Undersampling Genomes has Biased Time and Rate Estimates Throughout the Tree of Life

Julie Marin*[,1,2] and S. Blair Hedges[1]

[1]Center for Biodiversity, 502 SERC Building, Temple University, Philadelphia, PA

[2]Institut de Systématique, Evolution, Biodiversité UMR 7205, Département Systématique et Evolution, Muséum national d'Histoire naturelle, Sorbonne-Universités, Paris Cedex 05 75231, France

***Corresponding author:** E-mail: juliemarin46@gmail.com.

Associate editor: Koichiro Tamura

Abstract

Genomic data drive evolutionary research on the relationships and timescale of life but the genomes of most species remain poorly sampled. Phylogenetic trees can be reconstructed reliably using small data sets and the same has been assumed for the estimation of divergence time with molecular clocks. However, we show here that undersampling of molecular data results in a bias expressed as disproportionately shorter branch lengths and underestimated divergence times in the youngest nodes and branches, termed the small sample artifact. In turn, this leads to increasing speciation and diversification rates towards the present. Any evolutionary analyses derived from these biased branch lengths and speciation rates will be similarly biased. The widely used timetrees of the major species-rich studies of amphibians, birds, mammals, and squamate reptiles are all data-poor and show upswings in diversification rate, suggesting that their results were biased by undersampling. Our results show that greater sampling of genomes is needed for accurate time and rate estimation, which are basic data used in ecological and evolutionary research.

Key words: **diversification rate, molecular clock, small sample artifact, speciation rate, timetree.**

Introduction

The evaluation of evolutionary rates has taken a central place in evolutionary biology. These include micro-evolutionary rates such as those of nucleotide and amino acid substitution (Nei and Kumar 2000) as well as macro-evolutionary rates such as those of speciation, extinction, and diversification (speciation minus extinction) (Ricklefs 2007). All of these rates rely on estimates of molecular change (number of substitutions per site per unit time) being reliable and unbiased, whether the estimation of nucleotide substitution is the target of analysis or the data for the estimation of time. However, recently it was shown that rates can be biased when the number of variable sites is too small, causing underestimates of time and leading to an upturn in speciation, or diversification, rate towards the present (Hedges et al. 2015, Marin et al. 2017).

This "small sample artifact" is caused by the reduction of signal and increasing coarseness of the data available to estimate low sequence divergence values. It is a sampling bias that stems from the fact that any given study usually only samples (once) a small part of the entire genome. As zero is approached in true sequence divergence, it is then increasingly likely that rare variable sites will be omitted, underestimating divergence. The bias is negligible in the largest data sets and for the highest divergence estimates (deepest nodes) in most trees, but it is likely to be evident in small data sets or as estimates approach zero (shallow nodes) in any data set.

Many studies with large taxonomic coverage, including widely used tetrapod timetrees, have reported a rate increase towards the present and many have ascribed biological significance to it (Bininda-Emonds et al. 2007; Pyron and Wiens 2011; Jetz et al. 2012; Pyron et al. 2013; Claramunt and Cracraft 2015; Nürk et al. 2015) (fig. 1). Considering that the small sample artifact can produce the same pattern, and that most of these studies had large fractions of missing data and a relatively small number of variable sites, we raise the possibility here that those rate increases were artifacts, not of biological significance. If true, the hundreds of evolutionary studies that have used the results of those major tetrapod studies, in turn, may have been impacted by the artifact.

In two previous studies focused on building a timetree of eukaryotes and prokaryotes (fig. 1), we drew attention to this artifact (Hedges et al. 2015) and conducted analyses to understand it (Marin et al. 2017) using the time estimation method RELTIME (Tamura et al. 2012). Here, we present a more expanded study, taxonomically and methodologically, to further explore the scope and implications of the small sample artifact, with the widely used estimation method BEAST (Drummond et al. 2012). We evaluate the pairwise-distances and the evolutionary rate patterns of simulated and empirical data-poor and data-rich timetrees. Finally, we review the large empirical studies of tetrapods to determine if the results of those studies could be alternatively, or at least in part, explained by the small sample artifact. We conclude that this statistical artifact probably has had a wide impact across evolutionary and ecological studies that made use of

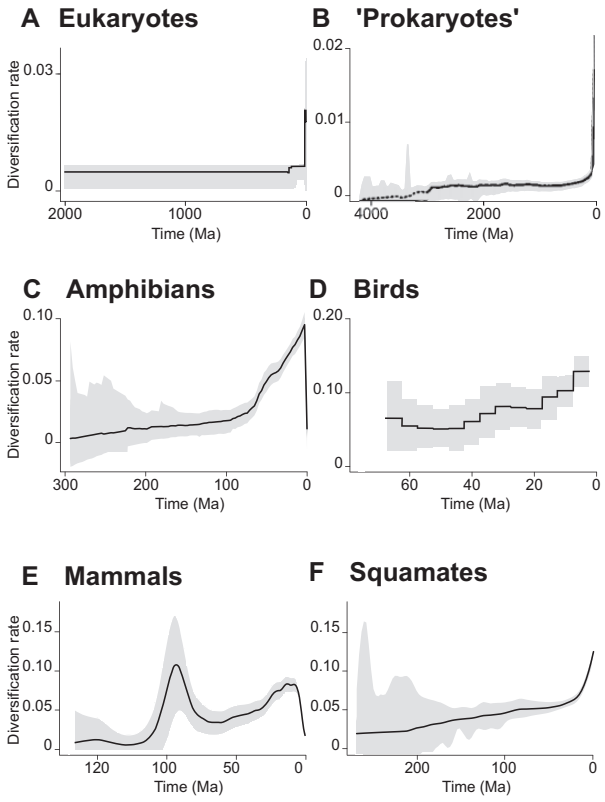

FIG. 1. Evolutionary rate increases in eukaryotes. Diversification rate through time of (*A*) eukaryotes (modified from Hedges et al. 2015), (*B*) "prokaryotes" (modified from Marin et al. 2017), (*C*) amphibians (timed from Pyron and Wiens 2011), (*D*) birds (modified from Jetz et al. 2012), (*E*) mammals (modified from Bininda-Emonds et al. 2007), and (*F*) squamates (timed from Pyron et al. 2013). The gray area represents the 95% confidence interval.

