Running head: Missing data and topology in Total Evidence method

Effects of missing data on topological inference using a Total Evidence approach

Thomas Guillerme[a,b*], and Natalie Cooper[a,b,1]

[a] *School of Natural Sciences, Trinity College Dublin, Dublin 2, Ireland.*

[b] *Trinity Centre for Biodiversity Research, Trinity College Dublin, Dublin 2, Ireland.*

[*] *Corresponding author. Zoology Building, Trinity College Dublin, Dublin 2, Ireland; E-mail: guillert@tcd.ie; Fax: +353 1 6778094; Tel: +353 1 896 2571.*

[1] *Present address: Department of Life Sciences, Natural History Museum, Cromwell Road, London, SW7 5BD, UK. E-mail: nhcooper12@gmail.com*

Abstract

To fully understand macroevolutionary patterns and processes, we need to include both extant and extinct species in our models. This requires phylogenetic trees with both living and fossil taxa at the tips. One way to infer such phylogenies is the Total Evidence approach which uses molecular data from living taxa and morphological data from living and fossil taxa.

Although the Total Evidence approach is very promising, it requires a great deal of data that can be hard to collect. Therefore this method is likely to suffer from missing data issues that may affect its ability to infer correct phylogenies.

Here we use simulations to assess the effects of missing data on tree topologies inferred from Total Evidence matrices. We investigate three major factors that directly affect the completeness and the size of the morphological part of the matrix: the proportion of living taxa with no morphological data, the amount of missing data in the fossil record, and the overall number of morphological characters in the matrix. We infer phylogenies from complete matrices and from matrices with various amounts of missing data, and then compare missing data topologies to the "best" tree topology inferred using the complete matrix.

We find that the number of living taxa with morphological characters and the overall number of morphological characters in the matrix, are more important than the amount of missing data in the fossil record for recovering the "best" tree topology. Therefore, we suggest that sampling effort should be focused on morphological data collection for living species to increase the accuracy of topological inference in a Total Evidence framework. Additionally, we find that Bayesian methods consistently outperform other tree inference methods. We therefore recommend using Bayesian consensus trees to fix the tree topology prior to further analyses.

Keywords: morphological characters, Bayesian, Maximum Likelihood, topology, fossil, living.

1 Introduction

Although most species that have ever lived are now extinct (Novacek and Wheeler, 1992; Raup, 1981), the majority of macroevolutionary studies focus solely on living species (e.g. Meredith et al., 2011; Jetz et al., 2012). Ignoring fossil taxa may lead to misinterpretation of macroevolutionary patterns and processes such as the timing of diversification events (e.g. Pyron, 2011), relationships among lineages (e.g. Manos et al., 2007) or niche occupancy (e.g. Pearman et al., 2008). This has led to increasing consensus among evolutionary biologists that fossil taxa should be included in macroevolutionary studies (Jackson and Erwin, 2006; Quental and Marshall, 2010; Dietl and Flessa, 2011; Slater and Harmon, 2013; Fritz et al., 2013). However, to do this we need to be able to place living and fossil taxa into the same phylogenies; a task that remains difficult despite recent methodological developments (e.g. Pyron, 2011; Ronquist et al., 2012a; Matzke, 2014).

Up to now, three main approaches have been used to place both living and fossil taxa into phylogenies. These approaches differ mainly in whether they treat fossil taxa as tips or as nodes in the phylogeny, and in which part of the available fossil data is used (i.e. the age of the fossil only or both its age and morphology). Classical cladistic methods use matrices containing morphological data from both living and fossil taxa and treat each taxon as a tip in the phylogeny. Relationships among the taxa are then inferred using optimality criteria such as maximum parsimony (Simpson, 1944). This approach is commonly used by paleontologists but it ignores the additional molecular data available from living species and does not allow use of probabilistic methods for dealing with phylogenetic uncertainty. Neontologists, on the other hand, more commonly use probabilistic approaches (e.g. Maximum Likelihood or Bayesian methods) based on matrices containing only molecular data from living species. Because fossil taxa do not usually have available DNA, fossils are used as nodes rather than tips in these phylogenies and their occurrence dates are used to time calibrate phylogenies (Zuckerkandl and Pauling, 1965). There have been great improvements in the theory and application of these two approaches (e.g. Bapst, 2013; Stadler and Yang, 2013; Heath et al., 2014) as well as much debate about the "best" approach to use (e.g. Spencer and Wilberg, 2013; Wright and Hillis, 2014). However neither approach uses all the available data.

A final approach, known as the Total Evidence method, uses matrices containing molecular data from living taxa and morphological data from both living and fossil taxa (Eernisse and Kluge, 1993). This approach treats every taxon as a tip in the phylogeny, uses the occurrence age of the fossils to time calibrate the phylogeny (known as tip-dating Ronquist et al., 2012a), and allows the use of probabilistic methods for estimating phylogenetic uncertainty (Ronquist et al., 2012a). Total Evidence methods have been successfully applied to empirical data (e.g. Pyron, 2011; Ronquist et al., 2012a; Schrago et al., 2013; Slater, 2013; Beck and Lee, 2014), and are becoming an increasingly popular way of adding fossil taxa to phylogenies. However, although the Total Evidence approach seems very promising, there is one big drawback in using this approach: it requires both molecular and morphological data, both of which can be difficult (or impossible) to collect for every living and fossil taxon in the tree. Morphological data for living taxa are rarely collected when molecular data are available (e.g. O'Leary et al., 2013 *vs.* Meredith et al., 2011), and for fossil taxa, data can only be collected from features preserved in the fossil record (for example, in vertebrates, the hardest parts of the skeleton are more often preserved than soft parts (Sansom and Wills, 2013); and molecular data is (nearly) always unavailable. Therefore Total Evidence matrices are likely to contain a lot of missing data that may affect the method's ability to infer correct topologies, branch lengths and support values (Salamin et al., 2003).

Although missing data does not appear be a major problem in molecular and morphological matrices separately (as long as enough data overlaps in each case, and missing data is not biased towards particular subclades (Wiens, 2003, 2006; Wiens and Moen, 2008; Lemmon et al., 2009; Sanderson et al., 2011; Roure and Philippe, 2011; Pattinson et al., 2014), it may become more of an issue in Total Evidence matrices containing both molecular and morphological data for living and fossil taxa. This may be particularly problematic as fossil taxa (generally) do not have molecular data, resulting in a large section of missing data in Total Evidence matrices. Until now, no attempt has been made to study the impact of this issue on phylogenetic inference using Total Evidence methods.

In this study, we focus on the effect of missing data on our ability to recover a "best" tree topology because it is a crucial aspect of a phylogeny in many macroevolutionary studies, for example when trying to elucidate the evolutionary relationships among species (e.g. Meredith et al., 2011; Jetz et al., 2012), or for studying evolutionary transitions (e.g. Friedman, 2010). Although branch length estimation is also important (namely for timing extinction and/or speciation events - e.g. Ronquist et al., 2012a), we do not consider branch lengths in this study. This is partially due to difficulties with simulating branch lengths and topology simultaneously, but also because previous studies have already empirically assessed the effect of the Total Evidence method on branch length variation but with a fixed topology approach (Ronquist et al., 2012a; Schrago et al., 2013; Slater, 2013; Beck and Lee, 2014). Thus understanding the sensitivity of topology to missing data is important for assessing the accuracy of tree estimation in the Total Evidence framework. To our knowledge, this question has never been formally assessed.

