MBE

Article

Theoretical foundation of the RelTime method for estimating divergence times from variable evolutionary rates

Koichiro Tamura[1,2], Qiqing Tao[3,4], and Sudhir Kumar[3,4]

[1]Department of Biological Sciences, Tokyo Metropolitan University, Hachioji, Tokyo, Japan

[2]Research Center for Genomics and Bioinformatics, Tokyo Metropolitan University, Hachioji, Tokyo, Japan

[3]Institute for Genomics and Evolutionary Medicine, Temple University

[4]Department of Biology, Temple University

Correspondence to:

Sudhir Kumar

Temple University

Philadelphia, PA 19122, USA

E-mail: s.kumar@temple.edu

Abstract

The RelTime approach estimates timetrees from molecular data when evolutionary rates vary from branch to branch. It has been shown to perform well in analyses of simulated and empirical datasets where evolutionary rates vary extensively. RelTime is computationally efficient and scales well with increasing volumes of data. Consequently, it is being used for estimating divergence time from large datasets. Until now, RelTime has been used without a mathematical foundation. Here, we show that a relative rate framework (RRF) with a principle of minimum rate change is the basis of RelTime. Under RRF, we present analytical solutions for estimating relative rates and divergence times. For both real and simulated datasets, RRF produces estimates similar to those from Bayesian analyses, but RRF provides orders of magnitude increases in computational speed. These gains rise with increasing volumes of data. The mathematical foundation and computational efficiency of RRF makes it suitable for analysis not only of molecular sequence datasets, but also evolutionary trees where the branch lengths reflect the amount of non-molecular (e.g., morphological and traits) evolutionary changes.

INTRODUCTION

Inference of divergence times requires either an assumption of a constant rate throughout the tree (a molecular clock) or a statistical distribution to model the variation of evolutionary rates among lineages (Ho and Duchêne 2014; Kumar and Hedges 2016; dos Reis et al. 2016). Widely used Bayesian methods require specification of a probability distribution of evolutionary rates in the tree (e.g., lognormal distribution) and whether the rates are correlated among lineages (Thorne et al. 1998) or independent (Drummond and Rambaut 2007). In contrast, RelTime approach does not require such a probability distribution (Tamura et al. 2012); it estimates relative rates throughout the tree to generate relative node ages that can be transformed into absolute dates by using temporal constraints for one or more nodes (Tamura *et al.* 2012; Tamura *et al.* 2013). RelTime has been found to perform well in analyses of many large empirical datasets (Mello et al. 2017), and it shows high accuracy in analyses of simulated datasets where true times are known (Tamura et al. 2012; Filipski et al. 2014).

RelTime is much faster than current Bayesian methods and is computationally feasible for very large datasets (Tamura et al. 2012) (**Fig. 1a**). Consequently, it is being used by many researchers for estimating divergence times, especially for large datasets, e.g. Mahler et al. (2013), Bond et al. (2014) and Bonaldo et al. (2016). However, a mathematical foundation for the RelTime method is lacking, which is important to not only reveal its relationship with other methods (Ho and Duchêne 2014; Kumar and Hedges 2016; dos Reis et al. 2016), but also to avoid misunderstanding its relationship with a strict molecular clock locally or globally (Lozano-Fernandez et al. 2017). In the following, we present the theoretical foundation of the RelTime method. Then, we present analysis of datasets generated by computer simulation, where sequences were evolved according to the independent rate model (Drummond and Rambaut 2007), autocorrelated rate model (Thorne et al. 1998), and hybrid rate models (Beaulieu et al. 2015). We compare the true and simulated rates and divergence times to assess the performance of RelTime and Bayesian methods.

MATHEMATICAL THEORY

Theoretical analysis for three sequences with an outgroup

We begin with the simplest case where the evolutionary tree contains a clade with three ingroup taxa (subtree at node 5) and one outgroup taxon (**Fig. 2a**). In this tree, b_1 and b_2 represent the numbers of substitutions per site that have occurred in lineages leading to taxon 1 and taxon 2

from node 4. We assume that taxon 1 and 2 are sampled at the same evolutionary time (t_1 = 0 and t_2 = 0), which is frequently the case in molecular phylogenetic studies. In RelTime, the assumption of contemporaneous sampling of data from taxa allows us to treat the sampling times (equal to 0) as calibration points (Tamura et al. 2012). We are then able to estimate the relative evolutionary rates (r's) for all the branches as well as relative times (t's) of nodes 4 and 5. The following system of equations formalizes the RelTime method mathematically, where we relate relative rates for lineages (r_i's) and branch lengths (b_i's) in **Fig. 2a**.

We write,

$$r_1/r_2 = b_1/b_2, \text{ and} \quad [1]$$

$$r_3/r_{1,2} = b_3/\ell_{1,2}, \quad [2]$$

where r_{12} is the overall evolutionary rate of the lineage from node 5 that leads to sequences 1 and 2, and ℓ_{12} is average divergence of taxon 1 and 2 from node 5. Here, r_3 is the rate and b_3 is the branch length of the lineage leading to taxon 3 from node 5.

Also,

$$r_4 = ½(r_1 + r_2), \text{ and} \quad [3]$$

$$r_5 = ½(r_3 + r_4), \quad [4]$$

where r_4 is the rate for branch b_4 and the descendant clade at node 4. r_5 is the rate at node 5. We do not assume r's to be equal to each other in the above equations. Instead, equations [3] and [4] express a preference for positing a minimum change in rate between an ancestral branch and its immediate descendant clades.

We now have five unknowns (r_1, r_2, r_3, r_4, and r_5) and four equations. We can reduce one unknown by assuming that the evolutionary rates will be scaled such that the rate at the most recent common ancestor (node 5) is 1, *i.e.*,

$$r_5 = 1. \quad [5]$$

Using these five equations, we get:

$$r_1 = 4b_1(b_1 + b_2 + 2b_3 + 2b_4)/(b_1 + b_2)(b_1 + b_2 + 2b_4), \quad [6]$$

$$r_2 = 4b_2(b_1 + b_2 + 2b_3 + 2b_4)/(b_1 + b_2)(b_1 + b_2 + 2b_4), \quad [7]$$

$$r_3 = 4b_3/(b_1 + b_2 + 2b_3 + 2b_4), \text{ and} \quad [8]$$

$$r_4 = 2(b_1 + b_2 + 2b_4)/(b_1 + b_2 + 2b_3 + 2b_4). \quad [9]$$

The estimate of relative rate for a branch yields the time elapsed on that branch from its length.

This produces an ultrametric tree with relative times for nodes 4 (t_4) and node 5 (t_5):

$$t_4 = (b_1 + b_2)(b_1 + b_2 + 2b_3 + 2b_4)/4(b_1 + b_2 + 2b_4), \quad [10]$$

$$t_5 = (b_1 + b_2 + 2b_3 + 2b_4)/4, \quad [11]$$

The above equations ([1] - [11]) establish the relative rate framework (RRF) for the RelTime approach.

