Nature Communications

Article

Pervasive correlation of molecular evolutionary rates in the tree of life

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

[1]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA19122

[2]Department of Biology, Temple University, Philadelphia, PA19122

[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, MI48309

Correspondence to:

Sudhir Kumar
Temple University
Philadelphia, PA 19122, USA
E-mail: s.kumar@temple.edu

New species arise from pre-existing species and inherit similar genomes and environments. This predicts greater similarity of mutation rates and the tempo of molecular evolution between direct ancestors and descendants, resulting in correlation of evolutionary rates within lineages in the tree of life. Surprisingly, molecular sequence data have not confirmed this expectation, possibly because available methods lack power to detect correlated rates. Here we present an accurate machine learning method used to detect correlation of rates in large phylogenies. By applying this method to multigene and genome-scale sequence alignments from mammals, birds, insects, metazoans, plants, fungi, and prokaryotes, we discover extensive correlation in molecular evolutionary rates throughout the tree of life in both DNA and protein sequences. These findings suggest concordance between molecular and non-molecular evolutionary patterns and will foster unbiased and precise dating of the tree of life.

Phylogenomics has revolutionized our understanding of the patterns and timescale of the tree of life[1,2]. Genome-scale data has revealed that rates of molecular sequence change vary extensively among species[3–5]. The causes and consequences of evolutionary rate variation are of fundamental importance in molecular phylogenetics and systematics[6–8], 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[3,5].

Three decades ago, Gillespie[9] 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[3,5,10–12]. However, the independent branch rate (IBR) model has emerged as a strong alternative to the correlated branch rate (CBR) model. IBR posits that 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[5,13]. IBR model is now widely used in estimating

divergence times from molecular data for diverse groups of species, including mammals[13], birds[14–16], amphibians[17], plants[18–24], and viruses[13,25,26]. 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[6,27,28] and are correlated with taxonomic or geographic distance[29,30].

Alternatively, the widespread use of the IBR model[13–20,22,23,25,26] may be explained by the fact that the currently available statistical tests lack sufficient power to reject the IBR model[31]. This may also explain why some studies report finding extensive branch rate correlation in many datasets (e.g., Lepage et al.[32]), but others cannot confirm this using the same tests (e.g., Linder et al.[19]). Consequently, many researchers use both CBR and IBR models for the same species groups[13,23,33–43], a practice that often generates controversy via widely differing time estimates[34,37,42,44–46].

Therefore, we need a powerful method to accurately test whether evolutionary rates are correlated among branches. This method should then be applied to molecular datasets representing taxonomic diversity across the tree of life to assess the ubiquity of correlated rates in nature. 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 establishes the pervasiveness of rate correlation in the tree of life.

RESULTS

A machine learning approach for detecting rate correlation

Machine learning is widely used to solve problems in many fields, but has not yet been used to address challenges in molecular phylogenetics. We employed a supervised machine learning (McL) framework[47] to build a predictive model that distinguishes 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 lengths[48] (**Fig. 1e** and **1f**) and use these lineage rates to generate informative features. An evolutionary lineage includes all the branches in the descendant subtree, which is distinct from a branch that only connects an ancestor to one of its immediate descendants. One cannot use branch rates as features, because their computation requires the knowledge of node times in the phylogeny, which cannot be estimated without prior assignment of a branch rate model.

The features that we selected for our McL predictive model were the correlation between ancestral and descendant lineage rates (ρ_{ad}), the correlation between the sister lineages (ρ_s), and the decay in ρ_{ad} when one and two parents are skipped (d_1 and d_2, respectively). We selected correlation between ancestral and descendant lineage rates (ρ_{ad}) as a feature because our analyses of simulated data showed that ρ_{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**; **Supplementary information**). While "independent rate" should imply a lack of correlation, ρ_{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. While ρ_{ad} is greater than zero[48], this feature shows distinct patterns for both CBR and IBR models and is thus a good candidate feature for McL (**Fig. 2a**). As our second feature, we selected the correlation between the sister lineages (ρ_s), because ρ_s was higher for the CBR model (0.89) than the IBR model (0.00, **Fig. 2b**; **Supplementary information**). Although our extensive simulations produced some scenarios in which ρ_s was greater than 0.4 for datasets that evolved with the IBR model (because ancestral lineage rates include descendant evolutionary rates) ρ_s remains a highly discriminating feature for McL. Two additional features included in McL measure the decay in ρ_{ad} when one and two parents are skipped (d_1 and d_2), respectively, in ρ_{ad} calculations (**Supplementary information**). We expect that ρ_{ad} will decay slower under CBR than IBR, which was consistent with our observations (**Fig. 2c**).

