Molecular Phylogenetics and Evolution 78 (2014) 277–289

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

The impact of calibration and clock-model choice on molecular estimates of divergence times

Sebastián Duchêne [a],*, Robert Lanfear [b], Simon Y.W. Ho [a]

[a] School of Biological Sciences, University of Sydney, NSW 2006, Australia
[b] Centre for Macroevolution and Macroecology, Research School of Biology, Australian National University, Canberra, ACT 0200, Australia

ARTICLE INFO

Article history:
Received 17 March 2014
Revised 16 May 2014
Accepted 28 May 2014
Available online 6 June 2014

Keywords:
Relaxed molecular clock
Calibration
Bayesian phylogenetics
Evolutionary rate
Divergence times
Date estimation

ABSTRACT

Phylogenetic estimates of evolutionary timescales can be obtained from nucleotide sequence data using the molecular clock. These estimates are important for our understanding of evolutionary processes across all taxonomic levels. The molecular clock needs to be calibrated with an independent source of information, such as fossil evidence, to allow absolute ages to be inferred. Calibration typically involves fixing or constraining the age of at least one node in the phylogeny, enabling the ages of the remaining nodes to be estimated. We conducted an extensive simulation study to investigate the effects of the position and number of calibrations on the resulting estimate of the timescale. Our analyses focused on Bayesian estimates obtained using relaxed molecular clocks. Our findings suggest that an effective strategy is to include multiple calibrations and to prefer those that are close to the root of the phylogeny. Under these conditions, we found that evolutionary timescales could be estimated accurately even when the relaxed-clock model was misspecified and when the sequence data were relatively uninformative. We tested these findings in a case study of simian foamy virus, where we found that shallow calibrations caused the overall timescale to be underestimated by up to three orders of magnitude. Finally, we provide some recommendations for improving the practice of molecular-clock calibration.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

Our understanding of the tempo and mode of evolution has been transformed by the study of molecular data. One of the most illuminating fields of research has been the use of molecular clocks to estimate evolutionary rates and timescales. There has been much progress in this area, with sophisticated methods being able to handle large, multilocus data sets and to model various patterns of rate variation among lineages (dos Reis and Yang, 2011; Drummond et al., 2006; Rannala and Yang, 2007). However, all molecular clocks need to be calibrated so that estimates of rates and timescales are given in units of absolute time. Accordingly, identifying and dealing with sources of error in calibrations is a crucial component of molecular-clock analyses (Ho and Phillips, 2009; Inoue et al., 2010; Parham et al., 2012).

The most common method for calibrating molecular clocks is to use independent information to constrain the age of one or more nodes in the phylogenetic tree. We refer to these as the 'calibrating nodes' throughout this article. Calibrations are often based on a biogeographic event or on fossil evidence that can provide an estimate of when two lineages last shared a common ancestor. In the tree in Fig. 1, for example, a paleontological estimate of the divergence time of species 1 and 2 can be used to calibrate node A. By analysing the DNA sequences of these two species, we can estimate the absolute rate of molecular evolution along the two lineages descending from node A. The ages of other nodes in the tree can then be inferred by assuming some relationship among the substitution rates along different branches. A common strategy is to use several calibrating nodes, but this is only possible in taxonomic groups with a sufficient paleontological or biogeographic record. Although calibrations are often specified as point values, it is more appropriate to take into account their associated uncertainty (Ho and Phillips, 2009).

In all molecular-clock analyses, the strongest assumption about the substitution rate is that it is homogeneous across the tree, which is known as a 'strict' molecular clock (Zuckerkandl and Pauling, 1962). However, many empirical data sets fail to meet this assumption, with important consequences for estimates of divergence times (Yoder and Yang, 2000). As a response, various methods that can account for rate variation among lineages have been implemented (see reviews by Rutschmann, 2006; Welch and Bromham, 2005). These can be broadly classified as either

* Corresponding author. Address: School of Biological Sciences, Edgeworth David Building A11, University of Sydney, NSW 2006, Australia. Fax: +61 2 93514771.
E-mail address: sebastian.duchene@sydney.edu.au (S. Duchêne).

http://dx.doi.org/10.1016/j.ympev.2014.05.032
1055-7903/© 2014 Elsevier Inc. All rights reserved.

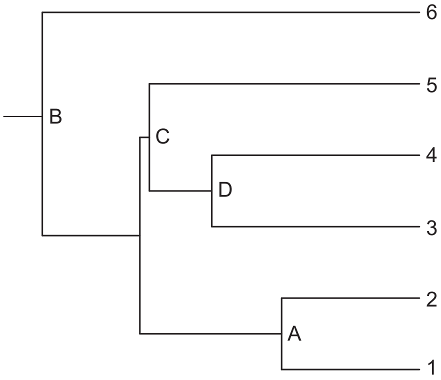

Fig. 1. Illustration of calibrating nodes in a phylogenetic tree. The shallowest node is A, whereas node B is the root. Note that only two lineages descend from node A, whereas deeper nodes are ancestral to a greater proportion of the tree.

uncorrelated or autocorrelated relaxed-clock models. In uncorrelated models, the rate along each branch of the phylogeny is an independent sample from a chosen probability distribution (Drummond et al., 2006; Rannala and Yang, 2007). The autocorrelated models assume that rates vary gradually throughout the phylogeny, so that the rates along neighbouring branches have some degree of correlation (Kishino et al., 2001; Sanderson, 2002, 1997; Thorne et al., 1998). The inclusion of calibrations can have an important impact on clock-model selection. In particular, informative calibration(s) can allow the pattern of rate variation among lineages to be resolved more reliably (Brandley et al., 2011; Lukoschek et al., 2012).

Molecular-clock estimates can be sensitive to the positions of the calibrations in the phylogenetic tree, especially when only a single or very few calibrations are available (Lee, 1999; Near and Sanderson, 2004). In general, calibrations at the root (node B in Fig. 1) or at deeper nodes are preferred over those at shallower nodes (e.g., nodes A and D in Fig. 1) (Hug and Roger, 2007; Sauquet et al., 2012; van Tuinen and Hedges, 2004). The estimate of the substitution rate is primarily based on the branches that lie between the calibrating nodes and the tips, so that deeper calibrations capture a larger proportion of the overall genetic variation.

Studies of various data sets have shown that analyses using multiple calibrations tend to produce more reliable estimates than those based on a single or few calibrations (Conroy and Van Tuinen, 2003; Smith and Peterson, 2002; Soltis et al., 2002). A possible explanation for this pattern is that the inclusion of only a small number of calibrations can lead to a biased estimate of the substitution rate if there is substantial among-lineage rate variation. Additionally, the use of multiple calibrations reduces the average genetic distance between the calibrating nodes and the nodes that are not calibrated (Marshall, 2008; Rutschmann et al., 2007). Another benefit of multiple calibrations is that they can improve the accuracy of date estimates in the presence of taxon undersampling (Linder et al., 2005).

In Bayesian molecular-clock analyses, calibrations can be specified in the form of prior probability densities for node ages (Drummond et al., 2006; Yang and Rannala, 2006). In some Bayesian implementations of relaxed clocks, these calibration priors, chosen by the user, interact with each other and with the prior distribution of the tree to give the marginal priors for the node ages (Heled and Drummond, 2012; Ho and Phillips, 2009; Kishino et al., 2001). This can lead to differences between the user-specified and marginal calibration priors, with unexpected impacts on the resulting estimates of divergence times (Heled and Drummond, 2012; Warnock et al., 2012). In practice, one can evaluate the extent of the problem by comparing the marginal and the user-specified priors, which is typically done by running a Bayesian analysis without sequence data. There are ongoing efforts to provide a more direct solution to this problem (Heled and Drummond, 2013).

Most research into molecular-clock calibrations has focussed on empirical data. A potential limitation of these studies is that the true divergence times and rates of evolution are unknown, making it impossible to assess the accuracy of the phylogenetic estimates. Here we perform an extensive simulation study to assess the impact of different calibration practices on the estimation of evolutionary timescales. By analysing data that were generated under known conditions, we are able to measure the error in the estimates of divergence times and substitution rates. We evaluate the impact of the number and position of calibrations, and investigate how these effects vary with sequence length, substitution rate, and misspecification of the molecular-clock model. We also test whether the correct distribution of rates among branches can be recovered using a Bayesian model-averaging approach. Finally, we examine the interactions among calibrations that lead to differences between the user-specified and marginal calibration priors. Our study provides insights into the effects of using different calibration strategies and offers a number of guidelines for future studies of evolutionary timescales.

2. Materials and methods

We simulated nucleotide sequence evolution to produce a large number of datasets, which we used to test hypotheses about calibration practices. The main advantage of using simulated data is that we have complete knowledge of the evolutionary parameters, including the phylogenetic tree, the node ages, the pattern of rate variation among lineages, and the substitution model. Therefore, assessing the impact of different assumptions in the analysis is much easier than with empirical data. However, we note that simulated data are ideal in the sense that stochastic deviation from the models used for the simulation is trivial, compared with the complex evolutionary dynamics of real data. For this reason, we also conducted an empirical case study using a simian foamy virus data set. This data set is well suited to test our findings because there are several calibrations available across the phylogeny of the virus.

