Evaluating the Adequacy of Molecular Clock Models Using Posterior Predictive Simulations

David A. Duchêne,*[,†,1] Sebastian Duchêne,[†,2,3] Edward C. Holmes,[2,3] and Simon Y.W. Ho[2]

[1]Research School of Biology, Australian National University, Canberra, ACT, Australia

[2]School of Biological Sciences, University of Sydney, Sydney, NSW, Australia

[3]Marie Bashir Institute for Infectious Diseases and Biosecurity, Charles Perkins Centre, Sydney Medical School, University of Sydney, Sydney, NSW, Australia

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: david.duchene@anu.edu.au.

Associate editor: Jeffrey Thorne

Abstract

Molecular clock models are commonly used to estimate evolutionary rates and timescales from nucleotide sequences. The goal of these models is to account for rate variation among lineages, such that they are assumed to be adequate descriptions of the processes that generated the data. A common approach for selecting a clock model for a data set of interest is to examine a set of candidates and to select the model that provides the best statistical fit. However, this can lead to unreliable estimates if all the candidate models are actually inadequate. For this reason, a method of evaluating absolute model performance is critical. We describe a method that uses posterior predictive simulations to assess the adequacy of clock models. We test the power of this approach using simulated data and find that the method is sensitive to bias in the estimates of branch lengths, which tends to occur when using underparameterized clock models. We also compare the performance of the multinomial test statistic, originally developed to assess the adequacy of substitution models, but find that it has low power in identifying the adequacy of clock models. We illustrate the performance of our method using empirical data sets from coronaviruses, simian immunodeficiency virus, killer whales, and marine turtles. Our results indicate that methods of investigating model adequacy, including the one proposed here, should be routinely used in combination with traditional model selection in evolutionary studies. This will reveal whether a broader range of clock models to be considered in phylogenetic analysis.

Key words: **model adequacy, posterior predictive simulations, Bayesian phylogenetics, molecular clock, evolutionary rates, model selection.**

Article

Introduction

Analyses of nucleotide sequences can provide a range of valuable insights into evolutionary relationships and timescales, allowing various biological questions to be addressed. The problem of inferring phylogenies and evolutionary divergence times is a statistical one, such that inferences are dependent on reliable models of the evolutionary process (Felsenstein 1983). Bayesian methods provide a powerful framework for estimating phylogenetic trees and evolutionary rates and timescales using parameter-rich models (Huelsenbeck et al. 2001; Yang and Rannala 2012). Model-based phylogenetic inference in a Bayesian framework has several desirable properties: It is possible to include detailed descriptions of molecular evolution (Dutheil et al. 2012; Heath et al. 2012); many of the model assumptions are explicit (Sullivan and Joyce 2005); large parameter spaces can be explored efficiently (Nylander et al. 2004; Drummond et al. 2006); and uncertainty is naturally incorporated in the estimates. As a consequence, the number and complexity of evolutionary models for Bayesian inference has grown rapidly, prompting considerable interest in methods of model selection (Xie et al. 2011; Baele et al. 2013).

Evolutionary models can provide useful insight into biological processes, but they are incomplete representations of molecular evolution (Goldman 1993). This can be problematic in phylogenetic inference when all the available models are poor descriptions of the process that generated the data (Gatesy 2007). Traditional methods of model selection do not allow the rejection, or falsification, of every model in the set of candidates being considered. Gelman and Shalizi (2013) recently referred to this as a critical weakness in current practice of Bayesian statistics. A different approach to model selection is to evaluate the adequacy, or plausibility (following Brown 2014a), of the model. This involves testing whether the data could have been generated by the model in question (Gelman et al. 2014).

Assessment of model adequacy is a critical step in Bayesian inference in general (Gelman and Shalizi 2013), and phylogenetics in particular (Brown 2014a). One method of evaluating the adequacy of a model is to use posterior predictive checks (Gelman et al. 2014). Among the first of such methods in phylogenetics was the use of posterior predictive simulations, proposed by Bollback (2002). The first step in this approach is to conduct a Bayesian phylogenetic analysis of the empirical

data. The second step is to use simulation to generate data sets with the same size as the empirical data, using the values of model parameters sampled from the posterior distribution obtained in the first step. The data generated via these posterior predictive simulations are considered to represent hypothetical alternative or future data sets, but generated by the model used for inference.

If the process that generated the empirical data can be described with the model used for inference, the posterior predictive data sets should resemble the empirical data set (Gelman et al. 2014). Therefore, the third step in assessing model adequacy is to perform a comparison between the posterior predictive data and the empirical data. This comparison must be done using a test statistic that quantifies the discrepancies between the posterior predictive data and the empirical data (Gelman and Meng 1996). The test statistic is calculated for each of the posterior predictive data sets to generate a distribution of values. If the test statistic calculated from the empirical data falls outside this distribution of the posterior predictive values, the model in question is considered to be inadequate.

Previous studies using posterior predictive checks of nucleotide substitution models have implemented a number of different test statistics. Some of these provide descriptions of the sequence alignments, such as the homogeneity of base composition (Huelsenbeck et al. 2001; Foster 2004), site frequency patterns (Bollback 2002; Lewis et al. 2014), and unequal synonymous versus nonsynonymous substitution rates (Nielsen 2002; Rodrigue et al. 2009). Brown (2014b) and Reid et al. (2014) introduced test statistics based on phylogenetic inferences from posterior predictive data sets. Some of the characteristics of inferred phylogenies that can be used as test statistics include the mean tree length and the median Robinson–Foulds distance between the sampled topologies in the analysis (Brown 2014b). Although several test statistics are available for assessing models of nucleotide substitution (Brown and ElDabaje 2009; Brown 2014a; Lewis et al. 2014), there are no methods available to assess the adequacy of molecular clock models.

Molecular clocks have become an established tool in evolutionary biology, allowing the study of molecular evolutionary rates and divergence times between organisms (Kumar 2005; Ho 2014). Molecular clock models describe the pattern of evolutionary rates among lineages, relying on external temporal information (e.g., fossil data) to calibrate estimates of absolute rates and times. The primary differences among the various clock models include the number of distinct substitution rates across the tree and the degree to which rates are treated as a heritable trait (Thorne et al. 1998; Drummond et al. 2006; Drummond and Suchard 2010; for a review see Ho and Duchêne 2014). For example, the strict clock assumes that the rate is the same for all branches, whereas some relaxed clock models allow each branch to have a different rate. We refer to models that assume a large number of rates as being more parameter rich than models with a small number of rates (Ho and Duchêne 2014). Although molecular clock models are used routinely, the methods of assessing their efficacy are restricted to estimating and comparing their statistical fit. For example, a common means of model selection is to compare marginal likelihoods in a Bayesian framework (Baele et al. 2013). However, model selection can only evaluate the relative statistical fit of the models, such that it can lead to false confidence in the estimates if all the candidate models are actually inadequate.