Here we use a simulation approach to assess the effect of missing data on tree topologies inferred from Total Evidence matrices. Since the molecular part of a Total Evidence matrix acts like a "classical" molecular matrix containing only the living taxa (Ronquist et al., 2012a), the effect of missing data on such matrices is well known (Wiens, 2006; Wiens and Moen, 2008; Lemmon et al., 2009; Roure and Philippe, 2011). Therefore, we focus only on missing data in the morphological part of the matrix. We investigate three major parameters that directly affect the completeness and size of the morphological part of the matrix, and reflect empirical biases in data availability: (i) the proportion of living taxa with no morphological data; (ii) the proportion of missing data in the fossil taxa; and (iii) the proportion of missing morphological characters for both living and fossil taxa in the matrix (i.e. the size of the matrix). We remove data from a Total Evidence matrix by changing the values of these three parameters and then assess how this affects the topology of trees inferred using Maximum Likelihood and Bayesian methods. We find that minimizing the number of living taxa with no morphological data and the number of missing morphological characters improves the ability of Total Evidence methods to recover the "best" tree topology more so than minimizing the amount of missing data in the fossil record. Additionally, we find that the ability of Total Evidence methods to recover the "best" tree topology is increased when using Bayesian methods.

2 Materials and Methods

To explore how missing data in the morphological sections of Total Evidence matrices influences tree topology, we used the following protocol (note that we explain each step in detail below this general outline; Figure 1).

1. Generating the matrix:

   We randomly generated a birth-death tree (hereafter called the "true" tree) and used it to simulate a matrix containing both molecular and morphological data for living and fossil taxa (hereafter called the "complete" matrix).

2. Removing data:

   We removed data from the morphological part of the "complete" matrix to simulate the effects of missing data by modifying three parameters (i) the proportion of living taxa with no morphological data (M_L), (ii) the proportion of missing data in the fossil taxa (M_F) and (iii) the proportion of missing morphological characters (M_C). We call the resulting 125 matrices "missing-data" matrices.

3. Estimating phylogenies:

   We inferred phylogenetic trees from the "complete" matrix and from the 125 "missing-data" matrices resulting in one tree generated from a matrix containing no missing data (hereafter called the "best" tree) and 125 trees inferred from the matrices with missing morphological data (hereafter called the "missing-data" trees). Phylogenies were inferred via both Maximum Likelihood and Bayesian approaches.

4. Comparing topologies:

   We compared the "best" tree to the "missing-data" trees to assess the influence of each parameter (M_L, M_F, M_C) and their interactions on the topologies of our phylogenies

We repeated these four steps 50 times to account for variation in our random parameters in the simulations.

2.1 Generating the matrix

First we randomly generated a "true" tree of 50 taxa in R v. 3.0.2 (R Core Team, 2014) using the package diversitree v. 0.9-6 (FitzJohn, 2012). We generated the tree using a birth death process by sampling speciation (λ) and extinction (μ) rates from a uniform distribution (bounded between 0 and 1) but maintaining $\lambda > \mu$ (Paradis, 2011). Empirical Total Evidence matrices vary in whether they have more fossil than living taxa or vice versa. For example, fossil taxa make up 88% (Beck and Lee, 2014), 58% (Schrago et al., 2013), 48% (Pyron, 2011), 31% (Ronquist et al., 2012a) and 31% (Slater, 2013) of taxa in various studies. To avoid biasing our simulations towards either living or fossil taxa and to make each simulation comparable, we implemented a rejection sampling algorithm to select only trees with 25 living and 25 fossil taxa. The fossil taxa were considered as unique tips at the end of extinct lineages. We then added an outgroup to the tree, using the mean branch length of the tree to separate the outgroup from the rest of the taxa, and with the branch length leading to the outgroup set as the sum of the mean branch length and the longest root-to-tip length of the tree.

Next, we generated a molecular and a morphological matrix from the "true" tree. The molecular matrix was inferred from the "true" tree using the R package phyclust v. 0.1-14 (Chen, 2011). The matrix contained 1000 character sites for 51 taxa and was generated using the seqgen algorithm (Rambaut and Grassly, 1997) and using the HKY model (Hasegawa et al., 1985) with random base frequencies (sampled from a uniform probability distribution bounded between 0 and 1 with the total frequency for the four bases equal to 1) and transition/transversion rate of two (Douady et al., 2003). The substitution rates were distributed following a gamma distribution with an alpha (α) shape of 0.5 (Yang, 1996). We chose a low value of α to reduce the number of sites with high substitution rates, thus avoiding too much homoplasy and a decrease in phylogenetic signal. We selected the parameters above to generate data with no special assumption about how the characters evolved, and to reduce the computational time required if these parameters were estimated rather than defined in the tree building part of the analysis (even with the parameters defined, total computational time for the whole analysis was around 150 CPU years). All the molecular information for fossil taxa was replaced by missing data ("?").

We inferred the morphological matrix using the R package ape v. 3.0-11 (Paradis et al., 2004) to generate a matrix of 100 character sites for 51 taxa. We assigned the number of character states (either two or three) for each morphological character by sampling with a probability of 0.85 for two states characters and 0.15 for three state characters. These probabilities were selected using the overall distribution of character states extracted from 100 published empirical morphological matrices (see Appendix A and Fig. A1 within). We then ran an independent discrete character simulation for each character using the "true" tree with the character's randomly selected number of states (two or three) and assuming an equal rate of change (i.e. evolutionary rate) from one character state to another (Pagel, 1994). This method allows us to have only two parameters for each character: the number of states and the evolutionary rate. For each character, the evolutionary rate was sampled from a gamma distribution with $\alpha = 0.5$. We used low evolutionary rate parameters to avoid homoplasy in the morphological part of the matrix and create a clear phylogenetic signal (Wagner, 2000; Dávalos et al., 2014; Wright and Hillis, 2014).

Finally, we combined the morphological and molecular matrices obtained from the "true" tree. Hereafter we call this the "complete" matrix, i.e. the matrix with no missing data except for the molecular data of the fossil taxa.

2.2 Removing data

We modified the "complete" matrix to get matrices with missing data by randomly replacing data with "?" in the morphological part of the matrices according to the following parameters:

1. M_L, the proportion of living taxa with no morphological data: 0%, 10%, 25%, 50% or 75%. This parameter illustrates the number of living taxa that are present in the molecular part of the matrix but not in the morphological part. This reflects the fact that the increased availability of molecular data means that morphological data for living species is rarely collected, and few people have the skills to identify characters needed for detailed phylogenetic analysis.

2. M_F, the proportion of missing data in the fossil taxa: 0%, 10%, 25%, 50% or 75%. This parameter illustrates the quality of the fossil record, a common problem due to preservation biases (Sansom and Wills, 2013).

3. M_C, the proportion of missing morphological characters for both living and fossil taxa: 0%, 10%, 25%, 50% or 75%. This parameter illustrates the number of available morphological characters for both living and fossil taxa. This reflects the amount of effort put into identifying the characters (e.g. O'Leary et al., 2013).

In practice, each parameter represents a different way of removing data from the matrix: M_L removes rows from the living taxa's data; M_F removes cells from the fossil taxa's data; and M_C removes columns across both living and fossil taxa's data. Note that M_L and M_F differ not only because of the region of the matrix affected: for M_L all the morphological data of a percentage of living taxa are removed, whereas for M_F a percentage of the data are removed at random from across the whole of the morphological matrix for fossil taxa. We first applied the parameters M_L and M_F to the matrix and then applied the M_C parameter. Therefore, when 10% data was missing for both M_L and M_F, 10% data was missing in the morphological part of the matrix. However, when applying the M_C parameter with the same percentage (10%) the resulting matrix potentially had more than 10% data missing.