The selected set of candidate features (ρ_s, ρ_{ad}, d_1, and d_2) can be measured for any phylogeny with branch lengths (e.g., derived from molecular data) 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 can build 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 real world training datasets are few in number and often containing high levels of error or uncertainty[49,50]. 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 (**Supplementary information**). The machine learning process generated a predictive model with an associated correlation score (CorrScore).

We evaluated the sensitivity and specificity of our model using standard receiver operating characteristic (ROC) curves, which show the sensitivity of our method 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 the best, which led to the inclusion of all four features in the predictive model (**Fig. 2d; Supplementary information**). 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. We also performed standard cross-validation tests and found that that the predictive models retained high accuracy (>92%, **Fig. 1k** and **Supplementary information**).

We developed a conventional statistical test (CorrTest) based on CorrScore (**Fig. 2e**) that will provide 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. We found that these CorrScore score thresholds were applicable even when predictive models were developed separately and when the number of sequences in the dataset were small (≤100), medium (100 – 200), large (200 – 300), and very large (> 300) (**Supplementary information**). The accuracy obtained using these models (**Fig. S1a-c**) is similar to those presented in **figure 3d** - **f**. Therefore, we suggest using the general model in CorrTest analysis.

CorrTest performs well in computational tests

We tested the performance of CorrTest on a simulated dataset where the correct rate model is known (**Fig. 1l**). This dataset used 91 angiosperms as a model system for simulating sequence evolution with IBR models (**supplementary information**)[51]. CorrTest correctly diagnosed 95% of these datasets to be evolving with independent rates. We also tested a large collection of datasets[52] generated using diverse evolutionary parameters including both CBR and IBR models (**supplementary information**). 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 best on datasets that contain more and longer sequences (**Fig. 3d**). In these analyses, we used the correct tree topology and nucleotide substitution model. We relaxed this requirement and evaluated CorrTest by first inferring a phylogeny using a dataset[53] with an oversimplified substitution model[54]. Naturally, many inferred phylogenies contained topological errors, but we found the accuracy of CorrTest to still be 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.

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 (**Supplementary information**). We computed Bayes factors (BF) by using the stepping-stone sampling (SS) method (see **Materials and Methods**). 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 ref. [31]). 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[43] 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.

Correlation of rates is 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 encompassing diverse groups across the tree life. This included nuclear, mitochondrial and plastid DNA, and protein sequences from mammals, birds, insects, metazoans, plants, fungi, and prokaryotes (**Table 1**). 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[57]. 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[13–24,41]. In fact, all but three datasets[33,55,56] 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[33] 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 scores decreased when taxon sampling and density were lowered (**Fig. 5a**). Overall, CorrTest detected rate correlation in all the empirical datasets.

Our results establish that the correlated rate model should be the default in molecular clock 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 [44]. For example, the use of IBR model produces estimates of divergence time of two major groups of grasses that are 66% older[46] and origin of a major group of mammal (Erinaceidea) to be 30% older[35] than estimates under CBR model. In fact, substantial differences between node age estimates under IBR and CBR models have been reported in many studies[23,34,37,42,44,46]. Thus, the use of an incorrect rate model has a large impact on time estimates, which may not be alleviated by adding

calibrations[44]. Knowledge that evolutionary rates are generally correlated within lineages will foster unbiased and precise dating of the tree of life.

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[58]. In this model, a single parameter (v) captures the degree of autocorrelation among branches in a phylogenetic tree. A low value of v indicates high autocorrelation, so, we use the inverse of v to represent the degree of rate autocorrelation. MCMCTree[59] analyses of simulated datasets confirmed that the estimated v is linearly related to the true value (**Fig. 5b**). 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[35], which are largely neutral and are the best reflection of mutation rates[60], 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**). These results establish that molecular and non-molecular evolutionary patterns are concordant, because morphological characteristics are also found to be similar between closely-related species[6,27,28] and correlated with taxonomic or geographic distance[29,30].

In conclusion, we have successfully addressed an enduring question in evolutionary biology: are the molecular rates of change between species correlated or

independent? We have shown that the evolutionary rates of change among closely related species are correlated in diverse species groups. That is, evolutionary rate correlation is likely universal, suggesting concordance between the patterns of evolutionary changes in genomes and higher-level biological attributes. Furthermore, revealing the existence of pervasive correlation in molecular rates throughout the tree of life 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.