Theoretical considerations with four sequences and an outgroup

Next we consider the case of four ingroup taxa (1-4) and an outgroup (**Fig. 2b**). Here, we need to estimate six evolutionary rates ($r_1 - r_6$) using branch lengths ($b_1 - b_6$) for the given topological configuration. Following the case of 3-taxa above, we can write a set of equations:

$$r_1/r_2 = b_1/b_2, \quad [12]$$

$$r_3/r_4 = b_3/b_4, \quad [13]$$

$$r_5 = ½(r_1 + r_2), \quad [14]$$

$$r_6 = ½(r_3 + r_4), \quad [15]$$

$$r_5/r_6 = \ell_{12}/\ell_{34}, \quad [16]$$

$$r_7 = ½ (r_5 + r_6), \text{ and} \quad [17]$$

$$r_7 = 1 \quad [18]$$

Here, equation [14] would lead to a preference of minimum change in rates between the ancestral branch and its immediate descendant clades, which is also the case for equations [15], and [17]. However, r_i's are not required to be equal and, thus, no molecular clock is assumed.

Analytical estimates of relative rates and divergence times derived from the equations above are:

$$r_1 = 4b_1(b_1 + b_2 + 2b_5)/(b_1 + b_2)(b_1 + b_2 + b_3 + b_4 + 2b_5 + 2b_6), \quad [19]$$

$$r_2 = 4b_2(b_1 + b_2 + 2b_5)/(b_1 + b_2)(b_1 + b_2 + b_3 + b_4 + 2b_5 + 2b_6), \quad [20]$$

$$r_3 = 4b_3(b_3 + b_4 + 2b_6)/(b_1 + b_2)(b_1 + b_2 + b_3 + b_4 + 2b_5 + 2b_6), \quad [21]$$

$$r_4 = 4b_4(b_3 + b_4 + 2b_6)/(b_1 + b_2)(b_1 + b_2 + b_3 + b_4 + 2b_5 + 2b_6), \quad [22]$$

$$r_5 = 2(b_1 + b_2 + 2b_5)/(b_1 + b_2 + b_3 + b_4 + 2b_5 + 2b_6) \text{ and} \quad [23]$$

$$r_6 = 2(b_3 + b_4 + 2b_6)/(b_1 + b_2 + b_3 + b_4 + 2b_5 + 2b_6). \quad [24]$$

Using these equations for relative rates, we can derive the following equations to estimate relative times t_5, t_6, t_7 for nodes 5, 6, and 7, respectively:

$$t_5 = (b_1 + b_2)(b_1 + b_2 + b_3 + b_4 + 2b_5 + 2b_6)/4(b_1 + b_2 + 2b_5), \quad [25]$$

$$t_6 = (b_3 + b_4)(b_1 + b_2 + b_3 + b_4 + 2b_5 + 2b_6)/4(b_3 + b_4 + 2b_6) \text{ and} \quad [26]$$

$$t_7 = (b_1 + b_2 + b_3 + b_4 + 2b_5 + 2b_6)/4. \quad [27]$$

Note that a similar analysis can be carried out for the alternative unlabeled topological configuration of four-ingroup taxa.

Relative rates framework for a general case

Now we consider a general case of a phylogeny with more than four taxa. In this case, two factors eliminate the need to derive additional equations. First, in any phylogeny, an ingroup clade has either a 3- or a 4-clade configuration (marked by a star) with an immediate outgroup clade (**Fig. 2c** and **2d**, respectively). We apply equations [6] – [9] for the 3-clade case and equations [19] – [24] for the 4-clade case to compute relative rates for local branches in individual configurations. In the final step, we start recursively from the branch tips and scale node rates by multiplying them by their ancestral node rate, which is needed to generate the final relative rates for all the branches in the ingroup clade (Tamura et al. 2012). The use of recursive computation is analogous to the way of calculating likelihood value for a tree in the maximum likelihood method (Felsenstein 1981; Felsenstein 2004).

Relative rate framework with geometric means

In the original RelTime algorithm (Tamura et al. 2012) and the mathematical formulations above, we considered arithmetic mean when averaging branch lengths to minimize evolutionary rate changes. (Note that this is not an equal rate assumption). We have also developed RRF in which the geometric mean is used to better balance the rate changes between two descendant lineages. For example, if b_1 = 1 and b_2 = 4 in **Fig. 2a**, then the arithmetic mean will give t_4 = 2.5. Thus, the evolutionary rate r_1 is 2.5 times slower and r_2 is 1.6 times faster as compared to their ancestral lineages. The difference in rate change (2.5 and 1.6 in the present case) becomes larger as the difference between b_1 and b_2 becomes larger. In contrast, the geometric mean would give t_4 = 2.0, which results in a two-times slower rate in b_1 and a two-times faster rate in b_2, as compared to the ancestral lineage. That is, the difference from ancestor to descendant taxa is always balanced between sister lineages. The analytical solution of t_4 and t_5 as well as r_1, r_2, r_3 and r_4 is given by the following set of equations.

$$r_1 = \sqrt{b_1}\sqrt{\sqrt{b_1 b_2} + b_4}/\sqrt{b_2 b_3}\ , \quad [28]$$

$$r_2 = \sqrt{b_2}\sqrt{\sqrt{b_1 b_2} + b_4}/\sqrt{b_1 b_3}, \quad [29]$$

$$r_3 = \sqrt{b_3}/\sqrt{\sqrt{b_1 b_2} + b_4}, \text{ and} \quad [30]$$

$$r_4 = \sqrt{\sqrt{b_1 b_2} + b_4}/\sqrt{b_3}. \quad [31]$$

$$t_4 = \sqrt{b_1 b_2 b_3}/\sqrt{\sqrt{b_1 b_2} + b_4}, \quad [32]$$

$$t_5 = \sqrt{b_3}\sqrt{\sqrt{b_1 b_2} + b_4}, \quad [33]$$

For the 4-taxon case in **Fig. 2b**, the equations are as follows:

$$t_7 = \sqrt{\left(\sqrt{b_1 b_2} + b_5\right)\left(\sqrt{b_3 b_4} + b_6\right)}\ , \quad [34]$$

$$r_1 = \sqrt{b_1}\sqrt{\sqrt{b_1 b_2} + b_5}/\sqrt{b_2}\sqrt{\sqrt{b_3 b_4} + b_6}\ , \quad [35]$$

$$r_2 = \sqrt{b_2}\sqrt{\sqrt{b_1 b_2} + b_5}/\sqrt{b_1}\sqrt{\sqrt{b_3 b_4} + b_6}\ , \quad [36]$$

$$r_3 = \sqrt{b_3}\sqrt{\sqrt{b_3 b_4} + b_6}/\sqrt{b_4}\sqrt{\sqrt{b_1 b_2} + b_5}\ , \quad [37]$$

$$r_4 = \sqrt{b_4}\sqrt{\sqrt{b_3 b_4} + b_6}/\sqrt{b_3}\sqrt{\sqrt{b_1 b_2} + b_5}\ , \quad [38]$$

$$r_5 = \sqrt{\sqrt{b_1 b_2} + b_5}/\sqrt{\sqrt{b_3 b_4} + b_6} \quad \text{and} \quad [39]$$

$$r_6 = \sqrt{\sqrt{b_3 b_4} + b_6}/\sqrt{\sqrt{b_1 b_2} + b_5}\ . \quad [40]$$

$$t_5 = \sqrt{b_1 b_2}\sqrt{\sqrt{b_3 b_4} + b_6}/\sqrt{\sqrt{b_1 b_2} + b_5}\ , \quad [41]$$

$$t_6 = \sqrt{b_3 b_4}\sqrt{\sqrt{b_1 b_2} + b_5}/\sqrt{\sqrt{b_3 b_4} + b_6}\ , \quad [42]$$

