Effect of missing data in matrices containing living and fossil taxa on topological accuracy

Thomas Guillerme[1,2], other authors [3], and Natalie Cooper[1,2]

[1]*School of Natural Sciences, Trinity College Dublin, Dublin 2, Republic of Ireland;*

[2]*Trinity Centre for Biodiversity Research, Trinity College Dublin, Dublin 2, Republic of Ireland;*

[3]*Somewhere else*

Corresponding author: Thomas Guillerme, School of Natural Sciences, Trinity College Dublin, Dublin 2, Republic of Ireland; E-mail: guillert@tcd.ie.

Abstract.— Living species represent less than 1% of all species that have ever lived. Ignoring fossil taxa may 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 fossil and living taxa must be included in macroevolutionary studies. One approach, the total evidence approach, uses molecular data from living taxa and morphological data from both living and fossil taxa to infer phylogenies with both fossil and living 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, in a Bayesian framework, difficulties in recovering a stable topology are mainly driven by the missing data in the molecular part of the matrix (for which fossil taxa have no data). In a Maximum Likelihood framework, however, topology is not directly affected by missing data *per se*, but by the number of morphological characters shared among the taxa. Therefore, the two main drivers of incorrect topologies are the overall number of morphological characters and the number of living species with no morphological data.

Our results suggest that, in order to use total evidence approaches, one should reduce the missing data in the morphological part of the matrix for living species and use a Maximum Likelihood framework to fix the topology prior to the overall Bayesian phylogenetic inference process.
(Keywords: missing data, total evidence, Bayesian, Maximum Likelihood, topology)

Living species represent less than 1% of all species that have ever lived (Novacek and Wheeler 1992; Raup 1993). However, the majority of macroevolutionary studies focus solely on living species (Cooper and Purvis 2009; Meredith et al. 2011; Healy et al. 2014). Ignoring fossil taxa may lead to misinterpretation of macroevolutionary patterns and processes such as 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). These factors have led to increasing consensus among scientists that fossil taxa must 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 which still remains complex despite an increasing number of significant efforts (Pyron 2011; Ronquist et al. 2012a; Schrago et al. 2013).

Up to now, three main approaches have been used for combining fossil and living taxa data in phylogenies. These approaches differ mainly in whether they treat fossil taxa as tips or as nodes in the phylogeny and on which part of the available data of the fossil is used (i.e. the fossil occurrence age only or both the age and the morphology). Classical cladistic methods uses matrices containing morphological data from both fossil and living taxa and allows to treat each taxon as a tip in the phylogeny; the relation between the taxa can be inferred using optimal criteria such as the maximum parsimony criterion (Simpson 1945). This approach is commonly used by paleontologists but it ignores the additional molecular data available from living species and doesn't allow to use probabilistic methods to deal with inference uncertainty (but see Spencer and Wilberg (2013). Neontologists, on the other hand, more commonly use probabilistic methods (e.g. Maximum Likelihood or Bayesian) 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 analyses and only their occurrence age can be 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). However none of them is optimal and uses all the available data.

A final approach, known as total evidence approach use matrices containing molecular data from living taxa and morphological data from both living and fossil taxa. It allows to use probabilistic methods, to treat every taxon as a tip in the phylogeny and to use the occurrence age of the fossil to time calibrate the phylogeny (Eernisse and Kluge 1993). Here we focus on this total evidence approach because they have been recently successfully developed and applied to empirical data (Pyron 2011; Ronquist et al. 2012a; Schrago et al. 2013). Although the total evidence approach seems very promising, there is one big drawback in using this approach: it requires a lot of data. In particular it requires morphological data form both living and fossil taxa, both of which are known to be scarce. Therefore total evidence approach is likely to suffer from having lots of missing data which may affect their ability to infer correct topology, branch length and support values (Salamin et al. 2003).

The effect of missing data on phylogenetic inferences has been widely studied (Wiens 2003, 2006; Wiens and Moen 2008; Lemmon et al. 2009; Roure and Philippe 2011; Sansom and Wills 2013). Missing molecular data has been seen by some authors as an issue because it can decrease the phylogenetic signal in some parts of the tree, especially when using large matrices (Lemmon et al. 2009). However other authors do not see missing molecular data as a major issue because the phylogenetic signal, rather than by reducing the amount of missing data, is more likely to increase by having: (i) at least a "modest" number of highly covered genes (i.e. approximatively half of the genes - Roure and Philippe (2011)); (ii) a higher number of taxa (especially slowly evolving taxa or taxa close to the outgroup - Roure and Philippe (2011)) and (iii) by choosing more adequate models

of sequence evolution rather than by reducing the amount of missing data (Wiens 2006; Wiens and Moen 2008; Roure and Philippe 2011).

Similarly, depending on the study, missing morphological data might be seen as either a major or minor issue for accurately inferring phylogenies depending on the study (Wiens 2003; Sansom and Wills 2013). Because soft-tissues characters are rarely preserved in the fossil record, missing data is mainly found in soft tissues characters, and is therefore not randomly distributed, which can lead to biased placement of fossil taxa in phylogenies (Sansom and Wills 2013). However, the phylogenetic signal is not related to the level of missing data *per se* but to the number of informative characters per taxa, therefore missing data is less an issue than the number of shared informative characters (Wiens 2003). Although missing data as been shown to be no major problem separately in molecular and morphological matrices (Wiens 2003, 2006; Wiens and Moen 2008; Roure and Philippe 2011) it may become an issue when combined in a total evidence type matrices. Because the nature of total evidence matrices, combining both molecular and morphological information for both living and fossil taxa, a big part of the matrix contains only missing data - the molecular data for fossil taxa. Until now, no attempt has been made to study the impact of this issue on phylogenetic inference and especially on topological recovery.

Here we assess the effect of missing data on tree topology inferred from total evidence matrices. The molecular part for living taxa of a total evidence matrix acts like a "classical" molecular matrix (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 only focus on the missing data issue in the morphological part of the matrix. 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 taxa (i.e. the preservation quality of the fossil record) and (3) the overall number of morphological characters for both

living and fossil taxa in the matrix. We assess how removing data from a total evidence type matrix by changing the values of these three parameters affects the topology of the inferred tree using Maximum Likelihood and Bayesian method.

We found that whatever the amount of missing data, total evidence matrices dramatically reduce the performance of Bayesian inference. However, when using a Maximum Likelihood approach, missing data decreases gradually the ability of recovering the right tree topology. We propose that this drastic difference between Bayesian and Maximum Likelihood methods is due to a flattening of the likelihood landscape because of the intrinsic amount of missing data in a total evidence matrix (i.e. the missing molecular data for the fossil taxa). This problem leads to an inability of Bayesian chains to pick up the sub and near optimal topologies leading to high variance in the posterior tree distribution and really low support in the consensus tree. These results should encourage authors in not using Bayesian inference to explore the topological landscape of their phylogeny. Instead, we suggest that authors should first find the Maximum Likelihood tree and use it's topology as a fix prior in their Bayesian inference. We also propose a new method for comparing Bayesian posterior tree distribution's topology available in a wonderful R package.

Methods

To explore the effect of missing data in total evidence matrices on 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 birth death tree (hereafter called the "true" tree, Table 1) to infer a matrix containing both molecular and morphological data for living and fossil taxa (hereafter called the "complete" matrix, Table 1).

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 (1) the proportion of missing living taxa (M_L), (2) the proportion of missing data in the fossil taxa (M_F) and (3) the proportion of missing morphological characters (M_C) (the resulting matrices are called hereafter the "missing-data" matrix, Table 1).

3. Building phylogenies:

   We inferred Bayesian phylogenetic trees from the "complete" matrix and from the "missing-data" matrices resulting in respectively a tree generate from a matrix containing no missing data in the morphological part (hereafter called the "best" tree, Table 1) and trees inferred from matrices with missing data in their morphological part (hereafter called the "missing-data" trees, Table 1).

4. Comparing topologies:

   We then 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 the phylogenies we estimated.

To measure the effect of missing data distribution, we repeated steps 1 to 4 with the exact same fixed parameters 51 (3×17) times. A list of all the terms used in this paper is available in table 1.

Generating the matrix

First we randomly generated a "true" tree of 50 taxa in R v3.0.2 (R Core Team 2014) using the package diversitree v0.9-6 (FitzJohn 2012). We generated the tree using a Birth Death process by sampling the values of the speciation events (λ) and extinction events (μ)
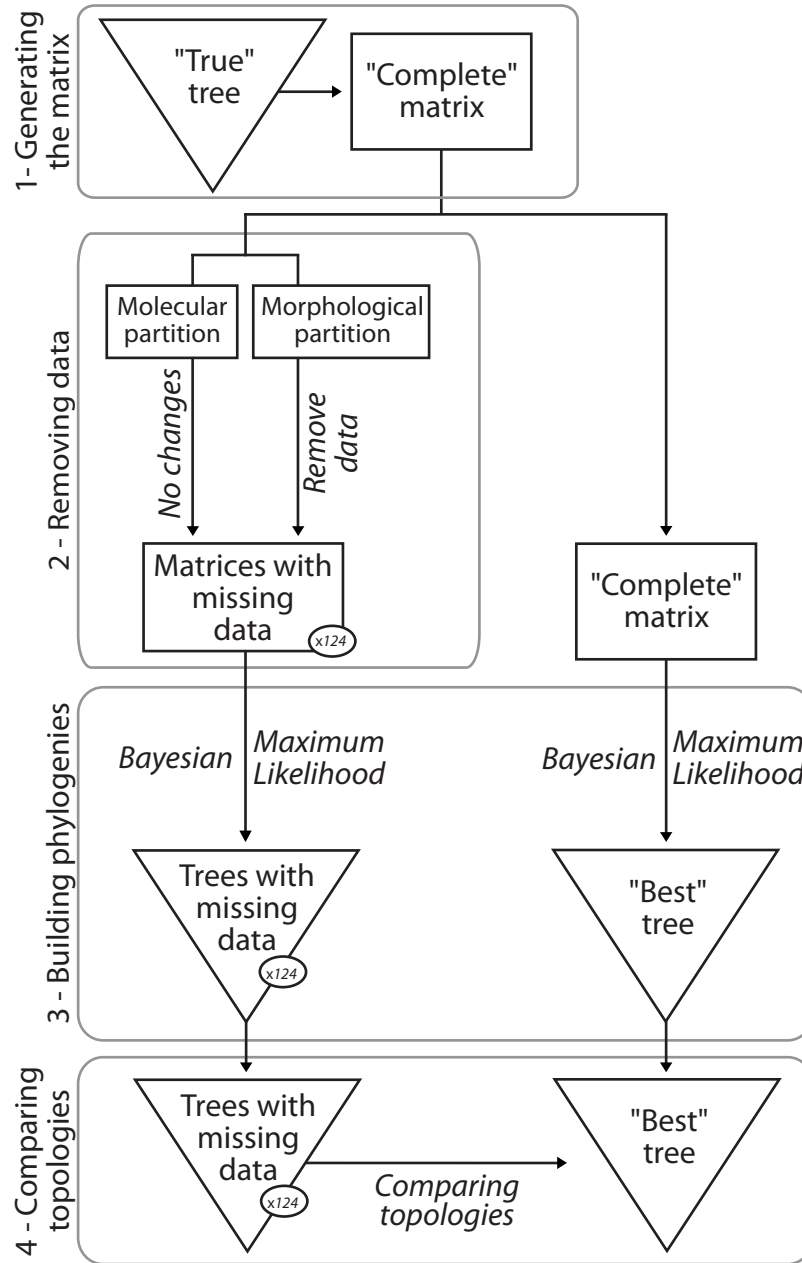
Figure 1: Protocol outline. (1) We generated a random tree (the "true" tree) to infer a matrix with no missing data (the "complete" matrix). (2) We removed data from the morphological partition of the "complete" matrix resulting in 125 "missing-data" matrices. (3) We infer phylogenetic trees from each matrix in both ML and Bayesian framework. (4) We then compared the "missing-data" trees to the "best" tree. We repeated step 1 to 4 51 (3×17) times.

