Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

*Original Paper*

# CorrTest: A new method for detecting correlation of evolutionary rates in a phylogenetic tree

Qiqing Tao[1,2], Koichiro Tamura[3,4], Fabia Battistuzzi[5], and Sudhir Kumar[1,2*]

[1]Institute for Genomics and Evolutionary Medicine, [2]Department of Biology, Temple University, Philadelphia, PA, USA. [3]Department of Biological Sciences, Tokyo Metropolitan University, Tokyo, Japan. [4]Research Center for Genomics and Bioinformatics, Tokyo Metropolitan University, Tokyo, Japan. [5]Department of Biological Sciences, Oakland University, Rochester, Michigan, USA.

*To whom correspondence should be addressed.

## Abstract

**Motivation**: Knowledge of the models of evolutionary rate variation in a phylogeny is of fundamental importance in molecular phylogenetics and systematics, not only to inform about the relationship among molecular, biological, and life history traits, but also for reliable estimation of divergence times among species and genes. Correlated and independent branch rates have emerged as two major competing models. The independent branch rate (IBR) model posits that evolutionary rates vary randomly throughout a phylogeny, in contrast to the alternative that these rates are correlated (CBR). However, currently available statistical tests lack sufficient power to reject the IBR model, which has caused many controversies because very different biological inferences are produced by the use of these models.

**Results**: We have developed a new method (CorrTest) to accurately detect the correlation of branch rates in large phylogenies. CorrTest is computationally efficient, and it performs better than the available state-of-the-art methods. CorrTest's application to multigene and genome-scale sequence alignments from mammals, birds, insects, metazoans, plants, fungi, and prokaryotes, suggests that DNA and amino acid sequence evolutionary rates are correlated throughout the tree of life. These findings suggest concordance between molecular and non-molecular evolutionary patterns and will foster unbiased and precise dating of the tree of life.

**Availability and Implementation:** The R source code of CorrTest is freely available for download at https://github.com/cathyqqtao/CorrTest.

**Contact:** s.kumar@temple.edu

**Supplementary information:** All empirical datasets, results, and source code for generating each figure are available at https://github.com/cathyqqtao/CorrTest. All simulated datasets are available on request.

## 1 Introduction

Phylogenomics has revolutionized our understanding of the patterns and timescale of the tree of life (Hedges et al., 2015; Marin et al., 2017). Genome-scale data has revealed that rates of molecular sequence change vary extensively among species (Kumar and Hedges, 2016; dos Reis et al., 2016; Ho and Duchêne, 2014). The causes and consequences of evolutionary rate variation are of fundamental importance in molecular phylogenetics and systematics (Lanfear et al., 2010; Lynch, 2010; Kimura, 1983), not only to inform about the relationship among molecular, biological, and life history traits, but also as a prerequisite for reliable estimation of divergence times among species and genes (Kumar and Hedges, 2016; Ho and Duchêne, 2014).

Three decades ago, Gillespie (1984) proposed that molecular evolutionary rates within a phylogeny will be correlated due to similarities in genomes, biology and environments between ancestral species and their immediate progeny. This idea led to statistical modelling of the variability of evolutionary rates among branches and formed the basis of the earliest relaxed clock methods for estimating divergence times without assuming a strict molecular clock (Kumar and Hedges, 2016; Kumar, 2005; Ho and Duchêne, 2014; Sanderson, 1997; Thorne et al., 1998). However, the independent branch rate (IBR) model has emerged as a strong alternative to the correlated branch rate (CBR) model. In the IBR model, rates vary randomly throughout the tree such that the evolutionary rate similarity between an ancestor and its descendant is, on average, no more than that between more distantly-related branches in a phylogeny (Drummond et al., 2006; Ho and Duchêne, 2014).

The IBR model is now widely used in estimating divergence times from molecular data for diverse groups of species, including mammals (Drummond et al., 2006), birds (Brown et al., 2008; Prum et al., 2015; Claramunt and Cracraft, 2015), amphibians (Feng et al., 2017), plants (Moore and Donoghue, 2007; Linder et al., 2011; Lu et al., 2014; Barreda et al., 2015; Smith et al., 2010; Bell et al., 2010; Barba-Montoya et al., 2018), and viruses (Drummond et al., 2006; Metsky et al., 2017; Buck et al., 2016). If the IBR model best explains the variability of evolutionary rates, then we must infer a decoupling of molecular and biological evolution, because morphology, behavior, and other life history traits are more similar between closely-related species (Sargis and Dagosto, 2008; Cox and Hautier, 2015; Lanfear et al., 2010) and are correlated with taxonomic or geographic distance (Wyles et al., 1983; Shao et al., 2016).

Alternatively, the widespread use of the IBR model (Drummond et al., 2006; Metsky et al., 2017; Brown et al., 2008; Prum et al., 2015; Claramunt and Cracraft, 2015; Linder et al., 2011; Bell et al., 2010; Smith et al., 2010; Lu et al., 2014; Moore and Donoghue, 2007; Feng et al., 2017; Buck et al., 2016) may be explained by the fact that the currently available statistical tests lack sufficient power to reject the IBR model (Ho et al., 2015). This may also explain why some studies report finding extensive branch rate correlation in many datasets (e.g., Lepage et al. (2007)), but others cannot confirm this using the same tests (e.g., Linder et al. (2011)). Consequently, many researchers use both CBR and IBR models for the same species groups (Erwin et al., 2011; dos Reis et al., 2015; Drummond et al., 2006; Meredith et al., 2011; dos Reis et al., 2012; Foster et al., 2016; Magallón et al., 2013; Bell et al., 2010; Wikström et al., 2001; Hertweck et al., 2015; Jarvis et al., 2014; Liu et al., 2017; dos Reis et al., 2018), a practice that often generates controversy via widely differing time estimates (Battistuzzi et al., 2010; dos Reis et al., 2014; Christin et al., 2014; Foster et al., 2016; dos Reis et al., 2015; Liu et al., 2017).

Therefore, a powerful method is needed to accurately test whether evolutionary rates are correlated among branches. Here, we introduce a new machine learning approach (CorrTest) with high power to detect correlation between molecular rates. CorrTest is computationally efficient, and its application to a large number of datasets enables an assessment of the presence of rate correlation in the tree of life.

In the following, we present a detailed description of CorrTest method and its performance on synthetic datasets, which is followed by a comparison with the Bayes factor method. We then present results from empirical analyses and discuss the pervasiveness of rate correlation throughout the tree of life.

## 2 Methods

### 2.1 The new CorrTest method

We employed a supervised machine learning (McL) framework (Bzdok et al., 2018) to build a predictive model to distinguish between CBR and IBR models. In our McL approach, the input is a molecular phylogeny with branch lengths (often derived from a multiple sequence alignment), and the output is a classification that corresponds to whether or not the evolutionary rates are correlated (CBR or IBR, respectively). We used a logistic regression to build a predictive model. An overview of our McL approach is presented in **Figure 1**.

