Molecular Phylogenetics and Evolution 94 (2016) 146–158

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

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

Thomas Guillerme [a,b,*], Natalie Cooper [a,b,c]

[a] School of Natural Sciences, Trinity College Dublin, Dublin 2, Ireland
[b] Trinity Centre for Biodiversity Research, Trinity College Dublin, Dublin 2, Ireland
[c] Department of Life Sciences, Natural History Museum, Cromwell Road, London SW7 5BD, UK

ARTICLE INFO

Article history:
Received 8 June 2015
Revised 24 August 2015
Accepted 26 August 2015
Available online 31 August 2015

Keywords:
Morphological characters
Bayesian
Maximum Likelihood
Topology
Fossil
Living

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.

1. Introduction

Although most species that have ever lived are now extinct (Novacek and Wheeler, 1992; Raup, 1981), many large-scale 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). To do this, however, 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 how they treat fossil taxa and their data. One can use fossils as tips or as nodes in the phylogeny, and can use only the age of the fossils, only the morphology of the fossils, or age and morphology jointly. 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

☆ This paper was edited by the Associate Editor Liliana M. Davalos.

* Corresponding author at: Zoology Building, Trinity College Dublin, Dublin 2, Ireland. Fax: +353 1 6778094.

E-mail address: guillert@tcd.ie (T. Guillerme).

http://dx.doi.org/10.1016/j.ympev.2015.08.023
1055-7903/© 2015 The Authors. Published by Elsevier Inc.

as maximum parsimony (Hennig, 1966; Felsenstein, 2004). This approach is commonly used by palaeontologists 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, only fossil 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). Neither approach, however, 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 taxa 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). The Total Evidence method is becoming an increasingly popular way of adding fossil taxa to phylogenies (e.g. Pyron, 2011; Ronquist et al., 2012a; Schrago et al., 2013; Slater, 2013; Beck and Lee, 2014; Arcila et al., 2015). 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 are (nearly) always unavailable. Therefore Total Evidence matrices are likely to contain a large proportion 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 do not appear be a major problem in molecular and morphological matrices separately (as long as enough data overlap in each case, and missing data are not phylogenetically biased; Wiens, 2003; Wiens et al., 2005; Wiens, 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, few attempts have been made to study the impact of this missing data issue on phylogenetic inference in a Total Evidence framework (i.e. using both molecular and morphological data; Wiens et al., 2005; Manos et al., 2007; Pattinson et al., 2014). These previous studies assessed the effect of missing data on topology by either (1) comparing a dataset with missing data to subsets without missing data (Wiens et al., 2005); or (2) removing both molecular and some morphological data from living taxa to create artificial fossils (Manos et al., 2007; Pattinson et al., 2014). Both approaches have shown that missing data are not a major problem and should not be an obstacle to combining both living and fossil species in the same phylogenies. The way these studies were conducted, however, means that their conclusions are not generally applicable across all scenarios involving missing data in Total Evidence phylogenies. For example, using an empirical (rather than simulation based) approach limits their conclusions to studies with similar distributions of data across species in the phylogeny. Additionally, one of the three previous studies did not include fossil taxa in their analyses, so their results cannot be used to make conclusions about how missing data may influence the placement of fossils (Wiens, 2003). The other two studies did include fossil taxa, but used the patchiness of the fossil record to determine how to remove data from their matrices (Manos et al., 2007; Pattinson et al., 2014). Data for living species are unlikely to be missing in this patchy way, instead full molecular data with the complete absence of morphological data is a likely pattern (Guillerme and Cooper, 2015). Finally, these previous studies mainly focused on how missing data in fossil taxa affect the placement of fossils, ignoring the effects of missing data in living species (Manos et al., 2007; Pattinson et al., 2014).

In this study, we propose a theoretical assessment of the effect of missing data in the Total Evidence method by removing living taxa with morphological data, fossil data, all data for certain characters and the combination of these three aspects. This is an advance on previous studies because we use large-scale simulations and analyse the effects of three distinct aspects of missing data thus focusing on both neontological and palaeontological parts of the matrix. In addition, we test the effect of missing data by measuring two crucial aspects of topology in both Maximum Likelihood and Bayesian phylogenies: (i) the conservation of clades (based on the Robinson–Foulds distance; Robinson and Foulds, 1981) and (ii) the displacement of wild-card taxa (based on the Triplets distance; Critchlow et al., 1996) rather than just a single measure of clade conservation or clade support (cf. Wiens et al., 2005; Pattinson et al., 2014).

We focus on the effects of missing data on our ability to recover 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 using topological constraints (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 amount of 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 resulting tree topology. We infer the topology from the matrices using both Maximum Likelihood and Bayesian inference methods and measure the differences in topology using two different topological distance metrics as proxies for clade conservation and for wild-card taxa placement. We find that minimising 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 minimising 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 partition of Total Evidence matrices influences tree topology, we used the following protocol (Fig. 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 number of morphological characters (N_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 with 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.