from a uniform distribution but maintaining $\lambda > \mu$ (Paradis 2011). We implemented a rejection sampling algorithm to select only random trees with 25 living and 25 fossil taxa. We then added a taxa to the resulting Birth-Death tree as the outgroup of the tree. The mean branch length of the tree was used to separate the outgroup from the rest of the taxa and the branch length leading to the outgroup was set as the sum of the mean branch length and the longest root-to-tip length of the tree.

Next, we created a molecular and a morphological matrix from the "true" tree. The molecular matrix was inferred from the "true" tree using the package phyclust v0.1-14 (Chen 2011). The matrix was made of 1000 characters sites for 51 taxa and generated using the seqgen algorithm (Rambaut and Grassly 1997). We used the HKY model (Hasegawa et al. 1985) with a random base frequencies and with the transition/transversion rate of 2 (Douady et al. 2003) as parameters for generating the matrix. The substitution rates were distributed following a gamma distribution with an alpha (α) shape of 0.5 (Yang 1996). We chose a low value of α to reduce the number of sites with high substitution rates, thus avoiding too much homoplasy and a decrease in phylogenetic signal. These parameters were selected to generate data with no special assumption about how the characters evolved as well as to reduce the computational time required if these parameters were estimated rather than defined (total computational time > 65 CPU years).

We inferred the morphological matrix using the ape package v3.0-11 (Paradis et al. 2004) to generate a matrix of 100 character sites for 51 taxa. We assigned the number of character states (either 2 or 3) 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 characters states extracted from 100 published empirical morphological matrices (See supplementaries). We then ran an independent discrete character simulation for each character using the "true" tree branch length and topology with the randomly selected number of states (2 or 3) and assuming an

equal rate of change (i.e. evolutionary rate) from one character state to an other (Pagel 1994). This method allows us to have only two parameters per 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 a low evolution rate parameter (i.e. α) in order to avoid homoplasy in the morphological part of the matrix and create a clear phylogenetic signal (Wagner 2000; Davalos et al. 2014).

All the molecular information for fossil taxa was replaced by missing data (”?”). Finally, we combined the morphological and molecular matrices obtained from the ”true” tree. Hereafter we call this the ”complete” matrix: the matrix with no missing data except for the molecular data of the fossil taxa.

Removing data

Once we obtained the ”complete” matrix we modified it to get a set of matrices with missing data. We randomly replaced data with ”?” in the morphological part of the matrices according to the following parameters (Fig. 2):

1. The proportion of living taxa with no morphological data (M_L): 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 one. Because of the increasing facility to sequence DNA for living taxa, the number of living taxa with molecular data is highly superior the the number of taxa with molecular and morphological data.

2. The proportion of missing morphological data across all fossil taxa (M_F): 0%, 10%, 25%, 50% or 75%. This parameter illustrates the quality of the fossil record.

3. The proportion of missing morphological characters across all taxa (living and fossil -
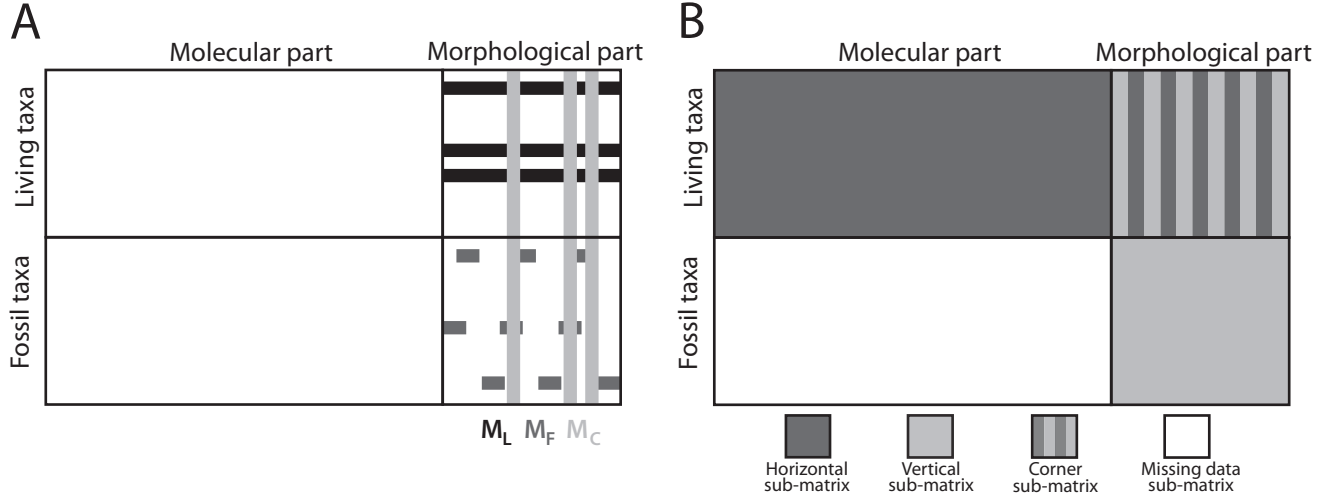
Figure 2: Names of the different parts of the matrix. A: Missing data parameters: Missing living - The proportion of living taxa with no morphological data (M_L); Missing fossil - The proportion of missing morphological data across all fossil taxa (M_F); Missing character - The proportion of missing morphological characters across all taxa (living and fossil) (M_C). B: Different parts of the matrix: The "horizontal" sub-matrix (orange) contains molecular and morphological data for living taxa; The "vertical" sub-matrix (blue) contains morphological data for living and fossil taxa; The "corner" sub-matrix (orange and blue striped) contains morphological data for living taxa; The "missing-data" sub-matrix (grey) is the molecular part of the fossil taxa and contains no data.

M_C): 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 represent a different way of removing data in the morphological part of the matrix: M_L removes a proportion of rows from the living taxa; M_F removes a proportion of cells from the fossil taxa; and M_C removes a proportion of columns across both living and fossil taxa (see Fig. 2). Note that M_L is different to M_F not only because of the region of the matrix affected: for M_L, all the morphological data of a proportion of the living taxa is removed (i.e. removing rows), as for M_F, a proportion of data is removed across the whole of the morphological matrix for fossil taxa (i.e. removing

cells).

We tested all parameters combinations resulting in 125 (5^3) matrices. Because some parameter combinations introduce a lot of missing (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 every taxa had at least 5% data across the whole matrix.

Building phylogenies

From the resulting matrices we generated two types of trees, the "best" tree that is 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 the 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 biologists have to work with.

Maximum Likelihood.— The "best" tree and the "missing-data" trees were inferred using RAxML v8.0.20 (Stamatakis 2014). For the molecular data, we used the GTR + Γ_4 model (Tavaré (1986); default GTRGAMMA in RAxML v8.0.20; Stamatakis (2014)) as a generalisation of the HKY + Γ_4 model (Hasegawa et al. 1985) for the molecular data. The GTR model can be seen as a generalisation of the HKY model (the 2 parameters from the HKY model are implicitly included in the 6 from GTR model - Stamatakis et al. (2008)). For the morphological data, we used the implemented Markov k state model (Lewis 2001)

which is a generalisation of the JC69 model (Jukes and Cantor 1969) with $k \geq 2$ assuming an equal state frequency and a unique overall substitution rate (μ) following a gamma distribution of the rate variation with four distinct categories ($\mathrm{M}k + \Gamma_4$; -K MK option in RAxML v8.0.20; Stamatakis (2014)). We used the fast bootstrap algorithm and performed 1000 bootstraps per tree inference to assess the topological support. The bootstrap algorithm used in RAxML is the Lazy Sub-tree Rearrangement (LSR) which consists in pruning one sub-tree from the tree and subsequently reinserting it to al neighbouring branches (Stamatakis et al. 2008). Sub-tree Pruning and Reinserting methods (SPR) have been demonstrated as being better than others (e.g. Nearest Neighbouring Interchange - NNI) in recovering good bootstrap values (Salamin et al. 2003).

Bayesian.— The "best" tree and the "missing-data" trees were inferred using MrBayes v3.2.2 (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 Markov k state model (Lewis 2001), with equal state frequency and a unique overall substitution rate (μ) with four distinct rates categories ($\mathrm{M}k + \Gamma_4$). 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 50×1^6 generations. 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). The effective sample size (ESS) was also checked on a random sub-sample of runs in each simulation to ensure that ESS $>> 200$ (Drummond et al. 2006). For each run, we removed 25% of the iterations as burn-in. We

used the following prior for each tree (see supplementaries):

1. the "true" trees topology as a starting tree (with a starting value for each branch length of 1),

2. an exponential prior on the shape of the gamma distribution of $\alpha=0.5$ for both partitions

3. and a transition/transversion ratio prior of 2 sampled from a strong beta distribution ($\beta(80,40)$).