In this study, we introduce a method for assessing the adequacy of molecular clock models. Using simulated and empirical data, we show that our approach is sensitive to underparameterization of the clock model, and that it can be used to identify the branches of the tree that are in conflict with the assumed clock model. In practice, our method is also sensitive to other aspects of the hierarchical model, such as misspecification of the node-age priors. We highlight the importance of methods of evaluating the adequacy of substitution models in molecular clock analyses.

New Approaches

A Method of Assessing Clock Model Adequacy

To evaluate the adequacy of molecular clock models, we propose a method of generating and analyzing posterior predictive data. In this method, the posterior predictive data sets are generated using phylogenetic trees inferred from branch-specific rates and times from the posterior samples (fig. 1). Because this method uses branch-specific estimates, it requires a fixed tree topology.

The first step in our method is to conduct a Bayesian molecular clock analysis of empirical data. We assume that this analysis obtains samples from the posterior distribution of branch-specific rates and times. These estimates are given in relative time, or in absolute time if calibration priors are used. In the second step, we take a random subset of these samples. For each of these samples, we multiply the branch-specific rates and times to produce phylogenetic trees in which the branch lengths are measured in substitutions per site (subs/site), known as phylograms. To assess model adequacy, we randomly select 100 samples from the posterior, excluding the burn-in. From these samples, posterior predictive data sets are generated by simulation along the phylograms and using the estimates of the parameters in the nucleotide substitution model. The third step in our approach is to use a clock-free method to estimate a phylogram from each of the posterior predictive data sets and from the empirical data set. For this step, we find that the maximum likelihood approach implemented in phangorn (Schliep 2011) is effective.

To compute our adequacy index, we consider the branch lengths estimated from the posterior predictive data sets under a clock-free method, such that there is a distribution of length estimates for each branch. We calculate a posterior predictive *P* value for each branch using the corresponding distribution obtained with the posterior predictive data sets. This value is important for identifying the length estimates for individual branches that are in conflict with the clock model. Our 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

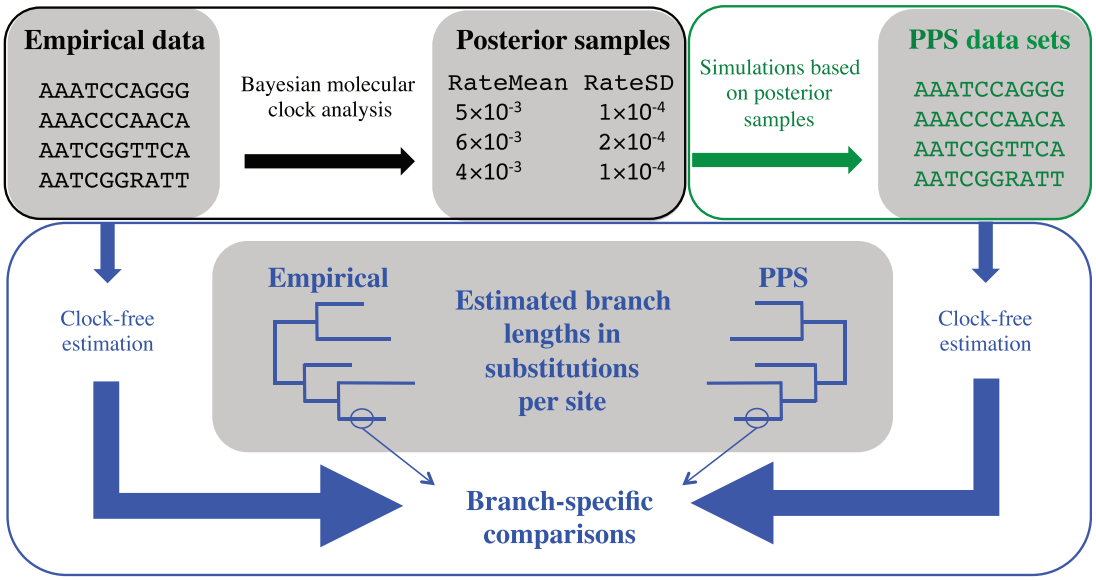

FIG. 1. Procedure for assessing the adequacy of molecular clock models. The top left box shows the components of a Bayesian clock analysis of empirical data, including samples from the posterior of the mean estimates and standard deviation of the substitution rates. The top right box shows the first step in assessing model adequacy using PPS. In our analyses, this step is performed using branch-specific rates and times. The bottom box shows our procedure for testing the clock model, which is based on the clock-free posterior predictive distribution of the length of each branch. The thin arrows indicate that the test statistic is the posterior predictive *P* value for each branch. PPS, posterior predictive simulations.

the posterior predictive data sets. We refer to our index as *A*, or overall plausibility of branch length estimates.

We also provide a measure of the extent to which the branch length estimates from the clock-free method differ from those obtained using the posterior predictive simulations. To do this, we calculate for each branch the absolute difference between the empirical branch length estimated using a clock-free method and the mean branch length estimated from the posterior predictive data. We then divide this value by the empirical branch length estimated using a clock-free method. This measure corresponds to the deviation of posterior predictive branch lengths from the branch length estimated from the empirical data. For simulations and analyses of empirical data, we present the median value across branches to avoid the effect of extreme values. We refer to this measure as "branch length deviation," of which low values represent high performance.

We also investigated the uncertainty in the estimates of posterior predictive branch lengths. This is useful because it provides insight into the combined uncertainty in estimates of rates and times. The method we used was to take the width of the 95% quantile range from the posterior predictive data sets, divided by the mean length estimated for each branch. This value, along with the width of the 95% credible interval of the rate estimate from the original analysis, can then be compared among clock models to investigate the increase in uncertainty that can occur when using complex models.

