Fast and Accurate Estimates of Divergence Times from Big Data

Beatriz Mello,[1] Qiqing Tao,[1,2] Koichiro Tamura,[3,4] and Sudhir Kumar*[,1,2,5,6]

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

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

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

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

[5]Center for Biodiversity, Temple University, Philadelphia, PA

[6]Center for Excellence in Genome Medicine and Research, King Abdulaziz University, Jeddah, Saudi Arabia

*Corresponding author: E-mail: s.kumar@temple.edu

Associate editor: Naoko Takezaki

Abstract

Ongoing advances in sequencing technology have led to an explosive expansion in the molecular data available for building increasingly larger and more comprehensive timetrees. However, Bayesian relaxed-clock approaches frequently used to infer these timetrees impose a large computational burden and discourage critical assessment of the robustness of inferred times to model assumptions, influence of calibrations, and selection of optimal data subsets. We analyzed eight large, recently published, empirical datasets to compare time estimates produced by RelTime (a non-Bayesian method) with those reported by using Bayesian approaches. We find that RelTime estimates are very similar to Bayesian approaches, yet RelTime requires orders of magnitude less computational time. This means that the use of RelTime will enable greater rigor in molecular dating, because faster computational speeds encourage more extensive testing of the robustness of inferred timetrees to prior assumptions (models and calibrations) and data subsets. Thus, RelTime provides a reliable and computationally thrifty approach for dating the tree of life using large-scale molecular datasets.

Key words: **molecular clocks, computational time, Bayesian, timetree.**

Introduction

Progress in sequencing technology has led to a two-dimensional expansion of datasets being used for dating evolutionary divergences, because both the number of sites in the sequence alignment and the number of included taxa are increasing quickly (Perelman et al. 2011; Pyron 2014; Cannon et al. 2016; Kumar and Hedges 2016; Tarver et al. 2016). Large time-calibrated phylogenies generated using these data provide new insight into evolutionary patterns and the underlying processes responsible for the biological diversity around us (Kumar 2005; Ho and Duchêne 2014; Kumar and Hedges 2016; Dos Reis et al. 2016).

With the rise of big data, the application of Bayesian approaches for dating cladogenetic events is becoming computationally demanding. Computational time requirements increase exponentially with increases in the number of species and the sequence length (Battistuzzi et al. 2011; Tamura et al. 2012). For example, it takes almost half a day to compute divergence times in a dataset with 43 species (~55k sites); and multiple days to estimate a timetree for 274 mitochondrial sequences (first and second codon positions) on a personal computer (Intel® Core i7® CPU @ 4.0GHz). Such slow speeds hinder discovery and can lead to suboptimal scientific practices, because they discourage tests of the robustness of inferred timetrees to the model and calibration assumptions.

Recently, ultra-fast non-Bayesian dating methods have been developed, which incorporate the possibility of rate variation from branch to branch and include multiple calibration points (Tamura et al. 2012; To et al. 2015). RelTime first transforms an evolutionary tree with branch lengths, in the units of number of substitutions per site, into an ultrametric tree with relative times. This is accomplished by estimating branch-specific relative rates for descendants of each internal node, by using the fact that the time elapsed from the most recent common ancestor of two sister lineages is equal when all the taxa are contemporaneous (Tamura et al. 2012). The final timetree is obtained by converting the ultrametric tree into a timetree using one or more calibrations points (Kumar and Hedges 2016).

Ultra-fast non-Bayesian methods have already been shown to produce excellent time estimates for simulated data (Tamura et al. 2012; Filipski et al. 2014; To et al. 2015). For example, the performance of RelTime was comparable to Bayesian approaches in computer simulations where large sequence datasets were generated under conditions with autocorrelated and independent rates among lineages [e.g., Fig. 5A–C in Tamura et al. (2012)]. Also, RelTime sometimes produced estimates that were frequently more accurate than Bayesian and other approaches (Sanderson 2003; Yang 2007), especially when there was a 50% rate increase in a specific clade [Fig. 5E–F in Tamura et al. (2012)]. We have found RelTime estimates to be robust to missing data, even when sequences from a majority of genes