To build a predictive model, we need measurable properties (features, **Fig. 1g** and **h**) that are derived from the input data. The output is ultimately the assignment of input data as most consistent with either CBR or IBR models. The selection of informative and discriminating features is critical for the success of McL. In CorrTest, we derive relative lineage rates using a given molecular phylogeny with branch ("edge") lengths (Tamura et al., 2018) (**Fig. 1e** and **f**) and use these lineage rates to generate informative features. One cannot use branch rates as features, because their computation requires the knowledge of node times in the phylogeny. In fact, IBR vs. CBR model selection is an early step in molecular dating by using Bayesian analyses. The use of relative lineage rates does not require the knowledge of divergence times, because an evolutionary lineage includes all the branches in the descendant subtree and the relative rate between lineages is simply the ratio of the evolutionary depths (sequence divergence) of the two descendants of a node (Tamura et al., 2018).

**Feature selection and acquisition.** We selected many possible features for use in our McL predictive model. These included, the correlation between ancestral and descendant lineage rates ($\rho_{ad}$), the correlation between the sister lineages ($\rho_s$), and the decay in $\rho_{ad}$ when one and two intervening branches are skipped ($d_1$ and $d_2$, respectively). For the given phylogeny, lineage-specific rate estimates ($r_i$'s) were obtained using equations [28] - [31] and [34] - [39] in Tamura et al. (2018). We then extracted the relative rates of ancestral clade ($r_a$) and two direct descendant clades ($r_1$ and $r_2$) of

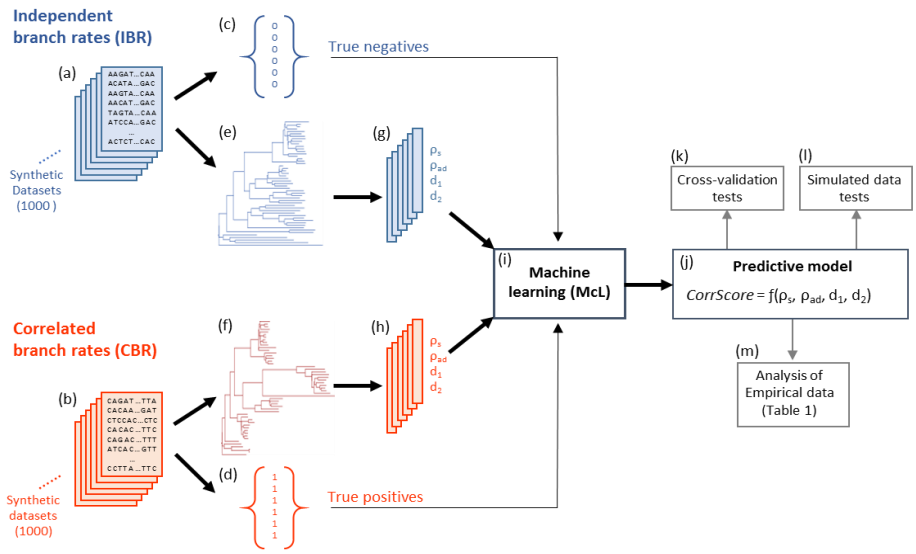


**Figure 1: A flowchart showing an overview of the machine learning (McL) approach applied to develop the predictive model (CorrTest).** We generated **(a)** 1,000 synthetic datasets that were evolved using an IBR model and **(b)** 1,000 synthetic datasets that were evolved using a CBR model. The numerical label (**c**) for all IBR datasets was 0 and (**d**) for all CBR datasets was 1. For each dataset, we estimated a molecular phylogeny with branch lengths (**e** and **f**) and computed $\rho_s$, $\rho_{ad}$, $d_1$, and $d_2$ (**g** and **h**) that served as features during the supervised machine learning. **(i)** Supervised machine learning was used to develop a predictive relationship between the input features and labels. **(j)** The predictive model produces a CorrScore for an input phylogeny with branch lengths. The predictive model was **(k)** validated with 10-fold and 2-fold cross-validation tests, **(l)** tested using external simulated data, and then **(m)** applied to real data to examine the prevalence of rate correlation in the tree of life.

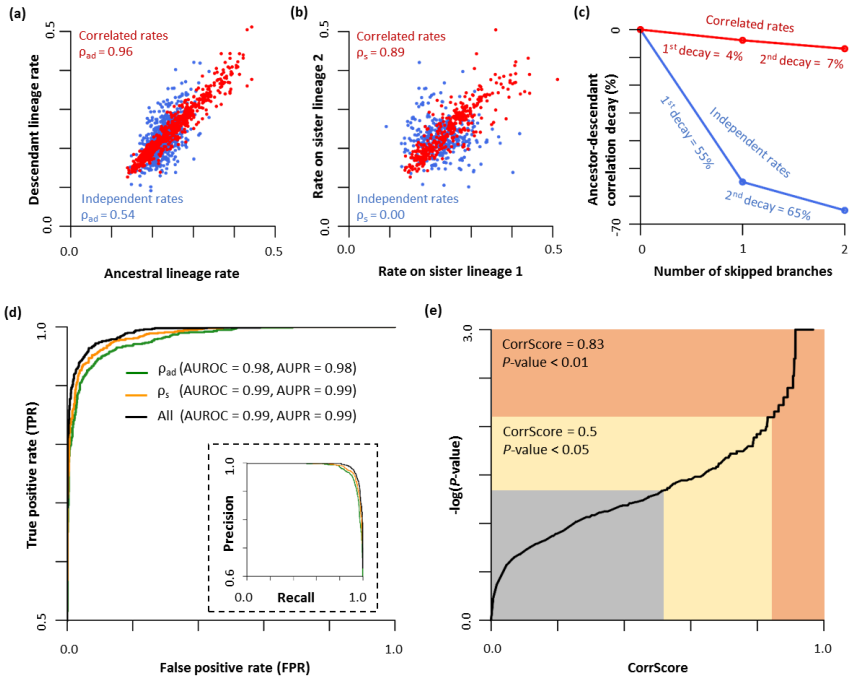


**Figure 2:** The relationship of **(a)** ancestral and direct descendent lineage rates and **(b)** sister lineage rates when the simulated evolutionary rates were correlated with each other (red) or varied independently (blue). The correlation coefficients are shown. **(c)** The decay of correlation between ancestral and descendant lineages when we skip one intervening branch (1st decay, $d_1$) and when we skip two intervening branches (2nd decay, $d_2$). Percent decay values are shown. **(d)** Receiver Operator Characteristic (ROC) and Precision Recall (PR) curves (inset) of the CorrTest for detecting branch rate model by using only ancestor-descendant lineage rates ($\rho_{ad}$, green), only sister lineage rates ($\rho_s$, orange), and all four features (all, black). The area under the curve is provided. **(e)** The relationship between the CorrScore produced by the machine learning model and the *P*-value. The null hypothesis of rate independence can be rejected when the CorrScore is greater than 0.83 at a significant level of $P < 0.01$, or when the CorrScore is greater than 0.5 at $P < 0.05$.

