

RH: Missing data and topology in total evidence approach

## **Effect of missing data on topological inference using a total evidence approach**

THOMAS GUILLERME<sup>1,2</sup>, AND NATALIE COOPER<sup>1,2</sup>

<sup>1</sup>*School of Natural Sciences, Trinity College Dublin, Dublin 2, Ireland;*

<sup>2</sup>*Trinity Centre for Biodiversity Research, Trinity College Dublin, Dublin 2, Ireland;*

**Corresponding author:** Thomas Guillerme, School of Natural Sciences, Trinity College Dublin, Dublin 2, Ireland; E-mail: guillert@tcd.ie; Fax: +353 1 6778094; Tel: +353 1 896 2571.

## Abstract

Living species represent only a small proportion of all species that have ever lived. Ignoring fossil taxa may therefore lead to misinterpretation of macroevolutionary patterns and processes such as trends in species richness, biogeographical history or paleoecology. This fact has led to an increasing consensus among scientists that both living and fossil taxa must be included in macroevolutionary studies. One approach called Total Evidence, uses molecular data from living taxa and morphological data from both living and fossil taxa to infer phylogenies with both living and fossil taxa at the tips. Although the Total Evidence approach seems very promising, it requires a lot of data and is therefore likely to suffer from missing data issues which may affect its ability to infer correct phylogenies.

In this study we assess the effect of missing data on tree topologies inferred from Total Evidence matrices. Using simulations we investigate three major factors that directly affect the completeness of the morphological part of the matrix: (1) the proportion of living taxa with no morphological data, (2) the amount of missing data in the fossil record, and (3) the overall number of morphological characters in the matrix.

We find that, when using a clade conservative metric such as Robinson-Foulds distance, Bayesian consensus trees recovers the right topology better than Maximum Likelihood or than a subset of Bayesian posterior tree distributions and this regardless of the amount of missing data (minimal tree similarity of 0.69 in the worst scenario). Regarding the data, good topological recovery is not related to the amount of missing data *per se* but to the amount of data overlap. The two factors that influences the most this data overlap are the parameters 1 (proportion of coded living taxa) and 3 (number of morphological characters).

(Keywords: missing data, Total Evidence, Bayesian, Maximum Likelihood, topology)

## INTRODUCTION

Although most species that have ever lived are now extinct (Novacek and Wheeler 1992; Raup 1993), 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 1945). 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. 2013) 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 a lot of data that can be difficult (or impossible) to collect. The 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). 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) (Wiens 2003, 2006; Wiens and Moen 2008; 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; 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 of the morphological part of the matrix: (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. These parameters reflect empirical biases in data availability - the advent of molecular phylogenetics means that morphological data for living species is rarely collected, and few people have the skills to identify characters needed for detailed phylogenetic analysis, missing data in fossil taxa is very common due to preservation biases (Sansom and Wills 2013), and the overall number of characters depends on the effort of the people identifying them (e.g. O'Leary et al. 2013). 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 the ability of Total Evidence methods to recover the "best" tree topology is increased when using Bayesian methods. Also 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.

## 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; Fig. 1).

### 1. Generating the matrix:

We randomly generated a birth-death tree (hereafter called the "true" tree) and used it to infer 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. Inferring 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.

### *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 ( $\lambda$ ) and extinction ( $\mu$ ) rates from a uniform distribution but maintaining  $\lambda > \mu$  (Paradis 2011). We implemented a rejection sampling algorithm to select only trees with 25 living and 25 fossil taxa to ensure that we had enough taxa of each type for our missing data simulations to work. 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 and transition/transversion rate of two (Douady et al. 2003). The substitution rates were distributed following a gamma distribution with an alpha ( $\alpha$ ) shape of 0.5 (Yang 1996). We chose a low value of  $\alpha$  to reduce the number of sites with high substitution rates, thus avoiding too much homoplasy and a decrease in phylogenetic signal. We selected

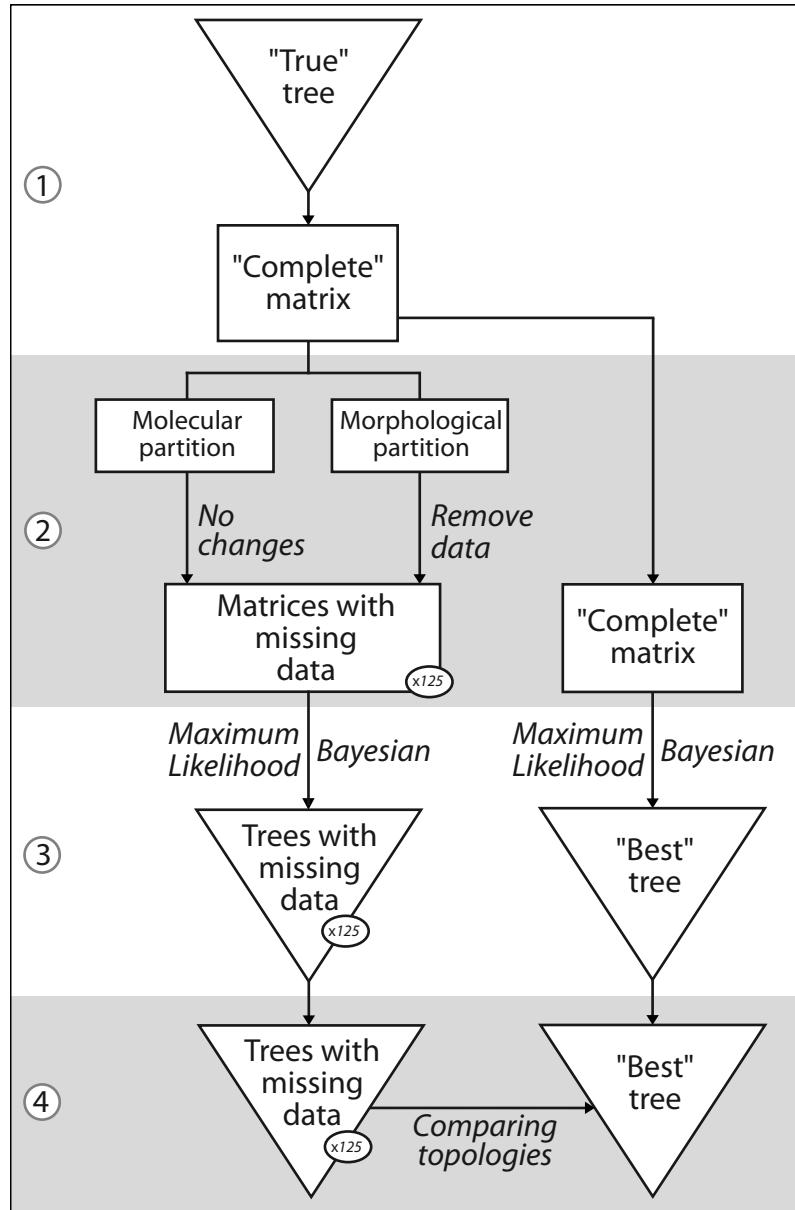


Figure 1: Protocol outline. (1) We randomly generated a birth-death tree (the "true" tree) and used it to infer a matrix with no missing data (the "complete" matrix). (2) We removed data from the morphological part of the "complete" matrix resulting in 125 "missing-data" matrices. (3) We built phylogenetic 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.

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 Supplementary Material Section 1). 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; Davalos 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. 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 detailed morphological data for living species are scarce.
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.
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.

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 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.

### *3. Building 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.

*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 +  $\Gamma_4$  model (Tavaré 1986; default GTRGAMMA in RAxML v. 8.0.20; Stamatakis 2014). For the morphological data, we used the implemented Markov  $k$  state model (Lewis 2001) assuming an equal state frequency and a unique overall substitution rate ( $\mu$ ) following a gamma distribution of the rate variation with four distinct categories (Mk +  $\Gamma_4$ ; -K MK option in RAxML v. 8.0.20; Stamatakis 2014).

To measure the phylogenetic signal of our simulations, 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  $50 \times 125$  "missing-data" matrices) with a relative good phylogenetic signal (median bootstrap  $> 50$ ). 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  $\sim 8$  CPU years to build 50 sets of 125 bootstrapped Maximum Likelihood trees (2.30GHz clock speed nodes).

