MACROEVOLUTION WITH FOSSIL AND LIVING TAXA

by

THOMAS GUILLERME

B.Sc., Université Montpellier 2, 2010
M.Sc., Université Montpellier 2, 2012

A thesis submitted in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

School of Natural Sciences
(Zoology)

Trinity College Dublin

SEPTEMBER 2015

DECLARATION

I declare that this thesis has not been submitted as an exercise for a degree at this or any other University and it is, unless otherwise referenced, entirely my own work. I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Thomas Guillerme

SUMMARY

Even if most of our current knowledge and tool-kits to study biodiversity focus on living species, the vast majority of the species that ever lived are long extinct. Therefore, to properly understand the drivers of biodiversity through time, it is crucial to combine data and methods from both living and fossil species in order to better assess macroevolutionary and macroecological patterns. My PhD focuses on ways to combine both living and fossil species into phylogenies and looks at how these phylogenies can be used for describing macroevolutionary patterns. I studied the use of both living and fossil species along two axes: firstly, the ability of modern phylogenetic methods to deal with molecular data for living species and morphological data for both living and fossil species; and secondly, the practicality of using such phylogenetic trees for more accurately describing patterns of diversification through time and space.

For the first part of this project, I ran extensive and thorough simulation analyses to test the effect of missing data on phylogenies topologies when using a combination of living and fossil data. I tested how multiple levels of missing data among living species, fossil species and the two combined affected our ability to recover the correct tree topology. I found that the amount of missing data among living species is the most crucial aspect for efficiently combining living and fossil species in the same phylogeny. Following these conclusions, I ran a thorough survey of the data available for living mammal species. I measured the amount of morphological data available within each mammalian order and tested whether this data was randomly distributed along the phylogeny or biased towards certain clades. The result of this analysis shows that although morphological data is scarce for living mammals, it is at least generally randomly distributed across the phylogeny.

For the second part of my PhD, I explored a way of using phylogenetic trees containing both living and fossil species to measure patterns of diversification among mammals through time. I measured changes in species richness as well as in morphological diversity (i.e. disparity) to describe mammalian diversification across the K-Pg boundary. I found that the K-Pg boundary had no significant effect on morphological diversification.

ACKNOWLEDGEMENTS

Thanks folks!

PREFACE

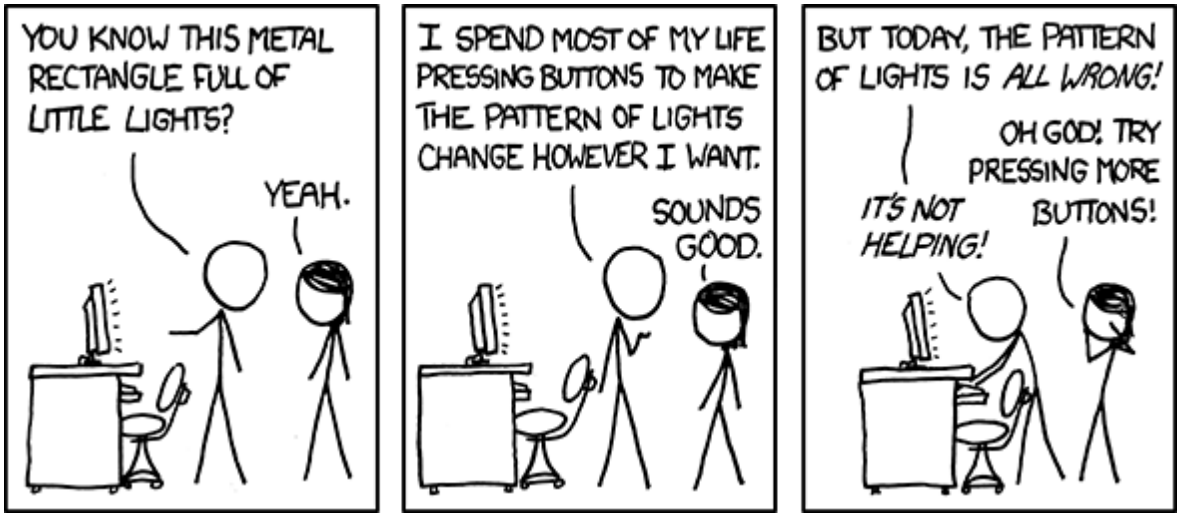

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

GENERAL INTRODUCTION

Today's amazing biodiversity represents only an overwhelmingly small fraction of the organisms that ever existed (Novacek and Wheeler, 1992; Raup, 1981). Even though it is widely accepted that the processes that shaped the patterns observed nowadays are influenced by evolutionary history (Fritz et al., 2013), most of the scientific endeavour in biology focus solely on living species. Ignoring this can lead to misinterpretation of macroevolutionary patterns and processes (Benton, 2015). For example, nowadays crocodilians constitute a species poor group (25 species; Uetz, 2010) with a low range of shapes and environments (marine or freshwater; Martin, 2008). Therefore when studying macroevolutionary patterns among all vertebrates, crocodilians will have a rather "marginal" effect. For example Wiens (2015) suggests that terrestriality is a driver of diversification among living vertebrates, a pattern essentially driven by Aves, Lepidosauria and Mammalia. However, crocodilians were much more diverse both in terms of species richness (244 species reported in Bronzati et al., 2015) or in terms shapes and environments (Stubbs et al., 2013). In the case of Wiens (2015), not including fossil species, conceal the true history of this clade, and thus, potentially biases the conclusions of the study.

Besides, including fossil species not only accounts for groups that were more diverse in the past, it also highly improves our descriptions of macroevolutionary patterns such as the timing of diversification events (e.g. significantly reducing node age confidence intervals; Ronquist et al., 2012a), the relationships among lineages (e.g. solving some controversial fossil placement; Dembo et al., 2015) or even gives a potential solution for understanding niche occupancy through time (e.g. Pearman et al., 2008). All this studies have led to a recent consensus among scientists that we need to combine both living and fossil species in macroevolutionary analysis (Jackson and Erwin, 2006; Quental and Marshall, 2010; Dietl and Flessa, 2011; Slater and Harmon, 2013; Fritz et al., 2013; Benton, 2015). Yet, in practice, only few studies have actively focused on combining them since the last three years (e.g. Ronquist et al., 2012a; Slater, 2013; Wood et al., 2013; Beck and Lee, 2014; Arcila et al., 2015; Dembo et al., 2015).

This scarcity is probably due to the fundamental differences between the two approaches to study macroevolution: using either living (neontological) or fossil (palaeontological) data.

The paleontological approachis based on cladistic data of the fossil record (i.e. discrete morphological observation). It relies on optimal criteria such as maximum parsimony (Hennig, 1966; Felsenstein, 2004) to resolve the relations among lineages and on stratigraphy to time such trees (Goloboff et al., 2008). This approach allows a direct interpretation of macroevolution in deep time and benefits from recent improvements both on data collection (e.g. 4541 characters in O'Leary et al., 2013, introducing the term "phenomics") and on dating method (e.g. the *cal3* method from Bapst, 2014). However, this approach does rarely takes into account full living diversity (e.g. only 38 out of 351 living primates for 119 fossil in Ni et al., 2013) and methods suffer from several biases (e.g. parsimony; Wright and Hillis, 2014).

Conversely, the neontological approach uses the vast amount of available molecular from living species and is based on probabilistic methods (e.g. Maximum Likelihood or Bayesian). This approach is based on evolutionary models that rely on the differences in DNA to resolve the relations among lineages and on some specific fossils' occurrence dates for timing the lineages divergence (i.e. the molecular clock; Zuckerkandl and Pauling, 1965). There has been enormous improvements of this approach in the last decade on both the evolutionary models (e.g. Bapst, 2013; Stadler and Yang, 2013; Heath et al., 2014) and on which fossils to use to calibrate the trees (Donoghue and Benton, 2007; Parham et al., 2012). However, this approach uses only the ages of certain fossils instead of the vast amount of informations available from the fossil record (e.g. species richness, traits, biogeography, etc).

1.1 PHYLOGENIES WITH LIVING AND FOSSIL TAXA

Nonetheless, the last three years have seen the development of the newly trending Total Evidence method (Ronquist et al., 2012a; Slater, 2013; Wood et al., 2013; Schrago et al., 2013; Beck and Lee, 2014; Arcila et al., 2015; Dembo et al., 2015). This methods allows to combine both molecular data from living taxa and morphological data from living and fossil taxa in the same phylogenetic matrices. It was first developed in the nineties (Eernisse and Kluge, 1993) but only recently successfully implemented in phylogenetic softwares (Ronquist et al., 2012b; Bouckaert et al.,

2014). By using both available neontological and palaeontological data, this methods allows to better study macroevolutionary patterns and processes. For example, it allowed great improvements on the estimation of divergence event (e.g. Ronquist et al., 2012a); evolutionary rates (e.g. Beck and Lee, 2014); topology (e.g. Dembo et al., 2015); traits evolution (e.g. Slater, 2013) or even speciation processes (e.g. Wood et al., 2013). There is, however, one drawback to this method: because it needs both molecular data for living taxa and morphological data for living and morphological taxa, it is susceptible to suffer from great amounts of missing data.

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

As a first part of this PhD thesis, in the second chapter, I tackled the problem of missing data in Total Evidence matrices. I ran long term and thorough simulations to test how the topologies inferred from Total Evidence matrices were sensitive to missing morphological data. I removed morphological data from Total Evidence matrices via three parameters where data are potentially missing: (1) the number of living taxa with molecular data but no morphological data; (2) the amount of missing data in the fossil record and (3) the number of overall morphological characters in the matrix. I modified the level of data in the three parameters and in their combination and then inferred the phylogenies using both Maximum Likelihood and Bayesian approach. Finally, I compared how the missing data parameters and their interactions as well as the phylogenetic inference method influenced the ability of estimating the correct tree topology. I found that the number of living taxa with both morphological and molecular data are essential to recover accurate topologies. This study rose the question of how can we improve Total Evidence topologies and especially, how much morphological data are available for living taxa?

1.1.2 *Morphological data availability in living mammals*

Following this question, in the third chapter of this thesis, I monitored how many morphological data were available in the literature for living mammals. I downloaded all the recent available morphological matrices and counted the number of living mammals with available morphological data at three different taxonomic levels (species, genus and family) for each mammalian order. For each order with missing data, I measured if the data weren't biased toward some specific clades in

each order using phylogenetic structure methods (Webb et al., 2002). I found that a lot of living mammals have no morphological data at the species or the genus level but, that at least most of the available data was randomly distributed. These results highlight the importance of cladistics and collecting morphological data, even in the age of genomics, especially for combining living and fossil data in the same phylogenies.

1.2 TOTAL EVIDENCE PHYLOGENIES APPLICATIONS

The two previous chapters only focused on the technical and practical side of combining living and fossil taxa into phylogenies and underlined the importance of a good data overlap between living and fossil taxa. However, many studies have been able to use the Total Evidence method even with low overlap between living and fossil taxa by using strong topological constraints (Ronquist et al. 2012a; Schrago et al. 2013; Slater 2013; Beck and Lee 2014; but see Arcila et al. 2015; Dembo et al. 2015). This resulted in Total Evidence phylogenies were the topology are based on strong but valid *a priori* topologies (e.g. based on Meredith et al. 2011 for Slater 2013). The observable patterns in these phylogenies can then be used by biologists to test some hypothesis relating to macroevolutionary processes.

One example of pattern observed in phylogenies can be the shift of ecological dominant species through time due to drastic biotic or abiotic changes in the biosphere (e.g. mass extinctions). For example, Brachiopoda was a dominant shelled filter feeding clade during the Paleozoic (514 to 252 million years ago; Mya) but was replaced by Bivalvia at the end Permian extinction event (252 Mya) which is now the dominant group (Sepkoski 1981; Clapham et al. 2006 but see Payne et al. 2014). This type of replacement pattern has also been observed in other groups such as Formaninifera (Coxall et al., 2006), Ichtyosauria (Thorne et al., 2011) or Plesiosauria Benson and Druckenmiller 2014 and are often related to competition (Brusatte et al., 2008) or adaptive radiations (Losos, 2010). Another classical example is the “replacement” of the dominant non-avian dinosaurs by mammals after the infamous Cretaceous-Paleogene (K-Pg) extinction 66 Mya...

1.2.1 *Cretaceous-Palaeogene extinction does not affect mammalian disparity*

In this fourth chapter, I studied the changes of morphological diversity (or disparity; Wills et al., 1994) through time using Total Evidence trees from Slater (2013) and

Beck and Lee (2014) to test whether the K-Pg extinction had an effect on mammal evolution. I propose a new approach to describe patterns of disparity through time base on the use of Total Evidence trees. This approach allows more precision in describing the changes through time as well as more freedom for choosing the underlining models of morphological evolution (e.g. punctuated or gradual; Hunt et al., 2015). Using this approach I found no evidence of changes in disparity in mammals around the K-Pg boundary, arguing that the extinction of non-avian dinosaurs had no direct effect on mammalian evolution.

This is just one example of the benefits of adding both living and fossil taxa in macroevolutionary studies. In the fifth chapter, I will discuss how the three previous chapters open new axis of research as well as the limitation of these studies. Finally I will present some concluding thoughts on the utility of combining data, methods and disciplines to better understand macroevlutionary patterns and processes.

CHAPTER 2

TOTAL EVIDENCE METHOD AND MISSING DATA

Effects of missing data on topological inference using a Total Evidence approach [1]

ABSTRACT

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

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

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

[1] A similar version of this chapter is currently (2015/09/30) under review in Molecular Phylogenetics and Evolution. T.G. and N.C. designed the experiments; T.G. ran the analysis and interpreted the results; T.G. and N.C. wrote the manuscripts. *Specific acknowledgements*: thanks to Gavin Thomas, Frédéric Delsuc, Emmanuel Douzery, Trevor Hodkinson, Andrew Jackson, Nick Matzke, and April Wright for useful comments on our simulation protocol and manuscript. Thanks to Paddy Doyle, Graziano D'Innocenzo and Sean McGrath for assistance with the computer cluster. Thanks to two anonymous reviewers for their useful and encouraging comments. *Data availability and reproducibility*: All the code used in this analysis is available on GitHub (goo.gl/4djNUf) with some information on how to use the various functions. Additionally all the simulated data is available on FigShare (dx.doi.org/10.6084/m9.figshare.1306861).

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

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

2.1 INTRODUCTION

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

Up to now, three main approaches have been used to place both living and fossil taxa into phylogenies. These approaches differ mainly in how they treat fossil taxa and their data. One can use fossils as tips or as nodes in the phylogeny, and can use only the age of the fossils, only the morphology of the fossils, or age and morphology jointly. Classical cladistic methods use matrices containing morphological data from both living and fossil taxa and treat each taxon as a tip in the phylogeny. Relationships among the taxa are then inferred using optimality criteria such as maximum parsimony (Hennig, 1966; Felsenstein, 2004). This approach is commonly used by paleontologists but it ignores the additional molecular data available from living species and does not allow use of probabilistic methods for dealing with phylogenetic uncertainty. Neontologists, on the other hand, more commonly use probabilistic approaches (e.g. Maximum Likelihood or Bayesian methods) based on matrices containing only molecular data from living species. Because fossil taxa do not usually have available DNA, only fossil occurrence dates are used to time calibrate phylogenies (Zuckerkandl and Pauling, 1965). There have been great improvements in the theory and application of these two approaches (e.g. Bapst, 2013; Stadler and Yang, 2013; Heath et al., 2014) as well as much debate about the “best” approach to use (e.g. Spencer and Wilberg, 2013; Wright and Hillis, 2014). Neither approach, however, uses all the available data.

A final approach, known as the Total Evidence method, uses matrices containing molecular data from living taxa and morphological data from both living and fossil

taxa (Eernisse and Kluge, 1993). This approach treats every taxon as a tip in the phylogeny, uses the occurrence age of the fossils to time calibrate the phylogeny (known as tip-dating; Ronquist et al., 2012a), and allows the use of probabilistic methods for estimating phylogenetic uncertainty (Ronquist et al., 2012a). The Total Evidence method is becoming an increasingly popular way of adding fossil taxa to phylogenies (e.g. Pyron, 2011; Ronquist et al., 2012a; Schrago et al., 2013; Slater, 2013; Beck and Lee, 2014; Arcila et al., 2015). Although the Total Evidence approach seems very promising, there is one big drawback in using this approach: it requires both molecular and morphological data, both of which can be difficult (or impossible) to collect for every living and fossil taxon in the tree. Morphological data for living taxa are rarely collected when molecular data are available (e.g. O'Leary et al., 2013 *vs.* Meredith et al., 2011), and for fossil taxa, data can only be collected from features preserved in the fossil record. For example, in vertebrates, the hardest parts of the skeleton are more often preserved than soft parts (Sansom and Wills, 2013); and molecular data are (nearly) always unavailable. Therefore Total Evidence matrices are likely to contain a large proportions of missing data that may affect the method's ability to infer correct topologies, branch lengths and support values (Salamin et al., 2003).

Although missing data do not appear be a major problem in molecular and morphological matrices separately (as long as enough data overlaps in each case, and missing data are not phylogeneticaly biased; Wiens, 2003; Wiens et al., 2005; Wiens, 2006; Wiens and Moen, 2008; Lemmon et al., 2009; Sanderson et al., 2011; Roure and Philippe, 2011; Pattinson et al., 2014), it may become more of an issue in Total Evidence matrices containing both molecular and morphological data for living and fossil taxa. This may be particularly problematic as fossil taxa (generally) do not have molecular data, resulting in a large section of missing data in Total Evidence matrices. Until now, few attempts have been made to study the impact of this missing data issue on phylogenetic inference in a Total Evidence framework (i.e. using both molecular and morphological data; Wiens et al., 2005; Manos et al., 2007; Pattinson et al., 2014). These previous studies assessed the effect of missing data on topology by either (1) comparing a dataset with missing data to subsets without missing data (Wiens et al., 2005); or (2) removing both molecular and some morphological data from living taxa to create artificial fossils (Manos et al., 2007; Pattinson et al., 2014). Both approaches have shown that missing data are not a major problem and should not be an obstacle to combining both living and fossil species in the

same phylogenies. The way these studies were conducted, however, means that their conclusions are not generally applicable across all scenarios involving missing data in Total Evidence phylogenies. For example, using an empirical (rather than simulation based) approach limits their conclusions to studies with similar distributions of data across species in the phylogeny. Additionally, Wiens et al. (2005) did not include fossil taxa in their analyses, so their results cannot be used to make conclusions about how missing data may influence the placement of fossils. Manos et al. (2007); Pattinson et al. (2014) include fossil taxa, but used the patchiness of the fossil record to determine how to remove data from their matrices. Data for living species are unlikely to be missing in this patchy way, instead full molecular data with the complete absence of morphological data is a likely pattern (Guillerme and Cooper, 2015). Finally, Manos et al. (2007) and Pattinson et al. (2014) mainly focused on how missing data in fossil taxa affect the placement of fossils, ignoring the effects of missing data in living species.

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

We focus on the effects of missing data on our ability to recover tree topology because it is a crucial aspect of a phylogeny in many macroevolutionary studies, for example when trying to elucidate the evolutionary relationships among species (e.g. Meredith et al., 2011; Jetz et al., 2012), or for studying evolutionary transitions (e.g. Friedman, 2010). Although branch length estimation is also important (namely for timing extinction and/or speciation events; e.g. Ronquist et al., 2012a), we do not consider branch lengths in this study. This is partially due to difficulties with simulating branch lengths and topology simultaneously, but also because previous studies have already empirically assessed the effect of the Total Evidence method

on branch length variation but using topological constraints (Ronquist et al., 2012a; Schrago et al., 2013; Slater, 2013; Beck and Lee, 2014). Thus understanding the sensitivity of topology to missing data is important for assessing the accuracy of tree estimation in the Total Evidence framework. To our knowledge, this question has never been formally assessed.

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

2.2 MATERIALS AND METHODS

To explore how missing data in the morphological partition of Total Evidence matrices influences tree topology, we used the following protocol (Fig. 2.1):

1. Generating the matrix:

   We randomly generated a birth-death tree (hereafter called the “true” tree) and

used it to simulate a matrix containing both molecular and morphological data for living and fossil taxa (hereafter called the "complete" matrix).

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

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

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

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

2.2.1 *Generating the matrix*

First we randomly generated a "true" tree of 50 taxa in R v. 3.0.2 (**?**) using the package diversitree v. 0.9-6 (FitzJohn, 2012). We generated the tree using a birth death process by sampling speciation (λ) and extinction (μ) rates from a uniform distribution (bounded between 0 and 1) but maintaining $\lambda > \mu$ (Paradis, 2011). Empirical Total Evidence matrices vary in whether they have more fossil than living taxa or vice versa. For example, fossil taxa make up 88% (Beck and Lee, 2014), 58% (Schrago et al., 2013), 48% (Pyron, 2011), 31% (Ronquist et al., 2012a) and 31% (Slater, 2013) of taxa in various studies. To avoid biasing our simulations towards either living or fossil taxa and to make each simulation comparable, we implemented a rejection sampling algorithm to select only trees with 25 living and 25 fossil taxa.