Results

Assessment of Clock Model Adequacy in Simulated Data

We first evaluated the accuracy and uncertainty of substitution rate estimates from simulated data. To do this, we compared the values used to generate the data with those estimated using each of three clock models: Strict clock, random local clocks (Drummond and Suchard 2010), and the uncorrelated lognormal relaxed clock (Drummond et al. 2006). We regarded the branch-specific rates as accurate when the rate used for the simulation was contained within the 95% credible interval. We found that rate estimates were frequently inaccurate under five circumstances: Clock model underparameterization; rate autocorrelation among branches (Kishino et al. 2001); uncorrelated beta-distributed rate variation among lineages; misleading node-age priors (i.e., node calibrations that differ considerably from the true node ages); and when data were generated under a strict clock but analyzed with an underparameterized substitution model (fig. 2*a*). When analyses were performed using the correct or an overparameterized clock model, more than 75% of branch rates were accurately estimated, such that the true value was contained within the 95% credible interval (fig. 2*a*). In most simulation schemes, the uncorrelated lognormal relaxed clock had high accuracy, at the expense of a small increase in the uncertainty compared with the other models (fig. 2*b*). These results are broadly similar to those of Drummond et al. (2006), who also found that underparameterization of the clock model resulted in low accuracy in rate estimates, whereas overparameterization had a negligible effect on accuracy.

We analyzed data generated by simulation to test our method of assessing the adequacy of molecular clock models. The *A* index was approximately proportional to the branch length deviation (fig. 3*a*). We found *A* to be $\geq$0.95 (indicating high performance) when the model used in the analyses matched that used to generate the data, or when it was overparameterized. When the assumed model was

underparameterized, *A* was ≤0.92. The uncertainty obtained using posterior predictive branch lengths was sensitive to the rate variance in the simulations. For this reason, estimates from data generated according to a strict clock or an uncorrelated lognormal relaxed clock had lower uncertainty than estimates from data generated under local clocks, regardless of the model used for analysis (fig. 3*b*). Estimates made using the uncorrelated lognormal relaxed clock had a larger variance in three analysis schemes: When data were generated with autocorrelated rates across branches; when data were generated with beta-distributed rates across branches; and when there was a misleading prior for the node ages. For analyses with substitution model underparameterization, our method incorrectly provided greater support for the more complex clock model, indicating that rate variation among lineages was overestimated (fig. 3).

We used our simulated data and posterior predictive simulations to investigate the performance of the multinomial test statistic for evaluating the adequacy of molecular clock models. This test statistic was originally designed to assess models of nucleotide substitution (Goldman 1993; Bollback 2002) and can perform well compared with some of the other existing test statistics (Brown 2014b). The multinomial test statistic for the empirical alignment can be compared with the distribution of test statistics from posterior predictive data sets to produce a posterior predictive *P* value. We find that the multinomial test statistic correctly identified when the substitution model was matched or underparameterized (fig. 4). The multinomial likelihood did not have the power to detect clock model adequacy, but it was sensitive to rate variation among lineages, primarily from the simulation involving autocorrelated rates and when the node-age prior was misleading (fig. 4).

Assessment of Clock Model Adequacy for Empirical Data

We used three clock models, as in our analyses of simulated data, to analyze a broad range of nucleotide sequence data

		Simulated													
		a. Accuracy							b. Uncertainty						
		SC	LOC	UCL	ACL	BIM	PRI	GTRG	SC	LOC	UCL	ACL	BIM	PRI	GTRG
Estimated	SC	1.00	0.18	0.31	0.04	0.03	0.10	<0.01	0.07	0.06	0.07	0.06	0.07	0.08	0.04
	RLC	0.82	0.78	0.35	0.19	0.20	0.06	0.01	0.20	0.23	0.20	0.31	0.38	0.30	0.14
	UCL	1.00	0.75	0.97	0.59	0.93	0.71	<0.01	0.16	1.40	0.31	0.88	1.45	2.74	0.23

Fig. 2. Mean values of (*a*) accuracy and (*b*) uncertainty of branch rate estimates from molecular clock analyses of simulated data. Each cell shows the results of 100 replicate analyses. Accuracy is measured as the proportion of data sets for which the rate used for simulation was contained in the 95% credible interval of the estimate. Darker shades in (*a*) represent high accuracy. Uncertainty is measured as the width of the 95% credible interval as a proportion of the mean rate. Dark shades in (*b*) represent small ranges in branch length estimates, and therefore low uncertainty. The initials stand for each of the schemes for estimation or simulation. SC, strict clock; LOC, local clock; UCL, uncorrelated lognormal relaxed clock; RLC, random local clock; ACL, autocorrelated relaxed clock; BIM, beta-distributed bimodal clock; PRI, misleading node-age prior; GTRG, data simulated under the parameter-rich general time-reversible substitution model with among-site rate heterogeneity.

		Simulated													
		a. Plausibility							b. Uncertainty						
		SC	LOC	UCL	ACL	BIM	PRI	GTRG	SC	LOC	UCL	ACL	BIM	PRI	GTRG
Estimated	SC	0.97 (0.08)	0.82 (0.19)	0.92 (0.10)	0.72 (0.19)	0.52 (0.33)	0.46 (0.40)	0.93 (0.08)	0.84	1.12	0.84	0.84	0.88	0.97	0.71
	RLC	0.97 (0.08)	0.95 (0.10)	0.92 (0.09)	0.87 (0.12)	0.84 (0.13)	0.88 (0.10)	0.94 (0.08)	0.85	1.13	0.84	0.86	0.94	0.93	0.72
	UCL	0.98 (0.07)	0.96 (0.04)	0.98 (0.07)	0.98 (0.03)	0.98 (0.03)	0.97 (0.02)	0.99 (0.05)	0.85	1.23	0.87	0.92	1.01	1.02	0.73

Fig. 3. Mean values of (*a*) plausibility, *A*, and (*b*) uncertainty as described by the posterior predictive simulations from clock analyses of simulated data. Each cell shows the results of 100 replicate analyses. Values in parentheses are the branch length deviations, of which lower values indicate good performance. The darker shades represent higher values of *A* and less uncertainty. High values of *A* represent good performance. In the case of uncertainty, small values indicate small ranges in posterior predictive branch lengths, and therefore low uncertainty. The initials stand for each of the schemes for estimation or simulation. SC, strict clock; LOC, local clock; UCL, uncorrelated lognormal relaxed clock; RLC, random local clock; ACL, autocorrelated relaxed clock; BIM, beta-distributed bimodal clock; PRI, misleading node-age prior; GTRG, data simulated under the parameter-rich general time-reversible substitution model with among-site rate heterogeneity.

sets: The *M* (matrix) gene of a set of coronaviruses; the *gag* gene of simian immunodeficiency virus (SIV; Wertheim and Worobey 2009); complete mitochondrial genomes of killer whales *Orcinus orca* (Morin et al. 2010); and 13 mitochondrial protein-coding genes of marine turtles (Duchene et al. 2012).