2.1. Position of calibrations

2.1.1. Simulations

We simulated sequence evolution along phylogenetic trees of 50 taxa, generated randomly using a Yule speciation process. This branching model assumes a constant speciation rate with no extinction and is commonly used for data sets that include different species. We scaled each tree so that the age of the root was 50 time units, then we multiplied the branch lengths by a random variable representing the rate of evolution (substitutions/site/time), drawn from either a lognormal or exponential distribution. We parameterized the lognormal distribution with a mean of either 0.01 or 0.001 substitutions/site/time and a standard deviation of 0%, 10%, or 50% of the mean. We parameterized the exponential distribution with a mean of either 0.01 or 0.001 substitutions/site/time (note that the mean and standard deviation are equal in the exponential distribution). These are similar to the uncorrelated lognormal and exponential relaxed-clock models described by Drummond et al. (2006). Multiplying the simulated branch lengths (in time units) by the rate yielded trees with branch lengths measured in substitutions/site. We simulated sequence evolution along these trees using the Jukes–Cantor model to generate alignments of 1000, 2000, and 5000 nucleotides.

Our simulations encompass a total of 24 scenarios, corresponding to six parameterizations of the lognormal relaxed clock and two of the exponential relaxed clock, along with three different sequence lengths. We carried out these simulations using custom functions and the packages APE 3 (Paradis et al., 2004), geiger 1.3 (Harmon et al., 2008), and phangorn 1.7 (Schliep, 2011) in the R 2.15 programming language (R Core Team, 2008). The custom R functions used for this project are available from figshare (http://bit.ly/15u4OUB).

2.1.2. Phylogenetic analyses

We analysed the data in a Bayesian phylogenetic framework using BEAST 1.7.2 (Drummond and Rambaut, 2007; Drummond et al., 2012). We used the uncorrelated lognormal and exponential relaxed-clock models, which assume that the substitution rate along each branch is drawn independently from one of these distributions (Drummond et al., 2006). In these models, rates along adjacent branches do not have an *a priori* correlation with each other. We used both of the relaxed-clock models to analyse each of the simulated data sets, so that in half of the analyses the model used for analysis did not match that used to generate the sequence data. Our choice of relaxed-clocks for all of our analyses is appropriate because these models have been shown to perform well, even in data sets with very low rate variation (Brown and Yang, 2011; Drummond et al., 2006; Ho et al., 2005b).

We fixed the tree topology to focus on the estimates of rates and node times without the confounding influence of phylogenetic uncertainty. We matched the substitution model in the analysis to that used for simulation to reduce the impact of the node-density effect (Venditti et al., 2006). This also minimises the effect of substitution-model misspecification, which is not the focus of this study. To match the simulation settings, we used a Yule prior for the relative node times in the tree. The parameters of the Yule and substitution models were estimated from the data.

To determine the effect of different calibration positions, we selected a single random node in each simulated phylogenetic tree and used it as the calibrating node. We specified the calibration as a normal prior distribution for the age of the node, with the mean set to the true (simulation) value and the standard deviation set to 10% of the mean. We chose the calibrating node randomly 100 times for each of the 24 scenarios, so that our analyses collectively included calibrations at various positions across the tree. In total, we performed 4800 Bayesian phylogenetic analyses, comprising two relaxed-clock analyses of each of the 100 replicates for each of the 24 simulation scenarios.

In each analysis, posterior distributions of parameters were estimated by Markov chain Monte Carlo sampling (MCMC). Lengths of MCMC analyses varied according to the size of the sequence alignment: 10^7 steps for 1000 nucleotides, 10^9 steps for 2000 nucleotides, and 5×10^9 for 5000 nucleotides. Acceptable sampling and convergence to the stationary distribution were checked using LogAnalyser in the BEAST package. If effective sample sizes of any of the estimated parameters were below 200, the analysis was conducted again with a tenfold increase in chain length.

2.1.3. Statistical analyses

In order to evaluate the impact of calibration placement on phylogenetic analysis, we focused on the estimates of several key parameters. These included the age of the root, the age of the shallowest node, and the mean substitution rate. Our evaluation of these parameters was based on the accuracy and precision of the posterior estimates. Accuracy was quantified using an error score, calculated as the absolute difference between the mean posterior estimate and the true value, divided by the true value. An error score of zero indicates that the posterior mean is identical to the true value, reflecting an accurate estimate. Precision was quantified as the width of the 95% credibility interval of the estimate, divided by the posterior mean. A very precise estimate would have a precision score close to zero, with higher values representing a decrease in precision. For the coefficient of rate variation we compared the mean estimate across analyses instead of the error. We used this approach because this parameter is calculated as the standard deviation of branch-specific rates divided by the mean rate and weighted by branch length in BEAST, so it is not comparable to the standard coefficient of variation.

For the key parameters, we fitted linear regressions for our error and precision scores as functions of the $\log_{10}$-transformed calibration age. The regressions were performed with calibration ages transformed logarithmically ($\log_{10}$), because these values spanned multiple orders of magnitude. For the estimate of the age of the shallowest node, we considered whether the node was nested within the calibration as a binary variable in the linear model. We define ‘nesting’ here as the situation in which the node of interest is a descendant of the calibrating node. In the tree in Fig. 1, for example, node D is nested within node C. The results of this analysis should be interpreted with the understanding that the data might not conform to some of the general assumptions of ordinary linear regression. Rather than providing a predictive model, however, the purpose of this analysis is to describe the effect of calibration age on the performance of molecular-clock estimates. In this respect, the interpretation of the slope coefficient is particularly useful because it is a straightforward indicator of the association between calibration age and the reliability of parameter estimates.

We note that estimates of various parameters in a given analysis do not all necessarily have the same levels of error and precision. For example, an analysis can yield a low-error estimate of the age of the root, but this does not imply that there is also low error in the age estimate of the shallowest node. To verify this, we investigated whether the error and precision scores were correlated between the key parameters by calculating Spearman’s correlation coefficient (ρ). For the coefficient of rate variation we used the mean estimate instead of the error score. If $\rho > 0$ for the precision of two parameters, such as the ages of the root and the shallowest node, one can infer that the precision of these two parameters is positively correlated.

2.2. Number of calibrations

Our simulations to examine the effect of the number of calibrations were similar to those described in the previous section. To restrict our study to a feasible number of scenarios, we focused on a subset of representative simulation scenarios and fixed the sequence length to 2000 nucleotides. Specifically, we simulated sequence evolution with a lognormal distribution of substitution rates among branches, with mean of 0.01 and a standard deviation of either 10% or 50% of the mean.

Bayesian phylogenetic analyses were performed in BEAST with the settings described in the previous section, but in this case the number of calibrating nodes was 1, 5, 10, 20, or 49. There were a total of 20 sets of analyses, based on the two values of the standard deviation of the rate, five numbers of calibrations, and two relaxed-clock models used for analysis. We analysed 100 replicates for each of these combinations.

We fitted linear regressions of the error and precision scores for each of the parameters of interest, listed in the previous section, as functions of the number of calibrations. The regression slopes represent the effect of the number of calibrations. As in the statistical analysis for the age of the calibration, we considered whether the shallowest node was nested within the calibration as a binary variable in the linear model. We also investigated the correlation between error and precision, and between the parameter estimates, by calculating Spearman's correlation coefficient (ρ).

2.3. Model averaging and maximum a posteriori clock-model selection

For a subset of our simulations, we tested a model-averaging approach recently implemented in BEAST (Li and Drummond, 2012). In this method, the MCMC draws samples from a set of candidate clock models. In the current implementation, samples can only be drawn from an uncorrelated lognormal or exponential clock. Each model is sampled in proportion to its posterior probability, so the resulting estimates are weighted according to the probabilities of the two models. Although the purpose of this method is not model selection, one can infer that the model with the best fit is that with the highest posterior probability, known as the maximum a posteriori (MAP) model (Baele et al., 2013).

We tested whether the MAP model corresponded to that used to generate the data by conducting ten replicates of 32 simulation scenarios, for a total of 320 analyses. We have used fewer replicates here to ease computational demand, allowing us to investigate the performance of the method under a wide range of simulation conditions. The scenarios included different numbers of calibrations, ranging from 1 to 49, rates sampled from a lognormal or an exponential distribution, and two levels of rate variation (s.d. 10% and 50% for the lognormal, and of 0.001 and 0.01 for the exponential). The settings used in the phylogenetic analyses were similar to those described above, but instead of fixing the clock model we allowed the MCMC to sample from the lognormal or exponential clock models.

2.4. Comparison of user-specified and marginal priors

For a subset of simulations, we compared the user-specified prior, marginal prior, and posterior distributions of the calibrating nodes. In this paper, the user-specified prior refers to the prior distribution that the user has chosen for the purposes of calibration, whereas the marginal prior is the resulting prior distribution for the age of the calibrating node after accounting for possible interactions with the remaining priors (such as the tree prior). As with our other simulations, this investigation involved sequences simulated along a Yule tree with a depth of 50 time units. Substitution rates among branches were assumed to follow a lognormal rate distribution with a mean of 0.01 substitutions/site/time and standard deviation of 10% of the mean. We simulated sequence evolution along the tree using the Jukes–Cantor model to generate alignments of 1000, 2000, and 5000 nucleotides.

