LEABHARLANN CHOLÁISTE NA TRÍONÓIDE, BAILE ÁTHA CLIATH | TRINITY COLLEGE LIBRARY DUBLIN
Ollscoil Átha Cliath | The University of Dublin

**Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin**

**Copyright statement**

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

**Liability statement**

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

**Access Agreement**

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

# MACROEVOLUTION WITH LIVING AND FOSSIL SPECIES

*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

DEAN OF GRADUATE STUDIES
11 NOV 2015
TRINITY COLLEGE, DUBLIN

Ph.D.
Natural Sciences

SEPTEMBER 2015

CC-BY-SA

TRINITY LIBRARY
27 JUL 2016
DUBLIN

Thesis 11022

## 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.

## SUMMARY

Although many biodiversity studies focus on living species, the vast majority of species that ever lived are long extinct. It is therefore crucial to combine data from both living and fossil species to fully understand macroevolutionary patterns and processes. This thesis focuses on ways to combine both living and fossil taxa into phylogenies and investigates how the resulting phylogenies can be used to investigate macroevolutionary questions.

In the first part of the thesis, I ran extensive simulation analyses to test the effect of missing data on phylogenetic topologies when using the Total Evidence method. This method builds phylogenies using both molecular data for living taxa and morphological data for living and fossil taxa. I tested how various proportions of missing morphological data among living taxa, fossil taxa, and the two combined, affected my ability to recover the correct tree topology. I found that the amount of missing morphological data among living taxa was the most crucial aspect for accurately placing living and fossil taxa in the same phylogeny. Following these conclusions, I performed a systematic review of the coded morphological data available for living mammal species. I recorded the amount of morphological data available for each mammalian order and tested whether this data was randomly distributed across the phylogeny or biased towards certain clades. The results of this analysis showed that although morphological data is scarce for living mammals, it is at least generally randomly distributed across the phylogeny and therefore should not bias the placement of fossil taxa towards particular clades.

For the second part of the thesis, I used Total Evidence phylogenies containing both living and fossil taxa to investigate whether mammals radiated during the Cenozoic in response to the infamous Cretaceous-Palaeogene (K-Pg) mass extinction event, 66 million years ago. Previous studies show support for an effect of the K-Pg extinction event on mammalian diversification when using palaeontological data but no support using neontological data. I used a novel time-slicing method for quantifying changes in morphological diversity (disparity) through time to describe the patterns of mammalian diversification across the K-Pg boundary. I found no significant difference in disparity before and after the K-Pg boundary. This suggests that, even though many terrestrial vertebrates (including the non-avian dinosaurs) went extinct during the K-Pg extinction event, it had no significant effect on mammalian morphological diversification. These results refute the popular belief that mammals only began diversifying after the extinction of the non-avian dinosaurs, and shows the advantage of using living and fossil species to answer macroevolutionary questions.

Finally, I discuss future avenues of research for improving analyses that include living and fossil species as well as the advantages of using both living and fossil taxa when investigating macroevolutionary questions. I argue that all macroevolutionary studies should include both types of data to advance our understanding of biodiversity.

# ACKNOWLEDGEMENTS

This thesis would not have been as enjoyable to put together without the help of the people mentioned below. It would have not been possible at all, however, without the devotion of my supervisor Natalie Cooper. Thank you for the countless time you spent patiently helping/coaching/teaching/mentoring me but also for being always motivated and positive about this whole project (especially when it wasn't going as planned) and all that without loosing your sanity. I've learned a lot from working with you both in terms of skills and of the attitude to have regarding academia and am ever grateful for both.

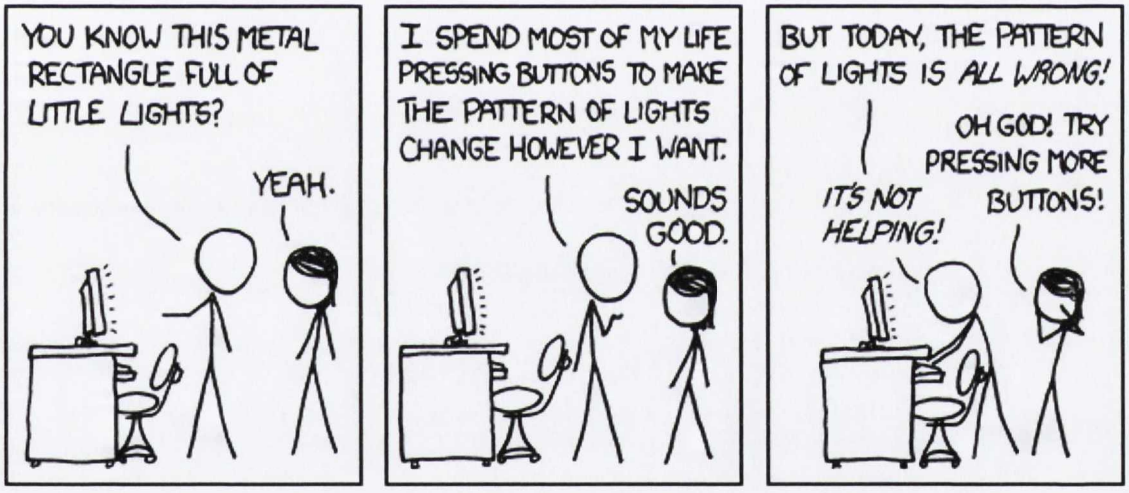
Other people also brought invaluable insight to this thesis: Andrew Jackson who was always around chatting about the latest Mortal Kombat test or a new smart statistical move (or the other way around); Gavin Thomas for kindly spending time commenting on my projects (especially Chapter 2); Graeme Lloyd for the fast and enthusiastic responses to my questions (especially Chapter 4); Rich FitzJohn for enforcing his strict, yet crucial, coding best practices (and for providing this thesis template); Trevor Hodkinson and Paula Murphy for following me through the thesis Steering Committee; Dave Bapst, Nick Matzke, April Wright and the people from the Macro JC for providing a fertile ground for new ideas.

A great thanks also to my office mates, Sive Finlay (PhD deirfiúr), Kevin Healy, Adam Kane and "Marvin" Qiang Yang for kindly supporting my random Frenchings. Excellent support also came from the staff, post-docs and students of the School of Natural Sciences who made me feel at home since day one in Ireland: Mirjam Ansorge, Yvonne Buckley, Joe Colgan, Luca Coscieme, Shaun Coutts, Ian Donohue, Alain Finn, Dave Kelly, Seán Kelly, Deirdre McClean, Keith McMahon, Darren O'Connell, Conor Owens, Marcin Penk, Maureen Williams, Mike Whitfield and all the NERD club members.

People from the Université Montpellier 2 are also to be accounted for much in my motivation to undertake a PhD: Julien Claude, Frédéric Delsuc, Emmanuel Douzery and Laurent Marivaux; and, of course, Franck Barbière, Sébastien Enault, Sylvain Monteux, Morgan Le Ny, Benjamin Ramassamy and all the other friends from the masters as well.

And finally, I'd like to thank my folks, Marja, Jean-Philippe, Nina and Daan for not always understanding what all that was about (and asking me to explain more clearly) but still always supporting my decision to undertake long and non-applied studies.

This PhD was funded by the Trinity Postgraduate Research Studentship.




xkcd.com/722 - CC BY-NC 2.5

# Table of Contents


Declaration . . . . . i

Summary . . . . . ii

Acknowledgements . . . . . iv

Table of Contents . . . . . vi

List of Tables . . . . . viii

List of Figures . . . . . ix

1 General Introduction . . . . . 1
- 1.1 Chapter outlines . . . . . 3
  - 1.1.1 Total Evidence method and missing data . . . . . 3
  - 1.1.2 Using Total Evidence phylogenies to ask macroevolutionary questions 3

2 Total Evidence method and missing data . . . . . 5
- 2.1 Introduction . . . . . 7
- 2.2 Methods . . . . . 10
  - 2.2.1 Generating the matrix . . . . . 12
  - 2.2.2 Removing data . . . . . 14
  - 2.2.3 Estimating phylogenies . . . . . 15
  - 2.2.4 Comparing topologies . . . . . 18
  - 2.2.5 Testing the effects of the missing data parameters on topological recovery 21
- 2.3 Results . . . . . 22
  - 2.3.1 Individual effects of missing data parameters . . . . . 23
  - 2.3.2 Combined effect of missing data parameters . . . . . 30
- 2.4 Discussion . . . . . 33
  - 2.4.1 Individual effects of missing data parameters . . . . . 33
  - 2.4.2 Combined effect of missing data parameters . . . . . 35
  - 2.4.3 Effects of tree inference methods . . . . . 35
  - 2.4.4 Practical implications . . . . . 37
  - 2.4.5 Conclusions . . . . . 38

3 Missing data in living mammals . . . . . 40
- 3.1 Introduction . . . . . 41
- 3.2 Methods . . . . . 42
  - 3.2.1 Data collection and standardisation . . . . . 42
  - 3.2.2 Data availability and distribution . . . . . 45
- 3.3 Results . . . . . 46
- 3.4 Discussion . . . . . 51

4 Spatio-temporal disparity in mammals at the K-Pg boundary . . . . . 54
- 4.1 Introduction . . . . . 56
- 4.2 Methods . . . . . 58
  - 4.2.1 Cladistic data and phylogenies . . . . . 58
  - 4.2.2 Estimating ancestral character states . . . . . 59
  - 4.2.3 Building the cladisto-space . . . . . 61
  - 4.2.4 Calculating disparity . . . . . 62
  - 4.2.5 Estimating disparity through time . . . . . 63

4.2.6 Testing the effects of the K-Pg extinction event on mammalian disparity 65
4.3 Results . . . 66
4.4 Discussion . . . 66
4.4.1 Methodological improvements for measuring disparity through time . . 72
4.4.2 Conclusion . . . 73

5 Discussion . . . 74
5.1 The future of the Total Evidence method . . . 74
5.2 The cladisto-space as a proxy for describing macroevolutionary changes . . 76
5.3 Future directions for combining living and fossil species in macroevolutionary studies . . . 78

Bibliography . . . 80

Appendices . . . 91

A Supplementary information to Chapter 2 . . . 91
A.1 Differences between the "true" and the inferred trees . . . 91
A.2 Tree inference software settings . . . 95

B Supplementary information to Chapter 3 . . . 96
B.1 Number of taxa with available cladistic data for mammalian orders without any character threshold . . . 96
B.2 Distribution of living species with available cladistic data for each mammalian order . . . 100

C Supplementary information to Chapter 4 . . . 103
C.1 Rarefied results for each dataset . . . 103
C.2 Comparison of different disparity metrics and time sampling methods . . . 106

# List of Tables


Table 2.1 Effect of using the "true" tree as a starting tree . . . . . 18
Table 2.2 Summary of the tree comparisons for the $M_L$, $M_F$ and $N_C$ parameters . . 27
Table 2.3 Bhattacharyya Coefficients of the pairwise method comparisons ($M_L$ and $M_F$) . . . . . 28
Table 2.4 Bhattacharyya Coefficients of the pairwise method comparisons ($N_C$ and all parameters combination) . . . . . 29

Table 3.1 Number of taxa with available cladistic data for mammalian orders . . . . 47

Table 4.1 Effect of the K-Pg boundary on mammlian disparity . . . . . 69
Table 4.2 Effect of the K-Pg boundary on mammlian disparity (rarefied) . . . . . 70

Table B.1 Number of taxa with available cladistic data for mammalian orders without any character threshold . . . . . 96

# List of Figures


Fig. 2.1 Protocol outline . . . . . 11
Fig. 2.2 Empirical proportion of morphological characters . . . . . 13
Fig. 2.3 Effect of using the "true" tree as a starting tree . . . . . 18
Fig. 2.4 Bhattacharyya Coefficient calculation outline 1 . . . . . 23
Fig. 2.5 Bhattacharyya Coefficient calculation outline 2 . . . . . 24
Fig. 2.6 Effects of increasing missing data on topological recovery . . . . . 25
Fig. 2.7 Effects of increasing missing data on topological recovery with parameters combination . . . . . 26
Fig. 2.8 Effects of missing data on topological recovery using Maximum Likelihood and Bayesian consensus trees . . . . . 31
Fig. 2.9 Effects of missing data on topological recovery using Maximum Likelihood bootstraps and Bayesian posterior distribution trees . . . . . 32

Fig. 3.1 Google searches additional OTUs rarefaction curve . . . . . 44
Fig. 3.2 Taxonomic matching algorithm used in this study . . . . . 45
Fig. 3.3 Phylogenetic distribution of species with available cladistic data across Primates and Carnivora . . . . . 51

Fig. 4.1 Mammaliaformes phylogeny . . . . . 59
Fig. 4.2 Eutheria phylogeny . . . . . 60
Fig. 4.3 Disparity through time in Eutheria and Mammaliaformes . . . . . 67
Fig. 4.4 Disparity through time in Eutheria and Mammaliaformes (rarefied) . . . . 68

Fig. A.1 Comparison between the "true" and the "best" trees . . . . . 92
Fig. A.2 Comparison between the "true" and the "missing-data" trees for the Maximum Likelihood an Bayesian consensus trees . . . . . 93
Fig. A.3 Comparison between the "true" and the "missing-data" trees for the bootstraps and posterior distribution trees . . . . . 94

Fig. B.1 Available data in Afrosoricida, Cetartiodactyla, Chiroptera and Dasyuromorphia 100
Fig. B.2 Available data in Dasyuromorphia, Didelphimorphia, Diprotodontia and Erinaceomorpha . . . . . 101
Fig. B.3 Available data in Pilosa, Rodentia, Scandentia and Soricomorpha . . . . 102

Fig. C.1 Mammaliaformes disparity (rarefied) . . . . . 104
Fig. C.2 Eutheria disparity (rarefied) . . . . . 105
Fig. C.3 Comparison of all the disparity metrics and all the time sampling methods for Mammaliaformes . . . . . 108
Fig. C.4 Comparison of all the disparity metrics and all the time sampling methods for Eutheria . . . . . 109
Fig. C.5 Comparison of all the disparity metrics and all the time sampling methods for Mammaliaformes (rarefied) . . . . . 110
Fig. C.6 Comparison of all the disparity metrics and all the time sampling methods for Eutheria (rarefied) . . . . . 111

## DATA AVAILABILITY AND REPRODUCIBILITY

### TOTAL EVIDENCE AND MISSING DATA

All the code used in this analysis is available on GitHub (`https://github.com/TGuillerme/Total_Evidence_Method-Missing_data`) with some information on how to use the various functions. Additionally all the simulated data is available on Figshare (`http://dx.doi.org/10.6084/m9.figshare.1306861`).

### MISSING DATA IN LIVING MAMMALS

All data and analysis code is available on GitHub (`https://github.com/TGuillerme/Missing_living_mammals`).

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

Data is available on Figshare (`http://dx.doi.org/10.6084/m9.figshare.1539545`). Code for reproducing the analysis is available on GitHub (`https://github.com/TGuillerme/SpatioTemporal_Disparity`).

CHAPTER 1

# General Introduction

Today's diversity of living organisms represents an overwhelmingly small fraction of the organisms that have ever existed (Raup, 1981; Novacek and Wheeler, 1992). Nonetheless, although it is widely accepted that the patterns of biodiversity we observe today are influenced by evolutionary history (Simpson, 1944; Gingerich, 1987; Archibald, 2011), much research focuses solely on living or fossil species separately. This narrow focus can lead to misinterpretation of macroevolutionary patterns and processes (Fritz et al., 2013; Benton, 2015). For example, Wiens (2015) suggest that terrestriality is a driver of diversification among living vertebrates, a pattern essentially driven by Aves (birds), Squamata (lizards and snakes) and Mammalia. Living crocodilians are also included in the analysis but because they constitute a species-poor group (25 species; Uetz, 2010), living in only a few types of environment (marine or freshwater; Martin, 2008), they have only a marginal effect on the conclusion of the study. However, extinct crocodilians were much more diverse than present-day species, both in terms of species richness (at least 244 species are reported in Bronzati et al., 2015) and the environments they lived in (extinct crocodilian species ranged from fully marine to fully terrestrial, and even included a few tree-dwelling species; Stubbs et al., 2013). Therefore, by not including fossil species, Wiens (2015) conceals the true history of this clade, and thus, potentially biases the conclusions of the study.

Including fossil species not only accounts for groups that were more diverse in the past, it also 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), speciation scenarios (e.g. revealing hidden vicariance patterns; Wood et al., 2013) or niche occupancy through time (e.g. Pearman et al., 2008). These studies have led to increasing consensus among evolutionary biologists that we need to combine both living and fossil species in macroevolutionary analyses (Quental and Marshall, 2010; Dietl and Flessa, 2011; Slater and Harmon, 2013; Fritz et al., 2013; Benton, 2015).

Testing macroevolutionary hypotheses on extinct and extant species simultaneously first requires a phylogenetic tree for the group of interest containing both fossil and living taxa. Unfortunately few studies have actively focused on building such trees and most have been published in just the last five years (e.g. Ronquist et al., 2012a; Slater, 2013; Beck and Lee, 2014). The scarcity of trees combining living and fossil species is probably due to the fact that palaeontologists and neontologists have a different approach for inferring phylogenies.

Palaeontological phylogenies are generally based on cladistic data from the fossil record (i.e. discrete morphological observations). Phylogenetic reconstructions then rely on optimality criteria such as maximum parsimony (Hennig, 1966; Felsenstein, 2004) to resolve relationships among lineages and on stratigraphy to date the trees (Goloboff et al., 2008). This allows a direct interpretation of macroevolution in deep time and benefits from recent increased data collection efforts (e.g. 4541 characters in O'Leary et al., 2013, introducing the term "phenomics") and improvements in tree dating methods (e.g. the *cal3* method from Bapst, 2014). However, palaeontological studies rarely take into account all of living diversity (e.g. only 38 out of 351 living primates are included with 119 fossils in Ni et al., 2013) and these methods suffer from several biases (e.g. evolution is not parsimonious; Wright and Hillis, 2014).

Conversely, neontological studies use the vast amount of available molecular data from living species and probabilistic methods (e.g. Maximum Likelihood or Bayesian Inference) to build phylogenies. These methods are based on evolutionary models that rely on the differences in DNA to resolve relationships among lineages and on some specific fossil occurrence dates for dating the trees (i.e. the molecular clock; Zuckerkandl and Pauling, 1965). There have been extensive improvements in these tree building methods in the last decade in both the evolutionary models (e.g. Bapst, 2013; Stadler and Yang, 2013; Heath et al., 2014) and in how fossils are used to time calibrate the trees (Donoghue and Benton, 2007; Parham et al., 2012). However, this approach uses only the ages of certain fossils instead of all the information available from the fossil record (e.g. species richness, traits, biogeography, etc.). What we really need to move the field of macroevolution forward are phylogenies containing both living and fossil taxa.

Encouragingly, the last three years have seen many improvements in the 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); a method that combines molecular data from living taxa and morphological data from living and fossil taxa in the same phylogenies. It was first developed in the nineties (Eernisse and Kluge, 1993) but only recently successfully implemented in user-friendly phylogenetic software (Ronquist et al., 2012b; Bouckaert et al., 2014). By using all the available neontological and palaeontological data, this method can greatly improve the estimation of divergence events (Ronquist et al., 2012a), evolutionary rates (Beck and Lee, 2014), tree topology (Dembo et al., 2015), trait evolution (Slater, 2013) and even speciation processes (Wood et al., 2013).

In this thesis, I explore the Total Evidence method in terms of its benefits and drawbacks for studying macroevolution. In the first part of the thesis (Chapters 2 and 3), I investigate the practical implications of combining paleontological and neontological data by focusing on the effect of missing data on tree topology using simulations. I also estimate how much

morphological data is missing from existing data on living mammals. In the second part of the thesis (Chapter 4), I use Total Evidence phylogenies to explore the effect of mass extinctions on morphological evolution, using mammals as an example.

## 1.1 CHAPTER OUTLINES

### 1.1.1 *Total Evidence method and missing data*

As introduced above, the Total Evidence method is a promising method for combining living and fossil species into phylogenies. There is, however, one drawback to this method: because it needs both molecular data for living taxa and morphological data for living and fossil taxa, Total Evidence phylogenies are likely to have a large proportion of missing data. In Chapter 2, I therefore investigate the problem of missing data in Total Evidence matrices. I perform extensive simulations to test how sensitive topologies inferred from Total Evidence matrices are to missing data in the morphological partition of the matrix, by removing data according to three parameters: (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 overall number of morphological characters in the matrix. I then build phylogenies from the complete matrices, and matrices with varying amounts of missing data, using both Maximum Likelihood and Bayesian inference methods. Finally, I compare how my missing data parameters and their interactions, as well as the phylogenetic inference method, influence the ability to estimate the correct tree topology.

One of the main conclusions of Chapter 2 is that to recover accurate topologies, we need as much morphological data for living species as possible. However, no estimates of the amount of morphological data already coded for living species exist. Therefore, in Chapter 3, I assess the availability of morphological data in the literature for living mammals. I download available morphological matrices and count the number of living mammals with available morphological data at three different taxonomic levels (species, genus and family) for each mammalian order. I then measure whether the missing data are biased toward specific clades in each order using methods adapted from community phylogenetics (Webb et al., 2002).

### 1.1.2 *Using Total Evidence phylogenies to ask macroevolutionary questions*

Chapters 2 and 3 focus on the technical and practical aspects of combining living and fossil taxa in the same phylogenies. However, ideally we not only want to build these phlyogenies, but also to use them to investigate interesting macroevolutionary questions. Until now, only a few studies have used Total Evidence phylogenies in macroevolutionary studies (e.g. Wood

et al., 2013; Slater, 2013; Dembo et al., 2015). Therefore, in Chapter 4 I tackle a classical macroevolutionary question using Total Evidence phylogenies.

One example of an interesting macroevolutionary pattern is the shift in ecologically dominant clades through time due to drastic biotic or abiotic changes (e.g. climate change, mass extinctions, land bridge formations etc.). For example, the Brachiopoda were the dominant shelled filter-feeding clade during the Paleozoic (514 to 252 million years ago; Ma) but were replaced by Bivalvia at the end Permian extinction event (252 Ma) so that Bivalvia is now the dominant group (Sepkoski 1981; Clapham et al. 2006; Liow et al. 2015 but see Payne et al. 2014). This type of replacement pattern has also been observed in other groups such as Foraminifera (Coxall et al., 2006), Ichthyosauria (Thorne et al., 2011) and Plesiosauria (Benson and Druckenmiller, 2014) and are often related to competition (Brusatte et al., 2008a) 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 Ma. In Chapter 4, I focus on this example, updating classical analyses using Total Evidence phylogenies and various methodological improvements.