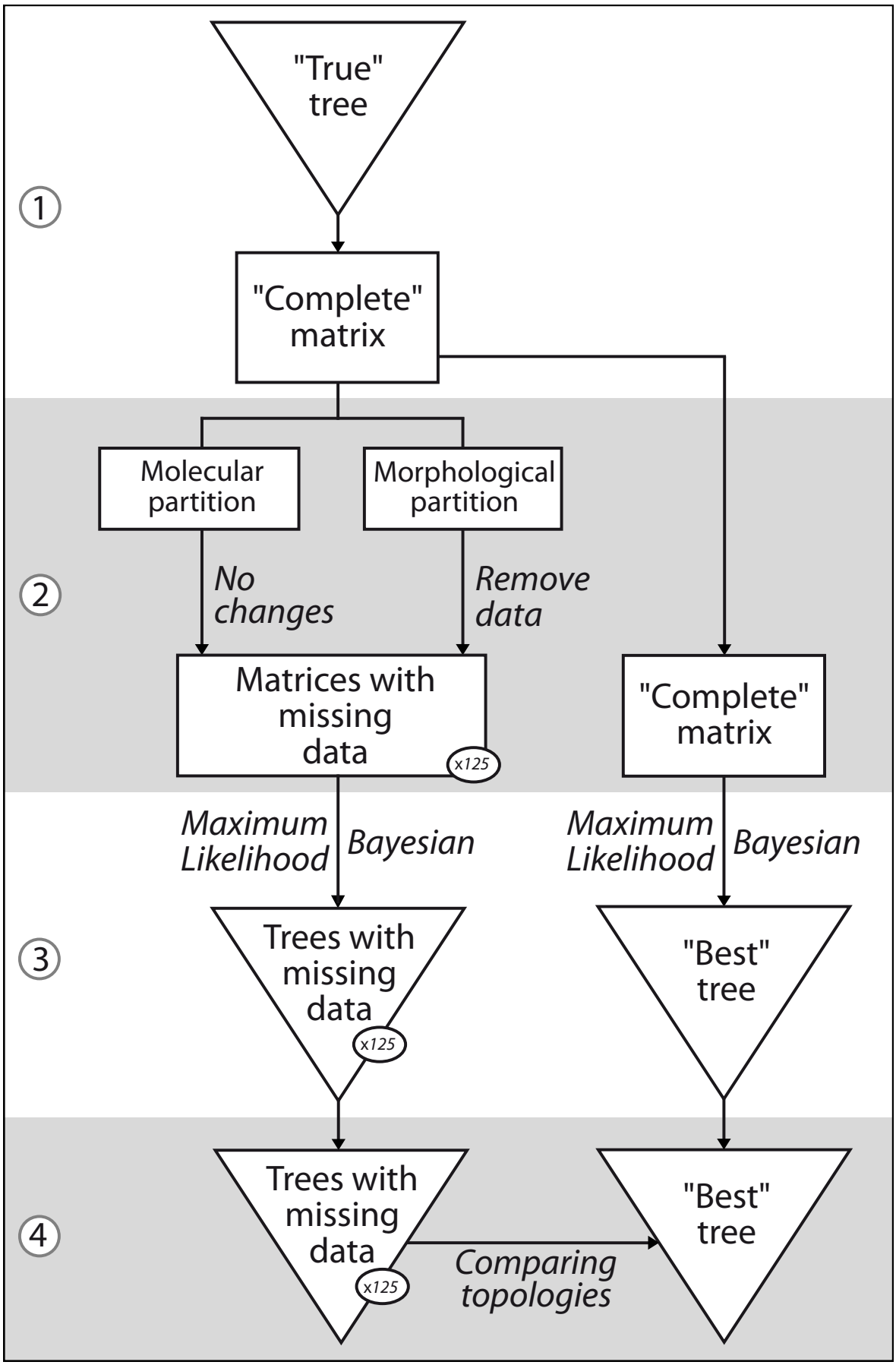

FIGURE 2.1: **Protocol outline.** (1) We randomly generated a birth-death tree (the "true" tree) and used it to simulate a matrix with no missing data (the "complete" matrix). (2) We removed data from the morphological part of the "complete" matrix resulting in 125 "missing-data" matrices. (3) We built phylogenetic trees from each matrix using both Maximum Likelihood and Bayesian methods. (4) We compared the "missing-data" trees to the "best" tree. We repeated these four steps 50 times.

The fossil taxa were considered as unique tips at the end of extinct lineages. We then added an outgroup to the tree, using the mean branch length of the tree to separate the outgroup from the rest of the taxa, and with the branch length leading to the outgroup set as the sum of the mean branch length and the longest root-to-tip length of the tree.

Next, we generated a molecular and a morphological matrix from the "true" tree. The molecular matrix was simulated from the "true" tree using the R package phyclust v. 0.1-14 (Chen, 2011). The matrix contained 1000 character sites for 51 taxa and was generated using the seqgen algorithm (Rambaut and Grassly, 1997) and using the HKY model (Hasegawa et al., 1985) with random base frequencies (sampled from a uniform probability distribution bounded between 0 and 1 with the total frequency for the four bases equal to 1) and transition/transversion rate of two (Douady et al., 2003). The substitution rates were selected from a gamma distribution with an (α) shape of 0.5 (Yang, 1996). In practice, a value of $\alpha < 1$ decreases the number of sites with high substitution rates, thus reducing homoplasic sites and increasing the phylogenetic signal (Hassanin et al., 1998; Estoup et al., 2002). Also, we chose this α value to be consistent with our protocol for simulating morphological characters (see below). This model and these parameter settings strike a balance between realism for empirical datasets (e.g. Douady et al., 2003; Kelly et al., 2014) and parameter richness with more complex models (e.g., GTR, multiple partitions with independent models), making them more suitable for our computational limitations (even with the parameters defined, the total computational time for the whole analysis was around 150 CPU years). All the molecular information for fossil taxa was replaced by missing data ("?").

We simulated the morphological matrix using the rTraitDisc function from the R package ape v. 3.0-11 (Paradis et al., 2004) to generate a matrix of 100 character sites for 51 taxa. We assigned the number of character states (either two or three) for each morphological character by sampling with a probability of 0.85 for two states characters and 0.15 for three state characters. We extracted these values from 100 random empirical matrices with more than 100 characters each downloaded from TreeBASE (`http://treebase.org/`). We selected matrices published between 1985 and 2013 and covering 19 taxonomic classes (Chordata, Arthropoda, Annelida, Angiosperm, Gymnosperm and Pteridophyta). These matrices contained a cumulative number of 22563 characters that had between two and 10 character states. We then extracted the proportion of characters with each number of states (two to 10)

to give us an empirical estimate of the average number of character states for each character, as shown in Fig. 2.2. Most morphological characters have two or three states, therefore we only simulate characters with two or three states, and sampled these in proportion to their occurrence in our empirical data (probability of 0.85 for two states characters and 0.15 for three state characters).

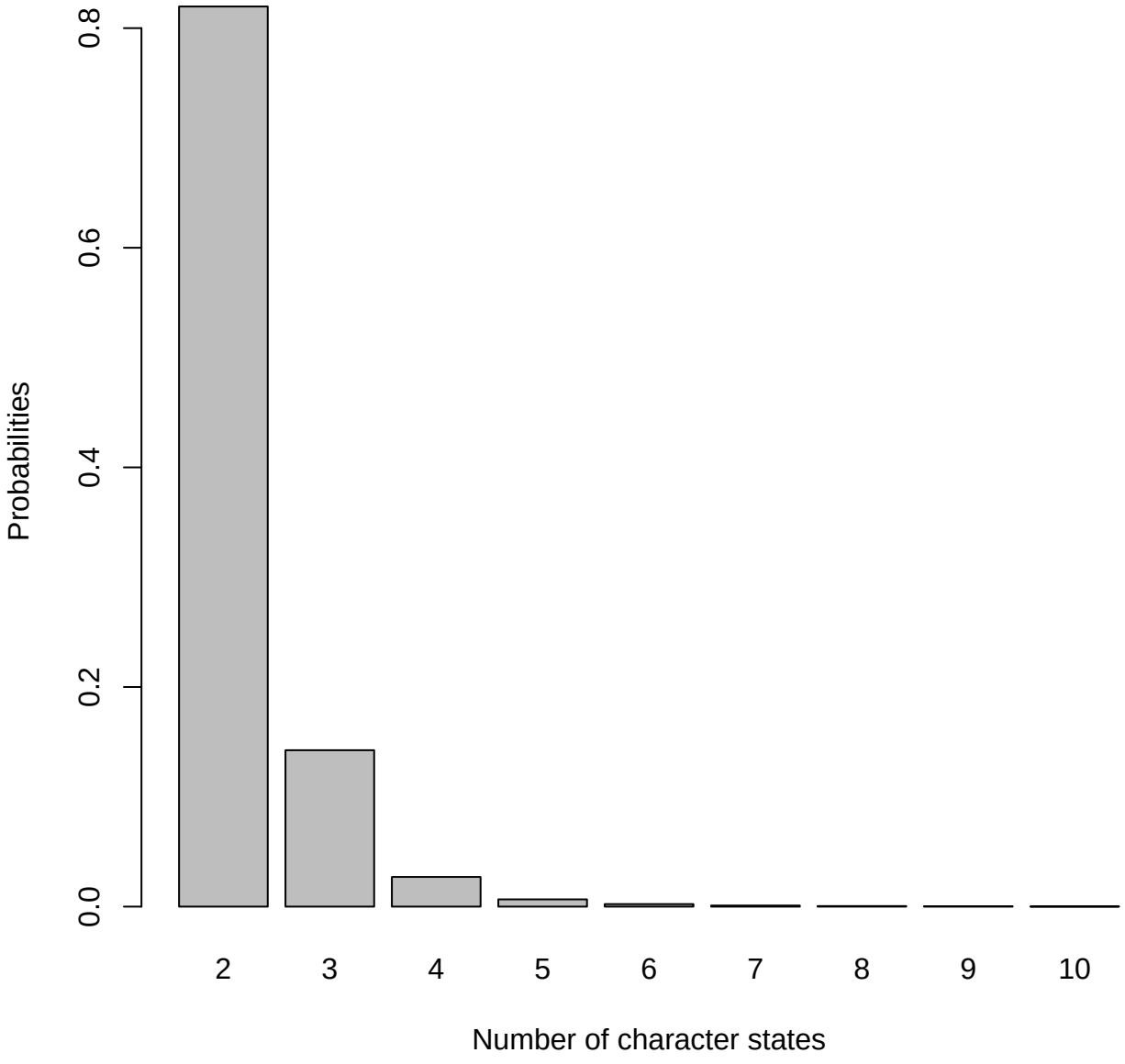

FIGURE 2.2: The proportion of morphological characters with between two and 10 character states extracted from 100 randomly selected empirical matrices downloaded from TreeBASE.

We then ran an independent discrete character simulation for each character using the “true” tree with the character’s randomly selected number of states (two

or three) and assuming an equal rate of change (i.e. evolutionary rate) from one character state to another (Pagel, 1994). This method allows us to have only two parameters for each character: the number of states and the evolutionary rate. For each character, the evolutionary rate was sampled from a gamma distribution with α = 0.5. We used low evolutionary rate parameters to be consistent with the molecular rate parameters, to avoid homoplasy in the morphological part of the matrix and create a clear phylogenetic signal (Wright and Hillis, 2014). Topological error has been shown to be minimal at a morphological rate of 0.5 when using the M*kv* model (Lewis, 2001; Wright and Hillis, 2014). Note, however, that Wright and Hillis (2014) have shown that low morphological rates (< 0.5) increase variance in topological error, but we discarded simulations with such topological error by selecting only matrices with a "fair" phylogenetic signal (see Estimating phylogenies section below; Zander, 2004) so this should not influence our results.

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

2.2.2 *Removing data*

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

1. The proportion of missing living taxa (M_L). This first missing-data parameter corresponds to the proportion of living taxa with no morphological data. It represents the number of living taxa that are present in the matrix but have only molecular data available. This reflects the fact that, because of the increasing ease of collecting molecular data, morphological data for living species are rarely collected (Guillerme and Cooper, 2015). Therefore, many living species will have only molecular data available. In practice, we removed all the morphological data from randomly chosen living taxa with five different proportions: 0%, 10%, 25%, 50% or 75% of living taxa with no morphological data.

2. The proportion of missing data in the fossil record (M_F). This missing data parameter represents the completeness of the fossil record. Due to preservation biases, missing data for fossil taxa are common (Sansom and Wills, 2013). In practice, we randomly removed a proportion of data from across the fossil taxa with five different proportions: 0%, 10%, 25%, 50% or 75% of overall missing data for the fossil taxa. Note that 50% missing data for fossil taxa does not mean that each fossil is missing 50% of its morphological data. Instead this 50% refers to missing fossil data across the whole matrix. Some fossils may retain 100% of their data and others may lose most of their data at this parameter value (down to a minimum threshold of 5% available data; see below).

3. The number of morphological characters for both living and fossil taxa (N_C). This parameter is not a missing data parameter *per se* but rather an indication of the size of the matrix. Any morphological matrix of any size has indeterminate missing data, given that the total number of characters is undefined, but presumably large. Therefore, this parameter corresponds to the overall number of characters available for both living and fossil taxa. In practice, we randomly removed entire characters from the morphological matrix reducing it to: 100, 90, 75, 50 or 25 characters. Note that these levels are equivalent to the two other parameters (i.e. 0%, 10%, 25%, 50% or 75% of "missing" morphological characters).

Each parameter represents a different way of removing data from the morphological part of the matrix: M_L removes entire rows from the living data; M_F removes cells from the fossil data; and N_C removes columns across both living and fossil data. Note that M_L and M_F differ not only because of the region of the matrix affected: for M_L all the morphological data of a percentage of living taxa are removed, whereas for M_F a percentage of the data are removed at random from across the whole of the morphological matrix for fossil taxa.

We created matrices using all parameter combinations resulting in 125 (5^3) "missing-data" matrices. Note that one of these combinations (M_L=0%; M_F=0% and N_C=100) has no missing data so is equivalent to the "complete" matrix, thus we have one effectively complete matrix in our 125 "missing-data" matrices. In practice, we first removed the data following the two missing data parameters M_L and M_F and then removed data following the N_C parameters. To avoid avoid matrices containing taxa

without any data (morphological or molecular), we repeated the random deletion until the matrices contained at least 5% of data for any taxon. Note that the living taxa always had at least 90% of data (the 1000 molecular characters).

2.2.3 *Estimating phylogenies*

From the resulting matrices we generated two types of trees: the "best" tree inferred from the "complete" matrix and the "missing-data" trees inferred from the 125 matrices with various amounts of missing data. The "true" tree was used to generate the "complete" matrix and reflects the "true" evolutionary history in our simulations. The "best" tree, on the other hand, is the best tree we can build using state-of-the-art phylogenetic methods. In real world situations, the "true" tree is never available to us because we cannot know the true evolutionary history of a clade (except in very rare circumstances, e.g. Rozen et al., 2005). We compare "best" trees to "missing data" trees but could also compare "true" trees to the "missing data" trees. In practice, the difference between the "best" trees and the "missing data" trees represents the effect of our missing data parameters and of the phylogenetic methods used to infer the "missing data" trees. The difference between the "true" and the "missing data" trees, however, represents the effect of our parameters used to generate the "true" tree and the algorithms used to generate the "complete" matrix as well as the effect of our missing data parameters and the phylogenetic methods used. Because the main aim of this study is to look at the effect our missing data parameters on topological recovery, we chose to represent only the comparisons between the "best" trees and "missing data" trees. The results of the comparisons of the "true" tree and the "missing data" trees are available in Appendix B. Note that this makes little difference to our overall results.

MAXIMUM LIKELIHOOD — The "best" tree and the "missing-data" trees were inferred using RAxML v. 8.0.20 (Stamatakis, 2014). For the molecular data, we used the GTR + Γ_4 model (Tavaré, 1986; default GTRGAMMA in RAxML v. 8.0.20; Stamatakis, 2014). For the morphological data, we used the M*kv* model (Lewis, 2001) assuming an equal state frequency and a unique overall substitution rate (μ) following a gamma distribution of the rate variation with four distinct categories (M*kv* + Γ_4; -K MK option in RAxML v. 8.0.20; Stamatakis, 2014). We used RAxML because it automatically corrects for acquisition bias (Lewis, 2001). It is also heavily used in the literature for Maximum Likelihood tree inference (e.g. Roure and Philippe, 2011;

Bogdanowicz et al., 2012; Springer et al., 2012; O'Leary et al., 2013; Kelly et al., 2014) and is one of the fastest methods available (Stamatakis et al., 2008).

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

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

Each Bayesian tree was estimated using two runs of four chains each for a maximum of 5×10^7 generations. For each estimation, we used the "true" tree's topology as a starting tree (with a starting value for each branch length of one). We used a fixed starting tree rather than a random starting tree (default MrBayes; Ronquist et al., 2012b) to speed up our Bayesian inferences. To assess if this had

an effect on the topology of the "best" tree, we ran a sub-sample of trees using a different random starting tree for the two MCMC chains (default MrBayes option; Ronquist et al., 2012b). We tested this effect on five trees with the five levels of missing data (i.e. first tree: M_L=0%, M_F=0% and N_C=100 (i.e. 0% "missing"); second tree: M_L=10%, M_F=10% and N_C=90, etc.) on the first 20 simulation chains. We then compared the trees inferred using a random starting tree to the "best" tree using the normalised Robinson-Foulds and Triplets metrics in an identical way as described below (Fig. 2.3).

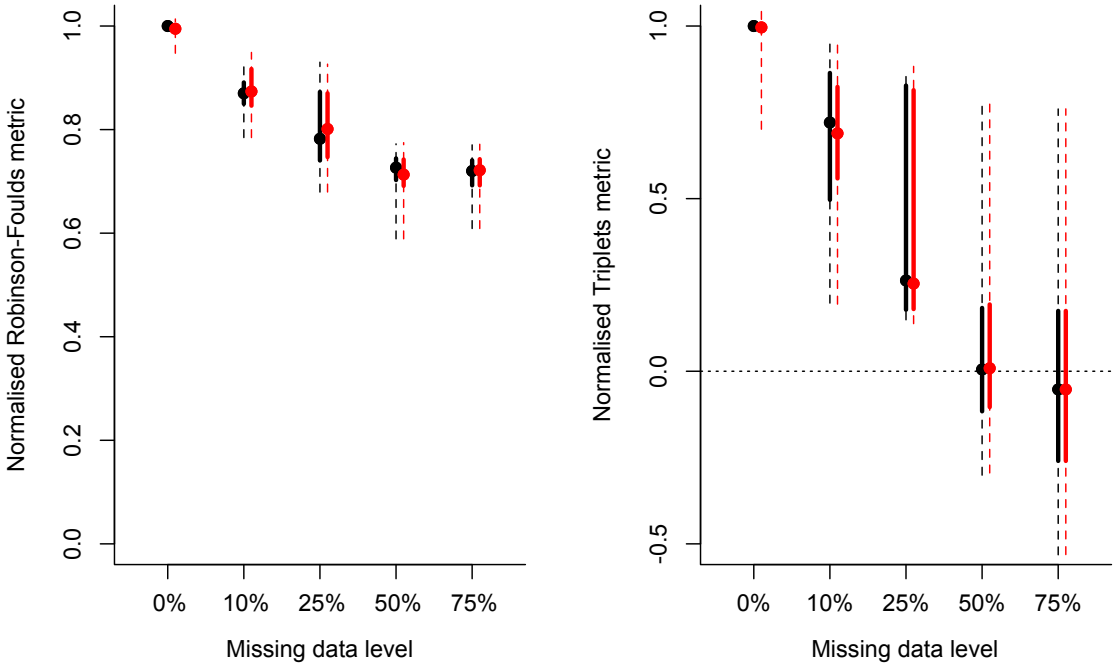

FIGURE 2.3: Effect of using the "true" tree (black) or a random tree (red) as the starting tree for the Bayesian inference. The x axis, represents the amount of missing data (see below).

We used a two-way ANOVA to test any significant effect of the starting tree ("true" or random) on the normalised Robinson-Foulds and Triplets metrics. We found no significant effect of using the "true" instead of a random tree as a starting tree on our ability to recover the "best" tree (Table 2.1).

Note that these results are not surprising since a starting tree is not a Bayesian prior on topology *per se*.

Additionally, we used two priors on the molecular part of the matrix: an exponential prior on the shape of the gamma distribution of α = 0.5, and a transition/transversion ratio prior of two sampled from a strong beta distribution (β(80,40));

TABLE 2.1: Test of the effect of using either a random tree or the "true" tree as a starting tree on two Normalised Robinson-Foulds (RF) and Triplets (Tr) metrics using a two-way ANOVA.

metric	terms	Df	Sum Sq	Mean Sq	F value	Pr(>F)
RF	starting	1	0.00	0.00	0.01	0.9125
	Residuals	198	2.97	0.01		
Tr	starting	1	0.01	0.01	0.07	0.7887
	Residuals	198	34.57	0.17		

and one prior on the morphological part of the matrix (exponential prior on the shape of the gamma distribution of α = 0.5). We used these priors to speed up the Bayesian estimation process. These priors biased the way the Bayesian process calculated branch lengths by giving non-random starting points and boundaries for parameter estimation however, here we are focusing on the effect of missing data on tree topology and not branch lengths. Even using these priors, it took 140 CPU years to build 50 sets of 125 Bayesian trees (2.30GHz clock speed nodes). The detailed MrBayes parameters are available in Appendix A. We also included an analysis showing the effect of missing data on the estimation of the shape parameter (α) of the morphological substitution rate distribution. This extra analysis, however, is beyond the scope of this paper so the results are not discussed further here.