every node in the phylogeny. The correlation between ancestral lineage and its direct descendant lineage rate to obtain estimates of ancestor-descendant rate correlation ($\rho_{ad}$). To avoid the assumption of linear correlation between lineages, we used Spearman rank correlation because it can capture both linear and non-linear correlation between two vectors. We selected $\rho_{ad}$ as a feature because our analyses of simulated data showed that $\rho_{ad}$ was much higher for phylogenetic trees in which molecular sequences evolved under CBR model (0.96) than the IBR model (0.54, **Fig. 2a**). While "independent rates" should imply a lack of correlation, $\rho_{ad}$ is not zero for sequences evolved under the IBR model because the evolutionary rate of an ancestral lineage is necessarily related to the evolutionary rates of its descendant lineages (Tamura et al., 2018). While $\rho_{ad}$ is greater than zero, this feature shows distinct patterns for both CBR and IBR models and is thus a good candidate feature for McL.

As our second feature, we selected the correlation between the sister lineages ($\rho_s$), which is the Spearman rank correlation between $r_1$ and $r_2$ for all the nodes, because $\rho_s$ was higher for the CBR model (0.89) than the IBR model (0.00, **Fig. 2b**). Although our extensive simulations produced some scenarios in which $\rho_s$ was greater than 0.4 for datasets that evolved with the IBR model (because ancestral lineage rates include descendant evolutionary rates), $\rho_s$ was a highly discriminating feature for McL. For estimating $\rho_s$, we labeled sister pairs randomly, a strategy that has a very small impact on $\rho_s$ when the number of sequences in the phylogeny is not too small (>50). For smaller datasets, we found that it is best to generate multiple $\rho_s$ estimates, each using randomly labelled sister pairs, in order to eliminate any bias that may result from the arbitrary designation of sister pairs during the correlation process. In this case, we use the mean $\rho_s$ from multiple replicates in the CorrTest analysis.

Two additional features included in McL measure the decay in $\rho_{ad}$ when one and two intervening branches are skipped ($d_1$ and $d_2$), respectively, in $\rho_{ad}$ calculations. We first estimated $\rho_{ad\_skip1}$ as the correlation between rates where the ancestor and descendant were separated by one intervening branch, and $\rho_{ad\_skip2}$ as the correlation between rates where the ancestor and descendant were separated by two intervening branches. This skipping reduces ancestor-descendant correlation, which we then used to derive the decay of correlation values by using equations $d_1 = (\rho_{ad} - \rho_{ad\_skip1})/\rho_{ad}$ and $d_2 = (\rho_{ad} - \rho_{ad\_skip2})/\rho_{ad}$. We expect that $\rho_{ad}$ will decay slower under CBR than IBR, which was consistent with our observations (**Fig. 2c**). The inclusion of $d_1$ and $d_2$ improved the accuracy of our model slightly.

**Training dataset.** The selected set of candidate features ($\rho_s$, $\rho_{ad}$, $d_1$, and $d_2$) can be measured for any phylogeny with branch lengths (e.g., derived from multispecies sequence alignments) and used to train the machine learning classifier (**Fig. 1i**). For this purpose, we need a large set of phylogenies in which branch rates are correlated (CBR = 1, **Fig. 1d**) and phylogenies in which the branch rates are independent (IBR = 0, **Fig. 1c**). By using the four selected features for each phylogeny and the associated numerical output state (0 or 1), we built a logistic regression that serves as the predictive model (**Fig. 1j**). However, there is a paucity of empirical data for which CBR and IBR rates are firmly established. We therefore trained our McL model on a simulated dataset, a practice that is now widely used in applications when there is a paucity of reliable real world training datasets (Ekbatani et al., 2017; Le et al., 2017).

We used computer simulations to generate 1,000 phylogenies that evolved with CBR models and 1,000 phylogenies that evolved with IBR models (**Fig. 1a** and **b**). To ensure the general utility of our model for analyses of diverse data, we sampled phylogenies with varying numbers of species, degrees of rate correlation, and degrees of independent rate variation. Specifically, we simulated nucleotide alignments under IBR and CBR models using the NELSI package (Ho et al., 2015).

In IBR, branch-specific rates were drawn from a lognormal distribution with a mean gene rate and a standard deviation (in log-scale) that varied from 0.1 to 0.4, previously used in a study simulating independent rates with different levels of variation (Ho et al., 2015). In CBR, branch-specific rates were simulated under an autocorrelated process (Kishino et al., 2001) with an initial rate set as the mean rate derived from an empirical gene and an autocorrelated parameter, *v*, that was randomly chosen from 0.01 to 0.3, previously used in a study simulating low, moderate and high degrees of autocorrelated rates (Ho et al., 2015). We used SeqGen (Grassly et al., 1997) to generate alignments under Hasegawa-Kishino-Yano (HKY) model (Hasegawa et al., 1985) with 4 discrete gamma categories by using

a master phylogeny, consisting of 60-400 ingroup taxa randomly sampled from the bony-vertebrate clade in the Timetree of Life (Hedges and Kumar, 2009). Mean evolutionary rates, G+C contents, transition/transversion ratios and numbers of sites for simulation were derived from empirical distributions (Rosenberg and Kumar, 2003). These 2,000 simulated datasets were used as training data in building the machine learning model.

**Developing the predictive model.** We trained a logistic regression model using the skit-learn module (Pedregosa et al., 2011), which is a python toolbox for data mining and data analysis using machine learning algorithms, with only $\rho_{ad}$, only $\rho_s$ or all four features ($\rho_{ad}$, $\rho_s$, $d_1$ and $d_2$) using 2,000 simulated training datasets (1,000 with CBR model and 1,000 with IBR model). A response value of 1 was given to true positive cases (correlated rates) and 0 was assigned to true negative cases (independent rates). Thus, the prediction scores (CorrScore) were between 0 and 1. A high score representing a higher probability that the rates are correlated. Then the global thresholds at 5% and 1% significant levels can be determined.

**Cross-validation tests.** We performed two cross-validation tests (**Fig. 1k**). In 10-fold cross-validation, the predictive model was developed using 90% of the synthetic datasets, and then its performance was tested on the remaining 10% of the datasets. The AUROC was greater than 0.99 and the accuracy was high (>94%). Even in the 2-fold cross-validation, where only half of the datasets were used for training the model and the remaining half were used for testing, the AUROC was still greater than 0.99 with an accuracy greater than 92%. This indicates that the features we used in building the machine learning model are powerful and ensures high accuracy even when the training data are limited.