We created matrices using all parameter combinations resulting in 125 (5^3) matrices. Note that one of these combinations has no missing data so is equivalent to the "complete" matrix, thus we have one effectively complete matrix in our 125 "missing-data" matrices. Because some parameter combinations introduce a lot of missing data (e.g. M_L = 75%, M_F = 75% and M_C = 75%), some matrices contained fossil taxa without any data at all. When this occurred we repeated the random deletion of characters until each taxon had at least 5% data across the whole morphological part of the matrix.

2.3 Estimating phylogenies

From the resulting matrices we generated two types of trees: the “best” tree inferred from the “complete” matrix and the “missing-data” trees inferred from the 125 matrices with various amounts of missing data. The “true” tree was used to generate the “complete” matrix and reflects the “true” evolutionary history in our simulations. The “best” tree, on the other hand, is the best tree we can build using state-of-the-art phylogenetic methods. In real world situations, the “true” tree is never available to us because we cannot know the true evolutionary history of a clade (except in very rare circumstances, e.g. Rozen et al., 2005). Therefore, here we focus on comparing the trees inferred from the matrices with missing data to the “best” tree, rather than the “true” tree, as the “best” tree is generally what we have to work with.

2.3.1 Maximum Likelihood

The “best” tree and the “missing-data” trees were inferred using RAxML v. 8.0.20 (Stamatakis, 2014). For the molecular data, we used the GTR + Γ_4 model (Tavaré, 1986) (default GTRGAMMA in RAxML v. 8.0.20 (Stamatakis, 2014)). For the morphological data, we used the implemented M*kv* model (Lewis, 2001) assuming an equal state frequency and a unique overall substitution rate (μ) following a gamma distribution of the rate variation with four distinct categories (M*k* + Γ_4; -K MK option in RAxML v. 8.0.20 (Stamatakis, 2014)). We used RAxML because it automatically corrects for acquisition bias (Lewis, 2001). It is also heavily used in the literature for Maximum Likelihood tree inference (e.g. Roure and Philippe, 2011; Bogdanowicz et al., 2012; Springer et al., 2012; O'Leary et al., 2013; Kelly et al., 2014) and is one of the fastest methods available (Stamatakis et al., 2008).

To measure the support for each branch in our simulated phylogenies we first ran a fast bootstrap analysis (Lazy Sub-tree Rearrangement) with 500 replicates on the "complete" matrix. We removed all the simulations with a median bootstrap support lower than 50 as a proxy for weak phylogenetic signal (Zander, 2004). We repeated this selection until we obtained 50 sets of simulations (i.e. 50 "complete" and 50 x 125 "missing-data" matrices) with a relatively strong phylogenetic signal (median bootstrap > 50). This step was implemented to make sure that the differences we observed in topologies (see below) were due to the amount of missing data for each parameter (M_L, M_F and M_C) and not simply to low branch support that is likely to lead to different topologies. On these selected simulations, we used the fast bootstrap algorithm and performed 1000 bootstraps for each tree inference to assess topological support (Pattengale et al., 2010). Using these parameters took ~8 CPU years to build 50 sets of 125 bootstrapped Maximum Likelihood trees (2.30GHz clock speed nodes). We performed this procedure to increase the resolution of our resulting trees.

2.3.2 Bayesian

The "best" tree and the "missing-data" trees were inferred using MrBayes v. 3.2.1 (Ronquist et al., 2012b). We partitioned the data to treat the molecular part as a non-codon DNA partition and the morphological part as a multi-state morphological partition. The molecular evolutionary history was inferred using the HKY model with a transition/transversion ratio of two (Douady et al., 2003) and a gamma distribution for the rate variation with four distinct categories (HKY + Γ_4). For the morphological data, we used the M*kv* model (Lewis, 2001), with equal state frequency and a unique overall substitution rate (μ) with four distinct rates categories (M*kv* + Γ_4). Note that MrBayes automatically corrects for acquisition bias in the morphological data partition (Nylander et al., 2004; Ronquist et al., 2012b). We chose these models to be consistent with the parameters used to generate the "complete" matrix.

Each Bayesian tree was estimated using two runs of four chains each for a maximum of 5×10^7 generations. For each estimation, we used the "true" tree's topology as a starting tree (with a starting value for each branch length of one). We also used two priors on the molecular part of the matrix: an exponential prior on the shape of the gamma distribution of $\alpha = 0.5$, and a transition/transversion ratio prior of two sampled from a strong beta distribution ($\beta(80,40)$); and one prior on the morphological part of the matrix (exponential prior on the shape of the gamma distribution of $\alpha = 0.5$). We used these priors to speed up the Bayesian estimation process. These priors biased the way the Bayesian process calculated branch lengths by giving non-random starting points and boundaries for parameter estimation however, here we are focusing on the effect of missing data on tree topology and not branch lengths. Even using these priors, it took 140 CPU years to build 50 sets of 125 Bayesian trees (2.30GHz clock speed nodes). The detailed MrBayes parameters are available in Appendix A.

We used the average standard deviation of split frequencies (ASDS) as a proxy to estimate the convergence of the chains and used a stop rule when the ASDS went below 0.01 (Ronquist et al., 2012b). We also checked the effective sample size (ESS) on a random sub-sample of runs in each simulation to ensure that ESS $>>$ 200 (Drummond et al., 2006). Finally we built a strict majority rule Bayesian consensus tree from the combined chains, excluding the 25% first iterations as burn-in (Ronquist et al., 2012b).

2.4 Comparing topologies

We compared the topology of the "missing-data" trees to the "best" tree to measure the effect of the three parameters M_L, M_F and M_C on tree topology. We used the Robinson-Foulds distance (Robinson and Foulds, 1981) to assess the amount of conserved clade positions and the Triplets distance (Dobson, 1975) to assess the amount of wildcard taxa (i.e. taxa that frequently change position in different trees (Kearney, 2002)). We used these two metrics because they illustrate two different aspects of tree topology (see Discussion) but also because their performance in measuring differences in topology is well described (Kuhner and Yamato, 2014) and well implemented (Bogdanowicz et al., 2012). We normalised both metrics using methods described in

Bogdanowicz et al. (2012) to generalize our results for any n number of taxa. These metrics are described in detail below.

2.4.1 Robinson-Foulds distance

The Robinson-Foulds distance (Robinson and Foulds, 1981), or "path difference", measures the difference between the number of clades and the twice number of shared clades across two trees. The metric reflects the distance between the distributions of tips among clades in the two trees (Robinson and Foulds, 1981) (see Appendix B for calculation details). This metric is bounded between zero, when the two trees are identical, and $2(n-2)$ (for two trees with n taxa) when there is no shared clade in the two trees. This metric is sensitive to minor changes in clade conservation: if the trees are composed of two clades of three taxa (*(((a,b),c),((d,e),f))*), the swapping of any two taxa will lead to a maximal score of the Robinson-Foulds distance indicating poor tree similarity. We normalised this metric following Bogdanowicz's Normalised Tree Similarity (NTS) method (Bogdanowicz et al., 2012). This methods scales any tree comparison metric using the mean distance between 1000 random trees (see Appendix B for the calculation details). This method is a generalisation of the topological accuracy method (Price et al., 2010) allowing to compare topological differences between any tree with any tree comparison metric.In practice when the Normalised Robinson-Foulds metric between two trees is equal to one, the trees are identical; if the metric is equal to zero, the trees are no more different than expected by chance; finally if the metric is less than zero, the trees are more different than expected by chance.