Materials and Methods

CorrTest analyses. All 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 when we tested the robustness of our model to topological error) in MEGA[61,62]. Then we used those branch lengths to compute relative lineages rates using RRF[48,52] and calculated the value of selected features (ρ_s, ρ_{ad}, and two decay measures) to obtain the CorrScore (see detail calculation in **Supplementary information**). We conducted CorrTest on the CorrScore to estimate the *P*-value of rejecting the null hypothesis of independent evolutionary rates. No calibration was needed for CorrTest analyses.

Bayes factor analyses. We computed the Bayes factor via stepping-stone sampling (BF-SS)[63] with n = 20 and a = 5 using mcmc3r package[43]. We chose BF-SS because the harmonic mean estimator it has many statistical shortcomings[32,63,64] and thermodynamic integration[43,65] is less efficient than BF-SS. Still, BF-SS requires a long computational time, we only finished analyses of 50% of synthetic datasets (**Supplementary information**). 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)[43]. 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[59]. For other priors, we used diffused distributions of

"rgene_gamma = 1 1", "sigma2_gamma=1 1" and "BDparas = 1 1 0". In all Bayesian analyses, two independent runs of 5,000,000 generations each were conducted, and results were checked in Tracer[66] for convergence. ESS values were higher than 200 after removing 10% burn-in samples for each run.

Analysis of empirical datasets

We used 16 datasets from 12 published studies of eukaryotes and 2 published studies of prokaryotes that cover the major groups in the tree of life (**Table 1**).These were selected because they did not contain too much missing data (<50%) and represented >80 sequences. When a phylogeny and branch lengths were available from the original study, we estimated relative rates directly from the branch lengths via the relative rate framework[48] 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[61,62].

To obtain the autocorrelation parameter (v), we used MCMCTree[59] with the same input priors as the original study, but no calibration priors were used in order to avoid undue influence of calibration uncertainty densities on the estimate of autocorrelation parameters. We did, however, provide a root calibration because MCMCTree requires a root calibration. For this purpose, we used the root calibration provided in the original article or selected the median age of the root node in the TimeTree database[67,68] ± 50My (soft uniform distribution) as the root calibration, as this does not impact the estimation of v. Bayesian analyses required long computational times, so we used the original alignments in MCMCTree analyses if alignments were shorter than 20,000 sites. If the alignments were longer than 20,000 sites, we randomly selected 20,000 sites from the original alignments to use in MCMCTree analyses. However, one dataset[69] 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 use in MCMCtree analyses (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[66] for convergence. ESS values were

higher than 200 after removing 10% burn-in samples for each run. All empirical datasets are available at https://github.com/cathyqqtao/CorrTest.

Code availability statement

The R source code of CorrTest is available at https://github.com/cathyqqtao/CorrTest

Data availability statement

All empirical datasets, results, and source code for generating each figure are available at https://github.com/cathyqqtao/CorrTest. All simulated datasets will be provided upon request.

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 critical comments and editorial suggestions. 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

1. Hedges, S. B., Marin, J., Suleski, M., Paymer, M. & Kumar, S. Tree of life reveals clock-like speciation and diversification. *Mol. Biol. Evol.* **32,** 835–845 (2015).

2. Marin, J., Battistuzzi, F. U., Brown, A. C. & Hedges, S. B. The Timetree of Prokaryotes: New Insights into Their Evolution and Speciation. *Mol. Biol. Evol.* **34,** 437–446 (2017).

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

4. Dos Reis, M., Donoghue, P. C. & Yang, Z. Bayesian molecular clock dating of species divergences in the genomics era. *Nat. Rev. Genet.* **17,** 71–80 (2016).

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

6. Lanfear, R., Welch, J. J. & Bromham, L. Watching the clock: studying variation in rates of molecular evolution between species. *Trends Ecol. Evol.* **25,** 495–503 (2010).

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

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

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

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

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

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

13. Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4,** 88–99 (2006).

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

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

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

17. Feng, Y.-J. *et al.* 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 (2017).

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

19. Linder, M., Britton, T. & Sennblad, B. Evaluation of Bayesian models of substitution rate evolution-parental guidance versus mutual independence. *Syst. Biol.* **60,** 329–342 (2011).

20. Lu, Y., Ran, J.-H., Guo, D.-M., Yang, Z.-Y. & Wang, X.-Q. Phylogeny and divergence times of gymnosperms inferred from single-copy nuclear genes. *PLoS One* **9,** e107679 (2014).