I investigate changes in 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 event had an effect on mammalian diversification. I propose a new approach to describe patterns of disparity through time based 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 underlying models of morphological evolution (e.g. punctuated or gradual; Hunt et al., 2015).

Finally, in Chapter 5, I draw together the results from Chapters 2, 3 and 4 and discuss how the results of this thesis opens new avenues for research. I then discuss the limitations of my analyses, and suggest improvements for future studies. I also present some concluding thoughts on the utility of combining palaeontological and neontological data to improve our understanding of macroevolutionary patterns and processes.

CHAPTER 2

TOTAL EVIDENCE METHOD AND MISSING DATA

# Effects of missing data on topological inference using a Total Evidence approach [1,2]


ABSTRACT

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

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

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

We find that the number of living taxa with morphological characters and the overall number of morphological characters in the matrix, are more important than the amount of missing data in the fossil record for recovering the "best" tree topology. Therefore, we suggest that sampling effort should be focused on morphological data collection for living species to increase the accuracy of topological inference in a Total Evidence framework. Additionally, we find that Bayesian method consistently outperforms other tree inference


[1] A similar version of this chapter has been published as: "Guillerme, T & Cooper, N. 2016. Effects of missing data on topological inference using a Total Evidence approach. **Molecular Phylogenetics and Evolution**, 94:146-158; doi: doi:10.1016/j.ympev.2015.08.023".

[2] *Author contributions*: I designed the study, ran the analyses and wrote the paper. NC helped design the study and commented on drafts of the manuscript.

methods. We therefore recommend using Bayesian consensus trees to fix the tree topology prior to further analyses.

## 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 proportion of missing data that may affect the method's ability to infer correct topologies, branch lengths and support values (Salamin et al., 2003).

Although missing data do not appear be a major problem in molecular and morphological matrices separately (as long as enough data overlap in each case, and missing data are not phylogenetically biased; Wiens, 2003; Wiens et al., 2005; Wiens, 2006; Wiens and Moen, 2008; Lemmon et al., 2009; Sanderson et al., 2011; Roure and Philippe, 2011; Pattinson et al., 2014), it may become more of an issue in Total Evidence matrices containing both molecular and morphological data for living and fossil taxa. This may be particularly problematic as fossil taxa (generally) do not have molecular data, resulting in a large section of missing data in Total Evidence matrices. Until now, few attempts have been made to study the impact of this missing data issue on phylogenetic inference in a Total Evidence framework (i.e. using both molecular and morphological data; Wiens et al., 2005; Manos et al., 2007; Pattinson et al., 2014). These previous studies assessed the effect of missing data on topology by either (1) comparing a dataset with missing data to subsets without missing data (Wiens et al., 2005); or (2) removing both molecular and some morphological data from living taxa to create artificial fossils (Manos et al., 2007; Pattinson et al., 2014). Both approaches have shown that missing data are not a major problem and should not be an obstacle to combining both living and fossil species in the same phylogenies. The way these studies were conducted, however, means that their conclusions are not generally applicable across all scenarios involving missing data in Total Evidence phylogenies. For example, using an empirical (rather than simulation based) approach limits their conclusions to studies with similar distributions of data across species in the phylogeny. Additionally, one of the three previous studies did not include fossil taxa in their analyses, so their results cannot be used to make conclusions about how missing data may influence the placement of fossils (Wiens, 2003). The other two studies did include fossil taxa, but used the patchiness of the fossil record to determine how to remove data from their matrices (Manos et al., 2007; Pattinson et al., 2014). Data for living species are unlikely to be missing in this patchy way, instead full molecular data with the complete absence of morphological data is a likely pattern (see Chapter 3; Guillerme and Cooper 2015). Finally, these previous studies mainly

focused on how missing data in fossil taxa affect the placement of fossils, ignoring the effects of missing data in living species (Manos et al., 2007; Pattinson et al., 2014).

In this study, we propose a theoretical assessment of the effect of missing data in the Total Evidence method by removing living taxa with morphological data, fossil data, all data for certain characters and the combination of these three aspects. This is an advance on previous studies because we use large-scale simulations and analyse the effects of three distinct aspects of missing data thus focusing on both neontological and 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 wild-card 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 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.

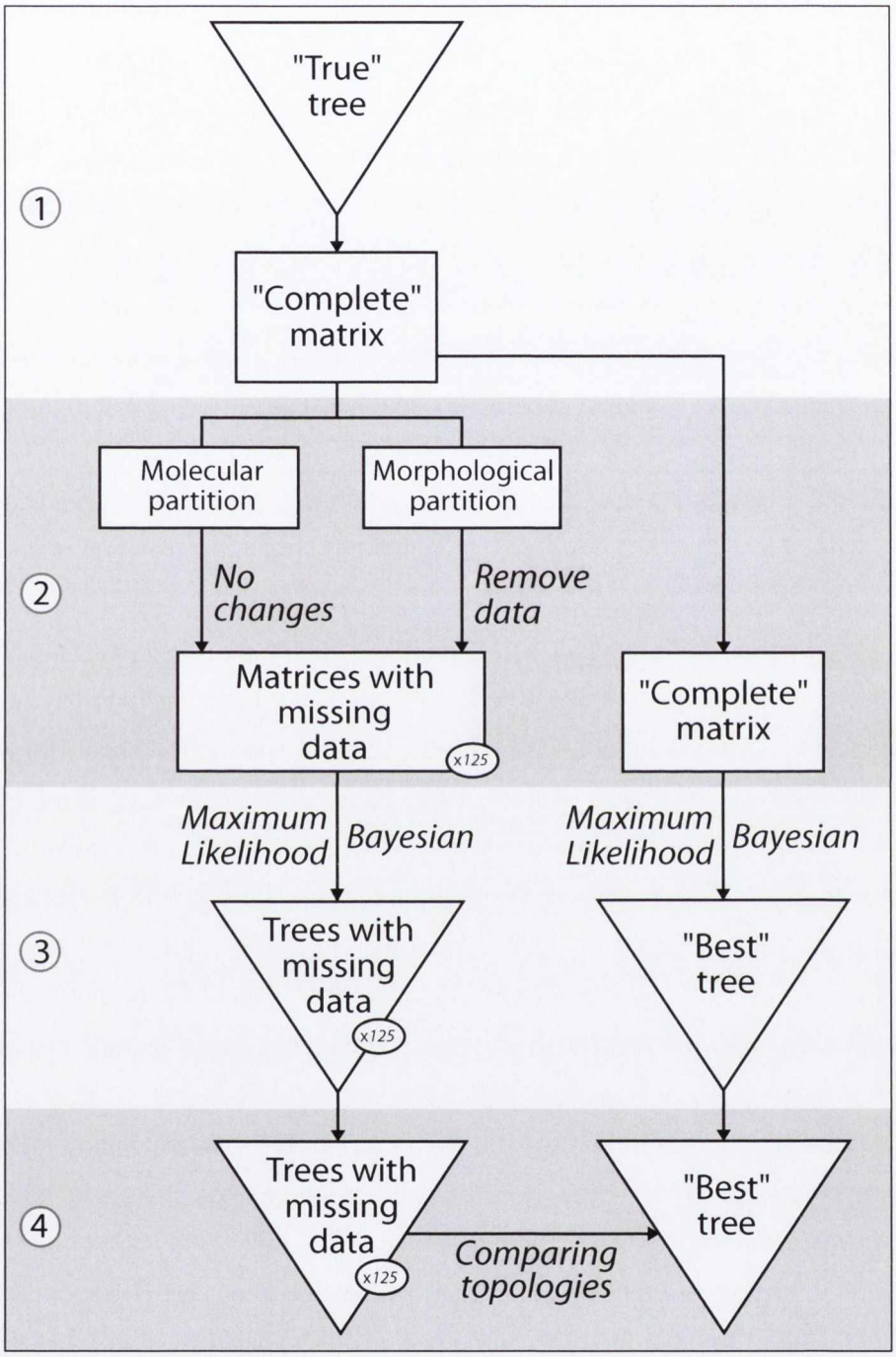



FIGURE 2.1: **Protocol outline.** (1) We randomly generated a birth-death tree (the "true" tree) and used it to simulate a matrix with no missing data (the "complete" matrix). (2) We removed data from the morphological part of the "complete" matrix resulting in 125 "missing-data" matrices. (3) We built 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.

### 2.2.1 *Generating the matrix*

First we randomly generated a "true" tree of 50 taxa in R v. 3.0.2 (R Core Team, 2015) using the package `diversitree` v. 0.9-6 (FitzJohn, 2012). We generated the tree using a birth death process by sampling speciation ($\lambda$) and extinction ($\mu$) rates from a uniform distribution (bounded between 0 and 1) but maintaining $\lambda > \mu$ (Paradis, 2011). Empirical Total Evidence matrices vary in whether they have more fossil than living taxa or vice versa. For example, fossil taxa make up 88% (Beck and Lee, 2014), 58% (Schrago et al., 2013), 48% (Pyron, 2011), 31% (Ronquist et al., 2012a) and 31% (Slater, 2013) of taxa in various studies. To avoid biasing our simulations towards either living or fossil taxa and to make each simulation comparable, we implemented a rejection sampling algorithm to select only trees with 25 living and 25 fossil taxa. The fossil taxa were considered as unique tips at the end of extinct lineages. We then added an outgroup to the tree, using the mean branch length of the tree to separate the outgroup from the rest of the taxa, and with the branch length leading to the outgroup set as the sum of the mean branch length and the longest root-to-tip length of the tree.

Next, we generated a molecular and a morphological matrix from the "true" tree. The molecular matrix was simulated from the "true" tree using the R package `phyclust` v. 0.1-14 (Chen, 2011). The matrix contained 1000 character sites for 51 taxa and was generated using the seqgen algorithm (Rambaut and Grassly, 1997) and using the HKY model (Hasegawa et al., 1985) with random base frequencies (sampled from a uniform probability distribution bounded between 0 and 1 with the total frequency for the four bases equal to 1) and transition/transversion rate of two (Douady et al., 2003). The substitution rates were selected from a gamma distribution with an ($\alpha$) 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 $\alpha$ 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 (whiting 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).

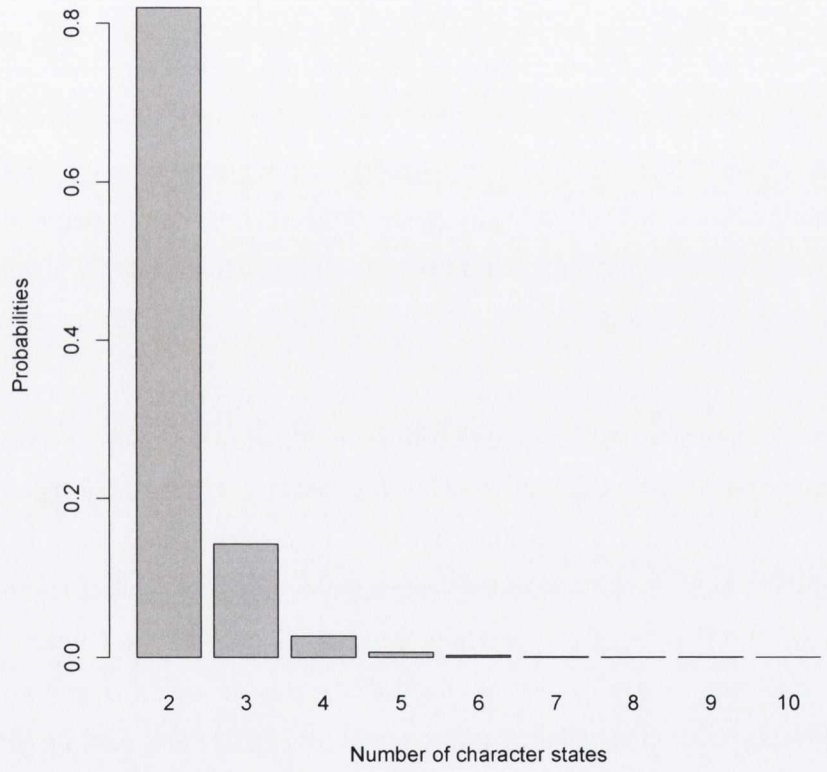



FIGURE 2.2: The proportion of morphological characters with between two and 10 character states extracted from 100 randomly selected empirical matrices downloaded from TreeBASE.

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 $\alpha = 0.5$. We used low evolutionary rate

parameters to be consistent with the molecular rate parameters, to avoid homoplasy in the morphological part of the matrix and create a clear phylogenetic signal (Wright and Hillis, 2014). Topological error has been shown to be minimal at a morphological rate of 0.5 when using the M*kv* model (Lewis, 2001; Wright and Hillis, 2014). Note, however, that Wright and Hillis (2014) have shown that low morphological rates ($< 0.5$) increase variance in topological error, but we discarded simulations with such topological error by selecting only matrices with a "fair" phylogenetic signal (see section 2.2.3; 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 (see Chapter 3; Guillerme and Cooper 2015). Therefore, many living species will have only molecular data available. In practice, we removed all the morphological data from randomly chosen living taxa with five different proportions: 0%, 10%, 25%, 50% or 75% of living taxa with no morphological data.

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

data at this parameter value (down to a minimum threshold of 5% available data; see below).

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

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

We created matrices using all parameter combinations resulting in 125 ($5^3$) "missing-data" matrices. Note that one of these combinations ($M_L$=0%; $M_F$=0% and $N_C$=100) has no missing data so is equivalent to the "complete" matrix, thus we have one effectively complete matrix in our 125 "missing-data" matrices. In practice, we first removed the data following the two missing data parameters $M_L$ and $M_F$ and then removed data following the $N_C$ parameters. To avoid matrices containing taxa without any data (morphological or molecular), we repeated the random deletion until the matrices contained at least 5% of data for any 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 Supplementary data A.1. 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 + $\Gamma_4$ 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 ($\mu$) following a gamma distribution of the rate variation with four distinct categories (M*kv* + $\Gamma_4$; -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 + $\Gamma_4$). For the morphological data, we used the M*kv* model (Lewis, 2001), with equal state frequency and a unique overall substitution rate ($\mu$) with four distinct rates categories (M*kv* + $\Gamma_4$). 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 $5\times10^7$ 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).

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 $\alpha$ = 0.5, and a transition/transversion ratio prior of two sampled from a strong beta distribution ($\beta$(80,40)); and one prior on the morphological part of the matrix (exponential prior on the shape of the gamma distribution of $\alpha$ = 0.5). We used these priors to speed up the Bayesian estimation process. These priors biased the way the Bayesian process calculated branch lengths by giving non-random starting points and boundaries for parameter estimation however, here we are focusing on the effect of missing data on tree topology and not branch lengths. Even using these priors, it took $\sim$ 140 CPU years to build 50 sets of 125 Bayesian trees (2.30GHz clock speed nodes). The detailed MrBayes parameters are available in Supplementary data A.2.

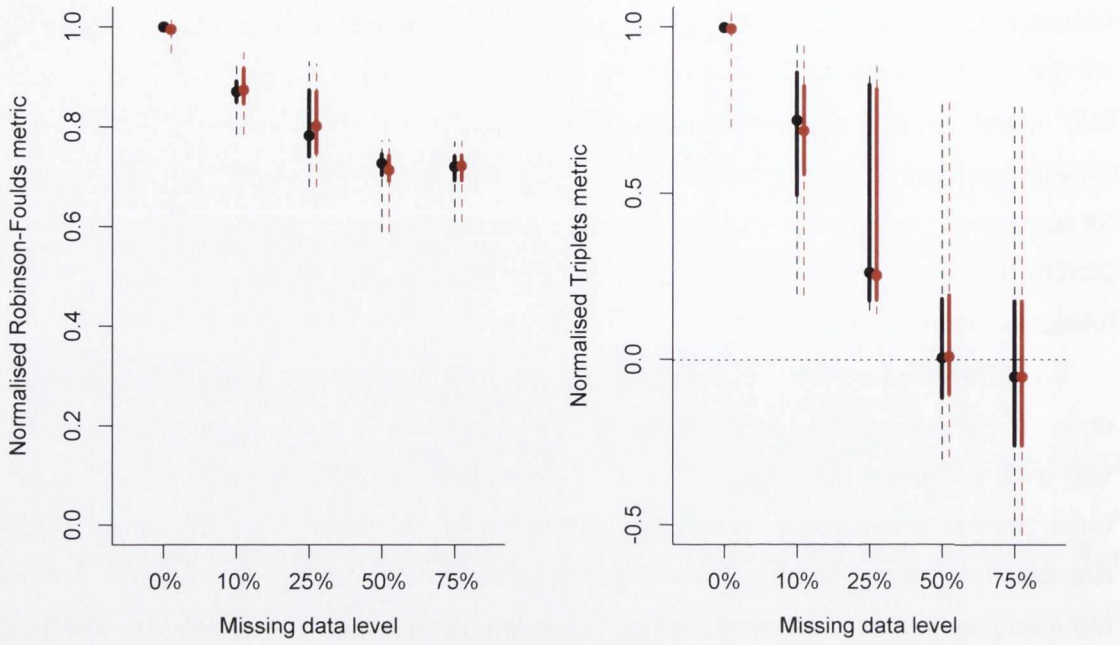



FIGURE 2.3: Effect of using the "true" tree (black) or a random tree (red) as the starting tree for the Bayesian inference. The x axis, represents the amount of missing data (see below).

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 | p value |
|---|---|---|---|---|---|---|
| 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 | | |

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, 2015) 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:

$$\sum_{ijk} = \binom{n}{3} = \frac{n!}{3!(n-3)!} \tag{2.4}$$

and where $n$ is the total number of taxa in both trees (Critchlow et al., 1996). If $S_n = 0$, the trees are identical; when $S_n = \binom{n}{3}$, 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 (full overlap; Bhattacharyya, 1943). 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 thresholds 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 and 2.9). 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 and 2.9).

To assess the effect of the different tree inference and uncertainty methods (i.e. Maximum Likelihood, Bayesian consensus, Maximum Likelihood bootstrap trees and Bayesian posterior tree distribution) on our ability to recover the "best" tree topology, we calculated the Bhattacharyya Coefficient between the distributions of the different metrics scores (Normalised Robinson-Foulds and Normalised Triplets metric) for each pairwise combination of tree inference and uncertainty methods, i.e. Maximum Likelihood *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 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).

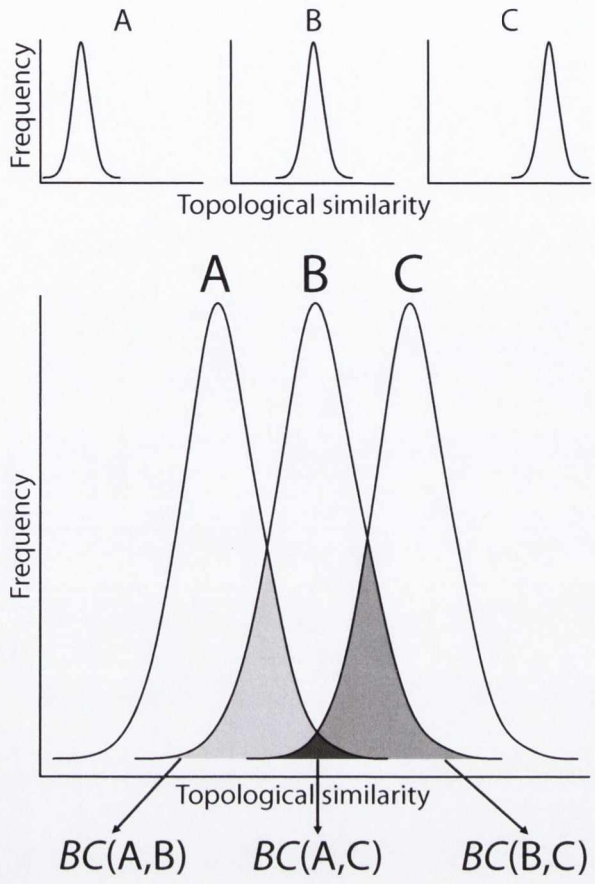



FIGURE 2.4: **Bhattacharyya Coefficient calculation outline 1.** A, B and C are distributions of tree similarity metrics (Normalised Robinson-Foulds or Normalised Triplets metrics) for any combination of missing data parameters (e.g. $M_L$ = 10%, $M_F$ = 50%, $N_C$ = 25). The Bhattacharyya Coefficient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

### 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-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; Table 2.2). 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; Table 2.2).

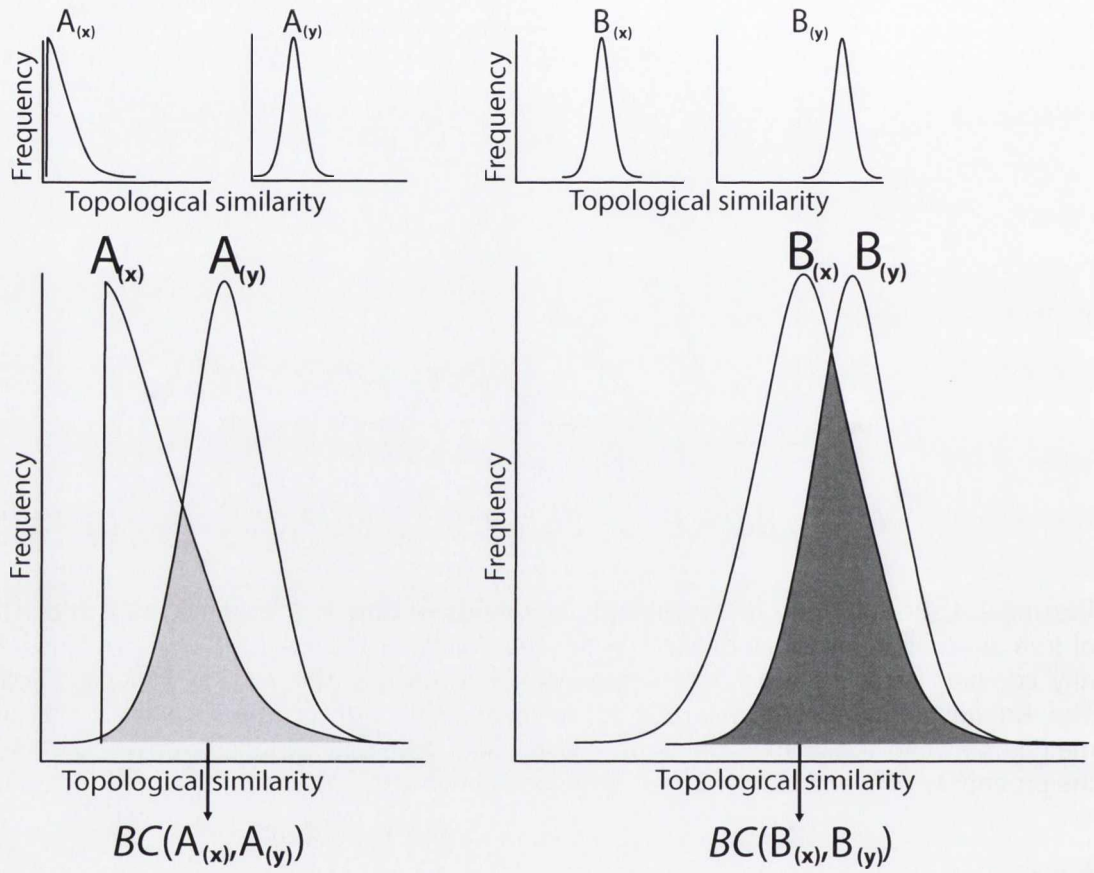



FIGURE 2.5: **Bhattacharyya Coefficient calculation outline 2.** A and B are distributions of tree similarity metrics (Normalised Robinson-Foulds or Normalised Triplets metrics) for any combination of missing data parameters (e.g. $M_L$ = 10%, $M_F$ = 50%, $N_C$ = 25). (x) and (y) are two different tree 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).