Note that once rescaled, the Normalised Robinson-Foulds metric is a measure of similarity, rather than of distance like the original Robinson-Foulds metric.

2.4.2 Triplets distance

The Triplets distance (Dobson, 1975) measures the number of sub-trees made up of three taxa that differ between two trees (Critchlow et al., 1996) (see Appendix B for calculation details). This metric measures the position of each taxon and clade in relation to its closest neighbours. It is bounded between zero when the two trees are identical and $\binom{n}{3}$ (for two trees with n taxa) when there is no shared taxa/clade position in the two trees. Therefore this metric sensitive to the conservation of wildcard taxa. We normalised this metric in the same way as for the Robinson-Foulds distance resulting in the Normalised Triplets metric.

2.4.3 Paired tree comparisons

For the Maximum Likelihood and Bayesian consensus trees we performed pairwise comparisons between the "best" tree and each "missing-data" tree using both the Normalised Robinson-Foulds and Normalised Triplets metrics with the TreeCmp java script (Bogdanowicz et al., 2012) resulting in 125 Normalised Robinson-Foulds metrics and 125 Normalised Triplets metric for each tree inference method. Also, to take into account the uncertainty of tree inference, we extracted 1000 random bootstrapped trees from the Maximum Likelihood analysis and 1000 trees from the posterior tree distribution of the Bayesian analysis for the "best" trees, and then did the same for the

125 “missing data” trees (resulting in 1000 “best” trees and 125×1000 “missing data” trees). For a given set of 1000 “missing data” trees and the 1000 “best” trees, we sampled one “missing data” tree and one “best” tree at random and compared them using both the Normalised Robinson-Foulds and Normalised Triplets metrics as described above. We repeated this 1000 times for each set of “missing data” trees resulting in 125×1000 values for each metric. We repeated all the paired tree comparisons described above for each of the 50 simulation runs. We then calculated the mode and the 50% and 95% confidence intervals from the resulting distribution using the hdrcde R package v. 3.1 (Hyndman et al., 2013).

2.5 Testing the effects of the missing data parameters on topological recovery

Finally, we tested the effects of our missing data parameters (M_L, M_F, M_C and their interactions) on our ability to recover the “best” tree topology in a Total Evidence framework. We also assessed the effect of our missing data parameters jointly with the effects of different tree inference and uncertainty methods (i.e. Maximum Likelihood, Bayesian consensus, Maximum Likelihood bootstrap trees and Bayesian posterior tree distribution).

We measured similarities among the distributions of the different metrics scores (Normalised Robinson-Foulds and Normalised Triplets metric) using the Bhattacharyya Coefficient (Bhattacharyya, 1943). The Bhattacharyya Coefficient is the probability of overlap between two distributions (Bhattacharyya, 1943) (see Appendix B for calculation details). Note that this is comparable to performing a two-sided t-test, but we use the Bhattacharyya Coefficient here because we are comparing whole distributions not just their means. To assess the effect of our missing data parameters, we calculated the Bhattacharyya Coefficient between the distributions of the different metrics scores (Normalised Robinson-Foulds and Normalised Triplets metric) for each pairwise combination of missing data parameters (M_L, M_F, M_C) and parameter states (0%, 10%, 25%, 50%, 75%), i.e. M_L = 0%, M_F = 0%, M_C = 0%; M_L = 10%, M_F = 0%, M_C = 0% etc. (see Figure 2 for more details). This resulted in 7875 pairwise comparisons (a triangular matrix with $3^5 \times 3^5$ cells). We performed this procedure separately for each tree inference and uncertainty method. When two combinations of missing data parameters have a similar ability to recover the "best" tree topology the Bhattacharyya Coefficient will be close to one. Conversely, if the two combinations of missing data parameters differ, the Bhattacharyya Coefficient will be close to zero.

To assess the effect of the different tree inference and uncertainty methods (i.e. Maximum Likelihood, Bayesian consensus, Maximum Likelihood bootstrap trees and Bayesian posterior tree distribution) on our ability to recover the "best" tree topology, we calculated the Bhattacharyya Coefficient between the distributions of the different metrics scores (Normalised Robinson-Foulds and Normalised Triplets metric) for each pairwise combination of tree inference and uncertainty methods, i.e. Maximum Likelihood *versus* Bayesian consensus; Maximum Likelihood *versus* Maximum

Likelihood bootstrap trees etc. (see Figure 3 for more details). Note that this procedure pools results from across all missing data parameter combinations so it results in just six pairwise comparisons. When two tree inference or uncertainty methods have a similar ability to recover the "best" tree topology the Bhattacharyya Coefficient will be close to one. Conversely, if the two tree inference or uncertainty methods differ, the Bhattacharyya Coefficient will be close to zero.

3 Results

As the amount of missing data in the morphological part of the Total Evidence matrix increases, our ability to recover the "best" tree topology decreases, regardless of the missing data parameter (M_L, M_F or M_C), the tree inference method (Maximum Likelihood or Bayesian) or the tree comparison metric used (Normalised Robinson-Foulds or Normalised Triplets metric). However, the different missing data parameters and tree inference methods do not affect the topology in the same way (Figure 4 and Figure 5).

3.1 Individual effects of missing data parameters

As the amount of missing data increases across all three parameters, our ability to recover the "best" tree topology decreases (Figure 4). The Normalised Robinson-Foulds metric is always lower for the Maximum Likelihood trees than for the Bayesian consensus trees (median Bhattacharrya Coefficient = 0.69, 0.48 and 0.66 for M_L, M_F and

M_C respectively; Figure 4; Tables C5, C6 and C7 in Appendix C). However, The Normalised Triplets metric is similar between the Maximum Likelihood trees and the Bayesian consensus trees for all the parameters (M_L, M_F and M_C) (median Bhattacharrya Coefficient = 0.84, 0.75 and 0.80 for M_L, M_F and M_C respectively; Figure 4; Tables C5, C6 and C7 in Appendix C).

3.2 Combined effect of missing data parameters

As expected, our ability to recover the "best" tree topology is worst when each parameter contains the maximum amount of missing data (i.e. M_L = 75%, M_F = 75% and M_C = 75%), and best when there is no missing data (i.e. M_L = 0%, M_F = 0%, M_C = 0%; Figure 5; Tables C2, C3 and C4 in Appendix C). Figure **??** shows the similarity of distributions of tree metrics in a triangular matrix with the values of each pairwise Bhattacharyya Coefficient coloured according to their values (orange when the distributions overlap completely, Bhattacharyya Coefficient = 1, and blue when they do not, Bhattacharyya Coefficient = 0; Figure 6).

Using both Normalised Robinson-Foulds and Normalised Triplets metrics from the Bayesian consensus trees, the parameter combination with no missing data (i.e. M_L = 0%, M_F = 0%, M_C = 0%) is always the most dissimilar to all the other parameter combinations (Figure 6). However, the Normalised Robinson-Foulds metric (median Bhattacharrya coefficient = 0.79; Figure 6A) displays more dissimilarities than the Normalised Triplets metric (median Bhattacharrya coefficient = 0.81; Figure 6B).

Additionally, when using the Normalised Robinson-Foulds metric, once $M_L \geq 50\%$, there is no additional affect of M_F and M_C, regardless of the amount of missing data in these parameters (Figure 6A). Likewise, once $M_C \geq 50\%$, there is no additional affect of M_L and M_F (Figure 6A).