molecular data to estimate divergence times, tree branch lengths, and rates of molecular change.

Results

Simulations

We simulated alignments of different lengths, corresponding to birth-death model trees of different sizes, using model parameters derived from empirical vertebrate sequence data (Supplementary Materials). We estimated divergence times with the program BEAST v. 1.8.3. (Drummond et al. 2012) from the simulated alignments. Theoretically, a constant speciation or diversification rate through time is expected under a birth-death model. However, we detected increasing speciation rate through time in our simulations. Those data sets with the fewest variable sites, the most missing data, and greatest number of tips led to the strongest bias (fig. 2*A*–*C*, and supplementary fig. S1, Supplementary Material online) even when constraining the topology (supplementary fig. S2, Supplementary Material online). For a 100-tip phylogeny, an average of ~410 variables sites over 1,000 sites for the ingroup was enough to correctly estimate speciation rate, that is, with a deviation in speciation rate from the reference timetree equal to zero (supplementary fig. S1, Supplementary Material online). A lower number of variable sites (~81 variables sites on average over 200 sites for the ingroup) resulted in increasing speciation rate towards the present and a wider confidence interval. On the other hand, ~413 variable sites over 1,000 sites was not enough to correctly estimate speciation rate of a larger phylogeny with 500 tips (supplementary fig. S1, Supplementary Material online). A larger number of variable sites (~830 on average over 2,000 sites for the ingroup) was needed to correctly estimate speciation rate for a 500-tip tree. For a 1,000-tip phylogeny, an average of ~1,225 variable sites over 3,000 sites for the ingroup was needed to correctly estimate speciation rate (fig. 2*A*). Fewer variable sites (~626 variable sites on average over 1,500 sites for the ingroup) resulted in an increasing speciation rate and a larger confidence interval (fig. 2*B*). Moreover, missing data influenced speciation rate estimates in the same way as data-poor alignments (~934 variable sites on average over 3,000 sites for the ingroup) (fig. 2*C*).

The comparison of timetree branch length variation between the simulated and model timetrees revealed shorter branches for data-poor and data-missing timetrees all along the trees for unconstrained (fig. 2*D*–*F*) and constrained timetrees (supplementary fig. S3, Supplementary Material online). For data-missing timetrees, we found a more scattered distribution, with both underestimated and overestimated node times along the tree (fig. 2*F*–*I*), but the major trend was clearly towards underestimated times as shown by the fitted curves (fig. 2*F*). We also found that the program BEAST introduces a small amount of sequence change, between 2.8e−04 and 1.2e−03, to branches (supplementary fig. S4, Supplementary Material online), explaining the pairwise difference increase (fig. 2*D*–*F*) and hence the speciation upturn attenuation near zero (fig. 2*B* and *C*). This is not explained in the BEAST manual, but is probably implemented to avoid zero-length branches, and also the result of an arbitrary resolution of polytomies in Bayesian analyses (Lewis et al. 2005). For deep nodes, impact of the "BEAST artifact" is small but it is proportionately larger in shallow nodes, where it causes deviations amounting to hundreds of percent. Therefore, in BEAST analyses, both statistical underestimation (small sample artifact) and program-induced overestimation (BEAST artifact) of sequence divergence occurs, with the strength and location of each bias depending on the number of variable sites (relative to the number of tips) and on the depth of nodes.

To further compare the influence of molecular sampling on divergence time estimates, we fitted polynomial equations to data-rich and data-poor pairwise distance proportions for unconstrained (fig. 2*D*–*F*) and constrained trees (supplementary fig. S3, Supplementary Material online). There was a significant difference in their fit to the data in each case, with data-poor analyses showing a stronger bias than data-rich analyses, as expected. The data-rich distributions were better-fitted by the corresponding data-rich polynomial equations for unconstrained (difference between the two correlations: P value = 0; $z = 45.86$) and constrained trees (P value = 0.02; $z = 2.26$). Similarly, the data-poor distributions were better fitted by the corresponding data-poor polynomial equations for unconstrained (P value = 0; $z = 56.65$) and constrained trees (P value = 0.01; $z = 2.53$). The jackknife