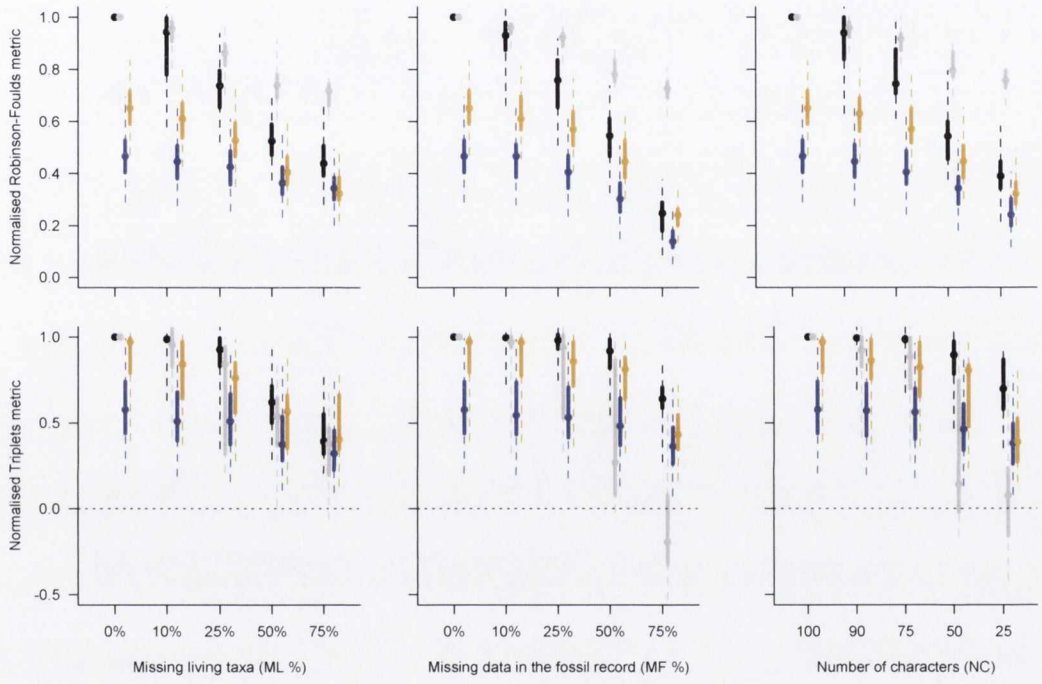



FIGURE 2.6: **The effects of increasing missing data on topological recovery** using Maximum Likelihood trees (black), Bayesian consensus trees (grey), Maximum Likelihood bootstrap trees (blue) and Bayesian posterior tree distributions (orange). The percentage of missing data for each parameter ($M_L$, $M_F$ 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.

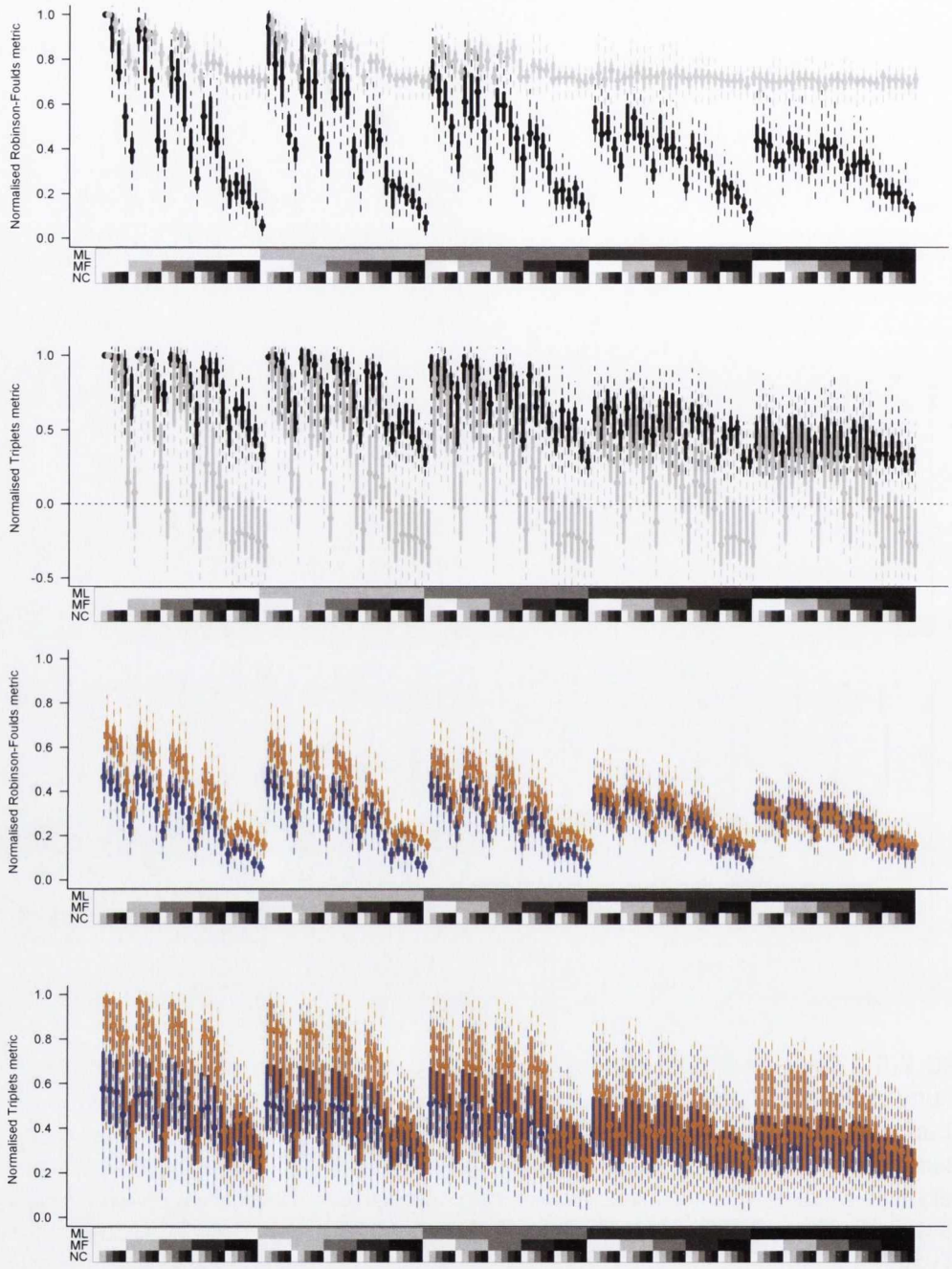



FIGURE 2.7: **The effects of increasing missing data on topological recovery** using Maximum Likelihood trees (black), Bayesian consensus trees (grey), Maximum Likelihood bootstrap trees (orange) and Bayesian posterior tree distribution (blue). The x axis shows the percentage of missing data from 0% (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.

TABLE 2.2: Summary of the comparisons between the "best" tree and the "missing-data" trees for each different tree inference method using either the Normalised Robinson-Foulds metric ($RF$) or the Normalised Triplets metric ($Tr$) for each parameters separately.

| Tree inference method | Metric | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|---|
| **$M_L$ missing data parameter** | | | | | | | |
| Maximum Likelihood | $RF$ | 0.44 | 0.51 | 0.63 | 0.66 | 0.78 | 0.95 |
| | $Tr$ | 0.45 | 0.56 | 0.76 | 0.74 | 0.93 | 0.99 |
| Bayesian consensus | $RF$ | 0.71 | 0.73 | 0.80 | 0.82 | 0.88 | 0.95 |
| | $Tr$ | 0.37 | 0.46 | 0.67 | 0.67 | 0.87 | 0.96 |
| Maximum Likelihood bootstraps | $RF$ | 0.34 | 0.37 | 0.42 | 0.41 | 0.44 | 0.46 |
| | $Tr$ | 0.32 | 0.40 | 0.51 | 0.46 | 0.51 | 0.57 |
| Bayesian posterior tree distributions | $RF$ | 0.33 | 0.41 | 0.52 | 0.50 | 0.60 | 0.65 |
| | $Tr$ | 0.41 | 0.56 | 0.76 | 0.71 | 0.84 | 0.98 |
| **$M_F$ missing data parameter** | | | | | | | |
| Maximum Likelihood | $RF$ | 0.23 | 0.46 | 0.64 | 0.61 | 0.79 | 0.93 |
| | $Tr$ | 0.65 | 0.84 | 0.95 | 0.89 | 0.99 | 1.00 |
| Bayesian consensus | $RF$ | 0.72 | 0.77 | 0.86 | 0.85 | 0.94 | 0.96 |
| | $Tr$ | -0.16 | 0.19 | 0.63 | 0.52 | 0.96 | 0.98 |
| Maximum Likelihood bootstraps | $RF$ | 0.14 | 0.30 | 0.40 | 0.35 | 0.45 | 0.46 |
| | $Tr$ | 0.37 | 0.49 | 0.54 | 0.51 | 0.56 | 0.57 |
| Bayesian posterior tree distributions | $RF$ | 0.24 | 0.45 | 0.57 | 0.51 | 0.63 | 0.65 |
| | $Tr$ | 0.44 | 0.81 | 0.86 | 0.82 | 0.98 | 0.98 |
| **$N_C$ missing data parameter** | | | | | | | |
| Maximum Likelihood | $RF$ | 0.40 | 0.50 | 0.64 | 0.65 | 0.79 | 0.94 |
| | $Tr$ | 0.70 | 0.84 | 0.93 | 0.89 | 0.99 | 1.00 |
| Bayesian consensus | $RF$ | 0.76 | 0.79 | 0.86 | 0.86 | 0.92 | 0.96 |
| | $Tr$ | 0.05 | 0.16 | 0.53 | 0.50 | 0.87 | 0.92 |
| Maximum Likelihood bootstraps | $RF$ | 0.25 | 0.34 | 0.42 | 0.38 | 0.45 | 0.46 |
| | $Tr$ | 0.38 | 0.47 | 0.55 | 0.51 | 0.57 | 0.58 |
| Bayesian posterior tree distributions | $RF$ | 0.32 | 0.44 | 0.58 | 0.52 | 0.62 | 0.65 |
| | $Tr$ | 0.39 | 0.78 | 0.82 | 0.79 | 0.98 | 0.98 |

TABLE 2.3: **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. |
|---|---|---|---|---|---|---|---|
| **$M_L$ missing data parameter** | | | | | | | |
| Maximum Likelihood *vs.* Bayesian consensus | $RF$ | 0.30 | 0.31 | 0.69 | 0.61 | 0.77 | **1.00** |
| | $Tr$ | 0.79 | 0.81 | 0.84 | 0.86 | 0.85 | **1.00** |
| Maximum Likelihood *vs.* Maximum Likelihood bootstraps | $RF$ | **0.03** | 0.22 | 0.29 | 0.36 | 0.54 | 0.69 |
| | $Tr$ | 0.08 | 0.42 | 0.53 | 0.51 | 0.74 | 0.78 |
| Maximum Likelihood *vs.* Bayesian posterior trees | $RF$ | **0.02** | 0.49 | 0.61 | 0.51 | 0.67 | 0.74 |
| | $Tr$ | 0.21 | 0.61 | 0.70 | 0.63 | 0.81 | 0.81 |
| Bayesian consensus *vs.* Maximum Likelihood bootstraps | $RF$ | **0.01** | **0.02** | **0.02** | **0.02** | **0.03** | **0.04** |
| | $Tr$ | 0.08 | 0.69 | 0.78 | 0.64 | 0.79 | 0.84 |
| Bayesian consensus *vs.* Bayesian posterior trees | $RF$ | **0.01** | **0.02** | **0.02** | **0.04** | 0.08 | 0.09 |
| | $Tr$ | 0.21 | 0.74 | 0.75 | 0.68 | 0.84 | 0.87 |
| Bayesian posterior tree *vs.* Maximum Likelihood bootstraps | $RF$ | 0.69 | 0.75 | 0.85 | 0.85 | **0.95** | **1.00** |
| | $Tr$ | 0.91 | 0.92 | **0.96** | **0.95** | **0.97** | **0.98** |
| **$M_F$ missing data parameter** | | | | | | | |
| Maximum Likelihood *vs.* Bayesian consensus | $RF$ | **0.00** | 0.25 | 0.48 | 0.50 | 0.76 | **1.00** |
| | $Tr$ | 0.38 | 0.69 | 0.75 | 0.72 | 0.80 | **1.00** |
| Maximum Likelihood *vs.* Maximum Likelihood bootstraps | $RF$ | **0.03** | 0.18 | 0.32 | 0.36 | 0.47 | 0.77 |
| | $Tr$ | 0.08 | 0.34 | 0.40 | 0.38 | 0.53 | 0.55 |
| Maximum Likelihood *vs.* Bayesian posterior trees | $RF$ | **0.02** | 0.47 | 0.71 | 0.60 | 0.86 | 0.94 |
| | $Tr$ | 0.21 | 0.54 | 0.62 | 0.56 | 0.64 | 0.80 |
| Bayesian consensus *vs.* Maximum Likelihood bootstraps | $RF$ | **0.00** | **0.00** | **0.01** | **0.01** | **0.01** | **0.03** |
| | $Tr$ | 0.08 | 0.38 | 0.54 | 0.49 | 0.70 | 0.75 |
| Bayesian consensus *vs.* Bayesian posterior trees | $RF$ | **0.00** | **0.02** | **0.02** | **0.02** | **0.04** | **0.04** |
| | $Tr$ | 0.21 | 0.29 | 0.66 | 0.54 | 0.72 | 0.82 |
| Bayesian posterior tree *vs.* Maximum Likelihood bootstraps | $RF$ | 0.69 | 0.69 | 0.72 | 0.71 | 0.72 | 0.72 |
| | $Tr$ | 0.91 | 0.91 | 0.91 | 0.93 | 0.92 | **0.98** |

TABLE 2.4: **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. |
|---|---|---|---|---|---|---|---|
| **$N_C$ missing data parameter** | | | | | | | |
| Maximum Likelihood *vs.* Bayesian consensus | $RF$ | **0.03** | 0.32 | 0.66 | 0.55 | 0.75 | **1.00** |
| | $Tr$ | 0.51 | 0.69 | 0.80 | 0.76 | 0.80 | **1.00** |
| Maximum Likelihood *vs.* Maximum Likelihood bootstraps | $RF$ | **0.03** | 0.17 | 0.21 | 0.31 | 0.46 | 0.68 |
| | $Tr$ | 0.08 | 0.31 | 0.39 | 0.39 | 0.56 | 0.61 |
| Maximum Likelihood *vs.* Bayesian posterior trees | $RF$ | **0.02** | 0.44 | 0.47 | 0.52 | 0.78 | 0.90 |
| | $Tr$ | 0.21 | 0.52 | 0.59 | 0.55 | 0.66 | 0.77 |
| Bayesian consensus *vs.* Maximum Likelihood bootstraps | $RF$ | **0.00** | **0.01** | **0.01** | **0.02** | **0.02** | **0.03** |
| | $Tr$ | 0.08 | 0.47 | 0.62 | 0.51 | 0.66 | 0.73 |
| Bayesian consensus *vs.* Bayesian posterior trees | $RF$ | **0.00** | **0.02** | **0.04** | **0.04** | 0.05 | 0.06 |
| | $Tr$ | 0.21 | 0.45 | 0.64 | 0.57 | 0.74 | 0.79 |
| Bayesian posterior tree *vs.* Maximum Likelihood bootstraps | $RF$ | 0.69 | 0.73 | 0.73 | 0.76 | 0.81 | 0.86 |
| | $Tr$ | 0.91 | 0.92 | 0.93 | 0.94 | **0.96** | **0.99** |
| **All missing data parameters combined** | | | | | | | |
| 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** |

### 2.3.2 *Combined effect of missing data parameters*

As expected, our ability to recover the "best" tree topology is worst when each parameter contains the maximum amount of missing data (i.e. $M_L$ = 75%, $M_F$ = 75% and $N_C$ = 75%), and best when there is no missing data (i.e. $M_L$ = 0%, $M_F$ = 0%, $N_C$ = 0%; Fig. 2.7; Table 2.2). Fig. 2.8 and 2.9 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. 2.8C-D). The Normalised Robinson-Foulds metric (median Bhattacharrya coefficient = 0.79; blue regions in Fig. 2.8C), however, displays more dissimilarities than the Normalised Triplets metric (median Bhattacharrya coefficient = 0.81; blue regions in Fig. 2.8D). The orange upper triangle in figure 2.8C shows a high probability of overlap of the Normalised Robinson-Foulds metric for the trees with the $M_L$ parameter $\geq$ 50% (Fig. 2.8C). 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. 2.8C). 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 in Fig. 2.8C). This can be interpreted as the overlap between the distributions once $M_L$=50% in figure 2.7 for the Normalised Robinson-Foulds metric (upper panel).

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; Tables 2.3 and 2.4). 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; Tables 2.3 and 2.4; Fig. 2.6).

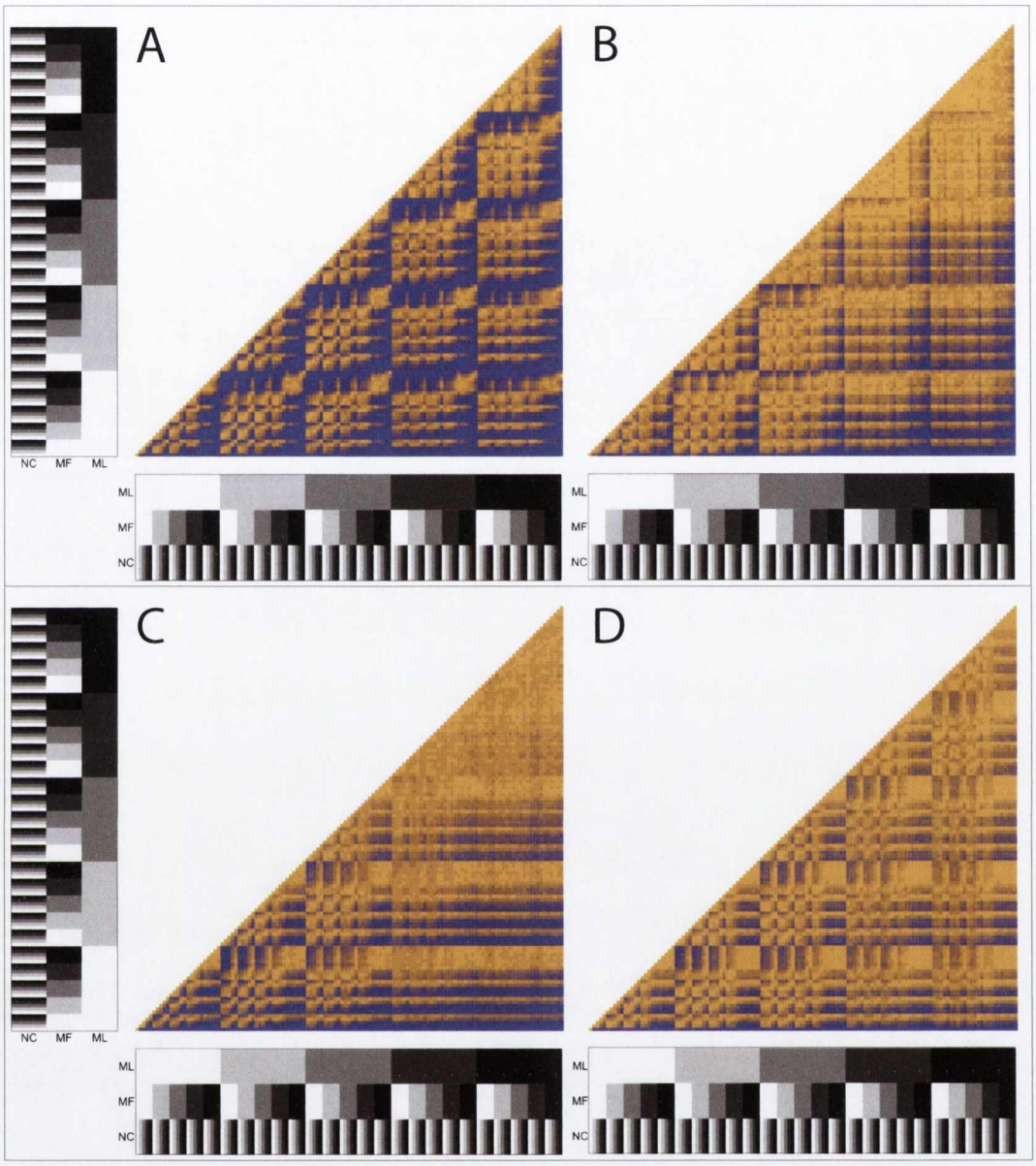



FIGURE 2.8: **The effects of missing data on topological recovery using Maximum Likelihood and Bayesian consensus trees.** Both axes show the percentage of missing data from 0% (white) to 75% (black) for the three parameters: $M_L$ (upper line), $M_F$ (middle line) and $N_C$ (lower line). The first row (A and B) corresponds to the Maximum Likelihood trees and the second row (C and D) to the Bayesian consensus trees. The topological recovery is measured as (A and C) the Normalised Robinson-Foulds metric and (B and D) 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).