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

22. Smith, S. A., Beaulieu, J. M. & Donoghue, M. J. An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants. *Proc. Natl. Acad. Sci. U.S.A.* **107,** 5897–5902 (2010).

23. Bell, C. D., Soltis, D. E. & Soltis, P. S. The age and diversification of the angiosperms re-revisited. *Am. J. Bot.* **97,** 1296–1303 (2010).

24. Barba-Montoya, J., Dos Reis, M., Schneider, H., Donoghue, P. C. J. & Yang, Z. Constraining uncertainty in the timescale of angiosperm evolution and the veracity of a Cretaceous Terrestrial Revolution. *New Phytol.* (2018). doi:10.1111/nph.15011

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

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

27. Sargis, E. J. & Dagosto, M. *Mammalian evolutionary morphology: a tribute to Frederick S. Szalay*. (New York: Springer Netherlands, 2008).

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

29. Wyles, J. S., Kunkel, J. G. & Wilson, A. C. Birds, behavior, and anatomical evolution. *Proc. Natl. Acad. Sci. U.S.A.* **80,** 4394–4397 (1983).

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

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

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

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

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

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

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

37. Foster, C. S. *et al.* Evaluating the impact of genomic data and priors on Bayesian estimates of the angiosperm evolutionary timescale. *Syst. Biol.* syw086 (2016).

38. Magallón, S., Hilu, K. W. & Quandt, D. 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 (2013).

39. Wikström, N., Savolainen, V. & Chase, M. W. Evolution of the angiosperms: calibrating the family tree. *Proc. R. Soc. B* **268,** 2211–2220 (2001).

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

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

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

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

44. Battistuzzi, F. U., Filipski, A., Hedges, S. B. & Kumar, S. Performance of relaxed-clock methods in estimating evolutionary divergence times and their credibility intervals. *Mol. Biol. Evol.* **27,** 1289–1300 (2010).

45. Dos Reis, M., Zhu, T. & Yang, Z. The impact of the rate prior on Bayesian estimation of divergence times with multiple loci. *Syst. Biol.* **64,** 555–565 (2014).

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

47. Bzdok, D., Krzywinski, M. & Altman, N. Machine learning: supervised methods. *Nat. Methods* **15,** 5 (2018).

48. Tamura, K., Tao, Q. & Kumar, S. Theoretical foundation of the RelTime method for estimating divergence times from variable evolutionary rates. *Mol. Biol. Evol.* msy044

(2018). doi:10.1093/molbev/msy044

49. Ekbatani, H. K., Pujol, O. & Segui, S. Synthetic Data Generation for Deep Learning in Counting Pedestrians. In *Pattern Recognition Applications and Methods (ICPRAM), 2017 The International Conference on* 318–323 (2017).

50. Le, T. A., Baydin, A. G., Zinkov, R. & Wood, F. Using synthetic data to train neural networks is model-based reasoning. In *Neural Networks (IJCNN), 2017 International Joint Conference on* 3514–3521 (2017).

51. Beaulieu, J. M., O'Meara, B. C., Crane, P. & Donoghue, M. J. Heterogeneous rates of molecular evolution and diversification could explain the Triassic age estimate for angiosperms. *Syst. Biol.* **64,** 869–878 (2015).

52. Tamura, K. *et al.* Estimating divergence times in large molecular phylogenies. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 19333–19338 (2012).

53. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4,** 406–425 (1987).

54. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16,** 111–120 (1980).

55. Calteau, A. *et al.* Phylum-wide comparative genomics unravel the diversity of secondary metabolism in Cyanobacteria. *BMC Genomics* **15,** 977 (2014).

56. Battistuzzi, F. U. & Hedges, S. B. A major clade of prokaryotes with ancient adaptations to life on land. *Mol. Biol. Evol.* **26,** 335–343 (2009).

57. Jayaswal, V., Wong, T. K., Robinson, J., Poladian, L. & Jermiin, L. S. Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. *Syst. Biol.* **63,** 726–742 (2014).

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

59. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24,** 1586–1591 (2007).

60. Kumar, S. & Subramanian, S. Mutation rates in mammalian genomes. *Proc. Natl. Acad. Sci. U.S.A.* **99,** 803–808 (2002).

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

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

63. Xie, W., Lewis, P. O., Fan, Y., Kuo, L. & Chen, M.-H. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.* **60,** 150–160 (2011).

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

65. Silvestro, D., Schnitzler, J. & Zizka, G. A Bayesian framework to estimate diversification rates and their variation through time and space. *BMC Evol. Biol.* **11,** 311 (2011).