**Estimating CorrTest *P*-value.** We developed a conventional statistical test (CorrTest) based on CorrScore (**Fig. 2e**) in order to generate a *P*-value for researchers to use when deciding whether they should reject a null hypothesis that branch rates within a phylogeny are uncorrelated (independent). A high CorrScore translates into a higher probability that the branch rates are correlated. At a CorrScore greater than 0.5, the Type I error (rejecting the null hypothesis of IBR when it was true) was less than 5%. Type I error of 1% (*P*-value of 0.01) was achieved with a CorrScore greater than 0.83.

## 2.2 Empirical datasets

We applied CorrTest to 16 large datasets, which included nuclear, mitochondrial and plastid DNA, and protein sequences from mammals, birds, insects, metazoans, plants, fungi, and prokaryotes (**Table 1**). These data were selected because they did not contain too much missing data (<50%) and represented >80 sequences, as a large amount of missing data (>50%) can result in unreliable estimates of branch lengths and other phylogenetic errors (Filipski et al., 2014; Xi et al., 2015; Lemmon et al., 2009; Wiens and Moen, 2008; Marin and Hedges, 2018) and potentially result in a biased test of evolutionary rate correlation. When a phylogeny and branch lengths were available from the original study, we estimated relative rates directly using the phylogeny with branch lengths via the relative rate framework (Tamura et al., 2018) and computed selected features to conduct CorrTest. Otherwise, maximum likelihood estimates of branch lengths were obtained using the published phylogeny, sequence alignments, and the substitution model specified in the original article (Kumar et al., 2012, 2016).

## 2.3 Software for data analysis

**CorrTest analyses.** All the CorrTest analyses were conducted using a customized R code (available from https://github.com/cathyqqtao/CorrTest). We estimated branch lengths of a tree topology on sequence alignments using maximum likelihood method (or Neighbor-Joining method where we tested the robustness of our model to topological error) in MEGA (Kumar et al., 2012, 2016). Then we used those branch lengths to compute relative lineages rates (Tamura et al., 2018, 2012) and calculated the value of selected features ($\rho_{ad}$, $\rho_s$, $d_1$ and $d_2$) to obtain the CorrScore. We conducted CorrTest on the CorrScore to estimate the *P*-value of rejecting the null hypothesis (IBR). No calibration was needed for CorrTest analyses.

**Bayes factor analyses.** We used stepping-stone sampling (BF-SS) (Xie et al., 2011) with n = 20 and a = 5 using mcmc3r package (dos Reis et al., 2018). We chose BF-SS because the harmonic mean estimator has many statistical shortcomings (Xie et al., 2011; Baele et al., 2013; Lepage et al., 2007) and thermodynamic integration (dos Reis et al., 2018; Silvestro et

**Table 1.** Results from the CorrTest analyses of datasets from a diversity of species.

| Group | Data type | Taxa number[a] | Sequence length | Substitution model | CorrTest score | P-value | 1/v[b] | Reference |
|---|---|---|---|---|---|---|---|---|
| Mammals | Nuclear 4-fold degenerate sites | 138 | 1,671 | GTR + Γ | 0.98 | < 0.001 | 3.21 | Meredith et al. (2011) |
| Mammals | Nuclear 3rd codon | 138 | 11,010 | GTR + Γ | 0.99 | < 0.001 | 4.42 | Meredith et al. (2011) |
| Mammals | Nuclear proteins | 138 | 11,010 | JTT + Γ | 0.99 | < 0.001 | 3.11 | Meredith et al. (2011) |
| Mammals | Mitochondrial DNA | 271 | 7,370 | HKY + Γ | 0.98 | < 0.001 | 3.77 | dos Reis et al. (2012) |
| Birds | Nuclear DNA | 198 | 101,781 | GTR + Γ | 1.00 | < 0.001 | 2.07 | Prum et al. (2015) |
| Birds | Nuclear 3rd codon | 222 | 1,364 | GTR + Γ | 1.00 | < 0.001 | 2.11 | Claramunt et al. (2015) |
| Birds | Nuclear 1st and 2nd codon | 222 | 2,728 | GTR + Γ | 1.00 | < 0.001 | 2.53 | Claramunt er al. (2015) |
| Insects | Nuclear proteins | 143 | 220,091 | LG + Γ | 1.00 | < 0.001 | 8.68 | Misof et al. (2014) |
| Metazoans | Mitochondrial & nuclear proteins | 113 | 2,049 | LG + Γ | 0.65 | < 0.05 | 40.0 | Erwin et al. (2011) |
| Plants | Plastid 3rd codon | 335 | 19,449 | GTR + Γ | 1.00 | < 0.001 | 2.28 | Ruhfel et al. (2014) |
| Plants | Plastid proteins | 335 | 19,449 | JTT + Γ | 1.00 | < 0.001 | 2.46 | Ruhfel et al. (2014) |
| Plants | Nuclear 1st and 2nd codon | 99 | 220,091 | GTR + Γ | 1.00 | < 0.001 | 5.50 | Wickett et al. (2014) |
| Plants | Chloroplast and nuclear DNA | 124 | 5,992 | GTR + Γ | 1.00 | < 0.001 | 2.64 | Beaulieu et al. (2015) |
| Fungi | Nuclear proteins | 85 | 609,772 | LG + Γ | 0.97 | < 0.001 | 3.78 | Shen et al. (2016) |
| Prokaryotes | Nuclear proteins | 197 | 6,884 | JTT + Γ | 0.79 | < 0.05 | 2.54 | Battistuzzi et al. (2009) |
| Prokaryotes | Nuclear proteins | 126 | 3,145 | JTT + Γ | 0.83 | < 0.05 | 1.23 | Calteau et al. (2014) |

[a]Taxa number is the number of ingroup taxa only.
[b]1/v is the inverse of the autocorrelation parameter that is estimated by MCMCTree with the autocorrelated rate model in the time unit of 100My.

al., 2011) is less efficient than BF-SS. Still, BF-SS requires a long computational time, we only finished analyses of 50% of synthetic datasets. For each dataset, we computed the log-likelihoods (*ln*K) of using IBR model and CBR model. The Bayes factor posterior probability for CBR was calculated as shown in dos Reis et al. (2018). We used only one calibration point at the root (true age with a narrow uniform distribution) in all the Bayesian analyses, as it is the minimum number of calibrations required by MCMCTree. For other priors, we used diffused distributions of "rgene_gamma = 1 1", "sigma2_gamma=1 1" and "BDparas = 1 1 0". In all Bayes factor analyses, two independent runs of 5,000,000 generations each were conducted, and results were checked in Tracer for convergence. ESS values were higher than 200 after removing 10% burn-in samples for each run.