Letter

Fast Track

Table 1. Detailed Information about the Large-Scale Datasets Analyzed

Data Name	Data Type[a]	Site Count	Taxa Count	Calibration Count	Substitution Model[b]	Software Used	Data Reference
Mammals (A)	N	20,593,949	36	25	HKY + G_4	MCMCTree	Dos Reis et al. (2012)[c]
Mammals (B)	M	7,370	274	33	HKY + G_4	MCMCTree	Dos Reis et al. (2012)[c]
Mammals (C)	A	11,010	162	64	JTT + G_4	MCMCTree	Meredith et al. (2011)
Spiders	A	55,447	43	8	WAG + G_5	RelTime	Bond et al. (2014)
Metazoans	A	38,577	54	33	LG + G_4 +F	MCMCTree	Dos Reis et al. (2015)
Insects	A	220,091	144	38	LG + G_4	BEAST	Misof et al. (2014)
Birds (A)	N	722,202	51	18	HKY + G_4	MCMCTree	Jarvis et al. (2014)
Birds (B)	N	101,781	200	20	GTR + G_4 + I	BEAST	Prum et al. (2015)

[a]N = nuclear DNA; M = mitochondrial DNA; A = amino acid (nuclear).
[b]The model that was used for the majority of partitions, if applicable. The number of discrete categories to approximate the Gamma distributions is shown.
[c]The study used two distinct mammalian data sets to estimate divergence times.

for many species were absent from the alignment (Filipski et al. 2014). Importantly, non-Bayesian approaches complete calculations thousands of times faster than the fastest Bayesian method (Dos Reis and Yang 2011), with even greater speed advantage for larger numbers of sequences (Tamura et al. 2012; To et al. 2015).

However, it is not clear if RelTime produces divergence time estimates are comparable to those obtained using Bayesian methods on empirical data, especially when the datasets are very large. If this is true, then RelTime would provide a computationally tractable alternative to Bayesian methods. Therefore, we directly compared Bayesian and RelTime methods by reanalyzing eight large-scale empirical datasets obtained from recently published studies (table 1). In these sequence alignments, the number of taxa ranged from 36 to 274 and the number of sites ranged from 7,370 to 20,593,949 (nucleotides or amino acids). These datasets represent some of the largest timetree analyses performed to date. To ensure comparability, we used identical substitution models, calibrations, and tree topologies in divergence time estimation by Bayesian and RelTime methodologies (see "Material and Methods" section).

Results

For each dataset, we found an extremely high correlation between Bayesian and RelTime estimates (0.88–0.99), with the slope of linear regression through the origin ranging from 0.91 to 1.07 (fig. 1; table 2). These correlations remain very high even when nodes with calibrations are excluded from the analysis (table 2), which is important to test because the inclusion of calibrated nodes inflates correlation by constraining the node age estimates to a narrow range in Bayesian and non-Bayesian methods. RelTime produced time estimates similar to Bayesian approaches for nuclear and mitochondrial DNA data and amino acid data from nuclear genes analyzed, in substantially less time. For example, MCMCTree (the fastest Bayesian implementation) took 650 times longer than RelTime to estimate evolutionary timing for a mitochondrial dataset of 274 taxa and 7,370 sites (Mammals A) on the same machine (Intel® Xeon® CPU @ 2.4GHz).

We also calculated the difference between RelTime and Bayesian estimates for each node and normalized it by the Bayesian time estimate for that node. The absolute mean of these normalized differences over the whole tree was fairly low (4.7–23.7%; table 2). These are relatively small differences when considering that Bayesian credibility intervals (CrIs) are generally much wider, e.g., the mean CrI width is 63% of the Bayesian times for three datasets for which CrIs of node times were available from published supplementary information (Meredith et al. 2011; Misof et al. 2014); or could be inferred (Bond et al. 2014). Therefore, concordances between RelTime and Bayesian time estimates are high.