66. Rambaut, A., Suchard, M., Xie, D. & Drummond, A. Tracer v1.6. (2014). at <http://beast.bio.ed.ac.uk/Tracer>

67. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol. Biol. Evol.* **34,** 1812–1819 (2017).

68. Hedges, S. B., Dudley, J. & Kumar, S. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22,** 2971–2972 (2006).

69. Ruhfel, B. R., Gitzendanner, M. A., Soltis, P. S., Soltis, D. E. & Burleigh, J. G. From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Evol. Biol.* **14,** 23 (2014).

70. Misof, B. *et al.* Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346,** 763–767 (2014).

71. Wickett, N. J. *et al.* Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl. Acad. Sci. U.S.A.* **111,** E4859–4868 (2014).

72. Shen, X.-X. *et al.* Reconstructing the Backbone of the Saccharomycotina Yeast Phylogeny Using Genome-Scale Data. *G3* **6,** 3927–3939 (2016).

Table 1. Results from the CorrTest analysis 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)[35]
Mammals	Nuclear 3rd codon	138	11,010	GTR + Γ	0.99	< 0.001	4.42	Meredith et al. (2011)[35]
Mammals	Nuclear proteins	138	11,010	JTT + Γ	0.99	< 0.001	3.11	Meredith et al. (2011)[35]
Mammals	Mitochondrial DNA	271	7,370	HKY + Γ	0.98	< 0.001	3.77	Dos Reis, et al. (2012)[36]
Birds	Nuclear DNA	198	101,781	GTR + Γ	1.00	< 0.001	2.07	Prum et al. (2015)[15]
Birds	Nuclear 3rd codon	222	1,364	GTR + Γ	1.00	< 0.001	2.11	Claramunt et al. (2015)[16]
Birds	Nuclear 1st and 2nd codon	222	2,728	GTR + Γ	1.00	< 0.001	2.53	Claramunt et al. (2015)[16]
Insects	Nuclear proteins	143	220,091	LG + Γ	1.00	< 0.001	8.68	Misof et al. (2015)[70]
Metazoans	Mitochondrial & nuclear proteins	113	2,049	LG + Γ	0.65	< 0.05	40.0	Erwin et al. (2011)[33]
Plants	Plastid 3rd codon	335	19,449	GTR + Γ	1.00	< 0.001	2.28	Ruhfel et al. (2014)[69]
Plants	Plastid proteins	335	19,449	JTT + Γ	1.00	< 0.001	2.46	Ruhfel et al. (2014)[69]
Plants	Nuclear 1st and 2nd codon	99	220,091	GTR + Γ	1.00	< 0.001	5.50	Wickett et al. (2014) [71]
Plants	Chloroplast and nuclear DNA	124	5,992	GTR + Γ	1.00	< 0.001	2.64	Beaulieu et al. (2015)[51]
Fungi	Nuclear proteins	85	609,772	LG + Γ	0.97	< 0.001	3.78	Shen et al. (2016)[72]
Prokaryotes	Nuclear proteins	197	6,884	JTT + Γ	0.79	< 0.05	2.54	Battistuzzi et al. (2009)[56]
Prokaryotes	Nuclear proteins	126	3,145	JTT + Γ	0.83	< 0.05	1.23	Calteau et al. (2014)[55]

[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.

