Molecular Phylogenetics and Evolution 85 (2015) 41–49

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Do missing data influence the accuracy of divergence-time estimation with BEAST?

Yuchi Zheng [a,b], John J. Wiens [b,*]

[a] Department of Herpetology, Chengdu Institute of Biology, Chinese Academy of Sciences, Chengdu 610041, China
[b] Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721-088, USA

ARTICLE INFO

Article history:
Received 8 October 2014
Revised 26 January 2015
Accepted 1 February 2015
Available online 11 February 2015

Keywords:
Accuracy
BEAST
Divergence dating
Fossil calibration
Missing data
Relaxed clock

ABSTRACT

Time-calibrated phylogenies have become essential to evolutionary biology. A recurrent and unresolved question for dating analyses is whether genes with missing data cells should be included or excluded. This issue is particularly unclear for the most widely used dating method, the uncorrelated lognormal approach implemented in BEAST. Here, we test the robustness of this method to missing data. We compare divergence-time estimates from a nearly complete dataset (20 nuclear genes for 32 species of squamate reptiles) to those from subsampled matrices, including those with 5 or 2 complete loci only and those with 5 or 8 incomplete loci added. In general, missing data had little impact on estimated dates (mean error of ~5 Myr per node or less, given an overall age of ~220 Myr in squamates), even when 80% of sampled genes had 75% missing data. Mean errors were somewhat higher when all genes were 75% incomplete (~17 Myr). However, errors increased dramatically when only 2 of 9 fossil calibration points were included (~40 Myr), regardless of missing data. Overall, missing data (and even numbers of genes sampled) may have only minor impacts on the accuracy of divergence dating with BEAST, relative to the dramatic effects of fossil calibrations.

1. Introduction

Assembling molecular datasets for phylogenetic analysis almost always requires dealing with the question of how missing data will influence the analysis, either implicitly or explicitly. For example, choosing to include only those genes with complete sampling among species is often based on the implicit assumption that the negative impacts of missing data will outweigh the positive impacts of increasing the number of genes sampled. Similarly, taxa may be excluded because they may be missing data for some genes, despite the potential benefits of increased taxon sampling. Therefore, it is critically important to understand if, when, and how missing data impact phylogenetic analyses, especially since eliminating missing data from a data matrix generally requires eliminating non-missing data as well (e.g. Jiang et al., 2014).

A growing number of studies have now addressed the potential consequences of missing data for phylogenetic analysis. These include studies that addressed the impact of including incomplete taxa, the impact of incomplete characters, and the impacts of missing data on branch-length estimation and support values (e.g. Wiens, 2003, 2005; Driskell et al., 2004; Philippe et al., 2004; Wiens and Moen, 2008; Burleigh et al., 2009; Lemmon et al., 2009; Sanderson et al., 2010, 2011; Cho et al., 2011; Pyron et al., 2011; Wiens and Morrill, 2011; Crawley and Hilu, 2012; Simmons, 2012, 2014; Wiens and Tiu, 2012; Hovmöller et al., 2013; Roure et al., 2013; Jiang et al., 2014). However, an important and largely unresolved question is how missing data impact divergence-time estimation. Divergence-time estimation has become a fundamental aspect of phylogenetic analysis, especially since many analyses of character evolution, diversification, and biogeography now utilize (or require) time-calibrated trees (e.g. Ricklefs, 2007; Ree and Smith, 2008; FitzJohn, 2010; Quintero and Wiens, 2013). Therefore, understanding how missing data impact divergence-time estimation is increasingly important and urgent.

Two studies have now addressed the impact of missing data on divergence-time estimation, but both were limited in some ways. First, Lemmon et al. (2009) stated that missing data were problematic for divergence-time estimation, although not based directly on analyses of estimated divergence dates. Instead, they investigated whether missing data caused incorrect acceptance or rejection of a molecular clock model in simulations of the four-taxon case. Second, Filipski et al. (2014) more directly analyzed the impact of missing data on divergence-time estimation, focusing on the performance of a relatively new dating method (RelTime; Tamura

* Corresponding author.
E-mail addresses: zhengyc@cib.ac.cn (Y. Zheng), wiensj@email.arizona.edu (J.J. Wiens).

http://dx.doi.org/10.1016/j.ympev.2015.02.002
1055-7903/

et al., 2012) using both simulated and empirical datasets. For example, they compared the accuracy of estimated divergence times using datasets with no missing data, limited missing data (20%), and extensive missing data (60%), using a simulated dataset of 426 taxa. They found little impact of 20% missing data on the accuracy of estimated divergence times, and that the impact of 60% missing data depended on the number of genes (less impact when more genes are sampled, and when at least one "backbone" gene had data in all taxa). These results parallel those on the impact of missing data on phylogeny estimation, in which missing data seem to become problematic primarily when the number of sampled characters is limited (review in Wiens and Morrill, 2011).

Overall, the analysis of Filipski et al. (2014) provided an invaluable contribution to this important question. However, the RelTime method is only one of many methods for analyzing divergence times. Even a cursory examination of the recent literature strongly suggests that the most widely used method for divergence-time estimation is the uncorrelated lognormal relaxed clock method implemented in BEAST (Drummond et al., 2006; Drummond and Rambaut, 2007). Therefore, a critical but unanswered question is how missing data impact divergence-date estimation using this approach.

Furthermore, in some ways, simply assessing how a method performs with extensive missing data is of somewhat limited value, even though this has been the focus of many studies (including some of our own; e.g. Wiens, 2003; Wiens and Moen, 2008). For example, no study is needed to decide whether it is better to have 10 genes for 20 taxa with 50% missing data in 5 of these genes, or the same number of genes and taxa with no missing data. In this situation, there is no reason to prefer the dataset with missing data. What is less clear is whether it is preferable to remove (for example) the 5 genes with missing data, or else include all 10 genes despite their missing data. In other words, the most relevant situation to address is the "gray zone" where eliminating missing data cells from a matrix requires eliminating substantial non-missing data as well (e.g. Jiang et al., 2014).

In this study, we analyze the impact of missing data on divergence-time estimation using the uncorrelated-lognormal relaxed-clock approach implemented in BEAST (Drummond et al., 2006; Drummond and Rambaut, 2007). For brevity, we call this approach "BEAST" hereafter (and in the title), with the understanding that we are referring to this particular relaxed-clock method that the software package BEAST is often used to implement. We specifically focus on the question of whether divergence-time estimates are more accurate using a dataset with many genes but some missing data or with fewer genes but little or no missing data.