The uncorrelated lognormal relaxed clock was the best-fitting clock model according to the marginal likelihood for the coronaviruses, SIV, and the killer whales (table 1). For the marine turtles, the random local clock provided the best fit. In all the analyses of empirical data sets, the uncorrelated lognormal relaxed clock had the best performance according to our *A* index. The highest *A* index was 0.78 for the SIV and the killer whales, and the lowest uncertainty in posterior predictive branch lengths was 0.7 for the killer whales. The uncertainty for all other data sets was above 1, indicating that it was larger than the mean of the posterior predictive branch lengths.

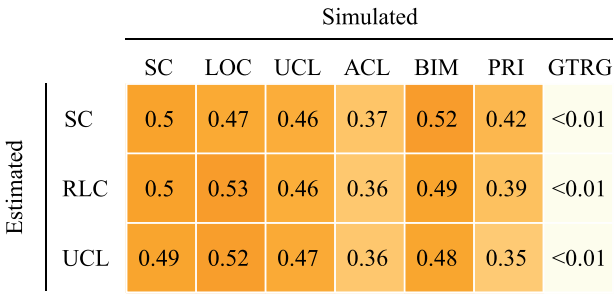

Fig. 4. Mean *P* values of the multinomial test statistic from posterior predictive simulations from simulated data. Each cell shows the results of 100 replicate analyses. Darker shades correspond to higher numbers. A value of 0.5 indicates that the model is adequate. The initials indicate the models for simulation and estimation. SC, strict clock; LOC, local clock; UCL, uncorrelated lognormal relaxed clock; RLC, random local clock; ACL, autocorrelated relaxed clock; BIM, beta-distributed bimodal clock; PRI, misleading node-age prior; GTRG, data simulated under the parameter-rich general time-reversible substitution model.

We calculated the multinomial test statistic for the empirical data sets using the posterior predictive data from a clock model analysis, as well as under a clock-free method. The multinomial test statistic from both methods suggested that the substitution model was inadequate for the SIV and the marine turtles, with posterior predictive *P* values below 0.05. The substitution model was identified as inadequate for the coronavirus data set by the multinomial test statistic estimated using posterior predictive data sets from a clock analysis ($P < 0.05$); however, it was identified as adequate when using a clock-free method ($P = 0.20$). The mitochondrial data set from killer whales represented the only case in which the substitution model was adequate according to both multinomial likelihood estimates. For the data sets from coronaviruses and killer whales, the clock models with the highest performance had *A* indices of 0.53 and 0.78, respectively (table 1). These indices are substantially lower than those obtained in analyses of simulated data when the clock model used for simulation and estimation was matched. However, we evaluated the posterior predictive *P* values for all branches in these empirical data sets and found that at least two-thirds of the incorrect estimates correspond to relatively short terminal branches (supplementary information, Supplementary Material online).

The branch length deviation in the empirical data ranged between 0.09 for the uncorrelated lognormal relaxed clock in the turtle data and 0.48 for the killer whale data analyzed with a strict clock (table 1). Low values for this metric indicate small differences between the posterior predictive and the empirical branch lengths. Although scores for this metric varied considerably between data sets, they were closely associated with the *A* indices for the different models for each data set individually. For example, in every empirical data set, the lowest branch length deviation was achieved by the model with the highest *A* index (indicative of higher performance). Importantly, the branch length deviation was not directly comparable with the *A* index between data sets.

Table 1. Statistical Fit and Performance of Three Molecular Clock Models in Analyses of Four Empirical Data Sets.

Data Set	Clock Model	Mean Number of Rate Changes (95% credible interval)	Mean Rate Estimate (95% credible interval)	Marginal Likelihood Estimate	Multinomial Test Statistic		*A* Index (mean branch-wise test statistic)	Uncertainty
					Clock	Clock-Free		
Coronaviruses	SC	—	2.04×10^{-5} (5.20×10^{-7} to 7.27×10^{-5})	−14,445.78	0.01	0.20	0.49 (0.24)	1.09
	RLC	0.80 (0–3)	2.48×10^{-5} (7.11×10^{-7} to 9.48×10^{-5})	−14,771.90	0.01	0.20	0.52 (0.24)	1.14
	UCL	—	2.18×10^{-5} (5.90×10^{-7} to 7.98×10^{-5})	−14,329.07	<0.01	0.20	0.53 (0.16)	1.11
SIV	SC	—	1.10×10^{-3} (8.22×10^{-4} to 1.45×10^{-3})	−3,275.23	0.01	<0.01	0.65 (0.46)	2.02
	RLC	2.43 (1–5)	1.10×10^{-3} (7.90×10^{-4} to 1.44×10^{-3})	−3,272.53	0.03	<0.01	0.65 (0.44)	2.03
	UCL	—	1.10×10^{-3} (7.90×10^{-4} to 1.56×10^{-3})	−3,256.00	0.04	<0.01	0.78 (0.23)	2.36
Killer whales	SC	—	3.78×10^{-3} (3.02×10^{-3} to 4.67×10^{-3})	−24,240.82	0.56	0.45	0.68 (0.48)	1.96
	RLC	0.79 (0–3)	3.77×10^{-3} (3.01×10^{-3} to 4.66×10^{-3})	−22,211.21	0.54	0.45	0.25 (0.48)	1.05
	UCL	—	3.90×10^{-3} (2.97×10^{-3} to 5.13×10^{-3})	−22,167.75	0.47	0.45	0.78 (0.47)	0.70
Marine turtles	SC	—	1.43×10^{-3} (1.34×10^{-3} to 1.52×10^{-3})	−37,505.44	<0.01	<0.01	0.66 (0.23)	4.14
	RLC	3.11 (1–6)	1.37×10^{-3} (1.15×10^{-3} to 1.56×10^{-3})	−37,454.97	<0.01	<0.01	0.68 (0.21)	3.96
	UCL	—	1.66×10^{-3} (1.39×10^{-3} to 1.90×10^{-3})	−37,488.56	<0.01	<0.01	0.70 (0.09)	4.17

Note.—The clock models are the strict clock (SC), uncorrelated lognormal relaxed clock (UCL), and the random local clock (RLC). For each data set, the number of rate changes is only estimated using the RLC. For the coronaviruses and SIV, the rate estimates are shown in subs/site/year, while those for the killer whales and marine turtles correspond to subs/site/My. Note that substitution model assessment under the clock-free method was conducted only once per data set. Rows in italics indicate the clock model with the lowest marginal likelihood estimate for each data set.