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

2.2.4 *Comparing topologies*

We compared the topology of the "missing-data" trees to the "best" tree to measure the effect of the three parameters M_L, M_F and N_C on tree topology. We used the Robinson-Foulds distance (Robinson and Foulds, 1981) to assess the number of conserved clade positions and the Triplets distance (Dobson, 1975) to assess the number of wildcard taxa (i.e. taxa that frequently change position in different trees Kearney, 2002). We used these two metrics because they illustrate two different aspects of tree topology (see Discussion) but also because their performance in

measuring differences in topology is well described (Kuhner and Yamato, 2014) and well implemented (Bogdanowicz et al., 2012). We normalised both metrics using methods described in Bogdanowicz et al. (2012) to generalize our results for any *n* number of taxa. These metrics are described in detail below.

ROBINSON-FOULDS DISTANCE — The Robinson-Foulds distance (Robinson and Foulds, 1981), or “path difference", measures the difference between the number of clades and twice the number of shared clades across two trees. The metric reflects the distance between the distributions of tips among clades in the two trees (Robinson and Foulds, 1981):

$$RF_{x,y} = N_x + N_y - 2C_{x,y} \tag{2.1}$$

where $C_{x,y}$ is the number of clades in common in the two trees. C is equal to one if the two trees have the same n taxa; and $C = n - 2$ when none of the n taxa are shared between the trees. This metric is bounded between zero, when the two trees are identical, and $2(n-2)$ (for two trees with n taxa) when there is no shared clade in the two trees. This metric is sensitive to minor changes in clade conservation: if the trees are composed of two clades of three taxa *(((a,b),c),((d,e),f))*, the swapping of any two taxa will lead to a maximal score of the Robinson-Foulds distance indicating poor tree similarity.

We normalised this metric following Bogdanowicz’s Normalised Tree Similarity (NTS) method (Bogdanowicz et al., 2012). For any tree with *n* taxa compared using a tree difference metric m, Normalized Tree Similarity, NTS_m, represents the similarity score for the two trees given the expected difference between 1000 random Yule trees (Bogdanowicz et al., 2012) with n taxa. If $\bar{d}_{m,n}$*(rand)* is the average difference between two random Yule trees with n taxa and $d_{m,n}$*(x,y)* the difference between the two trees *x* and *y* each containing n taxa, then:

$$NTS_{m,n}(x,y) = \frac{\bar{d}_{m,n}(rand) - d_{m,n}(x,y)}{\bar{d}_{m,n}(rand)} \tag{2.2}$$

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

This method is a generalisation of the topological accuracy method (Price et al., 2010) allowing to compare topological differences between any tree with any tree comparison metric. In practice when the Normalised Robinson-Foulds metric between two trees is equal to one, the trees are identical; if the metric is equal to zero, the trees are no more different than expected by chance; finally if the metric is less than zero, the trees are more different than expected by chance. Note that once rescaled, the Normalised Robinson-Foulds metric is a measure of similarity, rather than of distance like the original Robinson-Foulds metric.

TRIPLETS DISTANCE — The Triplets distance (Dobson, 1975) measures the number of sub-trees made up of three taxa that differ between two trees (Critchlow et al., 1996):

$$S_n = \sum_{ijk} I_{ijk} \tag{2.3}$$

where:

$$\sum_{ijk} = \binom{n}{4} = \frac{n!}{4!(n-4)!} \tag{2.4}$$

and where *n* is the total number of taxa in both trees (modified from Critchlow et al. (1996)). If S_n = 0, the trees are identical; when S_n = $\binom{n}{4}$, the trees are as different as possible (i.e. every taxon has a different placement in the two trees). This metric measures the position of each taxon and clade in relation to its closest neighbours. It is bounded between zero when the two trees are identical and $\binom{n}{3}$ (for two trees with n taxa) when there is no shared taxa/clade position in the two trees. Therefore this metric is sensitive to the conservation of wildcard taxa. We normalised this metric in the same way as for the Robinson-Foulds distance resulting in the Normalised Triplets metric.

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

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

2.2.5 *Testing the effects of the missing data parameters on topological recovery*

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

We measured similarities among the distributions of the different metrics scores (Normalised Robinson-Foulds and Normalised Triplets metric) using the Bhattacharyya Coefficient (Bhattacharyya, 1943). The Bhattacharyya Coefficient is the probability of overlap between two distributions bounded between 0 (no overlap) and 1 (Bhattacharyya, 1943, full overlap;). The coefficient is calculated as the sum of the square root of the relative counts shared in *n* bins among two distributions.

$$\text{Bhattacharyya Coefficient} = \sum_{i=1}^{n} \sqrt{\sum a_i \times \sum b_i} \tag{2.5}$$

where

$$a_i = \frac{\text{Number of counts in bin } i \text{ for the distribution } a}{\text{Total number of counts for the distribution } a} \tag{2.6}$$

and

$$b_i = \frac{\text{Number of counts in bin } i \text{ for the distribution } b}{\text{Total number of counts for the distribution } b} \tag{2.7}$$

The precision of the Bhattacharyya Coefficient is directly related to the number of bins, n. If n is low, the overlap will be overestimated and if n is too high, the overlap will be underestimated. In this analysis, we determined the number of bins using Silverman’s rule of thumb which states that n should be 0.9 times the minimum

of the standard deviation and the interquartile range of the distribution, divided by 1.34 times the sample size of the distribution to the negative one-fifth power (`bw.nrd0()` function in R (Silverman, 1986)). When the Bhattacharyya Coefficient between two distributions is $<$0.05, the distributions are significantly different. When this coefficient is $>$0.95 both distributions are significantly similar. Values in between these two threshold just show the probability of overlap between the distributions but are not conclusive to assess the similarity or differences between the distributions.

Note that this is comparable to performing a two-sided t-test, but we use the Bhattacharyya Coefficient here because we are comparing whole distributions not just their means. When the Bhattacharyya Coefficient between two distributions is $<$0.05, the distributions are significantly different. When this coefficient is $>$0.95, the distributions are significantly similar. Values between these two thresholds show the probability of overlap between the distributions but do not allow us to define the significance of the similarity or differences between distributions. To assess the effect of our missing data parameters, we calculated the Bhattacharyya Coefficient between the distributions of the different metrics scores (Normalised Robinson-Foulds and Normalised Triplets metric) for each pairwise combination of missing data parameters (M_L, M_F, N_C) and parameter states (0%, 10%, 25%, 50%, 75% and 100, 90, 75, 50, 25 characters), i.e. M_L = 0%, M_F = 0%, N_C = 100; M_L = 10%, M_F = 0%, N_C = 100 etc. (see Fig. 2.4 for more details). This resulted in 7875 pairwise comparisons (a triangular matrix with $3^5 \times 3^5$ cells). We performed this procedure separately for each tree inference and uncertainty method. When two combinations of missing data parameters have a similar ability to recover the "best" tree topology the Bhattacharyya Coefficient will be close to one. Conversely, if the two combinations of missing data parameters differ, the Bhattacharyya Coefficient will be close to zero. Because of the difficulties in representing so many pairwise comparisons in a meaningful way, we summarized these results as a heat map of Bhattacharyya Coefficients (see Fig. 2.8). In this type of figure, parameters that have similar effects on recovering the "best" topology (either positive or negative effects) will be denoted by similar colour patches in the heat map representation of these comparisons (see Fig. 2.8).

To assess the effect of the different tree inference and uncertainty methods (i.e. Maximum Likelihood, Bayesian consensus, Maximum Likelihood bootstrap trees and Bayesian posterior tree distribution) on our ability to recover the "best" tree topology, we calculated the Bhattacharyya Coefficient between the distributions of

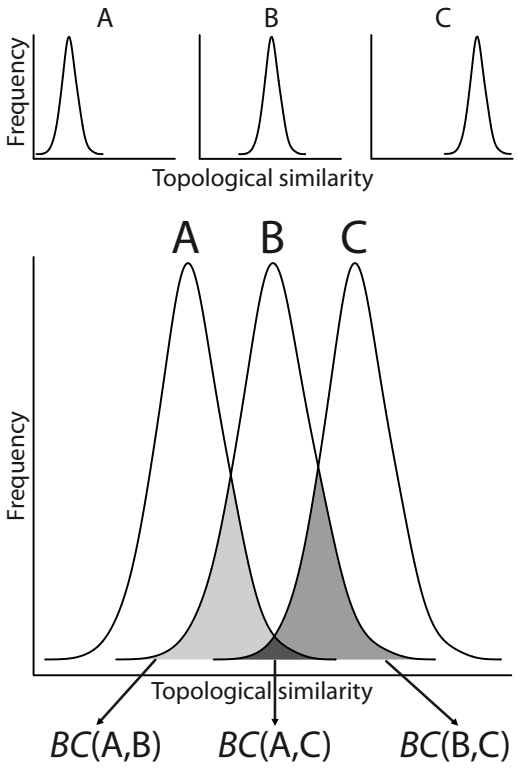

FIGURE 2.4: **Bhattacharyya Coefficient calculation outline 1.** A, B and C are distributions of tree similarity metrics (Normalised Robinson-Foulds or Normalised Triplets metrics) for any combination of missing data parameters (e.g. M_L = 10%, M_F = 50%, N_C = 25). The Bhattacharyya Coefficient (BC) is the overlap of the distribution of tree similarity metrics between two combinations of missing data parameters, for example, BC(A,B) is the probability of overlap between the distributions A and B.

the different metrics scores (Normalised Robinson-Foulds and Normalised Triplets metric) for each pairwise combination of tree inference and uncertainty methods, i.e. Maximum Likelihood *versus* Bayesian consensus; Maximum Likelihood *versus* Maximum Likelihood bootstrap trees etc. (see Fig. 2.5 for more details). Note that this procedure pools results from across all missing data parameter combinations so it results in just six pairwise comparisons. When two tree inference or uncertainty methods have a similar ability to recover the "best" tree topology the Bhattacharyya Coefficient will be close to one. Conversely, if the two tree inference or uncertainty methods differ, the Bhattacharyya Coefficient will be close to zero.

2.3 RESULTS

As the amount of missing data in the morphological part of the Total Evidence matrix increases, our ability to recover the "best" tree topology decreases, regardless of the

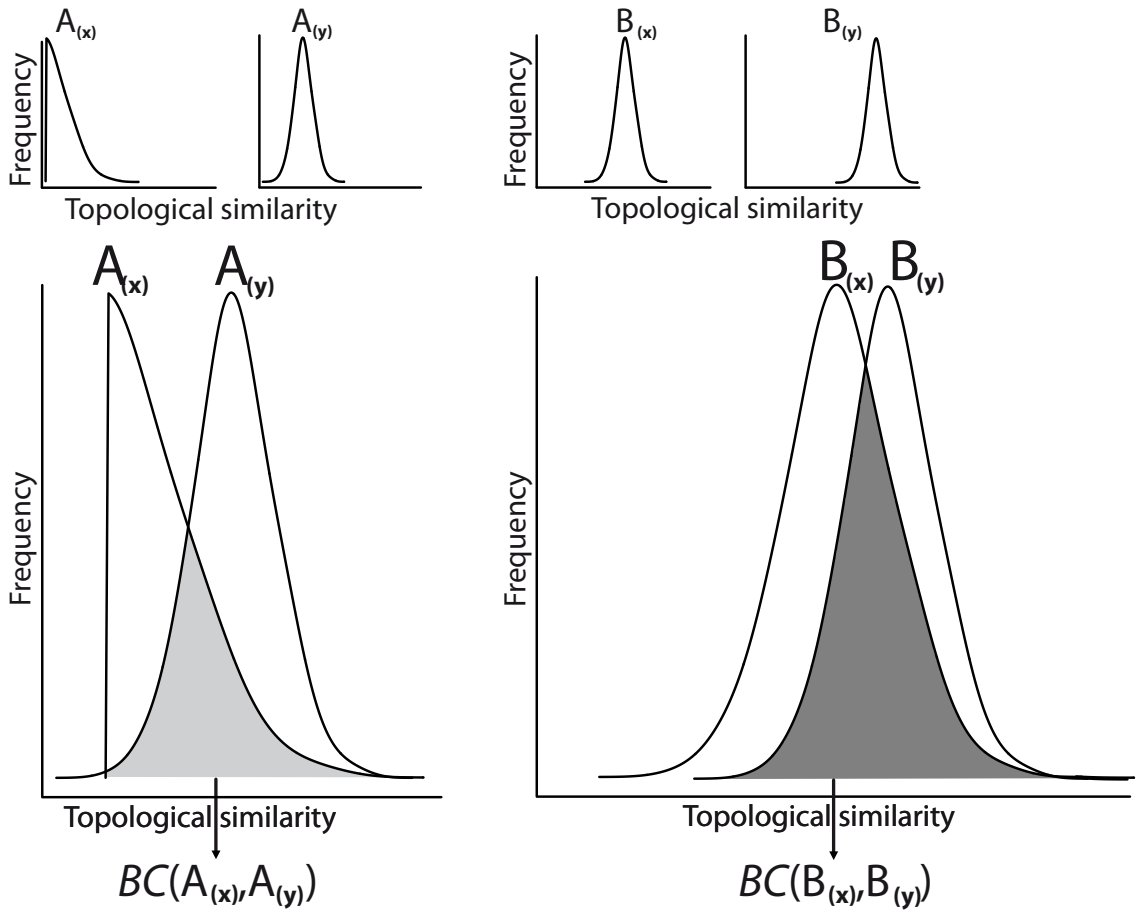

FIGURE 2.5: **Bhattacharyya Coefficient calculation outline 2.** A and B are distributions of tree similarity metrics (Normalised Robinson-Foulds or Normalised Triplets metrics) for any combination of missing data parameters (e.g. M_L = 10%, M_F = 50%, N_C = 25). **(x)** and **(y)** are two different tree inference methods (e.g. Maximum Likelihood or Bayesian). The Bhattacharyya Coefficient (BC) is the overlap of the distribution of tree similarity metrics between two methods for the same combination of missing data parameters, for example, BC(A_x,A_y) is the probability of overlap of the distribution A for methods x and y.

missing data parameter (M_L, M_F or N_C), the tree inference method (Maximum Likelihood or Bayesian) or the tree comparison metric used (Normalised Robinson-Foulds or Normalised Triplets metric). Nonetheless, the different missing data parameters and tree inference methods do not affect the topology in the same way (Fig. 2.6 and Fig. 2.7).

2.3.1 *Individual effects of missing data parameters*

As the amount of missing data increases across all three parameters, our ability to recover the "best" tree topology decreases (Fig. 2.6). The Normalised Robinson-

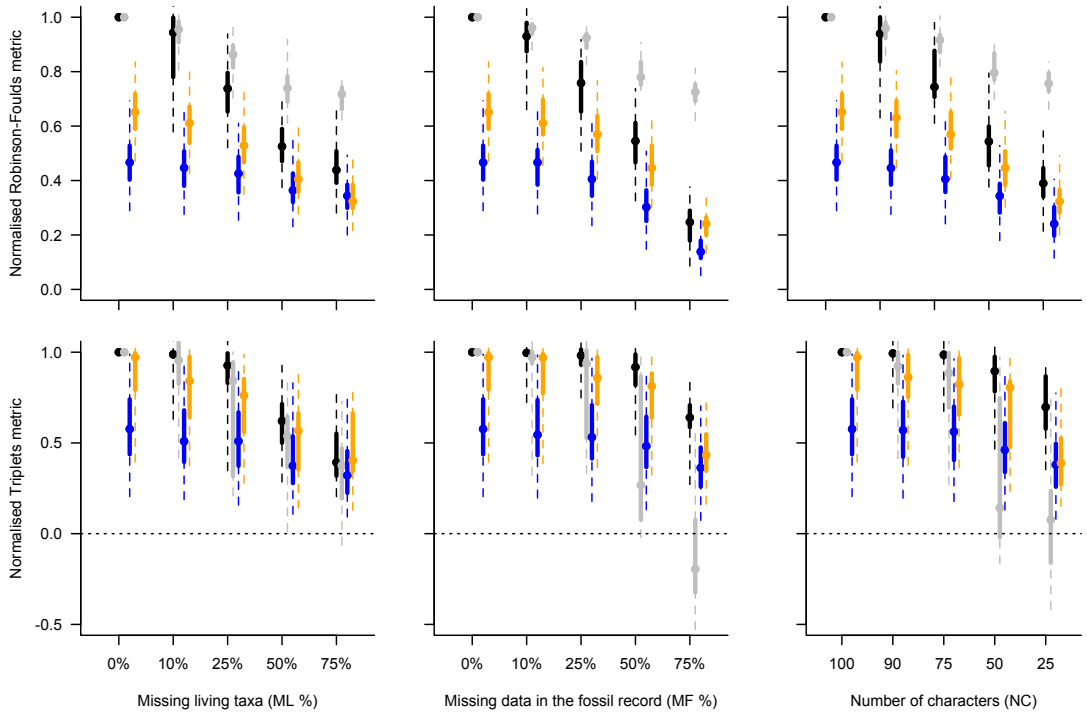

FIGURE 2.6: **The effects of increasing missing data on topological recovery** using Maximum Likelihood trees (black), Bayesian consensus trees (grey), Maximum Likelihood bootstrap trees (blue) and Bayesian posterior tree distributions (orange). The percentage of missing data for each parameter (M_L, M_F and N_C) is shown on the x axis. Topological recovery was measured using two different tree comparison metrics: Normalised Robinson-Foulds metric (upper row) and Normalised Triplets metric (lower row). The graph shows the modal value (points), and the 50% (thick solid lines) and 95% (thin dashed lines) confidence intervals of the distributions of the tree comparison metric for each missing data parameter and tree inference method.

Foulds metric is always lower for the Maximum Likelihood trees than for the Bayesian consensus trees (median Bhattacharrya Coefficient = 0.69, 0.48 and 0.66 for M_L, M_F and N_C respectively; Fig. 2.6; Tables C5, C6 and C7 in Appendix C). The Normalised Triplets metric, however, is similar when comparing the Maximum Likelihood trees and the Bayesian consensus trees for all the parameters (M_L, M_F and N_C) (median Bhattacharrya Coefficient = 0.84, 0.75 and 0.80 for M_L, M_F and N_C respectively; Fig. 2.6; Tables C5, C6 and C7 in Appendix C).

2.3.2 *Combined effect of missing data parameters*

As expected, our ability to recover the "best" tree topology is worst when each parameter contains the maximum amount of missing data (i.e. M_L = 75%, M_F =

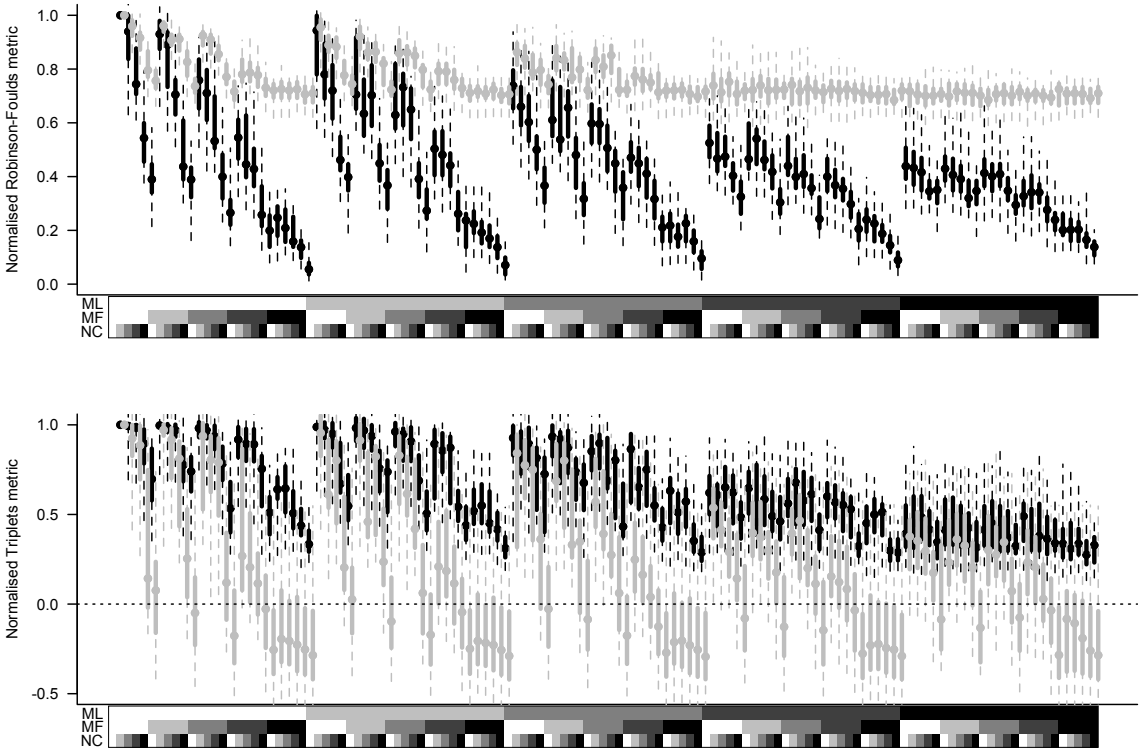

FIGURE 2.7: **The effects of increasing missing data on topological recovery using Maximum Likelihood trees (black) and Bayesian consensus trees (grey).** The x axis shows the percentage of missing data from 0% (white) to 75% (black) for the two parameters: M_L (upper line), M_F (middle line) and number of characters from 100 to 25 for the parameter N_C (lower line). Topological recovery was measured using two different tree comparison metrics: Normalised Robinson-Foulds metric (upper row) and Normalised Triplets metric (lower row). The graph shows the modal value (points), and the 50% (thick solid lines) and 95% (thin dashed lines) confidence intervals of the distributions of the tree comparison metric for each missing data parameter and tree inference method.