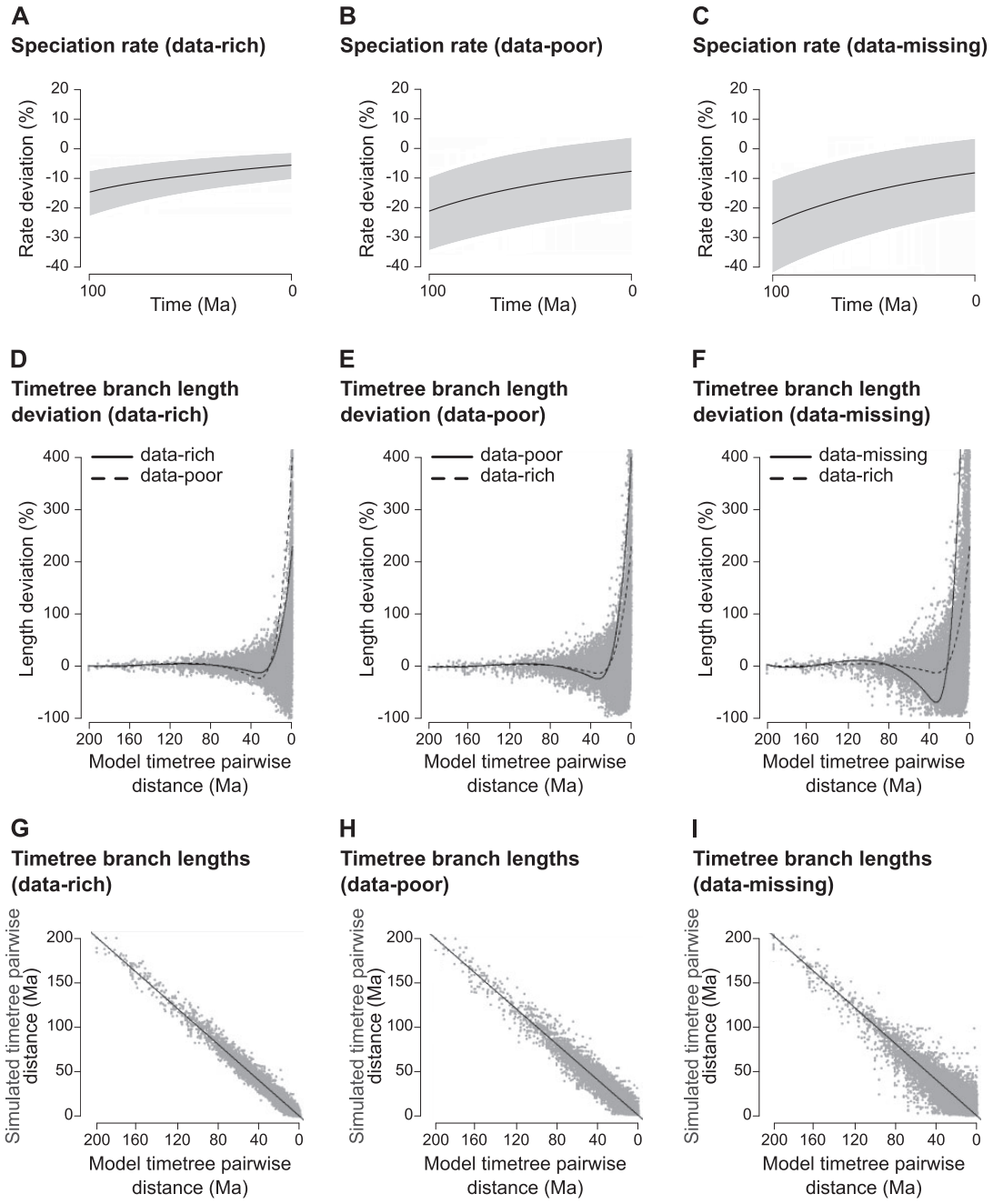

Fig. 2. Biases in divergence time and speciation rate estimation from undersampling of genomes. Ten alignments of 3,000 nucleotides (data-rich) and 1,500 nucleotides (data poor) were simulated from 1,000-tip timetrees (model timetrees) and used to estimate divergence times with the program BEAST (*A*, *B*, *D*, *E*, *G*, and *H*). We created data-missing alignments of 3,000 nucleotides by randomly replacing 70–100% of the nucleotides of 2,383 sites over 3,000 by missing nucleotides (*C*, *F*, and *I*). (*A*–*C*) Deviation of speciation rate through time of 1,000-tip simulations timed with BEAST. Percentage of rate deviation was estimated by subtracting the speciation rate of the reference timetree from the speciation rate of the timetrees built with different sequence lengths. The gray area represents the 95% confidence interval for each set of simulations. (*D*–*F*) Deviation percentage of branch length pairwise distance differences between unconstrained timetrees and model timetrees. The red lines correspond to the fitted polynomial line on data-rich differences and the blue lines correspond to the fitted polynomial line on data-poor differences and on data-missing differences. (*G*–*I*) Branch length pairwise distances comparison with the model timetree of data-rich, data-poor, and data-missing unconstrained timetrees. The black line represents the regression line through the origin.

procedure confirmed these results with significant differences between the two correlation values in 100% of the cases for unconstrained data-rich and data-poor distributions, and in 64% and 63% of the cases for the constrained data-rich and data-poor distributions, respectively.

However, estimated node ages deviate from the true times in data-rich trees (fig. 2*G*) even though they do not deviate in speciation rate (fig. 2*A*). This might be the consequence of the binning process of speciation rate analyses, or due to the stochastic models underlying these analyses. Indeed, different