*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 +  $\Gamma_4$ ). For the morphological data, we used the Markov  $k$  state model (Lewis 2001), with equal state frequency and a unique overall substitution rate ( $\mu$ ) with four distinct rates categories (Mk +  $\Gamma_4$ ). 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  $5 \times 10^7$  generations. We used the following priors for each tree: (i) the "true" trees topology as a starting tree (with a starting value for each branch length of one), (ii) an exponential prior on the shape of the gamma distribution of  $\alpha = 0.5$  for both partitions, and (iii) a transition/transversion ratio prior of two sampled from a strong beta distribution ( $\beta(80,40)$ ). 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  $\sim 140$  CPU years to build 50 sets of 125 Bayesian trees (2.30GHz clock speed nodes). The detailed MrBayes parameters are available in the Supplementary Material S1.

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).

#### *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 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.

*Robinson-Foulds distance.*— The Robinson-Foulds distance (Robinson and Foulds 1981), or "path difference", measures the 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 Supplementary Material S2 for calculation details). This metric is bounded between one, when the two trees are identical, and  $n - 2$  (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 by dividing the Robinson-Foulds distance by the mean Robinson-Foulds distance from 1000 random tree comparisons (Bogdanowicz et al. 2012). This allows us to compare the metric for any number of taxa by scaling it with the expected value of the metric when comparing two random trees (see Supplementary Material S2 for the calculation details). Therefore, when the Normalised Robinson-Foulds distance between two trees is equal to one, the trees are identical; if the distance is equal to zero, the trees are no more different than expected by chance; finally if the distance is less than zero, the trees are more different than expected by chance.

*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 Supplementary Material S2 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  $\binom{n}{4}$  (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 distance.

*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 Robison-Foulds distances and 125 Normalised Triplets distance 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  $125 \times 1000$  "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  $125 \times 1000$  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).

### *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 distance) using the Bhattacharyya Coefficient (Bhattacharyya 1943). The Bhattacharyya Coefficient is the probability of overlap between two distributions (Bhattacharyya 1943; see Supplementary Material Section 2 for calculation details). 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. 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 distance) 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 Fig. 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 distance) 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 Fig. 2 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.

## RESULTS

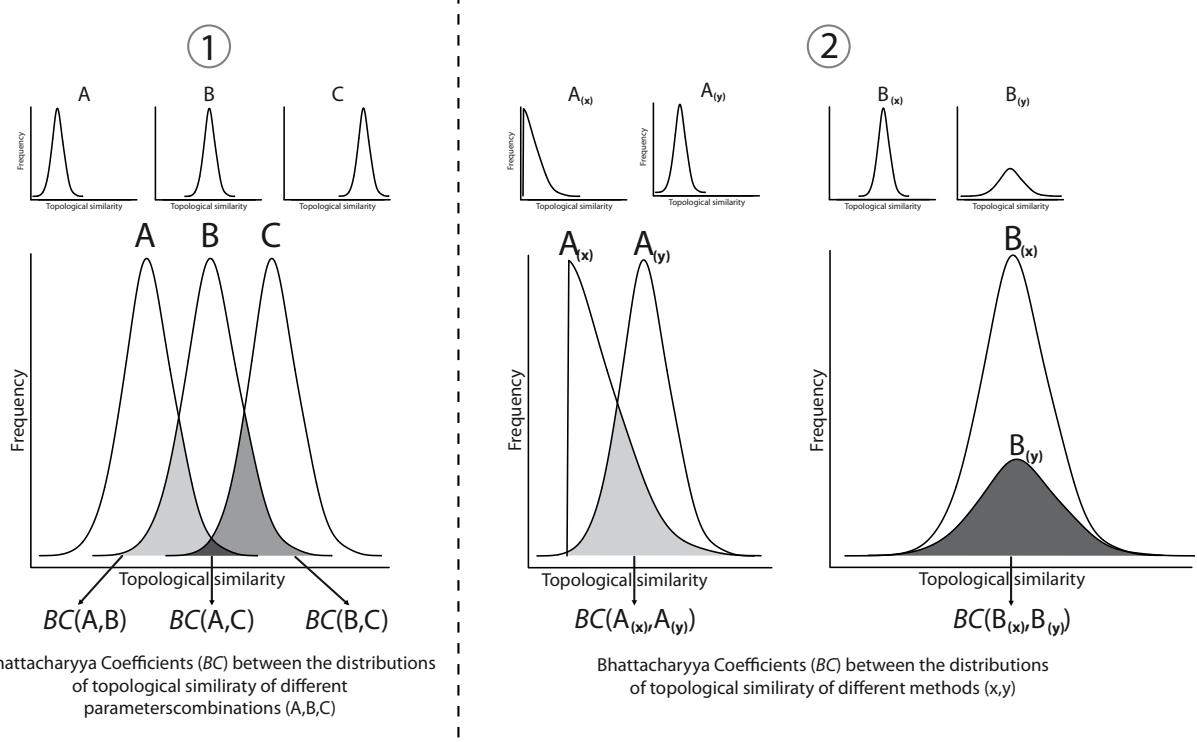


Figure 2: Bhattacharyya Coefficient calculation outline. A, B and C are distribution of tree similarity (Normalised Robinson-Foulds or Normalised Triplets distances) for any combination of missing data parameters (e.g.  $M_L = 10\%$ ,  $M_F = 50\%$ ,  $M_C = 75\%$ ). (x) and (y) are two different tree inference methods (e.g. Maximum Likelihood or Bayesian).

1: the Bhattacharyya Coefficient ( $BC$ ) is the overlap of the distribution of tree similarity between two combination of missing data parameters. 2: the Bhattacharyya Coefficient ( $BC$ ) is the overlap of the distribution of tree similarity between two methods for the same combination of missing data parameters.

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 distance). However, the different missing data parameters and tree inference methods do not affect the topology in the same way (Fig. 3 and Fig. 4).

### *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 (Fig. 3). The Normalised Robinson-Foulds distance 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; Fig. 3; Supplementary Material Section 3: Tables S2-S4). However, The Normalised Triplets distance 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; Fig. 3; Supplementary Material Section 3: Tables S2-S4).

### *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.  $M_L = 0\%$ ,  $M_F = 0\%$ ,  $M_C = 0\%$ ; Fig. 4; Supplementary Material Section 3). Fig. 5 shows the similarity of distributions of tree metrics in a triangular matrix with the values of each pairwise

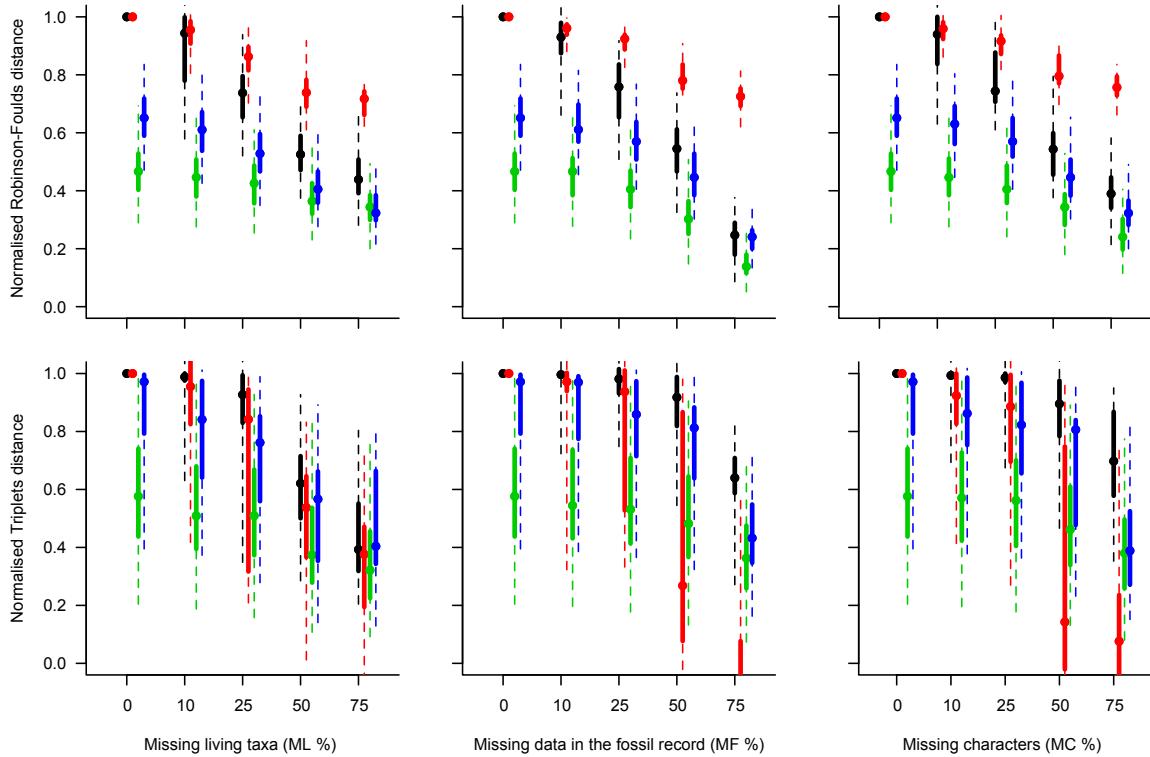


Figure 3: The effects of increasing missing data on topological recovery using Maximum Likelihood trees (black), Bayesian consensus trees (blue), Maximum Likelihood bootstrap trees (orange) and Bayesian posterior tree distributions (green). The percentage of missing data for each parameter ( $M_L$ ,  $M_F$  and  $M_C$ ) is shown on the x axis. Topological recovery was measured using two different tree comparison metrics: Normalised Robinson-Foulds distance (upper row) and Normalised Triplets distance (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.