We used these prior to speed up the Bayesian process. These prior biased the way the Bayesian process calculated the branch length by giving non-random starting points and boundaries for the parameters estimation process, however, in this study, we focused on the effect of missing data on the topology and not on the branch length. Even using these prior, it took 65 CPU years to build 51 sets of 125 Bayesian trees (8 core nodes 2.30GHz clock speed).

Comparing topologies

We compared the topology of the "missing-data" trees inferred from the matrices with missing data to the "best" tree to measure the effect of the three parameters M_L, M_F and M_C. Note that we only investigate differences in topology and not in branch length because the aim of this study is to look at the effect of missing data on the topology of trees inferred from total evidence type matrices. To compare the topology of the resulting trees, we used two metrics to assess number of conserved taxa and clades position using respectively the Triples (Dobson 1975) and the Robinson-Fould (Robinson and Foulds 1981) distance. We normalised the two metrics using the Normalised Tree Similarity index

(Bogdanowicz et al. 2012) to generalise our results for any n number of taxa. The three metrics are detailed below.

Triples distance ($T_{x,y}$) (Dobson 1975).— This metric measures the number of different sub-trees of three taxa between two given trees. Each triplet can be written as $I_{ijk}=(ijk)$. Where I_{ijk} is equal to 0 if the the two triplets (ijk) are the same in the two trees otherwise I_{ijk} is equal to 1. For any rooted binary tree there are only three possible combinations per triplets: $((j,k),i);$, $((i,k),j);$ and $((i,j),k);$ (Johnson and Soltis 1998). If the trees used are not fully binary, a fourth triplet combination is possible: $(i,j,k);$. One can calculate S_n, the triplet distance between two trees as:

$$S_n = \sum_{ijk} I_{ijk} \tag{1}$$

Where:

$$\sum_{ijk} = \binom{n}{4} = \frac{n!}{4!(n-4)!} \tag{2}$$

And where n is the number of taxa in both trees (modified from Critchlow et al. (1996)). If $S_n=0$, the trees are the same (i.e. no taxa as been displaced). When $S_n = \binom{n}{4}$, the trees are the most different possible (i.e. every taxa as been displaced). This metric therefore illustrates the amount of displayed taxa. It is less sensitive to the placement of individual taxa and to taxa of highly uncertain placement (e.g. fossil taxa) than the Robinson-Foulds metric (Critchlow et al. 1996; Johnson and Soltis 1998; Wiens 2003). Therefore we used this metric as a proxy to estimate the robustness of the tree to flying taxa (see supplementaries).

Robinson-Fould distance (RF) (Robinson and Foulds 1981).— This metric measures the number of shared clades among two trees and therefore illustrates the number of exactly conserved groups among the trees. The Robinson-Fould distance (also called path

difference) between two trees reflects the distance between the distributions of the tips among clades in the two trees (Robinson and Foulds 1981) and can be expressed as following:

$$RF_{x,y} = N_x + N_y - 2C_{x,y} \qquad (3)$$

Where $C_{x,y}$ is the number of clades in common in the two trees. The minimal value of C is 1 if the two trees have the same n taxa; the maximal value in $C=n-2$. This metric is more sensitive to taxa displacement than the Triples metric (i.e. if one taxa gets out of a clade, then the clades are no longer considered as similar - Critchlow et al. (1996); Johnson and Soltis (1998); Wiens (2003)). 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 (see supplementaries).

Normalised Tree Similarity NTS (Bogdanowicz et al. 2012).— For any tree with n taxa compared using a tree distance metric m, NTS_m represents the similarity score between the two trees given the expected distance between two random Yule trees with n taxa. Let $\bar{d}_{m,n}(rand)$ be 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:

$$NTS_{m,n}(x, y) = \frac{\bar{d}_{m,n}(rand) - d_{m,n}(x, y)}{\bar{d}_{m,n}(rand)} \qquad (4)$$

NTS ranges from 1 to -∞. For any m,n when $NTS=1$, the trees are the same; when $NTS=0$ the trees are not more different than expected by chance; when $NTS<0$, the trees are more different than expected by chance (see supplementaries).

We compared the "missing-data" tree to the "True" and the "Best" tree for each

chain. For the Maximum Likelihood trees we performed pairwise comparisons between the "True" and "Best" tree and the "missing-data" tree (see Table 1) for both the Robinson-Fould and the Triples metric. We calculated the difference between the trees using the metrics described above by using the TreeCmp java script (Bogdanowicz et al. 2012). For each metric, we normalized the value using the Normalised Tree Similarity scaled with the mean value of 1000 pairwise random tree comparisons for the same metric and the same number of taxa $n=51$ (see supplementaries). We ran the comparison for every "missing-data" tree in each chain resulting in 51 comparisons for every "missing-data" trees. We calculated the mode and the 50% and 95% confidence intervals from the resulting distribution using the hdrcde R package (v3.1 with contributions from Jochen Einbeck and Wand (2013)).

Bayesian tree inference allows to account for the statistical uncertainty of a phylogenetic tree by not using an optimal criterion (c.f. Maximum Likelihood) and gives a tree posterior distribution (c.f. the likeliest tree in ML). This method has the clear advantage of better dealing with error and uncertainty but it's output (a distribution rather than a single tree) is less practical to use for any further study (but see Healy et al. (2014)). To avoid this problem, people traditionally use the a consensus tree build on a majority rule in order to summarise a Bayesian tree posterior distribution. This results into a single tree containing both topological and branch length information as well as support information (i.e. posterior probabilities, e.g. Ronquist et al. (2012b)). However, using a Bayesian consensus tree has limitation especially if the resulting consensus tree is not well supported or resolved. Because the metrics used in this study are used to measure variation in topology between two trees (i.e. taxa placement or clade position), comparing two Bayesian consensus trees is not optimal in picking topological difference signal. For example, if one of the trees is not resolved at all (i.e. a star tree), comparing it to any other tree (even an other star tree) will not give useful results. Therefore we used the

entire Bayesian trees posterior distribution in order to perform the tree comparisons (i.e. the "Best" Bayesian tree posterior distribution vs. the "missing-data" Bayesian tree posterior distribution).

Random Pairwise Bayesian Tree Comparison (RPBTC).— We compared the Bayesian posterior distribution using a Random Pairwise Bayesian Tree Comparison (RPBTC) method. This method consists in comparing a series of randomly selected pairs of trees from two posterior distributions (on from each distribution) and use the mode to summarize the resulting distribution as a proxy of the difference between the two trees in a Bayesian framework.

$$RPBTC_{X,Y} = Mo[(]d_{m,n}(X_{i1}, Y_{j1}), d_{m,n}(X_{i2}, Y_{j2}), d_{m,n}(X_{i3}, Y_{j3}), ..., d_{m,n}(X_{ik}, Y_{jk})] \quad (5)$$

Where X and Y are two Bayesian tree posterior distributions; $d_{m,n}$ is the pairwise difference for any metric m and n taxa between X_i and Y_j which are two single trees randomly sampled respectively from X and Y; Mo is the mode and $Mo[(d_{m,n}(X_{i1}, Y_{j1}), d_{m,n}(X_{i2}, Y_{j2}), d_{m,n}(X_{i3}, Y_{j3}), ..., d_{m,n}(X_{ik}, Y_{jk})]$ is the mode of the pairwise difference repeated k times.

We used the mode to summarize the distributions because it is the value that is the most represented in the distribution and reflects what a consensus tree represents (the topology the most represented in the posterior distribution). Also, the mode is a value present in the distribution (which, depending on the metric, can be a series of discrete values) in contrast to the other distribution summary metrics, such as the mean or the median, which can take values that are not actually present in the distribution. For example, if comparing two posterior tree distribution, we obtain difference values of 10, 10, 10, 5, 5 and 1, using the mean (6.8) or the median (7.5) gives a values that actually doesn't exist.

In this study, we calculated RPBTC for both metrics (Robinson-Fould and Triples distance) for each pairs of Bayesian tree distributions (i.e. the "Best" Bayesian tree posterior distribution vs. the "missing-data" Bayesian tree posterior distribution) with 1000 random pairwise comparisons (k). Because the RPBTC uses random pairwise comparisons, there is a chance of comparing always trees that have the biggest or the smallest difference between two distributions, either deflating or inflating the RPBTC difference value. However, using 1000 random pairwise comparisons makes the difference stable (i.e. if repeated independently, no difference is detected - see supplementaries).

We calculated the NTS for the "Best" Bayesian tree vs. "missing-data" Bayesian tree comparison for both Robinson-Fould and Triples metric using the mode of the RPBTC for each chain resulting in a distribution of 51 modes per comparison. We then compared the NTS of the Bayesian trees to the NTS of the ML trees for both metrics for each comparison. This resulted in a distribution of 51 NTS values for the Bayesian trees and 51 NTS values for the ML trees for each tree comparison for which we calculated the mode and the 50% and 95% confidence intervals (see supplementaries - Fig. 3).