Each of the data sets was analysed using the lognormal relaxed clock in BEAST with 2, 10, 20, and 49 calibrating nodes selected at random. The calibrations were implemented as normal priors, with the mean chosen to match the true (simulation) value and with a standard deviation of 10% of the mean. To estimate the marginal prior distribution of the age of each calibrating node, each analysis was conducted without sequence data so that MCMC samples were drawn from the joint prior distribution. Simulations and phylogenetic analyses were replicated ten times. To compare the user-specified prior, marginal prior, and posterior distributions of the age of each calibrating node, we calculated the mean and the coefficient of variation. These measures can reveal differences related to the skewness and kurtosis of the distributions.

2.5. Case study: Simian foamy virus

2.5.1. Data collection

We analysed a data set comprising sequences from simian foamy virus (SFV). This single-stranded DNA virus is endemic to all primates, causing chronic and asymptomatic infections (Meiering and Linial, 2001). The host and virus tree topologies are congruent and are highly supported, with cross-species transmission thought to be rare and limited to closely related host species (Liu et al., 2008). The most well known cases involve humans infected through contact with chimpanzee blood (Sandstrom et al., 2000). Thus, there is strong support for long-term codivergence between SFV lineages and their primate hosts. Under this assumption, a previous study used the divergence times of the host species to calibrate the molecular clock of SFV (Switzer et al., 2005). The highly supported topology, reliable calibrations, and availability of sequence data make SFV a useful case for studying the effects of the placement of calibrating nodes.

We downloaded 18 complete genome sequences of SFV (11,954 nucleotides) from GenBank (Supplementary Table S1). Although this virus infects all primates screened so far, we chose SFV sequences from host species that were represented by multiple SFV samples. These host species were common chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), Bornean orangutan (*Pongo pygmaeus*), greater spot-nosed guenon (*Cercopithecus nictitans*), green monkey (*Chlorocebus sabaeus*), Barbary macaque (*Macaca sylvanus*), and common marmoset (*Callithrix jacchus*). The sequences were aligned using Clustal W2 v2.0 (Larkin et al., 2007) and visually inspected for possible frameshifts or other anomalies.

We included five calibrations based on SFV-primate codivergence events with the associated uncertainties, as reported by Switzer et al. (2005). We chose this set of calibrations to allow direct comparison with the results from the original study, rather than to provide an accurate estimate of the timescale of SFV or its primate host species. These calibrations were: (i) Cercopithecidae–Hominidae split, 28 Ma with standard deviation (s.d.) of 2.5; crown Cercopithecidae, 25 Ma with s.d. of 2.5; crown Hominidae, 16.52 Ma with s.d. of 2.5; (iv) *Pan/Homo-Gorilla* split, 13 Ma with s.d. of 2.5; and (v) crown of the most divergent *Pan* lineages at 1.2 Ma with s.d. of 1.

2.5.2. Phylogenetic analyses

We analysed the SFV data using similar settings to those in our simulation study. We fixed the topology of the SFV lineages to match that of their primate hosts, as reported by Switzer et al. (2005). The GTR + G substitution model was chosen according to the Bayesian information criterion. Bayesian phylogenetic analysis was conducted using BEAST. Samples from the posterior were drawn every 10^5 MCMC steps over a total of 10^9 steps. Convergence to the stationary distribution was checked by inspection of the MCMC trace, and effective sample sizes were >500 for all parameters. We ran five separate analyses, using each of the five calibrations in turn. We used a Yule prior for the tree and compared the fit of the uncorrelated lognormal and exponential relaxed clocks using Bayes factors with the stepping-stone estimator of the marginal likelihood (Xie et al., 2011). In most cases, the Bayes factor either indicated no strong preference for either of the two models, or favoured the exponential model. Therefore, we always used the exponential relaxed clock for our analyses because it has fewer parameters than the lognormal relaxed clock (Supplementary Table S2).

2.5.3. Statistical analyses

We used the same measures of precision and error as in our simulation study, with the 'true' values of the node ages considered to be the mean ages of the calibrations. The true substitution rate for this data set was unknown, so for this parameter we only analysed its precision.

Ordinary linear regressions are inappropriate for small data sets such as our SFV case study, with only five possible calibrating nodes. Therefore, we used Spearman's correlation coefficient (ρ) to investigate the association between the precision and error of the estimates of the rate and node ages, with the calibration age on a $\log_{10}$ scale.

3. Results

3.1. Position of calibrations

Our study of the impact of the position of calibrating nodes considered various values for substitution rates, among-lineage rate variation, sequence lengths, and relaxed-clock models. We found that misspecification of the relaxed-clock model always had a negative impact on the error and precision of estimates of rates and node times. Other treatment variables, such as among-lineage rate variation and sequence length, did not have a consistent impact on error and precision. For instance, estimates based on shorter sequences (1000 nucleotides) did not always have higher error and lower precision than the estimates based on longer sequences (2000 and 5000 nucleotides). This suggests that misspecification of the relaxed-clock model is the major cause of estimation unreliability in our analyses, and that its effect is not alleviated by using longer sequences or data with lower among-lineage rate variation.

We investigated the effect of the position of the calibration on parameter estimates in the 48 combinations of simulation scenarios and analysis settings. The results were similar for all settings; to simplify our discussion, we show the results for three of the 48 combinations in Figs. 2 and 3 and in Table 1. The error and precision significantly improved as a function of calibration age for three metrics: the age of the root, the age of the shallowest node, and the substitution rate (Table 1; Figs. 2 and 3). This improvement was greater for the analyses in which the molecular-clock model was misspecified, as shown by the regression coefficients (Table 1).

There were no significant differences when the shallow node was nested within the calibration, so we report the model coefficients without this term. Deeper calibrations appeared to be beneficial for estimates of all the parameters of interest, but the greatest improvements were found in the error of the estimate of the substitution rate. One interesting result is that the error in the estimate of the age of the shallowest node was sometimes very high, even when deep calibrations were used. The mean estimate of the coefficient of rate variation was not associated with the calibration age. However, this parameter appeared to be sensitive to clock-model misspecification. When the clock model was misspecified, the estimated coefficient of rate variation was higher and less precise than when the clock model matched that used to generate the data.

We found a correlation between error and precision scores for estimates of key parameters, including the substitution rate, age of the root, and the age of the shallow node. The ρ values for pairwise comparisons of parameters were between 0.72 and 0.97 for the error and were 0.99 for all comparisons of the precision score (Supplementary Table S3). Therefore, an analysis with low error and high precision for one of the parameter estimates would display similar relative levels of error and precision for the estimates of other parameters.

3.2. Number of calibrations

The association between the number of calibrations and the error and precision of parameter estimates was similar across the 20 analytical settings; we present detailed results for three representative cases (Table 2; Figs. 4 and 5). The error and precision scores were more variable when fewer calibrations were used. Slope terms for the regressions of estimation error and precision as a function of the number of calibrations were significant for all parameters of interest, including the coefficient of rate variation. For the estimates of the age of the shallow node, there were no significant differences when we considered whether or not it was nested within the calibrations, so the results are shown without including this term in the linear models. Although the slopes were always negative, their values ranged between -2.7×10^{-2}

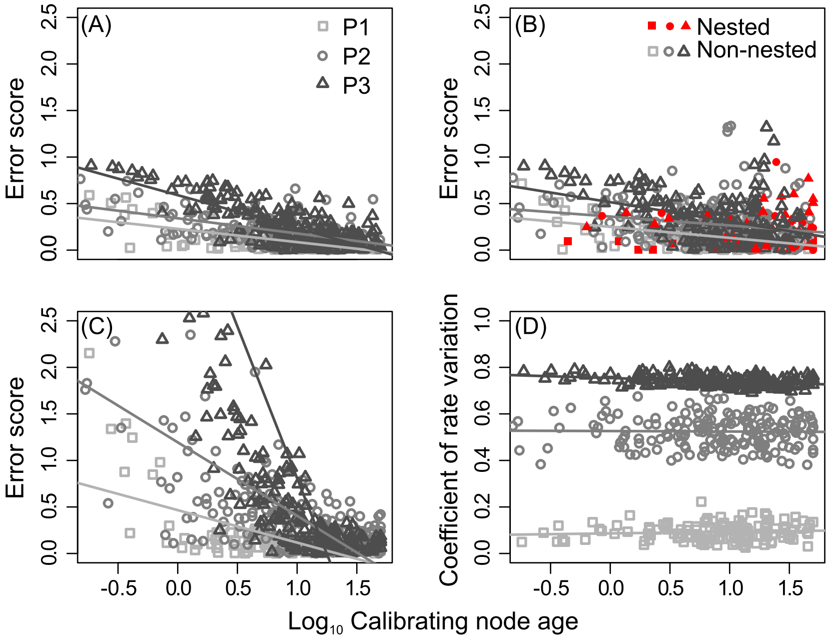

Fig. 2. Error score for the estimate of the age of the root (A), the age of the shallowest node (B), the substitution rate (C), and the mean of the coefficient of rate variation (D), as functions of the calibration age ($\log_{10}$ scale). In panels (A) through (C) the y-axis corresponds to the error score, where as in panel (D) it is the estimated mean of the coefficient of rate variation. Marker shape and shade represent the following analytical settings: (P1) sequence length of 5000 nucleotides, a simulated lognormal rate distribution with mean 10^{-3} and standard deviation of 10^{-4}, and analysed with a lognormal relaxed clock; (P2) sequence length of 1000 nucleotides, a simulated lognormal rate distribution with mean 10^{-2} and standard deviation 5×10^{-3}, and analysed with a lognormal relaxed clock; and (P3) sequence length of 2000 nucleotides, a simulated lognormal rate distribution with mean 10^{-3} and standard deviation 10^{-4}, and analysed with an exponential relaxed clock. The solid markers in panel (B) represent the simulations in which the shallow node was nested within the calibration.