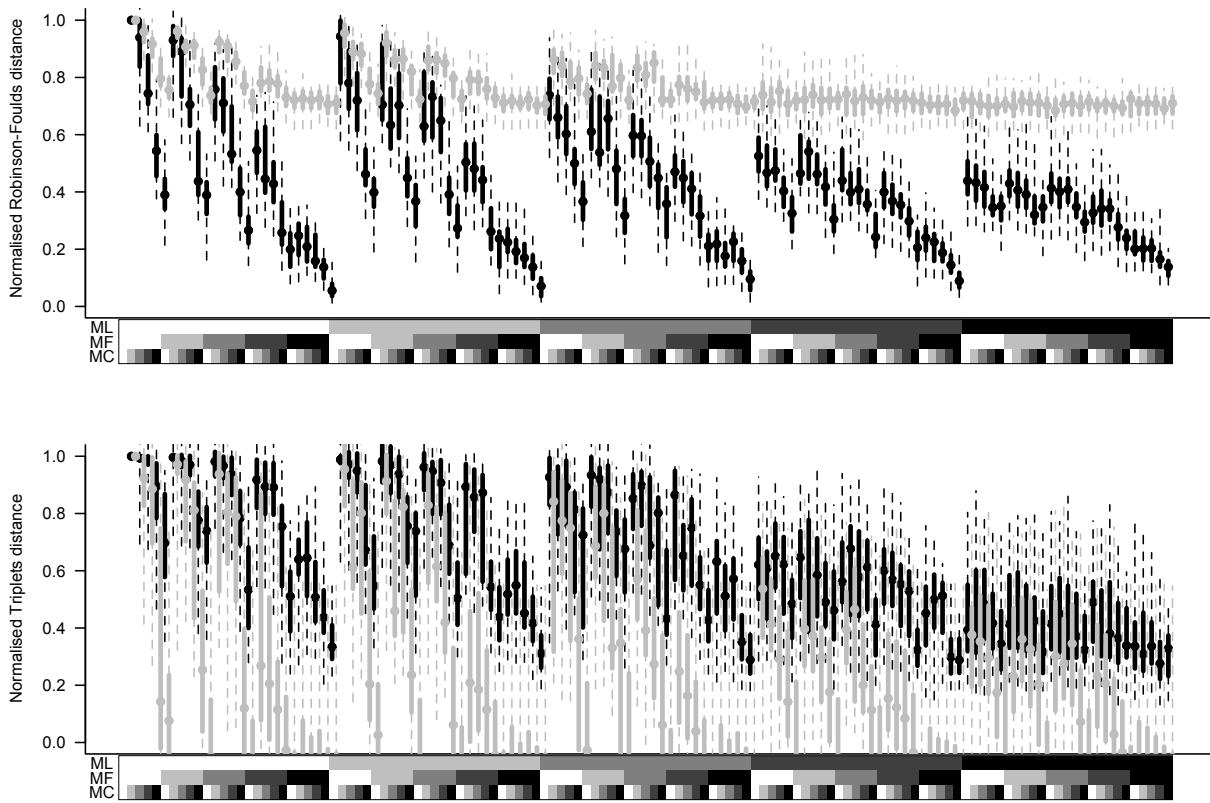


Figure 4: The effects of increasing missing data on topological recovery using Maximum Likelihood trees (black) and Bayesian consensus trees (grey). The x axis shows the percentage of missing data from 0% (white) to 75% (black) for the three parameters:  $M_L$  (upper line),  $M_F$  (middle line) and  $M_C$  (lower line). Topological recovery was measured using two different tree comparison metrics: Normalised Robinson-Foulds distance (upper row) and Normalised Triplets distance (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.

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; Fig. 5).

Using both Normalised Robinson-Foulds and Normalised Triplets distances 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 (Fig. 5). However, the Normalised Robinson-Foulds distance (median Bhattacharrya coefficient = 0.79; Fig. 5A) displays more dissimilarities than the Normalised Triplets distance (median Bhattacharrya coefficient = 0.81; Fig. 5B). Additionally, when using the Normalised Robinson-Foulds distance, 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 (Fig. 5A). Likewise, once  $M_C \geq 50\%$ , there is no additional affect of  $M_L$  and  $M_F$  (Fig. 5A).

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 distance or Normalised Triplets distance respectively; Table 1). However, these two methods perform worse than the Bayesian consensus trees using Normalised Robinson-Foulds distance (median Bhattacharrya Coefficient = 0 and 0.01, for the Maximum Likelihood bootstrap trees and the Bayesian posterior tree distribution respectively; Table 1; Fig. 3 and Supplementary Material Section 3).

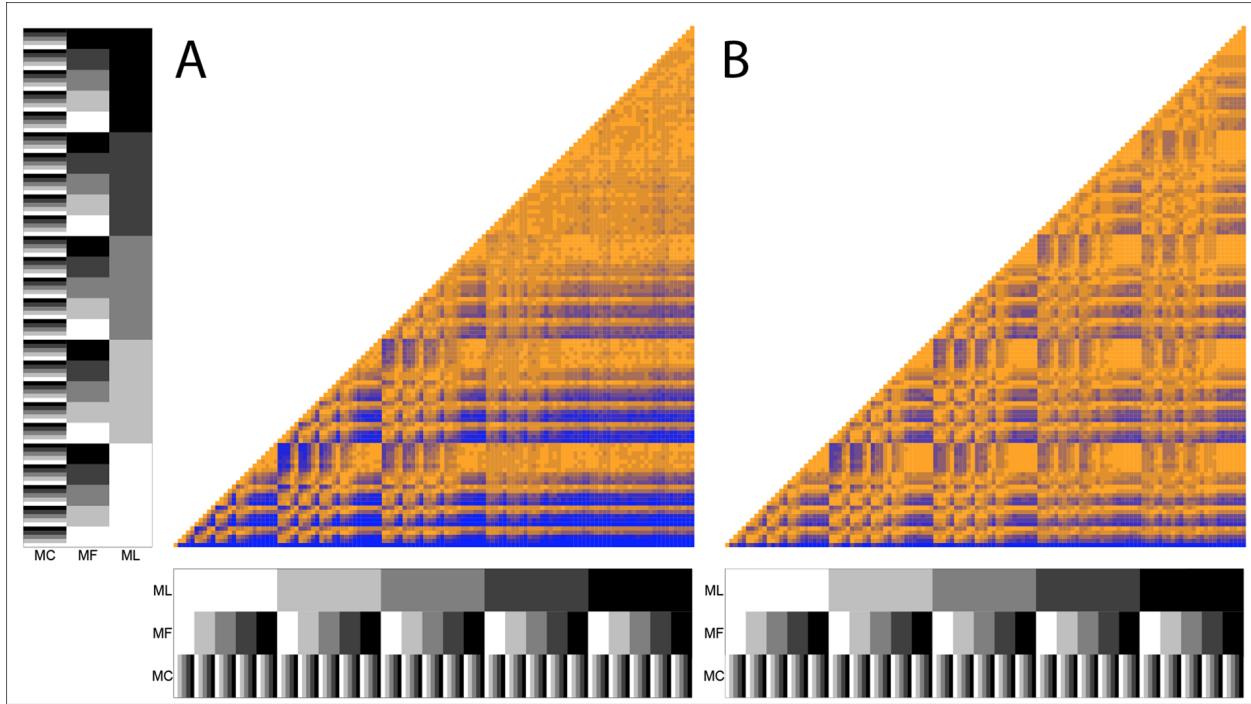


Figure 5: The effects of missing data on topological recovery using Bayesian consensus trees. Both axes show the percentage of missing data from 0% (white) to 75% (black) for the three parameters:  $M_L$  (upper line),  $M_F$  (middle line) and  $M_C$  (lower line). Topological recovery is represented by the probability of (A) Normalised Robinson-Foulds distance and (B) Normalised Triplets distance distributions overlapping with the "best" tree distribution, calculated using the Bhattacharyya Coefficient. The Bhattacharyya Coefficient values indicating low probability of overlap are shown in blue, those indicating a high probability of overlap are shown in orange.