## 3 Results and Discussion

### 3.1 CorrTest performs well on simulated datasets

We evaluated the sensitivity and specificity of our model using standard receiver operating characteristic (ROC) curves, which showed the performance of CorrTest to detect rate correlation when it is present (True Positive Rate, TPR) and when it was not present (False Positive Rate, FPR) at different CorrScore thresholds. The ROC curve for McL using all four features was slightly better than the use of only two features, which led to the inclusion of all four features in the predictive model (**Fig. 2d**). The area under the ROC (AUROC) was 99%, with a 95% TPR (i.e., CBR detection) achieved at the expense of only 5% FPR (**Fig. 2d**, black line). The area under the precision recall (AUPR) curve was also extremely high (0.99; **Fig. 2d** inset), which means that our predictive model detects correlation among branch rates with very high accuracy and precision.

In addition to the performance on the training dataset, we tested CorrTest on a large collection of simulated datasets from Tamura et al. (2012) in which different software and simulation schemes were used to generate a wide variety of large datasets (400 ingroup taxa). In these datasets, sequences evolved with different G+C contents, transition/transversion ratios, and evolutionary rates. CorrTest showed an accuracy greater than 94% in detecting rate autocorrelation for datasets that were simulated with low and high G+C contents (**Fig. 3a**), small and large substitution rate biases (**Fig. 3b**), and different levels of sequence conservation (**Fig. 3c**). As expected, CorrTest performed the best on datasets that contain more and longer sequences (**Fig. 3d**).

In the above analyses, we used the correct tree topology and nucleotide substitution model along with all the data. We relaxed this requirement and randomly sampled 50, 100, 200, and 300 sequences from the full datasets and conducted CorrTest by using phylogenies inferred using the Neighbor Joining method (Saitou and Nei, 1987) with an oversimplified substitution model (Kimura, 1980). Naturally, many inferred phylogenies contained topological errors, but we found that the accuracy of CorrTest is still high as long as the dataset contained >100 sequences of length >1,000 base pairs (**Fig. 3e**). CorrTest performed well even when 20% of the partitions were incorrect in the inferred phylogeny (**Fig. 3f**). Therefore, CorrTest will be most reliable for large datasets, but is relatively robust to errors in phylogenetic inference.

We also evaluated if higher accuracy could be achieved by building predictive models that were trained separately by using data with ≤ 100 (M100), 100 – 200 (M200), 200 – 300 (M300), and > 300 (M400) sequences. A specific threshold was determined for each training subset and then was tested using Tamura et al. (2012)'s data with the corresponding size. For example, we used the threshold determined by the model trained with small data (≤ 100 sequences) on the test data that contain less than 100 sequences, and used the threshold determined by the model trained with large data (>300 sequences) on the large test data (400 sequences). We found that the accuracy of using the specific thresholds (**Fig. S1a-c**) is similar to the accuracy when we used a global threshold (**Fig. 3d-f**). This is because the machine learning algorithm has automatically incorporated the impact of the number of sequences when it determined the relationship of four selected features ($\rho_{ad}$, $\rho_s$, $d_1$ and $d_2$). This suggests that our CorrTest model is appropriate for large and small datasets.

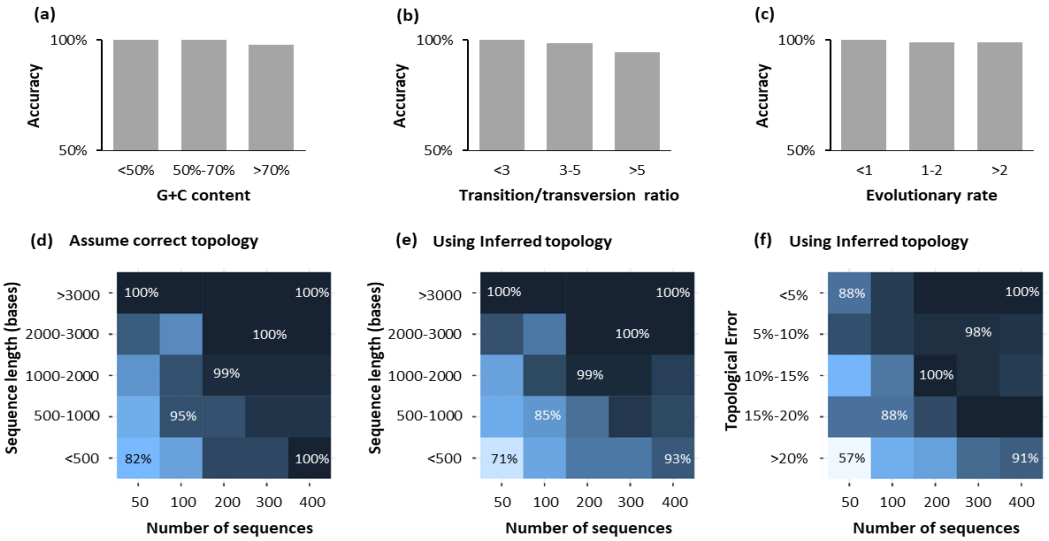


**Figure 3:** The performance of CorrTest in detecting rate correlation in the analysis of datasets (Tamura et al., 2012) that were simulated with different **(a)** G+C contents, **(b)** transition/transversion rate ratios, and **(c)** average molecular evolutionary rates. Darker color indicates higher accuracy. The evolutionary rates are in the units of $10^{-3}$ substitutions per site per million years. (**d – f**) Patterns of CorrTest accuracy for datasets containing increasing number of sequences. The accuracy of CorrTest for different sequence length is shown when (**d**) the correct topology was assumed and (**e**) the topology was inferred. (**f**) The accuracy of CorrTest for datasets in which the inferred the topology contained small and large number of topological errors.