For all combinations of missing data parameters and tree comparison metrics, the Maximum Likelihood bootstrap trees and the Bayesian posterior tree distributions perform very similarly (median Bhattacharrya Coefficient = 0.85 and 0.98, using Normalised Robinson-Foulds metric or Normalised Triplets metric respectively; Table 1). However, these two methods perform worse than the Bayesian consensus trees using Normalised Robinson-Foulds metric (median Bhattacharrya Coefficient = 0 and 0.01, for the Maximum Likelihood bootstrap trees and the Bayesian posterior tree distribution respectively; Table 1; Figure 4 and Figure C2 in Appendix C).

4 Discussion

Our results show that the ability to recover the "best" tree topology in a Total Evidence framework decreases as the amount of missing data increases, regardless of how data were removed or the method of tree inference used. However, these factors affected topological recovery in different ways and to different extents. Decreasing the amount of living taxa with morphological data (M_L) and the overall number of morphological characters in the matrix (M_C) had worst effects on topological recovery (Figure 6). Additionally, using Bayesian consensus trees recovered the "best" tree topology more consistently than using Maximum Likelihood trees or Bayesian posterior tree distributions (Figure 5, Figure 6, Table 1). As seen in previous studies, our results show that the amount of missing data is not a problem *per se* for Total Evidence methods, as long as enough living and fossil taxa in the matrix have data for overlapping morphological characters (e.g. Kearney, 2002; Wiens, 2003; Roure and Philippe, 2011; Pattinson et al., 2014).

4.1 Individual effects of missing data parameters

4.1.1 Missing data for living taxa (M_L)

When the number of living taxa with morphological data (M_L) decreases, entire rows of data are being removed from the living taxa part of the matrix. Because living taxa still have molecular characters available for phylogenetic inference (see Methods),even if they have no morphological data, the relationships among them will always be fairly well-resolved (depending on the phylogenetic signal from the molecular part of the matrix). However, this missing data parameter has a huge influence on the placement of fossil taxa because a decrease in the M_L parameter reduces the amount of overlapping data among the living and fossil taxa, meaning there is no part of the living taxa tree that the fossils can branch off.

4.1.2 Missing data for fossil taxa (M_F)

When the overall proportion of data for the fossil taxa (M_F) decreases, this also reduces the probability of morphological characters for fossil taxa overlapping with the ones for living taxa. This can lead to difficulties for the placement of certain taxa in the tree. However, it is important to note that even though the number of displaced wildcard taxa increases (i.e. decrease of Normalised Triplets metric) with increasing missing data in this parameter, clade conservation (i.e. Normalised Robinson-Foulds metric) is still relatively good (mode = 0.72) when the proportion of missing data is high (M_F = 75%).

The effect of the missing data in the fossil record (M_F) is less than the effect of the M_L parameter on clade conservation (Normalised Robinson-Foulds metric) but greater on the displacement of wildcard taxa (Normalised Triplets metric; Figure 4 and Figure 5). This is related to the fact that the Bayesian consensus tree is built using a majority consensus rule. When the fossil taxa have less data (e.g. M_F = 75%) they will tend to branch with any taxon in the clade that shares most characters with the fossils. Therefore a majority consensus position is unlikely to exist (i.e. every branching position is represented in $<$ 50% of the trees in the Bayesian posterior distribution) and the fossil taxa will form a polytomy at the base of the clade. In this case, the Normalised Robinson-Foulds metric will decrease when the fossil is present at a lower taxonomic level (i.e. separated with more nodes from the root) but affects the clade conservation less at higher taxonomic level (i.e. separated with less nodes from the root). Conversely, because a fossil in a high taxonomic level clade has many chances to branch on different nodes within the clade, it will be more likely to act as a wildcard taxa and decrease the Normalised Triplets metric. Therefore, the M_F parameter is likely to affect the Normalised Robinson-Foulds metric less than the Normalised Triplets metric for the Bayesian consensus trees. Conversely, the same scenario in a Maximum Likelihood framework will lead to a dichotomous branching of the fossils but with low bootstrap support (< 50). In other words, the Bayesian consensus tree allows a fossil taxon with few data to be placed with a higher confidence at a lower taxonomic level than the Maximum Likelihood tree, where the fossil will be placed with lower confidence at a higher taxonomic level. We argue that using the Bayesian consensus tree topology is preferable because it is more conservative (e.g. Pattinson et al., 2014).

4.1.3 Missing morphological characters (M_C)

Reducing the overall number of morphological characters reduces the probability of their overlap among the taxa in the matrix, and therefore decreases our ability to recover the "best" tree topology. We expected the decrease in this parameter to have an effect twice as large as that for the M_L and M_F parameters, because removing 10% of the data for the fossil or living taxa only removes 5% of data from the whole matrix (because this parameter affects only half of the taxa present in the matrix). Conversely, removing 10% of morphological characters (M_C) genuinely removes 10% of data in the matrix. However, the effect of removing characters on the ability to recover the "best" tree topology is of the same order of magnitude as for the other two parameters (Figure 4). We suspect this again reflects the importance of overlapping characters, as opposed to the number of characters *per se*.

Additionally, the number of morphological characters determines the size the matrix. This can affect our ability to recover the "best" tree topology through: (1) the incongruence of phylogenetic signal among morphological and molecular data; and/or (2) homoplasy. The incongruence of phylogenetic signal between morphological and molecular data has previously been demonstrated to be more important in small morphological matrices (Wagner, 2000). The size of our data matrices were constrained by the performance of our protocol: to reduce the computational time of our analysis to a reasonable level (150 CPU years), we ran our simulations on modestly-sized matrices of 1000 molecular characters and 100 morphological characters. Therefore, part of the decrease of the Normalised Robinson-Foulds metric and the Normalised Triplets metric in our simulations could be due to conflicting phylogenetic signal among morphological and molecular data in our matrices (Figure 4 and Figure 5). However, although these matrices are an order of magnitude smaller than some published matrices (e.g. Springer et al., 2012; Ni et al., 2013, they are still within the size range of more modestly-sized empirical matrices (e.g. Kelly et al., 2014; Sallam et al., 2011). Therefore, our simulations reflect realistic parameters. Homoplasy, on the other hand, is expected to increase with an increase in the number of morphological characters (Wright and Hillis, 2014). However, the use of probabilistic methods (i.e. Maximum Likelihood or Bayesian) and the M*kv* model (Lewis, 2001) has been previously demonstrated to partially resolve this issue (Wright and Hillis, 2014).

4.2 Combined effect of missing data parameters

As expected, when combining the missing data parameters, our ability to recover the "best" tree topology is affected in the same way as for the parameters individually: the Normalised Robinson-Foulds metric and the Normalised Triplets metric are higher when all the missing data parameters have few missing data (i.e. M_L = 0%, M_F = 0%, M_C = 0%) and lower when they have a lot of missing data (i.e. M_L = 75%, M_F = 75% and M_C = 75%; Figure 5). However, it is important to notice that the effect of each parameter is not additive. The number of missing living taxa with morphological data (M_L) and the overall number of missing morphological characters (M_C), have a bigger effect than the amount of missing data for the fossil taxa (M_F), and when both M_L and/or M_C reach 50% of missing data, the matrix does not contain enough phylogenetic information for the fossil taxa to be placed with confidence in the tree (Figure 6).