While summary statistics show broad agreement between RelTime and Bayesian methods, scatterplots in figure 1 do reveal notable differences. The biggest differences are observed for the Birds (B) dataset, which consists of 198 bird species (plus two outgroups) with a sequence alignment spanning 101,781 bases. Here, the correlation between RelTime and Bayesian dates is the lowest (0.88) and the relationship between RelTime and Bayesian dates is balloon-shaped. To better understand the factors contributing to this difference, we overlaid timetrees produced by RelTime and Bayesian methods. This comparison revealed one large clade (fig. 2A, clade 1) in which Bayesian time estimates were up to 54% older than RelTime estimates (35% older on average). This clade consists of 50 species of Neoaves (stem Psittaciformes node). We examined root-to-tip branch lengths in clade 1 (fig. 2B), because we have previously found that Bayesian methods may produce older dates when there are large clade-specific increases in evolutionary rates [Fig. 5E in Tamura et al. (2012)]. Indeed, clade 1 has experienced a significant rate acceleration, because root-to-tip lengths are 45% longer than the rest of the tree (fig. 2B and C). In fact, similar patterns are observed for other rate-accelerated clades highlighted in figure 2.

Another noticeable divergence between RelTime and Bayesian estimates is observed for one node in the metazoan dataset (fig. 1E). The RelTime estimate for the most recent common ancestor (MRCA) of Chaetognatha is 49% smaller than the Bayesian time. We found that the posterior distributions of times in Bayesian analyses are multi-modal for this node (fig. 3) and that the RelTime estimate for this node lies within one of the highest density peaks, which is younger than the other two. Multi-modality of posterior distributions can be caused by strong correlation of parameters and/or by

