Version dated: May 12, 2014

RH: Missing data in Total Evidence Supermatrices

# Effect of missing data in supermatrices 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 Method, 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 Method 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 supermatrices. Using simulations we investigate three major factors that directly affect the completeness of the morphological part of the supermatrix: (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 supermatrix. 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 methods, one should reduce the missing data in the morphological part of the supermatrix for living species and use a Maximum Likelihood framework to fix the topology prior to the overall Bayesian phylogenetic inference process. We apply our method to a comprehensive data set of both living and fossil primates. We find that using this integrative method modifies previous estimates of rates of body mass evolution within primates.
(Keywords: bla, bla, bla)

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). Ignoring fossil taxa may lead to misinterpretation of macroevolutionary patterns and processes such as species richness gradients, e.g. extant clade diversity may ignore past diversification patterns (Shoshani and Tassy 1996); biogeographical history, e.g. extant biogeographical patterns are improved when integrating fossil data (Metcalf et al. 2014); or paleoecology, e.g. extant clades niches might have differed greatly through time (Young et al. 2010). 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.

Three main approaches have been used for combining fossil and living taxa data in phylogenies. These approaches differ in whether they treat fossil taxa as nodes or tips and how much of the available data is used (i.e. age only or age and morphology). Classical cladistic methods use morphological data from both fossil and living taxa and treat each taxon as a tip (Simpson 1945). This approach is commonly used by paleontologists but it ignores the majority of the additional molecular data available from living species. Neontologists, on the other hand, more commonly use only molecular data from living species to build phylogenies. 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 is 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). A final class of methods, known as total evidence methods use molecular data from living taxa and morphological data from both

living and fossil taxa, and treats every taxon as a tip on the phylogeny (Eernisse and Kluge 1993).Here we focus on these total evidence methods 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 scarce. Therefore total evidence approaches are likely to suffer from having lots of missing data which may affect their ability to infer correct phyogenies.

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 supermatrices (Lemmon et al. 2009). However other authors do not see missing molecular data as a major issue because the phylogenetic signal is more likely to increase by having at least a "modest" number of highly covered genes ( 50% - Roure and Philippe (2011)), a higher number of taxa (especially slowly evolving taxa or taxa close to the outgroup) and 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, missing morphological data might be seen as major or minor issue for accurately infering phylogenies (Wiens 2003; Sansom and Wills 2013). Because soft-tissues characters are rarely preserved in the fossil record, missing data is found in soft tissues, i.e. it is 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 not a major problem separately (Wiens 2003, 2006; Wiens and

Moen 2008; Roure and Philippe 2011), missing data in molecular and morphological supermatrices may become an issue when combined for example in a total evidence type supermatrices and no attempt has been made to study the impact of this issue until now.

Here we assess the effect of missing data on tree topology infered from total evidence supermatrices. The molecular part of a total evidence supermatrix contains no fossil taxa so will act like a classical molecular supermatrix (Ronquist et al. 2012a). The effect of missing data on such matrices is well known, therefore, we only focus on the missing data issue in the morphological part of the supermatrix. Using simulations we investigate three major factors that directly affect the completeness of the morphological part of the supermatrix: (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 supermatrix. We assess how changing the values of these three parameters affects the topology of total evidence method phylogenies.

# Methods

To explore the effect of missing data on total evidence trees topologies we used the following protocol (note that we explain each step in detail below this general outline -Fig. 1):

1. Generating the matrix We built a randomly generated birth death tree to infer a "complete" matrix containing both molecular and morphological data for living and fossil taxa.

2. Removing data We removed data from 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 ($M_C$).

3. Building phylogenies We infered Bayesian phylogenetic trees from the "complete" matrix and from the matrices containing missing data.

4. Comparing topologies We then compared the trees obtained from the matrices containing missing data to the ones obtained from the "complete" matrix 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 datas distribution, we repeated steps 1 to 4 with the exact same fixed parameters 50 times.