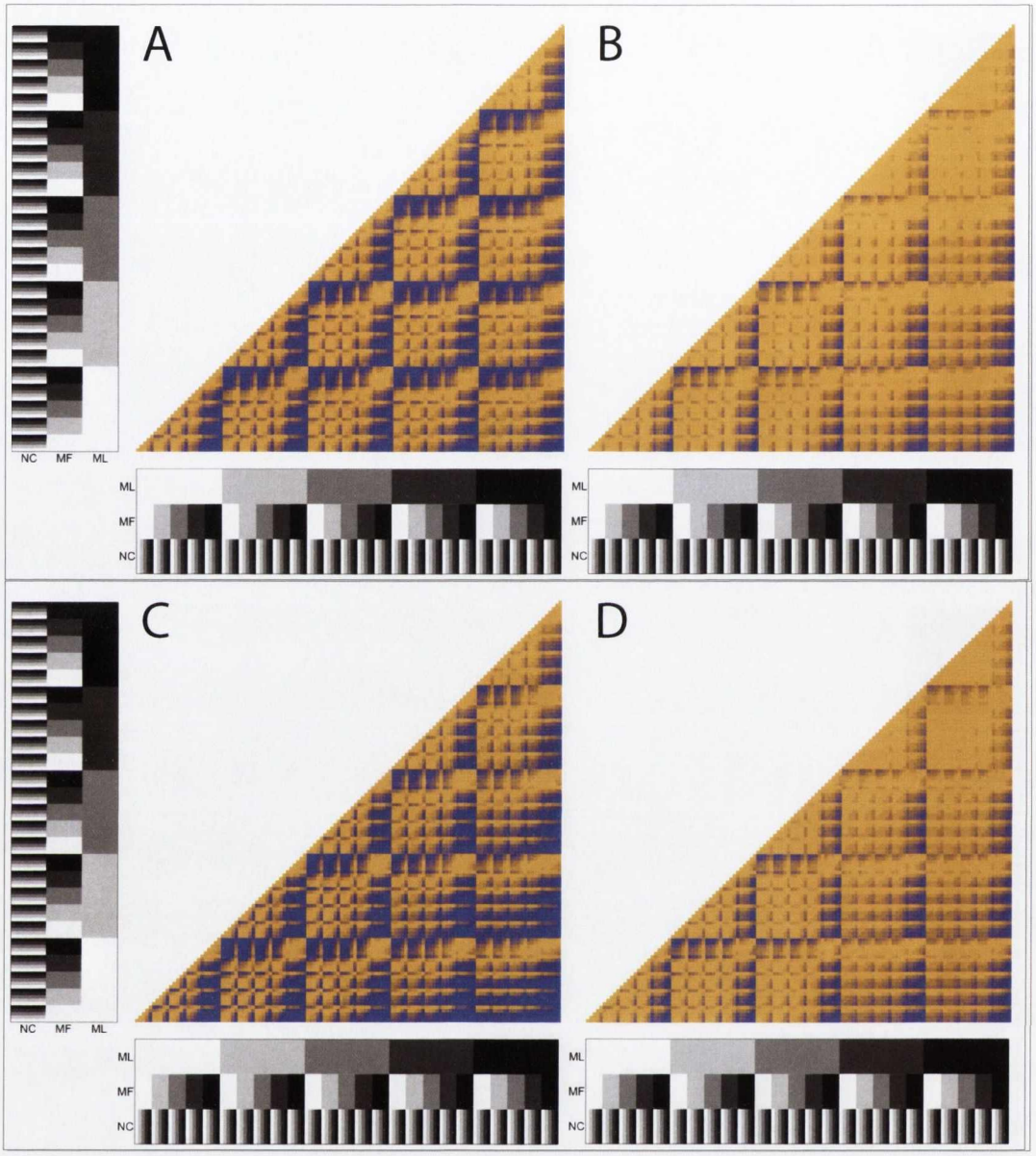



FIGURE 2.9: **The effects of missing data on topological recovery using Maximum Likelihood bootstraps and Bayesian posterior distribution trees.** Both axes show the percentage of missing data from 0% (white) to 75% (black) for the three parameters: $M_L$ (upper line), $M_F$ (middle line) and $N_C$ (lower line). The first row (A and B) corresponds to the Maximum Likelihood bootstrap trees and the second row (C and D) to the Bayesian posterior distribution trees. The topological recovery is measured as (A and C) the Normalised Robinson-Foulds metric and (B and D) 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).

## 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 and 2.9). 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, Tables 2.3 and 2.4). 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; Ni et al., 2013), they are still within the size range of more modestly-sized empirical matrices (e.g. Kelly et al., 2014; Sallam et al., 2011). Therefore, our simulations reflect realistic parameters. Nonetheless, the use of probabilistic methods (i.e. Maximum Likelihood or Bayesian) and the M*kv* model (Lewis, 2001) has been previously demonstrated to partially resolve this issue (Wright and Hillis, 2014).

### 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 Supplementary data A.1). 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, Tables 2.2, 2.3 and 2.4). 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. Ni et al., 2013). Fortunately, although these data are the most difficult to collect, they also have the least influence on whether our simulations recover the "best" tree topology (Fig. 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* Swofford et al., 2001). 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).

### 2.4.5 *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.

## Assessment of cladistic data availability for living mammals [3,4]

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.

[3] 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 as: "Guillerme, T & Cooper, N. 2015. Assessment of cladistic data availability for living mammals. **bioRxiv**; doi: dx.doi.org/10.1101/022970".

[4] *Author contributions*: I designed the study, collected the data, ran the analyses and wrote the paper. NC helped design the study and commented on drafts of the manuscript.

## 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 (Chapter 2; Guillerme and Cooper, 2016), 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 (Chapter 2; Guillerme and Cooper, 2016).

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 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 operational

taxonomic units (OTU) 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 all 601 matrices 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 SCHOLAR — 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.

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 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).

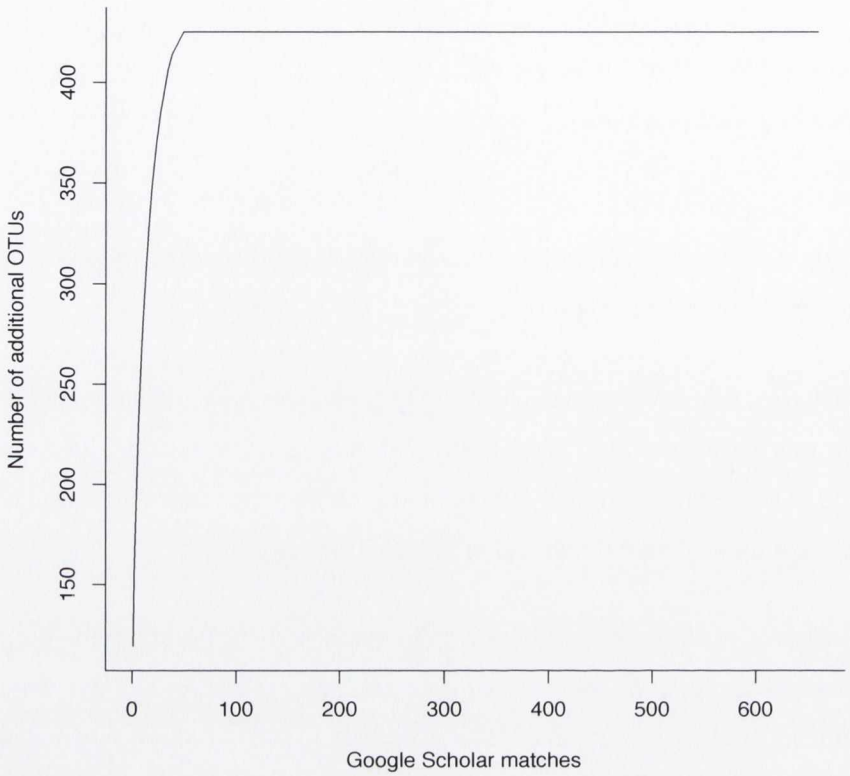



FIGURE 3.1: Google searches additional OTUs rarefaction curve. The x axis represent the number of google scholar matches (papers, books or abstracts) and the y axis represents the cumulative number of additional living OTUs per google scholar match.

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 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 Chapter 2 (Guillerme and Cooper, 2016) 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.

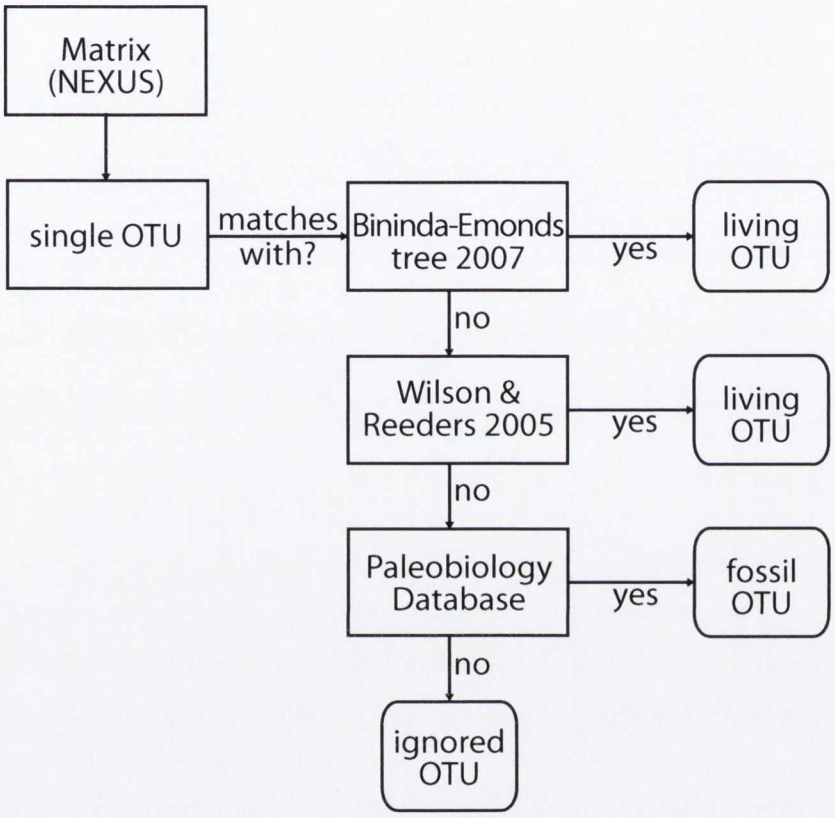



FIGURE 3.2: **Taxonomic matching algorithm used in this study**. For each matrix, each operational taxonomic units (OTU) is matched with the super tree from Bininda-Emonds et al. (2007). If the OTU matches, then it is classified as living. Else it is matched with the Wilson and Reeder (2005) taxonomy list. If the OTU matches, then it is classified as living. Else it is matched with the Paleobiology Database list of fossil 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 but more than one OTU at any taxonomic level, we investigated whether the available cladistic data was (i) randomly distributed, (ii) overdispersed 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:

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

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 $MNND$ 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:

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

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 one of the order pruned from (Bininda-Emonds et al., 2007).

## 3.3 RESULTS

Across the 126 cladistic matrices we analysed, 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, Chiroptera 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 figure 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).

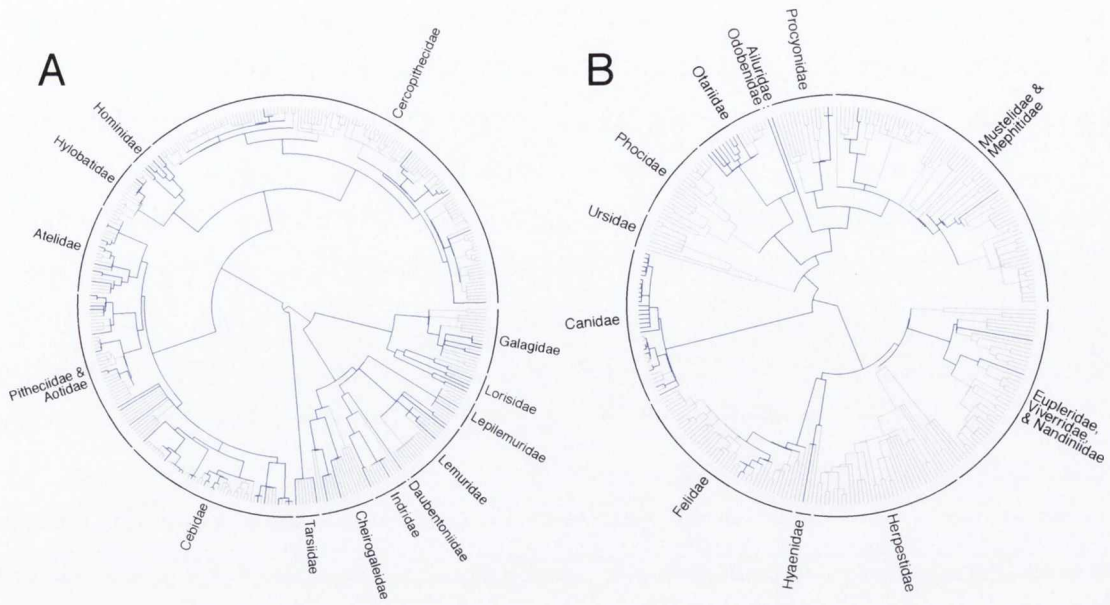



FIGURE 3.3: Phylogenetic distribution of species with available cladistic data across two mammalian orders (A: Primates; B: Carnivora). Edges are colored in grey when no cladistic data are available for a species and in blue when data are available.

## 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 (Chapter 2; Guillerme and Cooper, 2016). 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 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 (Chapter 2; Guillerme and Cooper, 2016). Beyond this threshold, there is considerable displacement of wildcard taxa (*sensu* Kearney, 2002) and decreases in clade conservation (Chapter 2; Guillerme and Cooper, 2016). 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 use 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 in the right clade. The second best scenario is that species with cladistic data are randomly distributed across the phylogenies. In this scenario we expect no special bias in the placement of the fossil (Chapter 2; Guillerme and Cooper, 2016), 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 living 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 be 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, 2016; 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 (Chapter 2; Guillerme and Cooper, 2016), 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

## Mammalian morphological diversity does not increase in response to the Cretaceous-Paleogene mass extinction and the extinction of the (non-avian) dinosaurs. [5,6]




smbc-comics.com/?id=1535 – ©SMBC Zach Weinersmith

"The most erroneous stories are those we think we know best - and therefore never scrutinize or question." S.J. Gould[7]

[5]A similar version of this chapter has been submitted for publication.

[6]*Author contributions*: I designed the study, collected the data, ran the analyses and wrote the paper. NC helped design the study and commented on drafts of the manuscript.

[7]Gould, S.J. 1997. *Full House: The Spread of Excellence from Plato to Darwin.* Harmony Books: New York. p 57.

ABSTRACT

Popular science accounts state that after the extinction of the non-avian dinosaurs at the Cretaceous-Paleogene (K-Pg) boundary 66 million years ago, mammals rapidly diversified to fill their empty ecological niches. However, evidence for this is mixed. Palaeontological analyses suggest that mammals radiated in response to the K-Pg extinction event, whereas neontological analyses suggest that mammals began to radiate before K-Pg and were not greatly affected by it. Here we aim to end this debate by looking at living and fossil taxa simultaneously.

We investigated the effect of the K-Pg extinction event on mammalian morphological diversity (disparity) using two Total Evidence tip-dated phylogenies of Mammaliaformes and Eutheria, containing both living and fossil taxa. Using a novel, continuous time-slicing method for measuring changes in disparity-through-time, we found no significant change in disparity before and after the K-Pg boundary, under either a gradual or punctuated model of evolution. This implies that the extinctions at the end of the Cretaceous did not affect mammalian morphological evolution. Our findings contradict the popular theory that the non-avian dinosaurs and other Mesozoic tetrapods were restricting mammalian evolution, and that their extinction liberated ecological niches for mammals to evolve into.

## 4.1 INTRODUCTION

Throughout history, life on Earth has suffered a series of mass extinction events resulting in drastic declines in global biodiversity (e.g. Raup, 1979; Benton and Twitchett, 2003; Renne et al., 2013; Brusatte et al., 2015). The long-term effects of mass extinctions, however, are more varied (Erwin, 1998), and include species richness increases in some clades (Friedman, 2010) and declines in others (Benton, 1985), changes in morphological diversity (Ciampaglio et al., 2001; Ciampaglio, 2004; Korn et al., 2013) and shifts in ecological dominance (e.g. Brusatte et al., 2008b; Toljagic and Butler, 2013; Benson and Druckenmiller, 2014). These shifts are characterised by the decline of one clade that is replaced by a different unrelated clade with a similar ecological role (e.g. Brachiopoda and Bivalvia at the end Permian extinction; Liow et al. 2015 but see Payne et al. 2014). Shifts in ecological dominance are of particular interest because they are a fairly common pattern observed in the fossil record (e.g. Foraminifera; Coxall et al. 2006; Icthyosauria; Thorne et al. 2011; Plesiosauria; Benson and Druckenmiller 2014) and are often linked to major macroevolutionary processes such as adaptive (Losos, 2010) or competitive (Brusatte et al., 2008b) radiations.

One classical example of a shift in ecological dominance is at the Cretaceous-Palaeogene (K-Pg) mass extinction 66 million years ago (Renne et al., 2013), where many terrestrial vertebrates (including the dominant non-avian dinosaur group; Archibald 2011; Renne et al. 2013; Brusatte et al. 2015) went extinct, allowing placental mammals to dominate the fauna (Archibald, 2011; Lovegrove et al., 2014). Some authors suggest this reflects placental mammals filling the "empty" niches left after the K-Pg extinction event (Archibald, 2011; O'Leary et al., 2013), others suggest it reflects a release from predation and/or competition (Slater, 2013; Lovegrove et al., 2014). However, evidence for the diversification of placental mammals being driven by the K-Pg extinction event is mixed. Thorough analysis of the fossil record (e.g. Goswami et al., 2011; O'Leary et al., 2013) supports the idea that placental mammals diversified after the K-Pg extinction event as there are no undebated placental mammal fossils before it and many afterwards (Archibald, 2011; Goswami et al., 2011; Slater, 2013; O'Leary et al., 2013; Wilson, 2013; Brusatte et al., 2015). Conversely, evidence from molecular data suggests that the diversification of placental mammals started prior to the K-Pg extinction event without being drastically affected by it (e.g. Douady and Douzery, 2003; Bininda-Emonds et al., 2007; Meredith et al., 2011; Stadler, 2011). Therefore, whether the diversification of placental mammals began before the K-Pg extinction event, or in response to the extinctions at K-Pg, is a matter of great debate (dos Reis et al., 2012; O'Leary et al., 2013; Springer et al., 2013; O'Leary et al., 2013; dos Reis et al., 2014).

There are two main reasons why there is still debate about the timing of the diversification of placental mammals. Firstly, palaeontological and neontological data show different

patterns; palaeontological data generally suggest that placental mammals diversified after K-Pg (e.g. O'Leary et al., 2013), whereas neontological data suggest that K-Pg extinction event had little to no effect on mammalian diversification (Bininda-Emonds et al., 2007; Meredith et al., 2011; Stadler, 2011). We can solve this issue by using both palaeontological and neontological data in our analyses. The Total Evidence method allows us to use cladistic data for both living and fossil taxa, along with molecular data for living taxa, to build phylogenies (Ronquist et al., 2012a). This method can also be combined with the tip-dating method (Ronquist et al., 2012a; Wood et al., 2013) to get more accurate estimates of diversification times for both living and fossil species (but see Arcila et al., 2015). Here we use two recent Total Evidence tip-dated phylogenies of mammals (Slater, 2013; Beck and Lee, 2014) to investigate palaeontological and neontological taxa simultaneously.

A second issue is that diversity can be defined in many different ways. In many studies it is measured as taxonomic diversity or species richness (Stadler, 2011; Meredith et al., 2011; O'Leary et al., 2013), but often the more interesting aspect of diversity is related to the ecological niches the species occupy (Wesley-Hunt, 2005; Brusatte et al., 2008b; Toljagic and Butler, 2013), particularly if we want to make hypotheses about macroevolutionary processes (Pearman et al., 2008; Olson and Arroyo-Santos, 2009; Losos, 2010; Glor, 2010; Benton, 2015). Sometimes taxonomic diversity is used as a proxy for other kinds of diversity, however, species richness can be decoupled from morphological diversity (e.g. Slater et al., 2010; Ruta et al., 2013; Hopkins, 2013), so it may not be the best proxy for ecological diversity. We can instead use morphological diversity, also known as disparity (e.g. Wills et al., 1994; Erwin, 2007; Hughes et al., 2013), as a way to quantify changes in mammalian morphology that should relate to the ecology of the species. However some methods for measuring disparity are outdated and make inappropriate assumptions. Many methods for quantifying changes in morphological diversity were proposed more than 20 years ago (Foote, 1994; Wills et al., 1994) and are sometimes used without modifications (e.g., Brusatte et al., 2008a,b; Cisneros and Ruta, 2010; Thorne et al., 2011; Prentice et al., 2011; Brusatte et al., 2012; Toljagic and Butler, 2013; Ruta et al., 2013; Benton et al., 2014; Benson and Druckenmiller, 2014). Additionally, previous methods are based on an underlying assumption that changes in disparity occur by punctuated evolution (e.g. Wesley-Hunt, 2005) which is not always the case (Hunt et al., 2015). Finally, most studies of disparity through time use unequal time units based on biostratigraphy (Brusatte et al., 2008b, 2012; Toljagic and Butler, 2013). This can be tautological as biostratigraphy is already based on changes in fossil assemblages and morphology through time. To deal with these issues, we propose an updated approach to test whether mammals diversified in response to the K-Pg event, using morphological disparity, measured as cladistic disparity (see Methods, section 4.2.4), as our proxy for diversity.

Here we measure the disparity of living and fossil mammals before or after K-Pg, using data taken from two previously published studies (Slater, 2013; Beck and Lee, 2014). Using a novel time-slicing approach, we produce fine-grained estimates of disparity through time under two different models of morphological character evolution (either gradual or punctuated). We also test whether mammals display significant changes in disparity between the end of the Cretaceous and throughout the Cenozoic.

Until now, this question has only been investigated using data from North American Therian mammals (excluding Monotremata) and without formally testing the effect of the K-Pg extinction event (Wilson, 2013). To our knowledge, this study is the first to approach the debate about the effects of the K-Pg extinction event on mammalian evolution using Total Evidence phylogenies and by calculating disparity through time in a continuous way. We find no significant changes in mammalian disparity between the end of the Cretaceous and any time during the Paleocene. These results suggest that the extinction of non-avian dinosaurs and other terrestrial vertebrate clades at the end of the Cretaceous did not affect mammalian morphological evolution.