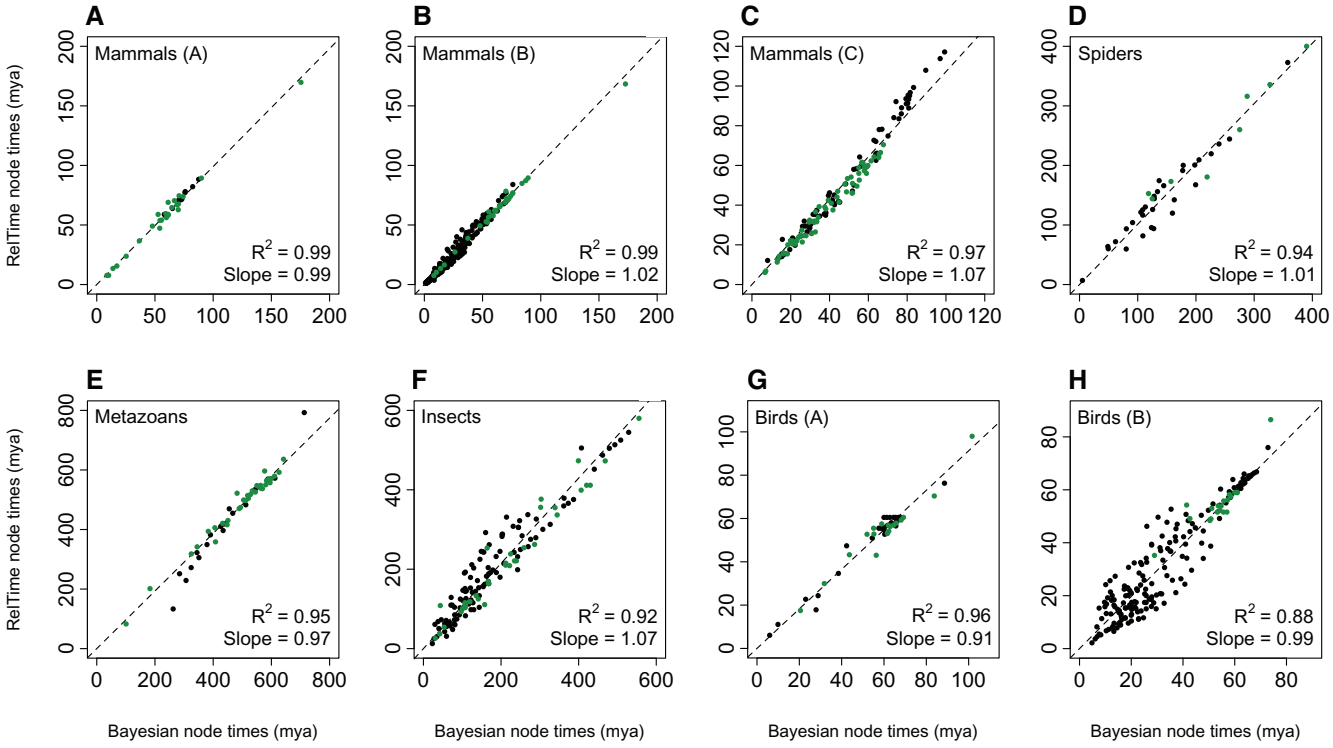

Fig. 1. Comparison of Bayesian and RelTime time estimates for each of the datasets analyzed. Each point represents an estimate of absolute time via Bayesian (*x*-axis) and RelTime (*y*-axis) methods. Each graph contains the linear regression through the origin (dashed line), and the slope and linear regression coefficient (R^2) values. Calibrated nodes are shown in green. The names inside panels refer to table 1: (*A*–*C*) Mammalian datasets; (*D*) Spider dataset; (*E*) Metazoan dataset; (*F*) Insect dataset; (*G*–*H*) Bird datasets.

Table 2. A Summary Comparison of Time Estimates Produced by Bayesian and RelTime Methods

Dataset	All Nodes			Nodes Without Calibrations		
	% Diff	R^2	Slope	% Diff	R^2	Slope
Mammals (A)	4.7	0.99	0.99	1.5	0.98	1.00
Mammals (B)	11.7	0.99	1.02	13.0	0.97	1.03
Mammals (C)	8.5	0.97	1.07	10.7	0.97	1.11
Spiders	14.4	0.94	1.01	15.1	0.92	1.00
Metazoans	6.4	0.95	0.97	10.1	0.91	0.96
Insects	20.7	0.92	1.07	24.2	0.90	1.10
Birds (A)	9.2	0.96	0.91	9.1	0.97	0.92
Birds (B)	23.7	0.88	0.99	25.7	0.86	0.98

Note.—% Diff, absolute mean of the normalized difference between time estimates from RelTime and Bayesian methods. R^2, linear fit coefficient values. Slope, those of the linear regression line through the origin.

lack of stationarity of MCMC calculations (Yang 2014, p. 229). For many nodes, time estimates themselves varied extensively when different calibration probability densities (e.g., uniform, skew-normal, and Cauchy) were applied in Bayesian analyses (Dos Reis et al. 2015). Therefore, we do not consider the differences between RelTime and Bayesian methods to be significant in this case.

Differences between RelTime and Bayesian estimates can arise for a number of reasons. For example, Beaulieu et al. (2015) have shown that heterogeneity of rate models among clades (including acceleration of rates and increased dispersion in rates) can lead to overestimates in Bayesian divergence times, because current Bayesian methods fit the same branch rates model (e.g., lognormal) to the whole tree. Consequently, differences between Bayesian and RelTime estimates could arise due to the fact that RelTime does not require the same statistical distribution with the same set of parameters to model rate heterogeneity in the tree. Differences could also arise because Bayesian methods incorporate a prior to describe branching process across the tree (e.g., birth–death or Yule process), which may not fit the whole tree, especially in a large phylogeny. RelTime does not assume an underlying diversification process, so a lack of consistent birth-death or Yule diversification among clades is unlikely to affect RelTime estimates.

Overall, we conclude that the concordance of time estimates between Bayesian and RelTime approaches is strong across datasets that vary extensively in numbers of taxa and length of the sequence alignment. Therefore, RelTime provides an accurate and computationally-efficient approach to estimate times when Bayesian methods are infeasible. Furthermore, achieving similar results from two distinct approaches increases our confidence in biological conclusions. As Bayesian methods require many more priors than RelTime (Kumar and Hedges 2016), we recommend that RelTime should be applied along with Bayesian and other approaches (Yang 2007; Drummond et al. 2012; Ronquist et al. 2012; Smith and O'Meara 2012). At the same time, it is important to note that the absolute times produced by