RESULTS

We first tested the RRF by conducting a simulation to generate 3-ingroup and 4-ingroup sequence datasets, where the sequences were evolved according to independent rate model (Drummond and Rambaut (2007). Relative rates estimated using RRF were similar to the simulated relative rates (true rates) (**Fig. 3**). Datasets with autocorrelated rates showed similar results. For comparison, we carried out Bayesian analyses using all the correct priors (based on simulation parameters) in the MCMCTree software (Yang 2007). Bayesian analyses also

produced excellent results when compared to the true values, and results from RRF and Bayesian analyses were generally similar. **Fig. 4** shows comparisons of estimates of time from RRF with true times and with Bayesian estimates. It is clear from these comparisons that RRF works well for 3- and 4-ingroup sequence datasets, because the correlation among RRF estimates, Bayesian estimates, and true parameters is very high. We next analyzed much larger datasets, all of which contained 100 ingroup sequences and were evolved over a range of randomly-selected rate variation parameters. Rate estimates from RRF were highly correlated with the true rates (**Fig. 5a**), and time estimates showed much better correspondence between the estimated and true values because the slope is close to 1.0 (**Fig. 5c**). Bayesian analyses produced results similar to RRF (**Fig. 5b**), but RRF computation was more than 1,000 times faster, on average, than the Bayesian computation for these datasets (**Fig. 1b**).

We examined why rate estimates sometimes show a more dispersed relationship (lower R^2) with the true values, and found that many large differences between estimated and true rates occurred when branch lengths were short. In fact, the correlation between estimated and true rates increased when we excluded short branches (length < 0.02; **Fig. 5a** dotted line). This result is expected, because rates are estimated with a large variance when there is only a small amount of evolutionary change on a branch, which is usually attributable to short time elapsed between the divergence events. Therefore, it remains difficult to reliably estimate evolutionary rates on short branches. However, we found that the mean and standard deviations of the distribution of estimated rates were similar to the simulated values (**Fig. 6**), and they were also similar to those produced by Bayesian methods when correct priors were used. These results are highly encouraging and suggest that rate distributions produced using RRF are likely to be suitable as informative priors for use in an Empirical Bayesian framework (Carlin and Louis 2000; Yang 2014).

The divergence time estimates showed an excellent linear relationship with the true times for both large (**Fig. 5c**) and small (**Fig. 4**) datasets, because the evolutionary rates are estimated for individual branches, whereas the divergence times span multiple lineages. Also, the estimation of divergence time does not require us to compute individual branch rates first in RRF, e.g., equations [10] – [11] and [25] – [27] are used to estimate times directly. Therefore, the diffused relationship between estimated and true rates does not significantly impact the estimation of times for RelTime and Bayesian methods.

DISCUSSION

We have presented a mathematical framework underlying the RelTime method, which scales well with increasing numbers of sequences and is much faster than Bayesian methods (**Fig. 1**). This increase in computational speed is due to the innovation that RelTime uses all the data first to map a large alignment onto a phylogeny, and then it uses the resulting branch lengths to generate relative divergence times and evolutionary rates. Results from analyses of simulated and empirical data clearly show that this decomposition is effective, because RelTime produces estimates that are similar to known values and to Bayesian methods. One major advantage of RelTime methodology is in its computational efficiency, which makes it very useful in the analysis of large datasets.

The Relative Rate Framework for the RelTime method is exclusively focused on comparing evolutionary rates among lineages, and it avoids making assumptions about the statistical distribution of evolutionary rates in the whole phylogeny. For example, in **Fig. 7a**, the rate of evolution is higher in the lineage leading from node 4 to taxon 2 than to taxon 1 ($r_2 > r_1$), because b_2 is longer than b_1. The ratio of evolutionary rates at node 4 is $R_4 = b_1/b_2$ ($= r_1/r_2$), which does not depend on t_4. That is, we can estimate R_4 without knowing anything about the probability distribution of evolutionary rates throughout the tree. Similarly, the ratio of evolutionary rates between the two descendant lineages of node 5 (composite taxon [1,2] and taxon 3, respectively) is $R_5 = [(b_1 + b_2)/2 + b_4]/b_3$. Again, this ratio does not depend on knowledge of distribution of rates among branches, so R_4 and R_5 can be computed using the branch lengths only.

However, in order to estimate relative times t_4 and t_5, we need to know the relationship of evolutionary rates on r_{1+2} and r_4, where r_{1+2} is the overall evolutionary rate of the clade originating at node 4 and consisting of taxon 1 and 2 and r_4 is the relative evolutionary rate for the branch b_4 (see **Fig. 2a**). Without assuming a specific distribution of rates, r_{1+2}/r_4 cannot be determined uniquely and t_4 can be at any point between 0 and t_5. **Figure 7c** and **7d** represent two extreme possibilities. In one, if the clade rate (r_{1+2}) is much higher after the divergence event at node 4 ($r_{1+2} >> r_4$), then the estimate of t_4 will be small and the divergence event recent (**Fig. 7c**). Alternately, if the clade rate is much slower after the divergence event at node 4 ($r_{1+2} << r_4$), then t_4 will be much more ancient (**Fig. 7d**). RRF prefers relative rate estimates that infer the smallest change in rates between ancestor and descendent clades (e.g., crown branch b_4 and clade 4 consisting of taxon 1 and 2), as shown in the example timetree (**Fig. 7b**). This is the principle of

minimum rate change, which is achieved by using an iterative approach presented by Tamura et al. (2012). Note that the probabilities of occurrence of the extreme rate assignments shown in **Fig. 7c** and **Fig. 7d** are expected to be rather low in the commonly used distributions (e.g., lognormal distribution), so Bayesian methods will also favor the smallest rate change needed to explain the data. This is supported by the strong correlation between Bayesian and RRF estimates (**Figs. 3** and **4**).

Relationship of RRF with other approaches

RRF does not assume a specific model for rate variation from branch to branch, which makes it different from many parametric approaches (Thorne et al. 1998; Huelsenbeck et al. 2000; Kishino et al. 2001; Drummond and Suchard 2010). RRF is also different from non-parametric and semi-parametric approaches based on the idea of Sanderson (1997), because RRF does not attempt to estimate a universal penalty for how quickly rates change from branch to branch throughout the tree. The principle of minimizing local evolutionary rate changes makes RRF conceptually closer to the autocorrelated rate model of Thorne et al. (1998), but RRF does not impose uniform autocorrelation throughout the tree.

Because the RRF approach above does not assume a specific model for rate variation from branch to branch, we examined the performance of RelTime in an analysis of simulated data from Beaulieu et al. (2015). They simulated two lognormal distributions for an angiosperm phylogeny in which herbaceous clades exhibited higher and more variable evolutionary rates than woody clades. They reported that single-model Bayesian methods produced considerably more ancient date estimates for the divergence of herbaceous and woody clades (**Fig. 8a**). This overestimation of divergence time became more severe as the difference between the two rate models increased (**Fig. 8b**). Application of RelTime produced divergence time estimates that were much closer to true times (**Fig. 8c** and **8d**), which shows that RelTime can be useful in cases where the rate distribution differs among clades (Smith and Donoghue 2008; Dornburg et al. 2011; Beaulieu et al. 2015) or when clocks are local (Drummond and Suchard 2010; Crisp et al. 2014). As a further example, Tamura et al. (2012) found that RelTime produced accurate time estimates in simulations with a very large number of sequences when one clade possessed accelerated evolutionary rates, but penalized likelihood did not perform well. In general, we expect that the limitation of single-model Bayesian analyses will be overcome by local clock methods are available to deal with such scenarios (Drummond and Suchard 2010; Höhna et al. 2016; Lartillot et al. 2016), but the computational times needs of these approaches can be

prohibitive for even modest sized data.