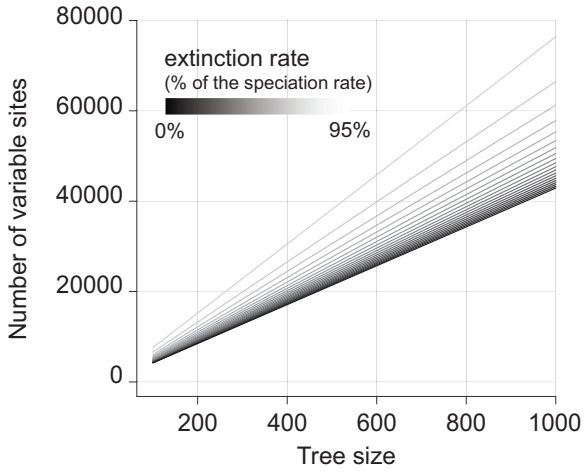

FIG. 3. Recommended number of variable sites depending on tree size. Lines represent the recommended number of variable sites for trees with an extinction rate equal to 0 (Yule tree; dark gray) to trees with an extinction rate equal to 95% of the speciation rate (light gray), spaced every 5%.

time estimates of nodes can still be in agreement with a constant speciation rate through time. For most research applications, that is a concern because accurate estimation of node age is important, regardless of whether speciation rate analysis is conducted. Based on our simulation results, and on terminal branch length and tree length formulas (Mooers et al. 2012), we find that at least 20 variable sites are needed to correctly estimate 95% of the terminal branch lengths (pendant edges), regardless of the speciation and extinction rates. Also the total number of variable sites needed to estimate all node ages in a tree increases with tree size and with the proportion of extinction rate compared with speciation rate (fig. 3). For example, ~50,000 variable sites are needed to estimate node ages of an 800-tip tree with an extinction rate equal to 85% of the speciation rate.

Empirical Studies

As a further test of the existence of the small sample artifact, we differentially sampled a vertebrate data-set and found an increasing speciation rate in data-poor timetrees (supplementary fig. S5, Supplementary Material online). Truncating the data set resulted in an increase in speciation rate towards the present compared with the complete, four-gene data set. Finally, we compared our simulation results with diversification plots from large, global data-sets of tetrapods that used thousands of species (Pyron and Wiens 2011; Jetz et al. 2012; Pyron et al. 2013), all containing upturns in diversification rate towards the present (fig. 1). One study was composed of a backbone and 129 clades of birds (Jetz et al. 2012). On average, the clades were defined by 774 variables sites for 154 species (table 1). Over the 26 clades analyzed containing >100 species, and after removing sites with >30% missing data, 27% of the clades did not have enough variable sites to avoid the small sample artifact according to our simulations (supplementary table S1, Supplementary Material online). For the two other large data-sets comprising 2,872 species of amphibians (Pyron and Wiens 2011) and 4,162 species of squamates (Pyron et al. 2013), only 771 and 0 variable sites were left, respectively, after removing sites with >30% missing data (table 1). Because the mammal tree (Bininda-Emonds et al. 2007) (fig. 1) is a combination of supertrees we were unable to evaluate this data set in an equivalent way. However, the data-set used for molecular dating strongly suggests that the mammal timetree was also influenced by the small sample artifact. For example, over the 68 molecular markers used, 24 did not have enough sites according to our simulations (supplementary table S2, Supplementary Material online), suggesting that most of this mammal data set (Bininda-Emonds et al. 2007) is susceptible to the small sample artifact.

Discussion

Divergence time estimates are used for many purposes (Hedges and Kumar 2009) and change in diversification rates are routinely interpreted as an adaptive response to changes in the environment, including the presence of other organisms. However, our results show that an insufficient number of variable sites could cause the underestimation of divergence times leading to artificial upturns in speciation (or diversification) rates. Given that the most species-rich studies regularly used to infer rate patterns show upturns in diversification rates (fig. 1), whereas harboring an insufficient number of variable sites according to our results, the small sample artifact likely has had wide-ranging impact across evolutionary biology. For example, the four major studies on tetrapod vertebrates alone (Bininda-Emonds et al. 2007; Pyron and Wiens 2011; Jetz et al. 2012; Pyron et al. 2013; fig. 1) have been cited 4,238 times.

Tree nodes are resolved by different combinations of variable sites, with the most recent nodes having the fewest variable sites. Because there were not enough variable sites to correctly infer substitution rates of data-poor trees, the shallowest branch lengths, in particular, were underestimated (fig. 2*D*–*F*). Consequently, node divergences were younger for data-poor timetrees (fig. 2*G*–*I*) explaining the artificial increase in speciation rate towards the recent (fig. 2*A*–*C*). Constraining the topology reduced the bias but did not eliminate it (supplementary figs. S2 and S3, Supplementary Material online), hence the upturn in speciation rate towards the present, as seen in data-poor timetrees, is an artifact that can be amplified if it also results in topological errors. Moreover, the small sample artifact could be the explanation for higher molecular rates (substitutions per site per unit of time) of data-poor trees observed in ancient DNA studies and others involving recent divergences (Debruyne and Poinar 2009), because of shorter branches and hence underestimates of time.

Elsewhere we have shown (Hedges et al. 2015; Marin et al. 2017) that another widespread bias, the taxonomic artifact, leads to a downturn in speciation, or diversification, rate towards the present in studies with incomplete sampling of lineages, regardless of the number of variable sites (fig. 4). This artifact likely explains the downturn in diversification

Table 1. Number of Tips and Variable Sites of Empirical Studies Showing Increasing Diversification Rate Through Time.[a]

Study	Number of Tips	Number of Variable Sites/Total Number of Sites			
		Full Alignment	<70% Missing Data	<50% Missing Data	<30% Missing Data
Jetz et al. (2012)[b]	154	2530/6744	1787/3800	1201/2270	774/1448
Pyron and Wiens (2011)	2872	9018/12712	2493/2615	1349/1410	771/822
Pyron et al. (2013)	4162	10502/12896	3353/3568	618/656	None

[a]The percentage of missing data corresponds to alignments for which sites are defined by at most this percentage of ambiguous sites.
[b]Jetz et al. (2012) built and dated a backbone and 129 clades. We analyzed the 26 data sets with >100 tips (supplementary table S1, Supplementary Material online) and reported the mean results here.