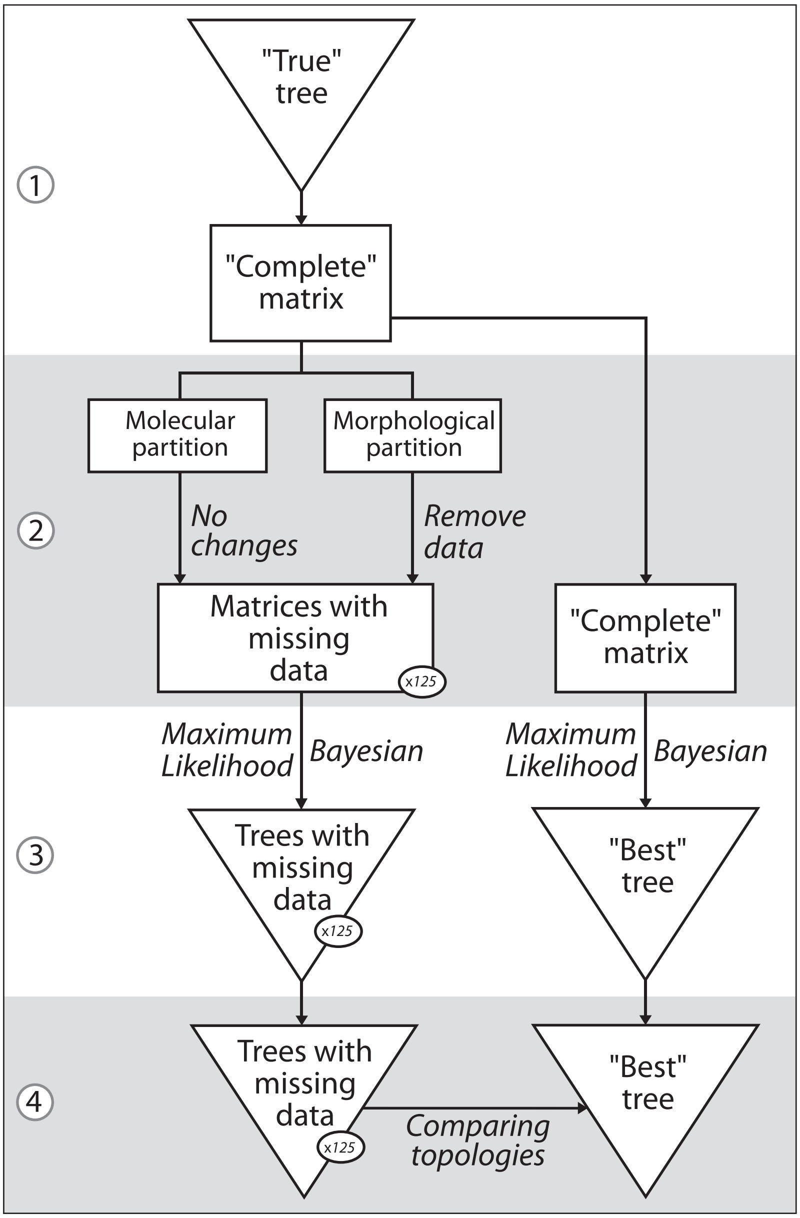

Fig. 1. Protocol outline. (1) We randomly generated a birth–death tree (the "true" tree) and used it to simulate a matrix with no missing data (the "complete" matrix). (2) We removed data from the morphological part of the "complete" matrix resulting in 125 "missing-data" matrices. (3) We built phylogenetic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

4. Comparing topologies:
   We compared the "best" tree to the "missing-data" trees to assess the influence of each parameter (M_L, M_F, N_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 simulated 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 selected from a gamma distribution with an (α) shape of 0.5 (Yang, 1996). In practice, a value of $\alpha < 1$ decreases the number of sites with high substitution rates, thus reducing homoplasic sites and increasing the phylogenetic signal (Hassanin et al., 1998; Estoup et al., 2002). Also, we chose this α value to be consistent with our protocol for simulating morphological characters (see below). This model and these parameter settings strike a balance between realism for empirical datasets (e.g. Douady et al., 2003; Kelly et al., 2014) and parameter richness with more complex models (e.g., GTR, multiple partitions with independent models), making them more suitable for our computational limitations (even with the parameters defined, the 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 simulated the morphological matrix using the rTraitDisc function from 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 (based on an empirical review of published 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 be consistent with the molecular rate parameters, to avoid homoplasy in the morphological part of the matrix and create a clear phylogenetic signal (Wright and Hillis, 2014). Topological error has been shown to be minimal at a morphological rate of 0.5 when using the M*kv* model (Lewis, 2001; Wright and Hillis, 2014). Note, however, that Wright and Hillis (2014) have shown that low morphological rates (<0.5) increase variance in topological error, but we discarded simulations with such topological error by selecting only matrices with a "fair" phylogenetic signal (see Estimating phylogenies section below; Zander, 2004) so this should not influence our results.

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

To explore the effect of missing morphological data on topological recovery, we removed various amounts of the "complete" matrix to obtain matrices with missing morphological data. Hereafter, we call these matrices with missing morphological data the "missing-data" matrices. Note that the amount of molecular data remained constant throughout our simulations: 1000 molecular characters for living taxa and no molecular data for fossil taxa (see above). We removed morphological data using three data incompleteness parameters:

1. The proportion of missing living taxa (M_L). This first missing-data parameter corresponds to the proportion of living taxa with no morphological data. It represents the number of living taxa that are present in the matrix but have only molecular data available. This reflects the fact that, because of the increasing ease of collecting molecular data, morphological data for living species are rarely collected (Guillerme and Cooper, 2015). Therefore, many living species will have only molecular data available. In practice, we removed all the morphological data from randomly chosen living taxa with five different proportions: 0%, 10%, 25%, 50% or 75% of living taxa with no morphological data.
2. The proportion of missing data in the fossil record (M_F). This missing data parameter represents the completeness of the fossil record. Due to preservation biases, missing data for fossil taxa are common (Sansom and Wills, 2013). In practice, we randomly removed a proportion of data from across the fossil taxa with five different proportions: 0%, 10%, 25%, 50% or 75% of overall missing data for the fossil taxa. Note that 50% missing data for fossil taxa does not mean that each fossil is missing 50% of its morphological data. Instead this 50% refers to missing fossil data across the whole matrix. Some fossils may retain 100% of their data and others may lose most of their data at this parameter value (down to a minimum threshold of 5% available data; see below).
3. The number of morphological characters for both living and fossil taxa (N_C). This parameter is not a missing data parameter *per se* but rather an indication of the size of the matrix. Any morphological matrix of any size has indeterminate missing data, given that the total number of characters is undefined, but presumably large. Therefore, this parameter corresponds to the overall number of characters available for both living and fossil taxa. In practice, we randomly removed entire characters from the morphological matrix reducing it to: 100, 90, 75, 50 or 25 characters. Note that these levels are equivalent to the two other parameters (i.e. 0%, 10%, 25%, 50% or 75% of "missing" morphological characters).

Each parameter represents a different way of removing data from the morphological part of the matrix: M_L removes entire rows from the living data; M_F removes cells from the fossil data; and N_C removes columns across both living and fossil 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) "missing-data" matrices. Note that one of these combinations ($M_L = 0\%$; $M_F = 0\%$ and $N_C = 100$) has no missing data so is equivalent to the "complete" matrix, thus we have one effectively complete matrix in our 125 "missing-data" matrices. In practice, we first removed the data following the two missing data parameters M_L and M_F and then removed data following the N_C parameters. To avoid matrices containing taxa without any data (morphological or molecular), we repeated the random deletion until the matrices contained at least 5% of data for any taxa. Note that the living taxa always had at least 90% of data (the 1000 molecular characters).