RRF for non-molecular data

Even though the RRF has been initially developed where branch lengths were obtained by the Maximum Likelihood method for DNA and protein sequence data (Tamura et al. 2012), RRF method can be applied to any phylogeny where branch lengths reflect the amount of change. For example, RRF is directly applicable when branch lengths are estimated by using pairwise evolutionary distances and a least squares approach for a given tree topology (Rzhetsky and Nei 1993), and also when Maximum Parsimony estimates of branch lengths are generated via molecular or other data. such as gene expression patterns, morphological, developmental, or life history characters, e.g., King et al. (2016) and Cooney et al. (2017). Of course, the accuracy of the relative rate inferences made for such data depend directly on the accuracy of the phylogenetic tree used and whether the estimates of branch lengths are unbiased.

Usefulness of relative times

RelTime's accurate estimation of relative node ages without assumption of a speciation-model or calibration priors can benefit many applications. For example, relative node ages can be directly compared with time estimates based on fossil data. This allows evaluation of biological hypotheses without the circularity created by the current use of calibration priors and densities inferred from molecular data (Battistuzzi et al. 2015; Gold et al. 2017). Along these lines, the RelTime method has been used to develop a protocol to identify calibration priors that have the strongest influence on the final time estimates in Bayesian dating (Battistuzzi et al. 2015), because the cross-validation methods are unlikely to be effective (Warnock et al. 2012; Warnock et al. 2015). We expect RRF to complement existing Bayesian approaches, particularly as a means to generate informative priors for use in an Empirical Bayesian framework (Carlin and Louis 2000; Yang 2014). Even when applying Bayesian methods for dating divergences, the RelTime method would be useful to assess *a priori* the heterogeneity of evolutionary rates.

In conclusion, we have presented a mathematical foundation for the RelTime method and elucidated its relationship with other methods that do not assume a molecular clock. The relative rate framework produces excellent estimates of evolutionary rates and divergence times for molecular datasets in which sequences have evolved with and without autocorrelation.

MATERIALS AND METHODS

Computer simulations and analysis. We simulated 200 sequence datasets evolved under a factorial combination of evolutionary rates and topologies: 50 replicates each for two models of evolutionary rates (independent rates and autocorrelated rates among lineages) and two topologies (three- and four-ingroup taxa topologies shown in **Fig. 2a** and **2b**, respectively). The node height of the ingroup clade was set to be 10 time units, while the node heights of all subclades varied independently from 0 to 10 time units. For each resulting model timetree, branch-specific rates were sampled from (1) an uncorrelated lognormal distribution, where the mean rate was drawn randomly from an empirical distribution (Rosenberg and Kumar 2003) and the standard deviation varied from 0.25 to 0.75; and (2) an autocorrelated lognormal distribution, where the initial rate was drawn randomly from an empirical distribution (Rosenberg and Kumar 2003) and the autocorrelation parameter varied from 0.01 to 0.1. This rate sampling resulted in a phylogram used for generating sequence alignments in SeqGen (Grassly et al. 1997). We used the Hasegawa-Kinshino-Yano (HKY) model (Hasegawa et al. 1985) with 4 gamma categories and empirically-derived GC content and transition/transversion ratio (Rosenberg and Kumar 2003) to generate data for 3,000 sites.

Using the same simulation strategy, we created 35 alignments each under independent and autocorrelated rate scenarios following a master phylogeny of 100 taxa that was sampled from the bony-vertebrate clade in the Timetree of Life (Hedges and Kumar 2009). In the independent rate case, the standard deviation varied from 0.3 to 0.5. In the autocorrelated rate case, the autocorrelation parameter varied from 0.01 to 0.04. All other simulation parameters (GC contents, transition/ transversion ratio and sequence length) were derived from empirical distributions (Rosenberg and Kumar 2003).

All simulated data were analyzed in MCMCTree (Yang 2007) using correct priors; two independent runs of 5,000,000 generations were carried out. Results were checked in Tracer (Rambaut et al. 2014) for convergence. ESS values were higher than 200 after removing 10% burn-in samples in each run. One root calibration (true age ± 0.1 time unit) was used in the MCMCTree analyses. All RelTime analyses were conducted within the MEGA software (Kumar et al. 2012; Kumar et al. 2016) using the correct substitution model and topology.

Relative rate analysis. For any branch, the evolutionary rate was obtained by dividing the estimate of branch length by the inferred time elapsed on that branch in the tree. The same procedure was used for RelTime and Bayesian methods. True rates for simulated data were

calculated by dividing the branch length realized during sequence evolution by actual time elapsed on each branch. These are the most accurate rates that any method can estimate. Inferred rates were calculated using branch lengths estimated from maximum likelihood framework divided by inferred time estimated by RelTime or Bayesian methods.

Analysis of hybrid rate models

Simulated datasets and BEAST results were provided by Beaulieu et al. (Beaulieu et al. 2015) and retrieved from the Dryad Repository. All outgroup and root calibrations were automatically disregarded in RelTime because the assumption of equal rates of evolution between the ingroup and outgroup sequences is not testable in any method (Kumar et al. 2016). Lognormal distributions with fixed median values of "true ages" were used as calibration densities in the original study (Beaulieu et al. 2015). Because RelTime doesn't require specific density distributions for calibrations, we used a 10 million years wide spectrum with mean values of "true ages" for all 15 ingroup calibrated nodes in the re-analysis in order to directly compare their time estimates with those from RelTime. These distributions had boundaries similar to 99% probability densities of lognormal distributions originally employed as calibrations. The estimates of angiosperm age were obtained by summarizing estimates of 100 datasets in 3x and 6x rate simulated datasets.

ACKNOWLEDGEMENTS

We thank Jeremy Beaulieu for providing simulated data and Bayesian results. We are thankful to Drs. Heather Rowe for editorial comments and Beatriz Mello for helpful discussions. This research was supported by grants from NSF (DBI 1356548), National Aeronautics and Space Administration (NASA, NNX16AJ30G), and Tokyo Metropolitan University (DB105).

Figure Legends

Figure 1. (**a**) Computational time taken by RelTime and a Bayesian method for datasets containing increasing number of sequences (*n*). Sequence alignment consisted of 4,493 sites in which sequences were evolved with extensive rate variation (RR50 data from Tamura *et al*. (2012)). RelTime speed advantage increases with data volume by $O(n^2)$. (**b**) Calculation speed difference between RelTime and MCMCTree for 70 datasets of 100 sequences (see **Methods** for details).