This is probably because the posterior predictive branch lengths have different amounts of uncertainty. In particular, the *A* index will tend to be low if the posterior predictive branch length estimates are similar to the empirical value but have low uncertainty. This would create a scenario with a small branch length deviation but also a low *A* index. This appears to be the case for the coronaviruses, for which all the clock models appear inadequate according to the *A* index, but with the uncorrelated lognormal relaxed clock having a small branch length deviation.

Discussion

Assessing the adequacy of models in phylogenetics is an important process that can provide information beyond that offered by traditional methods for model selection. Although traditional model selection can be used to evaluate the relative statistical fit of a set of candidates, model adequacy provides information about the absolute performance of the model, such that even the best-fitting model can be a poor predictor of the data (Gelman et al. 2014). There have been important developments in model adequacy methods and test statistics in the context of substitution models (Ripplinger and Sullivan 2010; Brown 2014b; Lewis et al. 2014) and estimates of gene trees (Reid et al. 2014). Here we have described a method that can be used for assessment of molecular clock models, and which should be used in combination with approaches for evaluating the adequacy of substitution models. The results of our analyses suggest that our method is able detect whether estimates of branch-specific rates and times are consistent with the expected number of substitutions along each branch. For example, in the coronavirus data set analyzed here, the best-fitting clock model was a poor predictor of the data, as was the substitution model. Our index is sensitive to underparameterization of clock models and has the benefit of being computationally efficient. In addition, our metric of uncertainty in posterior predictive branch lengths is sensitive to some cases of misspecification of clock models and node-age priors, but not to substitution model misspecification, as shown for our analyses of the coronavirus data set.

Analyses based on the random local clock and the data simulated under two local clocks generally produced low accuracy (fig. 2*a*), with lower *A* indices than the other models that were matched to the true model (fig. 3*a*). The substandard performance of the random local clock when it is matched to the true model is surprising. A possible explanation is that our simulations of the local clock represented an extreme scenario in which the rates of the local clocks differed by an order of magnitude. Previous studies based on simulations and empirical data demonstrated that this model can be effective when the rate differences are smaller (Drummond and Suchard 2010; Dornburg et al. 2012).

In our analyses of empirical data, even the highest values of our index were lower than the minimum value obtained in our analyses of simulated data when the three models matched those used for simulation. This is consistent with the results of previous studies of posterior predictive simulations, which have suggested that the proposed threshold for a test statistic using simulations is conservative for empirical data (Bollback 2002; Ripplinger and Sullivan 2010; Brown 2014b). It is difficult to suggest a specific threshold for our index to determine whether a model is inadequate. However, the interpretation is straightforward: A low *A* index indicates that a large proportion of branch rates and 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. However, our method allows the user to inspect the particular branches with inconsistent estimates, which can be useful for identifying regions of the tree that cause the clock model to be inadequate. Measuring the effect size of differences in the branch length estimates of the posterior predictive and empirical data can also be useful for quantifying potential errors in the estimates of node times and branch-specific rates.

An important finding of our study is that overparameterized clock models typically have higher accuracy than those that are underparameterized. This is consistent with a statistical phenomenon known as the bias–variance trade-off, with underparameterization leading to high bias, and overparameterization leading to high uncertainty. This was demonstrated for molecular clock models by Wertheim et al. (2009). Although our results show a bias when the model is underparameterized, we did not detect high uncertainty with increasing model complexity. This probably occurs because the models used here are not severely overparameterized. This is consistent with the fact that Bayesian analyses are robust to mild overparameterization because estimates are integrated over the uncertainty in additional parameters (Huelsenbeck and Rannala 2004; Lemmon and Moriarty 2004).

We note that our index is insensitive to the overparameterization in our analyses. This problem is also present in some adequacy statistics for substitution models (Bollback 2002; Ripplinger and Sullivan 2010). Identifying an overparameterized model is challenging, but a recent study proposed a method to do this for substitution models (Lewis et al. 2014). An equivalent implementation for clock models would also be valuable. Another potential solution is to select a pool of adequate models and to perform model selection using methods that penalize an excess of parameters, such as marginal likelihoods or information criteria.

We find that our assessment of clock model adequacy can be influenced by other components of the analysis. For example, multiple calibrations can create a misleading node-age prior that is in conflict with the clock model (Warnock et al. 2012; Duchêne et al. 2014; Heled and Drummond 2014). Although our simulations with misleading node calibrations were done using a strict clock, our method identified this scenario as clock model inadequacy when the models for estimation were the strict or random local clocks (fig. 3*a*). In the case of the uncorrelated lognormal relaxed clock, our method identified a misleading node-age prior as causing an increase in uncertainty (fig. 3*b*). This highlights the critical importance of selecting and using time calibrations appropriately, and we refer the reader to the comprehensive reviews of

this topic (Benton and Donoghue 2007; Ho and Phillips 2009). Another component of the analysis that can have an impact on the adequacy of the clock model is the tree prior, which can influence the estimates of branch lengths. Although one study suggested that the effect of the tree prior is not substantial (Lepage et al. 2007), its influence on divergence-time estimates remains largely unknown.

We found that substitution model underparameterization led to a severe reduction in accuracy. Overconfidence in incorrect branch lengths in terms of substitutions can cause bias in divergence-time estimates (Cutler 2000). However, this form of model inadequacy is incorrectly identified by the methods we used for estimation as a form of rate variation among lineages. For our data generated using a strict clock and an underparameterized substitution model, the *A* index rejected the strict clock and supported the overparameterized uncorrelated lognormal relaxed clock. On the other hand, the multinomial test statistic was sensitive to substitution model underparameterization, and to some forms of rate variation among lineages. The sensitivity of the multinomial likelihood to rate variation among lineages might explain why the substitution model was rejected for the coronavirus data set when using a clock model, but not when using a clock-free method. Due to this sensitivity and the substantial impact of substitution model misspecification, we recommend the use of a clock-free method to assess the substitution model before performing analyses using a clock model. Our results suggest that it is only advisable to perform a clock model analysis when an adequate substitution model is available. Other methods for substitution model assessment that are less conservative than the multinomial likelihood represent an interesting area for further research.

We find that the *A* index is sensitive to patterns of rate variation among lineages that conflict with the clock model used for estimation. This is highlighted in the simulations of rate variation among lineages under autocorrelated and the unusual beta-distributed rates. In these cases, the *A* index identified the uncorrelated lognormal clock as the only adequate clock model, despite an increase in uncertainty in both cases. Although other studies have also suggested that the uncorrelated lognormal relaxed clock can account for rate autocorrelation (Drummond et al. 2006; Ho et al. 2015), an increase in uncertainty can impair the interpretation of divergence-time estimates. We suggest caution when the uncertainty values are above 1, which occurs when the widths of the 95% credible intervals are greater than the mean parameter estimates.