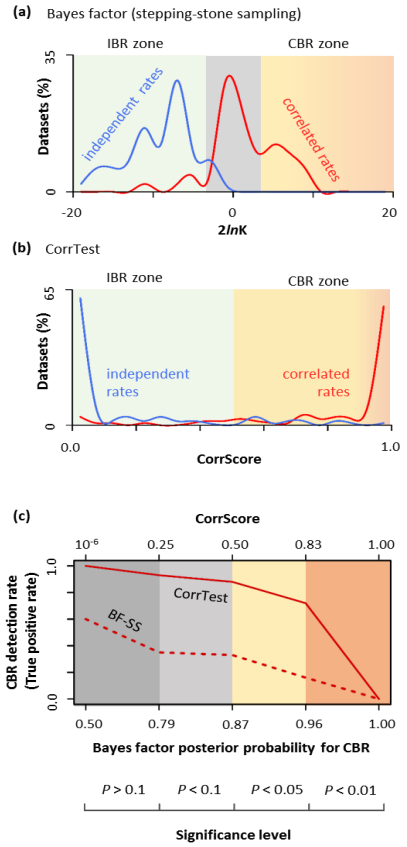


**Figure 4: Comparisons of the performance of CorrTest and Bayes Factor analyses. (a)** Distributions of 2 times the differences of marginal log-likelihood (2*ln*K) estimated via stepping-stone sampling method for datasets that were simulated with correlated branch rates (CBR, red) and independent branch rates (IBR, blue). CBR is preferred ($P < 0.05$) when 2lnK is greater than 3.841 (CBR zone), and IBR is preferred when 2lnK is less than -3.841 (IBR zone). When 2*ln*K is between -3.841 and 3.841, the fit of the two rate models is not significantly different (gray shade). **(b)** The distributions of CorrScores in analyses of CBR (red) and IBR (blue) datasets. Rates are predicted to be correlated if the CorrScore is greater than 0.5 ($P < 0.05$, CBR zone) and vary independently if the CorrScore is less than 0.5 (IBR zone). **(c)** The rate of detecting CBR model correctly (True Positive Rate) at different levels of statistical significance in Bayes factor (stepping-stone sampling) and CorrTest analyses. Posterior probabilities for CBR in BF-SS analysis are derived using the log-likelihood patterns in panel a. CorrTest *P*-values are derived using the CorrScore pattern in panel b.

## 3.2 CorrTest versus Bayes factor analysis

We compared the performance of CorrTest with that of the Bayes factor approach. Because the Bayes factor method is computationally demanding, we limited our comparison to 100 datasets containing 100 sequences each. For these simulations, a master phylogeny of 100 taxa was randomly sampled from the bony-vertebrate clade in the Timetree of Life (Hedges and Kumar, 2009). The computer simulations were conducted as generating the training data, and the 200 datasets produced were subject to CorrTest and Bayes factor analyses.

We computed Bayes factors (BF) by using the stepping-stone sampling (SS) method. BF-SS analysis detected autocorrelation ($P < 0.05$) for 32% of the datasets that actually evolved with correlated rates (**Fig. 4a**, red curve in the CBR zone). This is because the marginal log-likelihoods under the CBR model for 78% of these datasets were very similar to or lower than the IBR model. Therefore, BF was very conservative in rejecting the null hypothesis (see also in Ho et al. (2015)). In contrast, CorrTest correctly detected the CBR model for 88% of the datasets ($P < 0.05$; **Fig. 4b**, red curve in CBR zone). For datasets that evolved with IBR model, BF-SS correctly detected the IBR model for 92% (**Fig. 4a**, blue curves in the IBR zone), whereas CorrTest correctly detected 86% (**Fig. 4b**, blue curve in the IBR zone). Therefore, Bayes Factor analyses generally perform well in correctly classifying phylogenies evolved under IBR, but fail to detect the influence of CBR. The power of CorrTest to correctly infer CBR is responsible for its higher overall accuracy (87%, vs. 62% for BF). Such a difference in accuracy was observed at all levels of statistical significance (**Fig. 4c**). In the future, faster and more advanced BF implementations may allow extensive comparison of traditional Bayesian and CorrTest approaches, as the Bayesian approaches are still evolving (dos Reis et al., 2018) and currently require extensive computation time. Based on the limited comparisons presented here, we conclude that machine learning enables highly accurate detection of rate correlation in a given phylogeny and presents a computationally feasible alternative to Bayes Factor analyses for large datasets.

## 3.3 Correlated rates are common in molecular evolution

The high accuracy and fast computational speed of CorrTest enabled us to test the presence of autocorrelation in 16 large datasets from 12 published studies of eukaryotes and 2 published studies of prokaryotes encompassing diverse groups across the tree life (**Table 1**). These data were selected because they did not contain too much missing data (<50%). As we know, a large amount of missing data (>50%) can result in unreliable estimates of branch lengths and other phylogenetic errors (Filipski et al., 2014; Xi et al., 2015; Lemmon et al., 2009; Wiens and Moen, 2008; Marin and Hedges, 2018) and potentially bias CorrTest result. CorrTest rejected the IBR model for all datasets ($P < 0.05$). In these analyses, we assumed a time-reversible process for base substitution. However, the violation of this assumption may produce biased results in phylogenetic analysis (Jayaswal et al., 2014). We, therefore, applied an unrestricted substitution model for analyzing all the nuclear datasets and confirmed that CorrTest rejected the IBR model in every case ($P < 0.05$). This robustness stems from the fact that the branch lengths estimated under the time-reversible and the unrestricted model show an excellent linear relationship for these data ($r^2 > 0.99$). This is the reason why CorrTest produces reliable results even when an oversimplified model was used in computer simulations (**Fig. 3e** and **f**).

These results suggest that the correlation of rates among lineages is the rule, rather than the exception in molecular phylogenies. This pattern contrasts starkly with those reported in many previous studies (Linder et al., 2011; Brown et al., 2008; Drummond et al., 2006; Moore and Donoghue, 2007; Claramunt and Cracraft, 2015; Jarvis et al., 2014; Prum et al., 2015; Feng et al., 2017; Lu et al., 2014; Barreda et al., 2015; Barba-Montoya et al., 2018; Smith et al., 2010; Bell et al., 2010). In fact, all but three datasets (Erwin et al., 2011; Calteau et al., 2014; Battistuzzi and Hedges, 2009) received very high prediction scores in CorrTest, resulting in extremely significant *P*-values ($P < 0.001$). The IBR model was also rejected for the other three datasets ($P < 0.05$), but their test scores were not as high, likely because they sparsely sample a large phylogenetic space. For example, the metazoan dataset (Erwin et al., 2011) contains sequences primarily from highly divergent species that shared common ancestors hundreds of millions of years ago. In this case, tip lineages in the phylogeny are long and their evolutionary rates are influenced by many un-sampled lineages. Such sampling effects weaken the rate correlation signal. We verified this behavior via analyses of simulated data and found that CorrTest's prediction

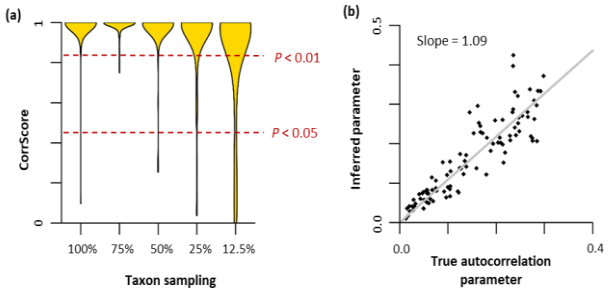


**Figure 5: (a)** The distribution of CorrScore when data have different taxon sampling densities. The CorrScore decreases when the density of taxon sampling is lower, as there is much less information to discriminate between CBR and IBR. Red, dashed lines mark two statistical significance levels of 5% and 1%. **(b)** The relationship between the inferred autocorrelation parameter from MCMCTree and the true value. The gray line represents the best-fit regression line, which has a slope of 1.09.

scores decreased when taxon sampling and density were lowered (**Fig. 5a**). Overall, CorrTest detected rate correlation in all the empirical datasets.