Table 1: Bhattacharyya Coefficients of the pairwise method comparisons, each of which corresponds to the normalised distance between the “best” tree and the “missing data” using either the Normalised Robinson-Foulds distance (RF) or the Normalised Triplets distance (Tr). The values in bold are the extreme values of high or low probability of overlap between two methods.

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

## 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 (Fig. 5). Additionally, using Bayesian consensus trees recovered the "best" tree topology more consistently than using Maximum Likelihood trees or Bayesian posterior tree distributions (Fig. 4, Fig. 5, 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).

### *Individual effects of missing data parameters*

*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.

*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 distance) with increasing missing data in this parameter, clade conservation (i.e. Normalised Robinson-Foulds distance) 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 distance) but greater on the displacement of wildcard taxa (Normalised Triplets distance; Fig. 3 and Fig. 4). 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. This does not decrease the Normalised Robinson-Foulds distance, because clades are conserved, but decreases the Normalised Triplets distance because the fossils act as wildcard taxa. 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).

*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 (Fig. 3). 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-Fould distance and the Normalised Triplets distance in our simulations could be due to conflicting phylogenetic signal among morphological and molecular data in our matrices (Fig. 3 and Fig. 4). 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 Mk model (Lewis 2001) has been previously demonstrated to partially resolve this issue (Wright and Hillis 2014).

### *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 distance and the Normalised Triplets distance 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  $M_C = 75\%$ ; Fig. 4). 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  $M_C$  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 (Fig. 5). This has important implications for building Total Evidence phylogenies (see below).

### *Effects of tree inference methods*

Variation in our ability to recover the "best" tree topology depends heavily on the tree inference method (Fig. 3 and Fig. 4). For morphological data, previous studies have shown some superiority of probabilistic tree inference methods with simple evolutionary models such as the Mk 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 Mk 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 (Fig. 4). 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.

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 ( $M_L = 0\%$ ;  $M_F = 0\%$ ,  $M_C = 0\%$ ; Fig. 3). 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.

### *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 (Fig. 5). 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* (OLeary 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). 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 discussion 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).

## CONCLUSION

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 in a Total Evidence analysis, for example when using the Tip-Dating method. 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 processes through both time and space.

## SUPPLEMENTARY MATERIAL

Supplementary material (R code, analyses, supplementary methods and full results) can be found in the Dryad data repository at <http://dx.doi.org/10.5061/dryad.XXXX>.

## ACKNOWLEDGMENTS

Thanks to Gavin Thomas, Frédéric Delsuc, Emmanuel Douzery, Trevor Hodkinson, Andrew Jackson, and April Wright for useful comments on our simulation protocol. 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.
- Beck, R. M. and M. S. Lee. 2014. Ancient dates or accelerated rates? morphological clocks and the antiquity of placental mammals. *Proceedings. Biological sciences / The Royal Society* 281.
- Bhattacharyya, A. 1943. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.* 35:99–109.
- Bogdanowicz, D., K. Giaro, and B. Wrbel. 2012. Treecmp: Comparison of trees in polynomial time. *Evolutionary Bioinformatics* 8:475–487 10.4137/EBO.S9657.
- 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.
- Davalos, 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* .

- Dietl, G. and K. Flessa. 2011. Conservation paleobiology: putting the dead to work. *Trends in Ecology & Evolution* 26:30–37.
- Dobson, A. J. 1975. Comparing the shapes of trees 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.
- 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., J. Schnitzler, J. Eronen, C. Hof, B. Katrin, and C. Graham. 2013. Diversity in time and space: wanted dead and alive. *Trends in Ecology & Evolution* .
- 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., J. Huelsenbeck, and T. Stadler. 2013. The fossilized Birth-Death process: a coherent model of fossil calibration for divergence time estimation .

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 & 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.

Johnson, L. A. and D. E. Soltis. 1998. Assessing congruence: empirical examples from molecular data chap. 11, Pages 297–348. Springer US.

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

Kelly, S. B., 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.

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.
- 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: phylogenomics 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* .
- Pearman, P., A. Guisan, O. Broennimann, and C. Randin. 2008. Niche dynamics in space and time. *Trends in Ecology & Evolution* 23:149–158.
- 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 & 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. 1993. Extinction: bad genes or bad luck? Oxford University Press.
- 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.
- 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.
- Sansom, R. and M. Wills. 2013. Fossilization causes organisms to appear erroneously primitive by distorting evolutionary trees. *Scientific Reports* 3.
- 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. G. 1945. Tempo and mode in evolution. *Transactions of the New York Academy of Sciences* 8:45–60.

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.

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.

Spencer, M. R. and E. W. Wilberg. 2013. Efficacy or convenience? model-based approaches to phylogeny estimation using morphological data. *Cladistics* 29.

Springer, M., R. Meredith, J. Gatesy, C. Emerling, J. Park, D. Rabosky, T. Stadler, C. Steiner, O. Ryder, J. Janeka, C. Fisher, and W. Murphy. 2012. Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix. *PLOS ONE* 7.

Stadler, T. and Z. Yang. 2013. Dating phylogenies with sequentially sampled tips. *Systematic Biology* .

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

Tavaré, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences vol. 17 of *Some Mathematical Questions in Biology*. American Mathematical Society.

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

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

Wiens, J. J. 2003. Missing data, incomplete taxa, and phylogenetic accuracy. *Systematic Biology* 52.

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.

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

Zander, R. 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.

## SUPPLEMENTARY MATERIAL

### *Tree Building*

## SUPPLEMENTARY MATERIAL SECTION 1

### *Morphological characters states*

In order to obtain a realistic probabilistic value for of  $k$  characters states for each simulated morphological character, we downloaded 100 random morphological characters (with more than 100 characters each) from TreeBASE database (<http://treebase.org/>) published between 1985 and 2013 and covering 19 taxonomic classes (Chordata, Arthropoda, Annelida, Angiosperm, Gymnosperm and Pteridophyta). We selected a total of 22563 characters ranging from 2 to 10 states. We calculated the proportion of characters with 2, 3, 4, 5, 6, 7, 8, 9 or 10 states. We then sampled 22563  $k$  values between 2 and 10 with the same proportion of characters from the empirical data. We then used a simple t-test to check if our simulation was equal to the empirical data. In this study, we only simulated characters with 2 or 3 states because of the high proportion of ordered characters encountered on characters with more than 3 states and the difficulties of simulate biologically sensible ordered characters.

### *Tree Building Software settings*

*Maximum Likelihood - RAxML v8.0.20 (Stamatakis 2014).*—

Model:

Molecular data:

GTR +  $\Gamma_4$  (-m GTRGAMMA)

Morphological data:

Mk +  $\Gamma_4$  (-K MK)

Support:

Rapid Bootstrap algorithm (LSR), 1000 replicates

*Bayesian - MrBayes v3.2.1 (Ronquist et al. 2012b).*—

Priors:

Molecular data:

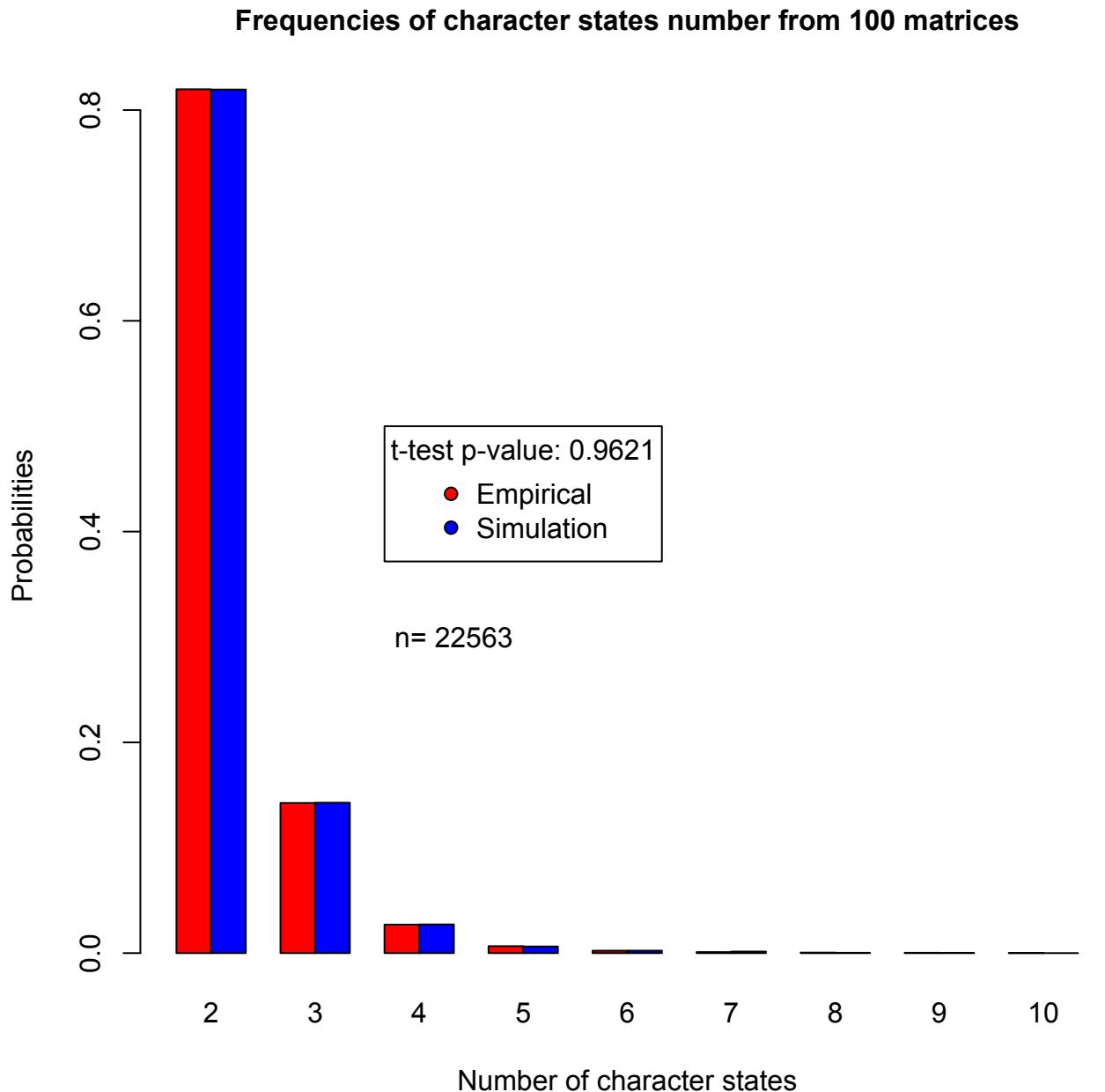


Figure 6: Character states distribution in empirical matrices. Characters states number distribution extracted from 100 random morphological matrices downloaded from ReeBase.

rates distribution shape ( $\alpha$ ) = 0.5

Transition/Transversion ratio = 2 ( $\beta(80,40)$ )

Starting tree: "True" tree topology with each branch length = 1

Morphological data:

rates distribution shape ( $\alpha$ ) = 0.5

Models:

Molecular data: HKY +  $\Gamma_4$

Morphological data: Mk +  $\Gamma_4$

MCMC:

2 runs

4 chains per run

generations ;  $50 \times 1^6$

sample frequency =  $1050 \times 1^3$

ASDS diagnosis frequency =  $50 \times 1^3$

ASDS < 0.01

ESS >> 200

Burnin = 25%

### *Tree Comparisons*

## SUPPLEMENTARY MATERIAL SECTION 2

### *Robinson-Foulds distance details*

Robinson-Foulds distance (RF; Robinson and Foulds 1981), or "path difference", measures the 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) and can be expressed as following:

$$RF_{x,y} = N_x + N_y - 2C_{x,y} \quad (1)$$

Where  $C_{x,y}$  is the number of clades in common in the two trees.  $C$  is equal to one if the two trees have the same  $n$  taxa; and  $C = n - 2$  when none of the  $n$  taxa are shared between the trees. This metric is more sensitive to taxon displacement than Triples distance (i.e. if one taxon moves out of a clade, then the clades are no longer considered similar; Critchlow et al. (1996); Johnson and Soltis (1998); Wiens (2003)). The minimal value of  $C$  is equal to 1 if the two trees have the same  $n$  taxa; the maximal value in  $C = n - 2$ . For a fully unresolved tree (star tree)  $N=1$  and for a fully resolved tree (binary tree)  $N = n - 2$ . The minimal and maximal topological distance for taxa is:

$$RF_{min} = 1 + 1 - 2C_{x,y} \quad (2)$$

And:

$$RF_{max} = 2(n - 2) - 2 \quad (3)$$

One can then rescale  $RF.scaled$  by using the maximal and minimal value for any  $n$  taxa:

$$RF.scaled_{x,y} = \frac{RF_{x,y} - RF_{min}}{RF_{max}} \quad (4)$$

This metric is more sensitive to taxa displacement than the Triplet distance (Critchlow et al. 1996; Johnson and Soltis 1998; Wiens 2003) and therefore a low value will show a good clade conservation between two trees and a high value will show a bad recovery of common clades.

### *Triplets distance details ( $T_{x,y}$ )*

Triples distance ( $T_{x,y}$ ; Dobson 1975) measures the number of sub-trees made up of three taxa (triplets) that differ between two given trees. Each triplet can be written as

$I_{ijk} = (ijk)$ . Where  $I_{ijk}$  is equal to zero if the the two triplets  $(ijk)$  are the same in the two trees otherwise  $I_{ijk}$  is equal to one. For any rooted binary tree there are only three possible combinations for each triplet:  $((j,k),i)$ ,  $((i,k),j)$ ; and  $((i,j),k)$ ; (Johnson and Soltis 1998). If the trees used are not fully binary, a fourth triplet combination is possible:  $(i,j,k)$ . We can calculate the triplet distance between two trees,  $S_n$ , as:

$$S_n = \sum_{ijk} I_{ijk} \quad (5)$$

Where:

$$\sum_{ijk} = \binom{n}{4} = \frac{n!}{4!(n-4)!} \quad (6)$$

And where  $n$  is the total number of taxa in both trees (modified from Critchlow et al. (1996)). If  $S_n = 0$ , the trees are identical; when  $S_n = \binom{n}{4}$ , the trees are as different as possible (i.e. every taxon has a different placement in the two trees). Because the possible number of triplets per clade is a finite number, the probability of two random trees with the same  $n$  taxa to have the same triplet is:

$$P(I_{ijk} = 0) = \frac{1}{4} \quad (7)$$

Therefore one can calculate the probability of two random trees having the same triplets:

$$P(S_n = 0) = \sum_{ijk} P_{I_{ijk}=0} \quad (8)$$

$$P(S_n = 0) = \frac{n!}{4(3!(n-3))!} \quad (9)$$

And in the same way:

$$P(S_n = 1) = \frac{3n!}{4(3!(n-3))!} \quad (10)$$

## *Normalised Tree Similarity*

For any tree with  $n$  taxa compared using a tree distance metric  $m$ , Normalized Tree Similarity,  $NTS_m$  (Bogdanowicz et al. 2012), represents the similarity score for the two trees given the expected distance between two random Yule trees with  $n$  taxa. If  $\bar{d}_{m,n}(rand)$  is the average distance between two random Yule trees with  $n$  taxa and  $d_{m,n}(x,y)$  the distance between the two trees  $x$  and  $y$  containing each  $n$  taxa, then:

$$NTS_{m,n}(x,y) = \frac{\bar{d}_{m,n}(rand) - d_{m,n}(x,y)}{\bar{d}_{m,n}(rand)} \quad (11)$$

$NTS$  ranges from one to  $-\infty$ . For any  $m, n$ , when  $NTS = 1$ , the trees are identical, when  $NTS = 0$  the trees are no more different than expected by chance, and when  $NTS < 0$ , the trees are more different than expected when comparing two random trees.

## *Tree comparisons*

*Random tree comparison scaling.*— We used the comparison of 1000 random trees to obtain the mean comparison value  $\bar{d}_{m,n}(rand)$  for the NTS metric. We randomly generated two sets of 1000 trees of  $n$  taxa using the rmtree function of ape package (v3.0-11 Paradis et al. (2004)) that generates a given number of random Yule trees. We calculated the  $\bar{d}_{m,n}(rand)$  value using an approach similar to the RPCBTC (described below) by performing 1000 random pairwise comparisons using the TreeCmp java script (Bogdanowicz et al. 2012).