75% and N_C = 75%), and best when there is no missing data (i.e. M_L = 0%, M_F = 0%, N_C = 0%; Fig. 2.7; Tables C2, C3 and C4 in Appendix C). Fig. 2.8 shows the similarity of distributions of tree metrics in a triangular matrix with the values of each pairwise Bhattacharyya Coefficient coloured according to their values (orange when the distributions overlap completely, Bhattacharyya Coefficient = 1, and blue when they do not, Bhattacharyya Coefficient = 0).

Using both Normalised Robinson-Foulds and Normalised Triplets metrics from the Bayesian consensus trees, the parameter combination with no missing data (i.e. M_L = 0%, M_F = 0%, N_C = 100) is always the most dissimilar to all the other parameter combinations (thin deep blue line at the base of Fig. 2.8). The Normalised Robinson-Foulds metric (median Bhattacharrya coefficient = 0.79; blue regions in Fig. 2.8A), however, displays more dissimilarities than the Normalised Triplets metric

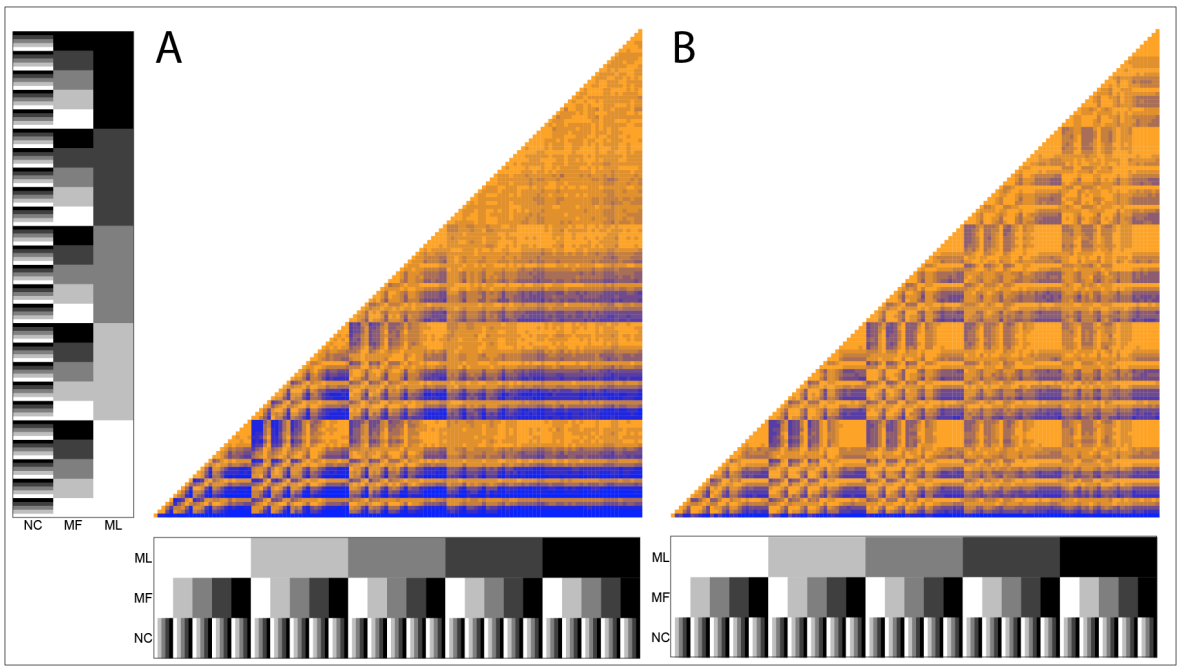

FIGURE 2.8: **The effects of missing data on topological recovery using Bayesian consensus trees.** Both axes show the percentage of missing data from 0% (white) to 75% (black) for the three parameters: M_L (upper line), M_F (middle line) and N_C (lower line). The topological recovery is measured as (A) the Normalised Robinson-Foulds metric and (B) the Normalised Triplets metric calculated using the Bhattacharyya Coefficient. The Bhattacharyya Coefficient values are indicated using a color gradient ranging from low probability of overlap in blue, to high probability of overlap in orange. Blue regions denote a poor overlap in Normalised metric values between the different parameter combinations (i.e. the parameters have a strong effect on the metric and thus the topological recovery). Conversely, orange regions denote a high overlap in Normalised metric values between the different parameter combinations (i.e. the parameters have a weak effect on the metric and thus the topological recovery).

(median Bhattacharrya coefficient = 0.81; blue regions in Fig. 2.8B). The orange upper triangle in Fig. 2.8A shows a high probability of overlap of the Normalised Robinson-Foulds metric for the trees with the M_L parameter $\geq$ 50% (Fig. 2.8A). Once $M_L \geq$ 50%, there is no additional effect of M_F and N_C, regardless of the amount of missing data in these parameters (Fig. 2.8A). Likewise, once $N_C < 50$, there is no additional effect of M_L and M_F as denoted by the high probability of Normalised Robinson-Foulds metric overlap (horizontal orange stripes between the blue regions Fig. 2.8A). In Fig. 2.7 for the Normalised Robinson-Foulds metric, this can be interpreted as the overlap between the distributions once M_L=50%.

For all combinations of missing data parameters and tree comparison metrics, the Maximum Likelihood bootstrap trees and the Bayesian posterior tree distributions perform very similarly (median Bhattacharrya Coefficient = 0.85 and 0.98, using Normalised Robinson-Foulds metric or Normalised Triplets metric respectively;

Table 2.2). These two methods, however, perform worse than the Bayesian consensus trees using Normalised Robinson-Foulds metric (median Bhattacharrya Coefficient = 0 and 0.01, for the Maximum Likelihood bootstrap trees and the Bayesian posterior tree distribution respectively; Table 2.2; Fig. 2.6 and Fig. C2 in Appendix C).

TABLE 2.2: Bhattacharyya Coefficients of the pairwise method comparisons. Each line summarizes the probabilities of overlap between the distributions of the "best" tree versus trees from each inference method (Maximum Likelihood; Bayesian consensus; Maximum Likelihood Bootstraps and Bayesian posterior trees) pooled across all combinations of missing data parameter values, using the Normalised Robinson-Foulds (RF) and Triplets (Tr) metrics. Values highlighted in bold are the extreme values of high or low probability of overlap between two methods. If two methods have a high probability of overlap, they have a similar ability to recover the "correct" tree topology. Values > 0.95 denote significantly similar distributions and values < 0.05 denote significantly different distributions.

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

2.4 DISCUSSION

Our results show that the ability to recover the "best" tree topology in a Total Evidence framework decreases as the amount of missing data increases, regardless of how data were removed or the method of tree inference used. These factors, however, affected topological recovery in different ways and to different extents. Decreasing the number of living taxa with morphological data (M_L) and the overall number of morphological characters in the matrix (N_C) had worst effects on topological recovery (Fig. 2.8). Additionally, using Bayesian consensus trees recovered the "best" tree topology more consistently than using Maximum Likelihood trees or Bayesian posterior tree distributions (Fig. 2.7, Fig. 2.8, Table 2.2). As seen in previous studies, our results show that the amount of missing data are not a problem *per se* for Total Evidence methods, as long as enough living and fossil taxa in the matrix have data for overlapping morphological characters (e.g. Kearney, 2002; Wiens, 2003; Roure and Philippe, 2011; Pattinson et al., 2014).

2.4.1 *Individual effects of missing data parameters*

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

MISSING DATA FOR FOSSIL TAXA (M_F) — When the overall proportion of data for the fossil taxa (M_F) decreases, this also reduces the probability of morphological characters for fossil taxa overlapping with the ones for living taxa. This can lead to difficulties for the placement of certain taxa in the tree. It is important, however, to note that even though the number of displaced wildcard taxa increases (i.e. decrease of Normalised Triplets metric) with increasing missing data in this parameter, clade conservation (i.e. Normalised Robinson-Foulds metric) is still relatively good

(mode = 0.72) when the proportion of missing data are high (M_F = 75%). These results are in agreement with Manos et al. (2007) where as few as 16 characters were sufficient for correctly assigning artificial fossils to their correct clade.

The effect of the missing data in the fossil record (M_F) is less than the effect of the M_L parameter on clade conservation (Normalised Robinson-Foulds metric) but greater on the displacement of wildcard taxa (Normalised Triplets metric; Fig. 2.6 and Fig. 2.7). This is related to the fact that the Bayesian consensus tree is built using a majority consensus rule. When the fossil taxa have less data (e.g. M_F = 75%) they will tend to branch with any taxon in the clade that shares most characters with the fossils. Therefore a majority consensus position is unlikely to exist (i.e. every branching position is represented in $<$ 50% of the trees in the Bayesian posterior distribution) and the fossil taxa will form a polytomy at the base of the clade. In this case, the Normalised Robinson-Foulds metric will decrease when the fossil is present near the tips but affects the clade conservation less when fossils are near the root. Conversely, because a fossil in a high taxonomic level clade has many chances to branch on different nodes within the clade, it will be more likely to act as a wildcard taxon and decrease the Normalised Triplets metric. Therefore, the M_F parameter is likely to affect the Normalised Robinson-Foulds metric less than the Normalised Triplets metric for the Bayesian consensus trees. Conversely, the same scenario in a Maximum Likelihood framework will lead to a dichotomous branching of the fossils but with low bootstrap support ($<$ 50). In other words, the Bayesian consensus tree allows a fossil taxon with few data to be placed with a higher confidence at a lower taxonomic level than the Maximum Likelihood tree, where the fossil will be placed with lower confidence at a higher taxonomic level. We argue that using the Bayesian consensus tree topology is preferable because it is more conservative (e.g. Pattinson et al., 2014).

NUMBER OF MORPHOLOGICAL CHARACTERS (N_C) — Reducing the overall number of morphological characters reduces the probability of their overlap among the taxa in the matrix, and therefore decreases our ability to recover the “best” tree topology. We expected the decrease in this parameter to have an effect twice as large as that for the M_L and M_F parameters, because removing 10% of the data for the fossil or living taxa only removes 5% of data from the whole matrix (because this parameter affects only half of the taxa present in the matrix). Conversely, removing 10% of morphological characters (i.e. N_C = 90) genuinely removes 10% of data in the matrix.

Nonetheless, the effect of removing characters on the ability to recover the “best” tree topology is of the same order of magnitude as for the other two parameters (Fig. 2.6). We suspect this again reflects the importance of overlapping characters, as opposed to the number of characters *per se*.

Additionally, the number of morphological characters determines the size of the matrix. This can affect our ability to recover the “best” tree topology through: (1) the incongruence of phylogenetic signal among morphological and molecular data; and/or (2) homoplasy. The incongruence of phylogenetic signal between morphological and molecular data has previously been demonstrated to be more important in small morphological matrices (Bremer and Struwe 1992; Patterson et al. 1993; see Masters and Brothers 2002 for an empirical example). The sizes of our data matrices were constrained by the performance of our protocol: to reduce the computational time of our analysis to a reasonable level (150 CPU years), we ran our simulations on modestly-sized matrices of 1000 molecular characters and 100 morphological characters. Therefore, part of the decrease of the Normalised Robinson-Foulds metric and the Normalised Triplets metric in our simulations could be due to conflicting phylogenetic signal among morphological and molecular data in our matrices (Fig. 2.6 and Fig. 2.6). Although these matrices are an order of magnitude smaller than some published matrices (e.g. Springer et al., 2012; **?**), they are still within the size range of more modestly-sized empirical matrices (e.g. Kelly et al., 2014; Sallam et al., 2011). Therefore, our simulations reflect realistic parameters. Nonetheless, the use of probabilistic methods (i.e. Maximum Likelihood or Bayesian) and the M*kv* model (Lewis, 2001) has been previously demonstrated to partially resolve this issue (Wright and Hillis, 2014).

2.4.2 *Combined effect of missing data parameters*

As expected, when combining the missing data parameters, our ability to recover the “best” tree topology is affected in the same way as for the parameters individually: the Normalised Robinson-Foulds metric and the Normalised Triplets metric are higher when all the missing data parameters have few missing data (i.e. M_L = 0%, M_F = 0%, N_C = 100) and lower when they have a larger proportion of missing data (i.e. M_L = 75%, M_F = 75% and N_C = 25; Fig. 2.7). It is important, however, to notice that the effect of each parameter is not additive. Surprisingly, the number of missing living taxa with morphological data (M_L) and the overall number of missing

morphological characters (N_C), have a bigger effect than the amount of missing data for the fossil taxa (M_F). For any additional missing living taxa with morphological data (M_L) beyond 50%, there is no difference among trees with any combination of the other parameters (M_F and N_C; Fig. 2.8). In other words, when the number of missing living taxa reaches 50%, neither the amount of missing data in the fossil record (M_F), nor the number of characters in the matrix (N_C) affect topology. A similar effect can be observed when the N_C parameter reaches 50 characters (Fig. 2.8). This has important practical implications, especially for the best strategy to improve topology by collecting more morphological data (see below).

2.4.3 *Effects of tree inference methods*

Variation in our ability to recover the "best" tree topology depends heavily on the tree inference method (Fig. 2.6 and Fig. 2.7). For morphological data, previous studies have shown some superiority of probabilistic tree inference methods with simple evolutionary models such as the M*kv* model (Lewis, 2001) over parsimony methods (Wright and Hillis, 2014; but see Spencer and Wilberg, 2013). This is, however, the first study, to our knowledge, to compare the performance of the M*kv* model (Lewis, 2001) for recovering the "best" tree topology using Maximum Likelihood and Bayesian methods in a Total Evidence framework. Our results show that the topology of the Bayesian consensus tree is always closer to the "best" tree topology than the "best" Maximum Likelihood tree (Fig. 2.7). Note that the methodological choice of using the "true" tree as a starting tree for the Bayesian Inference rather than a random starting tree (see Methods), had no significant effect on topological recovery (see Appendix A, section "Effect of the starting tree on Bayesian inference" for details). As described above, this is because the Bayesian consensus tree allows a fossil taxon with few data to be placed with a higher confidence at a lower taxonomic level than the Maximum Likelihood tree. This may also be because the "best" Bayesian consensus trees are not completely resolved, thus will always be more similar to the "missing data" trees than a completely resolved tree like the "best" Maximum Likelihood tree. Nonetheless, we minimized the probability of unresolved "best" trees in our Bayesian analyses by only using datasets with strong phylogenetic signal (see Methods).

The Bayesian consensus trees, however, perform poorly for the Normalised Triplets metric: some parameter combinations, especially when the M_F parameter reaches

75% missing data, lead to negative values (Fig. 2.7). A Normalised Triplets metric value below 0 means that the placement of some taxa is worse than expected by just randomly placing this taxon in the tree. This can be interpreted as the absence of comparable triplets between some of the “missing data" trees and “best” trees. Even if clades are conserved (Fig. 2.7), the resolution within them can be poor to non-existent when a large proportion of data are missing (i.e. 75%). In such cases, the fossil taxa are equally likely to be placed in any of the clades that they share the most characters with. These results are in agreement with previous studies that have showed that missing data can cause problems for recovering “correct” topologies, especially for small matrices of 100 characters (Wiens, 2003). It is important to note, however, that this effect can be reduced by increasing the number of characters (Wiens, 2003).

It is also worth noting that across all our analyses, the topologies of the Maximum Likelihood bootstrap trees and the Bayesian posterior trees distribution were always further from the “best” tree topology than Maximum Likelihood and Bayesian consensus trees. This was true even when no morphological data were missing (M_L = 0%; M_F = 0%, N_C = 100; Fig. 2.6). This reflects the fact that it is difficult to compare two distributions of trees, and each comparison between a set of “missing data" trees and a set of the “best” trees involved 1000 random pairwise comparisons rather than just one. Additionally, the Bayesian posterior trees performed more poorly than the Bayesian consensus tree (Fig. 2.6, Table 2.2 and Appendix C Fig. C5 and Tables C5, C6 and C7). This may be because the Bayesian posterior trees are always resolved and thus more likely to contain incorrectly resolved nodes (i.e. decreasing the Normalised Robinson-Foulds metric). Conversely, the Bayesian consensus trees might not resolve nodes that are poorly supported and thus are more likely to contain only correctly resolved nodes (i.e. increasing the Normalised Robinson-Foulds metric).

2.4.4 *Practical implications*

Our missing data parameters illustrate different sources of missing data in empirical matrices as follows: (M_L) the paucity of coded morphological characters for living taxa; (M_F) the missing data for fossils (or parts of fossils) that have not been preserved in the fossil record; and (N_C) characters that have not been coded across living and fossil species, perhaps due to difficulties in coding or poor preservation

of the feature in collections. Filling these gaps in empirical Total Evidence matrices should lead to a substantial increase in our ability to recover the "best" tree topology. We can increase the number of living taxa with coded morphological characters by increasing research efforts in this area, and encouraging use of our vast natural history collections. Increasing data for fossil species is harder, since it depends on fossil preservation biases and new fossil discoveries. Gaps in the matrix, however, can be filled with efforts in palaeontological field work that can potentially lead to future discoveries of exceptionally preserved fossils (e.g. **?**). Fortunately, although these data are the most difficult to collect, they also have the least influence on whether our simulations recover the "best" tree topology (Fig. 2.8). Finally, although increasing the number of coded characters is relatively straightforward, the amount of time it takes to build a morphological matrix increases directly with the number of characters involved. One solution to this problem may be to engage with collaborative data collection projects through web portals such as *MorphoBank* (O'Leary and Kaufman, 2011), so that no single individual collects all the data.

Another practical implication of our results regards the tree inference methods. Because the Bayesian consensus trees consistently recovered topologies closer to the "best" tree topology than the Maximum Likelihood trees, we advise that where a topological constraint is needed, Bayesian consensus trees should be used. This may apply to tree inferences using the Total Evidence method such as tip-dating (e.g. Ronquist et al., 2012a; Wood et al., 2013; Matzke, 2014). It is, however, possible that including dating information during tree inference could also improve the accuracy of the Bayesian posterior tree distribution, so a fixed topology should be used with caution. Using the Bayesian consensus tree rather than the Maximum Likelihood can also reduce the number of false positive topologies (*sensu* **?**). As shown in Fig. 2.7 and discussed in the section above (Effects of tree inference methods), the Bayesian consensus tree is more likely to not resolve poorly supported nodes due to missing data than the Maximum Likelihood tree that is more likely to incorrectly resolve such nodes (i.e. creating a false positive node). Note, however, that we do not suggest discarding the Bayesian posterior tree distributions even though they performed poorly in recovering the "best" tree topology in our simulations (this can probably be traced to the difficulties comparing distributions of trees; see above). These trees will be invaluable for phylogenetic comparative analyses. For example a sub-sample of posterior tree distributions can be used to assess macroecological

questions while better taking into account topological uncertainty (e.g. Fritz et al. 2013 and Jetz et al. 2012 used in Healy et al. 2014).

2.5 CONCLUSIONS

Previous studies have explored the effect of missing morphological data in Total Evidence matrices (Wiens et al., 2005; Manos et al., 2007; Pattinson et al., 2014). The conclusions of theses studies, however, were limited by their empirical approach making their results applicable only to similar missing data scenarios. Additionally, these studies focused either only on living taxa (Wiens et al., 2005) or on the patterns of missing data from the fossil record only (Manos et al., 2007; Pattinson et al., 2014). Here we instead used an approach where missing data were generated from simulated data and according to three clearly defined missing-data parameters (M_L, M_F or N_C) that removed data from both the living and fossil taxa. This allowed us to confirm previous results that missing data can be especially problematic in small matrices (Wiens, 2003), but also revealed the crucial importance of coding morphological data for living species in Total Evidence phylogenies. Missing data in Total Evidence matrices is not a problem for recovering the "best" tree topology as long as enough living and fossil taxa in the matrix have data for overlapping morphological characters. When missing data increases in any of our missing data parameters (M_L, M_F or N_C), it reduces support for the placement of fossil taxa and increases the displacement of wildcard taxa. Therefore we advise increased focus on coding morphological characters for a large number of the living taxa present in the matrix (i.e. at least 50%) if the goal is to accurately combine both living and fossil species in phylogenies. Doing so will increase overlap of morphological characters among living and fossil taxa, allowing the fossil taxa to be positioned relative to the living taxa based on their shared derived characters rather than simply on available data.

Additionally, the topologies of the Bayesian consensus trees, regardless of the amount of missing data, were always closer to the "best" tree topology than the Maximum Likelihood trees. This has also been observed in empirical data (e.g. Arcila et al., 2015) where Maximum Likelihood trees inferred from a Total Evidence matrix were less supported than the Bayesian consensus tree. This might have an important impact on estimating topologies in the Total Evidence framework, because previous studies had to rely either on molecular scaffolds (e.g. Slater, 2013),

taxonomic constraints (e.g. Slater, 2013; Beck and Lee, 2014) or even on fixing the topology (e.g. Ronquist et al., 2012a). Therefore, we suggest extracting such topological backbones from the Bayesian consensus tree if needed.