## *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 ($\lambda$) and extinction events ($\mu$) 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 species 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 infered 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

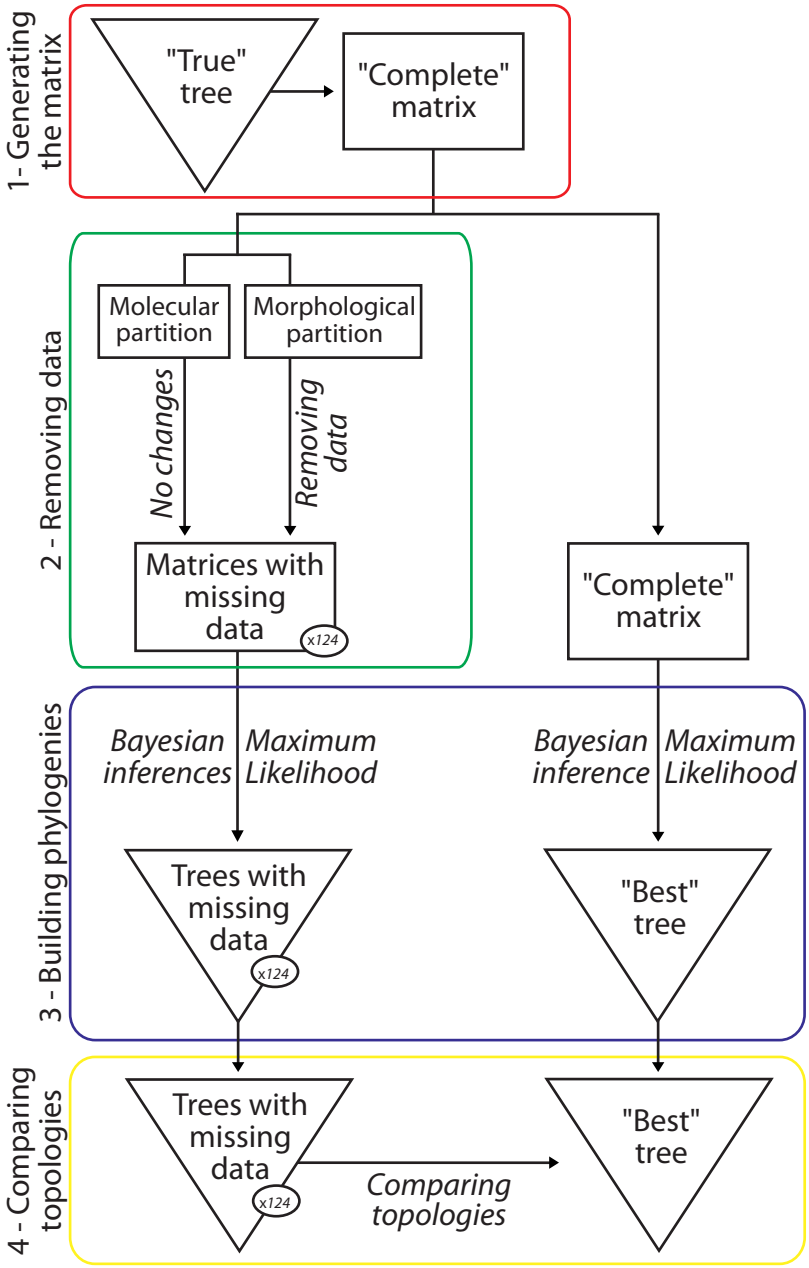


Figure 1: Protocol outline. (1) We generated a random tree (hereafter called the "true" tree) to infer a matrix with no missing data (hereafter, the "complete" matrix). (2) We removed data fron the morphological partition of the "Complete" matrix resulting in 124 matrices with missing data. (3) We build Bayesian phylogenetic trees from each matrix. (4) We then compared the trees with missing data (from the matrices with missing data) to the tree with no missing data (hereafter, the "best" tree infered from the "complete" matrix). We repeated step 1 to 4 50 times.

matrix. The substitution rates were distributed following a gamma distribution with an alpha ($\alpha$) shape of 0.5 (Yang 1996). We chose a low value of $\alpha$ to lower 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 (65 CPU years).

We infered the morphological matrix using the ape package v3.0-8 (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 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. This method allows us to have only two parameters per character i.e. the number of states and the evolutionary rate. For each character, the evolutionary rate was sampled from a gamma distribution with $\alpha = 0.5$. 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, i.e. 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):