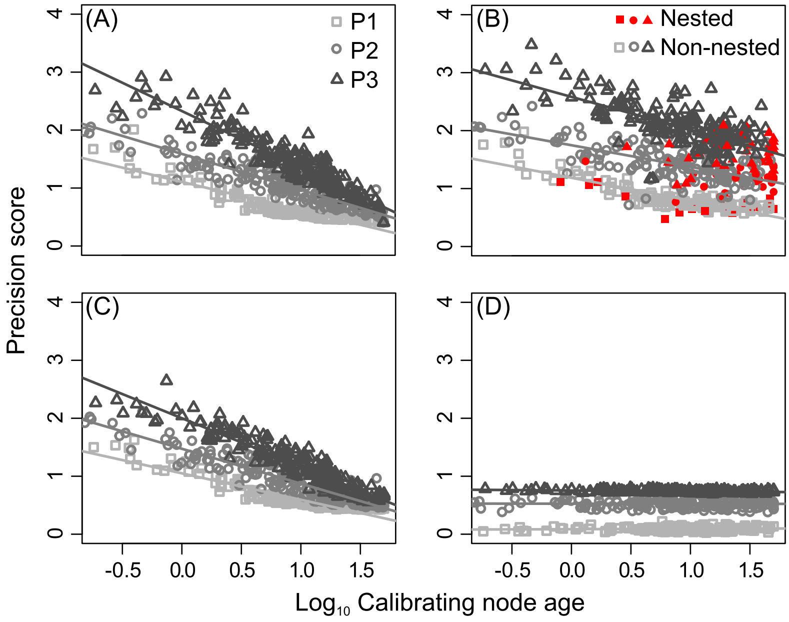

Fig. 3. Precision score for the estimate of the age of the root (A), the age of the shallowest node (B), the substitution rate (C), and the coefficient of rate variation (D), as functions of the calibration age ($\log_{10}$ scale). Note that lower values indicate higher precision. Marker shape and shade represent the analytical settings described in the caption for Fig. 2. The solid markers in panel (B) represent the simulations in which the shallow node was nested within one of the calibrations.

Table 1
Slopes and *P*-values for the error and precision of key parameters as functions of the age of the calibrating node (log10 scale). Low error and high precision correspond to lower values for these scores. In the case of the coefficient of rate variation, the mean estimate was used instead of the error. P1, P2, and P3 refer to the analyses described in Figs. 2 and 3. The estimate for the shallowest node corresponds to the node with the most recent estimated divergence time.

Estimated parameter per analysis		Error		Precision	
		Slope	*P*-value	Slope	*P*-value
Root node	P1	−0.142	<0.001	−0.493	<0.001
	P2	−0.163	<0.001	−0.621	<0.001
	P3	−0.355	<0.001	−0.976	<0.001
Shallowest node	P1	−0.118	<0.001	−0.394	<0.001
	P2	−0.096	<0.001	−0.373	<0.001
	P3	−0.207	<0.001	−0.568	<0.001
Substitution rate	P1	−0.349	<0.001	−0.456	<0.001
	P2	−0.789	<0.001	−0.602	<0.001
	P3	−3.262	<0.001	−0.834	<0.001
Coefficient of rate variation	P1	0.001*	0.277*	0.007	0.25
	P2	−0.002*	0.802*	−0.001	0.82
	P3	−0.003*	0.100*	−0.0014	<0.001

* These values for the coefficient of rate variation correspond to the estimated parameter value, instead of the error.

Table 2
Slopes and *P*-values for the error and precision scores for the key parameters as functions of the number of calibrations. N1, N2, and N3 refer to the analyses described in Figs. 4 and 5. The estimate for the shallowest node corresponds to the node with the most recent estimated divergence time.

Estimated parameter per analysis		Error		Precision	
		Slope	*P*-value	Slope	Slope
Root node	N1	−0.001	<0.001	−0.01	<0.001
	N2	−0.004	<0.001	−0.02	<0.001
	N3	−0.004	<0.001	−0.021	<0.001
Shallowest node	N1	−0.002	<0.001	−0.01	<0.001
	N2	−0.006	<0.001	−0.032	<0.001
	N3	−0.007	<0.001	−0.032	<0.001
Substitution rate	N1	−0.003	<0.001	−0.011	<0.001
	N2	−0.015	<0.001	−0.02	<0.001
	N3	−0.027	<0.001	−0.02	<0.001
Coefficient of rate variation	N1	<0.001*	0.023*	−0.0234	<0.001
	N2	−0.0006*	<0.001*	−0.012	<0.001
	N3	−0.0003*	<0.001*	−0.015	<0.001

* These values for the coefficient of rate variation correspond to the estimated parameter value, instead of the error.

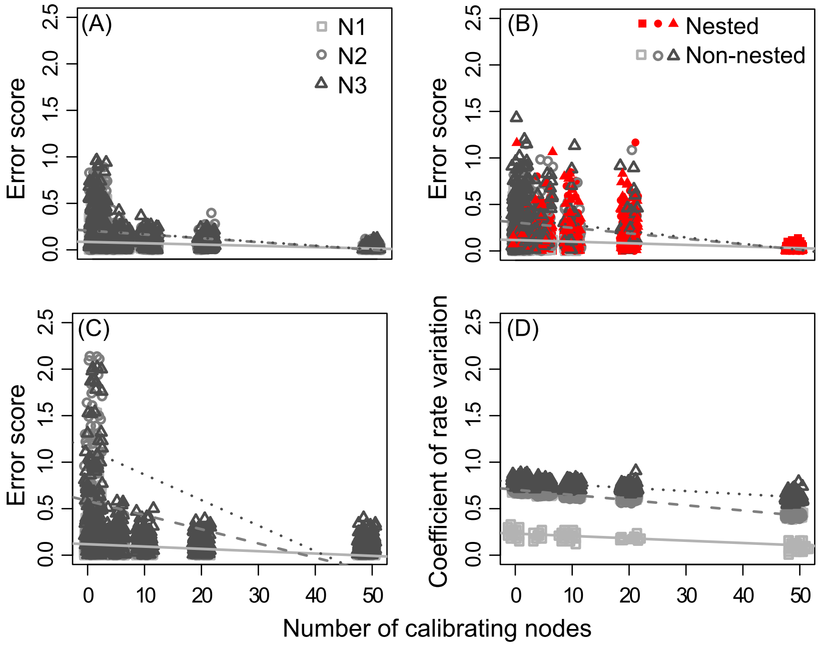

Fig. 4. Error score for the estimate of the age of the root (A), the age of the shallowest node (B), the substitution rate (C), and the mean of the coefficient of rate variation (D), as functions of the number of calibrations. In panels (A) through (C) the *y*-axis corresponds to the error score, where as in panel (D) it is the estimated mean of the coefficient of rate variation. Marker shape and shade represent the following analytical settings: (N1) simulated lognormal rate distribution with mean 10^{-2} and standard deviation 10^{-3}, analysed with a lognormal relaxed clock; (N2) simulated lognormal rate distribution with mean 10^{-2} and standard deviation 10^{-3}, analysed with an exponential relaxed clock; and (N3) simulated lognormal rate distribution with mean 10^{-2} and standard deviation 5×10^{-3}, analysed with an exponential relaxed clock. The solid markers in panel (B) represent the simulations in which the shallow node was nested within one of the calibrations. Note that the data have been jittered along the *x*-axis.

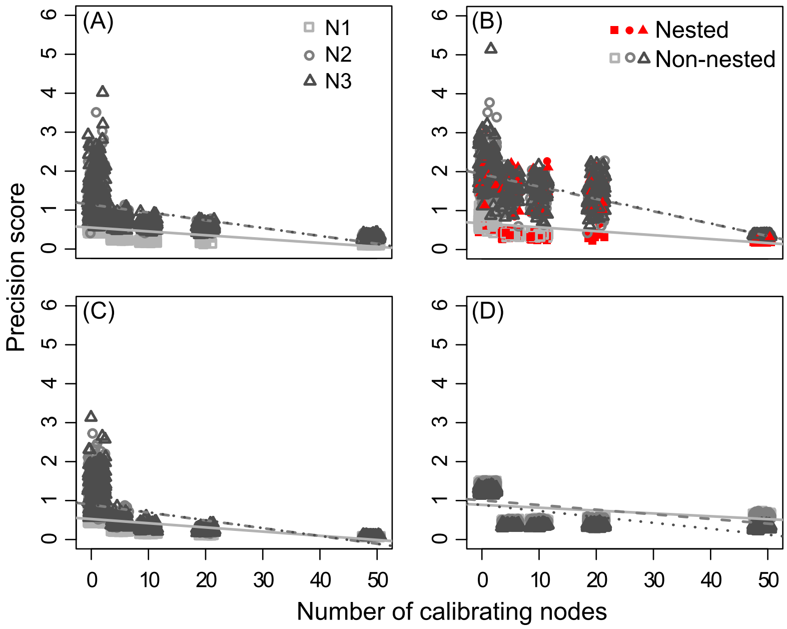

Fig. 5. Precision score for the estimate of the age of the root (A), the age of the shallowest node (B), the substitution rate (C), and the coefficient of rate variation (D), as functions of the number of calibrations. Note that lower values indicate higher precision. Marker shape and shade represent the analytical settings described in the caption for Fig. 4. Note that the data have been jittered along the *x*-axis.

and -1×10^{-4}, reflecting wide variation in the strength of this association (Table 2; Figs. 4 and 5).