## 4.2 METHODS

### 4.2.1 *Cladistic data and phylogenies*

We used the cladistic morphological matrices and the Total Evidence tip-dated trees (Ronquist et al., 2012a) from Slater (2013, 103 taxa with 446 morphological characters;) and Beck and Lee (2014, 102 taxa with 421 morphological characters). We chose these two datasets because they have a similar number of taxa and morphological characters. Slater (2013) ranges from 310 million years ago (Ma; Late Carboniferous) to the present and focuses on the clade Mammaliaformes at the family-level and is called hereafter the Mammaliaformes dataset. Beck and Lee (2014) ranges from 170 Ma (Middle Jurassic) to the present and focuses on Eutheria at the genus-level and is called hereafter the Eutheria dataset. We used the first and last occurrences reported in Slater (2013) and Beck and Lee (2014) as the temporal range of each taxon in our analysis. Both phylogenies are illustrated in figures 4.1 and 4.2. Both trees contain few taxa compared to the overall species richness of living and fossil mammals (Bininda-Emonds et al., 2007; Archibald, 2011). This is because Total Evidence trees need a lot of data, particularly morphological data for living taxa that can be hard to locate (Chapter 2 and 3; Guillerme and Cooper, 2015, 2016). Therefore, most Total Evidence studies to date contain one or two orders of magnitude fewer taxa than phylogenies based solely on molecular data (e.g. thousands of taxa in Bininda-Emonds et al. 2007; Meredith et al. 2011 *vs.* hundreds in Ronquist et al. 2012a; Slater 2013; Beck and Lee 2014).

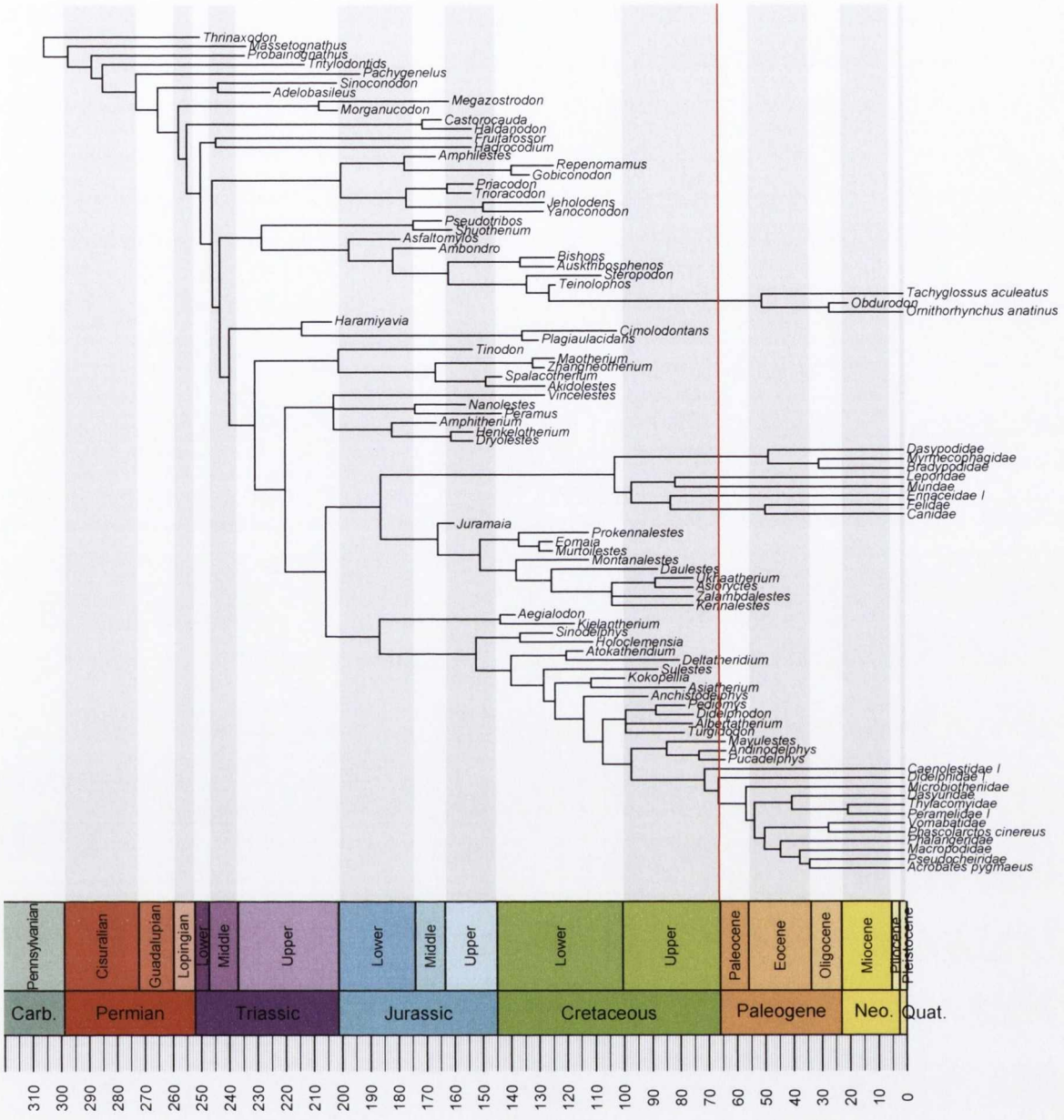



FIGURE 4.1: Mammaliaformes phylogeny from Slater (2013). The phylogeny only contains taxa with overlapping cladistic data. The vertical red line represents the K-Pg boundary. Carb. = Carboniferous; Quat. = Quaternary.

### 4.2.2 *Estimating ancestral character states*

For both datasets we used the re-rooting method (Yang et al., 1996; Garland and Ives, 2000) to get Maximum Likelihood estimates of the ancestral states for each character at every node in the tree, using the `rerootingMethod` function from the `R` package `phytools` version 0.4-45 (Revell, 2012; R Core Team, 2015). Where there was missing character data for a taxon we followed the method of Lloyd (2015) and treated missing data as any possible observed state for each character. For example, if a character had two observed states (0 and 1) across all taxa, we attributed the multi-state "0&1" value to the taxon with missing

data, representing an equal probability of being either $0$ or $1$. This allows the ancestral node of a taxon with missing data to be estimated with no assumptions other than that the taxon has one of the observed character states. To prevent poor ancestral state reconstructions from biasing our results, especially when a lot of error is associated with the reconstruction, we only included ancestral state reconstructions with a scaled Likelihood $\geq 0.95$. Ancestral state reconstructions with scaled Likelihoods below this threshold were replaced by missing data ("?").

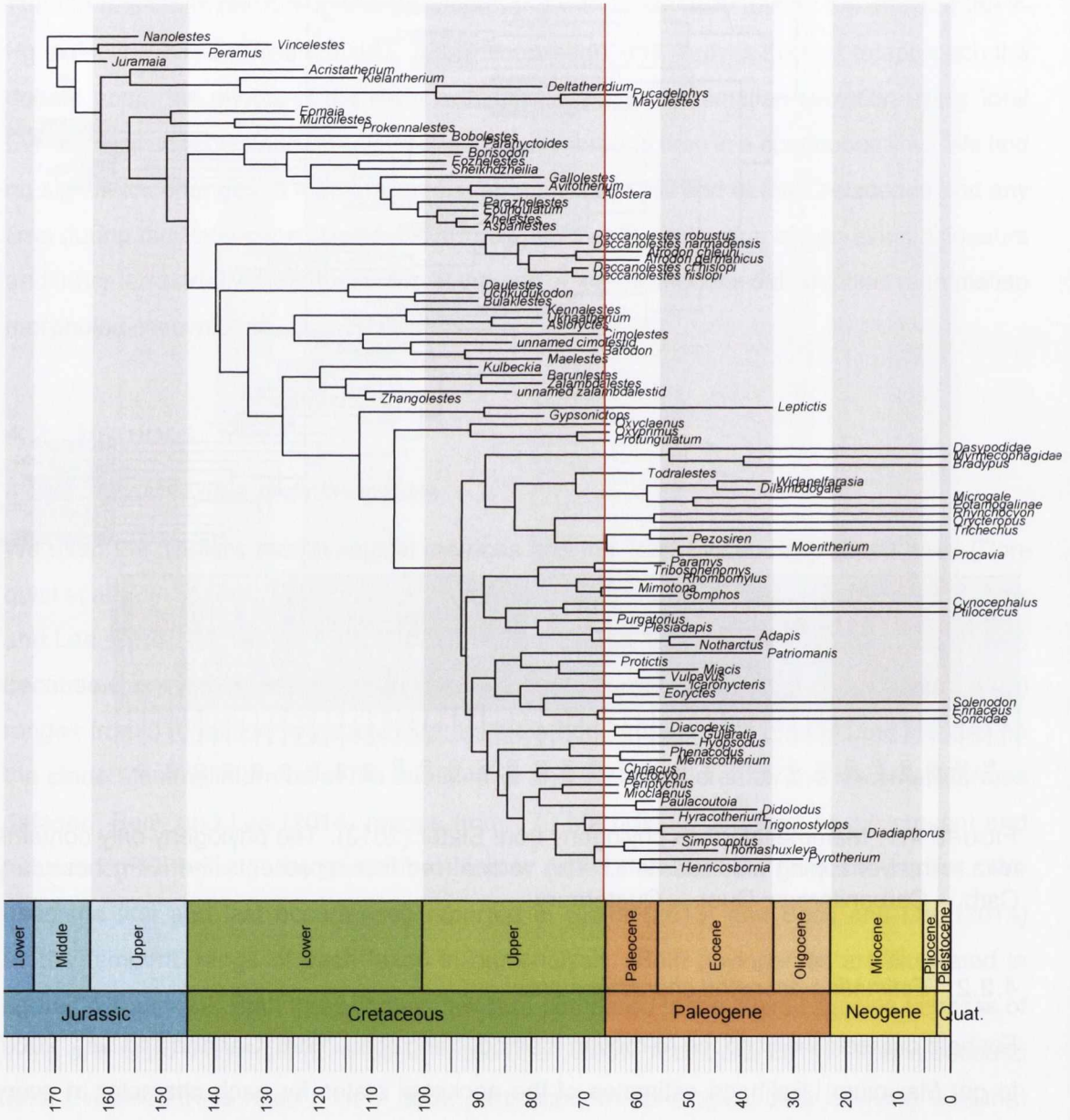



FIGURE 4.2: Eutheria phylogeny from Beck and Lee (2014). The phylogeny only contains taxa with overlapping cladistic data. The vertical red line represents the K-Pg boundary. Quat. = Quaternary.

### 4.2.3 *Building the cladisto-space*

To explore variations in mammalian disparity through time (defined here as the variation in morphologies through time), we used a cladisto-space approach (e.g. Foote, 1994, 1996; Wesley-Hunt, 2005; Brusatte et al., 2008b; Friedman, 2010; Toljagic and Butler, 2013; Hughes et al., 2013). This approach is similar to constructing a morphospace based on continuous morphological data (e.g. Friedman, 2010), except a cladisto-space is an approximation of the morphospace based on cladistic data (i.e. the discrete morphological characters used to build a phylogenetic tree). Mathematically, a cladisto-space is an $n$ dimensional object that summarises the cladistic distances between the taxa present in a cladistic matrix (see details below). Although empirically inter-taxon distances are the same in a morphospace or a cladisto-space (Foth et al., 2012; Hetherington et al., 2015), we prefer the term cladisto-space to make it clear that this space is estimated using cladistic data and not morphometric data and because both objects have slightly different properties. For example, because of its inherent combinatory properties, a cladisto-space is a finite theoretical object limited by the product of the number of character states, whereas a morphospace is an infinite theoretical object. Thus a cladisto-space will be overloaded if the number of taxa is higher than the product of the number of character states, although this is rarely an issue with empirical data (our cladisto-spaces have maximal capacities of $1.9\times10^{181}$ taxa for the Mammaliaformes dataset, i.e. 101 orders of magnitude more taxa than the number of particles in the universe; and $4.5\times10^{159}$ taxa for the Eutheria dataset).

To estimate the cladisto-spaces for each of our datasets we first constructed pairwise distance matrices of length $k$, where $k$ is the total number of tips and nodes in the datasets. For each dataset separately, we calculated the $k\times k$ distances using the Gower distance (Gower, 1971), i.e. the Euclidean distance between two taxa divided by the number of shared characters. This allows us to correct for distances between two taxa that share many characters and could be closer to each other than to taxa with fewer characters in common (i.e. because some pairs of taxa share more characters in common than others, they are more likely to be similar). For cladistic matrices, using this corrected distance is preferable to the raw Euclidean distance because of its ability to deal with discrete or/and ordinated characters as well as with missing data (Anderson and Friedman, 2012). However, the Gower distance cannot calculate distances when taxa have no overlapping data. Therefore, we used the `TrimMorphDistMatrix` function from the `Claddis R` package (Lloyd, 2015) to remove pairs of taxa with no cladistic characters in common. This led to us removing 11 taxa from the Mammaliaformes dataset but none from the Eutheria dataset.

After calculating our distance matrices we transformed them using classical multidimensional scaling (MDS; Torgerson, 1965; Gower, 1966; Cailliez, 1983). This method (also referred to as PCO; e.g. Brusatte et al. 2015; or PCoA; e.g. Paradis et al. 2004) is an eigen

decomposition of the distance matrix. Because we used Gower distances instead of raw Euclidean distances, negative eigenvalues can be calculated. To avoid this problem, we first transformed the distance matrices by applying the Cailliez correction (Cailliez, 1983) which adds a constant $c^*$ to the values in a distance matrix (apart from the diagonal) so that all the Gower distances become Euclidean ($d_{Gower} + c^* = d_{Euclidean}$; Cailliez 1983). We were then able to extract $n$ eigenvectors for each matrix (representing the $n$ dimensions of the cladisto-space) where $n$ is equal to $k - 2$, i.e. the number of taxa in the matrix ($k$) minus the last two eigenvectors that are always null after applying the Cailliez correction. Contrary to previous studies (e.g Brusatte et al., 2008a; Cisneros and Ruta, 2010; Prentice et al., 2011; Anderson and Friedman, 2012; Hughes et al., 2013; Benton et al., 2014), we use all $n$ dimensions of our cladisto-spaces and not a subsample representing the majority of the variance in the distance matrix (e.g. selecting only $m$ dimensions that represent up to 90% of the variance in the distance matrix; Brusatte et al. 2008b; Toljagic and Butler 2013).

Note that our cladisto-spaces represent an ordination of all possible mammalian morphologies coded in each study through time. It is unlikely that all morphologies will co-occur at each time point, therefore, the disparity of the whole cladisto-space is expected to be greater than the disparity at any specific point in time.

### 4.2.4 *Calculating disparity*

Disparity can be estimated in many different ways (e.g. Wills et al., 1994; Ciampaglio, 2004; Thorne et al., 2011; Hopkins, 2013; Huang et al., 2015), however most studies estimate disparity using four metrics: the sum and products of ranges and variances, each of which gives a slightly different estimate of how the data fits within the cladisto-space (Foote, 1994; Wills et al., 1994; Brusatte et al., 2008a,b; Cisneros and Ruta, 2010; Thorne et al., 2011; Prentice et al., 2011; Brusatte et al., 2012; Toljagic and Butler, 2013; Ruta et al., 2013; Benton et al., 2014; Benson and Druckenmiller, 2014). Nonetheless, these methods suffer several methodological caveats. First, the range metrics are affected by the uneven sampling of the fossil record (Butler et al., 2012). Second, because we include all $n$ dimensions in the analysis (see above), the products of ranges and variances will tend towards zero since the scores of the last dimension are usually really close to zero themselves. These features make using the sum and products of ranges and variances problematic in our study. Instead, we use a different metric that comes with no statistical assumptions for measuring the dispersion of the data in the cladisto-space: the median distance between tips and nodes and the centroid (similar but not equivalent to Wills et al. 1994; Korn et al. 2013; Huang et al. 2015) calculated as:

$$\text{Disparity} = median\sqrt{\sum(\mathbf{v}_n - Centroid_n)^2} \tag{4.1}$$

where:

$$Centroid_n = \frac{\sum(\mathbf{v}_n)}{k} \tag{4.2}$$

and $\mathbf{v}_n$ is any of the $n$ eigenvectors (i.e. any of the $n$ dimensions of the cladisto-space), $Centroid_n$ is the mean value of the $n^{th}$ eigenvector (equation 4.2) and $k$ is the total number of tips and nodes. Note that we also calculated the sum and products of ranges and variances and refer to these results in the appendix C (Fig. C.3, C.4, C.5, C.6).

### 4.2.5 *Estimating disparity through time*

Changes in disparity through time are generally investigated by calculating the disparity of taxa that occupy the cladisto-space during specific time intervals (e.g Cisneros and Ruta, 2010; Prentice et al., 2011; Hughes et al., 2013; Hopkins, 2013; Benton et al., 2014; Benson and Druckenmiller, 2014). These time intervals are usually defined based on biostratigraphy (e.g. Cisneros and Ruta, 2010; Prentice et al., 2011; Hughes et al., 2013; Benton et al., 2014) but can also be arbitrarily chosen time periods of equal duration (Butler et al., 2012; Hopkins, 2013; Benson and Druckenmiller, 2014). However, this approach suffers from two main biases. First, if biostratigraphy is used to determine the time intervals, disparity may be distorted towards higher differences between time intervals because biostratigraphical periods are geologically defined based on differences in the morphology of fossils found in the different strata. Second, this approach assumes that all characters evolve following a punctuated equilibrium model, because disparity is only estimated once for each interval resulting in all changes in disparity occurring between intervals, rather than also allowing for gradual changes within intervals (Hunt et al., 2015).

To address these issues, we used a "time-slicing" approach that considers subsets of taxa in the cladisto-space at specific equidistant points in time, as opposed to considering subsets of taxa between two points in time. This results in even-sampling of the cladisto-space across time and permits us to define the underlying model of character evolution (punctuated or gradual). In practice, time-slicing considers the disparity of any element present in the phylogeny (branches, nodes and tips) at any point in time. When the phylogenetic elements are nodes or tips, the eigenvector scores for the nodes (estimated using ancestral state reconstruction as described in section 4.2.2) or tips are directly used for estimating disparity. When the phylogenetic elements are branches we chose the eigenvector score for the branch using one of two evolutionary models:

1. **Punctuated evolution.** This model selects the eigenvector score from either the ancestral node or the descendant node/tip of the branch regardless of the position of the slice along the branch. Similarly to the time interval approach, this reflects a model of punctuated evolution where changes in disparity occur either at the start or

at the end of a branch over a relatively short time period and clades undergo long periods of stasis during their evolution (Gould and Eldredge, 1977; Hunt, 2007). We applied this model in three ways:

(i) selecting the eigenvector score of the ancestral node of the branch (ACCTRAN).

(ii) selecting the eigenvector score of the descendant node/tip of the branch (DELTRAN).

(iii) randomly selecting either the eigenvector score of the ancestral node or the descendant node/tip of the branch (random).

Method (i) assumes that changes always occur early on the branch (accelerated transition, ACCTRAN) and (ii) assumes that changes always occur later (delayed transition, DELTRAN). We prefer not to make either assumption so we report the results from (iii), although the ACCTRAN and DELTRAN results are available in the appendix C (Fig. C.3, C.4, C.5, C.6).

2. **Gradual evolution.** This model also selects the eigenvector score from either the ancestral node or the descendant node/tip of the branch, but the choice depends on the distance between the sampling time point and the end of the branch. If the sampling time point falls in the first half of the branch length the eigenvector score is taken from the ancestral node, conversely, if the sampling time point falls in the second half of the branch length the eigenvector score is taken from the descendant node/tip. This reflects a model of gradual evolution where changes in disparity are gradual and cumulative along the branch. Under this model, the gradual changes could be either directional or random, however, directional evolution has been empirically shown to be rare (only 5% of the time Hunt, 2007). We therefore considered that changes from a character state A to B were only dependent on the branch length.

We applied our time-slicing approach separately to the two cladisto-spaces calculated for the Mammaliaformes and Eutheria datasets, time-slicing the phylogeny every five million years from 170 Ma to the present resulting in 35 subsamples of the cladisto-space. For each subsample, we estimated its disparity assuming punctuated (ACCTRAN, DELTRAN and random) and gradual evolution as described above. To reduce the influence of outliers on our disparity estimates, we bootstrapped each disparity measurement by randomly resampling with replacement a new subsample of taxa from the observed taxa in the subsample 1000 times. We then calculated the median disparity value for each subsample along with the 50% and 95% confidence intervals. We also recorded the number of phylogenetic elements (nodes and tips) in each subsample as a proxy for taxonomic diversity. To compare our results to previous studies we also repeated our analyses using the time interval approach

based on biostratigraphy (e.g. Cisneros and Ruta, 2010; Prentice et al., 2011; Hughes et al., 2013; Benton et al., 2014) using each geological stage from the Middle Jurassic to the present. We report the results of these analyses in the appendix C (Fig. C.3, C.4, C.5, C.6).

### 4.2.6 *Testing the effects of the K-Pg extinction event on mammalian disparity*

If the K-Pg extinction event had a significant effect on mammalian disparity, we should see a significant difference between disparity at the end of the Cretaceous and disparity at the start of the Paleogene. To test this, we performed *t*-tests to look for differences in disparity between the time subsamples of interest (e.g. as used in Anderson and Friedman, 2012; Zelditch et al., 2012; Smith et al., 2014). We compared the last time subsample before the K-Pg boundary (70 Ma) to the first subsample of the Paleocene (65 Ma) for both the Mammaliaformes and Eutheria datasets and using both the gradual and punctuated evolutionary models. Even though one million years after the K-Pg event (66 to 65 Ma) seems to be a rather short geological time frame, effects on mammalian evolution have been detected as early as half a million year after K-Pg (Wilson, 2013). However, the effect of extinction on a group's evolution might not be detectable directly after the event due to delays in recovery (e.g. Chen and Benton, 2012, estimated that ecosystems only fully recovered 8-9 Ma after the Permo-Triassic mass extinction). Therefore, we also tested whether there was a significant difference in disparity between the end of the Cretaceous (70 Ma) and all subsamples from the Paleocene (65, 60 and 55 Ma). Additionally, some authors argue that the major diversification event in mammals took place during the Paleocene-Eocene Thermal Maximum (PETM; $\sim$ 56 Ma; Bininda-Emonds et al. 2007 but see Meredith et al. 2011 and Stadler 2011) with the extinctions at K-Pg providing the "empty" ecological space required for this diversification to occur. We therefore extended our comparisons between the last subsample of the Cretaceous (70 Ma) up to the late Eocene (35 Ma) to check for a delayed effect of the K-Pg extinction potentially allowing morphological diversification after the PETM. Because these analyses involved multiple comparisons, we used Bonferonni corrections (Holm, 1979) to ensure our significant results were robust to Type I error rate inflation.