Statistical difference between tree topologies.— To assess the difference between the tree topologies, whether they were inferred in a Bayesian or a ML framework, we performed a group comparison test for each set of configurations (e.g. effect of M_L in Bayesian inference, combined effect of M_F and M_C in ML inference, etc...). We used the pairwise tree comparisons as factors (i.e. the "Best" tree vs. one of each of the 125 "missing-data" trees) and the values of the 51 replicates from the pairwise comparison as a response variable. The 51 replicates values were either the replicates of the 51 pairwise tree comparisons in ML or the modes of the 51 RPBTC posterior distributions in Bayesian.
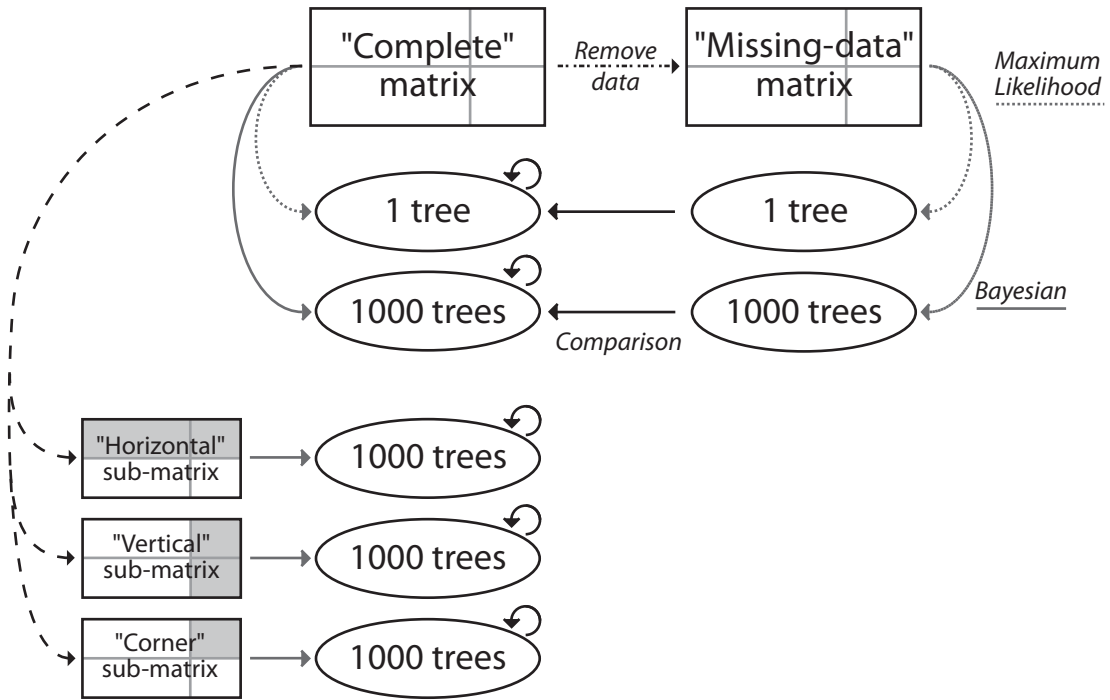
Figure 3: Tree comparison protocol. From each of the 125 parameters combination (M_L, M_F and M_C), we inferred a tree in Baysian or Maximum likelihood framework. We then compared this tree with or without fossil/living taxa to the "best/true" tree. In the Bayesian framework, 1000 trees from the posterior distribution where used for each comparison.

When we found a significant difference between the groups, we performed a pairwise comparison test to analyse which pairs of groups differ from the others. Depending on whether the data was normal with variance homoscedasticity or not (controlled by performing respectively a shapiro.test and a a bartlett.test in R{stats}), we used respectively parametric or non-parametric tests for the group comparison and the pairwise comparisons as suggested by Ruxton and Beauchamp (2008) (Table 2).

Effect of missing molecular characters for fossil taxa

To assess the effect of missing molecular characters for fossil taxa in a Bayesian framework, we split the "complete" matrix in sub-matrices containing no missing data at all (see Fig. 2):

1. A first containing both molecular and morphological data for living taxa only (hereafter called the "horizontal" sub-matrix, Table 1);

2. A second one containing morphological data for both living and fossil taxa (hereafter called the "vertical" sub-matrix, Table 1);

3. A third one containing morphological data for living taxa only (hereafter called the "corner" sub-matrix, Table 1);

We reran the Bayesian tree inference on the different sub-matrices with no missing data in the same way as described above. We then compared the resulting tree posterior distribution to itself (in the same way described above) to assess the ability to recover topology for each simulation when no missing data was involved in the phylogenetic inference process.

Empirical data

We also compared the results obtained from simulated data by using Ronquist et al. (2012a) empirical data. The matrix contains 67 living species plus one outgroup and 45 fossil species of Hymenopteras with 5097 molecular characters and 354 morphological characters. From the 68 living species used in the matrix, only 66 had molecular data, we therefore treated these 66 taxa as "living" taxa and all the other 47 as "fossil" taxa. We treated the matrix in the exact same way as described in step 2 and 3 resulting in 125 matrices with various amount of missing data and the same number of Maximum Likelihood and Bayesian trees. We used the same settings as for the simulated data in the Maximum Likelihood framework. For the Bayesian inferences however, we didnt used any priors except that we provided a starting tree with the topology of the 68 living species (topology with the highest posterior probability from non-clock analysis - Ronquist et al. (2012a)). Contrary to Ronquist et al. (2012a) analysis, we didnt performed any clock analysis since we were only interested in the topology of the inferred tree and not the branch length.

RESULTS

Building the trees

When generating the matrices using seqgen and rTraitDisc algorithms with low evolutionary rate parameter (α=0.5) we successfully generated evolutionary histories. In Maximum Likelihood, support values ranged from 0 to 100 with a median of 38 ($1^{s}t$ quartile = 14, $3^{r}d$ quartile = 72). Using Bayesian inference, all the chains converged with a significant effective sample size (ASSD < 0.01 and ESS >> 200 for each tree). Good enough phylogenetic signal.

Effect of the missing data parameters

Effect of M_L.— In Maximum Likelihood framework, the topological recovery quickly drops with the amount of missing data when using both metrics (see Table 5 - Fig. **??**). However, this negative effect of missing data seems to stabilize beyond 25% of missing data (no significant difference for trees with respectively 25%, 50% and 75% missing living taxa). With 10% of missing living taxa, there is a different topological recovery depending on the used metric. The Robinson-Fould metric shows a lower Tree similarity metric than the Triples metric meaning that individual taxa position are generally more conserved (i.e. low number of flying taxa) but that the full clades are less conserved (see Table 3).

Howerver, topological recovery was bad in Bayesian, the tree similarity score was only slightly above 0 which means that the mode of the RPBTC was only slightly different than expected by chance (see Table **??**). However, it seems that there is a small difference depending on the metric used. For the Triples metric there is no significant effect of the M_L parameter, in other words, if living taxa are removed, there is no effect on the placement of "flying" taxa. For the Robinson-Fould metric, there is a significant effect of the M_L leading small decrease in topological recovery when M_L increases. In other words, if living taxa are removed, clades are less conserved (see Table 8). However, in both cases, the effect of the parameter M_L is still minimal compared to trees generated in a ML framework since the NTS score was only slightly higher than expected by chance (see Table 5 - Fig. **??**).

Effect of M_F.— In Maximum Likelihood framework, the effect of missing data in the fossil record appeared to be constant and led to a constant decrease in topological recovery when the missing data increased (see Table 5 - Fig. **??**). The difference between the two metrics used is more important than for the M_L parameter: clades are less conserved than individual taxa placement as the amount of missing data increases. Interestingly, 10% or 25% missing data for the M_F parameter does not seems to affect the apparition of unstable

taxa (Triples NTS modes of respectively 0.97 and 0.93 - Table 3).

Similarly than for the effect of M_L, Bayesian tree recovery was only slightly better than expected by chance. And in the same way as described above, there seems to be a small difference depending on the metric used. There is no significant effect of the M_F parameter when using the Triplet metric but the effect is significant when using the Robinson-Fould metric (see Table 11). However, as mentioned previously the effect of the parameter M_F is still minimal in a Bayesian framework (see Table 5 - Fig. **??**).

Effect of M_C.— The number of missing morphological characters (M_C), however, seems to be more affecting topological recovery than M_L and M_F (see Table 5 - Fig. **??**). The Robinson-Fould metric shows a rapid drop in tree recovery from 10% of missing data (Robinson-Fould NTS mode $<$ to 0.5 from 10% missing data - Table 3). This decrease is however slowlier when using the Triples metric with still a good topological recovery at 10% (Triples NTS modes $=$ 0.95 - Table 3).

The effect of M_C in Bayesian framework is similar as the one described for M_L and M_F: tree recovery is only slightly better than expected by chance and only the Robinson-Fould metric shows a significant effect of the M_C parameter on tree recovery (see Table 8 and 5 - Fig. **??**).

The ability of recovering the "best" tree's topology in maximum likelihood method is function of the amount of data missing. The parameters M_C and M_L have more influence than M_F on decrease in topological recovery. The decrease in clade conservation (low Robinson-Fould NTS score) is faster than the increase in flying taxa (low Triples NTS score). However, in Bayesian, topology is badly (if not at all) recovered disregarding the amount of data missing.

Discussion

Building the trees

Hard to simulate morpho characters but see Pagel, April Wright, etc...

Comparing topologies

Hard to compare Bayesian tree posterior distributions topologies but gives a good idea.

ML vs Bayesian

Missing data is not a problem as long as random. However, in TEM, n.fossils*n.molecular data is missing and not random. Non optimal criterion approach (i.e. Bayesian) worse because of the number of considered "less"-likely trees (flat maximum likelihood landscapes). However, ML still good performances and can be used to fix topology.

Effect of missing data

Parameter blabla seems to be influencing the topology the most. But in practice, M_L is the only one realistically solvable. Go to museums!

What about the "best" trees with fairly bad bootstrap support values in ML? This is probably due to the low number of characters used in the simulation. The number of morphological characters (10^2) is in the same order of magnitudes of empirical studies (cite) but decreases unrealistically with the M_C parameter. The number of molecular characters (10^3) is at least two orders of magnitude lower than in empirical studies that can reach 10^6 characters (cite).

Two different matrices from the "True" tree gives two different phylogenetic interpretations (i.e. seqGen and ape performs badly at generating a data matrix from a tree)? That seems to not really be a problem since we are interested in recovering topology with missing data, not assessing if phylogenetic matrices generation software performs correctly. Also, the "noise" that might be created by this potential "bad" phylogenetic matrices generation software is probably diminished by the number of replicates (i.e. 50 chains).

Different phylogenetic signal between the morphological and the molecular matrix? That is not really a problem since the phenotypic evolution is not necessary correlated to the genotypic one. Also it is important to note that both information are not exactly the same. The main difference between molecular and morphological characters is that, because of historical and practical reasons, a morphological character matrix do no contains autapomorphies and invariant characters which are usually more common than synampomorphies in molecular character matrices (cite).