Increasing the number of calibrations was beneficial in all cases. However, the association between error and precision of parameter estimates and the number of calibrations was stronger in analyses with high among-lineage rate variation and when the clock model was misspecified (Table 2; Figs. 4 and 5). Although the estimates of all key parameters appeared to improve with an increasing number of calibrations, this was especially important for the substitution rate, which displayed the steepest slope. For the estimate of the coefficient of rate variation, the number of calibrations only marginally affected the estimate, with slope values of the order of 10^{-4}.

3.3. Model averaging and maximum a posteriori clock model selection

The posterior probabilities of the MAP clock models were always very high, ranging from 0.99 to 1.0. This implies that the estimates were obtained using one prevailing model, rather than a mixture of the two candidate models. In this sense, the estimates are equivalent to those obtained when the analysis is conditioned on a single clock model.

Model averaging performed well when the rates were simulated according to an exponential distribution. It identified the correct clock model in at least seven out of ten simulations. In some instances in which the mean rate was low (0.001), it chose the correct clock in all of the simulations (Supplementary Table S4).

The performance of model averaging in favouring the correct model was poor for the simulations with lognormal rate distributions. In one simulation scenario, it sampled from the correct clock model preferentially in six out of ten simulation replicates. In the remaining scenarios, it could not identify the correct model in the majority of the simulation replicates. In a few scenarios, it always misclassified the model. The performance of this method was independent of the number of calibrations (Supplementary Table S4).

3.4. Comparison of user-specified and marginal priors

Our results for the user-specified prior, marginal prior, and posterior distributions of node ages were similar across replicates. We focus our discussion on one shallow and one deep node for the analyses with 2, 10, 20, and 49 calibrations (Table 3; Figs. 6 and 7). The shallow and deep nodes corresponded to the most recent and oldest calibrating nodes, respectively.

The distributions of the user-specified and the marginal priors were nearly identical. Differences in the means were between 0.0078 and 0.98 time units, while those for the coefficient of variation ranged from 0.0001 to 0.07. These results indicate that there was no conflict among calibrations, even when the analysis included a very large number of calibrating nodes.

There was a consistent pattern in the posterior age distributions of calibrating nodes. In all cases, the posteriors for different sequence lengths largely overlapped, and the 95% credibility intervals were narrower than in the user-specified and marginal priors. However, this association was stronger for an increasing number of calibrations, rather than increasing sequence length. We attribute this to some degree of interaction among the large number of calibrations, resulting in a more informative analysis.

3.5. Case study: Simian foamy virus

In our case study of simian foamy virus, we were able to test whether our findings in the simulation analyses were replicated in empirical data. When either of the two shallowest calibrations was used, the ages of the nodes for which we had calibrations were estimated to be much younger than their corresponding calibration age. In fact, the estimated mean could be up to three orders of magnitude lower than the calibration age (Table 4). Substitution rates inferred using the two shallowest calibrations, which yielded lower estimates of node ages, were an order of magnitude higher than when the other calibrations were employed.

We found no association between precision in the estimates and calibration age, with correlation coefficients between 0.063 and 0.062 for all calibrating nodes. An exception to this was the estimates of the substitution rate, displaying a negative correlation with the age of the calibrating node ($\rho = -0.74$). This result is consistent with those of our simulation study.

The results for error score and calibration age stand in contrast with those for estimation precision. The error was lower with increasing calibration depth, with correlation coefficients in the range of −0.73 and −0.81 for most nodes. The exception was the shallowest node (*Pan* crown), for which we found that error increased with calibration depth (Supplementary Table S5). This can be explained by the fact that the age of this node was considerably overestimated when we employed the two deeper calibrations (Cercopithecidae–Hominidae split and Cercopithecidae crown) (Table 4). These overall results for estimation error and precision are consistent with our findings in our simulation study, which showed that the benefit of using deeper calibrations is greater for the estimation error than it is for the precision score.

4. Discussion

Our simulation study reveals a number of patterns that are informative for phylogenetic studies of evolutionary timescales. Importantly, we found that sequence length did not considerably affect the molecular-clock estimates. This is due to the nature of our simulations, which resulted in very informative sequence data. Previous studies have found that sequence length is an important factor with shallow phylogenies or uninformative sequences (Brown and Yang, 2010). As sequence length increases, however, the error in estimates of node ages declines to a theoretical, nonzero limit (dos Reis and Yang, 2013; Rannala and Yang, 2007).

4.1. Clock-model choice

The estimates of substitution rates and timescales were more strongly influenced by the choice of relaxed-clock model than by the level of rate variation among lineages. This is consistent with the results of previous simulation studies, which have shown that estimates of branch-specific rates are sensitive to the choice of relaxed-clock model (Drummond et al., 2006; Heath et al., 2012;

Table 3
Mean and coefficient of variation of user-specified prior distributions, marginal prior distributions, and posterior distributions of node ages for sequence lengths of 1000, 2000, and 5000 nucleotides. Values in the table describe the distributions shown in Figs. 6 and 7. The shallow and deep nodes corresponded to the calibrating nodes with the most recent and oldest estimated divergence times, respectively.

Node age estimate	Number of calibrations	User-specified prior		Marginal prior mean		Posterior with 1000 nucleotides		Posterior with 2000 nucleotides		Posterior with 5000 nucleotides	
		Mean	Coefficient of variation	Mean	Coefficient of variation	Mean	Coefficient of variation	Mean	Coefficient of variation	Mean	Coefficient of variation
Shallow node	2	4.49	0.10	4.505	0.10	4.659	0.08	4.653	0.08	4.632	0.08
	10	3.39	0.10	3.382	0.07	3.447	0.06	3.388	0.06	3.448	0.06
	20	4.01	0.10	4.03	0.10	4.079	0.08	3.988	0.07	3.87	0.07
	49	4.08	0.10	4.071	0.10	4.264	0.07	4.354	0.06	4.116	0.06
Deep node	2	27.97	0.10	26.984	0.11	25.715	0.09	25.825	0.09	26.03	0.09
	10	24.53	0.10	24.611	0.09	23.924	0.06	25.547	0.05	24.601	0.05
	20	20.96	0.10	20.435	0.10	20.618	0.04	20.818	0.04	20.573	0.04
	49	20.12	0.10	20.151	0.08	19.073	0.04	19.511	0.03	19.407	0.03

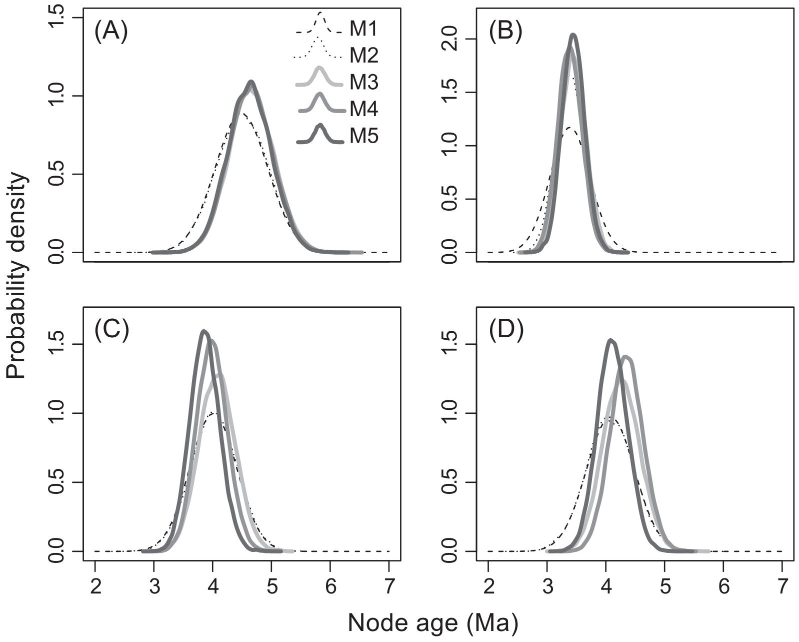

Fig. 6. Age distribution of a selected shallow node when the analysis includes 2 (A), 5 (B), 20 (C), and 49 (D) calibrating nodes. Line shades and patterns correspond to the user-specified prior distributions (M1), marginal prior distributions (M2), and posterior distributions for sequence lengths of 1000 (M3), 2000 (M4), and 5000 nucleotides (M5).