### 3.4 Magnitude of the rate correlation in molecular data

CorrScore is influenced by the size of the dataset in addition to the degree of correlation, so it is not a direct measure of the degree of rate correlation (effect size) in a phylogeny. Instead, one should use a Bayesian approach to estimate the degree of rate correlation, for example, under the Kishino et al.'s autocorrelated rate model (Kishino et al., 2001). In this model, a single parameter ($v$) captures the degree of autocorrelation among branches in a phylogenetic tree. MCMCTree (Yang, 2007) analyses of simulated datasets confirmed that the estimated $v$ is linearly related to the true value (**Fig. 5b**).

To obtain $v$ in empirical data, we used the same input priors as the original study and only one root calibration to avoid undue influence of calibration uncertainty densities on the estimate of $v$. We used the root calibration provided in the original article or selected the median age of the root node in the TimeTree database (Kumar et al., 2017; Hedges et al., 2006) ± 50My (soft uniform distribution) as the root calibration. Because Bayesian analyses require long computational times, we used either the original alignments or randomly selected 20,000 sites from the original alignments (if the alignments were longer than 20,000 sites) in MCMCTree analyses, except for Ruhfel et al. (2014). Ruhfel et al. (2014) contained more than 300 ingroup species, such that even alignments of 20,000 sites required prohibitive amounts of memory. In this case, we randomly selected 2,000 sites from the original alignments to estimate v (similar results were obtained with a different site subset). Two independent runs of 5,000,000 generations each were conducted, and results were checked in Tracer (Rambaut et al., 2018) for convergence. ESS values were higher than 200 after removing 10% burn-in samples for each run.

Because a low value of $v$ indicates high autocorrelation, we use the inverse of $v$ to represent the degree of rate autocorrelation. In empirical data analyses, we find that the inverse of $v$ is high for all datasets examined, which suggests ubiquitous high rate correlation across the tree of life. Many other interesting patterns emerge from this analysis. First, rate correlation is highly significant not only for mutational rates (= substitution rate at neutral positions), which are expected to be similar in sister species because they inherit cellular machinery from a common ancestor, but also amino acid substitution rates, which are more strongly influenced by natural selection (**Table 1**). For example, synonymous substitution rates in the third codon positions and the four-fold degenerate sites in mammals (Meredith et al., 2011), which are largely neutral and are the best reflection of mutation rates (Kumar and Subramanian, 2002), received high CorrScores of 0.99 and 0.98, respectively ($P < 0.001$). Second, our model also detected a strong signal of correlation for amino acid substitution rates in the same proteins (CorrScore = 0.99). Bayesian analyses showed that the degree of correlation is high in both cases: inverse of $v$ was 3.21 in 4-fold degenerate sites and 3.11 in amino acid sequences. Third, mutational and substitution rates in both nuclear and mitochondrial genomes are highly correlated (**Table 1**).

The above results establish that molecular and non-molecular evolutionary patterns are concordant, because morphological characteristics are also found to be similar between closely-related species (Sargis and Dagosto, 2008; Cox and Hautier, 2015; Lanfear et al., 2010) and correlated with taxonomic or geographic distance (Wyles et al., 1983; Shao et al., 2016). Therefore, we suggest the correlated rate model be the default in molecular dating analysis, and CorrTest can be used to test the independent rate model when sufficient numbers of sequences are available. Use of a correlated rate model is important because model selection has a strong influence on the posterior credible intervals of divergence times (Battistuzzi et al., 2010). For example, the use of IBR model produces estimates of divergence time of two major groups of grasses that are 66% older (Christin et al., 2014) and origin of a major group of mammal (Erinaceidea) to be 30% older (Meredith et al., 2011) than estimates under CBR model. In fact, substantial differences between node age estimates under IBR and CBR models have been reported in many studies (Battistuzzi et al., 2010; Christin et al., 2014; Foster et al., 2016; dos Reis et al., 2015; Bell et al., 2010; Liu et al., 2017). Thus, the use of an incorrect rate model has a large impact on time estimates, which may not be alleviated by adding calibrations (Battistuzzi et al., 2010). Knowledge that evolutionary rates are generally correlated within lineages will foster unbiased and precise dating of the tree of life.

## Conclusions

Excellent performance of CorrTest in detecting correlation among molecular evolutionary rates in a phylogeny is an early indication that the machine learning approaches will be useful in molecular phylogenetics, including model testing and pattern discovery. CorrTest is faster, scalable, and more accurate than the traditional Bayesian methods. CorrTest's application to a large number of biological datasets has successfully addressed an enduring question in evolutionary biology: are the molecular rates of change between species correlated or independent? Our data analysis suggests that the evolutionary rate correlation is universal, which will improve specification of correct rate models that are essential for molecular clock analyses to provide accurate estimates of evolutionary timing for use in studies of biodiversity, phylogeography, development, and genome evolution.

## Acknowledgements

We thank Xi Hang Cao for assisting on building the machine learning model, and Drs. Bui Quang Minh, Beatriz Mello, Heather Rowe, Ananias Escalante, Maria Pacheco, and S. Blair Hedges for comments and edits.

## Funding

This research was supported by grants from National Aeronautics and Space Administration (NASA NNX16AJ30G), National Institutes of

Health (GM0126567-01; LM012487-02), National Science Foundation (NSF DBI 1356548), and Tokyo Metropolitan University (DB105).

## References


Barba-Montoya,J. *et al.* (2018) Constraining uncertainty in the timescale of angiosperm evolution and the veracity of a Cretaceous Terrestrial Revolution. *New Phytol.* **218**, 819-834.

Barreda,V.D. *et al.* (2015) Early evolution of the angiosperm clade Asteraceae in the Cretaceous of Antarctica. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 10989–10994.

Battistuzzi,F.U. *et al.* (2010) Performance of relaxed-clock methods in estimating evolutionary divergence times and their credibility intervals. *Mol. Biol. Evol.*, **27**, 1289–1300.

Bell,C.D. *et al.* (2010) The age and diversification of the angiosperms re-revisited. *Am. J. Bot.*, **97**, 1296–1303.

Brown,J.W. *et al.* (2008) Strong mitochondrial DNA support for a Cretaceous origin of modern avian lineages. *BMC Biol.*, **6**, 6.

Buck,C.B. *et al.* (2016) The Ancient Evolutionary History of Polyomaviruses. *PLoS Pathog.*, **12**, e1005574.

Bzdok,D. *et al.* (2018) Machine learning: supervised methods. *Nat. Methods*, **15**, 5.

Christin,P.-A. *et al.* (2014) Molecular dating, evolutionary rates, and the age of the grasses. *Syst. Biol.*, **63**, 153–165.