Finally, disparity may be higher in subsamples with more phylogenetic elements simply because there are more taxa represented. To test whether this influenced our results, we repeated the *t*-tests using the rarefied Mammaliaformes and Eutheria disparities. In the Mammaliaformes, the minimum number of taxa in each subsample from 170 Ma to present was eight. In the Eutheria, the minimum number of taxa in each subsample was three, however, from 150 Ma until the present, the minimum number of taxa is eight. To make both datasets comparable, we used eight as a minimum number of taxa for the rarefied bootstrap

measurements, therefore in the Eutheria we ignored the subsample between 170 and 150 Ma that contained less than eight taxa.

## 4.3 RESULTS

Disparity in the Mammaliaformes reaches a plateau during the Middle Triassic around 240 Ma, and fluctuates slightly around this during the rest of the Mesozoic and the Cenozoic (Fig. 4.3 and Fig. C.1). The number of tips and nodes in each time subsample (a proxy for taxonomic richness), however, show a more idiosyncratic pattern with a steady increase until the Middle Jurassic around 170 Ma (Fig. S5) followed by random fluctuations during the rest of the Mesozoic and the Cenozoic (Fig. 4.3). Disparity in the Eutheria reaches a plateau at the end of the Jurassic around 150 Ma, whereas the number of tips and nodes increases up to the K-Pg boundary and then decreases throughout the Cenozoic (Fig. 4.3). For both Mammaliaformes and Eutheria the same patterns in changes in disparity appear in the rarefied analyses (Fig. 4.3 and 4.4). In both datasets the two evolutionary models (gradual or punctuated) also yield similar results (Fig. 4.3).

We found no significant differences in disparity between the last subsample of the Cretaceous (70 Ma) and the first subsample of the Paleogene (65 Ma; Table 4.1), using both datasets and under both evolutionary models. We also found no significant differences in disparity between the last subsample of the Cretaceous (70 Ma) and any subsamples of the Paleocene and Eocene in Mammaliaformes under both evolutionary models and in Eutheria under a gradual evolutionary model (Table 4.1). However, in Eutheria under the punctuated evolutionary model, we found a small significant difference (after applying Bonferonni corrections) in disparity between the last subsample of the Cretaceous (70 Ma) and the subsamples at 45 Ma (an increase in disparity of 0.17; Table 4.1). However, this result is not significant in the rarefied analyses (Table 4.2). Otherwise the results of the rarefied analyses are the same as when using the complete datasets.

## 4.4 DISCUSSION

Previous authors have suggested that the K-Pg extinction event released mammals from ecological pressures such as competition and predation, allowing them to radiate into newly available ecological niches (Archibald, 2011; O'Leary et al., 2013; Lovegrove et al., 2014; Slater, 2013). However, we did not detect any significant changes in mammalian disparity before and after K-Pg in either Mammaliaformes or Eutheria, under a model of punctuated or gradual evolution. Additionally, we tested whether the absence of a detectable effect might be due to a lag effect, with the effect only becoming obvious later in the Paleocene. Even when accounting for such a lag effect, we did not detect any significant effect of the K-Pg

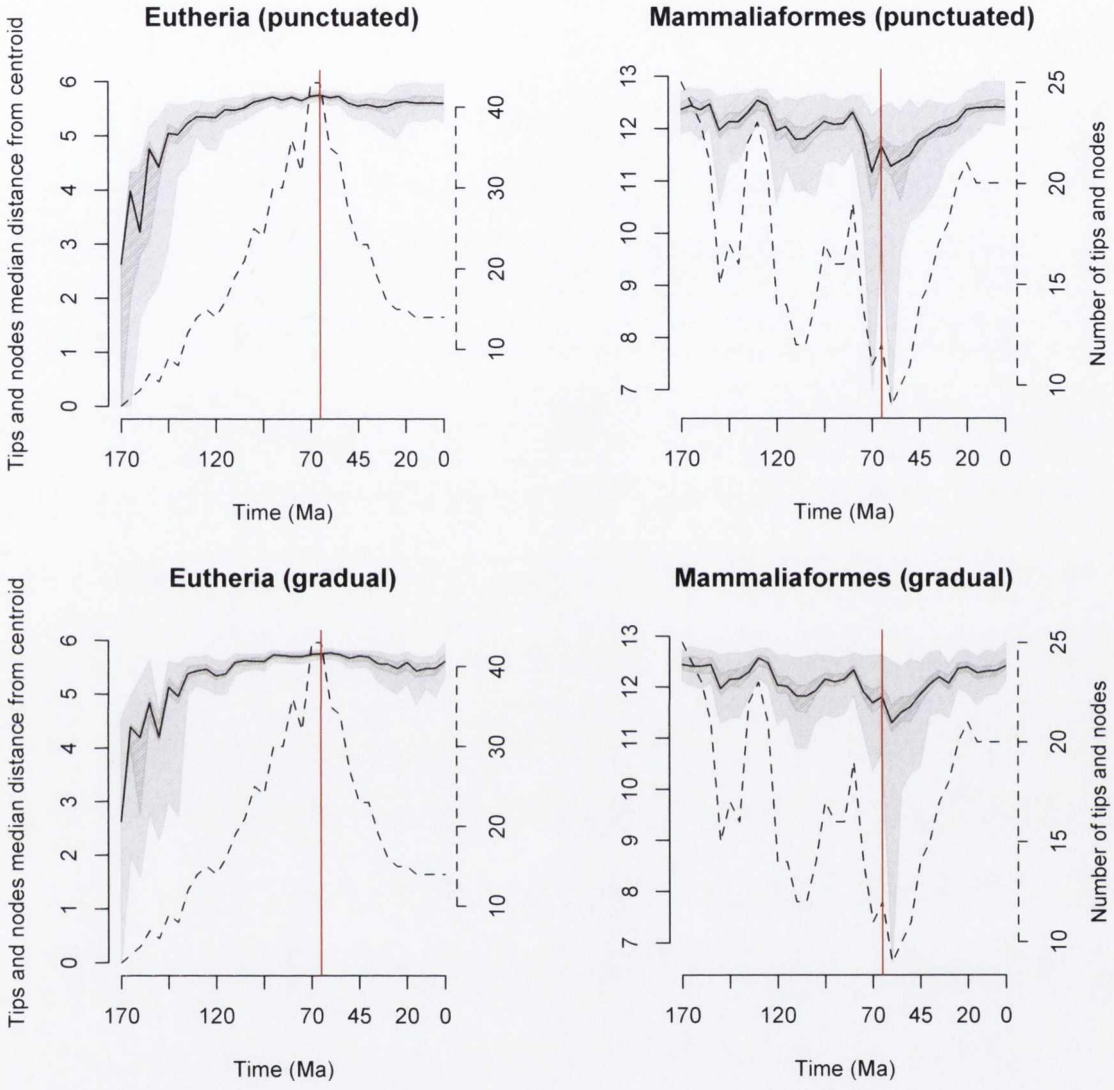



FIGURE 4.3: Disparity through time in Eutheria and Mammaliaformes calculated using a model of punctuated (upper panels) or gradual (lower panels) evolution. The x axis represents time in millions of years before the present (Ma). The y axis represents disparity, measured as the median distance between the centroid of the ordinated space and the tips/nodes in each time subsample. The solid black lines show the mean disparity estimated from 1000 bootstrapped pseudoreplicates and confidence intervals (CI) are represented by the grey polygons (50% CI in dark grey and 95% CI in light grey). The dashed line and the right hand axis represents the number of tips/nodes in each time slice. The red vertical line indicates the Cretaceous-Paleogene (K-Pg) boundary (66 Ma). Note that scale bars differ among panels.

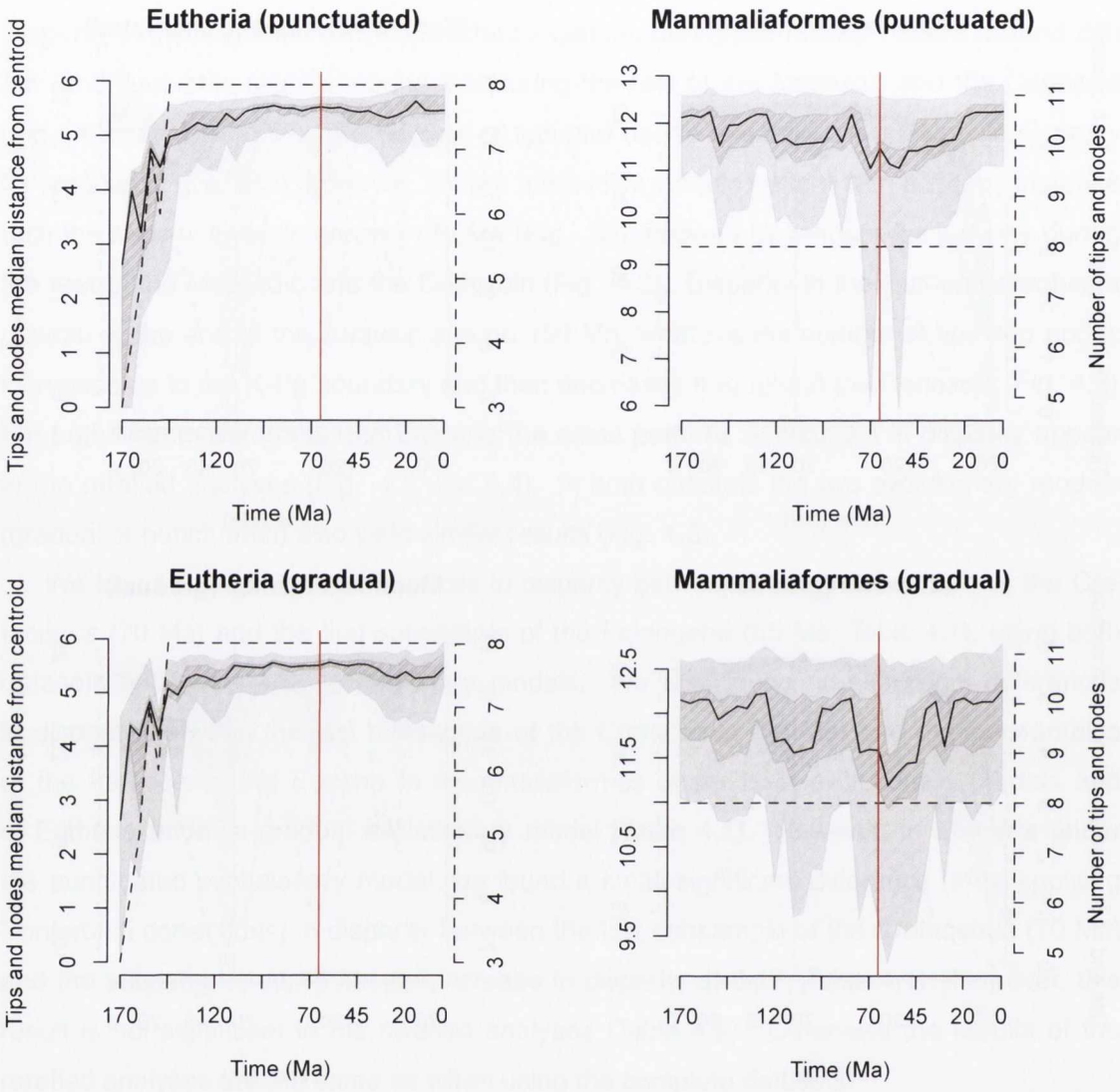



FIGURE 4.4: Rarefied disparity through time in Eutheria and Mammaliaformes calculated using a model of punctuated (upper panels) or gradual (lower panels) evolution. The x axis represents time in millions of years before the present (Ma). The y axis represents disparity, measured as the median distance between the centroid of the ordinated space and the tips/nodes in each time subsample. The solid black lines show the mean disparity estimated from 1000 bootstrapped pseudoreplicates and confidence intervals (CI) are represented by the grey polygons (50% CI in dark grey and 95% CI in light grey). The dashed line and the right hand axis represents the number of tips/nodes in each time slice. The red vertical line indicates the Cretaceous-Paleogene (K-Pg) boundary (66 Ma). Note that scale bars differ among panels.

extinction event on mammalian disparity. Our results imply that mammals did not diversify morphologically in response to the K-Pg extinction event. Instead, their diversification appears to have begun before the end of the Cretaceous (Fig. 4.3, Table 4.1 and see Meredith et al., 2011; dos Reis et al., 2014; Close et al., 2015; Lee and Beck, 2015).

TABLE 4.1: Results of *t*-tests comparing disparity at the last subsample of the Cretaceous (70 Ma) to subsamples of the Paleocene and Eocene, under both gradual and punctuated evolutionary models, in Mammaliaformes and Eutheria. Difference = mean difference in disparity between the two subsamples being compared; df = degrees of freedom; p value = original p value prior to Bonferonni correction. Significant differences (after applying Bonferonni corrections for multiple comparisons) are highlighted in bold.

| Subsamples compared | Gradual evolution model | | | | Punctuated evolution model | | | |
|---|---|---|---|---|---|---|---|---|
| | difference | df | t | p value | difference | df | t | p value |
| **Mammaliaformes** | | | | | | | | |
| 70 *vs.* 65 | -0.420 | 21 | -0.808 | 0.428 | -0.030 | 21 | -0.058 | 0.954 |
| 70 *vs.* 60 | 0.030 | 18 | 0.046 | 0.964 | 0.210 | 18 | 0.379 | 0.709 |
| 70 *vs.* 55 | 0.010 | 19 | 0.021 | 0.983 | 0.110 | 19 | 0.225 | 0.824 |
| 70 *vs.* 50 | -0.260 | 20 | -0.456 | 0.653 | 0.030 | 20 | 0.060 | 0.953 |
| 70 *vs.* 45 | -0.430 | 23 | -0.869 | 0.394 | 0.060 | 23 | 0.132 | 0.896 |
| 70 *vs.* 40 | -0.620 | 24 | -1.388 | 0.178 | -0.410 | 24 | -1.031 | 0.313 |
| 70 *vs.* 35 | -0.730 | 26 | -1.742 | 0.093 | -0.340 | 26 | -0.861 | 0.397 |
| **Eutheria** | | | | | | | | |
| 70 *vs.* 65 | -0.020 | 84 | -0.503 | 0.616 | 0.010 | 84 | 0.288 | 0.774 |
| 70 *vs.* 60 | 0.030 | 76 | 0.617 | 0.539 | 0.080 | 76 | 1.693 | 0.095 |
| 70 *vs.* 55 | 0.030 | 75 | 0.519 | 0.605 | 0.030 | 75 | 0.699 | 0.486 |
| 70 *vs.* 50 | 0.130 | 68 | 2.101 | 0.039[1] | 0.080 | 68 | 1.458 | 0.149 |
| 70 *vs.* 45 | 0.190 | 64 | 2.679 | 0.009[1] | 0.170 | 64 | 2.730 | **0.006**[2] |
| 70 *vs.* 40 | 0.160 | 64 | 2.249 | 0.028[1] | 0.130 | 64 | 2.084 | 0.041[1] |
| 70 *vs.* 35 | 0.190 | 60 | 2.358 | 0.022[1] | 0.120 | 60 | 1.893 | 0.063 |

[1]p value is non-significant after applying Bonferonni correction; [2]p value is **0.048** after applying Bonferonni correction.

We did, however, detect a small, yet significant, increase in disparity during the Eocene (45 Ma) under a punctuated evolutionary model using the Eutheria dataset. This might be due to a long lag effect of ~21 Ma after K-Pg. Note however, that this is double the lag time observed in other mass extinctions (Chen and Benton, 2012). Therefore, it may be more likely to be attributed to a lag effect of the Palaeocene-Eocene Thermal Maximum (PETM; ~11 Ma afterwards; Bininda-Emonds et al., 2007). However, this significant increase in disparity is only detected at 45 Ma but not afterwards (which would be expected if the increase was due to an evolutionary radiation) and is not seen under the gradual evolution model. This indicates that it is more likely due to differences in the evolutionary models rather than an actual increase in disparity. The 45 Ma subsample samples the long branch (~50 million years) leading to *Lepidictis* (33.9 to 33.3 Ma) that branches with its closest relative *Gypsonictops* (66.8 to 66 Ma) in the early Upper Cretaceous ~90 Ma (see Fig. 4.2). Therefore, in this time-slice under the gradual evolution model, the data for *Lepidictis* is always sampled, but under the punctuated evolution model the algorithm can also randomly

sample the data from its ancestor in the early Upper Cretaceous (see methods). This may inflate differences compared to other slices. Incidentally, this increase can also be linked to the number of tips and nodes used in the comparison (43 versus 23 tips and nodes at respectively 70 and 45 Ma), because the increase is not significant in the rarefied analysis (see Fig. 4.4 with only eight tips and nodes). Given these caveats we believe that no strong conclusions can be drawn from the increase in disparity during the Eocene.

TABLE 4.2: Results of rarefied *t*-tests comparing disparity at the last subsample of the Cretaceous (70 Ma) to subsamples of the Paleocene and Eocene, under both gradual and punctuated evolutionary models, in Mammaliaformes and Eutheria. Difference = mean difference in disparity between the two subsamples being compared; df = degrees of freedom; p value = original p value prior to Bonferonni correction.

| **Subsamples compared** | **Gradual evolution model** | | | | **Punctuated evolution model** | | | |
|---|---|---|---|---|---|---|---|---|
| | **difference** | **df** | **t** | **p value** | **difference** | **df** | **t** | **p value** |
| **Mammaliaformes** | | | | | | | | |
| 70 *vs.* 65 | -0.360 | 21 | -0.486 | 0.632 | 0.040 | 21 | 0.054 | 0.957 |
| 70 *vs.* 60 | -0.080 | 18 | -0.099 | 0.922 | 0.060 | 18 | 0.094 | 0.927 |
| 70 *vs.* 55 | -0.090 | 19 | -0.110 | 0.914 | 0.050 | 19 | 0.082 | 0.936 |
| 70 *vs.* 50 | -0.270 | 20 | -0.368 | 0.717 | 0.030 | 20 | 0.041 | 0.968 |
| 70 *vs.* 45 | -0.310 | 23 | -0.419 | 0.679 | 0.200 | 23 | 0.285 | 0.778 |
| 70 *vs.* 40 | -0.460 | 24 | -0.680 | 0.503 | -0.240 | 24 | -0.422 | 0.677 |
| 70 *vs.* 35 | -0.510 | 26 | -0.742 | 0.465 | -0.100 | 26 | -0.159 | 0.875 |
| **Eutheria** | | | | | | | | |
| 70 *vs.* 65 | -0.020 | 84 | -0.139 | 0.890 | 0.020 | 84 | 0.101 | 0.920 |
| 70 *vs.* 60 | 0.020 | 76 | 0.095 | 0.925 | 0.070 | 76 | 0.386 | 0.701 |
| 70 *vs.* 55 | 0.010 | 75 | 0.076 | 0.940 | 0.020 | 75 | 0.111 | 0.912 |
| 70 *vs.* 50 | 0.090 | 68 | 0.453 | 0.652 | 0.040 | 68 | 0.232 | 0.817 |
| 70 *vs.* 45 | 0.120 | 64 | 0.563 | 0.575 | 0.110 | 64 | 0.562 | 0.576 |
| 70 *vs.* 40 | 0.100 | 64 | 0.473 | 0.638 | 0.070 | 64 | 0.383 | 0.703 |
| 70 *vs.* 35 | 0.120 | 60 | 0.515 | 0.608 | 0.040 | 60 | 0.195 | 0.846 |

Our results differ from a previous study that found an increase in disparity in North American Theria as soon as ~0.5 Ma after K-Pg (Wilson, 2013). These differences are likely to be related to several methodological differences between the present study and the previous one (Wilson, 2013). Firstly, Wilson (2013) only measures disparity at a regional scale (North America) and proposes that the observed increases in disparity are linked to the immigration of new species into the study localities. This strongly implies that disparity was higher on a global scale. Secondly, most of the debate on mammalian diversification around the K-Pg boundary seems to be linked to the conflicting signal between palaeontological and neontological data (Meredith et al. 2011 *vs.* O'Leary et al. 2013 but see dos Reis et al. 2014). Therefore, an effect of the K-Pg extinction event might be detectable only when using just palaeontological data. In this study, however, we use Total Evidence tip-dated trees based on both palaeontological and neontological data (Slater, 2013; Beck and Lee, 2014), which may account for the differences between our study and that of Wilson (2013) who used only fossil data.

Interestingly, however, our results also differ from Slater (2013), the source of data for the Mammaliaformes dataset. Slater (2013) found support for a shift in the mode of body mass evolution (from an Ornstein–Uhlenbeck to a Brownian Motion model) directly after K-Pg suggesting a release in competition pressure or new ecological niches becoming availabile for mammals in the early Paleocene. Our studies may show different results due to the difference between changes observed in one continuous life-history trait (body mass; Slater, 2013) versus changes in an aggregate of 446 discrete morphological traits (the cladistic characters) in the present analysis. Body mass and disparity might be decoupled in a similar way to taxonomic diversity and disparity (e.g. Slater et al., 2010; Ruta et al., 2013; Hopkins, 2013) because the latter does not rely on size but rather on discrete morphological features. It is not unlikely that mammalian disparity increased rapidly early in their evolutionary history and then remained constant (Fig. 4.3; Close et al., 2015; Lee and Beck, 2015) while body mass variation continued to increase, especially after K-Pg (Slater, 2013). Note, however, that our methods did not investigate changes in body mass across the K-Pg boundary so they do not allow us to test this hypothesis. We remain confident in our results because we recovered the same pattern from two independent datasets (Slater, 2013; Beck and Lee, 2014).

There are several caveats to consider when interpreting our results. Firstly, both our datasets are limited taxonomically. They do not represent all known mammalian taxa, especially during the Neogene (23–2.58 Ma) where there are no fossil taxa in either dataset. Our study, however, focuses on changes in disparity around the K-Pg boundary and not during the whole Cenozoic. Besides, this might not cause a serious underestimation of disparity, at least for the Mammaliaformes, because their diversity peaked during the late Cretaceous (Campanian; 72.1–83.6 Ma; Newham et al., 2014) and mammalian diversification rates declined throughout the Cenozoic (Raia et al., 2012). Therefore, an effect of the K-Pg boundary would be more likely to be detected during the Paleogene when mammalian diversity was highest, so we do not believe that increasing taxon sampling would greatly alter our conclusions.

