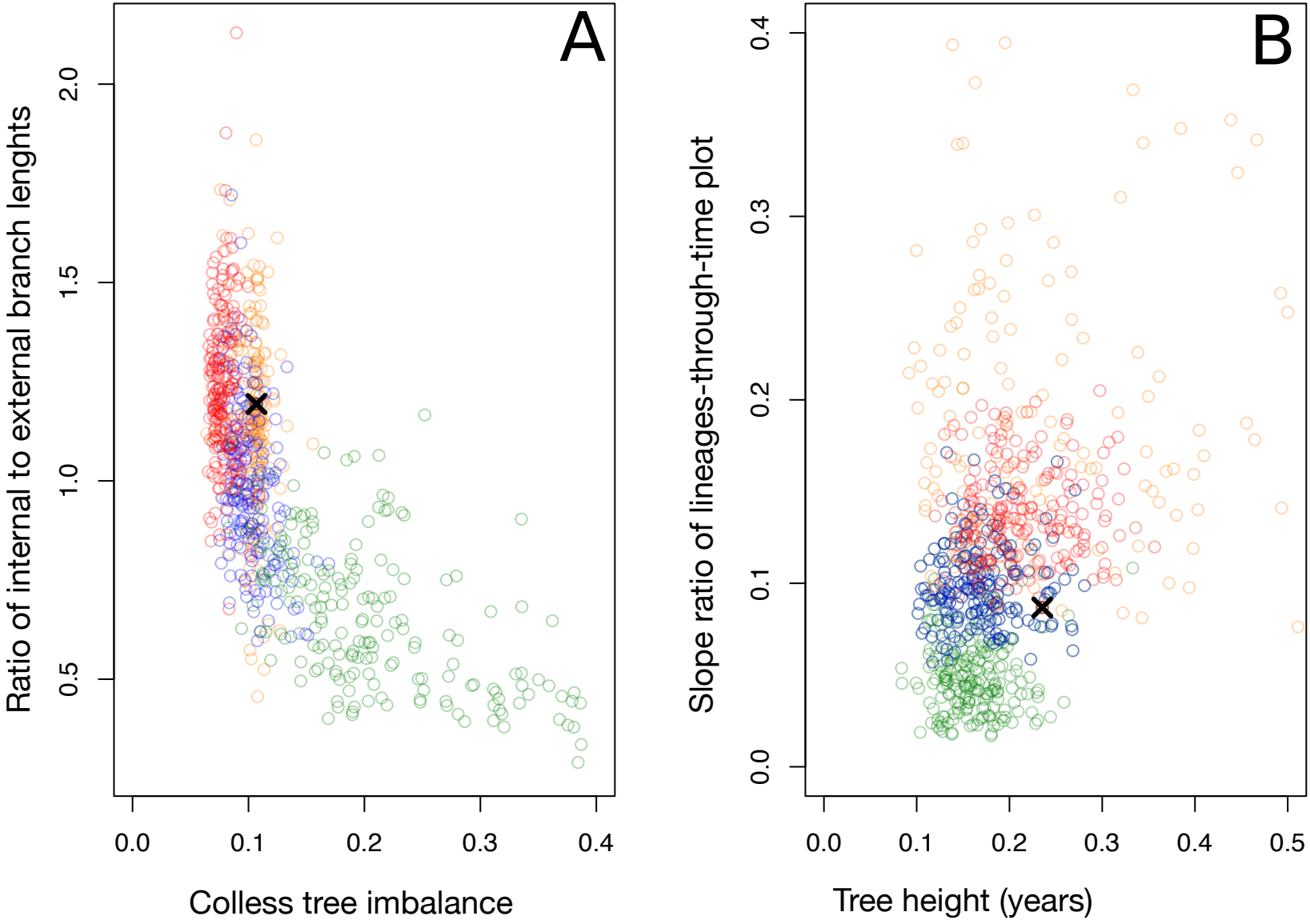
Stadler, T., Kühnert, D., Rasmussen, D. A., & du Plessis, L. (2014). Insights into the early epidemic spread of Ebola in Sierra Leone provided by viral sequence data. *PLoS Currents*, *6*.

Vaughan, T. G., & Drummond, A. J. (2013). A stochastic simulator of birth–death master equations with application to phylodynamics. *Molecular Biology and Evolution*, *30*(6), 1480–1493.

Volz, E. M., Koelle, K., & Bedford, T. (2013). Viral phylodynamics. *PLOS Computational Biology*, *9*(3), e1002947.

Volz, E., & Pond, S. (2014). Phylodynamic analysis of Ebola virus in the 2014 Sierra Leone epidemic. *PLoS Currents*, *6*.


Empirical data (tree)
BEAST
Bayesian analysis
(model M, with parameters η)
Posterior distribution
P(η|tree,M)
Posterior
probability
η
Sample P(η|tree,M) to
simulate trees
MASTER
TreeStat
TreeStat
Posterior predictive
simulations
T_a


A
Ratio of internal to external branch lenghts
0.5
1.0
1.5
2.0
0.0
0.1
0.2
0.3
0.4
Colless tree imbalance
B
Slope ratio of lineages-through-time plot
0.0
0.1
0.2
0.3
0.4
0.0
0.1
0.2
0.3
0.4
0.5
Tree height (years)