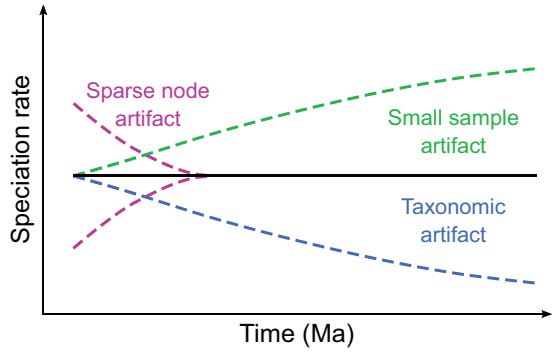

Fig. 4. Potential biases on speciation rate estimation. The thick continuous line represents constant speciation rate as expected if there are no artifacts or other factors affecting the rate. The small sample artifact (an insufficient number of variable sites) may impact all of the tree and diversification plot, resulting in a rate increase towards the present. The taxonomic artifact (incomplete sampling of taxa or lineages) also may impact all of the tree and diversification plot and results in a speciation rate decrease towards the present. The sparse nodes artifact (stochastic effect of a limited number of nodes) may impact the beginning of the diversification plot, causing decreases or increases in rate.

rate near zero observed in studies of amphibians (fig. 1*C*), mammals (fig. 1*E*), and eukaryotes (fig. 1*A*). Given the bird lineage trough time plot from Jetz et al. 2012, it was also likely present in the diversification rate plot for birds (fig. 1*D*) but those authors (Jetz et al. 2012) omitted the plot for the last 2.5 My, giving the reason as "difficulty of accounting for ongoing speciation events." The taxonomic artifact is a common feature in diversification studies because of the difficulty in sampling all individuals, populations, and species. For example, even in a complete species-level study, the absence of any lineage splits below the species level will be expected to cause a downturn in diversification rate towards the present, and this bias will penetrate back into the timetree for several million years or much more depending on the taxonomic rank sampled (Hedges et al. 2015). Finally, the sparse node artifact (fig. 4) occurs when tools incorrectly attribute statistical significance to deviations in speciation rate even when they are based on node counts as few as two, as we demonstrated elsewhere (Marin et al. 2017). There are also user-induced artifacts, such as wide branch length priors leading to topological errors (Yang and Rannala 2005), or incorrectly applied calibration constraints or topologies (Springer et al. 2013), that both could cause artificial excursions in molecular and speciation rates. But we focus here instead on largely unrecognized artifacts of commonly used data sets and methods.

Considering all of these potential biases, it is evident that nearly any change in speciation or diversification rate, in any given study regardless of data size, could be an artifact (fig. 4). This implies that many past studies have probably been impacted beyond errors in diversification results because timetrees are commonly used in biogeography, macro-evolution, and ecological studies. If many of the speciation rate upswings and downswings in past studies are artifactual, it may also mean that evolution is more constant and clock-like than previously thought (Hedges et al. 2015; Marin et al. 2017), something that future studies can test with larger data sets. Moreover, the small sample artifact has broader implications than just speciation or diversification rate, because it derives from biased timetrees having overestimates of node ages, particularly in the shallowest nodes. This would explain the prominent spike in speciation rate, near the recent, in the consensus timetree of life derived from thousands of published timetrees (Hedges et al. 2015). It also indicates that conclusions drawn from previous time estimation studies, especially of recent (Cenozoic) events, may need to be revisited. In addition, if calibrations were applied to biased node ages then other times in the timetree would, in turn, be biased.

For instance, the Cenozoic speciation rate increase in birds has been linked to key morphological and behavioral innovations or environmental opportunities (Jetz et al. 2012). Similarly for mammals, faunal and environmental changes have been invoked to explain a rate increase towards the present (Bininda-Emonds et al. 2007). However, those conclusions may not be justified given our simulation results whereby the small sample artifact created a large 78% rate deviation in trees with similar parameters as in those two empirical studies. For example, roughly 60% and 50% of the rate increase detected in birds and mammals, respectively, calculated between 60 Ma and present-day, could be explained by the small sample artifact alone.

Recommending a minimum number of variable sites to accurately estimate speciation rate is difficult because it scales among nodes in the tree, with deeper nodes having a greater number of variable sites. If the outgroup is distant, the basal node may have nearly all of the variable sites. In addition, missing data can lead to a complex patchwork of nodes, each with insufficient sampling of molecular data (Filipski et al.

2014). Branch length heterogeneity also might affect the accuracy of branch length estimation (Schwartz and Mueller 2010). Assuming a tree following a birth-death pattern of evolution, and given our simulations, at least 400 variable sites are needed to estimate speciation rate in 100-tip trees, and 1,200 variable sites are needed for 1,000-tip trees. These numbers of variable sites are typically found in data sets of 1,000–4,000 sites, although any type of nonstandard sampling (e.g., missing data, high numbers of shallow tips) could require ten or more times that number of sites. In addition, such recommendations are only for speciation rate.