Figure 2. The Relative Rate Framework for RelTime method. (**a**) A tree containing 3 ingroup sequences with an outgroup. Branch lengths are b_i's and branch rates are r_i's. Clade rates are shown by c_i's, and ℓ_{12} and r_{12} represent evolutionary rates on lineage emanating from node 5 and ending in sequences 1 and 2. Here, $\ell_{12} = b_4 + ½(b_1 + b_2)$. (**b**) The case of 4 ingroup sequences with an outgroup. Here, $\ell_{12} = b_4 + ½(b_1 + b_2)$. $\ell_{34} = b_6 + ½(b_3 + b_4)$. When applying the analytical solution to a larger phylogeny, all nodes in the tree have either a (**c**) 3-clades (**d**) or 4-clades configuration.

Figure 3. Comparison of RRF estimates of relative rates with true rates and with rates produced by MCMCTree Bayesian analyses. Results are from the analysis of datasets with three- or four-ingroup sequences, where sequence evolution was simulated under independent rate or autocorrelated rate models. Each panel contains results from 50 simulated datasets. All rates and divergence time estimates were normalized to allow direct comparison between true and estimated values. Slope through the origin and correlation coefficient (r^2) are shown for each panel.

Figure 4. Comparison of RRF estimates of relative divergence times with true times and with dates estimated via MCMCTree Bayesian analyses. Results are from the analysis of datasets with three- or four-ingroup sequences, where sequence evolution was simulated under independent rate or autocorrelated rate models. Each panel contains results from 50 simulated datasets. All rates and divergence time estimates were normalized to enable direct comparison between true and estimated values. Slope through the origin and correlation coefficient (r^2) are shown for each panel.

Figure 5. Performance of RRF in the analysis of datasets with 100-ingroup sequences and an outgroup. (**a**) Fraction of datasets for which inferred rates are correlated with true rates at different levels of correlation. Solid lines show result from all the rates, and the dotted lines show

results where rates for very short branches (< 0.02 substitutions per site) were excluded from correlation analysis. Results for datasets evolved with autocorrelated rates (blue) and with independent rates (red) are shown. (**b**) Relationship of RRF and Bayesian estimates of rates. Each circle represents correlation between the estimated and the true rate for one dataset. Results for datasets evolved with autocorrelated rates (blue) and with independent rates (red) are shown. (c) Distribution of the linear regression slopes of RRF estimates and true times for different datasets. Regression slopes were through the origin. Results for datasets evolved with autocorrelated rates (blue) and with independent rates (red) are shown. All the results are based on the analysis of 35 datasets that were evolved with autocorrelated rates and another 35 datasets that were evolved with independent rates.

Figure 6. Mean and standard deviations of the distribution of estimated and true rates for 100-sequence datasets. Each panel contains results for datasets that were evolved with autocorrelated rates (blue) and those with independent rates (red). For one dataset, the standard deviation of true and estimated rates was very large (0.009), which is not shown in order to properly display rest of the results.

Figure 7. A phylogenetic tree of three taxa (1, 2 and 3). (**a**) original phylogenetic tree with the observed branch lengths (*b*'s), which need to be used to estimate node times (*t*'s) shown in panel **b**. Evolutionary trees if the rate for clade containing taxon 1 and 2 is much (**c**) higher or (**d**) slower than that of its ancestor.

Figure 8. (**a**) Hybrid distribution of rates for branches leading to woody taxa (brown) and herbaceous taxa (green), with the former evolving 3-times slower than the latter. (**b**) Bayesian estimates reported by Beaulieu et al. (2015) when the rate difference between clades was 3-times (3x, solid line) and 6-times (6x, dashed line), with the simulated age of 140 million years ago shown by a red line. RelTime estimates of angiosperm age for Beaulieu et al. (2015)'s alignments with (**c**) 3x rate difference and (**d**) 6x mean rate difference. The medians and standard deviations are shown. Beaulieu et al. (2015) simulated 100 replicates (1000 bases) under GTR model in each scenario. Bayesian analyses were conducted using a single uncorrelated lognormal rate prior in Beaulieu et al. (2015). The same alignments, topology and ingroup calibrations were used in RelTime analyses.

REFERENCES

Battistuzzi FU, Billing-Ross P, Murillo O, Filipski A, Kumar S. 2015. A protocol for diagnosing the effect of calibration priors on posterior time estimates: A case study for the Cambrian explosion of animal phyla. Mol. Biol. Evol. 32:1907–1912.

Beaulieu JM, O'Meara BC, Crane P, Donoghue MJ. 2015. Heterogeneous rates of molecular evolution and diversification could explain the Triassic age estimate for angiosperms. Syst. Biol. 64:869–878.

Bonaldo MC, Ribeiro IP, Lima NS, Dos Santos AAC, Menezes LSR, da Cruz SOD, de Mello IS, Furtado ND, de Moura EE, Damasceno L, et al. 2016. Isolation of Infective Zika Virus from Urine and Saliva of Patients in Brazil. PLoS Negl. Trop. Dis. 10:e0004816.

Bond JE, Garrison NL, Hamilton CA, Godwin RL, Hedin M, Agnarsson I. 2014. Phylogenomics resolves a spider backbone phylogeny and rejects a prevailing paradigm for orb web evolution. Curr. Biol. 24:1765–1771.

Carlin BP, Louis TA. 2000. Bayes and empirical Bayes methods for data analysis. Hertfordshire: Chapman and Hall

Cooney CR, Bright JA, Capp EJR, Chira AM, Hughes EC, Moody CJA, Nouri LO, Varley ZK, Thomas GH. 2017. Mega-evolutionary dynamics of the adaptive radiation of birds. Nature 542:344–347.

Crisp MD, Hardy NB, Cook LG. 2014. Clock model makes a large difference to age estimates of long-stemmed clades with no internal calibration: a test using Australian grasstrees. BMC Evol. Biol. 14:263–279.

Dornburg A, Brandley MC, McGowen MR, Near TJ. 2011. Relaxed clocks and inferences of heterogeneous patterns of nucleotide substitution and divergence time estimates across whales and dolphins (Mammalia: Cetacea). Mol. Biol. Evol. 29:228–243.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214–221.

Drummond AJ, Suchard MA. 2010. Bayesian random local clocks, or one rate to rule them all. BMC Biol. 8:114–125.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mil. Evol. 17:368–376.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates

Filipski A, Murillo O, Freydenzon A, Tamura K, Kumar S. 2014. Prospects for building large timetrees using molecular data with incomplete gene coverage among species. Mol. Biol. Evol. 31:2542–2550.

Gold DA, Caron A, Fournier GP, Summons RE. 2017. Paleoproterozoic sterol biosynthesis and the rise of oxygen. Nature 543:420–423.

Grassly NC, Adachi J, Rambaut A. 1997. Seq-Gen: an application for the Monte Carlo simulation of protein sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

Hedges SB, Kumar S. 2009. The Timetree of Life. New York: Oxford University Press

Ho SY, Duchêne S. 2014. Molecular-clock methods for estimating evolutionary rates and timescales. Mol. Ecol. 23:5947–5965.

Huelsenbeck JP, Larget B, Swofford D. 2000. A compound poisson process for relaxing the molecular clock. Genetics 154:1879–1892.

Höhna S, Landis MJ, Heath TA, Boussau B, Lartillot N, Moore BR, Huelsenbeck JP, Ronquist F. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Syst. Biol. 65:726–736.