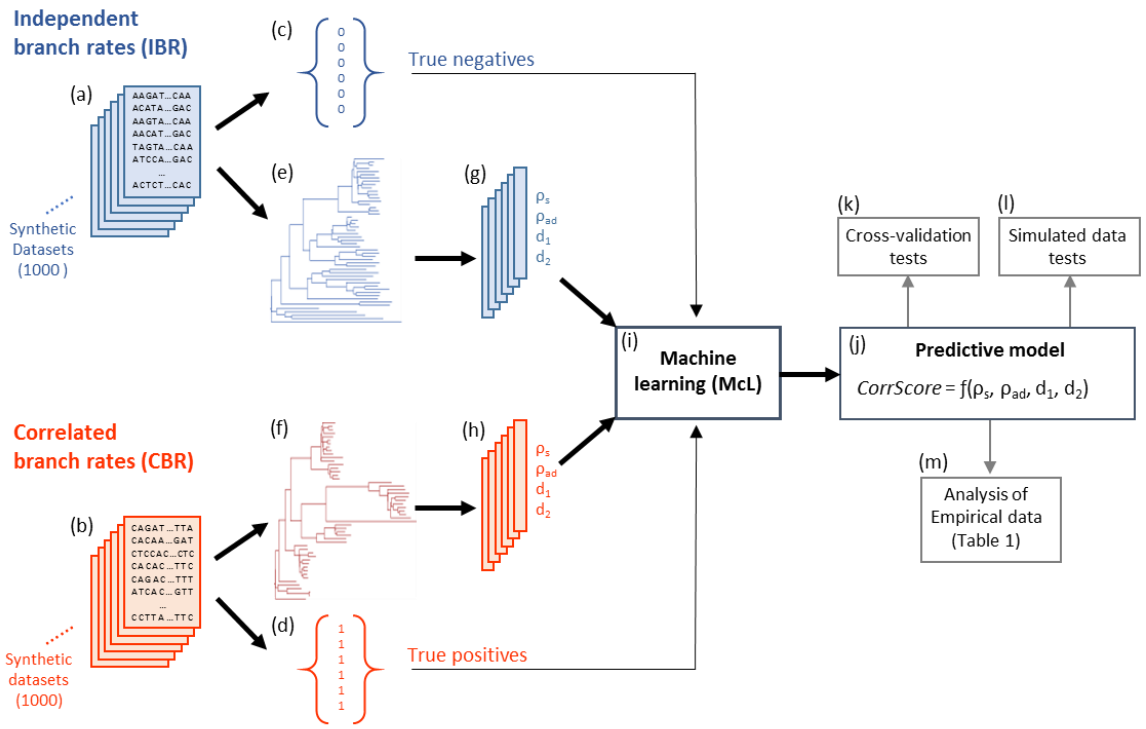

Figure 1. A flowchart showing an overview of the machine learning (McL) approach applied to develop the predictive model (CorrTest). We generated **(a)** 1,000 synthetic datasets that were evolved using an IBR model and **(b)** 1,000 synthetic datasets that were evolved using a CBR model. The numerical 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 ρ_s, ρ_{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.