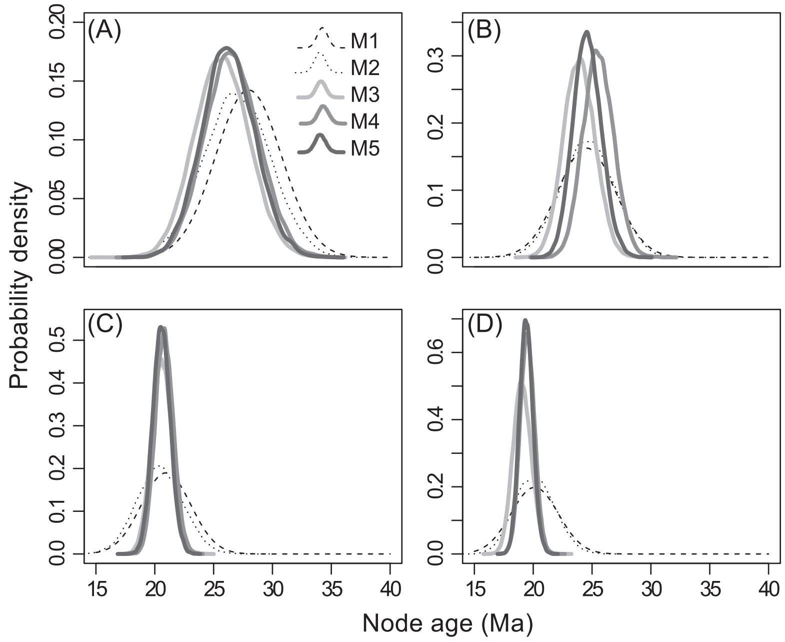

Fig. 7. Age distribution of a selected deep node when the analysis includes 2 (A), 5 (B), 20 (C), and 49 (D) calibrating nodes. Line shades and patterns represent the same distributions as in Fig. 6.

Ho et al., 2005b). Our findings emphasise the importance of rigorous selection of an appropriate relaxed-clock model, which can be done using various Bayesian methods (Baele et al., 2012; Lartillot and Philippe, 2006; Linder et al., 2005) or in a likelihood framework (Paradis, 2013). In this respect, a crucial limitation is that it is difficult to determine whether any of the available models provides an accurate reflection of the actual evolutionary process that produced the sequence data.

Model averaging is an attractive approach because the estimates of parameters are obtained from a mixture of models in proportion to their probabilities, rather than being conditioned on a single model (Li and Drummond, 2012). It can also be used to assess the fit of candidate models according to their posterior probabilities. Baele et al. (2013) found that this method often misclassified the clock model when the data were simulated according to a Yule speciation process with a lognormal rate distribution. We confirmed this result in our analyses, but also obtained the striking result that increasing the number of calibrations did not improve the performance of the model-averaging approach in favouring the correct clock model. If the model-averaging approach tends to sample the incorrect clock model, the errors in the estimates will be similar to those obtained when using the incorrect model

Table 4
Estimated node ages (Ma) and estimated rate of evolution (substitutions/site/Myr) for the different calibrations in the analysis of simian foamy virus. 95% CI denotes the width of the 95% credibility interval.

Calibrating node	Cercopithecidae–Hominidae (True age: mean = 28 and s.d = 2.5)		Crown Cercopithecidae (True age: mean = 25 and s.d = 2.5)		Crown Hominidae (True age: mean = 16.52 and s.d = 2.5)		*Pan/Homo-Gorilla* (True age: mean = 13 and s.d = 2.5)		Crown *Pan* (True age: mean = 1.2 and s.d = 1)		Rate of evolution	
	Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	95% CI
Cercopithecidae–Hominidae	27.31	8.36	26.08	9.76	19.88	13.86	13.58	12.43	4.74	7.13	0.02	0.009
Crown Cercopithecidae	25.16	9.89	24.21	8.38	22.65	10.33	20.79	11.17	3.35	8.04	0.02	0.009
Crown Hominidae	20.72	20.58	19.13	24.11	15.29	8.61	10.34	9.73	3.06	5.49	0.03	0.02
Pan/Homo-Gorilla	3.49	722.01	2.79	7160.12	3.06	684.55	2.48	505.78	2.09	3840.45	0.174	3258.34
Crown *Pan*	0.42	496.39	0.09	4917.16	0.37	459.89	0.22	178.01	0.34	140.10	0.176	3290.64

in a standard analysis. For empirical data, it is difficult to determine whether this method can recover the true rate distribution, so it may be necessary to extend the model-averaging approach to include a wider variety of rate distributions.

4.2. Calibration number and placement

The position and number of calibrating nodes are crucial components of molecular-clock analyses. This is relevant even when an informative sequence alignment, reliable calibrations, and appropriate relaxed-clock model are all at hand, which is the case for our simulations. In empirical studies, these factors can be largely uncertain, so the position and number of calibrations can have a stronger impact than that observed in our simulations. In all cases, increasing the number of calibrations and preferring calibrations closer to the root led to an improvement in estimates of substitution rates and divergence times. This is consistent with the findings of previous studies based on simulated and empirical data (Paradis, 2013; Sauquet et al., 2012). As a consequence, efforts to include a greater number of calibrations should be beneficial to molecular-clock analyses of empirical data, even when there is statistical support for rate homogeneity among lineages.

A practical concern is that there is a paucity of reliable calibrations in many taxonomic groups. It is not trivial to determine whether using a single reliable calibration is more appropriate than using several that are less reliable overall (Lee, 1999; Sanders and Lee, 2007). This is difficult to test because the accuracy and precision of fossil and biogeographic calibrations remain poorly understood (Andújar et al., 2014; Gandolfo et al., 2008; Magallón, 2004; Near et al., 2005). Our results show that levels of error and precision can vary widely when single calibrations are used. However, a single calibration can be effective if it is able to capture much of the among-lineage rate variation in the data, which is more likely to be achieved using calibrations at deep nodes in the tree.

The benefits of preferring deep over shallow calibrations, and multiple over few calibrating nodes, can be explained by the reliability of the estimate of the substitution rate. The estimate of this parameter improved considerably when deep or multiple calibrations were used. For instance, consider a molecular-clock analysis calibrated at node A in Fig. 1. If the two lineages descending from node A have a particularly high substitution rate, the mean rate across the tree will be overestimated and the timescale will be underestimated (Ho et al., 2005a; Lukoschek et al., 2012; Phillips, 2009). If the analysis is instead calibrated at the root (node B), the estimate of the substitution rate will be less subject to the vagaries of lineage-specific rates. Similarly, a rate estimated with two or more calibrating nodes will generally be more reliable than when node A is the only calibrating node.

In our simulations, the evolutionary rates were uncorrelated among branches. Although this model has been shown to be appropriate for many data sets (Brown et al., 2008; Drummond et al., 2006; Ho, 2009; Linder et al., 2011), some studies have found statistical support for models with rate autocorrelation (Lepage et al., 2007). When there are few or shallow calibrations, the risk of obtaining biased estimates of rates and divergence times can be higher when rates are autocorrelated among lineages than when they are uncorrelated. This is because the substitution rate would typically be more similar between parent and daughter branches than for a random pair of branches. For this reason, the reliability in the estimates of node ages might also depend on whether the nodes are nested within the clade defined by the calibrating node. To illustrate this point, a calibration at node C in Fig. 1 might be more effective for estimating the age of node D than for estimating the age of node A. Our recommendation of preferring deep and multiple calibrating nodes is also applicable in this case because it results in a higher proportion of all nodes being descendent from the calibrating node, when compared with strategies with few and shallow calibrating nodes. As we expected, we did not obtain more reliable estimates for nodes that were nested within the calibration because in our data the rates were uncorrelated, but we note that it may be an important aspect for empirical data sets that display rate autocorrelation.

When the focus is the estimate of the age of a shallow node, calibrations at the root or at very deep nodes can lead to a high error and low precision. This is demonstrated in our simulations, where the error in the estimate of the shallowest node was sometimes very high (Fig. 2B). Although analyses with deep calibrations can estimate many parameters more reliably than those with shallow calibrations, the rates along shallow branches are difficult to estimate with precision if rate variation is high. This would lead to unreliable estimates of the age of shallow nodes. This problem can be minimised by using an appropriate clock model or by including calibrations distributed throughout the tree.

Our case study of simian foamy virus shows that the impact of inadequate calibration strategies can be substantial, with effects that are difficult to predict in empirical data. In the case of viruses, an additional problem to the error in the rate estimate is that there is often substantial mutational saturation, especially in the basal branches of the tree. In some cases, even parameter-rich substitution models can fail to account for multiple substitutions adequately (Holmes, 2003). As a consequence, the lengths of deep branches in the tree will be underestimated, a phenomenon that has been referred to as ‘tree compression’ (Phillips, 2009). This is a probable cause of the dramatic underestimation of the timescale that we observed when analysing the simian foamy virus using shallow calibrations.

There are several strategies for mitigating the effect of tree compression on the estimate of the overall timescale of the tree. Rigorous selection of the best-fitting substitution model is crucial because branch-length estimation is sensitive to the parameters in the model, such as the proportion of invariant sites and the distribution of among-site rate heterogeneity (Sullivan and Joyce, 2005). When available substitution models underestimate lengths of basal branches, deep calibrations can be especially useful. A calibration at the root substantially reduces the potential for underestimating the ages of deep nodes. Recent work on empirical estimation of virus substitution models is of particular interest here, since empirical models have been shown to fit the data dramatically better than models estimated using standard parametric approaches (Bloom, 2014).