King B, Qiao T, Lee MSY, Zhu M, Long JA. 2016. Bayesian Morphological Clock Methods Resurrect Placoderm Monophyly and Reveal Rapid Early Evolution in Jawed Vertebrates.

Syst. Biol. 66:499–516.

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, Hedges SB. 2016. Advances in time estimation methods for molecular data. Mol. Biol. Evol. 33:863–869.

Kumar S, Stecher G, Peterson D, Tamura K. 2012. MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. Bioinformatics 28:2685–2686.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33:1870–1874.

Lartillot N, Phillips MJ, Ronquist F. 2016. A mixed relaxed clock model. Phil. Trans. R. Soc. B 371:20150132.

Lozano-Fernandez J, Dos Reis M, Donoghue PCJ, Pisani D. 2017. RelTime rates collapses to a strict clock when estimating the timeline of animal diversification. Genome Biol. Evol.

Mahler DL, Ingram T, Revell LJ, Losos JB. 2013. Exceptional convergence on the macroevolutionary landscape in island lizard radiations. Science 341:292–295.

Mello B, Tao Q, Tamura K, Kumar S. 2017. Fast and Accurate Estimates of Divergence Times from Big Data. Mol. Biol. Evol. 34:45–50.

Rambaut A, Suchard M, Xie D, Drummond A. 2014. Tracer v1.6. Available from: http://beast.bio.ed.ac.uk/Tracer

Dos Reis M, Donoghue PC, Yang Z. 2016. Bayesian molecular clock dating of species divergences in the genomics era. Nat. Rev. Genet. 17:71–80.

Rosenberg MS, Kumar S. 2003. Heterogeneity of nucleotide frequencies among evolutionary lineages and phylogenetic inference. Mol. Biol. Evol. 20:610–621.

Rzhetsky A, Nei M. 1993. Theoretical foundation of the minimum-evolution method of phylogenetic inference. Mol. Biol. Evol. 10:1073–1095.

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

Smith SA, Donoghue MJ. 2008. Rates of molecular evolution are linked to life history in flowering plants. Science 322:86–89.

Tamura K, Battistuzzi FU, Billing-Ross P, Murillo O, Filipski A, Kumar S. 2012. Estimating divergence times in large molecular phylogenies. Proc. Natl. Acad. Sci. U.S.A. 109:19333–19338.

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, Parham JF, Joyce WG, Lyson TR, Donoghue PC. 2015. Calibration uncertainty in molecular dating analyses: there is no substitute for the prior evaluation of time priors. Proc. R. Soc. B 282:20141013.

Warnock RC, Yang Z, Donoghue PC. 2012. Exploring uncertainty in the calibration of the molecular clock. Biol. Lett. 8:156–159.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Yang Z. 2014. Molecular Evolution: A Statistical Approach. Oxford: Oxford University Press