We address this question with a large empirical dataset previously published for squamate reptiles (lizards and snakes). We focus primarily on comparing dates estimated from a relatively complete dataset (all genes included, negligible missing data) to those from subsampled datasets with fewer genes, in which some genes have missing data artificially added, or in which these genes with missing data are removed. An obvious disadvantage of using empirical data is that the true phylogeny and dates are not actually known (in contrast to using simulated datasets). However, if we show that missing data have little impact on the estimated dates (and less impact than excluding these incomplete genes), this result would strongly suggest that dating analyses with BEAST are potentially robust to missing data, even if the true dates are unknown.

The disadvantages of using empirical data may be counterbalanced by the fact that divergence dating has many complexities that would be difficult to realistically capture with simulations. For example, a typical dating analysis includes multiple genes, each of which may have somewhat different rates (and different rates in different clades across the tree) and different underlying topologies (e.g. due to incomplete lineage sorting). Perhaps more importantly, divergence dating typically depends on the inclusion of one or more fossil calibration points. It is very unclear how to simulate the distribution of these calibration points in a realistic way, in terms of their number, age, and phylogenetic distribution.

2. Materials and methods

2.1. Design of subsampling experiments

Our subsampling experiments began with the dataset for squamate reptiles from Wiens et al. (2012). This dataset includes 161 species (plus 10 outgroup species) for up to 44 nuclear protein-coding loci (overall concatenated alignment of 33,717 base pairs), with ~20% missing data overall. Most extant families of squamate reptiles were included.

From this dataset, we then created a smaller dataset to use as our complete dataset. The original dataset was reduced in size for two major reasons. First, the reduced dataset allowed us to exclude some taxa that lacked data for many genes and some genes that lacked data for many taxa, so that the complete dataset included data for all species for all genes. Second, the full dataset with all taxa and genes would be computationally burdensome for divergence dating with BEAST, especially with the multiple replicates needed for the experimental analyses conducted here. Even with the reduced taxon sampling, these analyses were still very computationally intensive, and so only 10 replicates for each set of conditions were examined (but similar conditions provided additional replication). We also note that the complete dataset still contained small amounts of missing data (3.1% overall), due to gaps and to minor differences in sequenced lengths of genes in the final alignment, but we use the term "complete" for brevity and because it included data for all genes for all taxa.

We assembled the reduced dataset with 33 taxa (32 ingroup) and 20 genes such that no genes were lacking for any species, and so that all major clades of squamates were included. GenBank numbers, including some corrections relative to Wiens et al. (2012), are provided in Table S1. Not all families of squamates were included in this reduced dataset, but missing families were confined to the major clades Gekkota, Amphisbaenia, Serpentes, and Pleurodonta (within Iguania). The number of taxa (32) was also chosen so that sets of taxa were easily divisible by four for our experiments with missing data (see below). Relationships among the included taxa are also relatively well-supported given the complete data, with only a few poorly supported clades when all loci are included (see Results, Section 3). The included taxa also span a relatively large number of fossil calibration points for dating analyses (see below). One outgroup taxon was included (the rhynchocephalian, *Sphenodon punctatus*), and this species is widely recognized as the living sister group to extant Squamata (e.g. Hugall et al., 2007; Alfaro et al., 2009; Mulcahy et al., 2012). Previous studies have estimated the crown-group of living Squamata to be roughly 180–240 Myr old (review in Mulcahy et al., 2012).

The overall design of the missing-data experiments was as follows. First, we created 10 replicates, each with a different random selection of 10 loci from the original set of 20 loci (sampling without replacement, so that no gene was sampled twice in a given replicate). We then explored five sampling strategies to examine the effects of missing data.

Under sampling strategy 1, 5 of the 10 loci were randomly chosen to be incomplete, and each gene had a different set of species that were randomly chosen to be incomplete. This latter distribution of missing data was intended to mimic the situation in which random sets of taxa fail to amplify and/or be sequenced for each gene. Under this sampling strategy, we explored three different

levels of missing data: (a) missing data in 25% of the taxa (8 randomly selected ingroup taxa have missing data in 5 genes), with 12.5% missing data in the matrix overall; (b) missing data in 50% of the taxa (16 taxa in 5 genes), 25% overall; and (c) 75% of the taxa (24 taxa in 5 genes), 37.5% overall.

Sampling strategy 2 was the same as sampling strategy 1, except that 8 of the 10 loci were randomly chosen to be incomplete, and again each gene had a different set of species randomly chosen to have missing data. Again, we explored three different levels of missing data: (a) missing data in 25% of the taxa (8 randomly selected ingroup taxa have missing data in 8 genes), with 20% missing data overall; (b) missing data in 50% of the taxa (16 taxa in 8 genes), 40% overall; and (c) 75% of the taxa (24 taxa in 8 genes), 60% overall.

Sampling strategy 3 was similar to sampling strategy 1, with 5 loci chosen to be incomplete. However, it differed from strategies 1 and 2 in that instead of randomly selecting different sets of taxa to lack data for each incomplete gene, the same set of species was missing data across all incomplete genes. This sampling strategy was intended to mimic the creation of a supermatrix, in which sets of genes from different studies (which sampled different sets of species for different genes) are combined. For efficiency, we explored this strategy only under conditions with relatively extensive missing data. Thus, we examined the case with missing data in 75% of taxa (24 taxa), and 37.5% missing data overall.

Sampling strategy 4 was similar to sampling strategy 3, but with 8 loci chosen to be incomplete rather than 5. Again, the same randomly selected species were missing data for all incomplete genes, and we focused on the case with extensive missing data, with missing data in 75% of the taxa (24 taxa in 8 genes), and 60% missing data overall.

Under sampling strategy 5, every gene had a different set of species that were randomly chosen to be incomplete. This was intended to mimic the situation explored by Filipski et al. (2014) in which there are no complete "backbone" genes, a situation in which the divergence-dating method that they examined sometimes performed poorly. Here, three different levels of missing data were again explored: (a) missing data in 8 taxa, with 25% missing data in the matrix overall; (b) missing data in 16 taxa, 50% overall; and (c) missing data in 24 taxa, 75% overall. For scenario c, there were initially many replicates in which one or two species were missing data for all genes. In these cases, one gene was randomly selected and the random allocation of missing data was repeated, until every included species had non-missing data for at least one gene.

We analyzed 10 replicates for each of these 11 sets of conditions (each starting from the 10 subsampled sets of 10 loci). Also, for each set of conditions, we compared results including the incomplete genes to results excluding these genes (except for strategy 5, in which all genes are incomplete). For example, for sampling strategies 1 and 3, we analyzed datasets containing only the 5 complete genes, and for sampling strategies 2 and 4, we analyzed datasets including only the 2 complete genes.

2.2. Selection of models and partitions

Prior to conducting the BEAST analyses, we determined the best-fitting combination of partitions and models using Partition Finder version 1.1.1 (Lanfear et al., 2012). The best-fitting model was determined using the Bayesian Information Criterion (BIC). Branch lengths were linked across partitions. The set of models was restricted to those available in BEAST. The greedy search option was used. This analysis was conducted on the 20-locus dataset and the 10 complete 10-locus datasets. The partitions and models selected are listed in Table S2.