Although the strategy of using deep calibrations to alleviate tree compression can improve estimates of node ages, there are some potential repercussions that need to be considered. The mean substitution rate will be underestimated because the number of substitutions in the basal branches is underestimated, which in turn leads to an inflation of rate variation among lineages. Another consequence is overestimation of the ages of shallow nodes, a behaviour known as 'tree extension' (Phillips, 2009). Tree extension can be problematic in the analysis of taxonomic groups with few or no available calibrations. In such cases, it is common practice to include an outgroup taxon for which the divergence time with the ingroup is known (van Tuinen and Hedges, 2004). Although this brings the benefit of a calibration at the root, the inclusion of an outgroup taxon modifies the data set and the resulting estimates. If the outgroup taxon is very divergent from the ingroup, the basal branches of the tree will be prone to saturation. Furthermore, the substitution rates of the ingroup and outgroup might be very different, leading to high among-lineage rate variation, and low precision and accuracy in the estimates of node ages. This pattern is similar to some of our simulations, such as those with high rate variation, and those with low rate variation (simulation scenarios P1 and P2, respectively, in Figs. 2 and 3). We recommend that outgroup taxa should be chosen with the aim of minimising among-lineage rate variation. This is a critical point for population-level data, where including an outgroup taxon might have a particularly detrimental impact (Endicott et al., 2009; Ho and Larson, 2006).

4.3. Interaction among calibrations

A potential problem specific to Bayesian phylogenetic analyses is that the marginal priors for the ages of calibrating nodes can differ from the calibrations specified by the user. When multiple calibrations are employed, the user is incorporating information about the temporal order of the nodes and the phylogenetic relationships among lineages. Consequently, incompatibilities in topology or overlap between the calibration densities for different nodes will cause the marginal priors to differ from those specified by the user (Heled and Drummond, 2012; Kishino et al., 2001; Warnock et al., 2012). Our simulation study did not yield such differences between the user-specified and the marginal priors. However, our analyses were done using fixed tree topologies and our calibrations corresponded closely to the true node ages. These two factors minimise the possibility of conflict between calibrations. For real data sets, inspecting the marginal priors is essential because the phylogeny and actual divergence times are usually unknown. In addition, it should be borne in mind that multiple calibrations induce higher prior probabilities for the topologies that are consistent with the temporal order of nodes specified by the calibrations (Ho and Phillips, 2009).

We chose to use normally distributed calibrations in our study, but calibrations can also be modelled using lognormal, exponential, gamma, and uniform distributions, among others (Heath, 2012; Ho, 2007; Yang and Rannala, 2006). The choice of distribution is important because the information content of the prior depends on the concentration of the density around the central value. With calibrations of low information content, such as those modelled using wide uniform distributions, there is a higher probability that the user-specified and marginal densities will differ. This is due to maximum or minimum bounds being inadvertently enforced on some node ages because of overlap of calibration densities with low kurtosis. Our calibrations had a standard deviation of 10% of the mean, but in practice the uncertainty would often be much greater. Accordingly, we support recommendations by previous authors that the marginal priors should be inspected before proceeding with phylogenetic analysis (Heled and Drummond, 2012; Warnock et al., 2012). Heled and Drummond (2013) have recently implemented a potential solution for this problem.

4.4. Estimating among-lineage rate variation

Estimates of the coefficient of rate variation, which measures the degree of rate heterogeneity among lineages, did not seem to be influenced by the position and number of calibrations. This contradicts the expectation that among-lineage rate variation might be underestimated when there are shallow or few calibrating nodes (Benton and Donoghue, 2007; Ho and Phillips, 2009). We found that analyses involving misspecified clock models and shorter sequences led to a higher estimate of the coefficient of rate variation than did those with longer sequences and the correct clock model specified in the analysis, despite the fact that the estimate of this parameter should be the same. This pattern was consistent regardless of the position and number of calibrations. Although the coefficient of rate variation is useful for quantifying the degree to which the data conform to a strict clock, we recommend careful interpretation when comparing estimates across different data sets. According to our results, the estimate of this parameter might be highly sensitive to stochastic variation in the data, which is expected to increase with shorter sequence lengths. Therefore, high estimates for the coefficient of rate variation might indicate high stochastic variation, and not necessarily true among-lineage rate variation.

5. Conclusions

Overall, our study provides a number of insights into the impact of different calibration schemes on the estimates of evolutionary rates and timescales. The availability of reliable calibrations varies considerably among taxonomic groups. Some taxa have poor fossil and biogeographic records, leading to unreliable calibrations for molecular-clock analyses. In these cases, understanding the impact of different calibration strategies can help to improve estimates of evolutionary rates and timescales.

Our study addresses some general problems of molecular clock calibrations, but there remain some important areas for future study. Calibrations at the tips of the phylogenetic trees are effective when there is a measurable amount of evolutionary change between the collection time of samples (Drummond et al., 2003; Rambaut, 2000). They are mostly used for population-level data when ancient DNA sequences are available, or for organisms with very high rates of evolution, such as some viruses and bacteria. Previous studies have found that molecular-clock estimates are more reliable when tip calibrations have a wide temporal spread (Molak et al., 2013). The combination of calibrations at the tips and internal nodes is possible for some data sets, but further research is necessary to determine the effectiveness of this practice.

Misspecification of the clock model is a large source of error, especially when there are few informative calibrations. Although there has been substantial progress in this field, it may still be necessary to develop new models that accommodate among-lineage rate variation more accurately. Methods to explore model space, such as the model-averaging approach that we tested, can also be improved to include a wider array of models. Finally, the uncertainty in the topology and the use of different distributions for the calibrations is relevant for many empirical studies. Our simulation framework provides a starting point to explore this topic and test possible solutions. Future research in these directions will enable the continued development of molecular-clock models and will improve our estimates of the timescale of the tree of life.

Author contributions

S.D., S.Y.W.H., and R.L. designed the research. S.D. collected and analysed the data. S.D., S.Y.W.H., and R.L. wrote the paper.

Acknowledgments

We thank Joseph Heled and two anonymous reviewers for their helpful comments on our study. SD was supported by a Francisco José de Caldas Scholarship from the Colombian government and by a University of Sydney World Scholars Award. SYWH was supported by the Australian Research Council.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2014.05.032.

References

Andújar, C., Soria-Carrasco, V., Serrano, J., Gómez-Zurita, J., 2014. Congruence test of molecular clock calibration hypotheses based on Bayes factor comparisons. Methods Ecol. Evol..

Baele, G., Lemey, P., Bedford, T., Rambaut, A., Suchard, M.A., Alekseyenko, A.V., 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Mol. Biol. Evol. 29, 2157–2167.

Baele, G., Li, W.L.S., Drummond, A.J., Suchard, M.A., Lemey, P., 2013. Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. Mol. Biol. Evol. 30, 239–243.

Benton, M.J., Donoghue, P.C.J., 2007. Paleontological evidence to date the tree of life. Mol. Biol. Evol. 24, 26–53.

Bloom, J.D., 2014. An experimentally determined evolutionary model dramatically improves phylogenetic fit. bioRxiv. doi:http://dx.doi.org/10.1101/002899.

Brandley, M.C., Wang, Y., Guo, X., de Oca, A.N.M., Fería-Ortíz, M., Hikida, T., Ota, H., 2011. Accommodating heterogenous rates of evolution in molecular divergence dating methods: an example using intercontinental dispersal of Plestiodon (Eumeces) lizards. Syst. Biol. 60, 3–15.

Brown, R.P., Yang, Z., 2010. Bayesian dating of shallow phylogenies with a relaxed clock. Syst. Biol. 59, 119–131.

Brown, R., Yang, Z., 2011. Rate variation and estimation of divergence times using strict and relaxed clocks. BMC Evol. Biol. 11, 271.

Brown, J.W., Rest, J.S., García-Moreno, J., Sorenson, M.D., Mindell, D.P., 2008. Strong mitochondrial DNA support for a Cretaceous origin of modern avian lineages. BMC Biol. 6, 6.

Conroy, C.J., Van Tuinen, M., 2003. Extracting time from phylogenies: positive interplay between fossil and genetic data. J. Mammal. 84, 444–455.

Dos Reis, M., Yang, Z., 2011. Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. Mol. Biol. Evol. 28, 2161–2172.

Dos Reis, M., Yang, Z., 2013. The unbearable uncertainty of Bayesian divergence time estimation. J. Syst. Evol. 51, 30–43.

Drummond, A.J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7, 214.

Drummond, A.J., Pybus, O.G., Rambaut, A., Forsberg, R., Rodrigo, A.G., 2003. Measurably evolving populations. Trends Ecol. Evol. 18, 481–488.

Drummond, A.J., Ho, S.Y.W., Phillips, M.J., Rambaut, A., 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4, e88.

Drummond, A.J., Suchard, M.A., Xie, D., Rambaut, A., 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29, 1969–1973.

Endicott, P., Ho, S.Y.W., Metspalu, M., Stringer, C., 2009. Evaluating the mitochondrial timescale of human evolution. Trends Ecol. Evol. 24, 515–521.

Gandolfo, M.A., Nixon, K.C., Crepet, W.L., 2008. Selection of fossils for calibration of molecular dating models 1. Ann. Missouri Bot. Gard. 95, 34–42.

Harmon, L.J., Weir, J.T., Brock, C.D., Glor, R.E., Challenger, W., 2008. GEIGER: investigating evolutionary radiations. Bioinformatics 24, 129–131.

Heath, T.A., 2012. A hierarchical Bayesian model for calibrating estimates of species divergence times. Syst. Biol. 61, 793–809.