In our analyses of the two virus data sets, the multinomial test statistic suggested that the best-fitting substitution model was inadequate. In the analyses of the SIV data, our index of clock model adequacy was 0.78, similar to that of killer whales, for which the substitution model appeared adequate. We recommend caution when interpreting estimates of evolutionary rates and timescales when the substitution model is inadequate. This typically suggests that the substitution process is not being modeled correctly, which can affect inferences of branch lengths regardless of whether a clock model is used or not. For this reason, the *A* index of 0.78 for the SIV data set might be overconfident compared with the same index obtained for the killer whale data. Previous research has also suggested that there are processes in the evolution of SIV that are not accounted for by current evolutionary models (Wertheim and Worobey 2009).

We also found that all the clock models were inadequate for the coronavirus sequence data. Our results might provide an explanation for the lack of consensus over the evolutionary timescale of these viruses. For example, a study of mammalian and avian coronaviruses estimated that these viruses originated at most 5,000 years ago (Woo et al. 2012). This result stands in contrast with a subsequent study that suggested a much deeper origin of these viruses, in the order of millions of years (Wertheim et al. 2013). Our results suggest that estimating the timescale of these viruses might not be feasible with the current clock models.

Our analysis of mitochondrial genomes of killer whales shows that even if the clock model performance is not as high as that obtained in the simulations that match the models used for estimation, a large proportion of the divergence-time estimates can still be useful. Examining the estimates of specific branch lengths can indicate whether many of the node-age estimates are reliable, or whether important branches provide unreliable estimates. We recommend this practice when the substitution model has been deemed adequate and when a substantial proportion of the branch lengths are consistent with the clock model (i.e., when the *A* index is high). We note that the mitochondrial genomes of killer whales have the lowest *A* index of any data set when analyzed using a random local clock. This might occur because the model identified an average of 0–3 rate changes along the tree (0.79 rate changes; table 1). Although rate variation is likely to be higher in this data set, it might not be sufficiently high for the model to detect it.

Analyses of mitochondrial protein-coding genes from marine turtles identified the substitution model as inadequate using the multinomial test statistic. The clock model with the highest performance had an *A* index of 0.70, which might be considered sufficient to interpret the divergence-time estimates for at least some portions of the tree. Again, the fact that the substitution model is inadequate precludes further interpretation of the estimates of evolutionary rates and timescales. This is a surprising result for a mitochondrial data set with several internal-node calibrations. A potential solution is to assess substitution-model adequacy for individual genes and to conduct the molecular clock analysis using only those genes for which an adequate substitution model is available. We believe that, with the advent of genomic data sets, this will become a feasible strategy in the near future.

Some of the reasons for the paucity of studies that assess model adequacy in phylogenetics include computational demand and the lack of available methods. In this study, we have presented a method of evaluating clock model adequacy, using a simple test statistic that can be computed efficiently. Assessment of clock model adequacy is an important complement to traditional methods of model selection for two primary reasons: It allows the researcher to reject all the available models if they are inadequate; and, as

implemented in this study, it can be used to identify the branches with length estimates that are implausible under the assumed model. The results of our analyses of empirical data underscore the importance of evaluating the adequacy of the substitution and clock models. In some cases, several models might be adequate, particularly when they are overparameterized. In this respect, methods for traditional model selection are important tools because they can be used to select a single best-fitting model from a set of adequate models. Further research into methods, test statistics, and software for evaluating model adequacy is needed, both to improve the existing models and to identify data sets that will consistently provide unreliable estimates.

Materials and Methods

Analyses of Simulated Data

We generated 100 pure-birth trees with 50 tips and root-node ages of 50 My using BEAST v2.1 (Bouckaert et al. 2014). We then simulated branch-specific rates under five clock model treatments using the R package NELSI (Ho et al. 2015). This program simulates rates under a given model and multiplies rates by time to produce phylogenetic trees in which the branch lengths represent subs/site, known as phylograms. These phylograms were then used to simulate the evolution of DNA sequences of 2,000 nt in the R package phangorn.

The five clock model treatments included the following: 1) A strict clock with a rate of 5×10^{-3} subs/site/My; 2) an uncorrelated lognormal relaxed clock (Drummond et al. 2006), with a mean rate of 5×10^{-3} subs/site/My and a standard deviation of 0.1; 3) a treatment in which a randomly selected clade with at least ten tips experienced an increase in the rate, representing a scenario with two local clocks (Yoder and Yang 2000), with rates of 1×10^{-2} and 1×10^{-3} subs/site/My; 4) a treatment with rate autocorrelation, with an initial rate of 5×10^{-3} subs/site/My and a ν parameter of 0.3 (Kishino et al. 2001); and 5) a treatment with rate variation that followed a beta distribution with equal shape parameters of 0.4 and centered at 5×10^{-3} subs/site/My, resulting in a bimodal shape. In every simulation, the mean rate was 5×10^{-3} subs/site/My, which is approximately the mean mitochondrial evolutionary rate in mammals, birds, nonavian reptiles, and amphibians (Pereira and Baker 2006). We selected this mean rate instead of sampling from the prior because our estimation methods involved an uninformative rate prior, and random samples from this can produce data sets with high sequence saturation or with low information content. We used the Jukes–Cantor substitution model for simulation (Jukes and Cantor 1969). This model allows us to avoid making arbitrary parameterizations of more parameter-rich models, which is not the focus of this study.

To explore the effect of substitution model underparameterization, we simulated additional data sets under a strict clock and a general time-reversible model with gamma-distributed rates among sites, using parameters from empirical data (Murphy et al. 2001). We analyzed these data sets using the same method as for the rest of the simulated data, including the use of the simpler Jukes–Cantor substitution model. We also explored the effect of using misleading node-age priors. To do this, we placed two time calibrations with incorrect ages. One calibration was placed in one of the two nodes descending from the root selected at random, with an age prior of 0.1 times its true age (i.e., younger than the truth). The other calibration was placed on the most recent node in the other clade descending from the root, with an age of 0.9 of the root age (i.e., older than the truth). For this scenario, we only used trees with more than one descendant in each of the two oldest clades. We show an example of the simulated phylogeny compared with this kind of marginal prior on node ages in the supplementary information, Supplementary Material online. Our study had 100 simulated data sets for each simulation treatment, for a total of 700 simulated alignments.