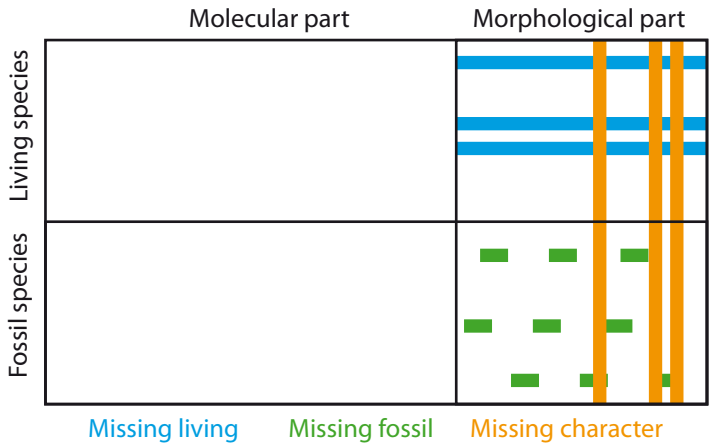


Figure 2: Data removal protocol: Missing living - The proportion of living taxa with no morphological data ($M_L$); Missing fossil - The proportion of missing morphological data across all fossil taxa ($M_F$); Missing character - The proportion of missing morphological characters across all taxa (living and fossil) ($M_C$).

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 species, the number of living species with molecular data is highly superior the the number of species 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 - $M_C$): 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 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. Note that $M_L$ is different to $M_F$ not only because of the region of the supermatrix 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 species had at least 5% data.

## *Building phylogenies*

From the resulting matrices we generated two types of trees, the "best" tree that is infered from the "complete" matrix and the trees with missing data infered 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 evolutionnary history of a clade (except in very rare circumstances, e.g. Rozen et al. (2005)). Therefore, here we focus on comparing the trees infered 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 infered using RAxML v7.0.4 (Stamatakis 2008). We used the GTR + $\Gamma_4$ model (Tavar (1986) default GTRGAMMA in RAxML v7.04) as a generalisation of the HKY + $\Gamma_4$ 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)). We used the fast bootstrap algorithm and performed 1000 bootstraps per tree inference to assess the topological support.

*Bayesian.*— The "best" tree and the "missing-data" trees were infered using MrBayes v3.2.2 (Ronquist et al. 2012b). We partitioned the data treating the molecular part as a non-codon DNA partition and the morphological part as a multi-state morphological partition. The molecular evolutionary history was infered using the HKY model with a transition/transversion ratio (Douady et al. 2003) of two and a gamma distribution for the rate variation with four distinct categories (HKY + $\Gamma_4$). For the morphological data, we used the Markov $k$ state model (Lewis 2001) wich is a generalisation of the JC69 model (Jukes and Cantor 1969) with $k$  2, assuming an equal state frequency and a unique overall substitution rate ($\mu$) following a gamma distribution of the rate variation with four distinct categories (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 $50\times1^6$ generations. We used the average standard deviation of split frequencies (ASDS) as a proxy to estimate the convergence of the chains and implemented 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 subsample of runs in each simulation to check that ESS >> 200 (Drummond et al. 2006). For each run, we removed 25% of the iterations as burnin. We used the following priors 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 priors to speed up the Bayesian process. These priors biased the way the Bayesian process calculated the branch length by giving non-random starting points and boundaries for the parameters estimation process, however, we are focusing on the effect of missing data on the topology and not on the branch length. Even using these priors, it took 65 CPU years to build 50 sets of 125 Bayesian trees (8 core nodes 2.30GHz clock speed).

## *Comparing topologies*

We compared the topology of the "missing-data" trees infered 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 infered from supermatrices with living and fossil species and molecular and morphological characters. To compare the topology of the resulting trees, we used two metrics to assess number of conserved taxa and clades position using respectively the Triple (Dobson 1975) and the Robison-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 two metrics and the index are detailed below.

*Triple distance ($T_{x,y}$) (Dobson 1975).*— This metric measures the number of different subtrees of three clades 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 displaxed taxa but is less sensitive to the placement of individual taxa and to taxa of highly uncertain placement (e.g. fossil taxa) than the Robinson-Foulds distance (Critchlow et al. 1996; Johnson and Soltis 1998; Wiens 2003) and can therefore be used 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 Robison-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:

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

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 triple metric (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)) 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 (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:

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

$NTS$ ranges from 1 to -$\infty$. 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 $Y$ (where $Y$ is one of the 125 trees resulting from the 125 super matrices including various amount of missing data) for both the Robinson-Fould and the Triple metric. We calculated the metrics 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 comparison for the same metric and the same number of taxa $n$=51 (see supplementaries). We rand the comparison for the 50 trees of each chain having the same combination of parameters values $M_L$, $M_F$ and $M_C$.