CHAPTER 3

MISSING DATA IN LIVING MAMMALS

Assessment of cladistic data availability for living mammals [2]

ABSTRACT

Analyses of living and fossil taxa are crucial for understanding changes in biodiversity through time. The Total Evidence method allows living and fossil taxa to be combined in phylogenies, by using molecular data for living taxa and morphological data for both living and fossil taxa. With this method, substantial overlap of morphological data among living and fossil taxa is crucial for accurately inferring topology. However, although molecular data for living species is widely available, scientists using and generating morphological data mainly focus on fossils. Therefore, there is a gap in our knowledge of neontological morphological data even in well-studied groups such as mammals.

We investigated the amount of morphological (cladistic) data available for living mammals and how this data was phylogenetically distributed across orders. 22 of 28 mammalian orders have <25% species with available morphological data; this has implications for the accurate placement of fossil taxa, although the issue is less pronounced at higher taxonomic levels. In most orders, species with available data are randomly distributed across the phylogeny, which may reduce the impact of the problem. We suggest that increased morphological data collection efforts for living taxa are needed to produce accurate Total Evidence phylogenies.

[2] A shorter version (2500 words) will be submitted under the same title to Biology Letters as an invited submission for a special issue on phylogenies with living and fossil species. This special issue is open to submission in December 2015. A pre-print is currently available at `http://dx.doi.org/10.1101/022970`. T.G. and N.C. designed the experiments; T.G. ran the analysis and interpreted the results; T.G. and N.C. wrote the manuscripts. *Specific acknowledgements*: thanks to David Bapst, Graeme Lloyd, Nick Matzke and April Wright. *Data availability and reproducibility*: all data and analysis code is available on GitHub (`https://github.com/TGuillerme/Missing_living_mammals`).

3.1 INTRODUCTION

There is an increasing consensus among evolutionary biologists that studying both living and fossil taxa is essential for fully understanding macroevolutionary patterns and processes (Slater and Harmon, 2013; Fritz et al., 2013; Wood et al., 2013). For example, including both living and fossil taxa in evolutionary studies can improve the accuracy of timing diversification events (e.g. Ronquist et al., 2012a), our understanding of relationships among lineages (e.g. Beck and Lee, 2014), and our ability to infer biogeographical patterns through time (e.g. Meseguer et al., 2015). To perform such analyses it is necessary to combine living and fossil taxa in phylogenetic trees. One increasingly popular method, the Total Evidence method (Eernisse and Kluge, 1993; Ronquist et al., 2012a), combines molecular data from living taxa and morphological data from both living and fossil taxa in a supermatrix (e.g. Pyron, 2011; Ronquist et al., 2012a; Schrago et al., 2013; Slater and Harmon, 2013; Beck and Lee, 2014; Meseguer et al., 2015), producing a phylogeny with living and fossil taxa at the tips. These phylogenies can be dated using methods such as tip-dating (Ronquist et al., 2012a; Wood et al., 2013) and incorporated into macroevolutionary studies (e.g. Ronquist et al., 2012a; Wood et al., 2013; Slater, 2013).

A downside of the Total Evidence method is that it requires a lot of data. One must collect molecular data for living taxa and morphological data for both living and fossil taxa; two types of data that require fairly different technical skills (e.g. molecular sequencing *vs.* anatomical description). Additionally, large chunks of this data can be difficult, or even impossible, to collect for every taxon present in the analysis. For example, fossils very rarely have molecular data and incomplete fossil preservation (e.g. soft *vs.* hard tissues) may restrict the amount of morphological data available (Sansom and Wills, 2013). Additionally, since the molecular phylogenetics revolution, it has become less common to collect morphological characters for living taxa when molecular data are available (e.g. in (Slater, 2013), only 13% of the 169 living taxa have coded morphological data). Unfortunately this missing data can lead to errors in phylogenetic inference; in fact, simulations show that the ability of the Total Evidence method to recover the correct phylogenetic topology decreases when there is a low overlap between morphological data in the living and fossil taxa (Guillerme and Cooper, In review), regardless the overall amount of morphological data available for the fossils (or the amount of molecular data available for the living species). The effect of missing data on topology is greatest when living taxa have

few morphological data. This is because (1) a fossil cannot branch in the correct clade if there is no overlapping morphological data in the clade; and (2) a fossil has a higher probability of branching within a clade with more morphological data available for living taxa, regardless of whether this is the correct clade or not (Guillerme and Cooper, In review).

The issues above highlight that it is crucial to have sufficient morphological data for living taxa in a clade before using a Total Evidence approach. However, it is unclear how much morphological data for living taxa is actually available (i.e. already coded from museum specimens and deposited in phylogenetic matrices accessible online), and how this data are distributed across clades. Intuitively, most people assume this kind of data has already been collected, but empirical data suggest otherwise (e.g. in (Ronquist et al., 2012a; Slater, 2013; Beck and Lee, 2014). To investigate this further, we assess the amount of available morphological data for living mammals to determine whether sufficient data exists to build reliable Total Evidence phylogenies in this group. We collected cladistic data (i.e. discrete morphological characters used in phylogenetics) from 286 phylogenetic matrices available online and measured the proportion of cladistic data available for each mammalian order. Additionally, because missing morphological data in living species can influence tree topology as described above, we determined whether the available cladistic data was phylogenetically overdispersed or clustered in the mammalian orders where data was missing. We find that available morphological data for living mammals is scarce but generally randomly distributed across phylogenies. We recommend that efforts be made to collect and share more cladistic data for living species to improve the accuracy of Total Evidence phylogenies.

3.2 MATERIALS AND METHODS

3.2.1 *Data collection and standardisation*

We downloaded all cladistic matrices containing any living and/or fossil mammal taxa from three major public databases (accessed 10th of June 2015): Morphobank (`http://www.morphobank.org/`) (O'Leary and Kaufman, 2011), Graeme Lloyd's website (`graemetlloyd.com/matrmamm.html`) and Ross Mounce's GitHub repository (`https://github.com/rossmounce/cladistic-data`). We also performed a systematic Google Scholar search (accessed 11th of June 2015) for matrices that were not uploaded to

these databases. We downloaded available matrices containing fossil and/or living mammal taxa from the three data bases using the following list of keywords:

`Mammalia; Monotremata; Marsupialia; Placentalia; Macroscelidea; Afrosoricida; Tubulidentata; Hyracoidea; Proboscidea; Sirenia; Pilosa; Cingulata; Scandentia; Dermoptera; Primates; Lagomorpha; Rodentia; Erinaceomorpha; Soricomorpha; Cetacea; Artiodactyla; Cetartiodactyla; Chiroptera; Perissodactyla; Pholidota; Carnivora; Didelphimorphia; Paucituberculata; Microbiotheria; Dasyuromorphia; Peramelemorphia; Notoryctemorphia; Diprotodontia`.

Note that some matrices have been downloaded from more than one database but that it is not an issue since we are interested in the total number of unique living OTUs and that if some where present in more than one matrix, they still only counted as one single OTU.

MORPHOBANK — We used the keywords listed above in the search menu of the Morphobank repository and downloaded the data associated with each project matching with the keywords.

GRAEME LLOYD — We downloaded all the matrices that were available with a direct download link in the mammal data section of Graeme Lloyd's website repository.

ROSS MOUNCE — We downloaded every 601 matrix from Ross Mounce's GitHub repository and then ran a shell script to select only the matrices that had any text element that match with one of the search terms. To make the matrix selection more thorough, we ignored the keywords case as well as the latin suffix (*ia*, *ata*, *ea*, and *a*).

GOOGLE SCHOLARS — To make sure we didn't miss any extra matrix that wasn't available on one of these repository, we ran a extra Google Scholar search. We downloaded the additional cladistic matrices from the 20 first search results matching with our selected keywords and with any of the 35 taxonomic levels (mammals Orders, Infraclasses and Class). We used the following key words:

order `("morphology" OR "morphological" OR "cladistic") AND characters matrix paleontology phylogeny`

were *order* was replaced by all the keywords listed above. For each 33 keywords, we selected the 20 first papers to match the Google search published since

2010 resulting in 660 papers. Among these papers, not all contained relevant data (discrete morphological characters AND mammalian data). We selected only the 20 first results per search term to avoid downloading articles that were to irrelevant. Among the 660 papers, only 50 contained a total of 425 extra living OTUs (Fig. 3.1). Also we decided to select only the articles published since 2010 because nearly every one of the recent published matrix contains both a fraction of morphological characters and OTUs from previous studies. For example in primates the character *p7* coded first by Ross et al. (1998) is reused with the same living species in Seiffert et al. (2003), Marivaux et al. (2005), Seiffert et al. (2005), Bloch et al. (2007), Bloch et al. (2007), Kay et al. (2008), Silcox (2008), Seiffert et al. (2009), Tabuce et al. (2009), Boyer et al. (2010), Seiffert et al. (2010), Marivaux et al. (2013) and Ni et al. (2013).

We transformed all the non-nexus matrices (tnt, word, excel, jpeg) to nexus format manually. In total, we downloaded 286 matrices containing a total of 11010 operational taxonomic units (OTUs) of which 5228 were unique. In this study, we refer to OTUs rather than species since the entries in the downloaded matrices were not standardised and ranged from specific individual specimen names (i.e. the name of a collection item) to the family-level. Where possible, we considered OTUs at their lowest valid taxonomic level (i.e. species) but some OTUs were only valid at a higher taxonomic level (e.g. genus or family). Therefore for some orders, we sampled more genera than species (Table 3.1).

To select the lowest valid taxonomic level for each OTU, we standardised their taxonomy by correcting species names so they matched standard taxonomic nomenclature (e.g., *H. sapiens* was transformed to *Homo sapiens*). We designated as "living" all OTUs that were either present in the phylogeny of (Bininda-Emonds et al., 2007) or the taxonomy of (Wilson and Reeder, 2005), and designated as "fossil" all OTUs that were present in the Paleobiology database (`https://paleobiodb.org/`).

For OTUs that did not appear in these three sources, we first decomposed the name (i.e. *Homo sapiens* became *Homo* and *sapiens*) and tried to match the first element with a higher taxonomic level (genus or family). Any OTUs that still had no matches in the sources above were designated as non-applicable (NA; see Fig. 3.2).

The number of characters in each matrix ranged from 6 to 4541. Matrices with few characters are problematic when comparing available data among matrices because (1) they have less chance of having characters that overlap with those of

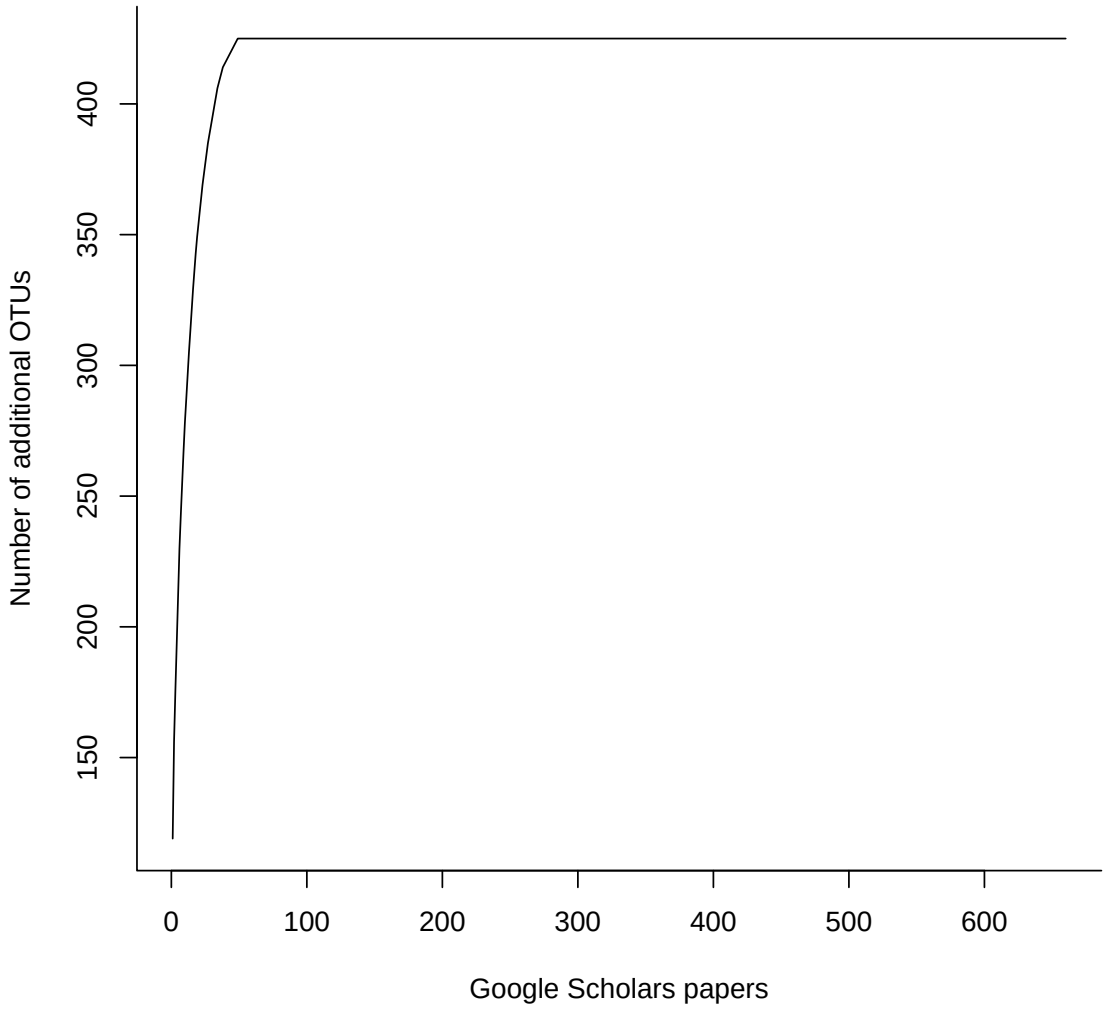

FIGURE 3.1: Google searches additional OTUs rarefaction curve. The x axis represent the number of google scholar matches (papers, books or abstracts) and the y axis represents the cumulative number of additional living OTUs per google scholar match.

other matrices (Wagner, 2000) and (2) they are more likely to contain a higher proportion of specific characters that are not-applicable across large clades (e.g. “antler ramifications” is a character that is only applicable to Cervidae not all mammals Brazeau, 2011). Therefore we selected only matrices containing >100 characters for each OTU. This threshold was chosen to correspond with the number of characters used in (Guillerme and Cooper, In review) and (Harrison and Larsson, 2015). Note that results of analyses with no character threshold are available in Supplementary Material. After removing all matrices with <100 characters, we retained 1074 unique living mammal OTUs from 126 matrices for our analyses.

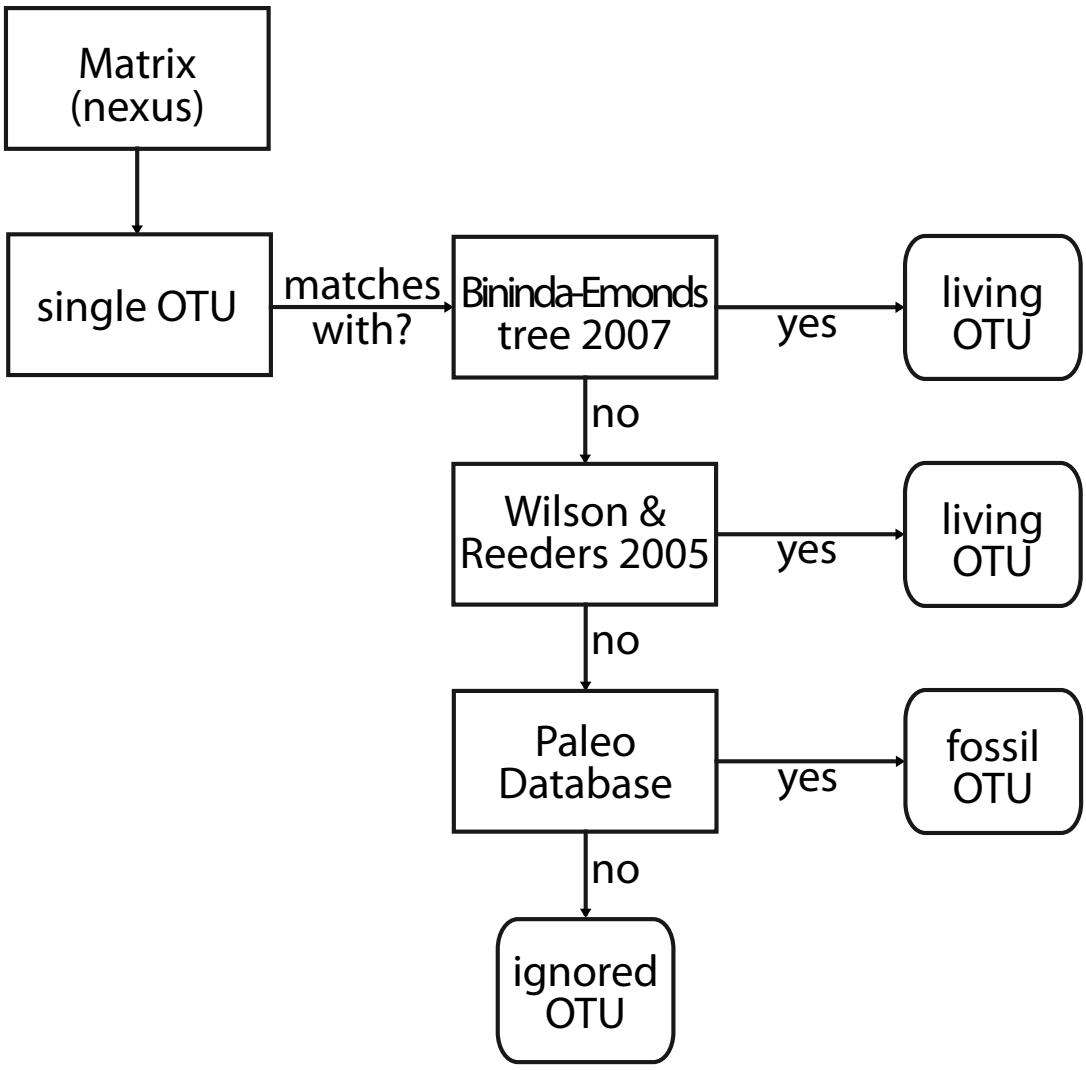

FIGURE 3.2: Taxonomic matching algorithm used in this study. For each matrix, each operational taxonomic units (OTU) is matched with the super tree from Bininda-Emonds 2007. If the OTU matches, then it is classified as living. Else it is matched with the Wilson & Reeders 2005 taxonomy list. If the OTU matches, then it is classified as living. Else it is matched with the Paleo Database list of mammals. If the OTU matches, then it is classified as fossil. Else it is ignored.

3.2.2 *Data availability and distribution*

To assess the availability of cladistic data for each mammalian order, we calculated the percentage of OTUs with cladistic data at three different taxonomic levels: family, genus and species. We consider orders with <25% of living taxa with cladistic data as having poor data coverage ("low" coverage), and orders with >75% of living taxa with cladistic data as having good data coverage (hereafter "high" coverage).

For orders with <100% cladistic data coverage at any taxonomic level, we investigated whether the available cladistic data was (i) randomly distributed, (ii) overdis-

persed or (iii) clustered, with respect to phylogeny, using two metrics from community phylogenetics: the Nearest Taxon Index (NTI; (Webb et al., 2002) and the Net Relatedness Index (NRI; (Webb et al., 2002). NTI is most sensitive to clustering or overdispersion near the tips, whereas NRI is more sensitive to clustering or overdispersion across the whole phylogeny (Cooper et al., 2008). Both metrics were calculated using the `picante` package in R (Kembel et al., 2010; R Core Team, 2015).

NTI (Webb et al., 2002) is based on mean nearest neighbour distance (MNND) and is calculated as follows:

$$NTI = -\left(\frac{\overline{MNND}_{obs} - \overline{MNND}_{n}}{\sigma(MNND_{n})}\right) \tag{3.1}$$

where $\overline{MNND}_{obs}$ is the observed mean distance between each of n taxa with cladistic data and its nearest neighbour with cladistic data in the phylogeny, $\overline{MNND}_{n}$ is the mean of 1000 mean MNND between n randomly drawn taxa, and $\sigma(MNND_{n})$ is the standard deviation of these 1000 random MNND values. NRI is similar but is based on mean phylogenetic distance (MPD) as follows:

$$NRI = -\left(\frac{\overline{MPD}_{obs} - \overline{MPD}_{n}}{\sigma(MPD_{n})}\right) \tag{3.2}$$

where $\overline{MPD}_{obs}$ is the observed mean phylogenetic distance of the tree containing only the n taxa with cladistic data, $\overline{MPD}_{n}$ is the expected random MPD for n taxa estimated by calculating the MPD from n taxa randomly drawn from the phylogeny and repeated 1000 times, and $\sigma(MPD_{n})$ is the standard deviation of the 1000 random MPD values. Negative NTI and NRI values show that the focal taxa are more overdispersed across the phylogeny than expected by chance, and positive values reflect significant clustering.