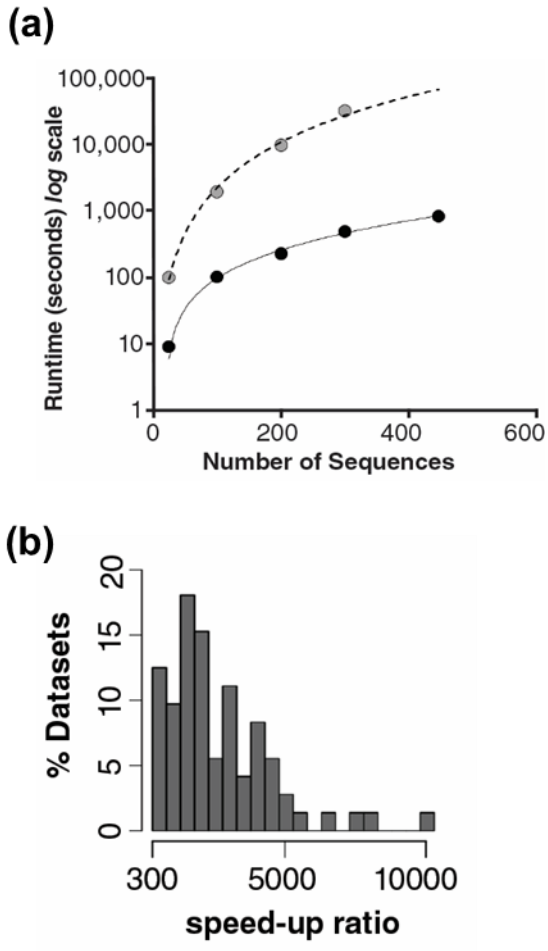

FIGURE 1

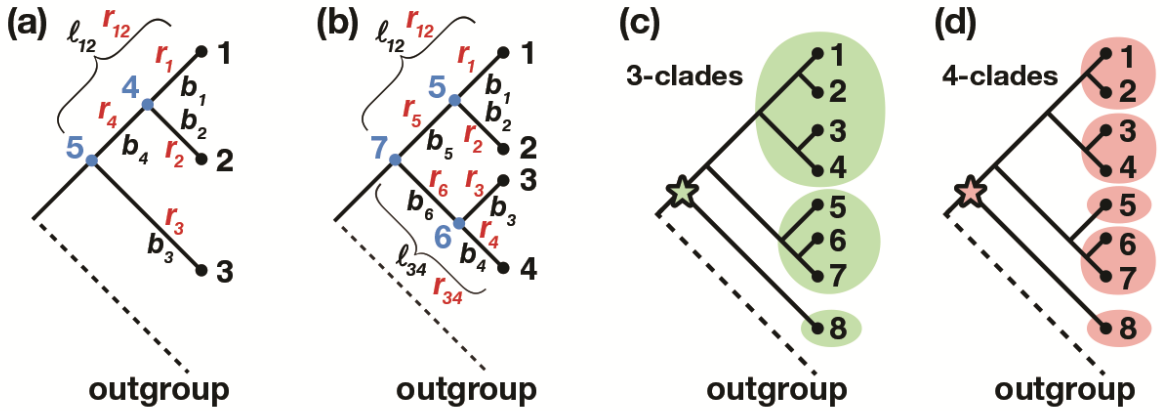

FIGURE 2

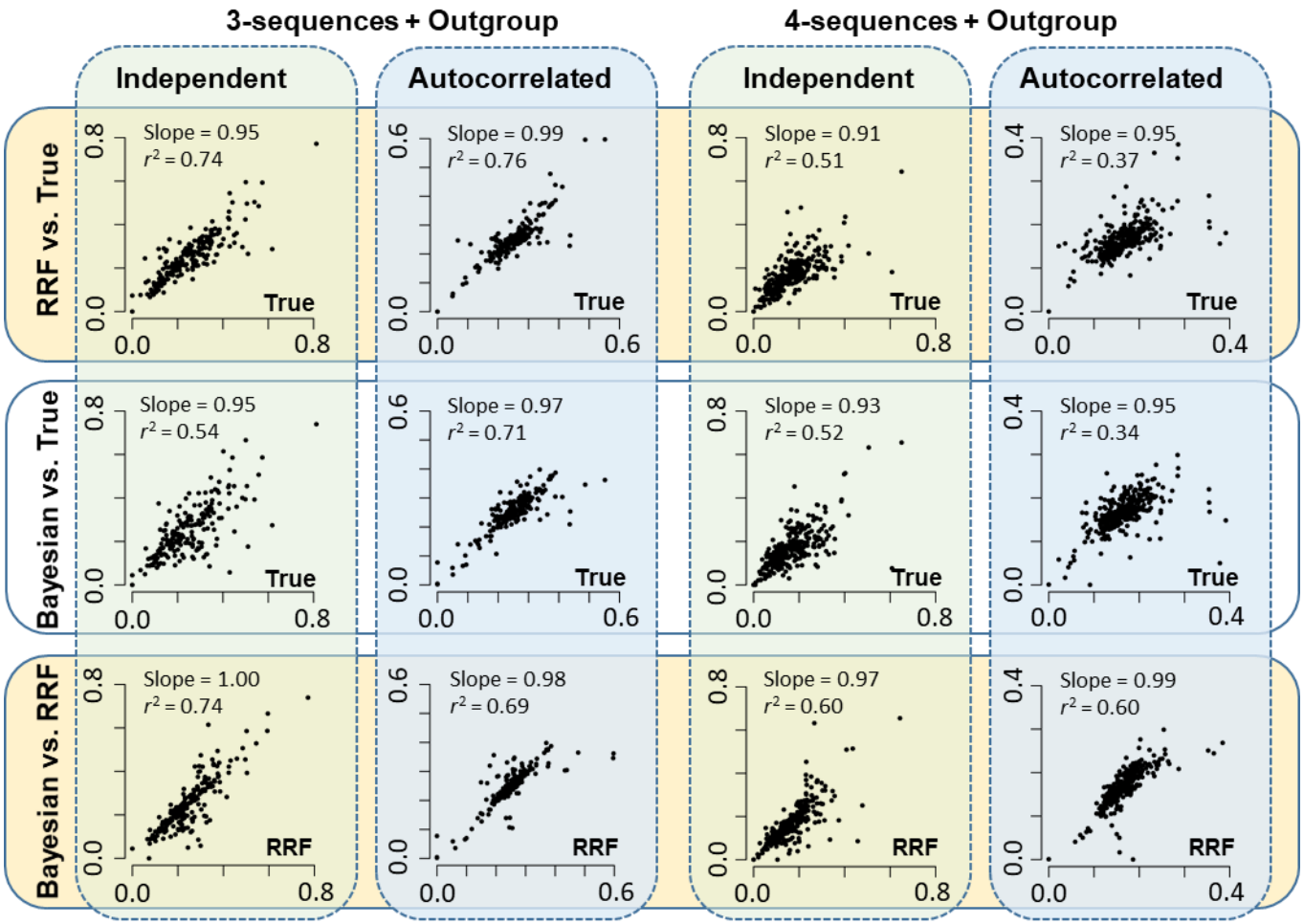

FIGURE 3

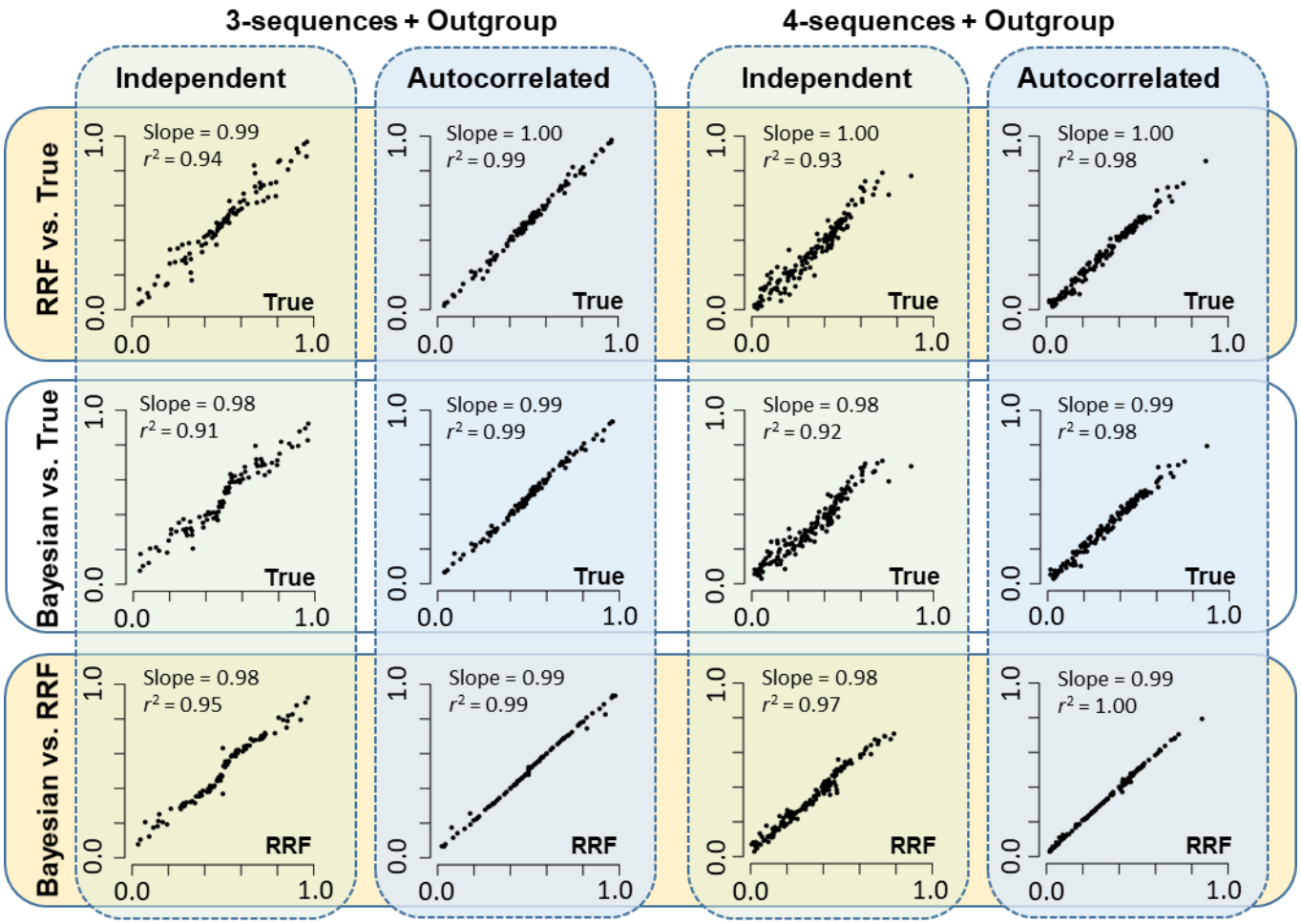

FIGURE 4