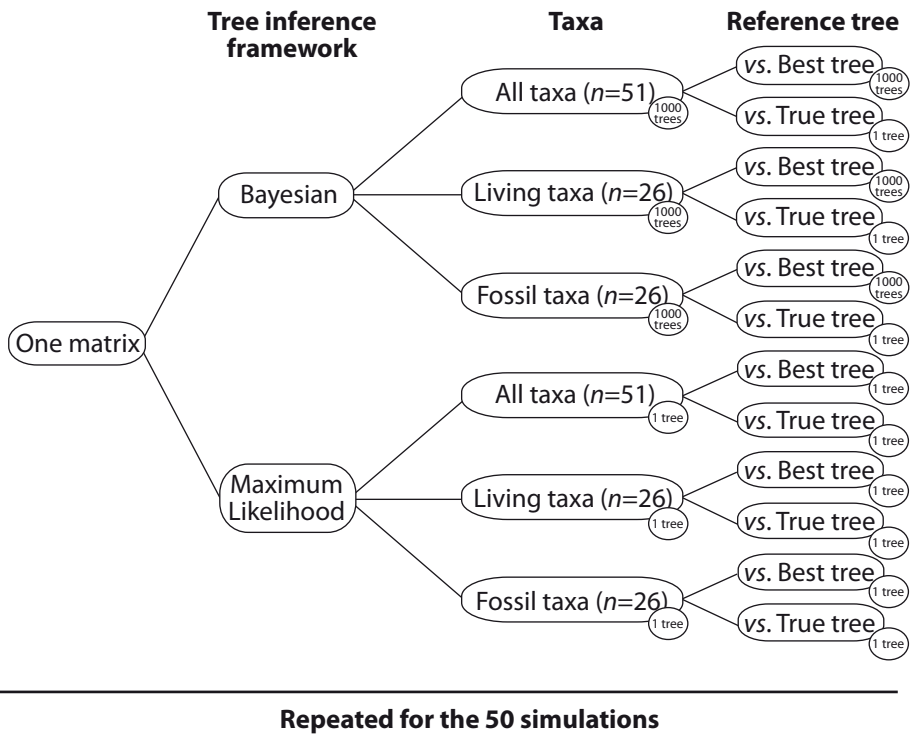


Figure 3: Tree comparison protocol. From each of the 125 parameters combination ($M_L$, $M_F$ and $M_C$), we infered a tree in Baysian or Maximum likelihood framework. We then compared this tree with or without fossil/living taxa to the "best/true" tree. In the Bayesian framework, 1000 trees from the posterior ditribution where used for each comparison.

We calculated the mode and the 50, 75 and 95 confidence intervals for each comparison of 50 trees using the hdrcde R package (v3.1 with contributions from Jochen Einbeck and Wand (2013)). For the Bayesian trees, we used the exact same approach described above but instead of comparing the "True" or "Best" tree to the tree $Y$, we compared 1000 trees from the "True" or "Best" tree posterior distribution to 1000 trees from the tree $Y$ posterior distribution resulting in 1000 pairwise comparison of the "True" or "Best" tree to the tree $Y$ for each 125 trees (see supplementaries - Fig. 3).

## *Empirical data*

We also compared the results obtained from simulated data by using Ronquist et al. (2012a) empirical data. The supermatrix 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 supermatrix, 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 supermatrices 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 infered tree and not the branch length.

# Results

# Discussion

# Supplementaries

blabla

*

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.

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.

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.

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.

Metcalf, J., S. Prost, N. David, E. Dechaine, C. Anderson, P. Batra, M. Araújo, A. Cooper, and R. Guralnick. 2014. Integrating multiple lines of evidence into historical biogeography hypothesis testing: a bison bison case study. Proceedings. Biological sciences / The Royal Society 281.

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

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.

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.

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.

Shoshani, J. and P. Tassy. 1996. The Proboscidea. Oxford University Press.

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. 2008. The RAxML 7.0. 4 manual. Department of Computer Science. Ludwig-Maximilians-Universitte Mnchen .

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

Tavar, S. 1986. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences vol. 17 of *Some Mathematical Questions in Biology*. American Mathematical Society.

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.

Young, M. T., S. L. Brusatte, M. Ruta, and M. B. de Andrade. 2010. The evolution of metriorhynchoidea (mesoeucrocodylia, thalattosuchia): an integrated approach using geometric morphometrics, analysis of disparity, and biomechanics. Zoological Journal of the Linnean Society 158.

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