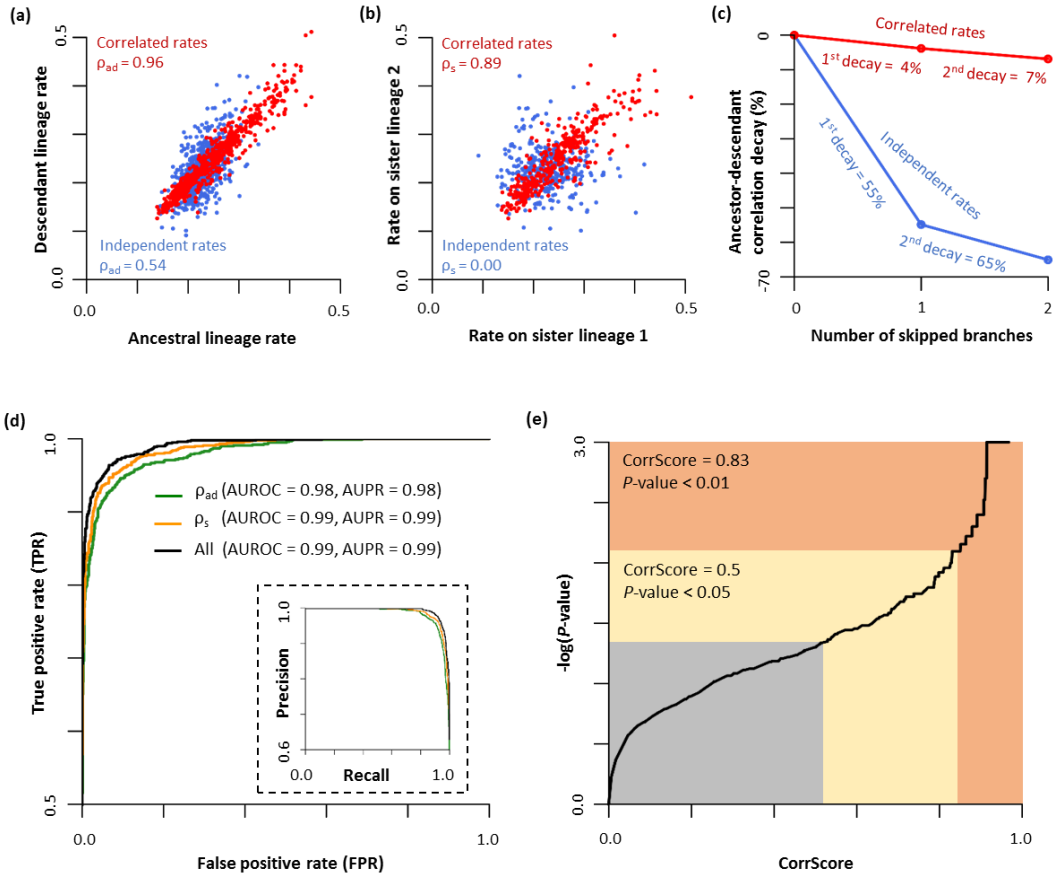

Figure 2. The relationship of **(a)** ancestral and direct descendent lineage rates and **(b)** sister lineage rates when the simulated evolutionary rates were correlated with each other (red) or varied independently (blue). The correlation coefficients are shown. **(c)** The decay of correlation between 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 (ρ_{ad}, green), only sister lineage rates (ρ_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 $P < 0.05$.

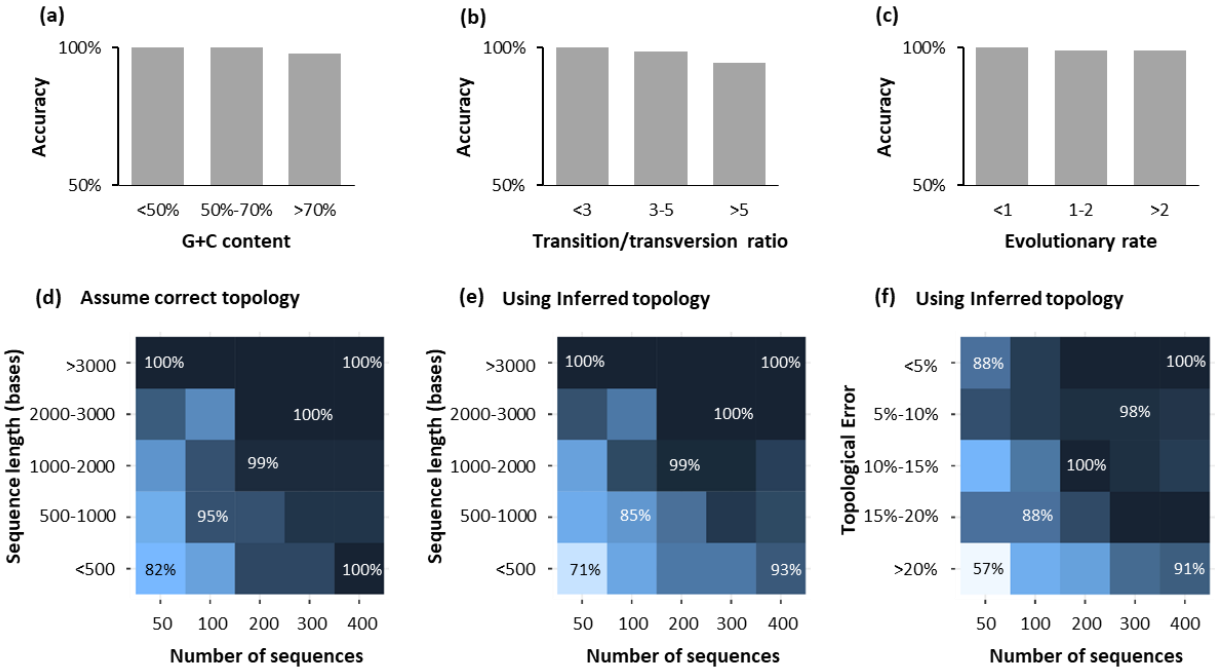

Figure 3. The performance of CorrTest in detecting rate correlation in the analysis of datasets[52] that were simulated with different **(a)** G+C contents, **(b)** transition/transversion rate ratios, and **(c)** average molecular evolutionary rates. Darker color indicates higher accuracy. The evolutionary 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.