4.3 Effects of tree inference methods

Variation in our ability to recover the "best" tree topology depends heavily on the tree inference method (Figure 4 and Figure 5). For morphological data, previous studies have shown some superiority of probabilistic tree inference methods with simple evolutionary models such as the M*kv* model (Lewis, 2001) over cladistic methods (Wright and Hillis, 2014) (but see Spencer and Wilberg, 2013). However, this is the first study, to our knowledge, to compare the performance of the M*kv* model (Lewis, 2001) for recovering the “best” tree topology using Maximum Likelihood and Bayesian methods in a Total Evidence framework. Our results show that the topology of the Bayesian consensus tree is always closer to the “best” tree topology than the one of the Maximum Likelihood tree (Figure 5). As described above, this is because the Bayesian consensus tree allows a fossil taxon with few data to be placed with a higher confidence at a lower taxonomic level than the Maximum Likelihood tree. This may also be because the “best” Bayesian consensus trees are not completely resolved, thus will always be more similar to the “missing data” trees than a completely resolved tree like the “best” Maximum Likelihood tree. However, we minimized the probability of unresolved “best” trees in our Bayesian analyses by only using datasets with strong phylogenetic signal (see Methods).

It is also worth noting that across all our analyses, the topologies of the Maximum Likelihood bootstrap trees and the Bayesian posterior trees distribution were always further from the “best” tree topology than Maximum Likelihood and Bayesian consensus trees. This was true even when no morphological data was missing (M_L = 0%; M_F = 0%, M_C = 0%; Figure 4). This reflects the fact that it is difficult to compare two distributions of trees, and each comparison between a set of “missing data” trees and a set of the “best” trees involved 1000 random pairwise comparisons rather than just one. This will obviously add noise to the results for these methods.

4.4 Practical implications

Our missing data parameters illustrate different sources of missing data in empirical matrices as follows: (M_L) the paucity of coded morphological characters for living taxa; (M_F) the missing data for fossils (or parts of fossils) that have not been preserved in the fossil record; and (M_C) characters that have not been coded across living and fossil species, perhaps due to difficulties in coding or poor preservation of the feature in collections. Filling these gaps in empirical Total Evidence matrices should lead to a substantial increase in our ability to recover the “best” tree topology. We can increase the number of living taxa with coded morphological characters by increasing research efforts in this area, and encouraging use of our vast natural history collections. Increasing data for fossil species is harder, since it depends on fossil preservation biases and new fossil discoveries. However, gaps in the matrix can be filled with efforts in palaeontological field work that can potentially lead to future discoveries of exceptionally preserved fossils (e.g. Ni et al., 2013). Fortunately, although this data is the most difficult to collect, it also has the least influence on whether our simulations recover the “best” tree topology (Figure 6). Finally, although increasing the number of coded characters is relatively straightforward, the amount of time it takes to build a morphological matrix increases directly with the number of characters involved. One solution to this problem may be to engage with collaborative data collection projects through web portals such as *Morphobank* (O’Leary and Kaufman, 2011), so that no one individual collects all the data.

Another practical implication of our results regards the tree inference methods. Because the Bayesian consensus trees consistently recovered topologies closer to the “best” tree topology than the Maximum Likelihood trees, we advise using Bayesian consensus trees to fix the topology where necessary for further steps in tree inference, for example in tip-dating (Ronquist et al., 2012a; Matzke, 2014); although it is possible that including dating information during tree inference could improve the accuracy of the Bayesian posterior tree distribution). Note that, although the Bayesian posterior tree distribution performed poorly in recovering the “best” tree topology, this is due to difficulties comparing distributions of trees (see above). Thus, we do not suggest discarding the Bayesian posterior tree distribution once the topology has been fixed, particularly because these trees will be invaluable for phylogenetic comparative analyses (e.g. Jetz et al., 2012).

5 Conclusions

Missing data in Total Evidence matrices is not a problem for recovering the “best” tree topology as long as enough living and fossil taxa in the matrix have data for overlapping morphological characters. When missing data increases in any of our missing data parameters (M_L, M_F or M_C), it reduces support for the placement of fossil taxa and increases the displacement of wildcard taxa. Therefore we advise filling as many gaps in Total Evidence matrices as possible. Because this is difficult, if not impossible, for fossil data, we recommend coding as many morphological characters, for as many living taxa, as possible. Additionally, the topology of the Bayesian consensus trees, regardless the amount of missing data, were always closer to the "best" tree topology than the Maximum Likelihood trees. Therefore, we advise using Bayesian consensus tree topologies if topology must be fixed for furhter analyses. The results of our analyses are encouraging and show that it is possible to combine both neontological and palaeontological data in the same phylogeny despite issues of missing data. Hopefully, using these approaches will greatly improve our understanding of macroevolutionary patterns and processes.

6 Acknowledgments

Thanks to Gavin Thomas, Frédéric Delsuc, Emmanuel Douzery, Trevor Hodkinson, Andrew Jackson, Nick Matzke, and April Wright for useful comments on our simulation protocol and manuscript. Thanks to Paddy Doyle, Graziano D'Innocenzo and Sean McGrath for assistance with the computer cluster. Simulations used the Lonsdale cluster maintained by the Trinity Centre for High Performance Computing and funded through grants from Science Foundation Ireland. This work was funded by a European Commission CORDIS Seventh Framework Programme (FP7) Marie Curie CIG grant (proposal number: 321696).

References

Bapst, D. W. 2013. A stochastic rate-calibrated method for time-scaling phylogenies of fossil taxa. Methods in Ecology and Evolution 4:724–733.

Beck, R. M. and M. S. Lee. 2014. Ancient dates or accelerated rates? Morphological clocks and the antiquity of placental mammals. Proceedings of the Royal Society B: Biological Sciences 281:1–10.

Bhattacharyya, A. 1943. On a measure of divergence between two statistical populations defined by their probability distributions. Bulletin of the Calcutta Mathematical Society 35:99–109.

Bogdanowicz, D., K. Giaro, and B. Wróbel. 2012. TreeCmp: Comparison of trees in polynomial time. Evolutionary Bioinformatics 8:475–487.

Chen, W.-C. 2011. Overlapping codon model, phylogenetic clustering, and alternative partial expectation conditional maximization algorithm. Ph.D. thesis.

Critchlow, D. E., D. K. Pearl, and C. Qian. 1996. The triples distance for rooted bifurcating phylogenetic trees. Systematic Biology 45:323–334.

Dávalos, L. M., P. M. Velazco, O. M. Warsi, P. D. Smits, and N. B. Simmons. 2014. Integrating incomplete fossils by isolating conflicting signal in saturated and non-independent morphological characters. Systematic Biology 63:582–600.

Dietl, G. P. and K. W. Flessa. 2011. Conservation paleobiology: putting the dead to work. Trends in Ecology and Evolution 26:30–37.

Dobson, A. J. 1975. Comparing the shapes of trees vol. 452 of *Lecture Notes in Mathematics* Pages 95–100. Springer Berlin Heidelberg.

Douady, C., F. Delsuc, Y. Boucher, W. Doolittle, and E. Douzery. 2003. Comparison of bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Molecular Biology and Evolution 20:248–254.

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

Eernisse, D. and A. Kluge. 1993. Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology. Molecular Biology and Evolution 10:1170–1195.

FitzJohn, R. G. 2012. Diversitree: comparative phylogenetic analyses of diversification in R. Methods in Ecology and Evolution 3:1084–1092.

Friedman, M. 2010. Explosive morphological diversification of spiny-finned teleost fishes in the aftermath of the end-Cretaceous extinction. Proceedings of the Royal Society B: Biological Sciences 277:1675–1683.

Fritz, S. A., J. Schnitzler, J. T. Eronen, C. Hof, K. Bhning-Gaese, and C. H. Graham. 2013. Diversity in time and space: wanted dead and alive. Trends in Ecology and Evolution 28:509 – 516.