2.3. Dating analyses

Dating analyses were conducted with BEAST version 1.8.0 (Drummond and Rambaut, 2007). Monophyly of the ingroup was constrained (and note that the outgroup had data for all genes). The uncorrelated lognormal model was used to describe the relaxed clock. A gamma prior (shape = 0.001, scale = 1000) was used for the mean of the branch rates. The standard Yule speciation process was specified for the tree prior. Clock models and topologies of individual data partitions were linked, whereas substitution parameters were unlinked across partitions. A few sequence ambiguities that were potentially the result of heterozygosity were included as such (i.e. with the setting "useAmbiguities = true").

A total of nine fossil calibration points were used (Table 1; Fig. 1), taken from those used by Mulcahy et al. (2012; see that paper for discussion and original references). For each calibration point, we used a lognormal prior, with the mean (in real space) of 5, the standard deviation set to 1, and an offset value equal to the minimum calibration age of the fossil. This combination of values allowed for the possibility that the actual age of the calibrated node was substantially older than the age of the oldest known fossil represented by the fossil calibration point (but most likely were only slightly older). Thus the 95% prior density interval extended roughly 15 million years before the minimum age of each fossil (Table 1). We recognize that other authors might prefer somewhat different options with regards to these settings, or even with regards to specific calibration points. However, the important point for our study was the comparison of estimated dates with and without missing data.

For each replicate and set of conditions, we ran four independent searches. For each search, the Markov chain was run for 300 million generations and sampled every 10,000 generations. Results of the four independent runs were then compared in Tracer version 1.5 (Rambaut and Drummond, 2007) to ensure that the chains were converging and mixing adequately. Then, results from the last 90% or 80% of the sampled generations from of each of the four runs were combined to achieve the recommended adequate effective sample size (>200; Drummond et al., 2006).

For the complete 20-locus dataset and some 10-locus datasets, the overall likelihood was unstable, apparently due to an overparameterized substitution model. These cases were resolved by replacing GTR models with HKY models and/or removing the invariable sites parameter of the substitution model (since this parameter partially overlaps the gamma parameter for among-site rate variation). Nevertheless, for each of the 10-locus replicates, the same combination of partitions and models was applied to all datasets (i.e. the same partitions and models were applied, both with and without missing data). However, some empty partitions (i.e. only missing data cells) were removed for the 5 and 2 locus datasets.

2.4. Evaluation of impacts of missing data

In this study, our primary interest was in the impacts of missing data on divergence-time estimation. Specifically, we tested whether these estimates were more accurate if one includes genes with missing data or excludes these genes and reduces the overall number of genes in the analysis. We also evaluated the impacts of missing data on the width of the posterior density intervals (i.e. the precision of these date estimates), and on the topology and support values (posterior probabilities) estimated using BEAST (in which topology and divergence times are often estimated simultaneously). Thus, we addressed the possibility that including genes with missing data might lead to much wider posterior density intervals, poor support values, and inaccurate estimates of topology.

Table 1
Fossil calibrations used for estimating divergence dates with BEAST, with time in millions of years.

Node in Fig. 1	Minimum date	Median (95% HPD)	Fossil calibration	Calibration number in Mulcahy et al. (2012)
C1	54.0	57.0 (54.4–75.5)	Gekkotan *Yantarogekko*	5
C2	70.0	73.0 (70.4–91.5)	*Contogenys*, *Sauriscus*	14
C3	65.2	68.2 (65.6–86.7)	*Konkasaurus*	6
C4	70.0	73.0 (70.4–91.5)	*Chamops*, *Haptosphenus*, *Letpochamops*, *Meniscognathus*	7
C5	92.7	95.7 (93.1–114.2)	*Coniophis*	9
C6	70.0	73.0 (70.4–91.5)	Priscagamines, iguanines, *Isodontosaurus*	12
C7	70.0	73.0 (70.4–91.5)	*Palaeosaniwa*, *Telmasaurus*, *Cherminotus*	10
C8	99.6	102.6 (100.0–121.1)	*Primaderma*	13
C9	70.0	73.0 (70.4–91.5)	*Odaxosaurus*	11