## *Bhattacharyya Coefficient*

The Bhattacharyya Coefficient calculates the probability of overlap of two distributions (Bhattacharyya 1943). When it is equal to 0, the probability of overlap of the

distributions is null and when it is equal to 1, the two distributions are entirely overlapping. It is an elegant and easy to compute continuous measurement of probability of similarity between two distributions. The coefficient is calculated as the sum of the square root of the relative counts shared in a  $n$  number of bins among two distributions.

$$\text{Bhattacharyya Coefficient} = \sum_{i=1}^n \sqrt{\sum a_i \times \sum b_i} \quad (12)$$

Where, for  $a$  and  $b$ :

$$\sum a_i = \frac{\text{Number of counts in bin } i \text{ for the distribution } a}{\text{Total number of counts for the distribution } a} \quad (13)$$

The precision of the Bhattacharyya Coefficient is directly related to the number of  $n$  bins. If  $n$  is low, the overlap will be overestimated and if  $n$  is too high, the overlap will be underestimated. In this analysis, we determined the number of bins Silverman's rule of thumb which determines the bins to be equal to 0.9 times the minimum of the standard deviation and the interquartile range divided by 1.34 times the sample size to the negative one-fifth power (Silverman (1986, page 48, eqn (3.31))).

### *Codes*

All codes are available at: [https://github.com/TGuillerme/Total\\_Evidence\\_Method-Missing\\_data/tree/master/Functions](https://github.com/TGuillerme/Total_Evidence_Method-Missing_data/tree/master/Functions)

The tree comparison results analysis can be repeated for more details at:  
[https://github.com/TGuillerme/Total\\_Evidence\\_Method-Missing\\_data/tree/master/Analysis](https://github.com/TGuillerme/Total_Evidence_Method-Missing_data/tree/master/Analysis)

### *Additional Results*

## SUPPLEMENTARY MATERIAL SECTION 3

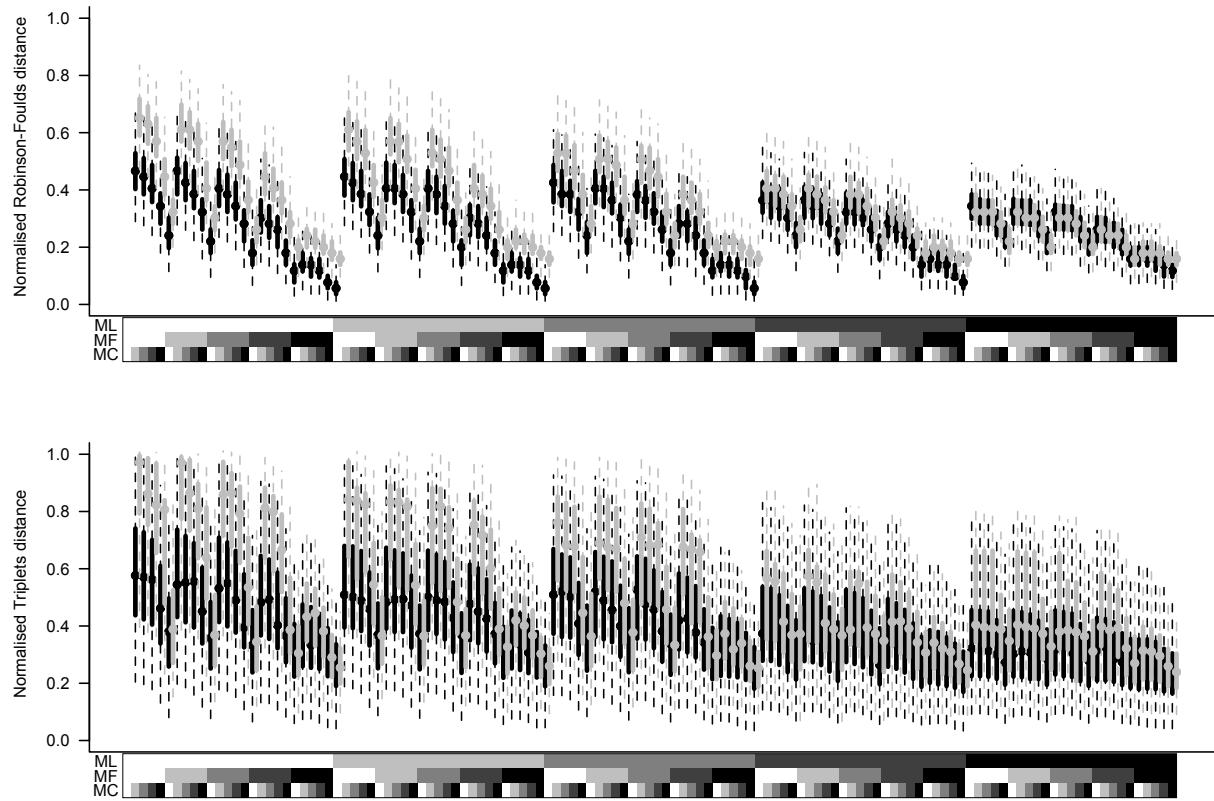


Figure 7: Effect of increasing missing data on topological recovery using Maximum Likelihood Bootstrap trees (black) and Bayesian posterior treee distribution (grey). The percentage of missing data (gradient colour from 0% (white), 10%, 25%, 50% and 75% (black)) for each parameter ( $M_L$  (upper line),  $M_F$  (middle line) and  $M_C$ (lower line)) is shown on the x axis. Topological recovery was measured using two different tree comparison metrics: Normalised Robinson-Foulds distance (upper row) and Normalised Triplets distance (lower row). The graph shows the modal value (points), 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.

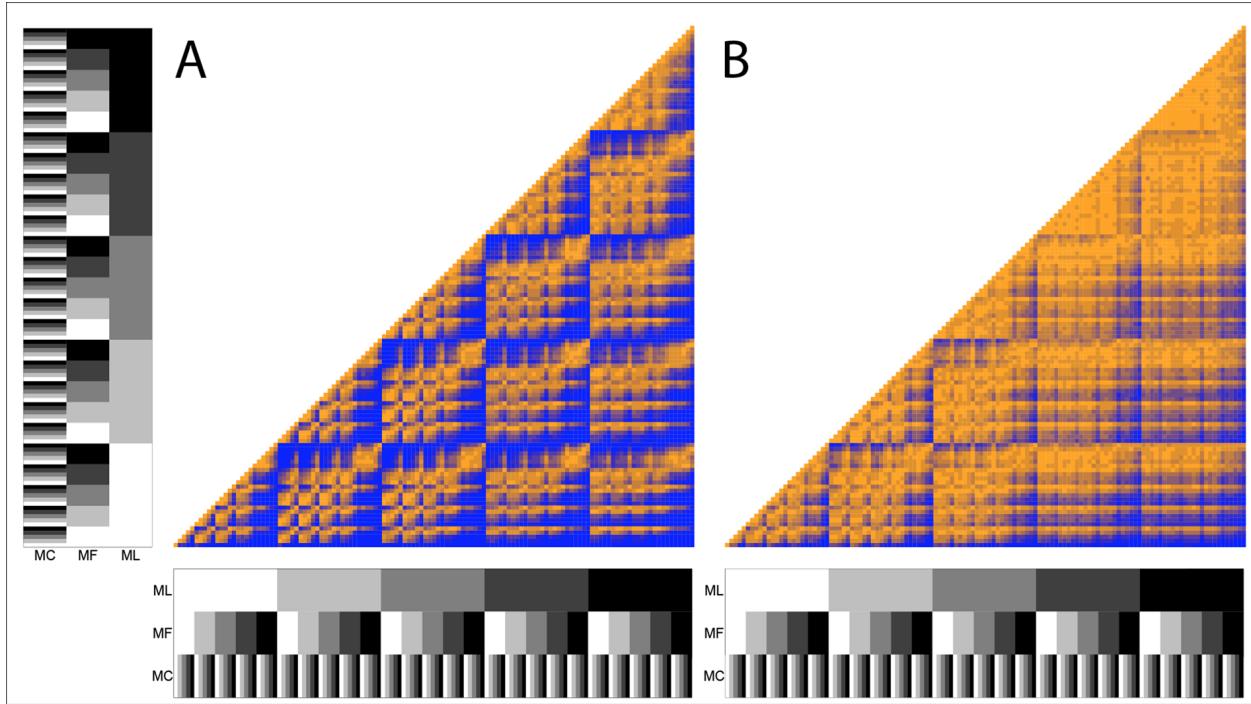


Figure 8: Effect of missing data on topological recovery using the Maximum Likelihood trees. The percentage of missing data (gradient colour from 0% (white), 10%, 25%, 50% and 75% (black)) for each parameter ( $M_L$  (upper line),  $M_F$  (middle line) and  $M_C$  (lower line)) is shown on both x and y axis. Topological recovery is represented by the probability of the two different tree comparison metrics (Normalised Robinson-Foulds distance (A - right) and Normalised Triplets distance (B - left)) distribution overlap with the "best" tree using the Bhattacharyya Coefficients (low probability of overlap - blue ; high probability of overlap - orange).