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). We compare "best" trees to "missing data" trees but could also compare "true" trees to the "missing data" trees. In practice, the difference between the "best" trees and the "missing data" trees represents the effect of our missing data parameters and of the phylogenetic methods used to infer the "missing data" trees. The difference between the "true" and the "missing data" trees, however, represents the effect of our parameters used to generate the "true" tree and the algorithms used to generate the "complete" matrix as well as the effect of our missing data parameters and the phylogenetic methods used. Because the main aim of this study is to look at the effect our missing data parameters on topological recovery, we chose to represent only the comparisons between the "best" trees and "missing data" trees. The results of the comparisons of the "true" tree and the "missing data" trees are available in Appendix B. Note that this makes little difference to our overall results.

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 M$k\nu$ 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$k\nu$ + 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 × 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 N_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 (2.30 GHz clock speed nodes). We performed this procedure to increase the resolution of our resulting trees.

2.3.2. Bayesian inference

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$k\nu$ model (Lewis, 2001), with equal state frequency and a unique overall substitution rate (μ) with four distinct rates categories (M$k\nu$ + 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 used a fixed starting tree rather than a random starting tree (default MrBayes; Ronquist et al., 2012b) to speed up our Bayesian inferences. Note that a starting tree is not a Bayesian prior on topology *per se* and using a fixed starting tree did not significantly affect topology compared to using random starting trees (see Appendix A, section "Effect of the starting tree on Bayesian inference"). 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 (β(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 (2.30 GHz clock speed nodes). The detailed MrBayes parameters are available in Appendix A. We also included an analysis showing the effect of missing data on the estimation of the shape parameter (α) of the morphological substitution rate distribution. This extra analysis, however, is beyond the scope of this paper so the results are not discussed further here.

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 $\gg$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 N_C on tree topology. We used the Robinson–Foulds distance (Robinson and Foulds, 1981) to assess the number of conserved clade positions and the Triplets distance (Dobson, 1975) to assess the number 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, 2015) and well implemented (Bogdanowicz et al., 2012). We normalised both metrics using methods described in Bogdanowicz et al. (2012) to generalise 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 twice 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 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 method 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 is 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, N_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 bounded between 0 (no overlap) and 1 (full overlap; 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. When the Bhattacharyya Coefficient between two distributions is <0.05, the distributions are significantly different. When this coefficient is >0.95, the distributions are significantly similar. Values between these two thresholds show the probability of overlap between the distributions but do not allow us to define the significance of the similarity or differences between distributions. 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, N_C) and parameter states (0%, 10%, 25%, 50%, 75% and 100, 90, 75, 50, 25 characters), i.e. $M_L = 0\%$, $M_F = 0\%$, $N_C = 100$; $M_L = 10\%$, $M_F = 0\%$, $N_C = 100$, 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. Because of the difficulties in representing so many pairwise comparisons in a meaningful way, we summarised these results as a

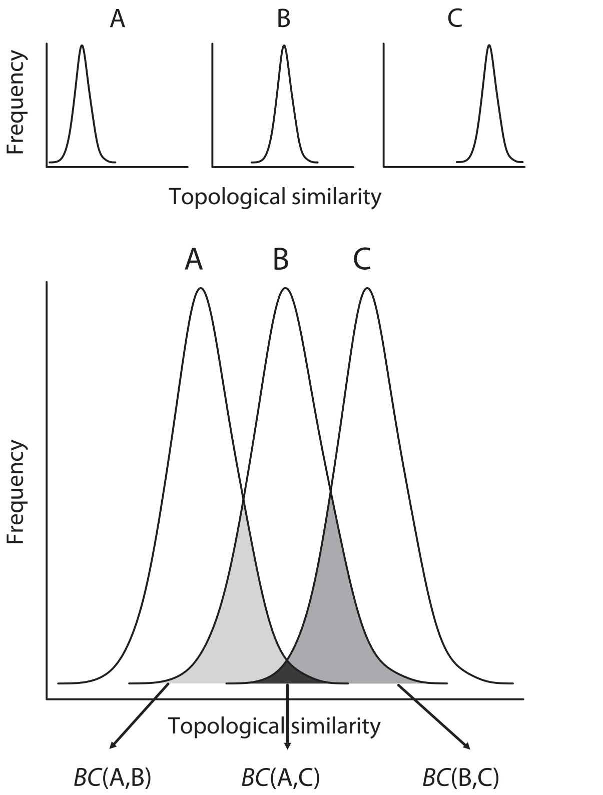

Fig. 2. Bhattacharyya Coefficient calculation outline 1. A, B and C are distributions of tree similarity metrics (Normalised Robinson–Foulds or Normalised Triplets metrics) for any combination of missing data parameters (e.g. $M_L = 10\%$, $M_F = 50\%$, $N_C = 25$). The Bhattacharyya Coefficient (BC) 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.

heat map of Bhattacharyya Coefficients (see Fig. 6). In this type of figure, parameters that have similar effects on recovering the "best" topology (either positive or negative effects) will be denoted by similar colour patches in the heat map representation of these comparisons (see Fig. 6).

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 *vs.* Bayesian consensus; Maximum Likelihood *vs.* Maximum Likelihood bootstrap trees, etc. (see Fig. 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 N_C), the tree inference method (Maximum Likelihood or Bayesian) or the tree comparison metric used (Normalised Robinson–Foulds or Normalised Triplets metric). Nonetheless, the different missing data parameters and tree inference methods do not affect the topology in the same way (Figs. 4 and 5).