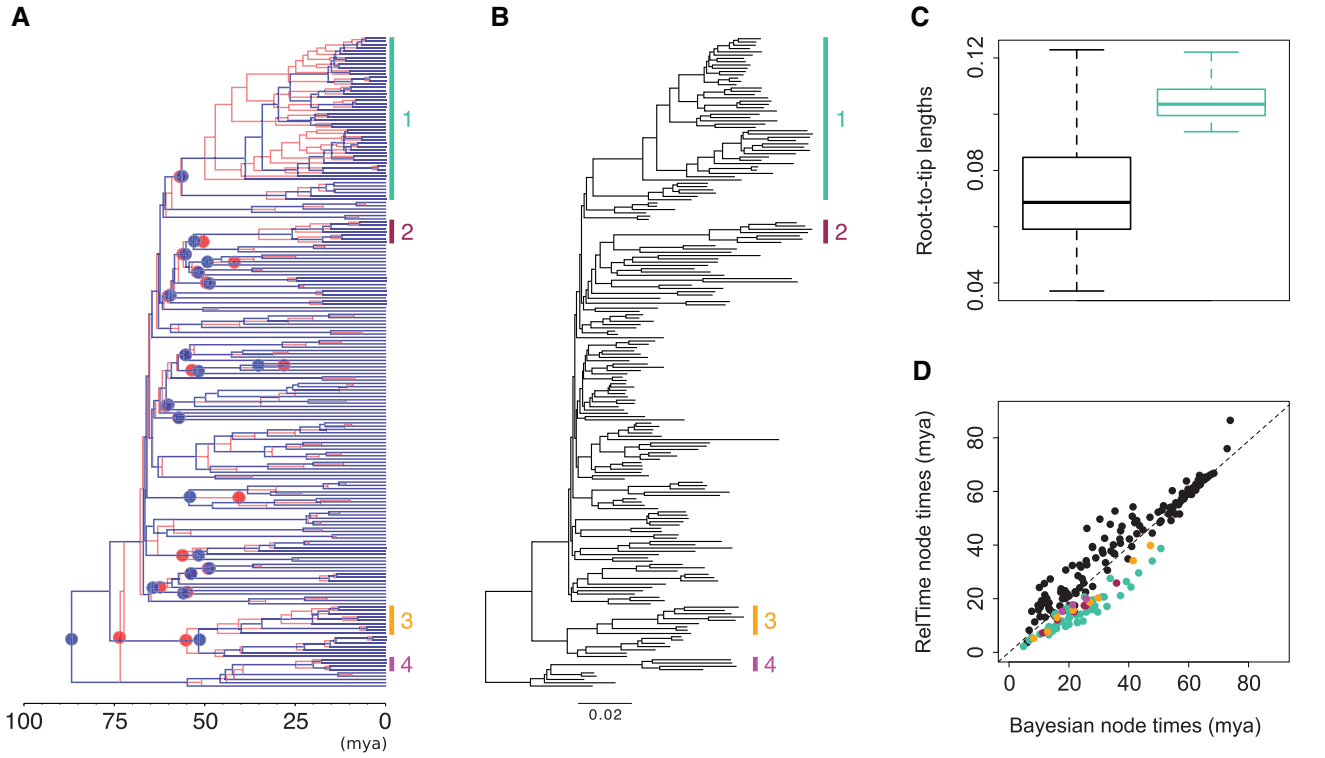

FIG. 2. (*A*) Bayesian and RelTime timescales for the Birds B dataset (Prum et al. 2015). The RelTime timetree is shown in blue, while the Bayesian is shown in red. Dots indicate the calibrated nodes. The colored vertical bars highlight clades that show discrepancies between RelTime and Bayesian analysis due to evolutionary rate acceleration (see text for discussion). (*B*) Phylogenetic tree with branch lengths estimated by the maximum likelihood method. (*C*) Root-to-tip distances obtained using maximum likelihood branch lengths for clade 1 (cyan bar) as compared with the rest of the tree (in black). (*D*) Scatter graph of Bayesian and RelTime estimates of node ages. Colored points correspond to discrepant clades highlighted in panels A and B.