Secondly, testing for significant changes in disparity through time is problematic. The disparity of each subsample can be dependent on disparity in the previous subsamples. For example, the tips and nodes used to estimate disparity are linked by common evolutionary history, therefore two tips or nodes sharing a close ancestor are more likely to have similar morphological features than more distantly related tips and nodes. Similarly, when looking at disparity through time, different subsamples are related by time, therefore, two subsamples closely together in time are more likely to have the same disparity value than more distant subsamples. Additionally, because disparity is a single value summarizing morphological disparity, its variance and mean were calculated by bootstrapping, thus the variances and

means used in our *t*-tests are calculated from non-independent pseudoreplicates rather than true replicates. A second caveat arising from using bootstraps is that using a large number of pseudoreplicates is likely to inflate Type I error rates. Currently, however, this method is still widely used in disparity analyses for lack of a better alternative (e.g. Anderson and Friedman, 2012; Zelditch et al., 2012; Smith et al., 2014).

### 4.4.1 *Methodological improvements for measuring disparity through time*

Our results may differ from previous studies because of our specific methodological choices. Throughout this paper, we propose several incremental changes to the classical ways of measuring disparity. Firstly we used all the axes of the cladisto-space, as opposed to previous studies that selected a subsample of the cladisto-space arguing that the $m$ first axes usually contain most of the dataset's variance (e.g Brusatte et al., 2008a; Cisneros and Ruta, 2010; Prentice et al., 2011; Anderson and Friedman, 2012; Hughes et al., 2013; Benton et al., 2014). We argue that even if the last dimensions of the cladisto-space contain a trivial amount of variance, there is no statistical justification for excluding them. However, by doing so, we included dimensions of the cladisto-space with near zero variance and range (the last dimension's variance was $2 \times 10^{-14}$ and $1.15 \times 10^{-15}$ and the range was $7.31 \times 10^{-7}$ and $3.33 \times 10^{-7}$ for respectively the Mammaliaformes and Eutheria datasets). An alternative method avoids this problem by simply not ordinating the data and using the raw inter-taxon distance matrix (e.g. Benson and Druckenmiller, 2014; Close et al., 2015). However, in both this method and our method, the calculation of the products of ranges and variances is impossible. Secondly, we used median distance between tips and nodes to centroid as a disparity metric, rather than the classical sum and products of ranges and variances (Wills et al., 1994). This metric is not affected by problems with using the last dimensions of the cladisto-space (see above). Also, it has several other advantages over other metrics. For example, it measures directly the median spread of the taxa in the cladisto-space unlike the sum and products of ranges and variances that measure the size of the cladisto-space dimensions (Wills et al., 1994). Additionally, it comes with no statistical caveats unlike the sums or products of variances that should also include covariances between axes to correctly assess the exact size of the cladisto-space (even though the covariance term is usually close to 0 because of the eigen decomposition; Gower, 1966).

Finally, we used a time-slicing method instead of binning the data into time intervals (e.g in: Cisneros and Ruta, 2010; Prentice et al., 2011; Hughes et al., 2013; Hopkins, 2013; Benton et al., 2014; Benson and Druckenmiller, 2014) thus allowing us to avoid two caveats of using the time intervals approach. Because time intervals are often based on biostratigraphy, which is in turn based on notable differences in fossil fauna and flora, this method is likely to artificially emphasise disparity differences among time intervals. It is also possible

to use arbitrary time bins of equal duration rather than biostratigraphy (Butler et al., 2012; Hopkins, 2013; Benson and Druckenmiller, 2014), but both approaches make the underlying assumption that disparity changes in a punctuated manner, i.e. changes occur only between time intervals. However, gradual evolution has been shown to be relatively common in the fossil record (Hunt, 2007; Hunt et al., 2015), so this assumption is unfounded. Our approach allowed us to fit different evolutionary models to our data - either assuming punctuated or gradual evolution. This is an improvement on previous approaches but could be improved further by implementing other common but more complex models for example, a combined stasis and random walk (Hunt et al., 2015) or models based on morphological rates rather than just branch lengths.

### 4.4.2 *Conclusion*

Evidence for whether mammals diversified before or after the K-Pg boundary is mixed (Meredith et al., 2011; O'Leary et al., 2013; dos Reis et al., 2014; Beck and Lee, 2014), and appears to be related to the kind of data used (paleontological or neontological) and how the analyses were conducted. Using both living and fossil taxa, and investigating morphological disparity through time rather than taxonomic diversity, we find no direct effect of the K-Pg extinction event on the diversity of mammals. We therefore suggest that, contrary to popular belief, the extinction of many terrestrial vertebrates including the non-avian dinosaurs 66 million years ago, did not significantly affect the evolution of mammals throughout the Cenozoic.

CHAPTER 5

# DISCUSSION

In this final chapter of my thesis, I first discuss the implications of the results of Chapters 2 and 3 and then of Chapter 4. In both sections, I discuss the methodological caveats of my approaches and propose future research avenues to solve these problems and expand into new areas. Finally, I discuss the overall importance of combining both living and fossil species into macroevolutionary studies.

## 5.1 THE FUTURE OF THE TOTAL EVIDENCE METHOD

The Total Evidence method is a promising way of testing macroevolutionary hypotheses using extinct and extant taxa (e.g. Ronquist et al., 2012a; Slater, 2013; Wood et al., 2013; Beck and Lee, 2014; Dembo et al., 2015). However, as shown in the Chapters 2 and 3, this method is sensitive to missing data. As discussed in both chapters, increasing the number of morphological characters and the number of living taxa with coded morphological characters is the most effective way to recover the correct topology. This data is available and reasonably easy to access in natural history museum collections worldwide. In addition, software infrastructure has been developed to facilitate collaborative data collection of this kind (*MorphoBank*; O'Leary and Kaufman, 2011). Therefore, hopefully the scarcity of morphological data for living species will be gradually solved with time, provided the funding exists to make such efforts possible. However, there are other general problems with Total Evidence phylogenies that were not developed in the discussions of Chapters 2 and 3.

Firstly, dating Total Evidence phylogenies is difficult. Because these phylogenies contain both living and fossil taxa, the preferred way to date them is to use the tip-dating method (e.g. Ronquist et al., 2012a; Wood et al., 2013; Dembo et al., 2015). This method relies on the age of the fossil taxa (treated as tips) to date the divergence times of the nodes, rather than defining node calibrations *a priori* (cf. node-dating; Ronquist et al., 2012a). The tip-dating method has two main advantages: (1) it has been observed using empirical data that it improves the ability to recover the correct topology because it can use the stratigraphic age of the fossil taxa to favour some topological solutions more than others, typically by minimising implied ghost lineages (Matzke, 2014, and personal communications); and (2) it reduces the confidence intervals of the node ages compared to a classic node-dating approach (Ronquist et al., 2012a). This second point, however, has been questioned by

Arcila et al. (2015) who compared the tip-dating method and the latest models for the node-dating method, for example, the fossilised birth-death model (Heath et al., 2014), and found the opposite effect, i.e. an increase in node age confidence interval with the tip-dating method. It would therefore be interesting to run a similar analysis to the simulations in Chapter 2 but adding a dating aspect to it. By comparing dated Total-Evidence matrices using both node-dating and tip-dating, one could formally test the two advantages outlined above as well as their resilience to missing data.

Secondly, a more general problem is that the Total Evidence method relies on the M*k* model (Lewis, 2001) to measure the morphological distances among taxa. This model is a generalisation of the Jukes-Cantor model (JC69; Jukes and Cantor, 1969) that allows a single mutation rate $\mu$ among all character states. The JC69 model is a great simplification of reality and has been replaced in molecular phylogenetics with more complex models that are closer to biological reality (e.g. the GTR model that allows a different rate for each different type of nucleotide change; Tavaré, 1986). It is therefore likely that the M*k* model is also a crude underestimation of the complex reality of morphological evolution, especially as the assumption that there is a unique transition rate between character states has been shown to be wrong in at least some cases (e.g. for Dollo traits that are irreversible; Wright et al., 2015). Spencer and Wilberg (2013) even demonstrated that non-probabilistic methods such as maximum parsimony outperform the M*k* model for the placement of contentious fossils such as *Archaeopteryx*. However, more recent and thorough simulations have demonstrated the opposite and have shown that using the M*k* model consistently outperforms maximum parsimony, especially in the presence of missing data (Wright and Hillis, 2014). Additionally, incremental improvements on the M*k* model are currently being made to solve at least the fact that the some character states are irreversible once evolved (Klopfstein et al., 2015, and personal communications).

Despite the three major caveats discussed above (missing data, dating methods, and models for morphological evolution), the Total Evidence method remains one of the best methods for including the diversity of life both past and present into phylogenies. Even though this method was first proposed decades ago (e.g. Eernisse and Kluge, 1993), recent successful software implementations (Ronquist et al., 2012b; Bouckaert et al., 2014) have allowed it to be more widely used in the last three years (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; Klopfstein et al., 2015; Carrizo and Catalano, 2015; Wittenberg et al., 2015; Gavryushkina et al., 2015). This increasing number of studies using the Total Evidence method will probably result in further improvements and popularity as time goes on. This is really encouraging because it will eventually result in more accurate phylogenies (based on

molecular and morphological data), including both living and fossil species, being available for macroevolutionary studies in the near future.

## 5.2 THE CLADISTO-SPACE AS A PROXY FOR DESCRIBING MACROEVOLUTIONARY CHANGES

One important point to keep in mind when building phylogenies, however, is that phylogenetic trees (whether they use all the available data or not) are tools for observing evolutionary patterns and testing evolutionary hypotheses, not the end point of the scientific endeavour. Until now, only few studies have used a Total Evidence phylogeny as a tool for specifically testing macroevolutionary hypotheses (e.g. Slater, 2013; Wood et al., 2013; Dembo et al., 2015). In Chapter 4, I used two tip-dated Total Evidence trees for testing whether the K-Pg mass extinction event influenced the diversity of mammals. Rather than using taxonomic richness as my proxy for diversity, I used disparity, i.e. morphological diversity (or in this case cladistic diversity), and tested whether mammalian disparity increased after the K-Pg extinction. Although it is becoming increasingly common to estimate the disparity of clades in palaeobiology (e.g. Butler et al., 2012; Brusatte et al., 2012; Toljagic and Butler, 2013; Brusatte et al., 2014; Benson and Druckenmiller, 2014; Lloyd, 2015; Close et al., 2015), methods for estimating disparity suffer from several biases that I did not fully discuss in Chapter 4.

Describing the shape or form of an organism is not straightforward and many mathematical methods exist for doing it (e.g. Elliptic Fourier; Kuhl and Giardina 1982; Procrustes; Rohlf and Marcus 1993; Convex Hull; Andrew 1979). By form here, I am referring to biological variations in the morphology of organisms (cf. the shape defined as the 2D outlines of an individual). In biology, one approach is to describe an organism's form by using geometric morphometrics (Zelditch et al., 2012). This involves collecting Cartesian coordinates of a series of discrete spatial points (landmarks), transforming them to remove the effect of size differences via Procrustes transformation, and then ordinating these transformed coordinates into a smaller number of variables, for example by using Principal Components Analysis (PCA). The organism's form can then be described as a single value that summarises the matrix (e.g. the sum of the ranges of each PCA axis; Zelditch et al., 2012). When using this approach, form is approximated by actual continuous measurements collected from the organisms (e.g. Friedman, 2010; Hopkins, 2013; Finlay and Cooper, 2015). In our case, however, we instead used inter-taxon distances based on discrete morphological features to describe form (i.e. the cladistic disparity method; e.g. Foote, 1997; Wills, 2001; Wesley-Hunt, 2005). This method is sometimes criticised because the morphological features are not randomly collected: cladistic characters are usually collected to resolve relationships among lineages (O'Leary et al., 2013). This might

distort reality because some authors will emphasise differences in the taxonomic groups of interest (Hopkins and Smith, 2015). Additionally, when including fossil data to such analyses, the available characters are highly dependent on the quality of the fossil record. This can be biased against certain type of characters (e.g. soft tissues; Sansom and Wills, 2013) or towards geological strata with more fossils (e.g. *Lagerstätten*; Butler et al., 2012). Nonetheless, these two biases are overshadowed by the advantages of having a great deal of easily available morphological data (some morphological matrices have more than 1000 characters; e.g. O'Leary et al., 2013; Ni et al., 2013) that can be compared among many taxa across many taxonomic levels (e.g. across all mammals; O'Leary et al., 2013; Slater, 2013; Beck and Lee, 2014). Furthermore, empirical studies have shown that the same signal seems to be captured when using either geometric morphometric or cladistic methods to describe disparity (Foth et al., 2012; Hetherington et al., 2015).

Secondly, disparity is an abstraction of morphological diversity; it is a unique value that describes the cladisto-space, that is in turn based on a multidimensional transformation of the discrete form differences among the taxa in the analysis (Wills et al., 1994; Foote, 1997). There are many different ways of doing this, though generally people have used the four metrics proposed by Wills et al. (1994) (sum and products of ranges and variances) to calculate disparity. I discussed the caveats of using such metrics in Chapter 4 but I did not explore several other, more general, problems. In macroevolutionary studies, one of the purposes of disparity metrics is to describe biological changes in the cladisto-space: for example, when species go extinct the occupancy of the cladisto-space can decrease, suggesting a loss of biological forms, or stay constant, suggesting random loss of species across cladisto-space. It is not clear what variations in these metrics really reflect biologically apart from variation in the range or variance of each dimension of the cladisto-space. These metrics are actually only describing the dimensions of the ordinated matrix, and are not directly describing the relative placement of the tips or nodes in that space, unlike measures such as the distance between taxa and the centroid. The latter may be of more interest when answering particular macroevolutionary questions such as the one tested in Chapter 4, thus disparity metrics based on the placement of the taxa in the cladisto-space should be developed further. Even though some attempts have been made for measuring the effectiveness of disparity metrics (Ciampaglio et al., 2001) there has been no systematic assessment of the statistical power of each metric for describing multidimensional space occupancy. Future developments of disparity-through-time studies require a better understanding of the statistical performance of disparity metrics and which metrics are most appropriate in specific empirical situations. In the future, I plan to assess the power of the different available disparity metrics (e.g. Wills et al., 1994; Ciampaglio, 2004; Hughes et al., 2013; Huang et al., 2015) through simulation studies by testing their performance at assessing various types of changes in ordinated

matrices such as the distribution of the taxa in the cladisto-space (being randomly- or evenly-distributed or clustered) or changes in cumulative (i.e. explanatory) variance among the dimensions of the cladisto-space.

One final crucial aspect of disparity-through-time studies regards the use of both living and fossil taxa in these studies. As discussed in Chapter 4, palaeontological or neontological data suggest different patterns of diversification in mammals, with diversification either occurring just after the K-Pg event when fossil species are used (suggesting an effect of K-Pg; O'Leary et al., 2013) or before when living species are used (rejecting an effect of K-Pg; Meredith et al., 2011; dos Reis et al., 2014). Because it is impossible to be certain which scenario is correct, I argue that using all the available data, i.e. both living and fossil taxa (Slater, 2013; Beck and Lee, 2014) is the best way to describe the observed patterns with more confidence. Future disparity studies should include both living and fossil taxa to get a true understanding of patterns in disparity-through-time.

## 5.3 FUTURE DIRECTIONS FOR COMBINING LIVING AND FOSSIL SPECIES IN MACROEVOLUTIONARY STUDIES

Using data from both living and fossil species has lead to substantial improvements in macroevolutionary studies (e.g. improving ancestral characters estimations; Finarelli and Flynn 2006; mode of evolution testing; Slater 2013; Pant et al. 2014; or disparity-through-time analyses; Mitchell 2015). However, combining living and fossil species in such studies often requires extra work and specific expertise (e.g. the study by Ronquist et al., 2012a, gathered experts in statistics, bioinformatics, phylogenetics and palaeontology) and can suffer from more problems than those arising from using living and fossil data separately (e.g. the missing data problem described in Chapters 2 and 3). Additionally, it could be counterproductive to add living species when studying fossil clades that have very few or no living relatives (e.g. Trilobita; Hopkins 2013; Pterosauria; Butler et al. 2012; etc.). Likewise, when studying clades that underwent a recent explosive radiation (e.g. Cichlidae; Genner et al., 2007), it may not be sensible to add fossil species. In both cases, adding fossil or living species would not significantly change the observed macroevolutionary patterns, but would take a lot of extra time, effort and expertise. Several studies, however, have suggested that it is still important to combine living and fossil species in analyses with predominantly living or fossil taxa. For example, when studying Ostracoderma (armoured jawless fishes that went extinct during the Devonian; 358 Ma), some morphological features can only be interpreted in the light of living species and it is therefore important to compare them to living vertebrates to understand vertebrate evolution (Janvier, 2015). On the other hand, even recent groups often have recently extinct members that may change the conclusions of our analyses if we choose to ignore them. For example, when studying Strepsirrhini (lemurs and lorises), it

may be important to include subfossil giant lemurs that went extinct only 600 years ago and were two orders of magnitude bigger than living lemurs (Goodman and Benstead, 2003), especially when studying body size evolution.

In general, discarding part of your data on an arbitrary basis is not good scientific practice. In addition, the effect of adding fossil species to analyses of living species (and vice versa) cannot be known *a priori*. Therefore I argue that we should always use both living and fossil species in our analyses wherever feasible. Our knowledge in biology has tremendously advanced since the last half century ranging from the amazing revelation of the few glimpses of the deep past provided by the fossil record to the understanding of the complexity and dynamics of modern ecosystems. Because the inherent characteristic of the deep past is to be unknown and mysterious, it is therefore crucial to incorporate all of this knowledge into macroevolutionary studies to continue revealing the complexities of biodiversity.

# Bibliography


Anderson, P. S. and M. Friedman. 2012. Using cladistic characters to predict functional variety: experiments using early Gnathostomes. Journal of Vertebrate Paleontology 32:1254–1270.

Andrew, A. 1979. Another efficient algorithm for convex hulls in two dimensions. Information Processing Letters 9:216–219.

Archibald, J. D. 2011. Extinction and radiation: how the fall of dinosaurs led to the rise of mammals. JHU Press: Baltimore.

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. 1985. Mass extinction among non-marine tetrapods. Nature 316:811–814.

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

Benton, M. J., J. Forth, and M. C. Langer. 2014. Models for the rise of the dinosaurs. Current biology 24:R87–R95.

Benton, M. J. and R. J. Twitchett. 2003. How to kill (almost) all life: the end-Permian extinction event. Trends in Ecology and Evolution 18:358–365.

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., M. Cardillo, K. E. Jones, R. D. MacPhee, R. M. 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.

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. Kühnert, 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.

Brazeau, M. D. 2011. Problematic character coding methods in morphology and their effects. Biological Journal of the Linnean Society 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., R. J. Butler, A. Prieto-Márquez, and M. A. Norell. 2012. Dinosaur morphological diversity and the end-Cretaceous extinction. Nature Communications 3:804–804.

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

Brusatte, S. L., M. J. Benton, M. Ruta, and G. T. Lloyd. 2008b. Superiority, competition, and opportunism in the evolutionary radiation of dinosaurs. Science 321:1485–1488.

Brusatte, S. L., R. J. Butler, P. M. Barrett, M. T. Carrano, D. C. Evans, G. T. Lloyd, P. D. Mannion, M. A. Norell, D. J. Peppe, P. Upchurch, and T. E. Williamson. 2015. The extinction of the dinosaurs. Biological Reviews 90:628–642.

Brusatte, S. L., G. T. Lloyd, S. C. Wang, and M. A. N. Norell. 2014. Gradual assembly of Avian body plan culminated in rapid rates of evolution across the Dinosaur-Bird transition. Current Biology 24.

Butler, R. J., S. L. Brusatte, B. Andres, and R. B. J. Benson. 2012. How do geological sampling biases affect studies of morphological evolution in deep time? a case study of Pterosaur (Reptilia: Archosauria) disparity. Evolution 66:147–162.

Cailliez, F. 1983. The analytical solution of the additive constant problem. Psychometrika 48:305–308.

Carrizo, L. V. and S. A. Catalano. 2015. First phylogenetic analysis of the tribe Phyllotini (Rodentia: Sigmodontinae) combining morphological and molecular data. Cladistics Pages 1–28.

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

Chen, Z.-Q. and M. J. Benton. 2012. The timing and pattern of biotic recovery following the end-Permian mass extinction. Nature Geoscience 5:375–383.

Ciampaglio, C. N. 2004. Measuring changes in articulate brachiopod morphology before and after the permian mass extinction event: do developmental constraints limit morphological innovation? Evolution and Development 6:260–274.

Ciampaglio, C. N., M. Kemp, and D. W. McShea. 2001. Detecting changes in morphospace occupation patterns in the fossil record: characterization and analysis of measures of disparity. Paleobiology 27:695–715.

Cisneros, J. C. and M. Ruta. 2010. Morphological diversity and biogeography of Procolophonids (Amniota: Parareptilia). Journal of Systematic Palaeontology 8:607–625.

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.

Close, R. A., M. Friedman, G. T. Lloyd, and R. B. Benson. 2015. Evidence for a mid-Jurassic adaptive radiation in mammals. Current Biology 25:2137–2142.

Cooper, N., J. Rodríguez, and A. Purvis. 2008. A common tendency for phylogenetic overdispersion in mammalian assemblages. Proceedings of the Royal Society of London B: Biological Sciences 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.

dos Reis, M., P. C. J. Donoghue, and Z. Yang. 2014. Neither phylogenomic nor palaeontological data support a Palaeogene origin of placental mammals. Biology Letters 10.

dos Reis, M., J. Inoue, M. Hasegawa, R. J. Asher, P. C. J. Donoghue, and Z. Yang. 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. Proceedings of the Royal Society of London B: Biological Sciences 279.

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.

Douady, C. J. and E. J. Douzery. 2003. Molecular estimation of eulipotyphlan divergence times and the evolution of "Insectivora". Molecular Phylogenetics and Evolution 28:285–296.

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.

Erwin, D. H. 1998. The end and the beginning: recoveries from mass extinctions. Trends in Ecology and Evolution 13:344–349.

Erwin, D. H. 2007. Disparity: Morphological pattern and developmental context. Palaeontology 50:57–73.

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 Sinauer Associate: Massachusetts.

Finarelli, J. A. and J. J. Flynn. 2006. Ancestral state reconstruction of body size in the Caniformia (Carnivora, Mammalia): The effects of incorporating data from the fossil record. Systematic Biology 55:301–313.