Claramunt,S. and Cracraft,J. (2015) A new time tree reveals Earth history's imprint on the evolution of modern birds. *Sci Adv*, **1**, e1501005.

Cox,P.G. and Hautier,L. (2015) Evolution of the Rodents: Volume 5: Advances in Phylogeny, Functional Morphology and Development Cox, P.G. and Hautier, L. (ed) Cambridge: Cambridge University Press.

Dos Reis,M. *et al.* (2016) Bayesian molecular clock dating of species divergences in the genomics era. *Nat. Rev. Genet.*, **17**, 71–80.

Dos Reis,M. *et al.* (2012) Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proc. R. Soc. B*, **279**, 3491–3500.

Dos Reis,M. *et al.* (2014) The impact of the rate prior on Bayesian estimation of divergence times with multiple loci. *Syst. Biol.*, **64**, 555–565.

Dos Reis,M. *et al.* (2015) Uncertainty in the Timing of Origin of Animals and the Limits of Precision in Molecular Timescales. *Curr. Biol.*, **25**, 1–12.

Dos Reis,M. *et al.* (2018) Using phylogenomic data to explore the effects of relaxed clocks and calibration strategies on divergence time estimation: Primates as a test case. *Syst. Biol.*, **67**, 594–615.

Drummond,A.J. *et al.* (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol.*, **4**, 88–99.

Erwin,D.H. *et al.* (2011) The Cambrian conundrum: early divergence and later ecological success in the early history of animals. *Science*, **334**, 1091–1097.

Feng,Y.-J. *et al.* (2017) Phylogenomics reveals rapid, simultaneous diversification of three major clades of Gondwanan frogs at the Cretaceous-Paleogene boundary. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, E5864–E5870.

Filipski,A. *et al.* (2014) Prospects for building large timetrees using molecular data with incomplete gene coverage among species. *Mol. Biol. Evol.*, **31**, 2542–2550.

Foster,C.S. *et al.* (2017) Evaluating the impact of genomic data and priors on Bayesian estimates of the angiosperm evolutionary timescale. *Syst. Biol.*, **66**, 338-351

Gillespie,J.H. (1984) The molecular clock may be an episodic clock. *Proc. Natl. Acad. Sci. U.S.A.*, **81**, 8009–8013.

Hedges,S.B. *et al.* (2015) Tree of life reveals clock-like speciation and diversification. *Mol. Biol. Evol.*, **32**, 835–845.

Hertweck,K.L. *et al.* (2015) Phylogenetics, divergence times and diversification from three genomic partitions in monocots. *Bot. J. Linn. Soc.*, **178**, 375–393.

Ho,S.Y. *et al.* (2015) Simulating and detecting autocorrelation of molecular evolutionary rates among lineages. *Mol. Ecol. Resour.*, **15**, 688–696.

Ho,S.Y. and Duchêne,S. (2014) Molecular-clock methods for estimating evolutionary rates and timescales. *Mol. Ecol.*, **23**, 5947–5965.

Jarvis,E.D. *et al.* (2014) Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, **346**, 1320–1331.

Kimura,M. (1983) The neutral theory of molecular evolution Cambridge: Cambridge University Press.

Kumar,S. (2005) Molecular clocks: four decades of evolution. *Nat. Rev. Genet.*, **6**, 654–662.

Kumar,S. and Hedges,S.B. (2016) Advances in time estimation methods for molecular data. *Mol. Biol. Evol.*, **33**, 863–869.

Lanfear,R. *et al.* (2010) Watching the clock: studying variation in rates of molecular evolution between species. *Trends Ecol. Evol.*, **25**, 495–503.

Lemmon,A.R. *et al.* (2009) The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. *Syst. Biol.*, **58**, 130–145.

Lepage,T. *et al.* (2007) A general comparison of relaxed molecular clock models. *Mol. Biol. Evol.*, **24**, 2669–2680.

Linder,M. *et al.* (2011) Evaluation of Bayesian models of substitution rate evolution-parental guidance versus mutual independence. *Syst. Biol.*, **60**, 329–342.

Liu,L. *et al.* (2017) Genomic evidence reveals a radiation of placental mammals uninterrupted by the KPg boundary. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, E7282–E7290.

Lu,Y. *et al.* (2014) Phylogeny and divergence times of gymnosperms inferred from single-copy nuclear genes. *PLoS One*, **9**, e107679.

Lynch,M. (2010) Evolution of the mutation rate. *Trends Genet.*, **26**, 345–352.

Magallón,S. *et al.* (2013) Land plant evolutionary timeline: gene effects are secondary to fossil constraints in relaxed clock estimation of age and substitution rates. *Am. J. Bot.*, **100**, 556–573.

Marin,J. *et al.* (2017) The Timetree of Prokaryotes: New Insights into Their Evolution and Speciation. *Mol. Biol. Evol.*, **34**, 437–446.

Marin,J. and Hedges,S.B. (2018) Undersampling genomes has biased time and rate estimates throughout the tree of life. *Mol. Biol. Evol.*, msy103.

Meredith,R.W. *et al.* (2011) Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. *Science*, **334**, 521–524.

Metsky,H.C. *et al.* (2017) Zika virus evolution and spread in the Americas. *Nature*, **546**, 411–415.

Moore,B.R. and Donoghue,M.J. (2007) Correlates of diversification in the plant clade Dipsacales: geographic movement and evolutionary innovations. *Am. Nat.*, **170 Suppl 2**, S28–55.

Prum,R.O. *et al.* (2015) A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature*, **526**, 569–578.

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

Sargis,E.J. and Dagosto,M. (2008) Mammalian evolutionary morphology: a tribute to Frederick S. Szalay Sargis, Eric J. and Dagosto, Marian (ed) New York: Springer Netherlands.

Shao,S. *et al.* (2016) Evolution of body morphology and beak shape revealed by a morphometric analysis of 14 Paridae species. *Front. Zool.*, **13**, 30.

Smith,S.A. *et al.* (2010) An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 5897–5902.

Tamura,K. *et al.* (2018) Theoretical foundation of the RelTime method for estimating divergence times from variable evolutionary rates. *Mol. Biol. Evol.*, **35**, 1170–1782.

Thorne,J.L. *et al.* (1998) Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.*, **15**, 1647–1657.

Wiens,J.J. and Moen,D.S. (2008) Missing data and the accuracy of Bayesian phylogenetics. *Journal of Systematics and Evolution*, **46**, 307–314.

Wikström,N. *et al.* (2001) Evolution of the angiosperms: calibrating the family tree. *Proc. R. Soc. B*, **268**, 2211–2220.

Wyles,J.S. *et al.* (1983) Birds, behavior, and anatomical evolution. *Proc. Natl. Acad. Sci. U.S.A.*, **80**, 4394–4397.

Xi,Z. *et al.* (2015) The impact of missing data on species tree estimation. *Mol. Biol. Evol.*, **33**, 838–860.