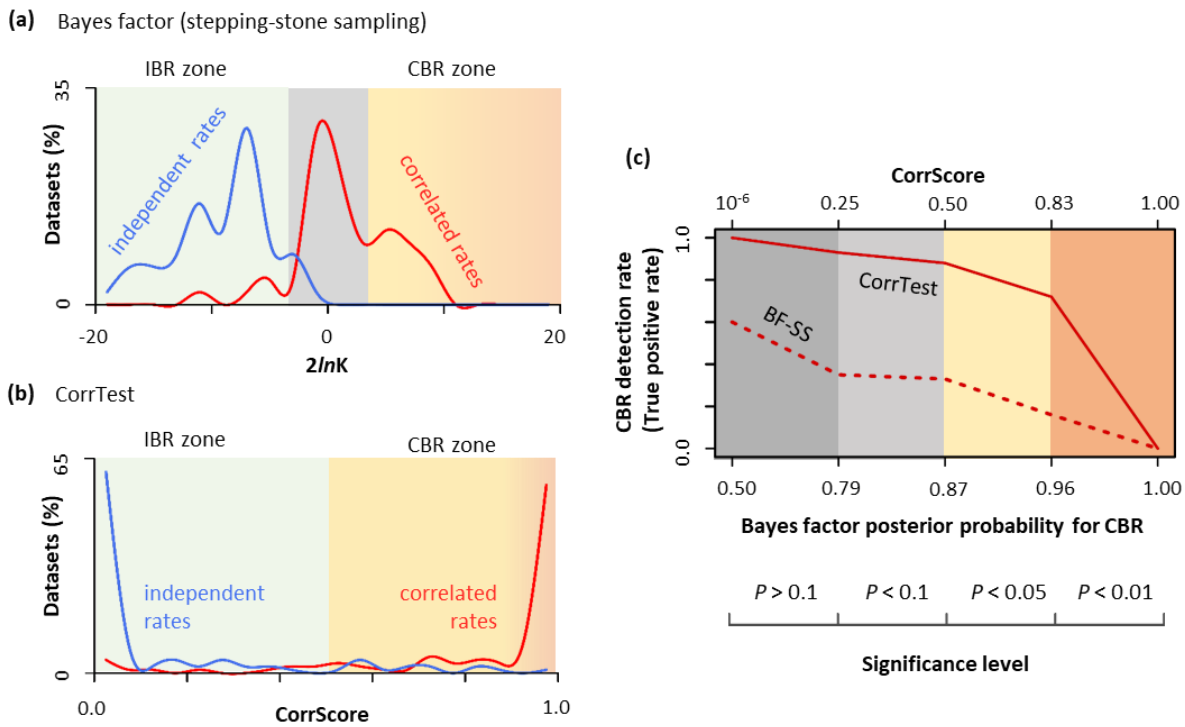

Figure 4. Comparisons of the performance of CorrTest and Bayes Factor analyses. **(a)** Distributions of 2 times the differences of marginal log-likelihood (2*ln*K) estimated via stepping-stone sampling method for datasets that were simulated with correlated branch rates (CBR, red) and independent branch 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.

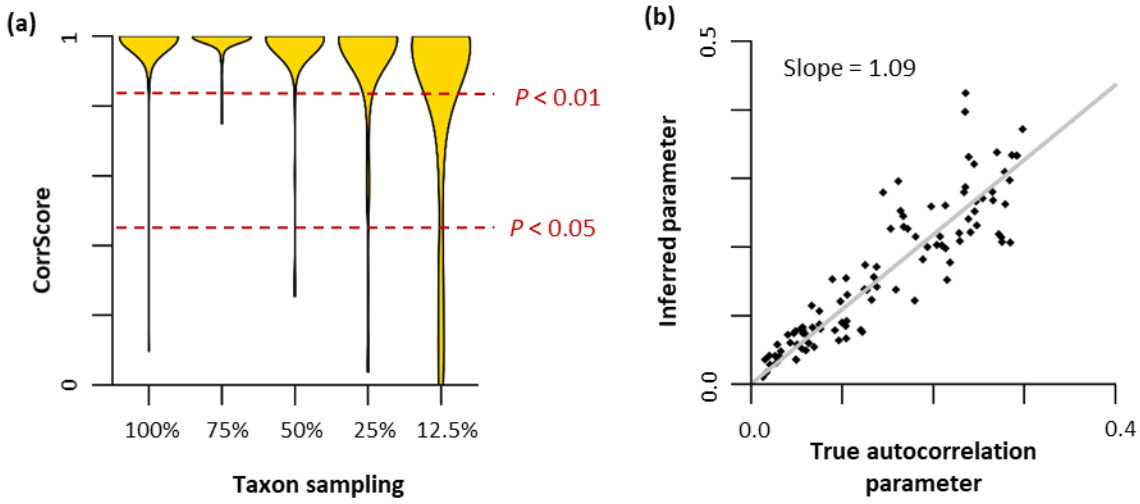

Figure 5. (a) The distribution of CorrScore when data have different taxon sampling densities. The CorrScore decreases when the density of taxon sampling is lower, as there is much less information to discriminate between CBR and IBR. Red, dashed lines mark two statistical significance levels of 5% and 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.