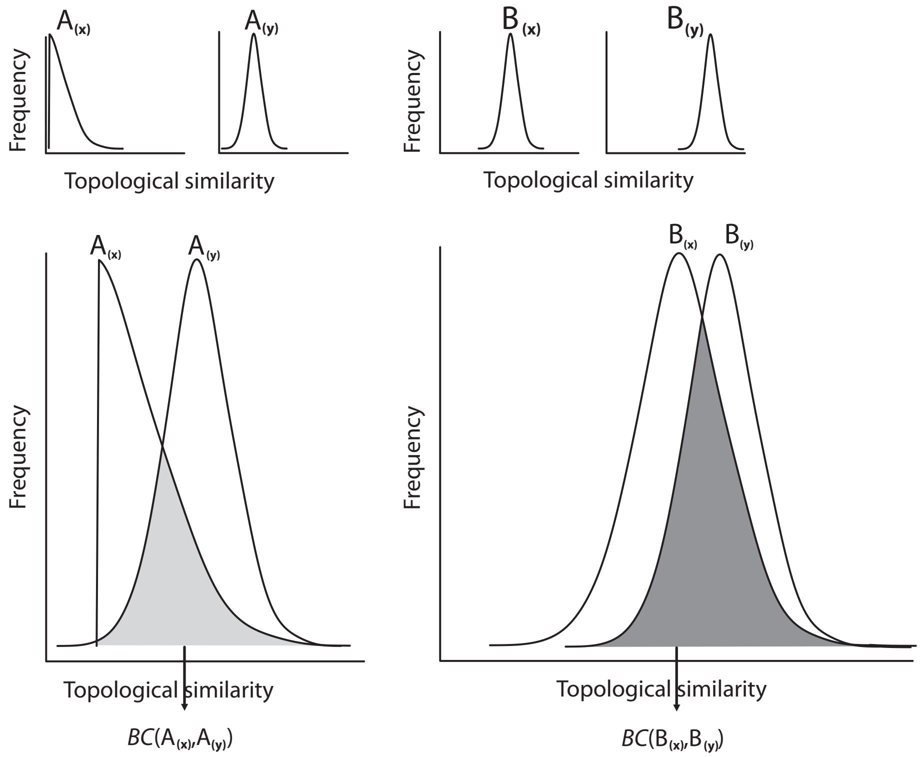

Fig. 3. Bhattacharyya Coefficient calculation outline 2. A and B are distributions of tree similarity metrics (Normalised Robinson–Foulds or Normalised Triplets metrics) for any combination of missing data parameters (e.g. $M_L = 10\%$, $M_F = 50\%$, $N_C = 25$). **(x)** and **(y)** are two different 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*.

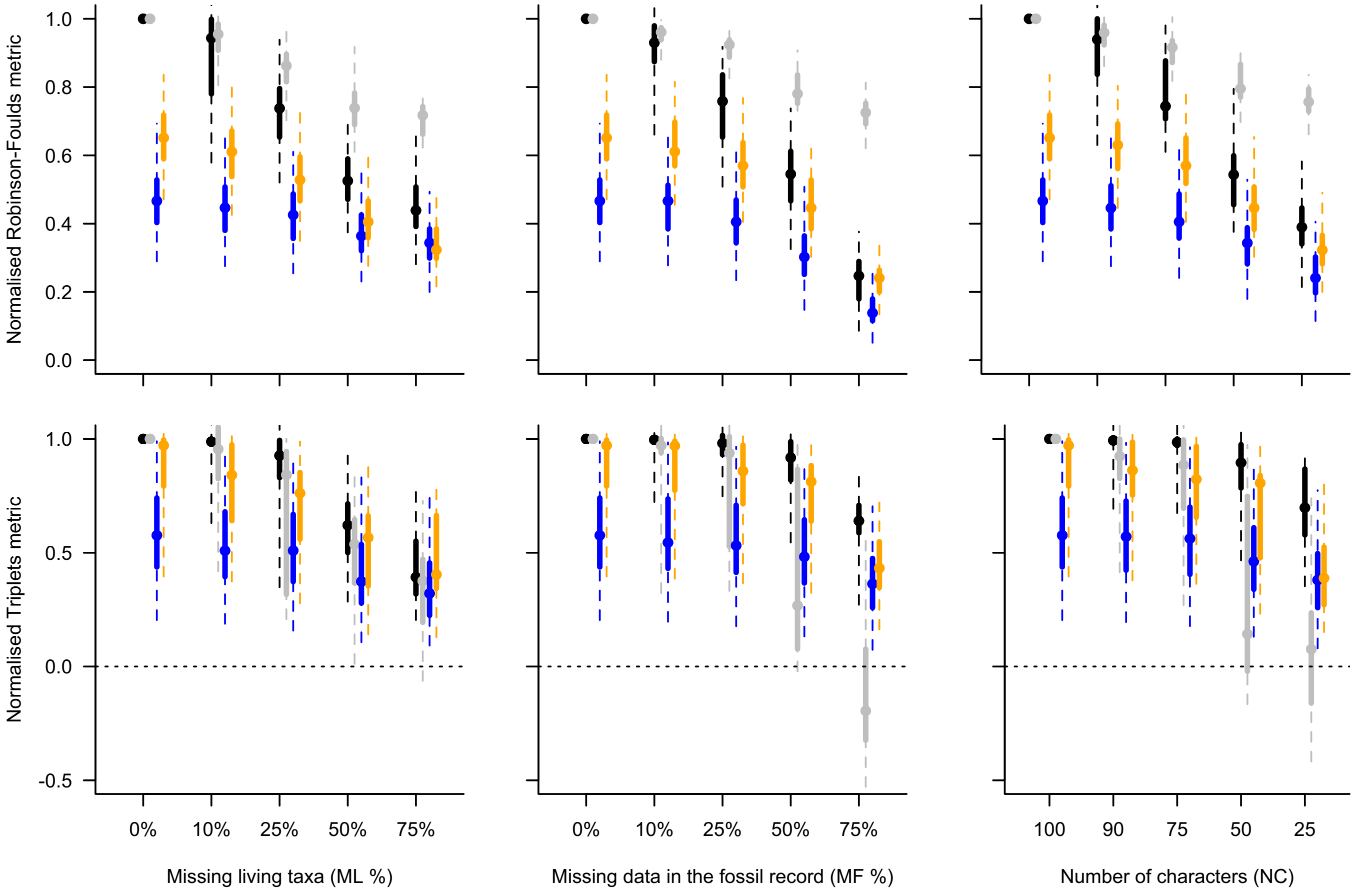

Fig. 4. The effects of increasing missing data on topological recovery using Maximum Likelihood trees (black), Bayesian consensus trees (grey), Maximum Likelihood bootstrap trees (blue) and Bayesian posterior tree distributions (orange). The percentage of missing data for each parameter (M_L, M_F and N_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. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

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 (Fig. 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 N_C respectively; Fig. 4; Tables C5, C6 and C7 in Appendix C). The Normalised Triplets metric, however, is similar when comparing the Maximum Likelihood trees and the Bayesian consensus trees for all the parameters (M_L, M_F and N_C) (median Bhattacharrya Coefficient = 0.84, 0.75 and 0.80 for M_L, M_F and N_C respectively; Fig. 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 $N_C = 75\%$), and best when there is no missing data (i.e. $M_L = 0\%$, $M_F = 0\%$, $N_C = 0\%$; Fig. 5; Tables C2, C3 and C4 in Appendix C). Fig. 6 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).