We calculated NTI and NRI values for each mammalian order separately, at each different taxonomic level. For each analysis our focal taxa were those with available cladistic data at that taxonomic level and the phylogeny was the phylogeny of the order pruned from (Bininda-Emonds et al., 2007).

3.3 RESULTS

Across the 126 cladistic matrices we extracted, 22 out of 28 mammalian orders have low coverage (<25% of species with cladistic data) and six have high coverage (>75% of species with cladistic data) at the species-level. At the genus-level, three orders have low coverage and 12 have high coverage, and at the family-level, no orders have low coverage and 23 have high coverage (Table 3.1).

TABLE 3.1: Number of taxa with available cladistic data for mammalian orders at three taxonomic levels. The left vertical bar represents "low" coverage (<25%); the right vertical bar represents "high" coverage (>75%). A negative Net Relatedness Index (NRI) and Nearest Taxon Index (NTI) shows more phylogeneticaly dispersed taxa than expected by chance; a positive value shows more phylogeneticaly clustered taxa than expected by chance. Significant NRI or NTI values are highlighted in bold. *p <0.05; **p <0.01; ***p <0.001.

Order	Taxonomic level	Proportion of taxa	Coverage	NRI	NTI
Afrosoricida	family	2/2			
Afrosoricida	genus	17/17			
Afrosoricida	**species**	**23/42**		**1.89***	1.19
Carnivora	family	11/15		0.43	1.68
Carnivora	**genus**	**30/125**		**4.14****	**1.81***
Carnivora	**species**	**42/283**		**18.64****	**3.02****
Cetartiodactyla	family	21/21			
Cetartiodactyla	**genus**	**77/128**		0.87	**1.77***
Cetartiodactyla	**species**	**129/310**		**2.72***	0.04
Chiroptera	family	13/18		0.55	0.63
Chiroptera	**genus**	**85/202**		**16.91****	**2.85****
Chiroptera	**species**	**165/1053**		**14.55****	**3.44****
Cingulata	family	1/1			
Cingulata	genus	8/9		1.49	-1.63

Cingulata	species	6/29	1.43	0.36
Dasyuromorphia	family	2/2		
Dasyuromorphia	genus	7/22	-1	-1.45
Dasyuromorphia	species	8/64	-1.15	-0.62
Dermoptera	family	1/1		
Dermoptera	genus	1/2		
Dermoptera	species	1/2		
Didelphimorphia	family	1/1		
Didelphimorphia	genus	16/16		
Didelphimorphia	species	40/84	-0.94	0.36
Diprotodontia	family	9/11	-0.8	0.56
Diprotodontia	genus	20/38	-1.36	-0.73
Diprotodontia	species	16/126	-2.29	-1.55
Erinaceomorpha	family	1/1		
Erinaceomorpha	genus	10/10		
Erinaceomorpha	species	21/22	-1.1	-0.3
Hyracoidea	family	1/1		
Hyracoidea	genus	1/3		
Hyracoidea	species	1/4		
Lagomorpha	family	1/2		
Lagomorpha	genus	1/12		
Lagomorpha	species	1/86		
Macroscelidea	family	1/1		
Macroscelidea	genus	4/4		

Macroscelidea	species	5/15	-0.98	-1.38
Microbiotheria	family	1/1		
Microbiotheria	genus	1/1		
Microbiotheria	species	1/1		
Monotremata	family	2/2		
Monotremata	genus	2/3	-0.71	-0.71
Monotremata	species	2/4	-1.01	-1.03
Notoryctemorphia	family	1/1		
Notoryctemorphia	genus	1/1		
Notoryctemorphia	species	0/2		
Paucituberculata	family	1/1		
Paucituberculata	genus	2/3	0	0
Paucituberculata	species	2/5	-0.64	-0.65
Peramelemorphia	family	2/2		
Peramelemorphia	genus	7/7		
Peramelemorphia	species	16/18	-0.09	1
Perissodactyla	family	3/3		
Perissodactyla	genus	6/6		
Perissodactyla	species	7/16	0.62	-2.5
Pholidota	family	1/1		
Pholidota	genus	1/1		
Pholidota	**species**	**3/8**	**2.64***	**2.23***
Pilosa	family	3/5	0.94	0.93
Pilosa	genus	3/5	-0.36	-0.31

Pilosa	species	3/29	0.33	0.79
Primates	family	15/15		
Primates	genus	48/68	-0.41	-1.4
Primates	species	56/351	-1.6	-2.04
Proboscidea	family	1/1		
Proboscidea	genus	1/2		
Proboscidea	species	1/3		
Rodentia	family	11/32	-0.46	-1.91
Rodentia	genus	21/450	-2.11	0.3
Rodentia	species	15/2094	-1.65	-2.55
Scandentia	family	2/2		
Scandentia	genus	2/5	-0.77	-0.76
Scandentia	species	2/20	-1.79	-1.99
Sirenia	family	2/2		
Sirenia	genus	2/2		
Sirenia	species	4/4		
Soricomorpha	family	3/4	-0.93	-0.92
Soricomorpha	**genus**	**19/43**	**6.98****	**2.49***
Soricomorpha	**species**	**19/392**	**13.19****	**3.89****
Tubulidentata	family	1/1		
Tubulidentata	genus	1/1		
Tubulidentata	species	1/1		

Among the mammalian orders containing OTUs with no available cladistic data, only six orders had significantly clustered data (Carnivora, Cetartiodactyla, Chi-

roptera and Soricomorpha at both species- and genus-level and Afrosoricida and Pholidota at the species-level only) and no order had significantly overdispersed data at any taxonomic level (Table 3.1).

Two contrasting results are shown in Fig. 3.3 with randomly distributed OTUs with cladistic data in Primates (Fig. 3.3A) and phylogenetically clustered OTUs with cladistic data in Carnivora (mainly Canidae; Fig. 3.3B).

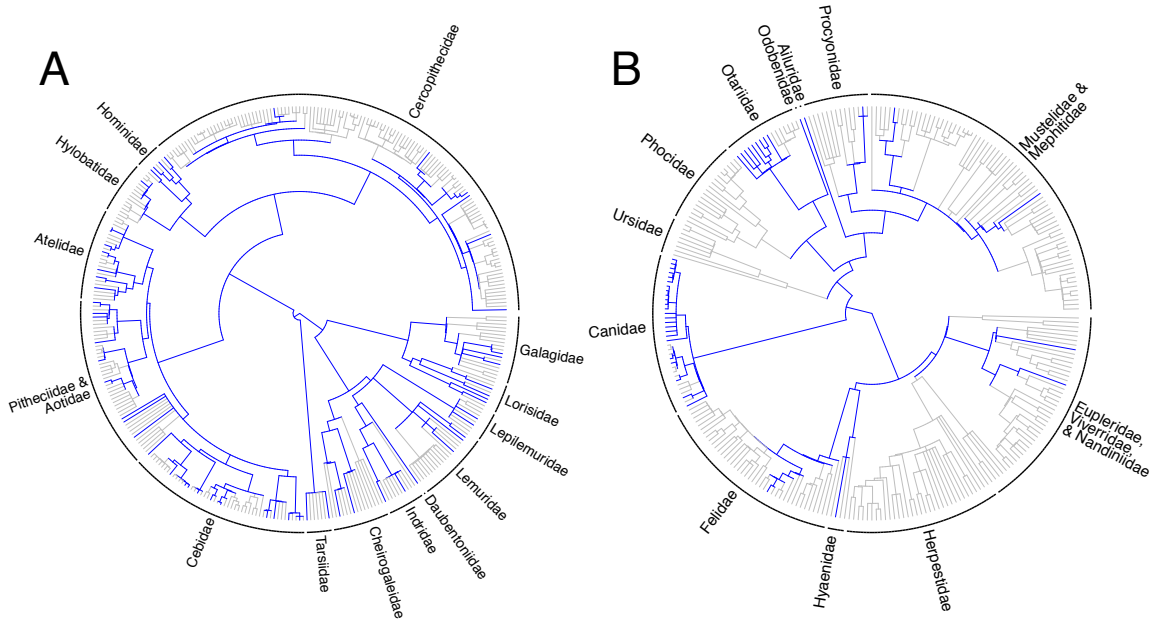

FIGURE 3.3: Phylogenetic distribution of species with available cladistic data across two mammalian orders (A: Primates; B: Carnivora). Edges are colored in grey when no cladistic data are available for a species and in blue when data are available.

3.4 DISCUSSION

Our results show that although phylogenetic relationships among living mammals are well-resolved (e.g. Bininda-Emonds et al., 2007; Meredith et al., 2011) , most of the data used to build these phylogenies is molecular, and very little cladistic data are available for living mammals compared to fossil mammals (e.g. O'Leary et al., 2013; Ni et al., 2013). This has implications for building Total Evidence phylogenies containing both living and fossil mammals, as without sufficient cladistic data for living species, fossil placements in these trees are very uncertain (Guillerme and Cooper, In review). Cladistic data coverage in living mammals varies across taxonomic levels and in its phylogenetic distribution. Higher taxonomic levels are always better sampled than lower ones and within these taxonomic levels, the available data are mostly randomly distributed across the phylogeny, apart from in six orders).

The number of living mammalian taxa with no available cladistic data was surprisingly high at the species-level: only six out of 28 orders have a high coverage of taxa with available cladistic data (and two of the 28 orders are monospecific!). This high coverage threshold of 75% of taxa with available cladistic data represents the minimum amount of data required before missing data has a significant effect on the topology of Total Evidence trees (Guillerme and Cooper, In review). Beyond this threshold, there is considerable displacement of wildcard taxa (*sensu* Kearney, 2002) and decreases in clade conservation (Guillerme and Cooper, In review). Therefore we expect a high probability of topological artefacts for the placement of fossil taxa at the species-level in most mammalian orders. However, data coverage seems to be less of an issue at higher taxonomic levels (i.e. genus- and family-level). This point is important from a practical point of view because of the slight discrepancy between the neontological and palaeontological concept of species. While neontological species are described using morphology, genetic distance, spatial distribution and even behaviour, palaeontological species can be based only on morphological, spatial and temporal data (e.g. Ni et al., 2013). Because of this, most palaeontological studies are using the genus as their smallest OTU (e.g. Ni et al., 2013; O'Leary et al., 2013). Thus data availability at the genus-level in living mammals should be our primary concern when aiming to build phylogenies of living and fossil taxa.

When only a few species with cladistic data are available, the ideal scenario is for them to be phylogenetically overdispersed (i.e. that there is data for at least every sub-clade) to maximize the possibilities of a fossil branching from the right clade. The second best scenario is that species with cladistic data are randomly distributed across the phylogeny. In this scenario we expect no special bias in the placement of the fossil (Guillerme and Cooper, In review), it is therefore encouraging that for most orders, species with cladistic data were randomly distributed across the phylogeny of each order. The worst case scenario for fossil placement is that species with cladistic data are phylogenetically clustered. In this situation we expect two major biases to occur: first, the fossil will not be able to branch within a clade containing no data, and second, the fossil will have a higher probability, at random, of branching within the clade containing most of the available data. This means that fossils with uncertain phylogenetic affinities (*incertae sedis* will have a higher probability of branching within the most sampled clade just by chance). Our results suggest that this may be an issue, at the genus-level, in Carnivora, Cetartiodactyla,

Chiroptera and Soricomorpha. For example, a Carnivora fossil will be unable to branch in the Herpestidae that has no species with cladistic data, and will also have more chance to branch, randomly, within the Canidae clade than any other clade in Carnivora (Fig. 3.3B). Thus, in Total Evidence trees, placements of some carnivoran fossils should considered with caution. In this study, we treated all cladistic matrices as equal in a similar way to molecular matrices. For example, if matrix A contained 100 characters for taxa X and Y, and matrix B contained 50 different characters for taxa X and Z, we assumed that both matrices can be combined in a supermatrix containing 150 independent characters for taxon X, 100 for taxon Y and 50 for taxon Z. Unfortunately, cladistic data cannot always be treated in this way because some characters may overlap. For example, if matrix A has a character coding for the shape of a particular morphological feature and matrix B has a character coding for the presence of this same morphological feature and a second character coding for its shape, then these three characters are non-independent compound characters (Brazeau, 2011). However, in reasonably sized matrices (>100 characters; Guillerme and Cooper, In review; Harrison and Larsson, 2015) it is more likely that a number of characters are consistently conserved among the different matrices and thus easily combinable. For example, within the Primate cladistic literature, the character *p7* - the size of the 4^{th} lower premolar paraconid - has been used consistently for >15 years (e.g. Ross et al., 1998; Marivaux et al., 2005; Ni et al., 2013) and can therefore be combined among the matrices. A conservative approach to avoid compound characters would be to select only the most recent matrix for each group, but this would result in the loss of a lot of data.

Despite the absence of good cladistic data coverage for living mammals, the Total Evidence methods still seems to be the most promising way of combining living and fossil data for macroevolutionary analyses. Following the recommendations in (Guillerme and Cooper, In review), we need to code cladistic characters for as many living species possible. Fortunately, data for living mammals is usually readily available in natural history collections, therefore, we propose that an increased effort be put into coding morphological characters from living species, possibly by engaging in collaborative data collection projects through web portals such as *MorphoBank* (O'Leary and Kaufman, 2011). Such an effort would be valuable not only to phylogeneticists, but also to any researcher focusing understanding macroevolutionary patterns and processes.

CHAPTER 4

SPATIO-TEMPORAL DISPARITY IN MAMMALS AT THE K-PG BOUNDARY

> "The most erroneous stories are those we think we know best - and therefore never scrutinize or question."
>
> S.J. Gould

Cretaceous-Palaeogene extinction does not affect mammalian disparity [3]

[3] A similar version of this chapter will be submitted to Evolution soon. T.G. and N.C. designed the experiments; T.G. ran the analysis and interpreted the results; T.G. and N.C. wrote the manuscripts. *Specific acknowledgements*: thanks to Graeme Lloyd, Andrew Jackson, Gavin Thomas and Sive Finlay. *Data availability and reproducibility*: Data will be available on Dryad or Figshare. Code for reproducing the analysis is available on GitHub (`https://github.com/TGuillerme/SpatioTemporal_Disparity`).

CHAPTER 5

DISCUSSION

5.1 THE FUTURE OF THE TOTAL EVIDENCE METHOD

Combined with tip-dating is super interesting but we need more data. To do so we can use plateforms such as morphobank and foster collaboration on big projects. Also we can make all the data available blablabla.

However, there are some limitations: Maybe tip-dating isn't that good? Compared to the nice recent node dating models... (Arcila) Also the Mk model is really crude and overly simplistic.

One way to improve could be a REAL total evidence dating using also trait data, biogeography, etc... In reality, all this parameters have an influence of lineages history and should technically be taken into account. But data problem is likely to increase, an needs models need to be improved as well. And in the end, how many parameters do we want?

5.2 DIVERSITY IS MULTIDIMENSIONAL

Diversity is often just seen as the sheer number of species. However, the processes that led to this pattern is fundamentaly intangled with all the other aspects of diversity. For example, specious rich groups have also so traits, etc... It is important to disentangle. But other dimensions as well: Ecological, life history, etc. We need to take into account more of these "disparity" patterns to really understand what happened. Especially when combining living and fossil, species richness is a really poor indicator of diversity.

However, this is more complex, species diversity is easy to interprate (many populations isolations through time) but disparity is a bit harder. What IS disparity? What metric to use? How to express the changes etc... Also, all these metrics are just using proxies.

But The statistician George Box wrote "essentially, all models are wrong, but some are useful" (Box and Draper, 1987). This is still really promising and can be

improved first by underestanding how all this works in a theoretical way (building the models). And only then apply it to observed patterns.

5.3 WHAT IS THE REAL EFFECT OF COMBINING?

Maybe only important when groups have actually a complex history? Old clades might have no living descendants and the question is therefore N/A Recent subclades maybe not have changed much in diversity so adding fossils might not change much. But we never know! Example of the giant lemur (recently extinct).

BIBLIOGRAPHY

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

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

Bapst, D. W. 2014. Assessing the effect of time-scaling methods on phylogeny-based analyses in the fossil record. Paleobiology 40:331–351.

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

Benson, R. B. J. and P. S. Druckenmiller. 2014. Faunal turnover of marine tetrapods during the Jurassic—Cretaceous transition. Biological Reviews 89:1–23.

Benton, M. J. 2015. Exploring macroevolution using modern and fossil data. Proceedings of the Royal Society of London B: Biological Sciences 282.

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

Bininda-Emonds, O. R. P., M. Cardillo, K. E. Jones, R. D. E. MacPhee, R. M. D. Beck, R. Grenyer, S. A. Price, R. A. Vos, J. L. Gittleman, and A. Purvis. 2007. The delayed rise of present-day mammals. Nature 446:507–512.

Bloch, J. I., M. T. Silcox, D. M. Boyer, and E. J. Sargis. 2007. New paleocene skeletons and the relationship of plesiadapiforms to crown-clade primates. Proc. Nat. Acad. Sci. 104:1159–1164.

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

Bouckaert, R., J. Heled, D. KÃijhnert, T. Vaughan, C.-H. Wu, D. Xie, M. A. Suchard, A. Rambaut, and A. J. Drummond. 2014. Beast 2: A software platform for bayesian evolutionary analysis. PLoS Comput Biol 10:e1003537.

Box, G. E. and N. R. Draper. 1987. Empirical model-building and response surfaces. John Wiley & Sons.

Boyer, D. M., E. R. Seiffert, and E. L. Simons. 2010. Astragalar morphology of afradapis, a large adapiform primate from the earliest late eocene of egypt. Am. J. Phys. Anthropol. 143:383–402.

Brazeau, M. D. 2011. Problematic character coding methods in morphology and their effects. Biol. J. Linn. Soc. 104:489–498.

Bremer, B. and L. Struwe. 1992. Phylogeny of the rubiaceae and the loganiaceae: Congruence of conflict between morphological and molecular data? American Journal of Botany Pages 1171–1184.

Bronzati, M., F. C. Montefeltro, and M. C. Langer. 2015. Diversification events and the effects of mass extinctions on crocodyliformes evolutionary history. Royal Society Open Science 2.

Brusatte, S. L., M. J. Benton, M. Ruta, and G. T. Lloyd. 2008. The first 50âĂŁmyr of dinosaur evolution: macroevolutionary pattern and morphological disparity. Biology Letters 4:733–736.

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

Clapham, M. E., D. J. Bottjer, C. M. Powers, N. Bonuso, M. L. Fraiser, P. J. Marenco, S. Q. Dornbos, and S. B. Pruss. 2006. Assessing the ecological dominance of phanerozoic marine invertebrates. PALAIOS 21:431–441.

Cooper, N., J. Rodríguez, and A. Purvis. 2008. A common tendency for phylogenetic overdispersion in mammalian assemblages. P. Roy. Soc. B-Biol. Sci. 275:2031–2037.

Coxall, H. K., S. D'Hondt, and J. C. Zachos. 2006. Pelagic evolution and environmental recovery after the cretaceous-paleogene mass extinction. Geology 34:297–300.

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

Dembo, M., N. J. Matzke, A. Ø. Mooers, and M. Collard. 2015. Bayesian analysis of a morphological supermatrix sheds light on controversial fossil hominin relationships. Proceedings of the Royal Society of London B: Biological Sciences 282.

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

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

Donoghue, P. C. and M. J. Benton. 2007. Rocks and clocks: calibrating the tree of life using fossils and molecules. Trends in Ecology and Evolution 22:424 – 431.

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

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

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

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

Felsenstein, J. 2004. Inferring phylogenies. Sunderland, Massachusetts: Sinauer Associate.

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

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

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

Goloboff, P. A., J. S. Farris, and K. C. Nixon. 2008. Tnt, a free program for phylogenetic analysis. Cladistics 24:774–786.

Guillerme, T. and N. Cooper. 2015. Assessment of cladistic data availability for living mammals. bioRxiv .

Guillerme, T. and N. Cooper. In review. Effects of missing data on topological inference using a total evidence approach,. Molecular Phylogenetics and Evolution .

Harrison, L. B. and H. C. E. Larsson. 2015. Among-character rate variation distributions in phylogenetic analysis of discrete morphological characters. Systematic Biology 64:307–324.

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.

Hassanin, A., G. Lecointre, and S. Tillier. 1998. The 'evolutionary signal' of homoplasy in proteincoding gene sequences and its consequences for a priori weighting in phylogeny. Comptes Rendus de l'Académie des Sciences - Series {III} - Sciences de la Vie 321:611 – 620.

Healy, K., T. Guillerme, S. Finlay, A. Kane, S. B. A. Kelly, D. McClean, D. J. Kelly, I. Donohue, A. L. Jackson, and N. Cooper. 2014. Ecology and mode-of-life explain lifespan variation in birds and mammals. Proceedings of the Royal Society of London B: Biological Sciences 281.

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

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

Hunt, G., M. J. Hopkins, and S. Lidgard. 2015. Simple versus complex models of trait evolution and stasis as a response to environmental change. Proceedings of the National Academy of Sciences 112:4885–4890.

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

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

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

Kay, R. F., J. Fleagle, T. Mitchell, M. Colbert, T. Bown, and D. W. Powers. 2008. The anatomy of dolichocebus gaimanensis, a stem platyrrhine monkey from argentina. J. Hum. Evol. 54:323–382.

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

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

Kembel, S., P. Cowan, M. Helmus, W. Cornwell, H. Morlon, D. Ackerly, S. Blomberg, and C. Webb. 2010. Picante: R tools for integrating phylogenies and ecology. Bioinformatics 26:1463–1464.