We analyzed the simulated alignments using Bayesian Markov chain Monte Carlo (MCMC) sampling as implemented in BEAST. We used three different clock models to analyze each of the simulated alignments: The strict clock, uncorrelated lognormal relaxed clock (Drummond et al. 2006), and random local clock (Drummond and Suchard 2010). We used the same tree prior and substitution model for estimation as those used for simulation. We fixed the age of the root to 50 My and fixed the tree topology to that used to simulate sequence evolution in every analysis. We analyzed the simulated data with an MCMC chain length of 2×10^7 steps, with samples drawn from the posterior every 2×10^3 steps. We discarded the first 10% of the samples as burn-in, and assessed satisfactory sampling from the posterior by verifying that effective sample sizes for all parameters were above 200 using the R package CODA (Plummer et al. 2006). We performed analyses using each of the three clock models for each of the 300 simulated data sets, for a total of 900 clock analyses.

We assessed the accuracy and uncertainty of the estimates made using each of the analysis schemes (fig. 2). To do this, we compared the simulated rates with the branch-specific rates in the posterior. Next, we tested the power of our method for assessing clock model adequacy using the simulated data under each of the scenarios of simulation and analysis. We provide example code and results in a public repository in GitHub (https://github.com/duchene/modad-clocks, last accessed July 1, 2015). We also tested the power of the multinomial test statistic to assess clock model adequacy in each of the 900 analyses. This test statistic quantifies the frequency of site patterns in an alignment and is appropriate for testing the adequacy of models of nucleotide substitution (Bollback 2002; Brown 2014b).

Analyses of Empirical Data

We used four published data sets to investigate the performance of our method of assessing clock model adequacy in empirical data. For each data set, we performed analyses in BEAST using each of the three clock models used to analyze the simulated data sets. To select the substitution model for each empirical data set, we used the Bayesian information criterion as calculated in the R package phangorn.

For each analysis of the empirical data sets, we ran the MCMC chain for 10^8 steps, with samples drawn from the posterior every 10^3 steps. We discarded the first 10% of the samples as burn-in and assessed satisfactory sampling from the posterior by verifying that the effective sample sizes for all parameters were above 200 using the R package CODA. We used stepping-stone sampling to estimate the marginal likelihood of the clock model (Gelman and Meng 1998; Lartillot and Philippe 2006; Xie et al. 2011). For each Bayesian analysis, we performed posterior predictive simulations as done for the simulated data sets, and assessed the substitution model using the multinomial test statistic. In addition, to estimate the clock-free multinomial test statistic, we analyzed each of the empirical data sets using MrBayes 3.2 (Ronquist et al. 2012). For these analyses we used the same chain length, sampling frequency, sampling verification method, and substitution model as in the analyses using clock models.

Our empirical data sets included nucleotide sequences of coronaviruses. This data set contained 43 sequences of 638 nt of a portion of the *M* (matrix) gene, as used by Wertheim et al. (2013). These sequences were sampled between 1941 and 2011. The best-fitting substitution model for this data set was GTR+Γ. We also used a data set of the *gag* gene of SIVs, which comprised 78 sequences of 477 nt, sampled between 1983 and 2004 (Wertheim and Worobey 2009). The best-fitting substitution model for this data set was GTR+Γ. We used the Bayesian skyline demographic model (Drummond et al. 2005) for the analyses of both of the virus data sets, and used the sampling times for calibration.

We analyzed a data set of the killer whale (*O. orca*), which contained 60 complete mitochondrial genome sequences of 16,386 nt (Morin et al. 2010). We calibrated the age of the root using a normal distribution with mean of 0.7 and a standard deviation of 5% of the mean, as used in the original study. The best-fitting substitution model for this data set was HKY+Γ. Finally, we analyzed a data set of several genera of marine turtles, which comprised 24 sequences of the 13 mitochondrial protein-coding genes (Duchene et al. 2012), and we selected the GTR+Γ substitution model. Following the scheme in the original study, we used calibrations at four internal nodes. The pure-birth process was used to generate the tree prior in the analyses of the killer whales and the marine turtles.

Supplementary Material

Supplementary information is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank the editor, Tracy Heath, and an anonymous reviewer for suggestions and insights that helped improve this article. This research was undertaken with the assistance of resources from the National Computational Infrastructure, which is supported by the Australian Government. D.D. was supported by an Australian National University HDR Merit Scholarship. E.C.H. was supported by a National Health and Medical Research Council Australia Fellowship (AF30). S.Y.W.H. was supported by the Australian Research Council (grant DP110100383).

References

Baele G, Li WLS, Drummond AJ, Suchard MA, Lemey P. 2013. Accurate model selection of relaxed molecular clocks in bayesian phylogenetics. *Mol Biol Evol.* 30:239–243.

Benton MJ, Donoghue PCJ. 2007. Paleontological evidence to date the tree of life. *Mol Biol Evol.* 24:26–53.

Bollback JP. 2002. Bayesian model adequacy and choice in phylogenetics. *Mol Biol Evol.* 19:1171–1180.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol.* 10:e1003537.

Brown JM. 2014a. Predictive approaches to assessing the fit of evolutionary models. *Syst Biol.* 63:289–292.

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

Brown JM, ElDabaje R. 2009. PuMA: Bayesian analysis of partitioned (and unpartitioned) model adequacy. *Bioinformatics* 25:537–538.

Cutler DJ. 2000. Estimating divergence times in the presence of an overdispersed molecular clock. *Mol Biol Evol.* 17:1647–1660.

Dornburg A, Brandley MC, McGowen MR, Near TJ. 2012. Relaxed clocks and inferences of heterogeneous patterns of nucleotide substitution and divergence time estimates across whales and dolphins (Mammalia: Cetacea). *Mol Biol Evol.* 29:721–736.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLOS Biol.* 4:e88.

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol.* 22:1185–1192.

Drummond AJ, Suchard MA. 2010. Bayesian random local clocks, or one rate to rule them all. *BMC Biol.* 8:114.

Duchene S, Frey A, Alfaro-Núñez A, Dutton PH, Gilbert TP, Morin PA. 2012. Marine turtle mitogenome phylogenetics and evolution. *Mol Phylogenet Evol.* 65:241–250.

Duchêne S, Lanfear R, Ho SYW. 2014. The impact of calibration and clock-model choice on molecular estimates of divergence times. *Mol Phylogenet Evol.* 78:277–289.

Dutheil JY, Galtier N, Romiguier J, Douzery EJP, Ranwez V, Boussau B. 2012. Efficient selection of brach specific models of sequence evolution. *Mol Biol Evol.* 24:1–15.