Using both Normalised Robinson–Foulds and Normalised Triplets metrics from the Bayesian consensus trees, the parameter combination with no missing data (i.e. $M_L = 0\%$, $M_F = 0\%$, $N_C = 100$) is always the most dissimilar to all the other parameter combinations (thin deep blue line at the base of Fig. 6). The Normalised Robinson–Foulds metric (median Bhattacharrya coefficient = 0.79; blue regions in Fig. 6A), however, displays more dissimilarities than the Normalised Triplets metric (median Bhattacharrya coefficient = 0.81; blue regions in Fig. 6B). The orange upper triangle in Fig. 6A shows a high probability of overlap of the Normalised Robinson–Foulds metric for the trees with the M_L parameter $\geqslant 50\%$ (Fig. 6A). Once $M_L \geqslant 50\%$, there is no additional effect of M_F and N_C, regardless of the amount of missing data in these parameters (Fig. 6A). Likewise, once $N_C < 50$, there is no additional effect of M_L and M_F as denoted by the high probability of Normalised Robinson–Foulds metric overlap (horizontal orange stripes between the blue regions Fig. 6A). In Fig. 5 for the Normalised Robinson–Foulds metric, this can be interpreted as the overlap between the distributions once $M_L = 50\%$.

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). These two methods, however, perform worse than the Bayesian consensus trees using Normalised Robinson–Foulds metric (median Bhattacharrya Coefficient = 0

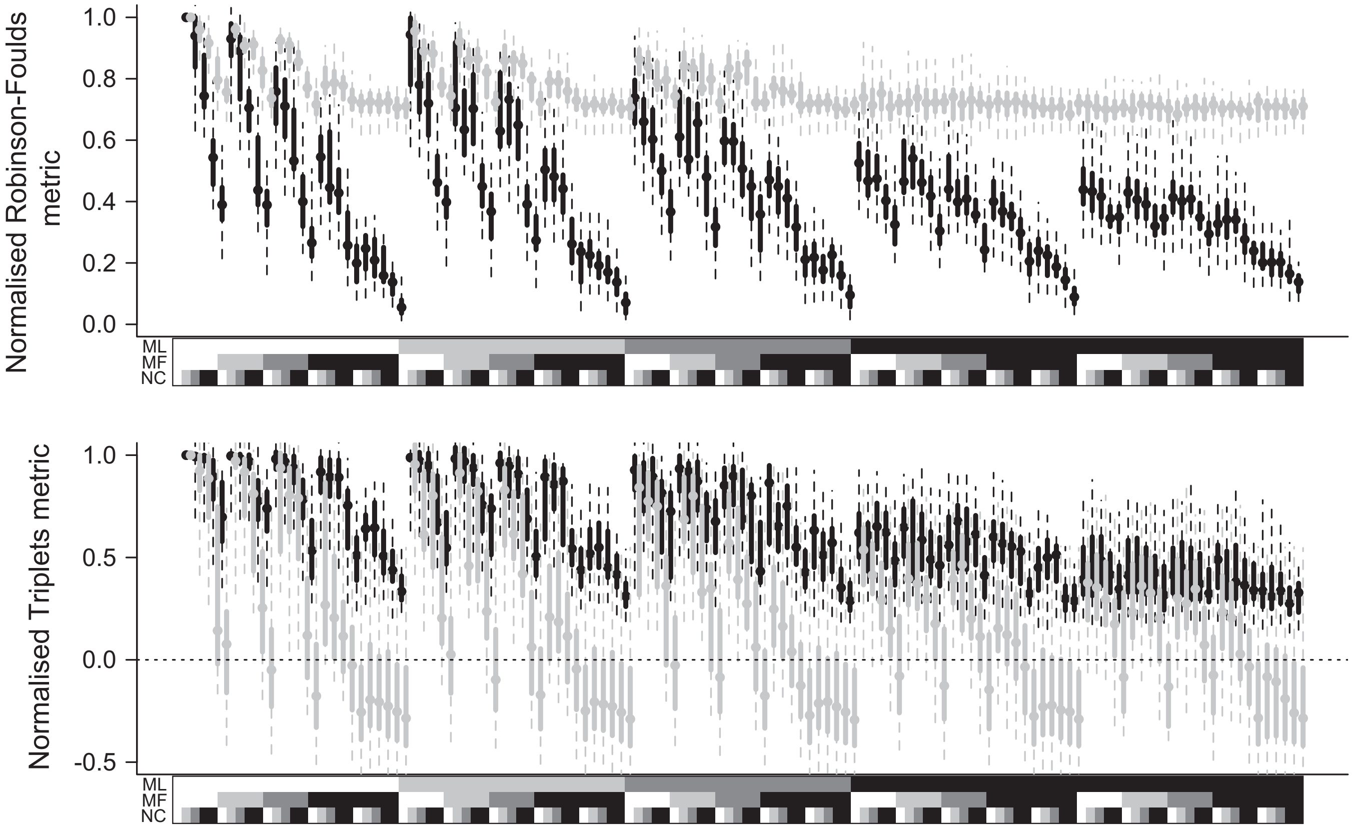

Fig. 5. The effects of increasing missing data on topological recovery using Maximum Likelihood trees (black) and Bayesian consensus trees (grey). The *x* axis shows the percentage of missing data from 0% (white) to 75% (black) for the two parameters: M_L (upper line), M_F (middle line) and number of characters from 100 to 25 for the parameter N_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.