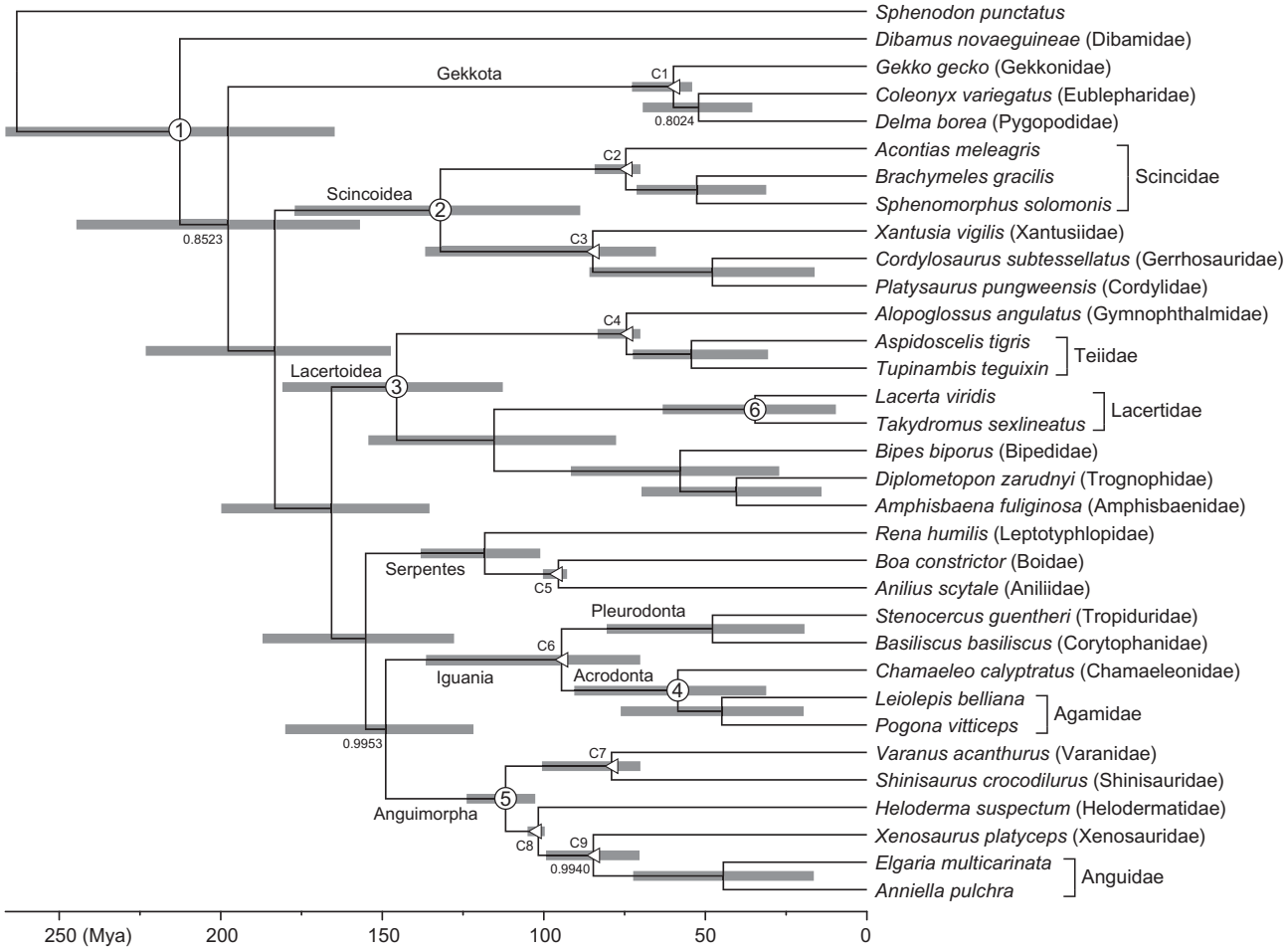

Fig. 1. Chronogram for 32 species of squamate reptiles and an outgroup (*Sphenodon punctatus*) estimated using the uncorrelated-lognormal approach in BEAST, based on 20 complete nuclear loci. Nodes indicated with open triangles and numbers (C1–C9) were associated with fossil calibration points (see Table 1). Nodes with numbers (1–6) in open circles were used to estimate errors in divergence-date estimates in a subset of analyses (most analyses used all nodes). Bayesian posterior probabilities less than 1 are shown beside nodes. The gray bars indicate 95% highest posterior densities for age estimates.

To assess the impact of missing data on divergence times, for each comparable node and each set of conditions and each replicate, we recorded the divergence date estimated from the analysis including the incomplete genes and that estimated for the same conditions but with the incomplete genes deleted. We then calculated the difference between the age for that node estimated from the complete dataset (all 32 taxa with data for all 20 genes) and the age estimated from the analysis including incomplete genes (and the analysis excluding these genes). We considered these differences in ages between the complete dataset (20 gene) and all others as errors in divergence-date estimation, either caused by including incomplete genes or excluding incomplete genes. We then calculated the average values of these errors across all nodes of the tree and then across all 10 replicates for each set of conditions. We calculated these errors as both mean differences from the complete dataset age estimates, and as absolute values of the differences. Thus, we evaluated whether missing data caused both error and bias in age estimates (large mean and absolute differences) or increased error but without bias (i.e. small mean differences, but large absolute differences), and whether these biases and errors were greater in magnitude than those caused by excluding these genes entirely. Note that because we summarized these error values by averaging across both nodes and replicates, we did not present standard errors on these values (or test for statistical differences between the means of means).

An important detail here is that all estimates of topology were generally very similar to those estimated from the 20-gene dataset (Fig. 1). Specifically, for almost all conditions, estimated trees had 85–95% of their nodes shared with the tree from the 20-gene dataset (see Results, Section 3). This was important because estimated dates can generally be compared only between nodes that appear in both trees. Nodes that were estimated in the subsampled datasets that were not present in the complete dataset were excluded from our summaries of mean age differences. We acknowledge that in theory, the estimated ages for these seemingly incorrect nodes could be inaccurate, and their exclusion might therefore cause us to underestimate errors in estimated ages overall. However, the number of nodes shared with the complete dataset is actually higher for datasets that include genes with missing data (see Results, Section 3), so the error caused by including these incomplete genes seems unlikely to be strongly underestimated by this potential source of bias, at least relative to the errors caused by excluding these genes. Similarly, we also excluded nodes in the subsampled datasets that had posterior probabilities less than 0.5. However, this set of excluded nodes was almost identical to the set of nodes already excluded because of incongruence with the complete dataset.

For most analyses, we summed these errors across all nodes of the entire tree. However, we did this with the caveat that the values at different nodes are not necessarily independent. Therefore, we also generated a small set of results focused on six focal nodes (Fig. 1): (1) root of squamates, (2) root of Scincoidea (Scincidae, Xantusiidae, Cordylidae, Gerrhosauridae), (3) root of Lacertoidea (Gymnophthalmidae, Teiidae, Lacertidae, Amphisbaenia), (4) root of Acrodonta (Chameleonidae, Agamidae), (5) root of Anguimorpha (Shinisauridae, Varanidae, Helodermatidae, Anguidae), and (6) the root of Lacertidae (*Lacerta*, *Takydromus*). We selected these nodes because they are strongly supported clades, do not correspond to fossil calibration points, are mostly phylogenetically independent and non-nested (except the root), and represent a diversity of tree depths and clades.

We also evaluated the width of the posterior density interval for each comparable node and averaged this width across the 10 replicates for each set of conditions. We then compared these values to those for the complete, 20-gene dataset for the analyses including and excluding incomplete genes. For each set of conditions, we also tallied the number of nodes that differed between the complete dataset tree and the trees from analyses including and excluding genes with missing data, and averaged these errors in topology across the 10 replicates. Similarly, we also compared the summed posterior probabilities across the trees including and excluding incomplete genes, to evaluate whether including or excluding incomplete genes tended to decrease clade support.

No single empirical study can represent all other empirical studies. One way in which our dataset may be unusual is the large number of fossil calibration points ($n = 9$) relative to the small number of ingroup taxa ($n = 32$). A potentially important consequence is that the large number of calibration points might potentially reduce the negative impacts of missing data (or excluding genes). Therefore, we performed an alternative set of analyses in which we randomly selected only 2 fossil calibration points from among the set of 9. Specifically, for each replicate, we used a different random selection of two calibration points. For these analyses, we considered one of those conditions with the most missing data (scenario "c" for strategy 2) and the complete 10 loci dataset.

We also performed statistical analyses of the relationships between the error in the estimated age of each node and the depth (age) of that node in the complete 20-locus dataset (i.e. do the ages of deeper nodes tend to be over- or underestimated?). Similarly, we addressed how the mean age of the two subsampled fossil calibration points in a given replicate was related to the mean errors in the estimated ages for that replicate (testing for a correlation between mean age of the fossil calibration points and raw and absolute errors in estimated clade ages across the tree, to see if sampling older calibration points led to overestimating clade ages). Finally, for the analyses of six selected nodes, we tested if greater amounts of missing data in a given replicate led to greater error in the estimated ages (for that specific node). Thus, we tested for a correlation between the percentage of genes missing in the taxa united by a given node and the error in the age estimated for that node, across the 10 replicates. Note that we did not include the root node here, since the percentage of missing genes would be the same across all replicates for this node. For all three of these correlation analyses, we first tested for the normality of the data using a Kolmogorov–Smirnov test and then performed Pearson correlation (for normal data) or non-parametric Spearman correlation (for non-normal data). Statistical analyses were performed using SPSS version 12.0.