What about homoplasy in the morphological matrix? Here, we made the assumption that theoretically, morphological characters are randomly distributed on an organism however it seems clear that empirical morphological data does not act randomly If one states that morphological characters accumulate through time in the same way as the majority of the molecular characters (neutral theory, cite) then, homoplasic characters are expected to appear randomly through time. Therefore, homoplasy is expected to be more important (by chance) in bigger morphological matrices (Davalos et al. 2014). After a certain amount of morphological characters, adding new ones adds homoplasy (Wagner 2000).

Difference between Robinson-Fould and Triples? Robison-Fould is a conservative tree topology metric, it is more sensible to single taxon displacement because it will count clades as similar only if they are composed of the same number of taxa with the same

topologies. When getting closer to the root of the tree, displacement of single taxon makes the clades not being exactly identical any more even if the clade still contains all the other taxa in both trees. On the other hand, the Triples method is measuring the position of each taxon towards to other reference taxa. It will penalise only trees where taxa get removed furthest from their original clade. Looking at our problematic (how does missing data influence topological recovery in trees containing both living and fossil taxa) we are more interested the placement of taxa (i.e. where does the fossil branch) than the exact conservation of clades.

What's the crucial factor? Regarding the fact that we are more interested in taxa placement than in clade conservation (Triples vs. Robinson-Fould), the parameter that affects the more the decrease in topological recovery if the number of missing living taxa in the morphological part of the matrix (M_L). This makes sense since the living taxa are bearing both the information to build the tree backbone (the molecular data) and the information used for branching the fossil on this backbone (the morphological information). Therefore, we advocate that an effort should be made in collecting all the available morphological information for living taxa in order to properly run Total Evidence analysis.

Conclusion

A Bayesian approach fails to recover accurate topology in a total evidence approach framework, whatever the amount of missing data. We think this failure is due to the intrinsic structure of the Total Evidence matrices that includes a vast amount of not randomly distributed missing data (i.e. the molecular part of the matrix for the fossil taxa). This missing data is equal to the number of fossil taxa × the number of molecular characters leading to a vast amount of likely trees to be sampled in the Bayesian mcmc tree building. However, one can use a Maximum Likelihood approach to obtain the likeliest

topology and to use this one as a fixed. If so, an effort should be made on collecting morphological data from living taxa in order to efficiently improve the quality of the trees.

Acknowledgements

*

References

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

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.

Cooper, N. and A. Purvis. 2009. What factors shape rates of phenotypic evolution? a comparative study of cranial morphology of four mammalian clades. Journal of evolutionary biology 22:1024–1035.

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 Biol 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.

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.

Healy, K., T. Guillerme, S. Finlay, A. Kane, S. Kelly, M. Deirdre, D. Kelly, I. Donohue, A. Jackson, and N. Cooper. 2014. Ecology and mode-of-life explain lifespan variation in birds and mammals. Proceedings. Biological sciences / The Royal Society 281.

Heath, T., J. Huelsenbeck, and T. Stadler. 2013. The fossilized Birth-Death process: A coherent model of fossil calibration for divergence time estimation .

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.

Johnson, L. and D. Soltis. 1998. Assessing Congruence: Empirical Examples from Molecular Data chap. 11, Pages 297–348. Springer US.

Jukes, T. H. and C. R. Cantor. 1969. Evolution of protein molecules vol. III Pages 21–132. Academic Press.

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

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

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

Meredith, R., J. Janeka, 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 (New York, N.Y.) 334:521–524.

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

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; international journal of organic evolution 65:661–672.

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: analyses of phylogenetics and evolution in r language. Bioinformatics (Oxford, England) 20:289–290.

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. Comput Appl Biosci 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. Syst Biol 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. J Mol Evol 61:171–80.

Ruxton, G. D. and G. Beauchamp. 2008. Time for some a priori thinking about post hoc testing. Behavioral Ecology 19.

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.

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. Trans N Y Acad Sci 8:45–60.

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.

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 .

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

Tavaré, S. 1986. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences 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. J Syst Evol 46:307–314.

with contributions from Jochen Einbeck, R. J. H. and M. Wand. 2013. hdrcde: Highest
density regions and conditional density estimation. R package version 3.1.

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

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

Supplementaries

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 taxomomic classess (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 sampeled 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
ecountered on characters with more than 3 states and the difficulties of simulate biologicaly
sensible ordered characters.

Tree Building Software settings
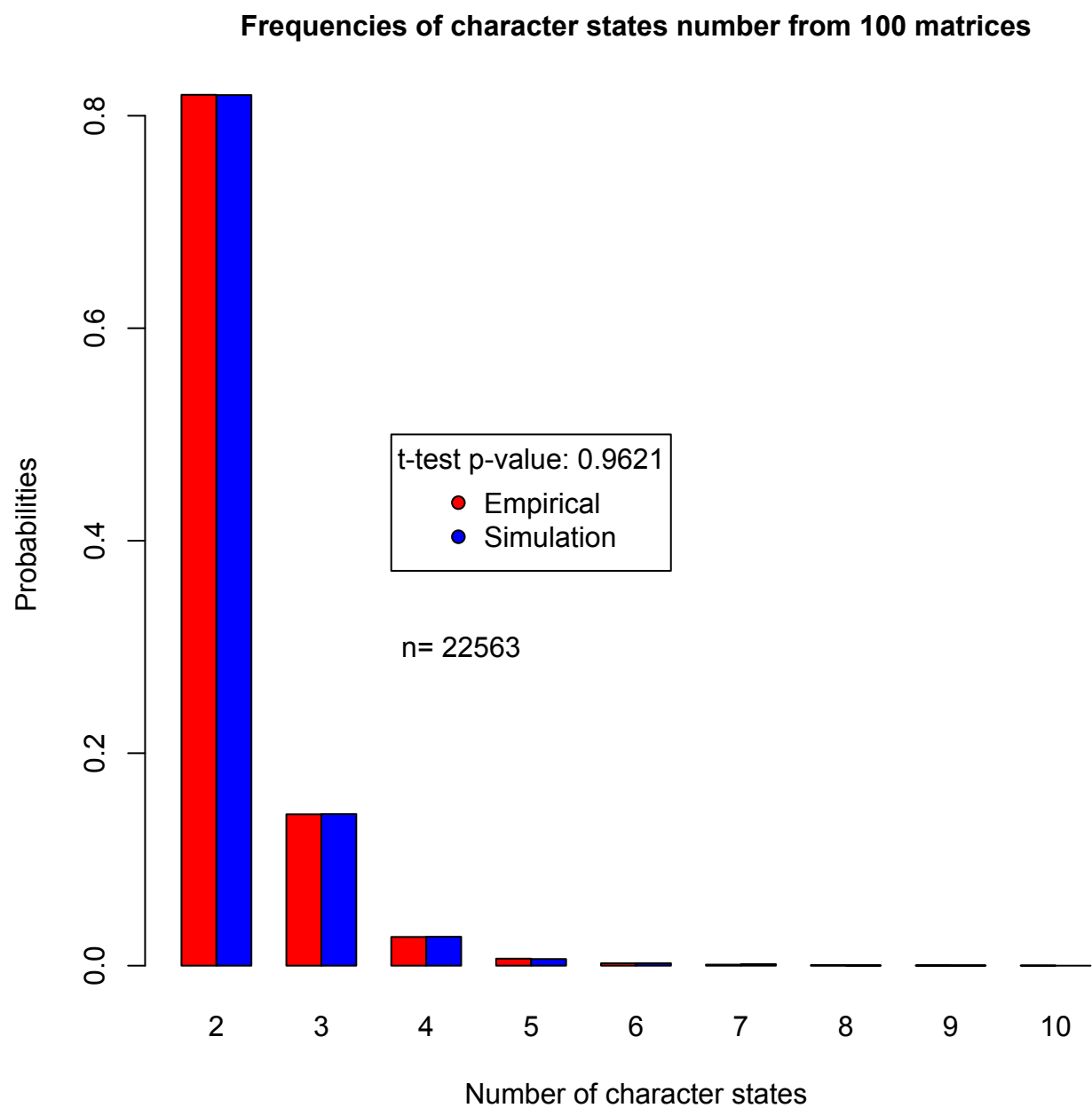
Figure 4: Character states distribution in empirical matrices. Characters states number distribution extracted from 100 random morphological matrices downloaded from RreeBase.

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

Model:

Molecular data:

GTR + Γ_4 (-m GTRGAMMA)

Morphological data:

Mk + Γ_4 (-K MK)

Support:

Rapid Boostrap algorithm (LSR), 1000 replicates

Bayesian - MrBayes v3.0.2 (Ronquist et al. 2012b).—

Priors:

Molecular data:

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 + Γ_4

Morphological data: Mk + Γ_4

MCMC:

2 runs

4 chains per run

generations ¡ 50×1^6

sample frequency $= 1050 \times 1^3$

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

ASDS < 0.01

ESS >> 200

Burnin = 25%

Triplets metric details ($T_{x,y}$)

Each triplet can be written as $I_{ijk}=(ijk))$. Where I_{ijk} is equal to 0 if the the two triplets (ijk) are the same in the two trees otherwise I_{ijk} is equal to 1. For any rooted tree there are only four possible combinations per triplets: $((j,k),i);, ((i,k),j);$ and $((i,j),k);$ and $(i,j,k);$ (Johnson and Soltis 1998). One can calculate S_n, the triplet distance between two trees as:

$$S_n = \sum_{ijk} I_{ijk} \tag{6}$$

Where:

$$\sum_{ijk} = \binom{n}{4} = \frac{n!}{4!(n-4)!} \tag{7}$$

And where n is the number of taxa in both trees (modified from Critchlow et al. (1996)). When all triplets across the two trees are the same, S_n is equal to 0 and when all the triplets are different S_n is equal to $\binom{n}{4}$. 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:

$$P(I_{ijk} = 0) = \frac{1}{4} \tag{8}$$

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

$$P(S_n = 0) = \sum_{ijk} P_{I_{ijk}=0} \tag{9}$$

$$P(S_n = 0) = \frac{n!}{4(3!(n-3)!)} \qquad (10)$$

And in the same way:

$$P(S_n = 1) = \frac{3n!}{4(3!(n-3)!)} \qquad (11)$$

RF metric details

The RF distance (or path difference) between two trees reflects the distance between the distributions of the tips among clades in the two trees (Robinson and Foulds 1981) and can be expressed as following:

$$RF_{x,y} = N_x + N_y - 2C_{x,y} \qquad (12)$$

Where $C_{x,y}$ is the number of clades in common in the two trees. 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 n taxa is:

$$RF_{min} = 1 + 1 - 2C_{x,y} \qquad (13)$$

And:

$$RF_{max} = 2(n-2) - 2 \qquad (14)$$

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

$$RF.scaled_{x,y} = \frac{RF_{x,y} - RF_{max}}{RF_{max}} \qquad (15)$$

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.