Felsenstein J. 1983. Statistical inference of phylogenies. *J R Stat. Soc A.* 146:246–272.

Foster P. 2004. Modeling compositional heterogeneity. *Syst Biol.* 53:485–495.

Gatesy J. 2007. A tenth crucial question regarding model use in phylogenetics. *Trends Ecol Evol.* 274:3–14.

Gelman A, Carlin JB, Stern HS, Rubin DB. 2014. Bayesian data analysis. New York: Chapman and Hall.

Gelman A, Meng X-L. 1996. Model checking and model improvement. In: Gilks WR, Richardson S, Spiegelhalter DJ, editors. Markov chain Monte Carlo in practice. New York: Chapman and Hall. p. 189–201.

Gelman A, Meng X-L. 1998. Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Stat Sci.* 13:163–185.

Gelman A, Shalizi CR. 2013. Philosophy and the practice of Bayesian statistics. *Br J Math Stat Psychol.* 66:8–38.

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

Heath TA, Holder MT, Huelsenbeck JP. 2012. A dirichlet process prior for estimating lineage-specific substitution rates. *Mol Biol Evol.* 29:939–955.

Heled J, Drummond AJ. 2014. Calibrated birth-death phylogenetic time-tree priors for Bayesian inference. *Syst Biol.* 64:369–383.

Ho SYW. 2014. The changing face of the molecular evolutionary clock. *Trends Ecol Evol.* 29:496–503.

Ho SYW, Duchêne S. 2014. Molecular-clock methods for estimating evolutionary rates and timescales. *Mol Ecol.* 23:5947–5965.
Ho SYW, Duchêne S, Duchêne D. 2015. Simulating and detecting autocorrelation of molecular evolutionary rates among lineages. *Mol Ecol Resour.* 15:688–696.
Ho SYW, Phillips MJ. 2009. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. *Syst Biol.* 58:367–380.
Huelsenbeck J, Rannala B. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Syst Biol.* 53:904–913.
Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294:2310–2314.
Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro H, editor. Mammalian protein metabolism. New York: Academic Press. p. 21–132.
Kishino H, Thorne JL, Bruno WJ. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol Biol Evol.* 18:352–361.
Kumar S. 2005. Molecular clocks: four decades of evolution. *Nat Rev Genet.* 6:654–662.
Lartillot N, Philippe H. 2006. Computing Bayes factors using thermodynamic integration. *Syst Biol.* 55:195–207.
Lemmon AR, Moriarty EC. 2004. The importance of proper model assumption in Bayesian phylogenetics. *Syst Biol.* 53:265–277.
Lepage T, Bryant D, Philippe H, Lartillot N. 2007. A general comparison of relaxed molecular clock models. *Mol Biol Evol.* 24:2669–2680.
Lewis PO, Xie W, Chen M-H, Fan Y, Kuo L. 2014. Posterior predictive Bayesian phylogenetic model selection. *Syst Biol.* 63:309–321.
Morin PA, Archer FI, Foote AD, Vilstrup J, Allen EE, Wade P, Durban J, Parsons K, Pitman R, Li L. 2010. Complete mitochondrial genome phylogeographic analysis of killer whales (*Orcinus orca*) indicates multiple species. *Genome Res.* 20:908–916.
Murphy JW, Eizirik E, O'Brien JS, Madsen O, Scally M, Douady CJ, Teeling E, Ryder OA, Stanhope MJ, de Jong WW, et al. 2001. Resolution of the early placental mammal radiation using bayesian phylogenetics. *Science* 294:2348–2351.
Nielsen R. 2002. Mapping mutations on phylogenies. *Syst Biol.* 51:729–739.
Nylander J, Ronqiuist F, Huelsenbeck J, Nieves-Aldery J. 2004. Bayesian phylogenetic analysis of combined data. *Syst Biol.* 53:47–67.
Pereira SL, Baker AJ. 2006. A mitogenomic timescale for birds detects variable phylogenetic rates of molecular evolution and refutes the standard molecular clock. *Mol Biol Evol.* 23:1731–1740.
Plummer M, Best N, Cowles K, Vines K. 2006. CODA: convergence diagnosis and output analysis for MCMC. *R News* 6:7–11.
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. 2010. Assessment of substitution model adequacy using frequentist and Bayesian methods. *Mol Biol Evol.* 27:2790–2803.
Rodrigue N, Kleinman CL, Philippe H, Lartillot N. 2009. Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons. *Mol Biol Evol.* 26:1663–1676.
Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol.* 61:539–542.
Schliep KP. 2011. Phangorn: phylogenetic analysis in R. *Bioinformatics* 27:592–593.
Sullivan J, Joyce P. 2005. Model selection in phylogenetics. *Annu Rev Ecol Evol Syst.* 36:445–466.
Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol Biol Evol.* 15:1647–1657.
Warnock RC, Yang Z, Donoghue P. 2012. Exploring uncertainty in the calibration of the molecular clock. *Biol Lett.* 8:156–159.
Wertheim JO, Chu DKW, Peiris JSM, Pond SLK, Poon LLM. 2013. A case for the ancient origin of coronaviruses. *J Virol.* 87:7039–7045.
Wertheim JO, Sanderson MJ, Worobey M, Bjork A. 2009. Relaxed molecular clocks, the bias–variance trade-off, and the quality of phylogenetic inference. *Syst Biol.* 58:1–8.
Wertheim JO, Worobey M. 2009. Dating the age of the SIV lineages that gave rise to HIV-1 and HIV-2. *PLoS Comput Biol.* 5:e1000377.
Woo PCY, Lau SKP, Lam CSF, Lau CCY, Tsang AKL, Lau JHN, Bai R, Teng JLL, Tsang CCC, Wang M, et al. 2012. Discovery of seven novel mammalian and avian coronaviruses in *Deltacoronavirus* supports bat coronaviruses as the gene source of *Alphacoronavirus* and *Betacoronavirus* and avian coronaviruses as the gene source of *Gammacoronavirus* and *Deltacoronavirus*. *J Virol.* 86:3995–4008.
Xie W, Lewis PO, Fan Y, Kuo L, Chen M-H. 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst Biol.* 60:150–160.
Yang Z, Rannala B. 2012. Molecular phylogenetics: principles and practice. *Nat Rev Genet.* 13:303–314.
Yoder AD, Yang ZH. 2000. Estimation of primate speciation dates using local molecular clocks. *Mol Biol Evol.* 17:1081–1090.