Hasegawa, M., H. Kishino, and T. A. Yano. 1985. Dating of the human ape splitting by a molecular clock of mitochondrial-DNA. Journal of Molecular Evolution 22:160–174.

Heath, T. A., J. P. Huelsenbeck, and T. Stadler. 2014. The fossilized birthdeath process for coherent calibration of divergence-time estimates. Proceedings of the National Academy of Sciences 111:E2957–E2966.

Hyndman, R. J., J. Einbeck, and M. Wand. 2013. hdrcde: Highest density regions and conditional density estimation. R package version 3.1.

Jackson, J. and D. Erwin. 2006. What can we learn about ecology and evolution from the fossil record? Trends in Ecology and Evolution 21:322–328.

Jetz, W., G. Thomas, J. Joy, K. Hartmann, and A. Mooers. 2012. The global diversity of birds in space and time. Nature 491:444–448.

Kearney, M. 2002. Fragmentary taxa, missing data, and ambiguity: mistaken assumptions and conclusions. Systematic Biology 51:369–381.

Kelly, S. B. A., D. J. Kelly, N. Cooper, A. Bahrun, K. Analuddin, and N. M. Marples. 2014. Molecular and phenotypic data support the recognition of the Wakatobi flowerpecker (*Dicaeum kuehni*) from the unique and understudied Sulawesi region. PLoS ONE 9:e98694.

Kuhner, M. K. and J. Yamato. 2014. Practical performance of tree comparison metrics. Systematic Biology .

Lemmon, A., J. Brown, S. Kathrin, and E. Lemmon. 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and bayesian inference. Systematic Biology 58:130–145.

Lewis, P. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Systematic Biology 50:913–925.

Manos, P., P. Soltis, D. Soltis, S. Manchester, S. Oh, C. Bell, D. Dilcher, and D. Stone. 2007. Phylogeny of extant and fossil Juglandaceae inferred from the integration of molecular and morphological data sets. Systematic Biology 56:412–430.

Matzke, N. J. 2014. Beastmaster: Automated conversion of nexus data to beast2 xml format, for fossil tip-dating and other uses. `http://phylo.wikidot.com/beastmaster`.

Meredith, R., J. Janečka, J. Gatesy, O. Ryder, C. Fisher, E. Teeling, A. Goodbla, E. Eizirik, T. L. Simão, T. Stadler, D. Rabosky, R. Honeycutt, J. Flynn, C. Ingram, C. Steiner, T. Williams, T. Robinson, B. Angela, M. Westerman, N. Ayoub, M. Springer, and W. Murphy. 2011. Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. Science 334:521–524.

Ni, X., D. Gebo, M. Dagosto, J. Meng, P. Tafforeau, J. Flynn, and K. Beard. 2013. The oldest known primate skeleton and early haplorhine evolution. Nature 498:60–64.

Novacek, M. J. and Q. Wheeler. 1992. Extinction and phylogeny. Columbia University Press.

Nylander, J. A. A., F. Ronquist, J. P. Huelsenbeck, and J. Nieves-Aldrey. 2004. Bayesian phylogenetic analysis of combined data. Systematic Biology 53:47–67.

O'Leary, M. A., J. I. Bloch, J. J. Flynn, T. J. Gaudin, A. Giallombardo, N. P. Giannini, S. L. Goldberg, B. P. Kraatz, Z.-X. Luo, J. Meng, X. Ni, M. J. Novacek, F. A. Perini, Z. S. Randall, G. W. Rougier, E. J. Sargis, M. T. Silcox, N. B. Simmons, M. Spaulding, P. M. Velazco, M. Weksler, J. R. Wible, and A. L. Cirranello. 2013. The placental mammal ancestor and the postK-Pg radiation of placentals. Science 339:662–667.

O'Leary, M. A. and S. Kaufman. 2011. Morphobank: phylophenomics in the cloud. Cladistics 27:529–537.

Pagel, M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proceedings of the Royal Society of London. Series B: Biological Sciences 255:37–45.

Paradis, E. 2011. Time-dependent speciation and extinction from phylogenies: a least squares approach. Evolution 65:661–672.

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.

Pattengale, N. D., M. Alipour, O. R. Bininda-Emonds, B. M. Moret, and A. Stamatakis. 2010. How many bootstrap replicates are necessary? Journal of Computational Biology 17:337–354.

Pattinson, D. J., R. S. Thompson, A. K. Piotrowski, and R. J. Asher. 2014. Phylogeny, paleontology, and primates: Do incomplete fossils bias the tree of life? Systematic Biology Pages 1–18.

Pearman, P., A. Guisan, O. Broennimann, and C. Randin. 2008. Niche dynamics in space and time. Trends in Ecology and Evolution 23:149–158.

Price, M. N., P. S. Dehal, and A. P. Arkin. 2010. Fasttree 2 approximately maximum-likelihood trees for large alignments. PLoS ONE 5:e9490.

Pyron, R. 2011. Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. Systematic Biology 60:466–481.

Quental, T. and C. Marshall. 2010. Diversity dynamics: molecular phylogenies need the fossil record. Trends in Ecology and Evolution 25:434–441.

R Core Team. 2014. R: a language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria.

Rambaut, A. and N. C. Grassly. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Computer Application in the Biosciences 13:235–8.

Raup, D. M. 1981. Extintion: bad genes or bad luck? Acta Geológica Hispánica 16:25–33.

Robinson, D. F. and L. R. Foulds. 1981. Comparison of phylogenetic trees. Mathematical Biosciences 53:131–147.

Ronquist, F., S. Klopfstein, L. Vilhelmsen, S. Schulmeister, D. Murray, and A. Rasnitsyn. 2012a. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. Systematic Biology 61:973–999.

Ronquist, F., M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Hohna, B. Larget, L. Liu, M. A. Suchard, and J. P. Huelsenbeck. 2012b. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology 61:539–42.

Roure, B. and H. Philippe. 2011. Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference. BMC Evolutionary Biology 11:17.

Rozen, D. E., D. Schneider, and R. E. Lenski. 2005. Long-term experimental evolution in *Escherichia coli*. XIII. Phylogenetic history of a balanced polymorphism. Journal of Molecular Evolution 61:171–80.

Salamin, N., M. W. Chase, T. R. Hodkinson, and V. Savolainen. 2003. Assessing internal support with large phylogenetic DNA matrices. Molecular Phylogenetics and Evolution 27:528–539.

Sallam, H. M., E. R. Seiffert, and E. L. Simons. 2011. Craniodental morphology and systematics of a new family of hystricognathous rodents (Gaudeamuridae) from the Late Eocene and Early Oligocene of Egypt. PloS ONE 6:e16525.

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

Sansom, R. S. and M. A. Wills. 2013. Fossilization causes organisms to appear erroneously primitive by distorting evolutionary trees. Scientific Reports 3:1–5.

Schrago, C., B. Mello, and A. Soares. 2013. Combining fossil and molecular data to date the diversification of New World Primates. Journal of Evolutionary Biology 26:2438–2446.

Simpson, G. 1944. Tempo and Mode in Evolution. Columbia University Biological Series Columbia University Press.

Slater, G. J. 2013. Phylogenetic evidence for a shift in the mode of mammalian body size evolution at the cretaceous-palaeogene boundary. Methods in Ecology and Evolution 4:734–744.

Slater, G. J. and L. J. Harmon. 2013. Unifying fossils and phylogenies for comparative analyses of diversification and trait evolution. Methods in Ecology and Evolution 4:699–702.