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).

Random Pairwise Bayesian Tree Comparison (RPBTC).— We assessed the power of the Random Pairwise Bayesian Tree Comparison (RPBTC) method by comparing 1000 random trees from a posterior distribution trees set to another 1000 random trees from the same posterior distribution trees set. We repeated this 100 times independently using the same posterior distribution trees set each time resulting in 100 replicates of the same posterior distribution trees set compared 1000 times. We used an anova to test if there was no significant difference between the replicates so that the RBTC can be replicated. We applied this protocol on a poorly resolved tree (Low Score), a resolved tree with low support value (Medium Score) and a resolved tree with high support values (High Score). Results are available in table .

Codes

All codes are available at: *https : //github.com/TGuillerme/Total_Evidence_Method − Missing_data/tree/master/Functions*

Full results

Bootstraps distribution

Bootstraps distribution
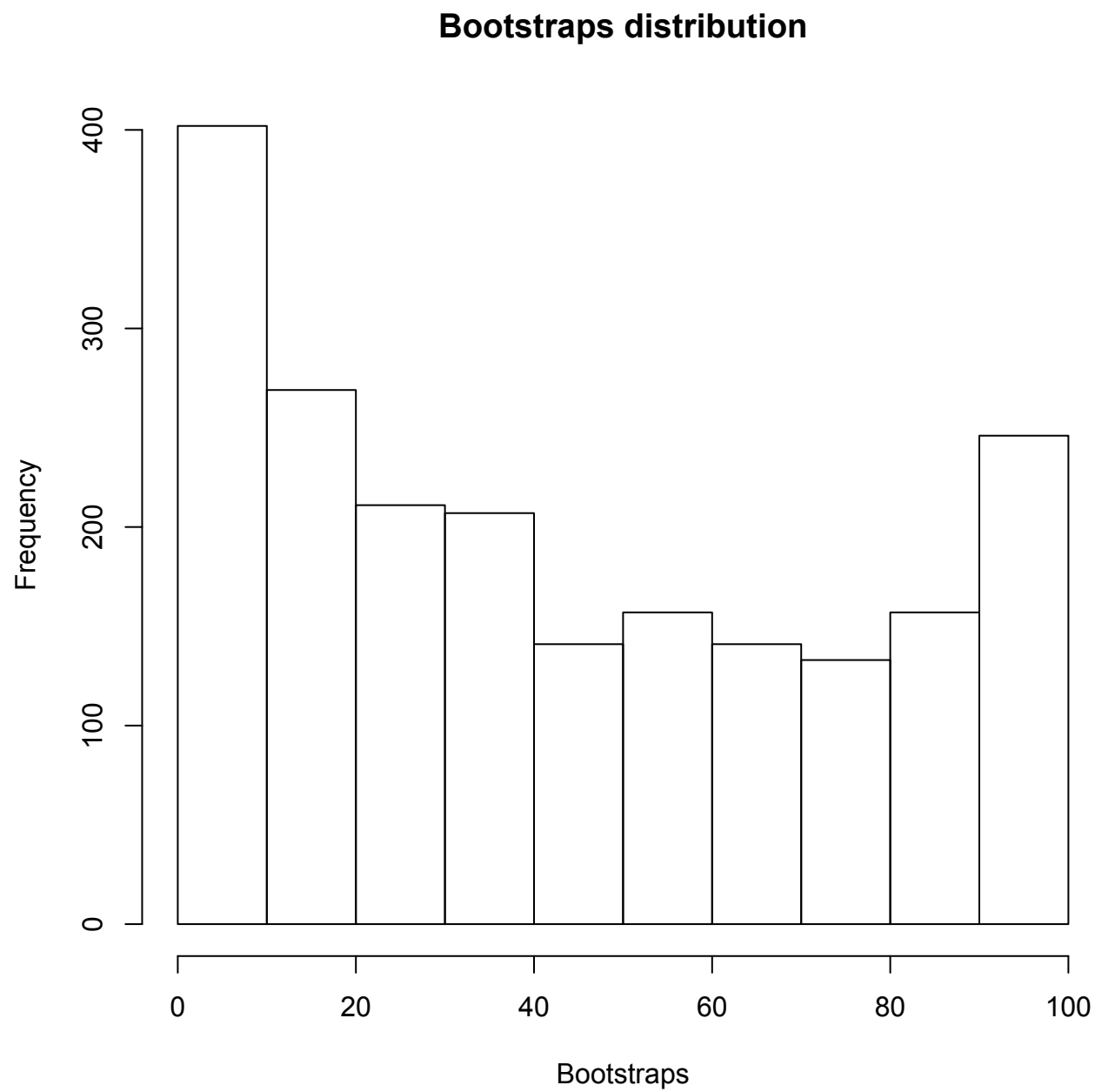
Figure 5: Bootstraps distribution across the "best" trees.

Table 1: Glossary

Term	Definition
living taxa	taxa with both molecular and morphological data available
fossil taxa	taxa with only morphological data available
"complete" matrix	matrix with no missing data except for the molecular part of the fossil taxa
"missing-data" matrix	matrix with various amount of missing data
M_L	missing living taxa in the morphological part of the matrix
M_F	missing data for the fossil taxa in morphological part of the matrix
M_C	missing morphological characters for both living and fossil taxa
"true" tree	tree used to simulate the matrix
"best" tree	tree inferred from the "complete" matrix
"missing-data" tree	tree inferred the "missing-data" matrices
RPBTC	Random pairwise Bayesian tree comparison
"horizontal" sub-matrix	sub-matrix with morphological and molecular characters for living taxa
"vertical" sub-matrix	sub-matrix with morphological data for both living and fossil taxa
"corner" sub-matrix	sub-matrix with morphological data for living taxa
"missing-data" sub-matrix	molecular part of the matrix for the fossil taxa (no data)

Table 2: Statistical tests

	test	statistics	code{package}
Groups comparisons	ANOVA	parametric	anova{stats}
	Kruskal-Wallis	non-parametric	kruskal.test{stats}
Pairwise comparisons	Tukey HSD	parametric	TukeyHSD{stats}
	Nemenyi-Damico-Wolfe-Dunn	non-parametric	kruskalmc{pgirmess}

Table 3: Tree similarity values per parameter in ML framework

Parameter	amount of missing data	metric	mode	50%CI lower	upper	95%CI lower	upper
M_L	0%	Robinson-Fould	1	NA	NA	NA	NA
		Triples	1	NA	NA	NA	NA
	10%	Robinson-Fould	0.671	0.569	0.789	0.468	1.036
		Triples	0.942	0.868	0.999	0.612	1.063
	25%	Robinson-Fould	0.516	0.440	0.630	0.266	0.892
		Triples	0.559	0.708	0.810	0.291	0.999
	50%	Robinson-Fould	0.432	0.364	0.521	0.231	0.642
		Triples	0.514	0.419	0.599	0.293	0.910
	75%	Robinson-Fould	0.372	0.323	0.434	0.203	0.577
		Triples	0.444	0.338	0.532	0.192	0.855
M_L	0%	Robinson-Fould	1	NA	NA	NA	NA
		Triples	1	NA	NA	NA	NA
	10%	Robinson-Fould	0.907	0.766	1.000	0.407	1.047
		Triples	0.976	0.944	1.003	0.715	1.104
	25%	Robinson-Fould	0.572	0.488	0.692	0.325	0.936
		Triples	0.937	0.820	0.993	0.539	1.086
	50%	Robinson-Fould	0.433	0.302	0.476	0.170	0.668
		Triples	0.732	0.594	0.873	0.341	1.025
	75%	Robinson-Fould	0.153	0.115	0.200	0.068	0.323
		Triples	0.458	0.390	0.611	0.290	0.856
M_L	0%	Robinson-Fould	1	NA	NA	NA	NA
		Triples	1	NA	NA	NA	NA
	10%	Robinson-Fould	0.473	0.384	0.835	0.356	0.958
		Triples	0.959	0.869	1.026	0.572	1.067
	25%	Robinson-Fould	0.454	0.357	0.630	0.242	0.916
		Triples	0.907	0.751	0.997	0.446	1.070
	50%	Robinson-Fould	0.424	0.375	0.466	0.249	0.828
		Triples	0.683	0.571	0.822	0.277	1.008
	75%	Robinson-Fould	0.230	0.177	0.302	0.108	0.567
		Triples	0.451	0.373	0.572	0.224	0.791

Table 4: Tree similarity values per parameter in Bayesian framework

Parameter	amount of missing data	mode		50%CI lower	50%CI upper	95%CI lower	95%CI upper
M_L	0%	Robinson-Fould	0.056	0.035	0.063	0.031	0.076
		Triples	0.041	0.008	0.054	-0.039	0.106
	10%	Robinson-Fould	0.035	0.035	0.035	0.028	0.061
		Triples	0.033	0.003	0.049	-0.049	0.117
	25%	Robinson-Fould	0.035	0.035	0.035	0.015	0.061
		Triples	0.035	0.002	0.049	-0.041	0.096
	50%	Robinson-Fould	0.035	0.035	0.035	0.015	0.056
		Triples	0.028	0.003	0.052	-0.038	0.092
	75%	Robinson-Fould	0.035	0.035	0.035	0.010	0.043
		Triples	-0.001	-0.018	0.040	-0.060	0.106
M_L	0%	Robinson-Fould	0.056	0.035	0.063	0.031	0.076
		Triples	0.041	0.008	0.054	-0.039	0.106
	10%	Robinson-Fould	0.035	0.031	0.056	0.027	0.076
		Triples	0.026	0.002	0.050	-0.046	0.105
	25%	Robinson-Fould	0.035	0.035	0.035	0.0153	0.068
		Triples	0.038	0.002	0.050	-0.046	0.105
	50%	Robinson-Fould	0.035	0.035	0.035	0.015	0.058
		Triples	0.014	-0.011	0.039	-0.038	0.124
	75%	Robinson-Fould	0.035	0.035	0.035	0.015	0.041
		Triples	0.029	0.019	0.042	-0.030	0.089
M_L	0%	Robinson-Fould	0.056	0.035	0.063	0.031	0.076
		Triples	0.041	0.008	0.054	-0.039	0.106
	10%	Robinson-Fould	0.056	0.035	0.059	0.0281	0.076
		Triples	0.030	0.011	0.048	-0.024	0.094
	25%	Robinson-Fould	0.035	0.035	0.035	0.023	0.076
		Triples	0.020	-0.001	0.037	-0.060	0.109
	50%	Robinson-Fould	0.035	0.035	0.035	0.030	0.059
		Triples	0.010	-0.005	0.045	-0.036	0.104
	75%	Robinson-Fould	0.035	0.035	0.035	0.011	0.043
		Triples	0.024	0.001	0.042	-0.037	0.104