Heath, T.A., Holder, M.T., Huelsenbeck, J.P., 2012. A dirichlet process prior for estimating lineage-specific substitution rates. Mol. Biol. Evol. 29, 939–955.

Heled, J., Drummond, A.J., 2012. Calibrated tree priors for relaxed phylogenetics and divergence time estimation. Syst. Biol. 61, 138–149.

Heled, J., Drummond, A.J., 2013. Calibrated birth-death phylogenetic time-tree priors for Bayesian inference. arXiv Prepr. arXiv1311.4921.

Ho, S.Y.M., 2007. Calibrating molecular estimates of substitution rates and divergence times in birds. J. Avian Biol. 38, 409–414.

Ho, S.Y.W., 2009. An examination of phylogenetic models of substitution rate variation among lineages. Biol. Lett. 5, 421–424.

Ho, S.Y.W., Larson, G., 2006. Molecular clocks: when times are a-changin'. Trends Genet. 22, 79–83.

Ho, S.Y.W., Phillips, M.J., 2009. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. Syst. Biol. 58, 367–380.

Ho, S.Y.W., Phillips, M.J., Cooper, A., Drummond, A.J., 2005a. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. Mol. Biol. Evol. 22, 1561–1568.

Ho, S.Y.W., Phillips, M.J., Drummond, A.J., Cooper, A., 2005b. Accuracy of rate estimation using relaxed-clock models with a critical focus on the early metazoan radiation. Mol. Biol. Evol. 22, 1355–1363.

Holmes, E.C., 2003. Molecular clocks and the puzzle of RNA virus origins. J. Virol. 77, 3893–3897.

Hug, L.A., Roger, A.J., 2007. The impact of fossils and taxon sampling on ancient molecular dating analyses. Mol. Biol. Evol. 24, 1889–1897.

Inoue, J., Donoghue, P.C.J., Yang, Z., 2010. The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. Syst. Biol. 59, 74–89.

Kishino, H., Thorne, J.L., Bruno, W.J., 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. Mol. Biol. Evol. 18, 352–361.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947–2948.

Lartillot, N., Philippe, H., 2006. Computing Bayes factors using thermodynamic integration. Syst. Biol. 55, 195–207.

Lee, M.S.Y., 1999. Molecular clock calibrations and metazoan divergence dates. J. Mol. Evol. 49, 385–391.

Lepage, T., Bryant, D., Philippe, H., Lartillot, N., 2007. A general comparison of relaxed molecular clock models. Mol. Biol. Evol. 24, 2669–2680.

Li, W.L.S., Drummond, A.J., 2012. Model averaging and Bayes factor calculation of relaxed molecular clocks in Bayesian phylogenetics. Mol. Biol. Evol. 29, 751–761.

Linder, H.P., Hardy, C.R., Rutschmann, F., 2005. Taxon sampling effects in molecular clock dating: an example from the African Restionaceae. Mol. Phylogenet. Evol. 35, 569–582.

Linder, M., Britton, T., Sennblad, B., 2011. Evaluation of Bayesian models of substitution rate evolution—parental guidance versus mutual independence. Syst. Biol. 60, 329–342.

Liu, W., Worobey, M., Li, Y., Keele, B.F., Bibollet-Ruche, F., Guo, Y., Goepfert, P.A., Santiago, M.L., Ndjango, J.-B.N., Neel, C., 2008. Molecular ecology and natural history of Simian foamy virus infection in wild-living chimpanzees. PLoS Pathog. 4, e1000097.

Lukoschek, V., Keogh, J.S., Avise, J.C., 2012. Evaluating fossil calibrations for dating phylogenies in light of rates of molecular evolution: a comparison of three approaches. Syst. Biol. 61, 22–43.

Magallón, S.A., 2004. Dating lineages: molecular and paleontological approaches to the temporal framework of clades. Int. J. Plant Sci. 165, S7–S21.

Marshall, C.R., 2008. A simple method for bracketing absolute divergence times on molecular phylogenies using multiple fossil calibration points. Am. Nat. 171, 726–742.

Meiering, C.D., Linial, M.L., 2001. Historical perspective of foamy virus epidemiology and infection. Clin. Microbiol. Rev. 14, 165–176.

Molak, M., Lorenzen, E.D., Shapiro, B., Ho, S.Y.W., 2013. Phylogenetic estimation of timescales using ancient DNA: the effects of temporal sampling scheme and uncertainty in sample ages. Mol. Biol. Evol. 30, 253–262.

Near, T.J., Sanderson, M.J., 2004. Assessing the quality of molecular divergence time estimates by fossil calibrations and fossil-based model selection. Philos. Trans. R. Soc. London. Ser. B: Biol. Sci. 359, 1477–1483.

Near, T.J., Meylan, P.A., Shaffer, H.B., 2005. Assessing concordance of fossil calibration points in molecular clock studies: an example using turtles. Am. Nat. 165, 137–146.

Paradis, E., 2013. Molecular dating of phylogenies by likelihood methods: a comparison of models and a new information criterion. Mol. Phylogenet. Evol..

Paradis, E., Claude, J., Strimmer, K., 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20, 289–290.

Parham, J.F., Donoghue, P.C.J., Bell, C.J., Calway, T.D., Head, J.J., Holroyd, P.A., Inoue, J.G., Irmis, R.B., Joyce, W.G., Ksepka, D.T., 2012. Best practices for justifying fossil calibrations. Syst. Biol. 61, 346–359.

Phillips, M.J., 2009. Branch-length estimation bias misleads molecular dating for a vertebrate mitochondrial phylogeny. Gene 441, 132–140.

R Core Team, 2008. R: A Language and Environment for Statistical Computing. R Found. Stat. Comput.

Rambaut, A., 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. Bioinformatics 16, 395–399.

Rannala, B., Yang, Z., 2007. Inferring speciation times under an episodic molecular clock. Syst. Biol. 56, 453–466.

Rutschmann, F., 2006. Molecular dating of phylogenetic trees: a brief review of current methods that estimate divergence times. Divers. Distrib. 12, 35–48.

Rutschmann, F., Eriksson, T., Salim, K.A., Conti, E., 2007. Assessing calibration uncertainty in molecular dating: the assignment of fossils to alternative calibration points. Syst. Biol. 56, 591–608.

Sanders, K.L., Lee, M.S.Y., 2007. Evaluating molecular clock calibrations using Bayesian analyses with soft and hard bounds. Biol. Lett. 3, 275–279.

Sanderson, M.J., 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol. Biol. Evol. 14, 1218–1231.

Sanderson, M.J., 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol. Biol. Evol. 19, 101–109.

Sandstrom, P.A., Phan, K.O., Switzer, W.M., Fredeking, T., Chapman, L., Heneine, W., Folks, T.M., 2000. Simian foamy virus infection among zoo keepers. Lancet 355, 551–552.

Sauquet, H., Ho, S.Y.W., Gandolfo, M.A., Jordan, G.J., Wilf, P., Cantrill, D.J., Bayly, M.J., Bromham, L., Brown, G.K., Carpenter, R.J., 2012. Testing the impact of calibration on molecular divergence times using a fossil-rich group: the case of Nothofagus (Fagales). Syst. Biol. 61, 289–313.

Schliep, K.P., 2011. Phangorn: phylogenetic analysis in R. Bioinformatics 27, 592–593.

Smith, A.B., Peterson, K.J., 2002. Dating the time of origin of major clades: molecular clocks and the fossil record. Annu. Rev. Earth Planet. Sci. 30, 65–88.

Soltis, P.S., Soltis, D.E., Savolainen, V., Crane, P.R., Barraclough, T.G., 2002. Rate heterogeneity among lineages of tracheophytes: integration of molecular and fossil data and evidence for molecular living fossils. Proc. Natl. Acad. Sci. USA 99, 4430–4435.

Sullivan, J., Joyce, P., 2005. Model selection in phylogenetics. Annu. Rev. Ecol. Evol. Syst., 445–466.

Switzer, W.M., Salemi, M., Shanmugam, V., Gao, F., Cong, M., Kuiken, C., Bhullar, V., Beer, B.E., Vallet, D., Gautier-Hion, A., 2005. Ancient co-speciation of simian foamy viruses and primates. Nature 434, 376–380.

Thorne, J.L., Kishino, H., Painter, I.S., 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15, 1647–1657.

Van Tuinen, M., Hedges, S.B., 2004. The effect of external and internal fossil calibrations on the avian evolutionary timescale. J. Paleontol. 78.

Venditti, C., Meade, A., Pagel, M., 2006. Detecting the node-density artifact in phylogeny reconstruction. Syst. Biol. 55, 637–643.

Warnock, R.C.M., Yang, Z., Donoghue, P.C.J., 2012. Exploring uncertainty in the calibration of the molecular clock. Biol. Lett. 8, 156–159.

Welch, J.J., Bromham, L., 2005. Molecular dating when rates vary. Trends Ecol. Evol. 20, 320–327.

Xie, W., Lewis, P.O., 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., 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Mol. Biol. Evol. 23, 212–226.

Yoder, A.D., Yang, Z., 2000. Estimation of primate speciation dates using local molecular clocks. Mol. Biol. Evol. 17, 1081–1090.

Zuckerkandl, E., Pauling, L., 1962. Molecular disease, evolution and genetic heterogeneity. In: Kasha, M., Pullman, B. (Eds.), New York.