RelTime and Bayesian methods rely strongly on the assumption that the calibrations are correct. Biological accuracy of time estimates for RelTime and Bayesian methods is directly dependent on calibration information and associated probability densities, which are usually based on empirical fossil-derived data. Therefore, it is advisable to test the impact of calibration information on the inference of divergence times by applying protocols such as those proposed by Battistuzzi et al. (2015).

Material and Methods

Data Acquisition

We obtained genomic datasets and timetrees from seven recently published studies, including Meredith et al. (2011) and Dos Reis et al. (2012) for mammals, Jarvis et al. (2014) and Prum et al. (2015) for birds, Dos Reis et al. (2015) for metazoans, Misof et al. (2014) for insects, and Bond et al. (2014) for spiders. Alignment sizes, data types, the number of terminal taxa, timing calibrations, and methodology originally employed are summarized in table 1. All divergence times were originally inferred using the Bayesian software packages MCMCTree (Yang 2007) or BEAST (Drummond et al. 2012), except for the study of Bond et al. (2014), which estimated divergence times via the RelTime method. Therefore, we used the published alignments as provided in the original studies to estimate divergence times. Whenever possible, timetrees were obtained from the same studies. When the study did not provide complete timetree data, we obtained the phylogenies instead and manually added divergence times based on the node ages displayed on the original studies.

RelTime Inference

We used the same alignment, topology, calibrations and substitution model as in the original studies to estimate absolute times in RelTime (Tamura et al. 2012). All RelTime calculations were carried out on the command line version of MEGA7 (Kumar et al. 2012; Kumar et al. 2016) on an Intel® Xeon® CPU E5-2665 @ 2.4 GHz machine. If distinct substitution models had been originally applied to different partitions, we used the model originally applied to the majority of partitions. Since RelTime only requires minimum and/or maximum boundaries of calibrations, we used the boundaries specified in original studies. If only the calibration density distributions were provided in an original study, we derived the boundaries based on the 95% cumulative probability of density distributions for those calibrations. All calibrations located on the root of the phylogenetic tree and on the outgroup clade were automatically removed because the assumption of equal rates of evolution between the ingroup and outgroup sequences is not testable (Kumar et al. 2016). In the primary analysis of the insect dataset (Misof et al. 2014), we noticed that the calibration assigned by the authors to the root node placed a strong constraint on their results, so we instead used the maximum boundary of this root calibration

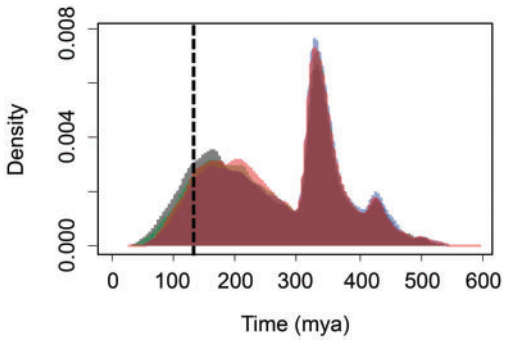

Fig. 3. Posterior distributions derived from distinct calibration strategies for Chaetoganatha MRCA node times (modified from the Figure S1A in Dos Reis et al. 2012). The dashed vertical line marks the RelTime estimate for the corresponding divergence. The color of the distributions refers to the calibration strategies used by the authors, which mainly differ in the probability densities employed. At the calibration scenario S1 (black), all the calibrations were assigned as uniform distributions with soft bounds. In S2 (red), S3 (green) and S4 (blue), distinct calibration densities were assigned for the phylum and superphylum crown nodes (skew-normal, and truncated Cauchy with long and short tails extending back in time, respectively), while the others were kept as uniform distributions. Additionally, the age of crown Metazoa had different minimum constraints between S1-S2 and S3-S4, based on distinct interpretations of the fossil record (Love et al. 2009; Antcliffe et al. 2014).

as a maximum time for the most recent common ancestor of the ingroup.