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

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

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

Losos, J. B. 2010. Adaptive radiation, ecological opportunity, and evolutionary determinism. The American Naturalist 175:pp. 623–639.

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.

Marivaux, L., P.-O. Antoine, S. R. H. Baqri, M. Benammi, Y. Chaimanee, J.-Y. Crochet, D. De Franceschi, N. Iqbal, J.-J. Jaeger, G. Métais, et al. 2005. Anthropoid primates from the oligocene of pakistan (bugti hills): data on early anthropoid evolution and biogeography. Proceedings of the National Academy of Sciences of the United States of America 102:8436–8441.

Marivaux, L., A. Ramdarshan, E. M. Essid, W. Marzougui, H. K. Ammar, R. Lebrun, B. Marandat, G. Merzeraud, R. Tabuce, and M. Vianey-Liaud. 2013. Djebelemur, a tiny pre-tooth-combed primate from the eocene of tunisia: a glimpse into the origin of crown strepsirhines. PloS ONE 8:e80778.

Martin, S. 2008. Global diversity of crocodiles (crocodilia, reptilia) in freshwater. Hydrobiologia 595:587–591.

Masters, J. C. and D. J. Brothers. 2002. Lack of congruence between morphological and molecular data in reconstructing the phylogeny of the galagonidae. American Journal of Physical Anthropology 117:79–93.

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

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

Meseguer, A. S., J. M. Lobo, R. Ree, D. J. Beerling, and I. Sanmartín. 2015. Integrating fossils, phylogenies, and niche models into biogeography to reveal ancient evolutionary history: The case of hypericum (hypericaceae). Systematic Biology 64:215–232.

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

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

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

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

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

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

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

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

Parham, J. F., P. C. J. Donoghue, C. J. Bell, T. D. Calway, J. J. Head, P. A. Holroyd, J. G. Inoue, R. B. Irmis, W. G. Joyce, D. T. Ksepka, J. S. L. PatanÃĺ, N. D. Smith, J. E. Tarver, M. van Tuinen, Z. Yang, K. D. Angielczyk, J. M. Greenwood, C. A. Hipsley, L. Jacobs, P. J. Makovicky, J. MÃijller, K. T. Smith, J. M. Theodor, R. C. M. Warnock, and M. J. Benton. 2012. Best practices for justifying fossil calibrations. Systematic Biology 61:346–359.

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

Patterson, C., D. M. Williams, and C. J. Humpries. 1993. Congruence between molecular and morphological phylogenies. Annual Review of Ecology and Systematics Pages 153–188.

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

Payne, J. L., N. A. Heim, M. L. Knope, and C. R. McClain. 2014. Metabolic dominance of bivalves predates brachiopod diversity decline by more than 150 million years. Proceedings of the Royal Society B: Biological Sciences 281.

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

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

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

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

R Core Team. 2015. R: a language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria.

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

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

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

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

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

Ross, C., B. Williams, and R. F. Kay. 1998. Phylogenetic analysis of anthropoid relationships. Journal of Human Evolution 35:221–306.

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

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

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

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

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

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

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

Seiffert, E. R., J. M. Perry, E. L. Simons, and D. M. Boyer. 2009. Convergent evolution of anthropoid-like adaptations in eocene adapiform primates. Nature 461:1118–1121.

Seiffert, E. R., E. L. Simons, and Y. Attia. 2003. Fossil evidence for an ancient divergence of lorises and galagos. Nature 422:421–424.

Seiffert, E. R., E. L. Simons, D. M. Boyer, J. M. Perry, T. M. Ryan, and H. M. Sallam. 2010. A fossil primate of uncertain affinities from the earliest late eocene of egypt. Proc. Nat. Acad. Sci. 107:9712–9717.

Seiffert, E. R., E. L. Simons, W. C. Clyde, J. B. Rossie, Y. Attia, T. M. Bown, P. Chatrath, and M. E. Mathison. 2005. Basal anthropoids from egypt and the antiquity of africa's higher primate radiation. Science 310:300–304.

Sepkoski, J., J. John. 1981. A factor analytic description of the phanerozoic marine fossil record. Paleobiology 7:pp. 36–53.

Silcox, M. T. 2008. The biogeographic origins of primates and euprimates: east, west, north, or south of eden? Pages 199–231 *in* Mammalian Evolutionary Morphology. Springer.

Silverman, B. 1986. Density Estimation for Statistics and Data Analysis. Chapman & Hall/CRC Monographs on Statistics & Applied Probability Taylor & Francis.

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

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

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

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

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

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

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

Stubbs, T. L., S. E. Pierce, E. J. Rayfield, and P. S. L. Anderson. 2013. Morphological and biomechanical disparity of crocodile-line archosaurs following the end-triassic extinction. Proceedings of the Royal Society of London B: Biological Sciences 280.

Tabuce, R., L. Marivaux, R. Lebrun, M. Adaci, M. Bensalah, P.-H. Fabre, E. Fara, H. G. Rodrigues, L. Hautier, J.-J. Jaeger, et al. 2009. Anthropoid versus strepsirhine status of the african eocene primates algeripithecus and azibius: craniodental evidence. P. Roy. Soc. B-Biol. Sci.s Page rspb20091339.

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

Thorne, P. M., M. Ruta, and M. J. Benton. 2011. Resetting the evolution of marine reptiles at the Triassic-Jurassic boundary. Proceedings of the National Academy of Sciences 108:8339–8344.

Uetz, P. 2010. The original descriptions of reptiles. Zootaxa 2334:59–68.

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

Webb, C. O., D. D. Ackerly, M. A. McPeek, and M. J. Donoghue. 2002. Phylogenies and community ecology. Annual review of ecology and systematics Pages 475–505.

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

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

Wiens, J. J. 2015. Explaining large-scale patterns of vertebrate diversity. Biology Letters 11.

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

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

Wills, M. A., D. E. G. Briggs, and R. A. Fortey. 1994. Disparity as an evolutionary index: A comparison of cambrian and recent arthropods. Paleobiology 20:93–130.

Wilson, D. E. and D. M. Reeder. 2005. Mammal species of the world: a taxonomic and geographic reference vol. 1. JHU Press.

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

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

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

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

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

APPENDIX A

SUPPLEMENTARY DATA TO CHAPTER 2

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

bigskip

The following section contains supplementary results to the chapter “Effects of missing data on topological inference using a Total Evidence approach”.

A.1 DIFFERENCES BETWEEN THE “TRUE” AND THE INFERRED TREES.

In our simulation protocol, we used the “true” tree to generate the molecular characters and the morphological characters for the living and fossil taxa (i.e. the “complete” matrix). Therefore, the “true” tree can be seen as a random seed for starting our simulations. The following analysis measures the performance of our parameter and algorithms choices to generate the “complete” matrix. To asses the performance of our simulation protocol, we compared our “true” trees (i.e. the trees **used to create** the “complete” matrices) to the “best” trees (i.e. the trees **inferred from** the “complete” matrices; Fig. A.1). Note that the difference between the “true” and the “best” trees represents the effect of the parameters choice and the algorithms used to create the “complete” matrix as well the as the capacity of RAxML and MrBayes to infer phylogenies from this particular matrices (i.e. small sized and generated using specific algorithms). This does not affect, however, the results of our analysis since we deliberately compared the the “missing-data” trees to the “best” tree rather than to the “true” tree.

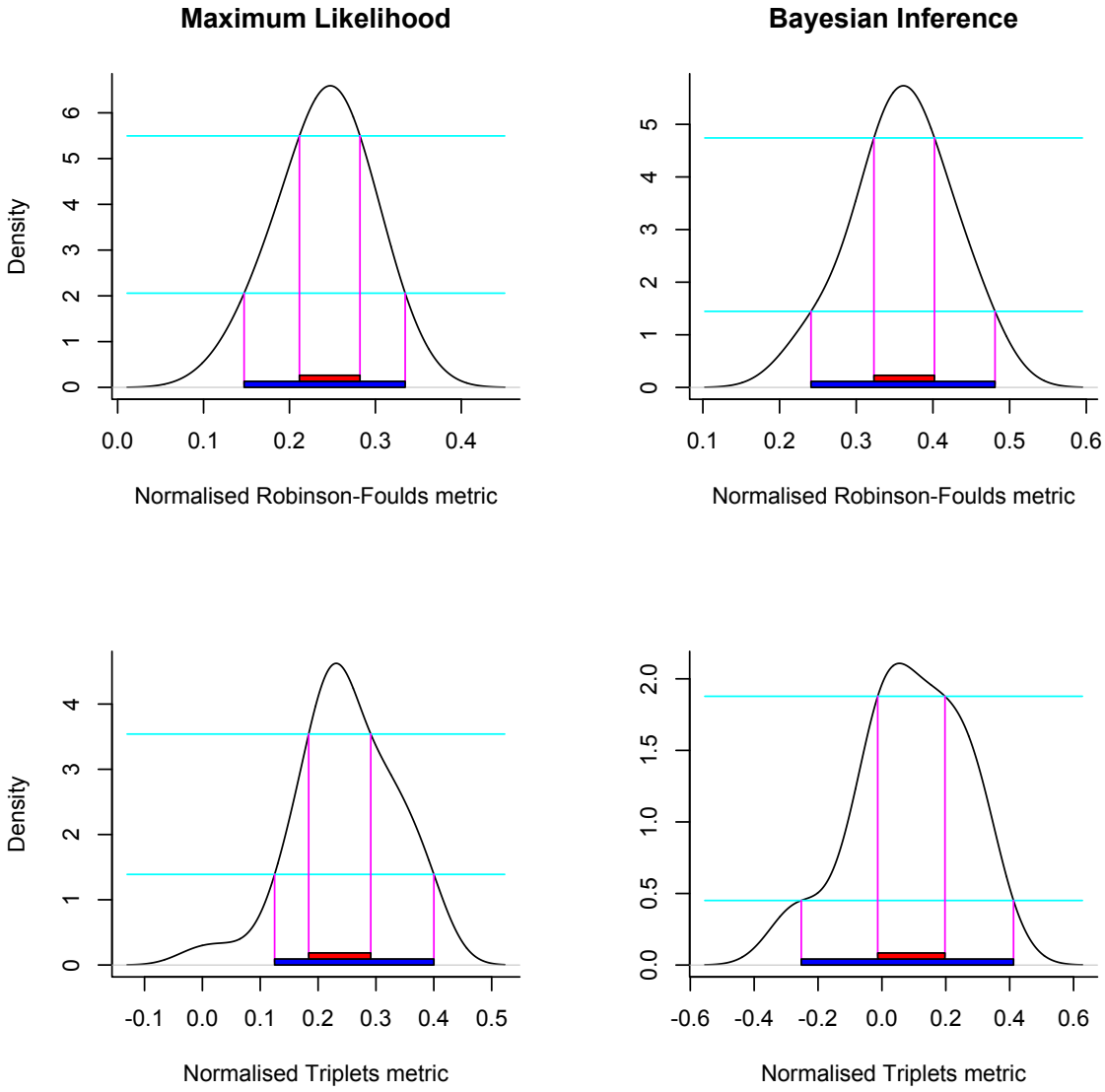

FIGURE A.1: Pairwise comparisons among the 50 “true” trees and the 50 “best” trees from the Maximum Likelihood and Bayesian inference methods. The horizontal blue and red lines represent, respectively, the 95% and 50% confidence intervals.

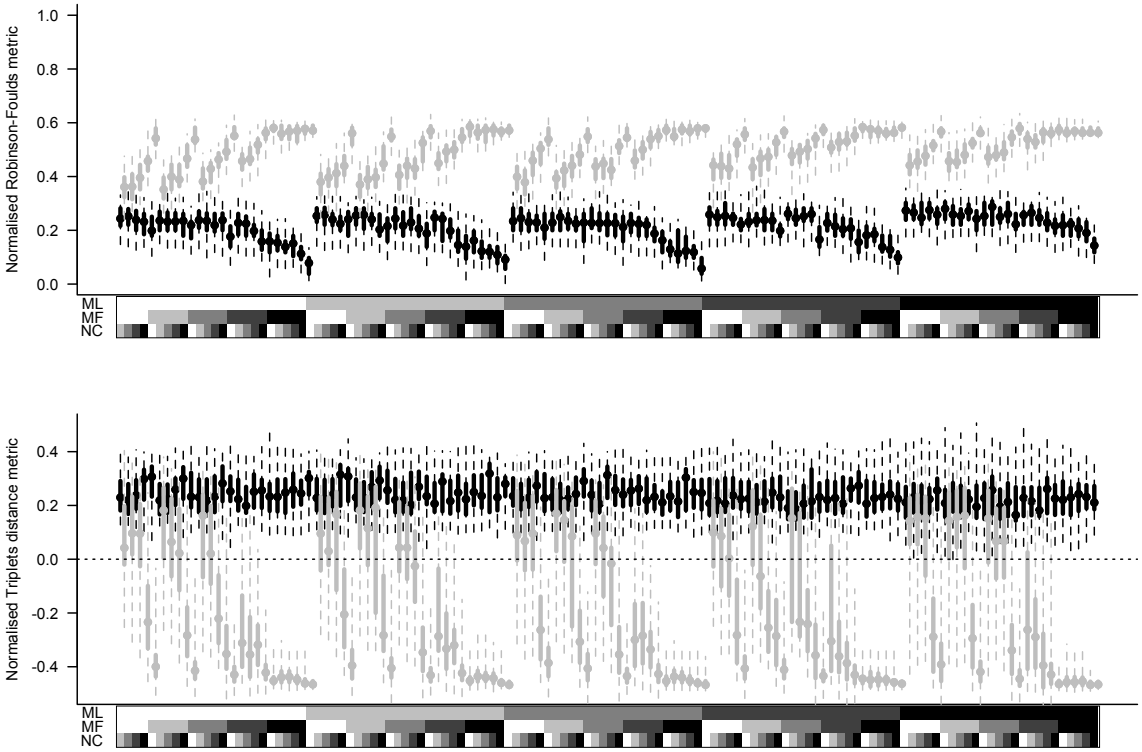

FIGURE A.2: Effect of increasing missing data on recovering the "true" tree topology (the tree used for starting our simulations) for the Maximum Likelihood trees (black) and Bayesian consensus trees (grey). The x axis shows the percentage of missing data from 0% (white) to 75% (black) for the two parameters: M_L (upper line), M_F (middle line) and number of characters from 100 to 25 for the parameter N_C (lower line). Topological recovery was measured using two different tree comparison metrics: Normalised Robinson-Foulds metric (upper row) and Normalised Triplets metric (lower row). The graph shows the modal value (points), and the 50% (thick solid lines) and 95% (thin dashed lines) confidence intervals of the distributions of the tree comparison metric for each missing data parameter and tree inference method.

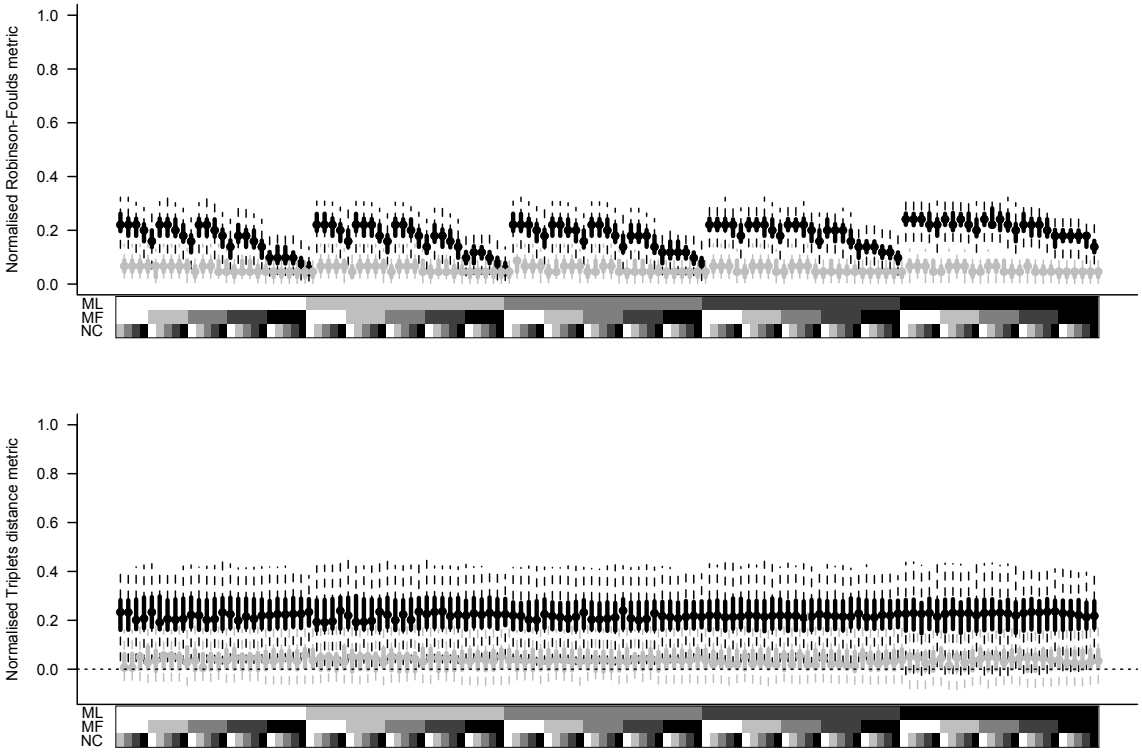

FIGURE A.3: Effect of increasing missing data on topological recovering the "true" tree topology (the tree used for starting our simulations) for the Maximum Likelihood Bootstrap trees (black) and Bayesian posterior tree distribution (grey). The x axis shows the percentage of missing data from 0% (white) to 75% (black) for the two parameters: M_L (upper line), M_F (middle line) and number of characters from 100 to 25 for the parameter N_C (lower line). Topological recovery was measured using two different tree comparison metrics: Normalised Robinson-Foulds metric (upper row) and Normalised Triplets metric (lower row). The graph shows the modal value (points), and the 50% (thick solid lines) and 95% (thin dashed lines) confidence intervals of the distributions of the tree comparison metric for each missing data parameter and tree inference method.

A.2 TREE INFERENCE SOFTWARE SETTINGS

For clarity we have provided the exact settings used in our tree building below.

Maximum Likelihood: RAxML version 8.0.20 Stamatakis (2014)

- Molecular data: GTR + Γ_4 (-m GTRGAMMA)
- Morphological data: M*kv* + Γ_4 (-K MK)
- Support: Rapid Boostrap algorithm (LSR), 1000 replicates

Bayesian: MrBayes version 3.2.1 Ronquist et al. (2012b)

- Priors: Molecular data
  - Rates distribution shape (α) = 0.5
  - Transition/Transversion ratio = 2 (β(80,40))
  - Starting tree: "True" tree topology with each branch length = 1
- Priors: Morphological data
  - rates distribution shape (α) = 0.5
- Models
  - Molecular data: HKY + Γ_4
  - Morphological data: M*kv* + Γ_4
- MCMC
  - Two runs
  - Four chains per run
  - Generations $< 5\times10^7$
  - Sample frequency = 1.05×10^4
  - ASDS diagnosis frequency = 5×1^4
  - ASDS $<$ 0.01
  - ESS $>>$ 200
  - Burnin = 25%

APPENDIX B

SUPPLEMENTARY DATA TO CHAPTER 3

Assessment of cladistic data availability for living mammals

bigskip

The following section contains supplementary results to the chapter “Assessment of cladistic data availability for living mammals”: the available data structure using the NTI and the PD metric; the proportion of available data and the data structure for all the matrices (including the matrices with less than 100 characters); and phylogenetical representation of the data availability per order (excluding Primates and Carnivora, present in the main body).

TABLE B.1: Number of taxa with available cladistic data for mammalian orders at three taxonomic levels (without any character threshold; results from the 286 matrices). The coverage represents the proportion of taxa with available morphological data. The left vertical bar represents 25% of available data (“low” coverage if <25%); The right vertical bar represents 75% of available data (“high” coverage if >75%). When the Net Relatedness Index (NRI) and the Nearest Taxon Index (NTI) are negative, taxa are more phylogenetically dispersed than expected by chance; when NRI or NTI are positive, taxa are more phylogenetically clustered by expected by chance. Significant NRI or NTI are highlighted in bold. *p <0.05; **p <0.01; ***p <0.001.