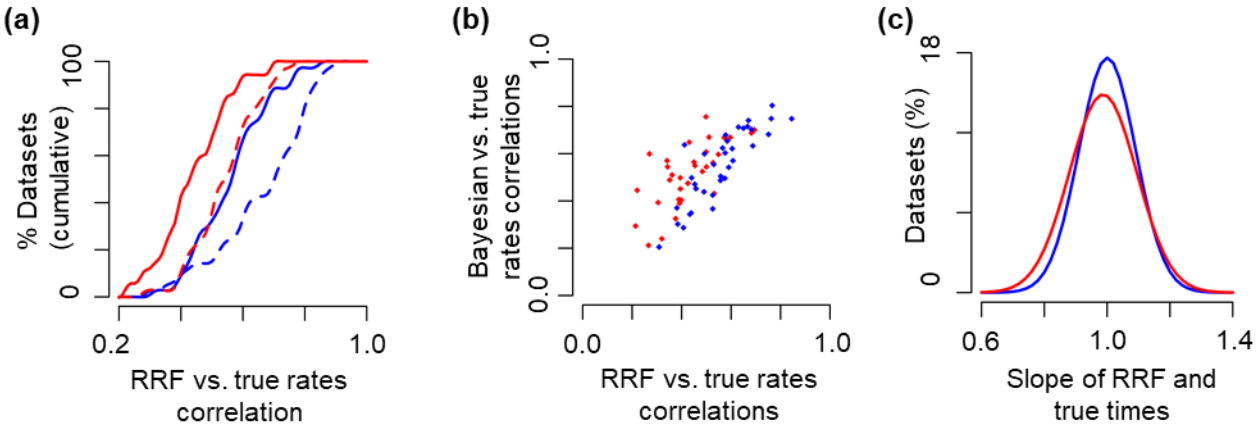

FIGURE 5

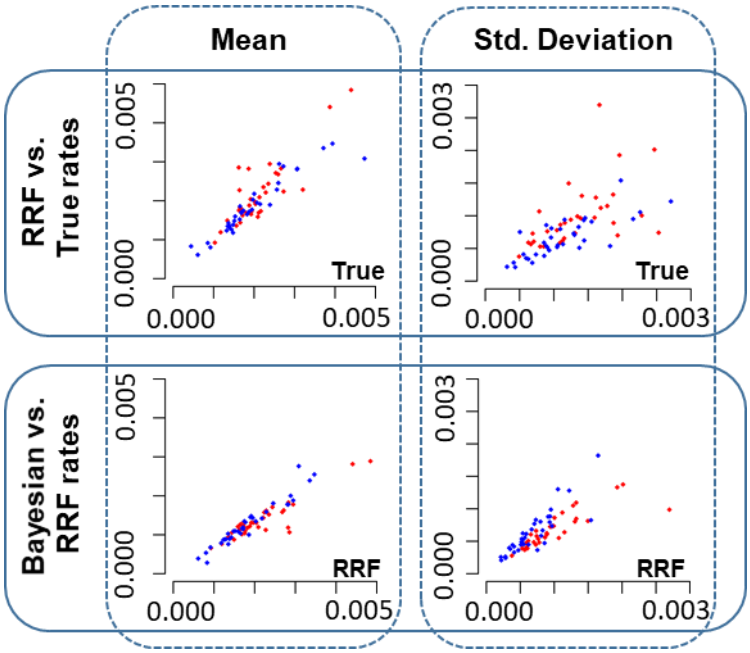

FIGURE 6

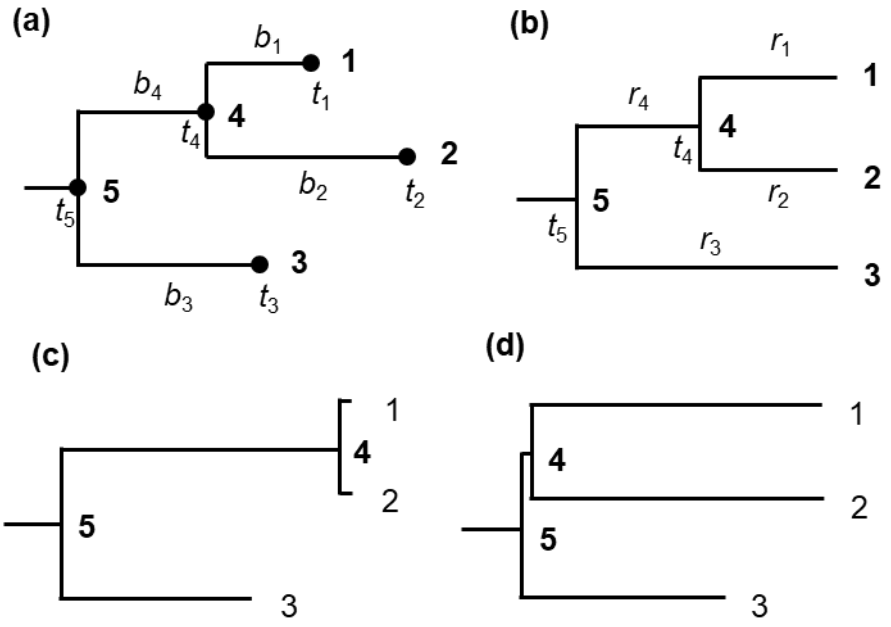

FIGURE 7

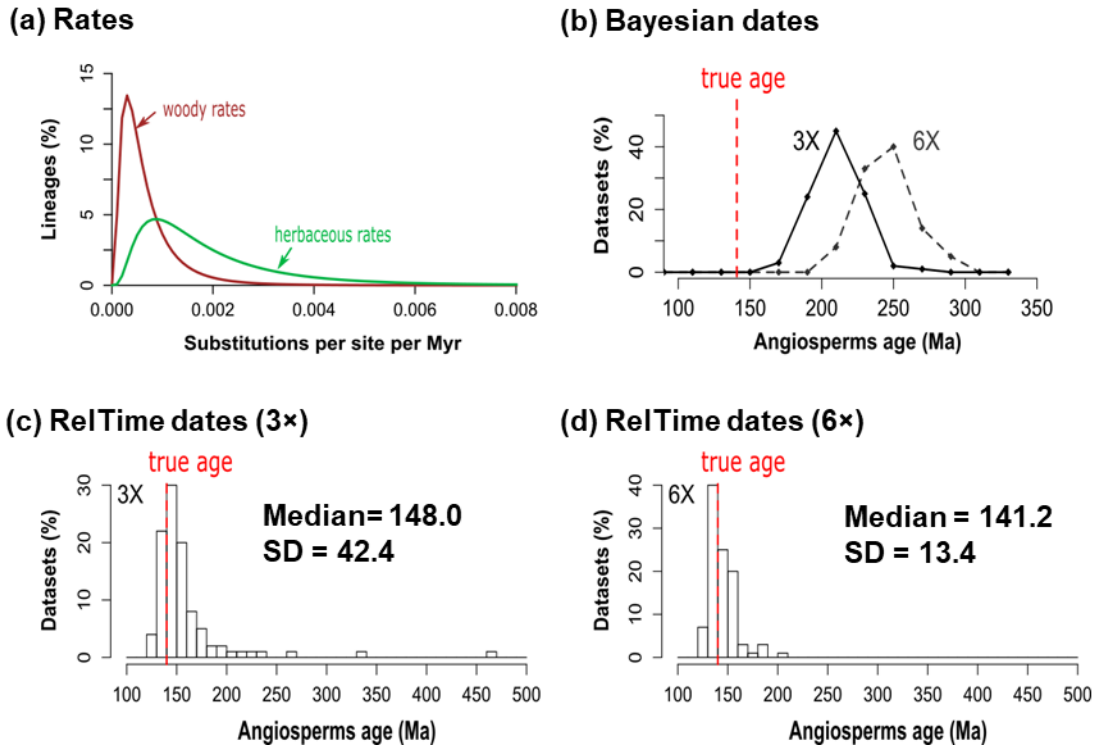

FIGURE 8