3. Results

The phylogeny and divergence times estimated for the complete dataset of 20 loci for 32 squamate taxa is shown in Fig. 1. Overall, the clade ages are broadly similar to those estimated in other recent studies (see review in Mulcahy et al., 2012), as is the overall topology (e.g. Mulcahy et al., 2012; Wiens et al., 2012; Pyron et al., 2013).

The results show that adding genes with missing data generally had relatively little negative impact on divergence-time estimation (Table 2). The differences between ages estimated with the complete data and those estimated with the subsampled data were smaller when genes with missing data were included, relative to analyses in which these genes were excluded (Table 2). The results showed that genes with less missing data (data absent in 25% of taxa) were more beneficial than those with more missing data (data absent in 75% of taxa), but in general, adding the genes with missing data did not have a negative impact (Table 2). On average, the absolute deviations from the dates estimated from the complete dataset are small, within about 5 Myr (Table 2; 5 Myr is especially small considering that the group is ~220 Myr old; Fig. 1). The values of raw mean error were relatively close to zero, suggesting that there was not a strong bias in the estimated ages to be either consistently older or younger than those for the complete data (Table 2). Overall, they tended to be slightly positive, indicating that the estimated ages were biased towards being older than those from the complete dataset.

The results above describe the case in which either 5 or 2 genes are complete, and the other genes are incomplete. We also explored the case in which all the genes were incomplete (Table 2). When each gene was only 25% or 50% incomplete, the mean errors remained relatively small (still close to 5 Myr). However, when each gene was 75% incomplete, the error was more substantial, with mean error close to 17 Myr and with a strong bias towards underestimated clade ages (relative to the complete dataset of 20 genes). These results emphasize the benefits of including genes that are sampled for all taxa (see also Filipski et al., 2014). However, in some cases, the errors were relatively small without these "backbone" genes, even when data were missing in 50% of the taxa in all genes.

The results also showed that, for these data, adding genes had relatively little impact on the accuracy of divergence time estimates (Table 2). For example, the mean error for 10 complete loci (2.10 Myr per node) was quite similar to that for only 2 loci (4.34 Myr). Again, the errors associated with subsampling genes (and missing data) were very small relative to the overall timescales involved (Fig. 1).

Table 2
Errors in divergence-time estimates associated with including loci with missing data, compared to the error from excluding incomplete loci. Missing data are distributed among a different set of randomly selected taxa in each incomplete gene. Results are given in terms of mean absolute error (for each node, the absolute value of the difference in the date estimated from these subsampled data relative to the date estimated from the dataset of 20 complete loci, then averaged across all nodes and all replicates) and mean raw error (the average value of these differences across nodes and replicates, with negative values indicating that dates are underestimated relative to the complete dataset). Results are also contrasted with those for 10 complete loci. Missing data are listed as "0%" for 2, 5, and 10 complete loci (all taxa have data for all genes), but the actual value is ~3%.

	Overall missing data (%)	Mean absolute error (Myr)	Mean raw error (Myr)
10 complete loci	0	2.10	0.26
5 complete loci	0	3.77	1.31
10 loci (5 loci with missing data in 8 taxa)	12.5	2.21	0.46
10 loci (5 loci with missing data in 16 taxa)	25	2.95	1.05
10 loci (5 loci with missing data in 24 taxa)	37.5	3.12	0.67
2 complete loci	0	4.34	−0.44
10 loci (8 loci with missing data in 8 taxa)	20	2.88	0.67
10 loci (8 loci with missing data in 16 taxa)	40	3.72	1.23
10 loci (8 loci with missing data 24 taxa)	60	3.93	0.34
10 loci (all loci with missing data in 8 taxa)	25	2.64	0.22
10 loci (all loci with missing data in 16 taxa)	50	5.31	0.15
10 loci (all loci with missing data in 24 taxa)	75	16.93	−16.11

We also tested whether errors in estimated node ages tended to be greater for deeper nodes. We explored cases in which there were 5 loci with missing data in 8 and 24 taxa, 8 loci with missing data in 8 and 24 taxa, and all 10 loci with missing data in 8 and 24 taxa (i.e. cases with both relatively little and relatively extensive missing data). We tested for a relationship between the depth of nodes (based on the age in the complete 20-locus dataset) and the mean error in the estimated ages of those nodes relative to the complete data, based on both the raw error values and the absolute error (i.e. treating both overestimates and underestimates as positive values). Across the six conditions and 10 replicates, we generally found only sporadic support for a significant relationship between node depth and error (Table S3). Specifically, we found support for a relationship between raw error and node depth in 20 (out of 60) replicates, and between absolute error and node depth in 17. The relationship with absolute error was always positive (greater error for deeper nodes) but the relationship with raw error tended to be negative (ages of deeper nodes tend to be underestimated; 17 of 20 replicates). Significant relationships between node depth and both absolute and raw error were most common (7 of 10 replicates) and strongest in magnitude when all 10 loci had missing data in 24 of the 32 taxa. These were also the conditions under which mean errors were greatest (~17 Myr per node; see above and Table 2). In general, greater errors should be expected for deeper nodes than for shallower nodes (i.e. shallow nodes can only take a limited range of values), but these results show underestimation of ages for deeper nodes when there are extensive missing data (75%) in all sampled genes.

Results for the 95% highest posterior density (HPD) intervals showed that sampling fewer loci generally led to only small differences in the width of the HPD relative to the complete dataset of 20 loci (Table 3). Adding loci with missing data generally reduced these differences, but the improvement was smaller with more extensive missing data (Table 3). Interestingly, neither reducing the number of loci nor increasing the amount of missing data appeared to consistently bias the width of the HPDs in a particular direction (i.e. intervals did not become wider with less data). However, the results with 75% missing data in all genes were again an outlier, with HPD widths that were substantially narrower. This pattern of narrower HPDs may occur because the estimated ages are substantially younger under these conditions (see above), limiting the range of possible dates that can be included in the HPD.

BEAST analyses also estimate topology and support values (posterior probabilities of clades). The results here showed that adding five loci with missing data had little impact on mean support values, but improved the accuracy of the estimated topology slightly (relative to excluding these loci), even when the missing data were extensive (Table 4). When adding eight loci with missing data, the improvement in posterior probabilities and topological accuracy was more substantial, especially when the added genes had only 25% or 50% missing data (Table 4). The results were generally similar when all genes were incomplete, but accuracy and support were both substantially reduced when all genes had 75% missing data.