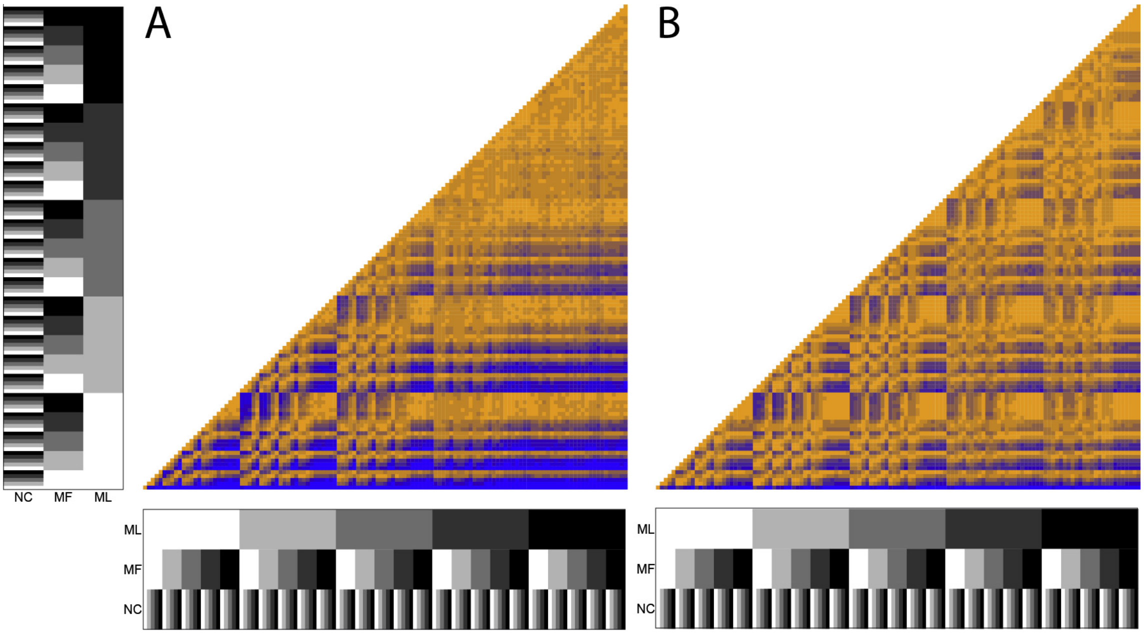

Fig. 6. The effects of missing data on topological recovery using Bayesian consensus trees. Both axes show the percentage of missing data from 0% (white) to 75% (black) for the three parameters: M_L (upper line), M_F (middle line) and N_C (lower line). The topological recovery is measured as (A) the Normalised Robinson–Foulds metric and (B) the Normalised Triplets metric calculated using the Bhattacharyya Coefficient. The Bhattacharyya Coefficient values are indicated using a colour gradient ranging from low probability of overlap in blue, to high probability of overlap in orange. Blue regions denote a poor overlap in Normalised metric values between the different parameter combinations (i.e. the parameters have a strong effect on the metric and thus the topological recovery). Conversely, orange regions denote a high overlap in Normalised metric values between the different parameter combinations (i.e. the parameters have a weak effect on the metric and thus the topological recovery). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Bhattacharyya Coefficients of the pairwise method comparisons. Each line summarises the probabilities of overlap between the distributions of the "best" tree *vs.* trees from each inference method (Maximum Likelihood; Bayesian consensus; Maximum Likelihood Bootstraps and Bayesian posterior trees) pooled across all combinations of missing data parameter values, using the Normalised Robinson–Foulds (RF) and Triplets (Tr) metrics. 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. Values >0.95 denote significantly similar distributions and values <0.05 denote significantly different distributions.

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**

and 0.01, for the Maximum Likelihood bootstrap trees and the Bayesian posterior tree distribution respectively; Table 1; Fig. 4 and Fig. 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. These factors, however, affected topological recovery in different ways and to different extents. Decreasing the number of living taxa with morphological data (M_L) and the overall number of morphological characters in the matrix (N_C) had worst effects on topological recovery (Fig. 6). Additionally, using Bayesian consensus trees recovered the "best" tree topology more consistently than using Maximum Likelihood trees or Bayesian posterior tree distributions (Figs. 5 and 6, Table 1). As seen in previous studies, our results show that the amount of missing data are 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). This missing data parameter, however, 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. It is important, however, 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 are high (M_F = 75%). These results are in agreement with Manos et al. (2007) where as few as 16 characters were sufficient for correctly assigning artificial fossils to their correct clade.

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; Figs. 4 and 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 taxa 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 near the tips but affects the clade conservation less when fossils are near 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 taxon 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. Number of morphological characters (N_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 (i.e. $N_C = 90$) genuinely removes 10% of data in the matrix. Nonetheless, 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. 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 of 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 (Bremer and Struwe, 1992; Patterson et al., 1993; see Masters and Brothers, 2002 for an empirical example). The sizes 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 (Figs. 4 and 5). 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. Nonetheless, 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. $M_L = 0\%$, $M_F = 0\%$, $N_C = 100$) and lower when they have a larger proportion of missing data (i.e. $M_L = 75\%$, $M_F = 75\%$ and $N_C = 25$; Fig. 5). It is important, however, to notice that the effect of each parameter is not additive. Surprisingly, the number of missing living taxa with morphological data (M_L) and the overall number of missing morphological characters (N_C), have a bigger effect than the amount of missing data for the fossil taxa (M_F). For any additional missing living taxa with morphological data (M_L) beyond 50%, there is no difference among trees with any combination of the other parameters (M_F and N_C; Fig. 6). In other words, when the number of missing living taxa reaches 50%, neither the amount of missing data in the fossil record (M_F), nor the number of characters in the matrix (N_C) affect topology. A similar effect can be observed when the N_C parameter reaches 50 characters (Fig. 6). This has important practical implications, especially for the best strategy to improve topology by collecting more morphological data (see below).

4.3. Effects of tree inference methods

Variation in our ability to recover the "best" tree topology depends heavily on the tree inference method (Figs. 4 and 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 parsimony methods (Wright and Hillis, 2014; but see Spencer and Wilberg, 2013). This is, however, 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 "best" Maximum Likelihood tree (Fig. 5). Note that the methodological choice of using the "true" tree as a starting tree for the Bayesian Inference rather than a random starting tree (see Methods), had no significant effect on topological recovery (see Appendix A, section "Effect of the starting tree on Bayesian inference" for details). 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. Nonetheless, we minimised the probability of unresolved "best" trees in our Bayesian analyses by only using datasets with strong phylogenetic signal (see Section 2).