Table 5: Group difference per configuration set

set	metric	framework	parametric	$stats^1$	df	p.value
M_L	Robinson-Fould	ML	no	151.438	4	**0.00**
		Bayesian	no	65.62	4	**0.00**
	Triples	ML	no	142.8329	4	**0.00**
		Bayesian	no	4.56	4	0.34
M_F	Robinson-Fould	ML	no	179.9238	4	**0.00**
		Bayesian	no	85.61	4	**0.00**
	Triples	ML	no	150.059	4	**0.00**
		Bayesian	no	3.01	4	0.56
M_C	Robinson-Fould	ML	no	151.3818	4	**0.00**
		Bayesian	no	64.23	4	**0.00**
	Triples	ML	no	138.374	4	**0.00**
		Bayesian	no	26.13	24	0.35
$M_L + M_F$	Robinson-Fould	ML	no	734.9213	24	**0.00**
		Bayesian	no	317.97	24	**0.00**
	Triples	ML	no	531.5558	24	**0.00**
		Bayesian	no	3.01	24	0.56
$M_L + M_C$	Robinson-Fould	ML	no	542.3233	24	**0.00**
		Bayesian	no	290.10	24	**0.00**
	Triples	ML	no	437.7413	24	**0.00**
		Bayesian	no	22.19	24	0.57
$M_F + M_C$	Robinson-Fould	ML	no	812.206	24	**0.00**
		Bayesian	no	385.96	24	**0.00**
	Triples	ML	no	528.9215	24	**0.00**
		Bayesian	no	20.23	24	0.68
$M_L + M_F + M_C$	Robinson-Fould	ML	no	3603.808	124	**0.00**
		Bayesian	no	1167.40	124	**0.00**
	Triples	ML	no	2047.923	124	**0.00**
		Bayesian	no	115.86	124	0.69

[1] F-value for parametric tests and Kruskal Wallis chi^2 for non parametric tests.

Table 6: M_L non-parametric pairwise difference for Robinson-Fould metric in ML framework

	$M_L00\%$	$M_L10\%$	$M_L25\%$	$M_L50\%$	$M_L75\%$
$M_L00\%$	-	**TRUE**	**TRUE**	**TRUE**	**TRUE**
$M_L10\%$		-	**TRUE**	**TRUE**	**TRUE**
$M_L25\%$			-	FALSE	**TRUE**
$M_L50\%$				-	FALSE
$M_L75\%$					-

pairwise comparison	obs.dif	critical.dif	difference
$M_L00\%$-$M_L10\%$	56.53	37.66	TRUE
$M_L00\%$-$M_L25\%$	96.85	37.66	TRUE
$M_L00\%$-$M_L50\%$	127.03	37.66	TRUE
$M_L00\%$-$M_L75\%$	144.58	37.66	TRUE
$M_L10\%$-$M_L25\%$	40.31	37.66	TRUE
$M_L10\%$-$M_L50\%$	70.50	37.66	TRUE
$M_L10\%$-$M_L75\%$	88.05	37.66	TRUE
$M_L25\%$-$M_L50\%$	30.19	37.66	FALSE
$M_L25\%$-$M_L75\%$	47.73	37.66	TRUE
$M_L50\%$-$M_L75\%$	17.55	37.66	FALSE

Table 7: M_L non-parametric pairwise difference for Triples metric in ML framework

	$M_L00\%$	$M_L10\%$	$M_L25\%$	$M_L50\%$	$M_L75\%$
$M_L00\%$	-	**TRUE**	**TRUE**	**TRUE**	**TRUE**
$M_L10\%$		-	**TRUE**	**TRUE**	**TRUE**
$M_L25\%$			-	FALSE	FALSE
$M_L50\%$				-	FALSE
$M_L75\%$					-

pairwise comparison	obs.dif	critical.dif	difference
$M_L00\%$-$M_L10\%$	57.98	37.66	TRUE
$M_L00\%$-$M_L25\%$	105.83	37.66	TRUE
$M_L00\%$-$M_L50\%$	120.86	37.66	TRUE
$M_L00\%$-$M_L75\%$	140.34	37.66	TRUE
$M_L10\%$-$M_L25\%$	47.85	37.66	TRUE
$M_L10\%$-$M_L50\%$	62.88	37.66	TRUE
$M_L10\%$-$M_L75\%$	82.36	37.66	TRUE
$M_L25\%$-$M_L50\%$	15.03	37.66	FALSE
$M_L25\%$-$M_L75\%$	34.51	37.66	FALSE
$M_L50\%$-$M_L75\%$	19.48	37.66	FALSE

Table 8: M_L non-parametric pairwise difference for Robinson-Fould metric in Bayesian framework

	$M_L00\%$	$M_L10\%$	$M_L25\%$	$M_L50\%$	$M_L75\%$
$M_L00\%$	-	FALSE	**TRUE**	**TRUE**	**TRUE**
$M_L10\%$		-	FALSE	**TRUE**	**TRUE**
$M_L25\%$			-	FALSE	**TRUE**
$M_L50\%$				-	FALSE
$M_L75\%$					-

pairwise comparison	obs.dif	critical.dif	difference
$M_L00\%$-$M_L10\%$	17.05	41.00	FALSE
$M_L00\%$-$M_L25\%$	43.79	41.00	TRUE
$M_L00\%$-$M_L50\%$	77.32	41.00	TRUE
$M_L00\%$-$M_L75\%$	103.99	41.00	TRUE
$M_L10\%$-$M_L25\%$	26.75	41.00	FALSE
$M_L10\%$-$M_L50\%$	60.27	41.00	TRUE
$M_L10\%$-$M_L75\%$	86.94	41.00	TRUE
$M_L25\%$-$M_L50\%$	33.53	41.00	FALSE
$M_L25\%$-$M_L75\%$	60.20	41.00	TRUE
$M_L50\%$-$M_L75\%$	26.67	41.00	FALSE

Table 9: M_F non-parametric pairwise difference for Robinson-Fould metric in ML framework

	$M_F00\%$	$M_F10\%$	$M_F25\%$	$M_F50\%$	$M_F75\%$
$M_L00\%$	-	**TRUE**	**TRUE**	**TRUE**	**TRUE**
$M_L10\%$		-	FALSE	**TRUE**	**TRUE**
$M_L25\%$			-	**TRUE**	**TRUE**
$M_L50\%$				-	**TRUE**
$M_L75\%$					-

pairwise comparison	obs.dif	critical.dif	difference
$M_F00\%$-$M_F10\%$	59.27	37.66	TRUE
$M_F00\%$-$M_F25\%$	79.72	37.66	TRUE
$M_F00\%$-$M_F50\%$	120.70	37.66	TRUE
$M_F00\%$-$M_F75\%$	167.81	37.66	TRUE
$M_F10\%$-$M_F25\%$	20.45	37.66	FALSE
$M_F10\%$-$M_F50\%$	61.43	37.66	TRUE
$M_F10\%$-$M_F75\%$	108.55	37.66	TRUE
$M_F25\%$-$M_F50\%$	40.98	37.66	TRUE
$M_F25\%$-$M_F75\%$	88.09	37.66	TRUE
$M_F50\%$-$M_F75\%$	47.12	37.66	TRUE

Table 10: M_F non-parametric pairwise difference for Triples metric in ML framework

	$M_F00\%$	$M_F10\%$	$M_F25\%$	$M_F50\%$	$M_F75\%$
$M_L00\%$	-	**TRUE**	**TRUE**	**TRUE**	**TRUE**
$M_L10\%$		-	FALSE	**TRUE**	**TRUE**
$M_L25\%$			-	FALSE	**TRUE**
$M_L50\%$				-	FALSE
$M_L75\%$					-

pairwise comparison	obs.dif	critical.dif	difference
$M_F00\%$-$M_F10\%$	64.78	37.66	TRUE
$M_F00\%$-$M_F25\%$	89.09	37.66	TRUE
$M_F00\%$-$M_F50\%$	123.20	37.66	TRUE
$M_F00\%$-$M_F75\%$	150.43	37.66	TRUE
$M_F10\%$-$M_F25\%$	24.31	37.66	FALSE
$M_F10\%$-$M_F50\%$	58.42	37.66	TRUE
$M_F10\%$-$M_F75\%$	85.65	37.66	TRUE
$M_F25\%$-$M_F50\%$	34.10	37.66	FALSE
$M_F25\%$-$M_F75\%$	61.34	37.66	TRUE
$M_F50\%$-$M_F75\%$	27.23	37.66	FALSE

Table 11: M_F non-parametric pairwise difference for Robinson-Fould metric in Bayesian framework