The preceding analyses were based on the case in which missing data cells had a different random distribution in each gene. We also explored the case in which all incomplete genes had missing data for the same taxa, focusing on the case with the maximum amount of missing data (such that different ways of distributing missing data should have maximum impact). We found that the results were very similar regardless of how missing data cells were distributed (Table 5), although errors in age estimates were slightly larger when the missing data were in the same species across genes.

We also examined the case in which only 2 fossil calibration points were used, instead of the full set of 9 (Fig. 2). For this analysis, we compared a situation with extensive missing data (8 loci with 24 taxa incomplete, incomplete taxa randomly chosen in each gene) to that with 10 genes with complete data. The results (Fig. 2; Table 6) showed that reducing the number of calibration points caused a dramatic increase in error in the estimated divergence dates, regardless of the amount of missing data. For the case with missing data, age estimates were off by (on average) nearly 40 Myr per node, with a substantial bias towards estimates being older than those from the full dataset (20 loci, 9 calibration points). These errors were even greater with datasets of 10 complete loci than the datasets with 8 incomplete genes. There were also dramatic changes in the HPD values for both datasets, with HPDs becoming much broader with only 2 fossil calibration points. However, reducing the number of calibration points seemed to have little impact on support values or topological accuracy (Table 6).

We tested whether the mean age of the two sampled fossil calibration points influenced the estimated ages. We confirmed that when older fossils were sampled (e.g. older than the mean of all nine, 76.5 Myr), the estimated ages tended to be older than those estimated from the complete set of nine calibration points, yielding a significant positive correlation between the mean age of the sampled calibration points and both the raw and absolute error in the estimated clade ages (Table S4). However, the errors were clearly biased towards overestimation of clade ages overall (mean raw error = 21.62 Myr), and the mean age of the subsampled calibration points varied along a fairly limited range (65–99 Myr).

Table 3
Impact of missing data on the width of the 95% highest posterior density interval, compared to the width when excluding incomplete genes. Missing data are distributed among a different set of randomly selected taxa in each incomplete gene. For each node, widths are compared to those estimated for the complete dataset of 20 loci, and then averaged across nodes and replicates. Values are given in terms of mean absolute differences in width (in millions of years) and the mean raw difference. Missing data are listed as "0%" for 2, 5, and 10 complete loci (all taxa have data for all genes), but the actual value is ~3%.

	Overall missing data (%)	Mean absolute difference (Myr)	Mean raw difference (Myr)
10 complete loci	0	2.73	0.39
5 complete loci	0	4.06	−0.06
10 loci (5 loci with missing data in 8 taxa)	12.5	2.78	−0.05
10 loci (5 loci with missing data in 16 taxa)	25	3.34	0.17
10 loci (5 loci with missing data in 24 taxa)	37.5	3.81	−0.13
2 complete loci	0	6.73	−0.35
10 loci (8 loci with missing data in 8 taxa)	20	3.49	0.35
10 loci (8 loci with missing data in 16 taxa)	40	4.31	−0.35
10 loci (8 loci with missing data 24 taxa)	60	5.35	0.44
10 loci (all loci with missing data in 8 taxa)	25	3.44	0.86
10 loci (all loci with missing data in 16 taxa)	50	5.90	−1.42
10 loci (all loci with missing data in 24 taxa)	75	17.45	−11.89

Table 4
Impact of missing data on mean support values (posterior probabilities) and accuracy of trees (proportion of nodes shared with complete 20-locus tree), compared to results when excluding incomplete genes. Missing data are distributed among a different set of randomly selected taxa in each incomplete gene. Posterior probabilities are averaged across all nodes within each tree and then averaged across all replicates for a given set of conditions. Missing data are listed as "0%" for 5, 10, and 20 complete loci (all taxa have data for all genes), but the actual value is ~3%.

	Overall missing data (%)	Mean Pp	Mean accuracy
20 complete loci	0	0.988	100
10 complete loci	0	0.966	94.3
5 complete loci	0	0.957	90.3
10 loci (5 loci with missing data in 8 taxa)	12.5	0.969	92.7
10 loci (5 loci with missing data in 16 taxa)	25	0.960	91.7
10 loci (5 loci with missing data in 24 taxa)	37.5	0.960	92.0
2 complete loci	0	0.891	85.3
10 loci (8 loci with missing data in 8 taxa)	20	0.957	91.3
10 loci (8 loci with missing data in 16 taxa)	40	0.953	90.0
10 loci (8 loci with missing data 24 taxa)	60	0.913	87.3
10 loci (all loci with missing data in 8 taxa)	25	0.954	91.3
10 loci (all loci with missing data in 16 taxa)	50	0.923	86.0
10 loci (all loci with missing data in 24 taxa)	75	0.594	48.0

Table 5
Comparison of results when missing data cells are distributed among a different set of randomly selected taxa in each gene (random), or are distributed in the same taxa across all incomplete genes (fixed). These comparisons are only for conditions with relatively extensive missing data. Five loci refers to the case in which five loci have missing data in 75% of the taxa (24 taxa), and 37.5% missing data overall. Eight loci refers to the case in which eight loci have missing data in 75% of the taxa (24 taxa in 8 genes), and 60% missing data overall.

	Random	Fixed
Age difference (absolute value)—5 loci	3.12	3.60
Age difference (raw value)—5 loci	0.67	1.26
Age difference (absolute value)—8 loci	3.93	4.27
Age difference (raw value)—8 loci	0.34	−0.13
HPD difference (absolute)—5 loci	3.81	4.02
HPD difference (raw)—5 loci	−0.13	0.07
HPD difference (absolute)—8 loci	5.35	6.49
HPD difference (raw)—8 loci	0.44	0.35
Mean Pp—5 loci	0.960	0.960
Mean Pp—8 loci	0.913	0.919
Accuracy—5 loci	92.0	90.0
Accuracy—8 loci	87.3	87.3

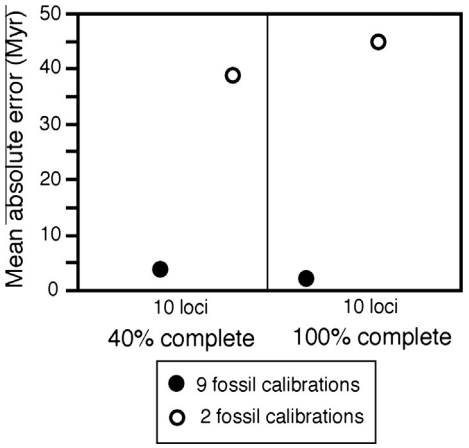

Fig. 2. Comparison of mean absolute error for analyses with missing data (60% overall, with 8 of 10 genes with 75% missing data each) and without (10 complete loci), and using the full set of 9 fossil calibration points or a randomly selected subset of 2 calibration points. Error is based on the difference between the ages estimated in these analyses and those estimated from the full set of 20 complete loci. Errors are averaged across all nodes of the tree and then across 10 replicates, and so error bars are not shown. See Table 6 for full set of results.