The Bayesian consensus trees, however, perform poorly for the Normalised Triplets metric: some parameter combinations, especially when the M_F parameter reaches 75% missing data, lead to negative values (Fig. 5). A Normalised Triplets metric value below 0 means that the placement of some taxa is worse than expected by just randomly placing this taxon in the tree. This can be interpreted as the absence of comparable triplets between some of the "missing data" trees and "best" trees. Even if clades are conserved (Fig. 5), the resolution within them can be poor to non-existent when a large proportion of data are missing (i.e. 75%). In such cases, the fossil taxa are equally likely to be placed in any of the clades that they share the most characters with. These results are in agreement with previous studies that have showed that missing data can cause problems for recovering "correct" topologies, especially for small matrices of 100 characters (Wiens, 2003). It is important to note, however, that this effect can be reduced by increasing the number of characters (Wiens, 2003).

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 were missing ($M_L = 0\%$; $M_F = 0\%$, $N_C = 100$; Fig. 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. Additionally, the Bayesian posterior trees performed more poorly than the Bayesian consensus tree (Fig. 4, Table 1 and Appendix C Fig. C5 and Tables C5, C6 and C7). This may be because the Bayesian posterior trees are always resolved and thus more likely to contain incorrectly resolved nodes (i.e. decreasing the Normalised Robinson–Foulds metric). Conversely, the Bayesian consensus trees might not resolve nodes that are poorly supported and thus are more likely to contain only correctly resolved nodes (i. e. increasing the Normalised Robinson–Foulds metric).

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 (N_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. Gaps in the matrix, however, 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 these data are the most difficult to collect, they also have the least influence on whether our simulations recover the "best" tree topology (Fig. 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 single 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 that where a topological constraint is needed, Bayesian consensus trees should be used. This may apply to tree inferences using the Total Evidence method such as tip-dating (e.g. Ronquist et al., 2012a; Wood et al., 2013; Matzke, 2014). It is, however, possible that including dating information during tree inference could also improve the accuracy of the Bayesian posterior tree distribution, so a fixed topology should be used with caution. Using the Bayesian consensus tree rather than the Maximum Likelihood can also reduce the number of false positive topologies (*sensu* Swofford et al., 2001). As shown in Fig. 5 and discussed in the section above (Effects of tree inference methods), the Bayesian consensus tree is more likely to not resolve poorly supported nodes due to missing data than the Maximum Likelihood tree that is more likely to incorrectly resolve such nodes (i.e. creating a false positive node). Note, however, that we do not suggest discarding the Bayesian posterior tree distributions even though they performed poorly in recovering the "best" tree topology in our simulations (this can probably be traced to the difficulties comparing distributions of trees; see above). These trees will be invaluable for phylogenetic comparative analyses. For example a sub-sample of posterior tree distributions can be used to assess macroecological questions while better taking into account topological uncertainty (e.g. Fritz et al., 2013 and Jetz et al., 2012 used in Healy et al., 2014).

5. Conclusions

Previous studies have explored the effect of missing morphological data in Total Evidence matrices (Wiens et al., 2005; Manos et al., 2007; Pattinson et al., 2014). The conclusions of theses studies, however, were limited by their empirical approach making their results applicable only to similar missing data scenarios. Additionally, these studies focused either only on living taxa (Wiens et al., 2005) or on the patterns of missing data from the fossil record only (Manos et al., 2007; Pattinson et al., 2014). Here we instead used an approach where missing data were generated from simulated data and according to three clearly defined missing-data parameters (M_L, M_F or N_C) that removed data from both the living and fossil taxa. This allowed us to confirm previous results that missing data can be especially problematic in small matrices (Wiens, 2003), but also revealed the crucial importance of coding morphological data for living species in Total Evidence phylogenies. 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 N_C), it reduces support for the placement of fossil taxa and increases the displacement of wildcard taxa. Therefore we advise increased focus on coding morphological characters for a large number of the living taxa present in the matrix (i.e. at least 50%) if the goal is to accurately combine both living and fossil species in phylogenies. Doing so will increase overlap of morphological characters among living and fossil taxa, allowing the fossil taxa to be positioned relative to the living taxa based on their shared derived characters rather than simply on available data.

Additionally, the topologies of the Bayesian consensus trees, regardless of the amount of missing data, were always closer to the "best" tree topology than the Maximum Likelihood trees. This has also been observed in empirical data (e.g. Arcila et al., 2015) where Maximum Likelihood trees inferred from a Total Evidence matrix were less supported than the Bayesian consensus tree. This might have an important impact on estimating topologies in the Total Evidence framework, because previous studies had to rely either on molecular scaffolds (e.g. Slater, 2013), taxonomic constraints (e.g. Slater, 2013; Beck and Lee, 2014) or even on fixing the topology (e.g. Ronquist et al., 2012a). Therefore, we suggest extracting such topological backbones from the Bayesian consensus tree if needed.

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. Thanks to one anonymous reviewer and Alex Pyron for their useful and encouraging comments. 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).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2015.08.023.

References

Arcila, D., Pyron, R.A., Tyler, J.C., Ort, G., Betancur-R, R., 2015. An evaluation of fossil tip-dating versus node-age calibrations in tetraodontiform fishes (teleostei: Percomorphaceae). Molec. Phylogenet. Evol. 82 (Part A), 131–145.

Bapst, D.W., 2013. A stochastic rate-calibrated method for time-scaling phylogenies of fossil taxa. Methods Ecol. Evol. 4, 724–733.

Beck, R.M., Lee, M.S., 2014. Ancient dates or accelerated rates? Morphological clocks and the antiquity of placental mammals. Proc. Roy. Soc. B: Biol. Sci. 281, 1–10.

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., Giaro, K., Wróbel, B., 2012. TreeCmp: comparison of trees in polynomial time. Evol. Bioinform. 8, 475–487.

Bremer, B., Struwe, L., 1992. Phylogeny of the rubiaceae and the loganiaceae: congruence of conflict between morphological and molecular data? Am. J. Botany, 1171–1184.

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

Critchlow, D.E., Pearl, D.K., Qian, C., 1996. The triples distance for rooted bifurcating phylogenetic trees. Syst. Biol. 45, 323–334.

Dietl, G.P., Flessa, K.W., 2011. Conservation paleobiology: putting the dead to work. Trends Ecol. Evol. 26, 30–37.

Dobson, A.J., 1975. In: Comparing the Shapes of Trees. Lecture Notes in Mathematics, vol. 452. Springer, Berlin Heidelberg, pp. 95–100.

Douady, C., Delsuc, F., Boucher, Y., Doolittle, W., Douzery, E., 2003. Comparison of bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Molec. Biol. Evol. 20, 248–254.