Order	Taxonomic level	Proportion of taxa	Coverage	NRI	NTI
Afrosoricida	family	2/2			
Afrosoricida	genus	17/17			
Afrosoricida	species	23/42		1.75	1.08
Carnivora	family	14/15		0.63	0.6
Carnivora	**genus**	**54/125**		**4.81****	**1.78***

Carnivora	**species**	**76/283**	**7.66****	**0.85**
Cetartiodactyla	family	21/21		
Cetartiodactyla	genus	100/128	0.85	0.94
Cetartiodactyla	**species**	**171/310**	**1.92***	**-0.46**
Chiroptera	family	15/18	-0.28	0.56
Chiroptera	**genus**	**93/202**	**13.47****	**1.1**
Chiroptera	**species**	**215/1053**	**8.82****	**1.22**
Cingulata	family	1/1		
Cingulata	genus	8/9	1.51	-1.57
Cingulata	**species**	**9/29**	**1.9***	**0.11**
Dasyuromorphia	family	2/2		
Dasyuromorphia	genus	8/22	-0.75	-1.07
Dasyuromorphia	species	9/64	-0.88	-0.34
Dermoptera	family	1/1		
Dermoptera	genus	1/2		
Dermoptera	species	1/2		
Didelphimorphia	family	1/1		
Didelphimorphia	genus	16/16		
Didelphimorphia	species	42/84	-1.65	0.2
Diprotodontia	family	11/11		
Diprotodontia	genus	25/38	-1.13	-1.31
Diprotodontia	species	31/126	0.48	-1.77
Erinaceomorpha	family	1/1		
Erinaceomorpha	genus	10/10		

Erinaceomorpha	species	21/22	-1.07	-0.2
Hyracoidea	family	1/1		
Hyracoidea	genus	1/3		
Hyracoidea	species	1/4		
Lagomorpha	family	2/2		
Lagomorpha	genus	5/12	-1.06	-0.95
Lagomorpha	species	12/86	-0.62	-1.88
Macroscelidea	family	1/1		
Macroscelidea	genus	4/4		
Macroscelidea	species	12/15	-1.3	-1.06
Microbiotheria	family	1/1		
Microbiotheria	genus	1/1		
Microbiotheria	species	1/1		
Monotremata	family	2/2		
Monotremata	genus	2/3	-0.72	-0.69
Monotremata	species	2/4	-0.97	-0.97
Notoryctemorphia	family	1/1		
Notoryctemorphia	genus	1/1		
Notoryctemorphia	species	0/2		
Paucituberculata	family	1/1		
Paucituberculata	genus	3/3		
Paucituberculata	species	5/5		
Peramelemorphia	family	2/2		
Peramelemorphia	genus	7/7		

Peramelemorphia	species	16/18	-0.13	0.97
Perissodactyla	family	3/3		
Perissodactyla	genus	6/6		
Perissodactyla	species	10/16	-0.07	-2.63
Pholidota	family	1/1		
Pholidota	genus	1/1		
Pholidota	species	4/8	1.18	0.94
Pilosa	family	4/5	1.87	2
Pilosa	genus	4/5	-0.96	0.36
Pilosa	**species**	**5/29**	**1.28**	**2.38***
Primates	family	15/15		
Primates	genus	48/68	-0.35	-1.33
Primates	species	64/351	-0.67	-1.27
Proboscidea	family	1/1		
Proboscidea	genus	2/2		
Proboscidea	species	2/3	-0.69	-0.69
Rodentia	family	18/32	0.66	-0.98
Rodentia	genus	82/450	-1.66	1.55
Rodentia	**species**	**90/2094**	**2.76***	**2.34***
Scandentia	family	2/2		
Scandentia	genus	2/5	-0.74	-0.74
Scandentia	species	3/20	-1.88	-0.84
Sirenia	family	2/2		
Sirenia	genus	2/2		

Sirenia	species	4/4		
Soricomorpha	family	3/4	-0.98	-0.99
Soricomorpha	**genus**	**19/43**	**7.11****	**2.59****
Soricomorpha	**species**	**21/392**	**10.65****	**3.56****
Tubulidentata	family	1/1		
Tubulidentata	genus	1/1		
Tubulidentata	species	1/1		

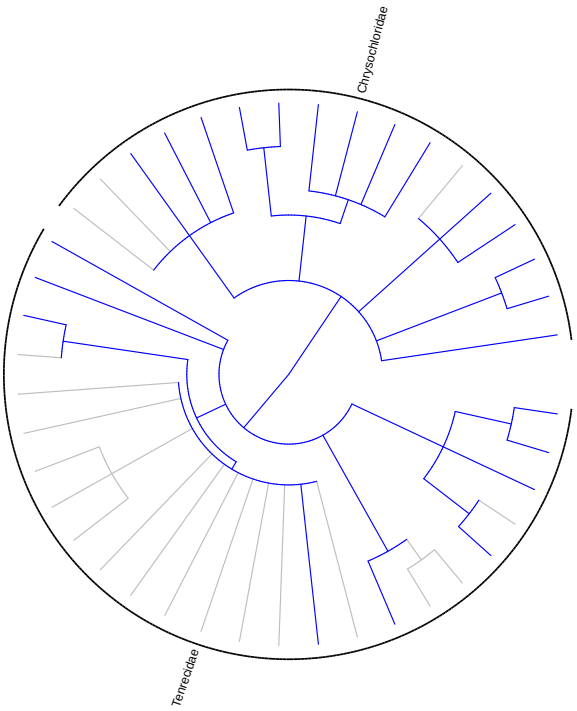

FIGURE B.1: Distribution of available morphological data across Afrosoricida. Edges are colored in grey when no morphological data is available or in blue when data is available.

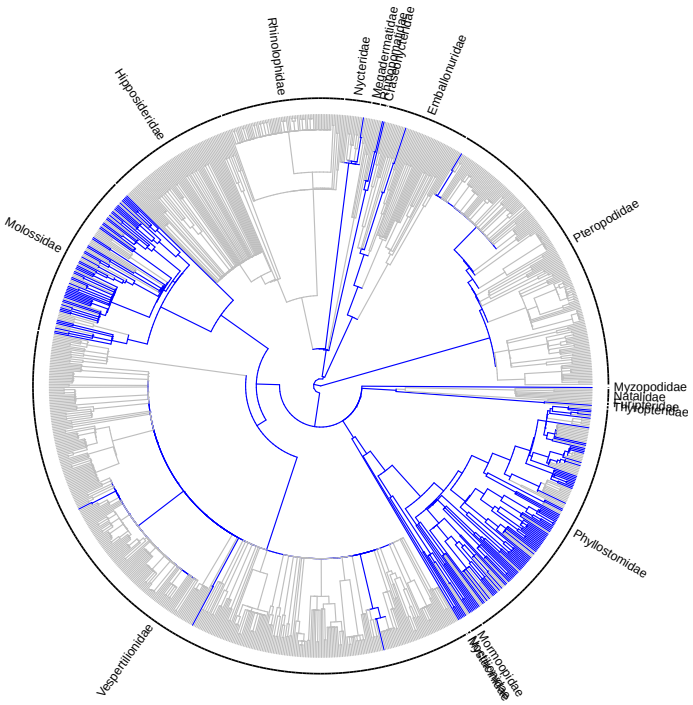

FIGURE B.2: Distribution of available morphological data across Chiroptera. Edges are colored in grey when no morphological data is available or in blue when data is available.

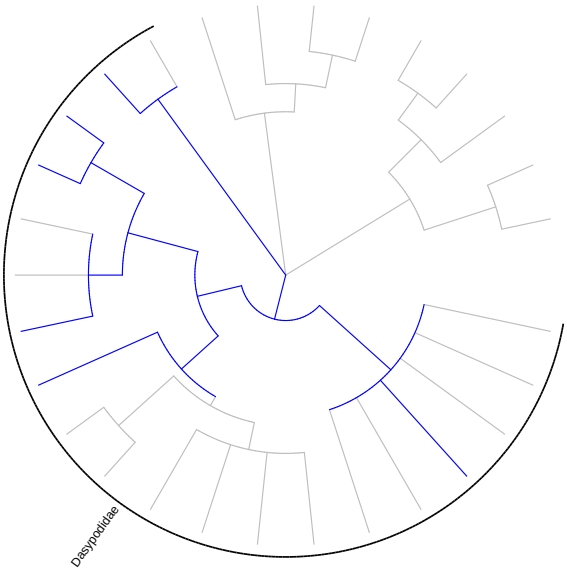

FIGURE B.3: Distribution of available morphological data across Cingulata. Edges are colored in grey when no morphological data is available or in blue when data is available.

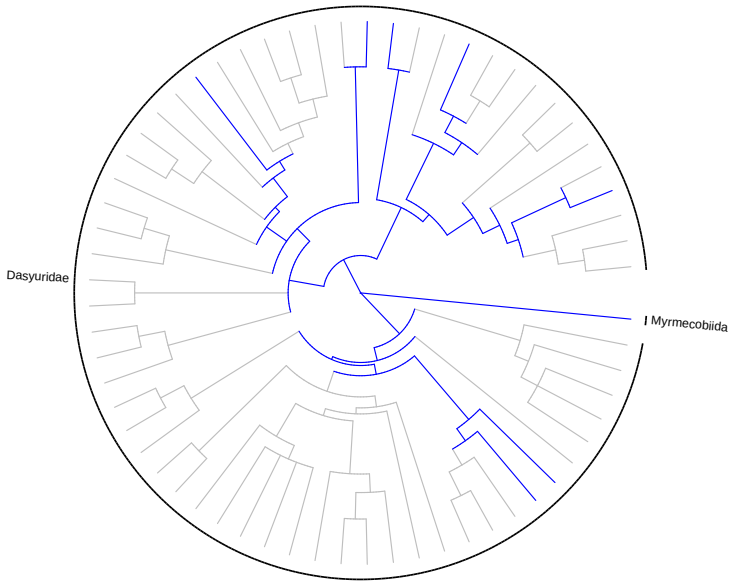

FIGURE B.4: Distribution of available morphological data across Dasyuromorphia. Edges are colored in grey when no morphological data is available or in blue when data is available.

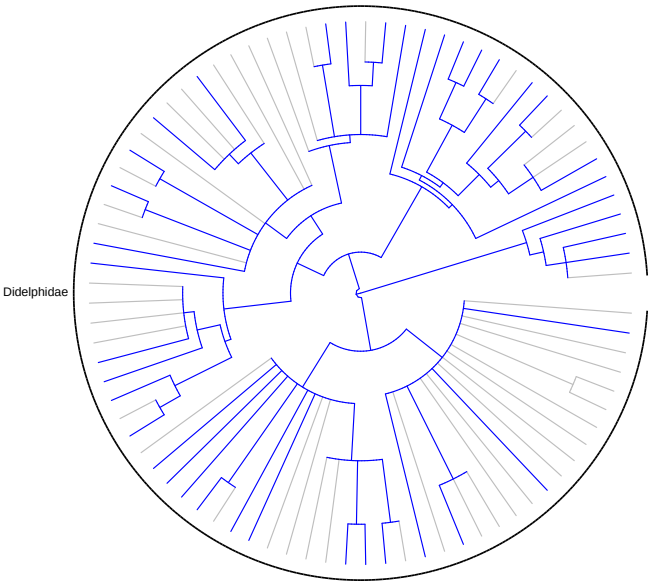

FIGURE B.5: Distribution of available morphological data across Didelphimorphia. Edges are colored in grey when no morphological data is available or in blue when data is available.

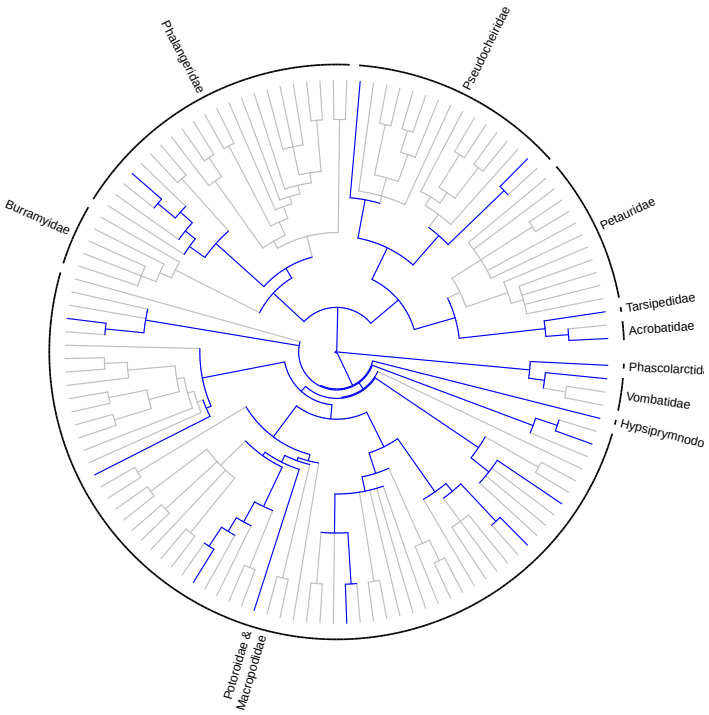

FIGURE B.6: Distribution of available morphological data across Diprotodontia. Edges are colored in grey when no morphological data is available or in blue when data is available.

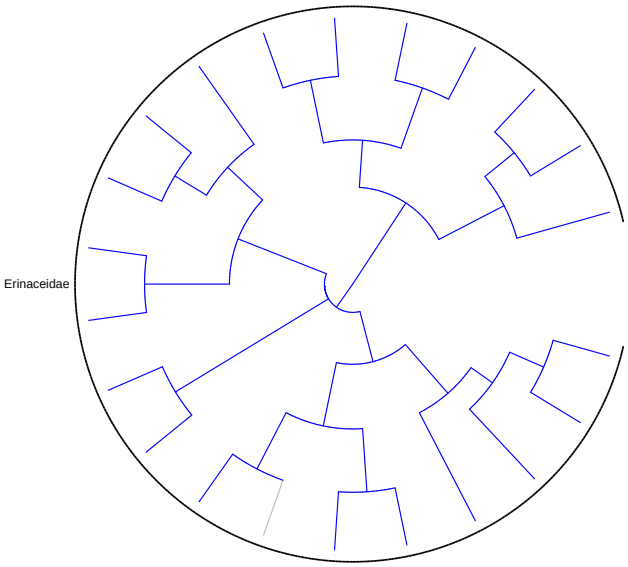

FIGURE B.7: Distribution of available morphological data across Erinaceomorpha. Edges are colored in grey when no morphological data is available or in blue when data is available.

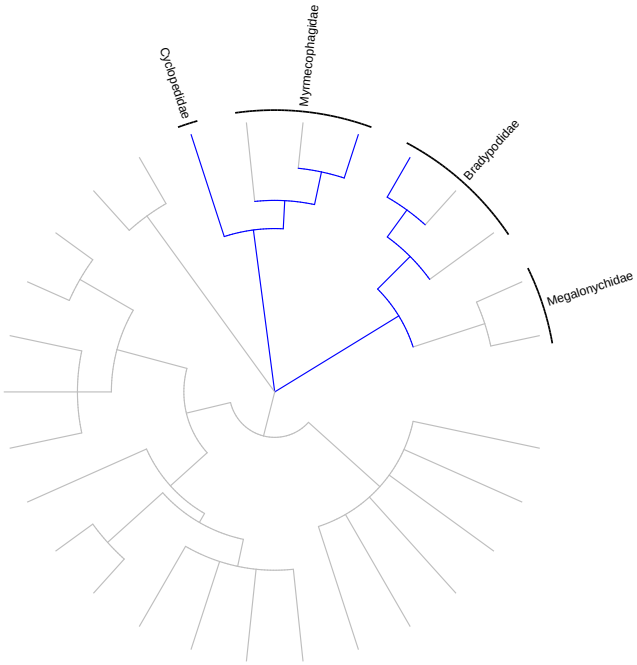

FIGURE B.8: Distribution of available morphological data across Pilosa. Edges are colored in grey when no morphological data is available or in blue when data is available.

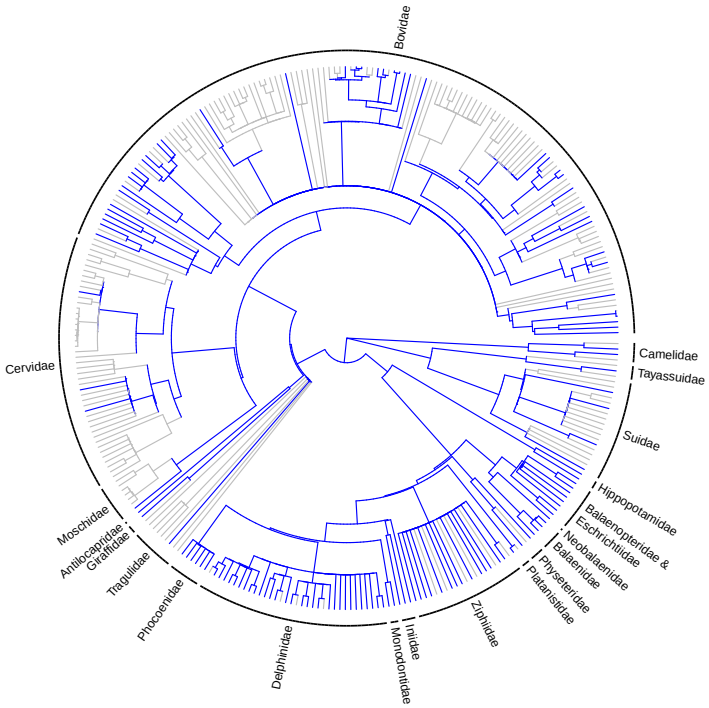

FIGURE B.9: Distribution of available morphological data across Cetartiodactyla. Edges are colored in grey when no morphological data is available or in blue when data is available.

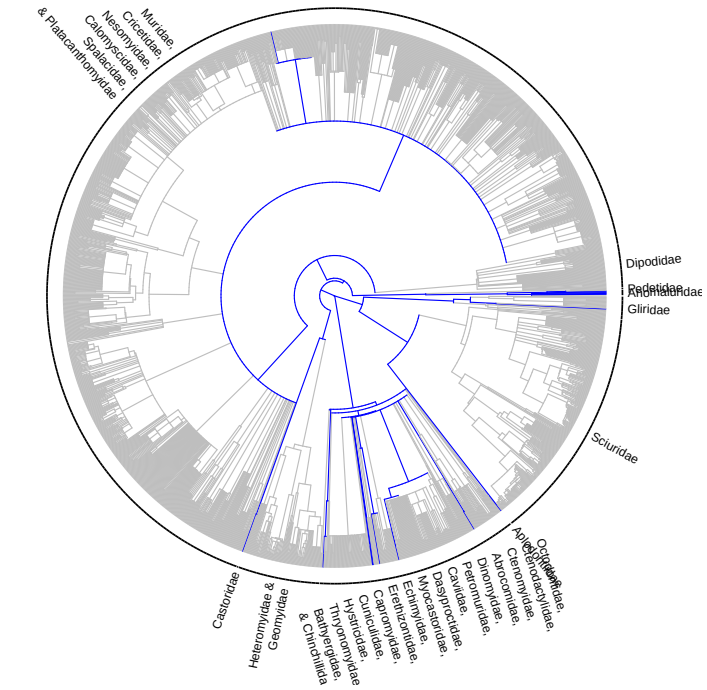

FIGURE B.10: Distribution of available morphological data across Rodentia. Edges are colored in grey when no morphological data is available or in blue when data is available.

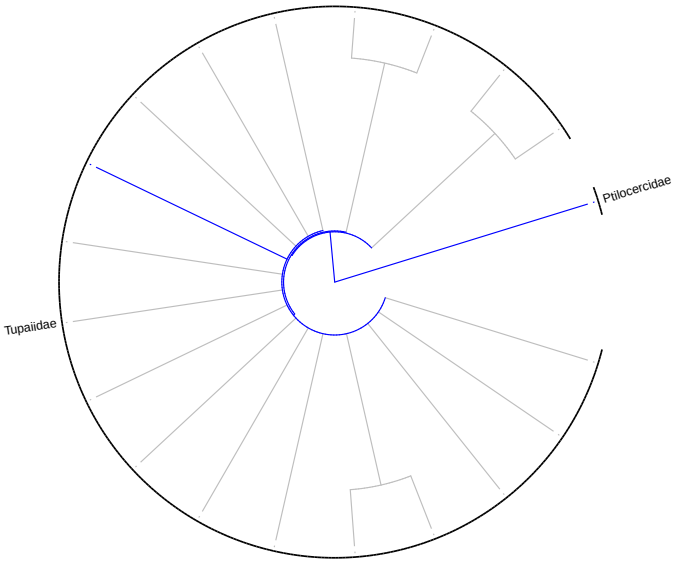

FIGURE B.11: Distribution of available morphological data across Scandentia. Edges are colored in grey when no morphological data is available or in blue when data is available.

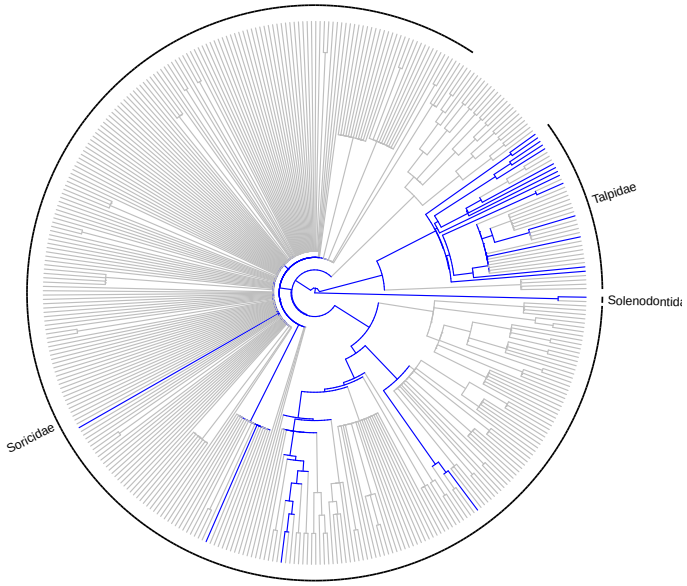

FIGURE B.12: Distribution of available morphological data across Soricomorpha. Edges are colored in grey when no morphological data is available or in blue when data is available.