	$M_F00\%$	$M_F10\%$	$M_F25\%$	$M_F50\%$	$M_F75\%$
$M_L00\%$	-	FALSE	FALSE	**TRUE**	**TRUE**
$M_L10\%$		-	FALSE	**TRUE**	**TRUE**
$M_L25\%$			-	FALSE	**TRUE**
$M_L50\%$				-	FALSE
$M_L75\%$					-

pairwise comparison	obs.dif	critical.dif	difference
$M_F00\%$-$M_F10\%$	16.03	41.00	FALSE
$M_F00\%$-$M_F25\%$	38.50	41.00	FALSE
$M_F00\%$-$M_F50\%$	61.52	41.00	TRUE
$M_F00\%$-$M_F75\%$	95.52	41.00	TRUE
$M_F10\%$-$M_F25\%$	22.47	41.00	FALSE
$M_F10\%$-$M_F50\%$	45.49	41.00	TRUE
$M_F10\%$-$M_F75\%$	79.49	41.00	TRUE
$M_F25\%$-$M_F50\%$	23.02	41.00	FALSE
$M_F25\%$-$M_F75\%$	57.02	41.00	TRUE
$M_F50\%$-$M_F75\%$	34.00	41.00	FALSE

Table 12: M_C non-parametric pairwise difference for Robinson-Fould metric in ML framework

	M_C00%	M_C10%	M_C25%	M_C50%	M_C75%
M_C00%	-	**TRUE**	**TRUE**	**TRUE**	**TRUE**
M_C10%		-	FALSE	**TRUE**	**TRUE**
M_C25%			-	FALSE	**TRUE**
M_C50%				-	**TRUE**
M_C75%					-

pairwise comparison	obs.dif	critical.dif	difference
M_C00%-M_C10%	67.99	37.66	TRUE
M_C00%-M_C25%	87.58	37.66	TRUE
M_C00%-M_C50%	116.16	37.66	TRUE
M_C00%-M_C75%	155.77	37.66	TRUE
M_C10%-M_C25%	19.59	37.66	FALSE
M_C10%-M_C50%	48.17	37.66	TRUE
M_C10%-M_C75%	87.78	37.66	TRUE
M_C25%-M_C50%	28.58	37.66	FALSE
M_C25%-M_C75%	68.19	37.66	TRUE
M_C50%-M_C75%	39.60	37.66	TRUE

Table 13: M_C non-parametric pairwise difference for Triples metric in ML framework

	M_C00%	M_C10%	M_C25%	M_C50%	M_C75%
M_C00%	-	**TRUE**	**TRUE**	**TRUE**	**TRUE**
M_C10%		-	FALSE	**TRUE**	**TRUE**
M_C25%			-	FALSE	**TRUE**
M_C50%				-	FALSE
M_C75%					-

pairwise comparison	obs.dif	critical.dif	difference
M_C00%-M_C10%	75.29	37.66	TRUE
M_C00%-M_C25%	88.58	37.66	TRUE
M_C00%-M_C50%	114.02	37.66	TRUE
M_C00%-M_C75%	149.60	37.66	TRUE
M_C10%-M_C25%	13.29	37.66	FALSE
M_C10%-M_C50%	38.73	37.66	TRUE
M_C10%-M_C75%	74.31	37.66	TRUE
M_C25%-M_C50%	25.44	37.66	FALSE
M_C25%-M_C75%	61.02	37.66	TRUE
M_C50%-M_C75%	35.58	37.66	FALSE

Table 14: M_C non-parametric pairwise difference for Robinson-Fould metric in Bayesian framework

	$M_C00\%$	$M_C10\%$	$M_C25\%$	$M_C50\%$	$M_C75\%$
$M_C00\%$	-	FALSE	FALSE	**TRUE**	**TRUE**
$M_C10\%$		-	FALSE	**TRUE**	**TRUE**
$M_C25\%$			-	FALSE	**TRUE**
$M_C50\%$				-	FALSE
$M_C75\%$					-

pairwise comparison	obs.dif	critical.dif	difference
$M_C00\%$-$M_C10\%$	16.03	41.00	FALSE
$M_C00\%$-$M_C25\%$	38.50	41.00	FALSE
$M_C00\%$-$M_C50\%$	61.52	41.00	TRUE
$M_C00\%$-$M_C75\%$	95.52	41.00	TRUE
$M_C10\%$-$M_C25\%$	22.47	41.00	FALSE
$M_C10\%$-$M_C50\%$	45.49	41.00	TRUE
$M_C10\%$-$M_C75\%$	79.49	41.00	TRUE
$M_C25\%$-$M_C50\%$	23.02	41.00	FALSE
$M_C25\%$-$M_C75\%$	57.02	41.00	TRUE
$M_C50\%$-$M_C75\%$	34.00	41.00	FALSE

Table 15: Group comparison results: difference between 100 replicates using the RPBTC method

Tree.Type	Used.metric	Replicates	Df	F.value	p.value
Low Score	RF	100.00	99.00	0.74	0.98
Low Score	Tr	100.00	99.00	0.97	0.58
Medium Score	RF	100.00	99.00	0.64	1.00
Medium Score	Tr	100.00	99.00	0.45	1.00
High Score	RF	100.00	99.00	0.20	1.00
High Score	Tr	100.00	99.00	0.37	1.00

Table 16: Tree similarity values per parameter in ML framework

Parameter	amount of missing data	metric	mode	50%CI lower	upper	95%CI lower	upper
M_L	0%	Robinson-Fould	1	NA	NA	NA	NA
		Triples	1	NA	NA	NA	NA
	10%	Robinson-Fould	0.6714878	0.5692308	0.7894649	0.4689069	1.0361514
		Triples	0.9424133	0.8683742	0.9998559	0.6121399	1.0635250
	25%	Robinson-Fould	0.5167595	0.4403147	0.6307692	0.2664429	0.8929104
		Triples	0.5599494	0.7084874	0.8105781	0.2912950	0.9990703
	50%	Robinson-Fould	0.4321852	0.3641026	0.5215627	0.2313246	0.6423846
		Triples	0.5143009	0.4197127	0.5991498	0.2991432	0.9104526
	75%	Robinson-Fould	0.3727090	0.3230769	0.4342249	0.2035474	0.5776286
		Triples	0.4449563	0.3385128	0.5329149	0.1929395	0.8551570
M_L	0%	Robinson-Fould	1	NA	NA	NA	NA
		Triples	1	NA	NA	NA	NA
	10%	Robinson-Fould	0.9072419	0.7668415	1.0000000	0.4070844	1.0478010
		Triples	0.9763617	0.9448100	1.0039460	0.7158118	1.1043330
	25%	Robinson-Fould	0.5722382	0.4883560	0.6923077	0.3253239	0.9367467
		Triples	0.9370152	0.8207407	0.9938846	0.5394443	1.0868073
	50%	Robinson-Fould	0.4338159	0.3025641	0.4760777	0.1704016	0.6683220
		Triples	0.7327464	0.5947211	0.8739630	0.3418679	1.0251540
	75%	Robinson-Fould	0.1534487	0.1155128	0.2000000	0.0683485	0.3230769
		Triples	0.4588292	0.3905631	0.6111484	0.2903676	0.8568274
M_L	0%	Robinson-Fould	1	NA	NA	NA	NA
		Triples	1	NA	NA	NA	NA
	10%	Robinson-Fould	0.473285	0.3842965	0.8358974	0.3564753	0.9589744
		Triples	0.9592557	0.8691580	1.026095	0.5720847	1.067524
	25%	Robinson-Fould	0.4548038	0.3572132	0.6307692	0.2425339	0.9161212
		Triples	0.9071925	0.7516934	0.9976224	0.4462168	1.0705291
	50%	Robinson-Fould	0.4244263	0.3755493	0.4666667	0.2494927	0.8287312
		Triples	0.6834366	0.5713927	0.8226139	0.2776690	1.0085944
	75%	Robinson-Fould	0.2306888	0.1774508	0.3025641	0.1087764	0.5673541
		Triples	0.4513642	0.3737397	0.5723858	0.2245224	0.7917946

Table 17: Tree similarity values per parameter in Bayesian framework

Parameter	amount of missing data	mode	50%CI lower	upper	95%CI lower	upper
M_L	0%	Robinson-Fould	0.05641026	0.03589744	0.06337695	0.03139209
		Triples	0.04104231	0.00830962	0.05441173	-0.03933377
	10%	Robinson-Fould	0.03584309	03584196	03589744	02825643
		Triples	0.03399849	0.003196469	0.04981018	-0.049913382
	25%	Robinson-Fould	0.03589744	0.03589744	0.03589744	0.01536739
		Triples	0.03510953	0.002113358	0.04939228	-0.041202264
	50%	Robinson-Fould	0.03589744	0.03589744	0.03589744	0.01538462
		Triples	0.02849433	0.003263098	0.05298036	-0.038344731
	75%	Robinson-Fould	0.03589744	0.03589744	0.03589744	0.01069992
		Triples	-0.0001940566	-0.01801301	0.04058783	-0.06030922
M_L	0%	Robinson-Fould	0.05641026	0.03589744	0.06337695	0.03139209
		Triples	0.04104231	0.008309622	0.05441173	-0.039333778
	10%	Robinson-Fould	0.03593712	0.03198256	0.05641026	0.02798864
		Triples	0.02639309	0.002185408	0.05061314	-0.046563042
	25%	Robinson-Fould	0.03588134	0.03587629	0.03589744	0.01538462
		Triples	0.03856296	0.002185408	0.05061314	-0.046563042
	50%	Robinson-Fould	0.03589744	0.03589744	0.03589744	0.01536289
		Triples	0.01403716	-0.01100608	0.03972324	-0.03800324
	75%	Robinson-Fould	0.03589206	0.03588825	0.03589744	0.01538462
		Triples	0.02939363	0.01902351	0.04238907	-0.03087905
M_L	0%	Robinson-Fould	0.05641026	0.03589744	0.06337695	0.03139209
		Triples	0.04104231	0.008309622	0.05441173	-0.039333778
	10%	Robinson-Fould	0.05637057	0.03589744	0.05927956	0.02813296
		Triples	0.03008132	0.01176800	0.04847400	-0.02451461
	25%	Robinson-Fould	0.03593743	0.03589744	0.03596018	0.02330790
		Triples	0.02010221	-0.001145118	0.0377779	-0.060183332
	50%	Robinson-Fould	0.03589744	0.03589744	0.03589744	0.03050289
		Triples	0.01053959	-0.005595947	0.04510444	-0.036278820
	75%	Robinson-Fould	0.03589744	0.03589744	0.03589744	0.01115983
		Triples	0.02427778	0.0007405737	0.04260522	-0.0373004106