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

Eernisse, D., Kluge, A., 1993. Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology. Molec. Biol. Evol. 10, 1170–1195.

Estoup, A., Jarne, P., Cornuet, J.-M., 2002. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. Molec. Ecol. 11, 1591–1604.

Felsenstein, J., 2004. Inferring Phylogenies. Sinauer Associate, Sunderland, Massachusetts.

FitzJohn, R.G., 2012. Diversitree: comparative phylogenetic analyses of diversification in R. Methods Ecol. Evol. 3, 1084–1092.

Friedman, M., 2010. Explosive morphological diversification of spiny-finned teleost fishes in the aftermath of the end-Cretaceous extinction. Proc. Roy. Soc. B: Biol. Sci. 277, 1675–1683.

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

Guillerme, T., Cooper, N., 2015. Assessment of cladistic data availability for living mammals. bioRxiv doi: http://dx.doi.org/10.1101/022970.

Hasegawa, M., Kishino, H., Yano, T.A., 1985. Dating of the human ape splitting by a molecular clock of mitochondrial-DNA. J. Molec. Evol. 22, 160–174.

Hassanin, A., Lecointre, G., Tillier, S., 1998. The evolutionary signal of homoplasy in proteincoding gene sequences and its consequences for a priori weighting in phylogeny. Comp. Rend. l'Acad. Sci. – Ser. III – Sci. Vie 321, 611–620.

Healy, K., Guillerme, T., Finlay, S., Kane, A., Kelly, S.B.A., McClean, D., Kelly, D.J., Donohue, I., Jackson, A.L., Cooper, N., 2014. Ecology and mode-of-life explain lifespan variation in birds and mammals. Proc. Roy. Soc. Lond. B: Biol. Sci. 281.

Heath, T.A., Huelsenbeck, J.P., Stadler, T., 2014. The fossilized birthdeath process for coherent calibration of divergence-time estimates. Proc. Natl. Acad. Sci. 111, E2957–E2966.

Hennig, W., 1966. Phylogenetic Systematics. University of Illinos Press, Urbana.

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

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

Jetz, W., Thomas, G., Joy, J., Hartmann, K., Mooers, A., 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. Syst. Biol. 51, 369–381.

Kelly, S.B.A., Kelly, D.J., Cooper, N., Bahrun, A., Analuddin, K., Marples, N.M., 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., Yamato, J., 2015. Practical performance of tree comparison metrics. Syst. Biol. 64, 205–214.

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

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

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

Masters, J.C., Brothers, D.J., 2002. Lack of congruence between morphological and molecular data in reconstructing the phylogeny of the galagonidae. Am. J. Phys. Anthropol. 117, 79–93.

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

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

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

Nylander, J.A.A., Ronquist, F., Huelsenbeck, J.P., Nieves-Aldrey, J., 2004. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53, 47–67.

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

O'Leary, M.A., Kaufman, S., 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. Proc. Roy. Soc. Lond. Ser. B: Biol. Sci. 255, 37–45.

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

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

Pattengale, N.D., Alipour, M., Bininda-Emonds, O.R., Moret, B.M., Stamatakis, A., 2010. How many bootstrap replicates are necessary? J. Comput. Biol. 17, 337–354.

Patterson, C., Williams, D.M., Humpries, C.J., 1993. Congruence between molecular and morphological phylogenies. Annu. Rev. Ecol. Syst., 153–188

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

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

Price, M.N., Dehal, P.S., Arkin, A.P., 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. Syst. Biol. 60, 466–481.

Quental, T., Marshall, C., 2010. Diversity dynamics: molecular phylogenies need the fossil record. Trends Ecol. Evol. 25, 434–441.

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

Rambaut, A., Grassly, N.C., 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13, 235–238.

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

Robinson, D.F., Foulds, L.R., 1981. Comparison of phylogenetic trees. Math. Biosci. 53, 131–147.

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

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

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

Rozen, D.E., Schneider, D., Lenski, R.E., 2005. Long-term experimental evolution in *Escherichia coli*. XIII. Phylogenetic history of a balanced polymorphism. J. Molec. Evol. 61, 171–180.

Salamin, N., Chase, M.W., Hodkinson, T.R., Savolainen, V., 2003. Assessing internal support with large phylogenetic DNA matrices. Molec. Phylogenet. Evol. 27, 528–539.

Sallam, H.M., Seiffert, E.R., Simons, E.L., 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., McMahon, M.M., Steel, M., 2011. Terraces in phylogenetic tree space. Science 333, 448–450.

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

Schrago, C., Mello, B., Soares, A., 2013. Combining fossil and molecular data to date the diversification of New World Primates. J. Evol. Biol. 26, 2438–2446.

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

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

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

Springer, M.S., Meredith, R.W., Gatesy, J., Emerling, C.A., Park, J., Rabosky, D.L., Stadler, T., Steiner, C., Ryder, O.A., Janecka, J.E., Fisher, C.A., Murphy, W.J., 2012. Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix. PLoS ONE 7, e49521.

Stadler, T., Yang, Z., 2013. Dating phylogenies with sequentially sampled tips. Syst. Biol. 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., Hoover, P., Rougemont, J., 2008. A rapid bootstrap algorithm for the raxml web servers. Syst. Biol. 57, 758–771.

Swofford, D.L., Waddell, P.J., Huelsenbeck, J.P., Foster, P.G., Lewis, P.O., Rogers, J.S., 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst. Biol. 50, 525–539.

Tavaré, S., 1986. Some probabilistic and statistical problems in the analysis of dna sequences. Lect. Math. Life Sci. 17, 57–86.

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

Wiens, J.J., 2006. Missing data and the design of phylogenetic analyses. J. Biomed. Inform. 39, 34–42.

Wiens, J.J., Fetzner, J.W., Parkinson, C.L., Reeder, T.W., 2005. Hylid frog phylogeny and sampling strategies for speciose clades. Syst. Biol. 54, 778–807.

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

Wood, H.M., Matzke, N.J., Gillespie, R.G., Griswold, C.E., 2013. Treating fossils as terminal taxa in divergence time estimation reveals ancient vicariance patterns in the palpimanoid spiders. Syst. Biol. 62, 264–284.

Wright, A.M., Hillis, D.M., 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 Ecol. Evol. 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., Pauling, L., 1965. Molecules as documents of evolutionary history. J. Theor. Biol. 8, 357–366.