A much larger number of variable sites is needed for accurate time estimation, and this is probably the major use of any given timetree. So rather than relying on a universal cutoff of sites in an alignment, it is best to verify that the shallowest nodes have a sufficient number (e.g., >20) of variable sites or nucleotide differences for precise estimates, and to avoid the small sample artifact. Moreover, because in practice it is difficult to assess the number of variable sites for pairs of terminal branches, we also provide here the recommended number of variable sites for the whole alignment (fig. 3). However, because a large fraction of the sites, between 30% and 80% (unpublished data), are usually invariant, the total sequence length should be roughly twice as long as the recommended number of variable sites. Additionally, diversification analysis tools, such as TreePar (Stadler 2013) and BAMM (Rabosky et al. 2014), need to be improved by incorporating node-based statistical error, which decreases with increasing data size (Filipski et al. 2014). This need for improvement applies also to the sparse-nodes artifact (fig. 4). Other methods of analyzing diversification rate change (Morlon et al. 2010) are not immune to these artifacts. Furthermore, polytomies can cause spikes in rate and be interpreted as real evolutionary events, as did happen, for example, with a 30–33 My speciation rate spike in mammals (Stadler 2011). Also, polytomies can be resolved arbitrarily with very high posterior probabilities in Bayesian analyses, leading to erroneous interpretation of lineage relationships, a problem that can be overcome with the modification of the Metropolis-Hasting algorithm (Lewis et al. 2005).

The increasing availability of genome-scale data sets will lessen the impact of the small sample artifact in future studies. Nonetheless, a greater taxonomic coverage requires a corresponding increase in the number of variable sites that may not be satisfied even in the largest phylogenomic data sets, and especially for recent divergences of hundreds or thousands of years. Indeed, our results show that a change in experimental design is needed in molecular phylogenetic studies. Previously, it was assumed that a data set used for phylogeny was sufficient for time estimation and speciation rate analysis. Now it is evident that these three different analyses each have different data set requirements. A small data set of one or a few genes may be sufficient to build a phylogeny with high support values on nodes, even when the data set has missing data. However, a much larger and cleaner data set is needed for accurate time estimates (fig. 3), especially of the shallow nodes. We found that the data size needs of speciation rate analysis are intermediate between those of phylogeny and time estimation, probably because of the stochastic flexibility of the underlying models in speciation rate analyses, buffering individual node age variation. For practical reasons, researchers will likely choose to sample the genome more broadly for all analyses.

Materials and Methods

Simulations

We simulated three sets of ten timetrees of 100, 500, and 1,000 tips with the function "sim.bdtree" (geiger package; Harmon et al. 2008) (b = 0.6, d = 0.4; 100, 500, and 1,000 tips and one outgroup manually added at 5 My from the ingroup), referred to as the model timetrees. Two alignments of different lengths were simulated from each tree, 200 and 1,000 nucleotides for the 100-tip timetrees, 1,000 and 2,000 for the 500-tip timetrees, and 1,500 and 3,000 for the 1,000-tip timetrees (PhyloSim package; Sipos et al. 2011). To define the model of DNA evolution for the simulations we used the parameters estimated by modeltest (function "modelTest", package phangorn in R; Schliep 2011) on a vertebrate alignment (Marin et al. 2013): model: GTR model; rate matrix: $a = 0.203$, $b = 0.029$, $c = 0.026$, $d = 0.048$, $e = 0.023$, $f = 0.343$; base frequencies: 0.244, 0.232, 0.280, and 0.244); shape parameter α set at 0.4; and proportion of invariable sites set at 0.28. We recorded the number of variable sites for each simulated alignment. From the simulated alignments, the tree construction was performed with RAxML 8.1.11 (Stamatakis 2014), assuming GTR (general time reversible) model with 1,000 bootstrap replicates. The divergence times were estimated with the program BEAST v. 1.8.3. (Drummond et al. 2012). The xml file was created using BEAUTi (v. 1.8.3) (Drummond et al. 2012) with the following parameters: GTR + G+I substitution model, relaxed uncorrelated lognormal clock; a birth-death process to model speciation events; ten million generations with sampling every 1,000 steps. We used the uncorrelated lognormal clock because we found no evidence for autocorrelation in our datasets. The mean covariances of parent and child branches ranged between −0.015 and 0.007, and the HPD (highest posterior density) minimum and maximum intervals ranged between −0.0637 and 0.05 under the lognormally distributed model of rate variation. Moreover, this model is the one traditionally used with the program BEAST. We used one calibration point, ingroup node, at 100 Ma (normal distribution, mean value: 100, standard deviation: 1). During the timing process, the topology was unconstrained using the RAxML tree as the imput tree, or constrained to follow the model tree topology.

We evaluated the relationship between tip pairwise distances of the timetrees and the corresponding model trees. Because the same amount of distance difference could have a limited importance for deep branches, but make a significant difference for shallow branches, we calculated the proportional pairwise distances to the branch lengths of the model tree. We then fitted polynomial curves to the resulting distributions, and compared the correlation coefficients of the data-rich and data-poor fitted curves to the data-rich and

data-poor distributions (paired.r test in R). In order to test the robustness of our results, we performed a jackknife procedure. We sampled, without replacement, 20% of each data-set (data-rich and data-poor proportional pairwise distances) 100 times, and compared, as we did previously, the correlation coefficients of the data-rich and data-poor fitted curves to the data-rich and data-poor distributions.

Because in many data sets the missing data proportion is important (table 1), and because missing data represent undersampling of sites, we evaluated the influence of missing data on speciation rate for the 1,000-tip timetrees. The number of missing sites corresponded to the proportion of missing data of the alignment used in Pyron and Wiens (2011) where 79.4% of the 12,712 sites have between 70% and 100% missing data. For the same proportion in our simulation (3,000 sites), we randomly replaced 70–100% (uniform distribution) of the nucleotides with missing nucleotides (N).