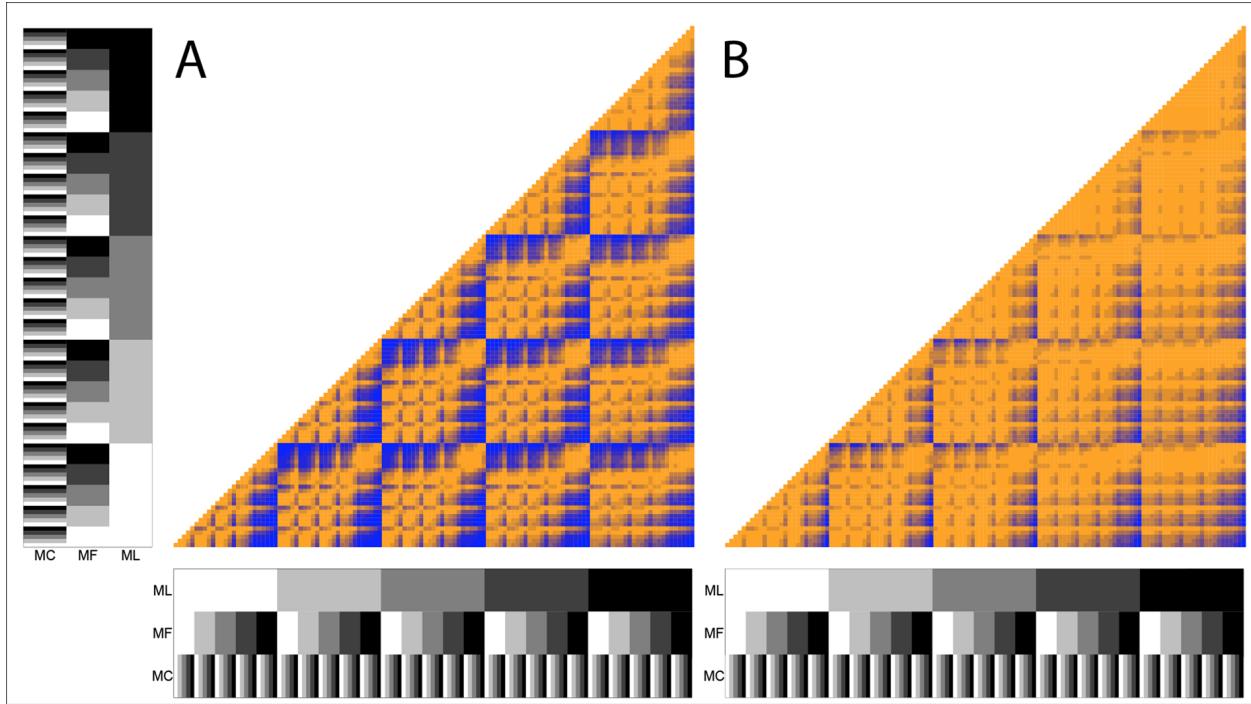


Figure 9: Effect of missing data on topological recovery using the Maximum Likelihood bootstrap trees. The percentage of missing data (gradient colour from 0% (white), 10%, 25%, 50% and 75% (black)) for each parameter ( $M_L$  (upper line),  $M_F$  (middle line) and  $M_C$  (lower line)) is shown on both x and y axis. Topological recovery is represented by the probability of the two different tree comparison metrics (Normalised Robinson-Foulds distance (A - right) and Normalised Triplets distance (B - left)) distribution overlap with the "best" tree using the Bhattacharyya Coefficients (low probability of overlap - blue ; high probability of overlap - orange).

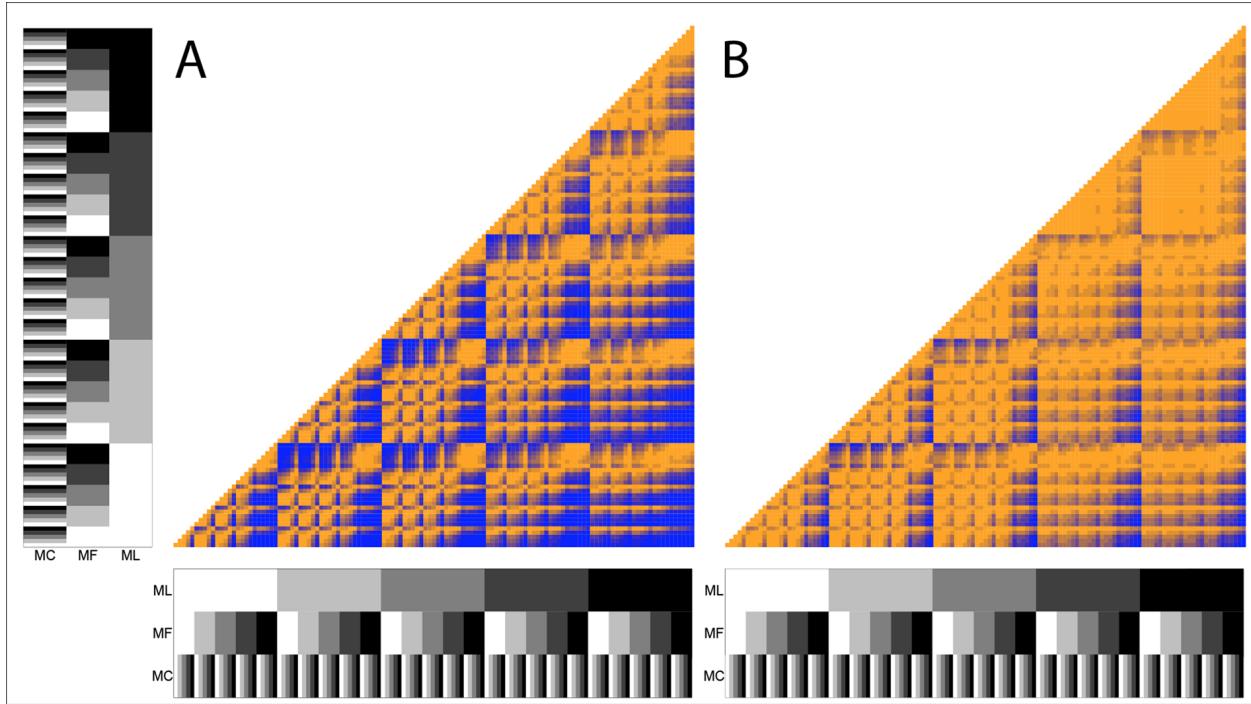


Figure 10: Effect of missing data on topological recovery using the Bayesian posterior tree distribution. The percentage of missing data (gradient colour from 0% (white), 10%, 25%, 50% and 75% (black)) for each parameter ( $M_L$  (upper line),  $M_F$  (middle line) and  $M_C$  (lower line)) is shown on both x and y axis. Topological recovery is represented by the probability of the two different tree comparison metrics (Normalised Robinson-Foulds distance (A - right) and Normalised Triplets distance (B - left)) distribution overlap with the "best" tree using the the Bhattacharyya Coefficients (low probability of overlap - blue ; high probability of overlap - orange).

Table 2: Summary of the of the comparisons between the "best" tree and the "missing data" trees for each different tree inference method using either the Normalised Robinson-Foulds distance (RF) or the Normalised Triplets distance (Tr).

Tree inference method	Metric	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Maximum Likelihood	<i>RF</i>	0.06	0.26	0.40	0.41	0.50	0.95
	<i>Tr</i>	0.29	0.45	0.59	0.63	0.84	1.00
Bayesian consensus	<i>RF</i>	0.69	0.71	0.72	0.76	0.79	0.96
	<i>Tr</i>	-0.28	-0.11	0.17	0.19	0.37	0.98
Maximum Likelihood bootstraps	<i>RF</i>	0.06	0.18	0.27	0.26	0.34	0.46
	<i>Tr</i>	0.23	0.31	0.35	0.38	0.45	0.58
Bayesian posterior tree distributions	<i>RF</i>	0.16	0.22	0.32	0.34	0.42	0.65
	<i>Tr</i>	0.24	0.35	0.40	0.50	0.67	0.98

<sup>1</sup>*RF*=Normalised Robison-Fould distance ; <sup>2</sup>*Tr*=Normalised Triplets distance

Table 3: Summary of the comparisons between the "best" tree and the "missing data" trees for each different tree inference method using either the Normalised Robinson-Foulds distance (RF) or the Normalised Triplets distance (Tr) for the  $M_L$  missing data parameter only.