The main results described above (Tables 2–4) were based on averaged differences across all nodes. The results remained similar when the differences were standardized (i.e. using % change in age rather than raw differences in Myr) and after excluding the nine fossil calibrated nodes and the two relatively poorly supported nodes (Bayesian posterior probabilities 0.8024 and 0.8523, Fig. 1). Details of comparisons of the mean value and HPD width of the divergence-time estimate, topology, and support values are presented in Tables S5–S7.

Table 6
Comparison of results based on 9 fossil calibration points (as in most of the results of this study) to those based on only 2 fossil calibration points (randomly selected for each replicate). Missing data are distributed randomly among 24 taxa for each of the 8 incomplete genes (60% missing data overall). Results are compared to the case in which 10 complete loci are sampled.

	9 points	2 points
2 complete loci, 8 incomplete		
Age differences (absolute value)	3.93	38.83
Age difference (raw value)	0.34	21.62
HPD difference (absolute)	5.35	49.46
HPD difference (raw)	0.44	42.12
Mean Pp	0.913	0.905
Accuracy	87.3	83.0
10 loci, all complete		
Age differences (absolute value)	2.10	44.97
Age difference (raw value)	0.26	27.88
HPD difference (absolute)	2.73	45.90
HPD difference (raw)	0.39	35.96
Mean Pp	0.966	0.966
Accuracy	94.3	94.0

Results were also similar when focusing on the select set of six nodes (Fig. 1) described in the methods (Table S8). Importantly, when focusing on the five nodes above the root, we found no tendency for the age estimates for these nodes to show greater error in those replicates in which there was a higher percentage of missing data (i.e. more genes lacking data) in the taxa united by that node (Table S9). Specifically, there was no significant relationship between mean error and percentage of missing genes for that node across the 10 replicates for a given set of conditions. Furthermore, in the seven cases that approached significance ($P < 0.05$, but not significant under a Bonferroni correction), the relationship between the amount of error and the percentage of missing genes was positive in three cases and negative in four. Thus, in the majority of these cases, there was actually less error in the age estimates for a given node when the percentage of missing genes was higher.

4. Discussion

Our results based on experimental analyses of empirical data in reptiles show that extensive missing data do not necessarily lead to misleading estimates of divergence dates using the relaxed lognormal dating approach in BEAST. In fact, adding genes that have 50% or even 75% missing data can improve estimates of divergence dates, relative to excluding these genes. More generally, we found that changing the number of genes sampled had relatively little impact on the accuracy of divergence-time estimates for these data. Thus, estimates based on 2 genes were (on average) within 5 Myr of estimates based on 20 genes (an especially small number considering the ~220 Myr timescale of squamate phylogeny; Fig. 1). Results were somewhat different when all genes were 75% incomplete (estimates off by ~17 Myr on average), but errors remained relatively small even when all genes were 50% incomplete.

In contrast, we found that reducing the number of fossil calibration points led to dramatic errors in divergence-time estimates, regardless of the amount of missing data (Fig. 2). For example, errors in divergence-times estimates with only two randomly selected calibration points were typically off by ~40 Myr, with similar levels of error given 10 complete loci or 2 complete loci with 8 highly incomplete loci (Fig. 2). Overall, our results suggest that the amount of missing data, or even the amount of sequence data overall, might be minor issues for dating analyses relative to the quality and quantity of fossil calibration points. The importance of fossil calibrations to dating analyses is well-recognized in general (e.g. Near and Sanderson, 2004; Linder et al., 2005; Near et al., 2005; Benton and Donoghue, 2006; Battistuzzi et al., 2010; Magallón et al., 2013), but our results provide a particularly striking empirical example based on controlled experiments. Our results are especially striking when compared to the seemingly minor importance of other factors (e.g. missing data, gene sampling). It should also be noted that having only two fossil calibration points is not an unusually small number in empirical studies.

We acknowledge several limitations to our study. First, our study is based on one empirical dataset. Thus, our results might not be applicable to all other studies. For example, other studies might involve much deeper (or shallower) divergence times or more rapidly evolving genes (e.g. mitochondrial DNA, as in Mulcahy et al., 2012). Nevertheless, our results clearly show that large amounts of missing data need not strongly impact divergence-dating analyses with BEAST. If the effects of missing data were strong, general, and misleading, then they should have appeared in our results. However, at the same time, our results do not guarantee that missing data will always be harmless in all divergence dating analyses (see below). Second, because our results are empirical, the actual divergence dates are unknown. Therefore, there is a certain "leap of faith" required to treat the results from 20 complete loci as known and to treat deviations from those results as errors. However, regardless of whether the dates from 20 loci are actually correct, our experiments clearly show that adding genes with extensive missing data does not necessarily lead to dramatic changes in the estimated dates. Third, our analyses are based on a finite number of replicates (i.e. 10 for each set of conditions). This was necessary given the computational intensity of BEAST. Fortunately, results were highly consistent across conditions as well as among replicates for a given set of conditions.

Our results are generally consistent with those of Filipski et al. (2014), who analyzed the performance of the dating method RelTime given different amounts and distributions of missing data. They found that RelTime was also relatively robust to extensive missing data, using both simulated and empirical data. The major exception was in cases in which a matrix was so sparse that no single gene spanned a given node in the tree (see their Fig. 5), such that there were no data available to estimate the length of that branch. We also found that missing data were most problematic when all loci were highly incomplete (75%) and there were no backbone genes that were present in all taxa.

Finally, we think it likely that our results (and those of Filipski et al., 2014) should apply to other dating methods as well. For example, other empirical analyses suggest that extensive missing data can have relatively little impact on estimated branch lengths (e.g. Wiens and Tiu, 2012; Jiang et al., 2014). Therefore, it seems that dating methods that rely on branch-length estimates from previous likelihood or Bayesian analyses should also be robust to missing data (e.g. penalized likelihood; Sanderson, 2002). Nevertheless, additional study of this issue would be useful.

In summary, our results show that divergence-date estimation with BEAST can be highly robust to extensive missing data. Perhaps more importantly, our results suggest that the number of fossil calibration points might be far more crucial for the accuracy of divergence-date estimates than the sparseness or even amount of sequence data in a matrix.

Acknowledgments

Y.Z. was supported by grants from the National Natural Science Foundation of China (NSFC-31372181) and Chinese Academy of Sciences (Y1C3051100). We thank A. Larson and two anonymous reviewers for helpful comments on the manuscript.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2015.02.002.

References

Alfaro, M.E., Santini, F., Brock, C.D., Alamillo, H., Dornburg, A., Carnevale, G., Rabosky, D., Harmon, L.J., 2009. Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. Proc. Natl. Acad. Sci. USA 106, 13410–13414.