Bayesian Inference

Because Bond et al. (2014) originally used RelTime to estimate divergence times, we used the same data in MCMCTree (Yang 2007; Dos Reis and Yang 2011) to obtain Bayesian estimates. Minimum and maximum times originally adopted as calibration constraints were used as minimum and maximum values to delimit uniform distributions with soft bounds (left and right tail probabilities equal 0.025) in MCMCTree analysis. The time unit was set to be 100 million years. We applied an independent rates model and the WAG $+\Gamma_5$ substitution model (Whelan and Goldman 2001) as authors originally did. We ran codeml under the global clock model with point calibrations to derive a rate estimate to use as the prior mean for the overall rate parameter (rgene_gamma = 0.8855 1). The rate drift parameter was "sigma2_gamma = 1 1" and the parameters of the birth-death process were "BDparas = 1 1 0". Markov chain Monte Carlo was sampled every 1000th generation until ESS values were higher than 200 (after removing the burn-in period). The analysis was carried out twice to check for convergence of the chains.

Supplementary Material

Supplementary information is available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank many reviewers for helpful comments on previous versions of this manuscript. This research was supported by grants from National Aeronautics and Space Administration (NASA, NNX16AJ30G) to SK, Tokyo Metropolitan University (DB105) to KT, and the Brazilian Research Council (CNPq, 233920/2014-5) to BM.

References

Antcliffe JB, Callow RHT, Brasier MD. 2014. Giving the early fossil record of sponges a squeeze. *Biol Rev Camb Philos Soc*. 89:972–1004.

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

Battistuzzi FU, Billing-Ross P, Paliwal A, Kumar S. 2011. Fast and slow implementations of relaxed-clock methods show similar patterns of accuracy in estimating divergence times. *Mol Biol Evol*. 28:2439–2442.

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

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

Cannon JT, Vellutini BC, Smith J, Ronquist F, Jondelius U, Hejnol A. 2016. Xenacoelomorpha is the sister group to Nephrozoa. *Nature* 530:89–93.

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

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

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

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SY, Faircloth BC, Nabholz B, Howard JT, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346:1320–1331.

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

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

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

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

Love GD, Grosjean E, Stalvies C, Fike DA, Grotzinger JP, Bradley AS, Kelly AE, Bhatia M, Meredith W, Snape CE, et al. 2009. Fossil steroids record the appearance of Demospongiae during the Cryogenian period. *Nature* 457:718–721.

Meredith RW, Janečka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simão TL, Stadler T, et al. 2011. Impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science* 334:521–524.

Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science* 346:763–767.

Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, Moreira MA, Kessing B, Pontius J, Roelke M, Rumpler Y, et al. 2011. A molecular phylogeny of living primates. *PLoS Genet*. 7:e1001342.

Prum RO, Berv JS, Dornburg A, Field DJ, Townsend JP, Lemmon EM, Lemmon AR. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* 526:569–578.

Pyron RA. 2014. Biogeographic analysis reveals ancient continental vicariance and recent oceanic dispersal in amphibians. *Syst Biol*. 63:779–797.

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

Dos Reis M, Inoue J, Hasegawa M, Asher RJ, Donoghue PC, Yang Z. 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, Thawornwattana Y, Angelis K, Telford MJ, Donoghue PC, Yang Z. 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, Yang Z. 2011. Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. *Mol Biol Evol*. 28:2161–2172.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol*. 61:539–542.

Sanderson MJ. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19:301–302.

Smith SA, O'Meara BC. 2012. treePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* 28:2689–2690.

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

Tarver JE, dos Reis M, Mirarab S, Moran RJ, Parker S, O'Reilly JE, King BL, O'Connell MJ, Asher RJ, Warnow T, et al. 2016. The interrelationships of placental mammals and the limits of phylogenetic inference. *Genome Biol Evol*. 8:330–344.

To TH, Jung M, Lycett S, Gascuel O. 2015. Fast dating using least-squares criteria and algorithms. *Syst Biol*. 65:82–97.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol*. 18:691–699.

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

Yang Z. 2014. Molecular Evolution: A Statistical Approach. Oxford University Press, Oxford, 229.