Tree inference method	Metric	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Maximum Likelihood	<i>RF</i>	0.44	0.51	0.63	0.66	0.78	0.95
	<i>Tr</i>	0.45	0.56	0.76	0.74	0.93	0.99
Bayesian consensus	<i>RF</i>	0.71	0.73	0.80	0.82	0.88	0.95
	<i>Tr</i>	0.37	0.46	0.67	0.67	0.87	0.96
Maximum Likelihood bootstraps	<i>RF</i>	0.34	0.37	0.42	0.41	0.44	0.46
	<i>Tr</i>	0.32	0.40	0.51	0.46	0.51	0.57
Bayesian posterior tree distributions	<i>RF</i>	0.33	0.41	0.52	0.50	0.60	0.65
	<i>Tr</i>	0.41	0.56	0.76	0.71	0.84	0.98

<sup>1</sup>*RF*=Normalised Robison-Fould distance ; <sup>2</sup>*Tr*=Normalised Triplets distance

Table 4: Summary of the comparisons between the "best" tree and the "missing data" trees for each different tree inference method using either the Normalised Robinson-Foulds distance (RF) or the Normalised Triplets distance (Tr) for the  $M_F$  missing data parameter only.

Tree inference method	Metric	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Maximum Likelihood	<i>RF</i>	0.23	0.46	0.64	0.61	0.79	0.93
	<i>Tr</i>	0.65	0.84	0.95	0.89	0.99	1.00
Bayesian consensus	<i>RF</i>	0.72	0.77	0.86	0.85	0.94	0.96
	<i>Tr</i>	-0.16	0.19	0.63	0.52	0.96	0.98
Maximum Likelihood bootstraps	<i>RF</i>	0.14	0.30	0.40	0.35	0.45	0.46
	<i>Tr</i>	0.37	0.49	0.54	0.51	0.56	0.57
Bayesian posterior tree distributions	<i>RF</i>	0.24	0.45	0.57	0.51	0.63	0.65
	<i>Tr</i>	0.44	0.81	0.86	0.82	0.98	0.98

<sup>1</sup>*RF*=Normalised Robison-Fould distance ; <sup>2</sup>*Tr*=Normalised Triplets distance

Table 5: Summary of the of the comparisons between the "best" tree and the "missing data" trees for each different tree inference method using either the Normalised Robinson-Foulds distance (RF) or the Normalised Triplets distance (Tr) for the  $M_C$  missing data parameter only.

Tree inference method	Metric	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Maximum Likelihood	<i>RF</i>	0.40	0.50	0.64	0.65	0.79	0.94
	<i>Tr</i>	0.70	0.84	0.93	0.89	0.99	1.00
Bayesian consensus	<i>RF</i>	0.76	0.79	0.86	0.86	0.92	0.96
	<i>Tr</i>	0.05	0.16	0.53	0.50	0.87	0.92
Maximum Likelihood bootstraps	<i>RF</i>	0.25	0.34	0.42	0.38	0.45	0.46
	<i>Tr</i>	0.38	0.47	0.55	0.51	0.57	0.58
Bayesian posterior tree distributions	<i>RF</i>	0.32	0.44	0.58	0.52	0.62	0.65
	<i>Tr</i>	0.39	0.78	0.82	0.79	0.98	0.98

Table 6: Bhattacharyya Coefficients of the pairwise method comparisons, each of which corresponds to the normalised distance between the "best" tree and the "missing data" using either the Normalised Robinson-Foulds distance (RF) or the Normalised Triplets distance (Tr).

Comparison	Metric	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Maximum Likelihood vs. Bayesian consensus	<i>RF</i>	0.30	0.31	0.69	0.61	0.77	1.00
	<i>Tr</i>	0.79	0.81	0.84	0.86	0.85	1.00
Maximum Likelihood vs. Maximum Likelihood bootstraps	<i>RF</i>	0.03	0.22	0.29	0.36	0.54	0.69
	<i>Tr</i>	0.08	0.42	0.53	0.51	0.74	0.78
Maximum Likelihood vs. Bayesian posterior trees	<i>RF</i>	0.02	0.49	0.61	0.51	0.67	0.74
	<i>Tr</i>	0.21	0.61	0.70	0.63	0.81	0.81
Bayesian consensus vs. Maximum Likelihood bootstraps	<i>RF</i>	0.01	0.02	0.02	0.02	0.03	0.04
	<i>Tr</i>	0.08	0.69	0.78	0.64	0.79	0.84
Bayesian consensus vs. Bayesian posterior trees	<i>RF</i>	0.01	0.02	0.02	0.04	0.08	0.09
	<i>Tr</i>	0.21	0.74	0.75	0.68	0.84	0.87
Bayesian posterior tree vs. Maximum Likelihood bootstraps	<i>RF</i>	0.69	0.75	0.85	0.85	0.95	1.00
	<i>Tr</i>	0.91	0.92	0.96	0.95	0.97	0.98

Table 7: Bhattacharyya Coefficients of the pairwise method comparisons, each of which corresponds to the normalised distance between the "best" tree and the "missing data" using either the Normalised Robinson-Foulds distance (RF) or the Normalised Triplets distance (Tr).

Comparison	Metric	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Maximum Likelihood vs. Bayesian consensus	<i>RF</i>	0.00	0.25	0.48	0.50	0.76	1.00
	<i>Tr</i>	0.38	0.69	0.75	0.72	0.80	1.00
Maximum Likelihood vs. Maximum Likelihood bootstraps	<i>RF</i>	0.03	0.18	0.32	0.36	0.47	0.77
	<i>Tr</i>	0.08	0.34	0.40	0.38	0.53	0.55
Maximum Likelihood vs. Bayesian posterior trees	<i>RF</i>	0.02	0.47	0.71	0.60	0.86	0.94
	<i>Tr</i>	0.21	0.54	0.62	0.56	0.64	0.80
Bayesian consensus vs. Maximum Likelihood bootstraps	<i>RF</i>	0.00	0.00	0.01	0.01	0.01	0.03
	<i>Tr</i>	0.08	0.38	0.54	0.49	0.70	0.75
Bayesian consensus vs. Bayesian posterior trees	<i>RF</i>	0.00	0.02	0.02	0.02	0.04	0.04
	<i>Tr</i>	0.21	0.29	0.66	0.54	0.72	0.82
Bayesian posterior tree vs. Maximum Likelihood bootstraps	<i>RF</i>	0.69	0.69	0.72	0.71	0.72	0.72
	<i>Tr</i>	0.91	0.91	0.91	0.93	0.92	0.98

Table 8: Bhattacharyya Coefficients of the pairwise method comparisons, each of which corresponds to the normalised distance between the "best" tree and the "missing data" using either the Normalised Robinson-Foulds distance (RF) or the Normalised Triplets distance (Tr).

Comparison	Metric	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Maximum Likelihood vs. Bayesian consensus	<i>RF</i>	0.03	0.32	0.66	0.55	0.75	1.00
	<i>Tr</i>	0.51	0.69	0.80	0.76	0.80	1.00
Maximum Likelihood vs. Maximum Likelihood bootstraps	<i>RF</i>	0.03	0.17	0.21	0.31	0.46	0.68
	<i>Tr</i>	0.08	0.31	0.39	0.39	0.56	0.61
Maximum Likelihood vs. Bayesian posterior trees	<i>RF</i>	0.02	0.44	0.47	0.52	0.78	0.90
	<i>Tr</i>	0.21	0.52	0.59	0.55	0.66	0.77
Bayesian consensus vs. Maximum Likelihood bootstraps	<i>RF</i>	0.00	0.01	0.01	0.02	0.02	0.03
	<i>Tr</i>	0.08	0.47	0.62	0.51	0.66	0.73
Bayesian consensus vs. Bayesian posterior trees	<i>RF</i>	0.00	0.02	0.04	0.04	0.05	0.06
	<i>Tr</i>	0.21	0.45	0.64	0.57	0.74	0.79
Bayesian posterior tree vs. Maximum Likelihood bootstraps	<i>RF</i>	0.69	0.73	0.73	0.76	0.81	0.86
	<i>Tr</i>	0.91	0.92	0.93	0.94	0.96	0.99