Spencer, M. R. and E. W. Wilberg. 2013. Efficacy or convenience? Model-based approaches to phylogeny estimation using morphological data. Cladistics 29:663–671.

Springer, M. S., R. W. Meredith, J. Gatesy, C. A. Emerling, J. Park, D. L. Rabosky, T. Stadler, C. Steiner, O. A. Ryder, J. E. Janeka, C. A. Fisher, and W. J. Murphy. 2012.

Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix. PLoS ONE 7:e49521.

Stadler, T. and Z. Yang. 2013. Dating phylogenies with sequentially sampled tips. Systematic Biology 62:674–688.

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

Stamatakis, A., P. Hoover, and J. Rougemont. 2008. A rapid bootstrap algorithm for the raxml web servers. Systematic Biology 57:758–771.

Tavaré, S. 1986. Some probabilistic and statistical problems in the analysis of dna sequences. Lectures on mathematics in the life sciences 17:57–86.

Wagner, P. J. 2000. Exhaustion of morphologic character states among fossil taxa. Evolution 54:365–386.

Wiens, J. J. 2003. Missing data, incomplete taxa, and phylogenetic accuracy. Systematic Biology 52:528–538.

Wiens, J. J. 2006. Missing data and the design of phylogenetic analyses. Journal of Biomedical Informatics 39:34–42.

Wiens, J. J. and D. S. Moen. 2008. Missing data and the accuracy of Bayesian phylogenetics. Journal of Systematic Evolution 46:307–314.

Wright, A. M. and D. M. Hillis. 2014. Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. PLoS ONE 9:e109210.

Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends in Ecology and Evolution 11:367–372.

Zander, R. H. 2004. Minimal values for reliability of bootstrap and jackknife proportions, decay index, and Bayesian posterior probability. Phyloinformatics 2:1–13.

Zuckerkandl, E. and L. Pauling. 1965. Molecules as documents of evolutionary history. Journal of Theoretical Biology 8:357–366.

Figure captions

Figure 1: Protocol outline. (1) We randomly generated a birth-death tree (the "true" tree) and used it to simulate a matrix with no missing data (the "complete" matrix). (2) We removed data from the morphological part of the "complete" matrix resulting in 125 "missing-data" matrices. (3) We built phylogenetic trees from each matrix using both Maximum Likelihood and Bayesian methods. (4) We compared the "missing-data" trees to the "best" tree. We repeated these four steps 50 times.

Figure 2: Bhattacharyya Coefficient calculation outline 1. A, B and C are distributions of tree similarity metrics (Normalised Robinson-Foulds or Normalised Triplets metrics) for any combination of missing data parameters (e.g. M_L = 10%, M_F = 50%, M_C = 75%). The Bhattacharyya Coefficient (BC) is the overlap of the distribution of tree similarity metrics between two combinations of missing data parameters, for example, BC(A,B) is the probability of overlap between the distributions A and B. Note that this is similar to performing a two-sided t-test, but we use the Bhattacharyya Coefficient here because we are comparing distributions not means.

Figure 3: Bhattacharyya Coefficient calculation outline 2. A and B are distributions of tree similarity metrics (Normalised Robinson-Foulds or Normalised Triplets metrics) for any combination of missing data parameters (e.g. M_L = 10%, M_F = 50%, M_C = 75%). **(x)** and **(y)** are two different tree inference methods (e.g. Maximum Likelihood or Bayesian). The Bhattacharyya Coefficient (BC) is the overlap of the distribution of tree similarity metrics between two methods for the same combination of missing data

parameters, for example, BC(A_x,A_y) is the probability of overlap of the distribution A for methods x and y. Note that this is similar to performing a two-sided t-test, but we use the Bhattacharyya Coefficient here because we are comparing distributions not means.

Figure 4: The effects of increasing missing data on topological recovery using Maximum Likelihood trees (black), Bayesian consensus trees (blue), Maximum Likelihood bootstrap trees (orange) and Bayesian posterior tree distributions (blue). The percentage of missing data for each parameter (M_L, M_F and M_C) is shown on the x axis. Topological recovery was measured using two different tree comparison metrics: Normalised Robinson-Foulds metric (upper row) and Normalised Triplets metric (lower row). The graph shows the modal value (points), and the 50% (thick solid lines) and 95% (thin dashed lines) confidence intervals of the distributions of the tree comparison metric for each missing data parameter and tree inference method.

Figure 5: The effects of increasing missing data on topological recovery using Maximum Likelihood trees (black) and Bayesian consensus trees (grey). The x axis shows the percentage of missing data from 0% (white) to 75% (black) for the three parameters: M_L (upper line), M_F (middle line) and M_C (lower line). Topological recovery was measured using two different tree comparison metrics: Normalised Robinson-Foulds metric (upper row) and Normalised Triplets metric (lower row). The graph shows the modal value (points), and the 50% (thick solid lines) and 95% (thin dashed lines) confidence intervals of the distributions of the tree comparison metric for

each missing data parameter and tree inference method.

Figure 6: The effects of missing data on topological recovery using Bayesian consensus trees. Both axes show the percentage of missing data from 0% (white) to 75% (black) for the three parameters: M_L (upper line), M_F (middle line) and M_C (lower line). Topological recovery is represented by the probability of (A) Normalised Robinson-Foulds metric and (B) Normalised Triplets metric distributions overlapping with the "best" tree distribution, calculated using the Bhattacharyya Coefficient. The Bhattacharyya Coefficient values are indicated using a color gradient ranging from low probability of overlap in blue, to a high probability of overlap in orange.

Tables

Table 1: Bhattacharyya Coefficients of the pairwise method comparisons. Each comparison which corresponds to the normalised metric between the "best" tree and the "missing data" trees using either the Normalised Robinson-Foulds metric (RF) or the Normalised Triplets metric (Tr). Each line summarizes the distribution of the probability of overlap between pairs of tree inference methods. Note that this is equivalent to performing a two-sided t-test, but we use the Bhattacharyya Coefficient here because we are comparing distributions not means. The values highlighted in bold are the extreme values of high or low probability of overlap between two methods. If two methods have a high probability of overlap, they have a similar ability to recover the "correct" tree topology.

Table 1

Comparison	Metric	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Maximum Likelihood *vs.* Bayesian consensus	*RF*	0.00	0.00	0.10	0.20	0.32	1.00
	Tr	0.34	0.49	0.61	0.62	0.75	1.00
Maximum Likelihood *vs.* Maximum Likelihood bootstraps	*RF*	0.03	0.54	0.69	0.64	0.77	0.98
	Tr	0.08	0.57	0.65	0.64	0.73	0.82
Maximum Likelihood *vs.* Bayesian posterior trees	*RF*	0.02	0.74	0.80	0.79	0.89	0.98
	Tr	0.21	0.67	0.73	0.72	0.77	0.84
Bayesian consensus *vs.* Maximum Likelihood bootstraps	*RF*	0.00	0.00	**0.00**	**0.01**	0.01	0.04
	Tr	0.08	0.38	0.59	0.57	0.73	0.84
Bayesian consensus *vs.* Bayesian posterior trees	*RF*	0.00	0.00	**0.01**	**0.02**	0.04	0.11
	Tr	0.21	0.36	0.56	0.55	0.74	0.87
Bayesian posterior tree *vs.* Maximum Likelihood bootstraps	*RF*	0.50	0.77	**0.85**	**0.85**	0.96	1.00
	Tr	0.91	0.96	**0.98**	**0.97**	0.99	1.00