Because an accurately estimated speciation rate does not imply that all underlying node ages are accurately estimated, we also evaluated the number of variables sites required to estimate node ages. For the 1,000-tip model trees we extracted the terminal branch lengths. After discarding the smallest 5% of those branches, we recorded the minimum branch length, representing the minimal length that we want to be able to estimate from the alignment. We assumed that one variable site is needed to estimate this minimum branch length. Then, proportionally, we first determined the number of variable sites needed to estimate pairs of terminal branch lengths using the formula for average terminal branch length from Mooers et al. (2012). Secondly, we determined the number of variable sites needed to estimate all of the node ages of a tree using total tree length formula from Mooers et al. (2012) while varying tree size, speciation rate, and extinction rate.

Empirical Studies

In order to test the effect of the number of variable sites on real data, we evaluated the speciation rate pattern of a vertebrate data set (Marin et al. 2013) using one or four genes. The data set comprised 107 Australian scolecophidian snakes (Squamata: Serpentes) and 16 outgroups. The four genes that were sequenced were: cytochrome b (cytb) with 432 variables sites among 678, prolactine receptor (PRLR) with 379 variables sites among 483 sites, brain-derived neurotrophic factor (BDNF) with 188 variables sites among 672, and bone morphogenetic protein 2 (BMP2) with 361 variables sites among 591. The tree construction was performed with RAxML 8.1.11 (Stamatakis 2014), assuming GTR (general time reversible) model with 1,000 bootstrap replicates. Following Marin et al. (2013), we treated each codon position of the cytb, BDNF, LRPR, and BMP2 genes as a separate partition so the combined data set included 11 partitions (because of saturation, the third position of cytb was excluded).

Divergence times were estimated by the program BEAST v. 1.8.3. (Drummond et al. 2012). The xml file was created using BEAUTi (v. 1.8.3) (Drummond et al. 2012) with the following parameters: unlinked substitution and clock models, GTR + G+I model, relaxed uncorrelated lognormal clock; a birth-death process to model speciation events; ten million generations with sampling every 1,000 steps. The RAxML tree was used as the input tree and we used seven calibrations (detailed in Marin et al. 2013). The same methodology was followed to reconstruct the scolecophidian timetree on a restricted molecular data set, using only the gene BDNF (105 Australian scolecophidian snakes and 16 outgroups).

Finally, we compared our simulation results with diversification plots from large, global data sets of tetrapods that used thousands of species (fig. 1). Because the eukaryote and mammal timetrees were built from many timetrees by consensus, we focused on the data sets of aligned sequences for amphibians (2,872 species; Pyron and Wiens 2011), birds (9,993 species; Jetz et al. 2012), and squamates (4,162 species; Pyron et al. 2013). The number of variable sites was recorded for the complete alignment and after the removal of sites with <70%, 50%, and 30% missing data. The amphibian and squamate data sets were analyzed and dated previously in Hedges et al. (2015). We estimated here the speciation rate patterns of these three data sets with BAMM (Rabosky et al. 2014) as described above. For the amphibian and the squamate trees, the number of described species (IUCN 2017 and Uetz et al. 2017, respectively) was used to calculate the sampling fraction parameter. The bird timetree is composed of a backbone and 129 clades. We evaluated the number of variables sites and the speciation rate pattern as described above for the 26 clades with >100 species [stage 1 data-sets from Jetz et al. (2012): species with genetic data]. We used the total number of species after the addition of species without genetic data [stage 2 data-sets from Jetz et al. [2012]] to calculate the sampling fraction parameter.

Diversification Analyses

Some concerns have been raised concerning the method BAMM (Moore et al. 2016), although those criticisms were later shown to be unfounded (Rabosky et al. 2017). We used this program to estimate evolutionary rates over time. Extinction rate estimates are not reliable in phylogenies that have diversification rate heterogeneity (Rabosky 2010, 2016). For this reason we only estimated speciation rates over time in our simulations. For the empirical data sets (amphibians and squamates), we estimated both speciation and diversification rate plots and found comparable results (fig. 1 and supplementary fig. S6, Supplementary Material online).

The BAMMtools package and BAMM program (Rabosky et al. 2014) were used to estimate evolutionary rate through time of the simulated and empirical timetrees. The function "setBAMMpriors" was used to generate a prior block that matched the "scale" (e.g., depth of the tree) of our data. Both λ and μ rates were allowed to vary through time and across lineages, and MCMC chains were run for 10,000,000,000 iterations. Because the program does not allow branch lengths equal to zero, we changed the length of these branches to 1e-07. Timetrees were still ultrametric after the change; this change is much smaller than the BEAST artifact. We discarded between 15% and 80% of the MCMC chains to reach the convergence for the timetrees, which was checked

by calculating the effective sample size of the log-likelihood and of the number of shifts events present in each sample that should be over 200 as recommended by the authors of the program. Diversification rate plots were obtained with the function "plotRateThroughTime."

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Author Contributions

J.M. and S.B.H conceived the research; J.M. analyzed the data; and J.M. and S.B.H. wrote the paper.

Acknowledgments

We thank Sayaka Miura and Sudhir Kumar for first pointing out the BEAST artifact, and to Sudhir Kumar and QiQing Tao for stimulating discussion and comments. This work was supported by grants from the U.S. National Science Foundation (grant numbers 1136590, 1455762) to S.B.H. We also thank Arne Mooers for providing us the bird alignments used in Jetz et al (2012).