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

Benton, M.J., Donoghue, P.C.J., 2006. Paleontological evidence to date the tree of life. Mol. Biol. Evol. 24, 26–53.

Burleigh, J.G., Hilu, K., Soltis, D., 2009. Inferring phylogenies with incomplete data sets: a 5-gene, 567-taxon analysis of angiosperms. BMC Evol. Biol. 9, 61.

Cho, S., Zwick, A., Regier, J.C., Mitter, C., Cummings, M.P., Yao, J., Du, Z., Zhao, H., Kawahara, A.Y., Weller, S., 2011. Can deliberately incomplete gene sample augmentation improve a phylogeny estimate for the advanced moths and butterflies (Hexapoda: Lepidoptera)? Syst. Biol. 60, 782–796.

Crawley, S.S., Hilu, K.W., 2012. Impact of missing data, gene choice, and taxon sampling on phylogenetic reconstruction: the Caryophyllales (angiosperms). Plant Syst. Evol. 298, 297–312.

Driskell, A.C., Ané, C., Burleigh, J.G., McMahon, M.M., O'Meara, B.C., Sanderson, M.J., 2004. Prospects for building the tree of life from large sequence databases. Science 306, 1172–1174.

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

Drummond, A.J., Ho, S.Y.W., Phillips, M.J., Rambaut, A., 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4, e88.

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.

FitzJohn, R.G., 2010. Quantitative traits and diversification. Syst. Biol. 59, 619–633.

Hovmöller, R., Knowles, L.L., Kubatko, L.S., 2013. Effects of missing data on species tree estimation under the coalescent. Mol. Phylogenet. Evol. 69, 1057–1062.

Hugall, A.F., Foster, R., Lee, M.S.Y., 2007. Calibration choice, rate smoothing, and the pattern of tetrapod diversification according to the long nuclear gene RAG-1. Syst. Biol. 56, 543–563.

Jiang, W., Chen, S.-Y., Wang, H., Li, D.-Z., Wiens, J.J., 2014. Should genes with missing data be excluded from phylogenetic analyses? Mol. Phylogenet. Evol. 80, 308–318.

Lanfear, R., Calcott, B., Ho, S.Y., Guindon, S., 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. 29, 1695–1701.

Lemmon, A.R., Brown, J.M., Stanger-Hall, K., Lemmon, E.M., 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst. Biol. 58, 130–145.

Linder, H.P., Hardy, C.R., Rutschmann, F., 2005. Taxon sampling effects in molecular clock dating: an example from the African Restionaceae. Mol. Phylogenet. Evol. 35, 569–582.

Magallón, S., Hilu, K.W., Quandt, D., 2013. Land plant evolutionary timeline: gene effects are secondary to fossil constraints in relaxed clock estimation of age and substitution rates. Am. J. Bot. 100, 556–573.

Mulcahy, D.G., Noonan, B.P., Moss, T., Townsend, T.M., Reeder, T.W., Sites Jr., J.W., Wiens, J.J., 2012. Estimating divergence dates and evaluating dating methods using phylogenomic and mitochondrial data in squamate reptiles. Mol. Phylogenet. Evol. 65, 974–991.

Near, T.J., Sanderson, M.J., 2004. Assessing the quality of molecular divergence time estimates by fossil calibrations and fossil based model selection. Philos. Trans. R. Soc. Lond. B 359, 1477–1483.

Near, T.J., Meylan, P.A., Shaffer, H.B., 2005. Assessing concordance of fossil calibration points in molecular clock studies: an example using turtles. Am. Nat. 165, 137–146.

Philippe, H., Snell, E.A., Bapteste, E., Lopez, P., Holland, P.W.H., Casane, D., 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol. Biol. Evol. 21, 1740–1752.

Pyron, R.A., Burbrink, F.T., Colli, G.R., Nieto Montes de Oca, A., Vitt, L.J., Kuczynski, C.A., Wiens, J.J., 2011. The phylogeny of advanced snakes (Colubroidea), with discovery of a new subfamily and comparison of support methods for likelihood trees. Mol. Phylogenet. Evol. 58, 329–342.

Pyron, R.A., Burbrink, F.T., Wiens, J.J., 2013. A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes. BMC Evol. Biol. 13, 93.

Quintero, I., Wiens, J.J., 2013. Rates of projected climate change dramatically exceed past rates of climatic-niche evolution among vertebrate species. Ecol. Lett. 16, 1095–1103.

Rambaut, A., Drummond, A.J., 2007. Tracer v1.4. Institute of Evolutionary Biology, University of Edinburgh, Edinburgh (United Kingdom), <http://beast.bio.ed.ac.uk/software/tracer>.

Ree, R.H., Smith, S.A., 2008. Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. Syst. Biol. 57, 4–14.

Ricklefs, R.E., 2007. Estimating diversification rates from phylogenetic information. Trends Ecol. Evol. 22, 601–610.

Roure, B., Baurain, D., Philippe, H., 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol. Biol. Evol. 30, 197–214.

Sanderson, M.J., 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol. Biol. Evol. 19, 101–109.

Sanderson, M.J., McMahon, M.M., Steel, M., 2010. Phylogenomics with incomplete taxon coverage: the limits to inference. BMC Evol. Biol. 10, 155.

Sanderson, M.J., McMahon, M.M., Steel, M., 2011. Terraces in phylogenetic tree space. Science 333, 448–450.

Simmons, M.P., 2012. Misleading results of likelihood-based phylogenetic analyses in the presence of missing data. Cladistics 28, 208–222.

Simmons, M.P., 2014. A confounding effect of missing data on character conflict in maximum likelihood and Bayesian MCMC phylogenetic analyses. Mol. Phylogenet. Evol. 80, 267–280.

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

Wiens, J.J., 2003. Missing data, incomplete taxa, and phylogenetic accuracy. Syst. Biol. 52, 528–538.

Wiens, J.J., 2005. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? Syst. Biol. 54, 731–742.

Wiens, J.J., Moen, D., 2008. Missing data and the accuracy of Bayesian phylogenetics. J. Syst. Evol. 46, 307–314.

Wiens, J.J., Morrill, M.C., 2011. Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. Syst. Biol. 60, 719–731.

Wiens, J.J., Tiu, J., 2012. Highly incomplete taxa can rescue phylogenetic analyses from the negative impacts of limited taxon sampling. PLoS ONE 7, e42925.

Wiens, J.J., Hutter, C.R., Mulcahy, D.G., Noonan, B.P., Townsend, T.M., Sites Jr., J.W., Reeder, T.W., 2012. Resolving the phylogeny of lizards and snakes (Squamata) with extensive sampling of genes and species. Biol. Lett. 8, 1043–1046.