References

Bininda-Emonds OR, Cardillo M, Jones KE, MacPhee RD, Beck RM, Grenyer R, Price SA, Vos RA, Gittleman JL, Purvis A. 2007. The delayed rise of present-day mammals. *Nature* 446(7135): 507–512.

Claramunt S, Cracraft J. 2015. A new time tree reveals Earth history's imprint on the evolution of modern birds. *Sci Adv*. 1(11): e1501005.

Debruyne R, Poinar HN. 2009. Time dependency of molecular rates in ancient DNA data sets, a sampling artifact? *Syst Biol*. 58(3): 348–360.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol*. 29(8): 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. *Mol Biol Evol*. 31(9): 2542–2550.

Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W. 2008. GEIGER: investigating evolutionary radiations. *Bioinformatics* 24(1): 129–131.

Hedges SB, Kumar S. 2009. Discovering the timetree of life. In: Hedges SB, Kumar S, editors. The timetree of life. New York: Oxford University Press. p. 3–18.

Hedges SB, Marin J, Suleski M, Paymer M, Kumar S. 2015. Tree of life reveals clock-like speciation and diversification. *Mol Biol Evol*. 32(4): 835–845.

IUCN. 2017. IUCN Redlist of Threatened Species. [cited 2017 May 3]. Available from: http://www.iucnredlist.org.

Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO. 2012. The global diversity of birds in space and time. *Nature* 491(7424): 444–448.

Lewis PO, Holder MT, Holsinger KE. 2005. Polytomies and Bayesian phylogenetic inference. *Syst Biol*. 54(2): 241–253.

Marin J, Battistuzzi FU, Brown AC, Hedges SB. 2017. The timetree of prokaryotes: new insights into their evolution and speciation. *Mol Biol Evol*. 34(2): 437–446.

Marin J, Donnellan SC, Hedges SB, Doughty P, Hutchinson MN, Cruaud C, Vidal N. 2013. Tracing the history and biogeography of the Australian blindsnake radiation. *J Biogeogr*. 40(5): 928–937.

Mooers A, Gascuel O, Stadler T, Li H, Steel M. 2012. Branch lengths on birth–death trees and the expected loss of phylogenetic diversity. *Syst Biol*. 61(2): 195–203.

Moore BR, Höhna S, May MR, Rannala B, Huelsenbeck JP. 2016. Critically evaluating the theory and performance of Bayesian analysis of macroevolutionary mixtures. *Proc Natl Acad Sci U S A*. 113(34): 9569–9574.

Morlon H, Potts MD, Plotkin JB. 2010. Inferring the dynamics of diversification: a coalescent approach. *PLoS Biol*. 8(9): e1000493.

Nei M, Kumar S. 2000. Molecular evolution and phylogenetics. Oxford: Oxford University Press.

Nürk NM, Uribe-Convers S, Gehrke B, Tank DC, Blattner FR. 2015. Oligocene niche shift, Miocene diversification—cold tolerance and accelerated speciation rates in the St. John's Worts (*Hypericum*, Hypericaceae). *BMC Evol Biol*. 15(1): 80.

Pyron RA, Wiens JJ. 2011. A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. *Mol Phylogenet Evol*. 61(2): 543–583.

Pyron RA, Frank TB, Wiens JJ. 2013. A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes. *BMC Evol Biol*. 13:93.

Rabosky DL. 2010. Extinction rates should not be estimated from molecular phylogenies. *Evolution* 64(6): 1816–1824.

Rabosky DL, Grundler M, Anderson C, Title P, Shi JJ, Brown JW, Huang H, Larson JG. 2014. BAMMtools: an R package for the analysis of evolutionary dynamics on phylogenetic trees. *Methods Ecol Evol*. 5(7): 701–707.

Rabosky DL. 2016. Challenges in the estimation of extinction from molecular phylogenies: a response to Beaulieu and O'Meara. *Evolution* 70(1): 218–228.

Rabosky DL, Mitchell JS, Chang J. 2017. Is BAMM flawed? Theoretical and practical concerns in the analysis of the multi-rate diversification models. *Syst Biol*. 66(4): 477–498.

Ricklefs RE. 2007. Estimating diversification rates from phylogenetic information. *Trends Ecol Evol*. 22(11): 601–610.

Schliep KP. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* 27(4): 592–593.

Schwartz RS, Mueller RL. 2010. Branch length estimation and divergence dating: estimates of error in Bayesian and maximum likelihood frameworks. *BMC Evol Biol*. 10:5.

Sipos B, Massingham T, Jordan GE, Goldman N. 2011. PhyloSim-Monte Carlo simulation of sequence evolution in the R statistical computing environment. *BMC Bioinformatics* 12(1): 104.

Springer MS, Meredith RW, Teeling EC, Murphy WJ. 2013. Technical comment on "the Placental mammal ancestor and the post-K-PG radiation of placentals". *Science* 341(6146): 613.

Stadler T. 2011. Mammalian phylogeny reveals recent diversification rate shifts. *Proc Natl Acad Sci U S A*. 108(15): 6187–6192.

Stadler T. 2013. TreePar: estimating birth and death rates based on phylogenies. Vienna (Austria): Comprehensive R Archive Network. Available from: https://CRAN.R-project.org/package=TreePar, last accessed June 22, 2017.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9): 1312–1313.

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(47): 19333–19338.

Uetz P, Freed P, Hošek J. 2017. The reptile database. [cited 2017 May 3]. Available from: http://www.reptile-database.org.

Yang Z, Rannala B. 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. *Syst Biol*. 54(3): 455–470.