Finlay, S. and N. Cooper. 2015. Morphological diversity in tenrecs (Afrosoricida, Tenrecidae): comparing tenrec skull diversity to their closest relatives. PeerJ 3:e927.

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

Foote, M. 1994. Morphological disparity in Ordovician-Devonian crinoids and the early saturation of morphological space. Paleobiology 20:320–344.

Foote, M. 1996. Ecological controls on the evolutionary recovery of post-Paleozoic Crinoids. Science 274:1492–1495.

Foote, M. 1997. The evolution of morphological diversity. Annual Review of Ecology and Systematics Pages 129–152.

Foth, C., S. Brusatte, and R. Butler. 2012. Do different disparity proxies converge on a common signal? Insights from the cranial morphometrics and evolutionary history of Pterosauria (Diapsida: Archosauria). Journal of evolutionary biology 25:904–915.

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.

Garland, J., Theodore and A. R. Ives. 2000. Using the past to predict the present: Confidence intervals for regression equations in phylogenetic comparative methods. The American Naturalist 155:346–364.

Gavryushkina, A., T. A. Heath, D. T. Ksepka, T. Stadler, D. Welch, and A. J. Drummond. 2015. Bayesian Total Evidence dating reveals the recent crown radiation of penguins. arXiv preprint:1506.04797 .

Genner, M. J., O. Seehausen, D. H. Lunt, D. A. Joyce, P. W. Shaw, G. R. Carvalho, and G. F. Turner. 2007. Age of Cichlids: New dates for ancient lake fish radiations. Molecular Biology and Evolution 24:1269–1282.

Gingerich, P. D. 1987. Evolution and the fossil record: patterns, rates, and processes. Canadian Journal of Zoology 65:1053–1060.

Glor, R. E. 2010. Phylogenetic insights on adaptive radiation. Annual Review of Ecology, Evolution, and Systematics 41:251–270.

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

Goodman, S. M. and J. P. Benstead. 2003. Natural history of Madagascar. University of Chicago Press: Chicago.

Goswami, A., G. V. Prasad, P. Upchurch, D. M. Boyer, E. R. Seiffert, O. Verma, E. Gheerbrant, and J. J. Flynn. 2011. A radiation of arboreal basal eutherian mammals beginning in the late Cretaceous of India. Proceedings of the National Academy of Sciences of the United States of America 108:16333–16338.

Gould, S. J. and N. Eldredge. 1977. Punctuated equilibria: The tempo and mode of evolution reconsidered. Paleobiology 3:115–151.

Gower, J. C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53:325–338.

Gower, J. C. 1971. A general coefficient of similarity and some of its properties. Biometrics 27:857–871.

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

Guillerme, T. and N. Cooper. 2016. Effects of missing data on topological inference using a Total Evidence approach. Molecular Phylogenetics and Evolution 94, Part A:146–158.

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 of the United States of America 111:E2957–E2966.

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

Hetherington, A. J., E. Sherratt, M. Ruta, M. Wilkinson, B. Deline, and P. C. Donoghue. 2015. Do cladistic and morphometric data capture common patterns of morphological disparity? Palaeontology 58:393–399.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics Pages 65–70.

Hopkins, M. 2013. Decoupling of taxonomic diversity and morphological disparity during decline of the Cambrian trilobite family Pterocephaliidae. Journal of Evolutionary Biology 26:1665–1676.

Hopkins, M. J. and A. B. Smith. 2015. Dynamic evolutionary change in post-Paleozoic echinoids and the importance of scale when interpreting changes in rates of evolution. Proceedings of the National Academy of Sciences of the United States of America 112:3758–3763.

Huang, S., K. Roy, and D. Jablonski. 2015. Origins, bottlenecks, and present-day diversity: Patterns of morphospace occupation in marine bivalves. Evolution .

Hughes, M., S. Gerber, and M. A. Wills. 2013. Clades reach highest morphological disparity early in their evolution. Proceedings of the National Academy of Sciences of the United States of America 110:13875–13879.

Hunt, G. 2007. The relative importance of directional change, random walks, and stasis in the evolution of fossil lineages. Proceedings of the National Academy of Sciences of the United States of America 104:18404–18408.

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 of the United States of America 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.

Janvier, P. 2015. Facts and fancies about early fossil chordates and vertebrates. Nature 520:483–489.

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.

Jukes, T. H. and C. R. Cantor. 1969. Evolution of protein molecules. Mammalian protein metabolism 3:21–132.

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.

Klopfstein, S., L. Vilhelmsen, and F. Ronquist. 2015. A non-stationary Markov model detects directional evolution in hymenopteran morphology. Systematic Biology .

Korn, D., M. J. Hopkins, and S. A. Walton. 2013. Extinction space-a method for the quantification and classification of changes in morphospace across extinction boundaries. Evolution 67:2795–2810.

Kuhl, F. P. and C. R. Giardina. 1982. Elliptic Fourier features of a closed contour. Computer Graphics and Image Processing 18:236–258.

Kuhner, M. K. and J. Yamato. 2015. Practical performance of tree comparison metrics. Systematic Biology 64:205–214.

Lee, M. S. and R. M. Beck. 2015. Mammalian evolution: A Jurassic spark. Current Biology 25:R759–R761.

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.

Liow, L. H., T. Reitan, and P. G. Harnik. 2015. Ecological interactions on macroevolutionary time scales: clams and brachiopods are more than ships that pass in the night. Ecology Letters 18:1030–1039.

Lloyd, G. T. 2015. Claddis: Measuring Morphological Diversity and Evolutionary Tempo. R package version 0.1.

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

Lovegrove, B. G., K. D. Lobban, and D. L. Levesque. 2014. Mammal survival at the Cretaceous–Palaeogene boundary: metabolic homeostasis in prolonged tropical hibernation in tenrecs. Proceedings of the Royal Society of London B: Biological Sciences 281.

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.

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. W., J. E. Janečka, J. Gatesy, O. A. Ryder, C. A. Fisher, E. C. Teeling, A. Goodbla, E. Eizirik, T. L. L. Simão, T. Stadler, D. L. Rabosky, R. L. Honeycutt, J. J. Flynn, C. M. Ingram, C. Steiner, T. L. Williams, T. J. Robinson, A. Burk-Herrick, M. Westerman, N. A. Ayoub, M. S. Springer, and W. J. 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.

Mitchell, J. S. 2015. Extant-only comparative methods fail to recover the disparity preserved in the bird fossil record. Evolution .

Newham, E., R. Benson, P. Upchurch, and A. Goswami. 2014. Mesozoic mammaliaform diversity: The effect of sampling corrections on reconstructions of evolutionary dynamics. Palaeogeography, Palaeoclimatology, Palaeoecology 412:32–44.

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: New York.

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 post-K-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.

Olson, M. E. and A. Arroyo-Santos. 2009. Thinking in continua: beyond the "adaptive radiation" metaphor. BioEssays 31:1337–1346.

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 B: Biological Sciences 255:37–45.

Pant, S. R., A. Goswami, and J. A. Finarelli. 2014. Complex body size trends in the evolution of sloths (Xenarthra: Pilosa). BMC evolutionary biology 14:184.

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. Patané, 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. Müller, 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. B., A. Guisan, O. Broennimann, and C. F. Randin. 2008. Niche dynamics in space and time. Trends in Ecology and Evolution 23:149–158.

Prentice, K. C., P. Ruta, and M. J. Benton. 2011. Evolution of morphological disparity in Pterosaurs. Journal of Systematic Palaeontology 9:337–353.

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.

Raia, P., F. Carotenuto, F. Passaro, P. Piras, D. Fulgione, L. Werdelin, J. Saarinen, and M. Fortelius. 2012. Rapid action in the Palaeogene, the relationship between phenotypic and taxonomic diversification in Coenozoic mammals. Proceedings of the Royal Society of London B: Biological Sciences 280.

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. 1979. Size of the Permo-Triassic bottleneck and its evolutionary implications. Science 206:217–218.

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

Renne, P. R., A. L. Deino, F. J. Hilgen, K. F. Kuiper, D. F. Mark, W. S. Mitchell, L. E. Morgan, R. Mundil, and J. Smit. 2013. Time scales of critical events around the Cretaceous-Paleogene boundary. Science 339:684–687.

Revell, L. J. 2012. phytools: An R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3:217–223.

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

Rohlf, F. J. and L. F. Marcus. 1993. A revolution morphometrics. Trends in Ecology and Evolution 8:129–132.

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.

Ruta, M., K. D. Angielczyk, J. Fröbisch, and M. J. Benton. 2013. Decoupling of morphological disparity and taxic diversity during the adaptive radiation of Anomodont Therapsids. Proceedings of the Royal Society of London B: Biological Sciences 280.

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.

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

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

Simpson, G. 1944. Tempo and Mode in Evolution. Columbia University Biological Series Columbia University Press: New York.

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.

Slater, G. J., S. A. Price, F. Santini, and M. E. Alfaro. 2010. Diversity versus disparity and the radiation of modern cetaceans. Proceedings of the Royal Society of London B: Biological Sciences .

Smith, A. J., M. V. Rosario, T. P. Eiting, and E. R. Dumont. 2014. Joined at the hip: Linked characters and the problem of missing data in studies of disparity. Evolution 68:2386–2400.

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.

Springer, M. S., R. W. Meredith, E. C. Teeling, and W. J. Murphy. 2013. Technical comment on "The placental mammal ancestor and the post-K-Pg radiation of placentals". Science 341:613.

Stadler, T. 2011. Mammalian phylogeny reveals recent diversification rate shifts. Proceedings of the National Academy of Sciences of the United States of America 108:6187–6192.

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.

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

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 of the United States of America 108:8339–8344.

Toljagic, O. and R. J. Butler. 2013. Triassic-Jurassic mass extinction as trigger for the Mesozoic radiation of Crocodylomorphs. Biology Letters 9.

Torgerson, W. S. 1965. Multidimensional scaling of similarity. Psychometrika 30:379–393.

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.

Wesley-Hunt, G. D. 2005. The morphological diversification of carnivores in North America. Paleobiology 31:35–55.

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. 2001. Morphological disparity: A primer. Pages 55–144 *in* Fossils, Phylogeny, and Form (J. M. Adrain, G. D. Edgecombe, and B. S. Lieberman, eds.) vol. 19 of *Topics in Geobiology*. Springer: New York.

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: Baltimore.

Wilson, G. P. 2013. Mammals across the K-Pg boundary in northeastern Montana, U.S.A.: dental morphology and body-size patterns reveal extinction selectivity and immigrant-fueled ecospace filling. Paleobiology 39:429–469.

Wittenberg, R., R. Jadin, A. Fenwick, and J. Gutberlet, RonaldL. 2015. Recovering the evolutionary history of Africa's most diverse viper genus: morphological and molecular phylogeny of *Bitis* (Reptilia: Squamata: Viperidae). Organisms Diversity & Evolution 15:115–125.

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.

Wright, A. M., K. M. Lyons, M. C. Brandley, and D. M. Hillis. 2015. Which came first: The lizard or the egg? robustness in phylogenetic reconstruction of ancestral states. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution 324:504–516.

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

Yang, Z., S. Kumar, and M. Nei. 1996. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141:1641–50.

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

Zelditch, M. L., D. L. Swiderski, and H. D. Sheets. 2012. Geometric morphometrics for biologists: a primer. Academic Press: Waltham.

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

APPENDIX A

SUPPLEMENTARY INFORMATION TO CHAPTER 2

# 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 parameters 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.

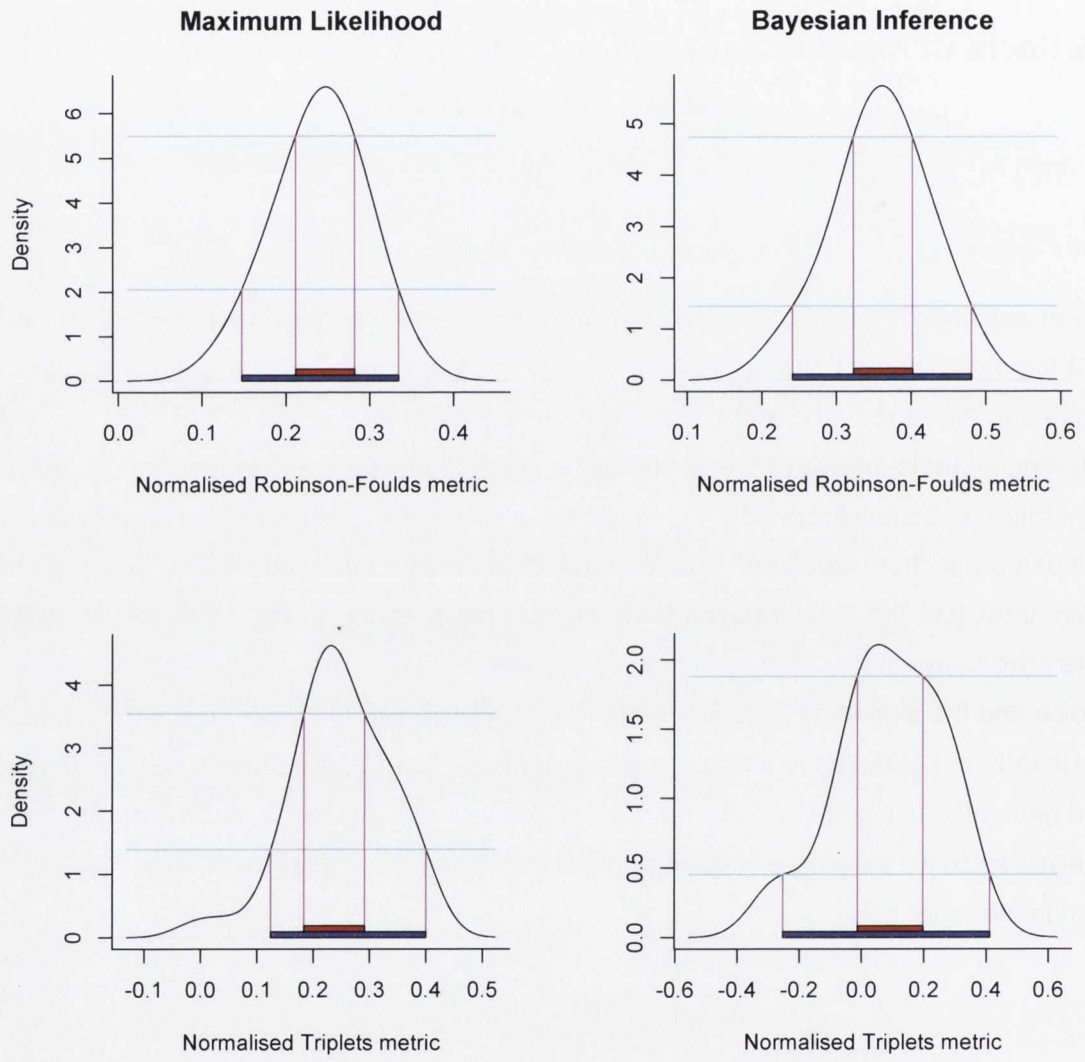



FIGURE A.1: Pairwise comparisons among the 50 "true" trees and the 50 "best" trees from the Maximum Likelihood and Bayesian inference methods. The horizontal blue and red lines represent, respectively, the 95% and 50% confidence intervals.

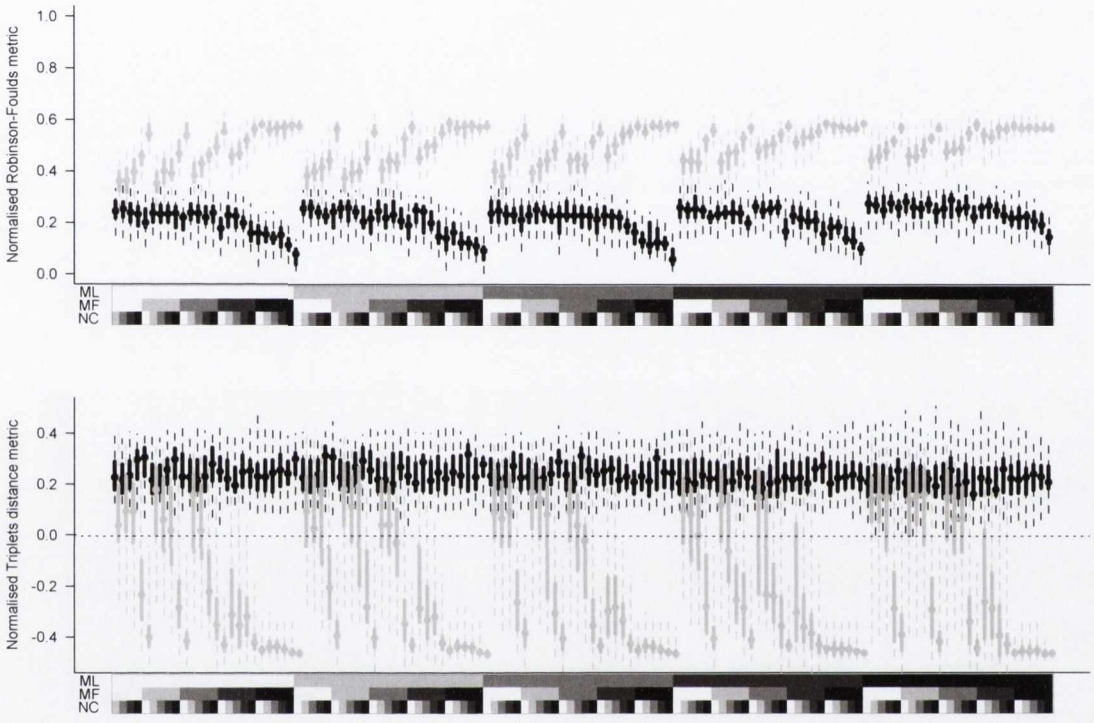



FIGURE A.2: **Effect of increasing missing data on recovering the "true" tree topology (the tree used for starting our simulations) for the Maximum Likelihood trees (black) and Bayesian consensus trees (grey).** The x axis shows the percentage of missing data from 0% (white) to 75% (black) for the two 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.

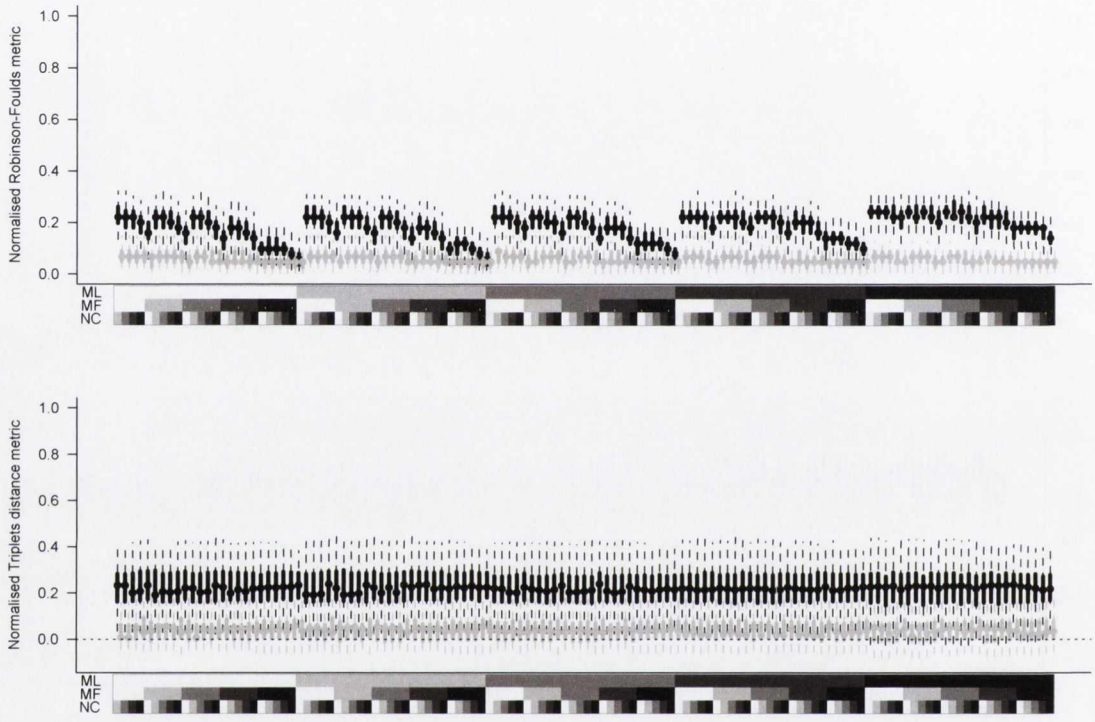



FIGURE A.3: **Effect of increasing missing data on topological recovering the "true" tree topology (the tree used for starting our simulations) for the Maximum Likelihood bootstrap trees (black) and Bayesian posterior tree distribution (grey).** The x axis shows the percentage of missing data from 0% (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 + $\Gamma_4$ (-m GTRGAMMA)
- Morphological data: M*kv* + $\Gamma_4$ (-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 ($\alpha$) = 0.5
  - Transition/Transversion ratio = 2 ($\beta$(80,40))
  - Starting tree: "True" tree topology with each branch length = 1
- Priors: Morphological data
  - rates distribution shape ($\alpha$) = 0.5
- Models
  - Molecular data: HKY + $\Gamma_4$
  - Morphological data: M*kv* + $\Gamma_4$
- MCMC
  - Two runs
  - Four chains per run
  - Generations $< 5\times10^7$
  - Sample frequency = $1.05\times10^4$
  - ASDS diagnosis frequency = $5\times1^4$
  - ASDS $< 0.01$
  - ESS $>>$ 200
  - Burnin = 25%

APPENDIX B

SUPPLEMENTARY INFORMATION TO CHAPTER 3

# Assessment of cladistic data availability for living mammals

## B.1 NUMBER OF TAXA WITH AVAILABLE CLADISTIC DATA FOR MAMMALIAN ORDERS WITHOUT ANY CHARACTER THRESHOLD

This table below shows the available morphological data for all living mammals, i.e. also including the living mammals from the matrices with fewer than 100 morphological characters (cf. Table 3.1; see methods in Chapter 3, section 3.2.2 for details). The phylogenetic structure of the available data was also calculated using the NTI and the NRI metrics as outlined in Chapter 3, section 3.2.2.

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 | |

## B.2 DISTRIBUTION OF LIVING SPECIES WITH AVAILABLE CLADISTIC DATA FOR EACH MAMMALIAN ORDER

The figures below show the distribution of living species with available cladistic data for each mammalian order excluding Primates and Carnivora that are illustrated in figure 3.3 (see methods in Chapter 3, section 3.2.2 for details).

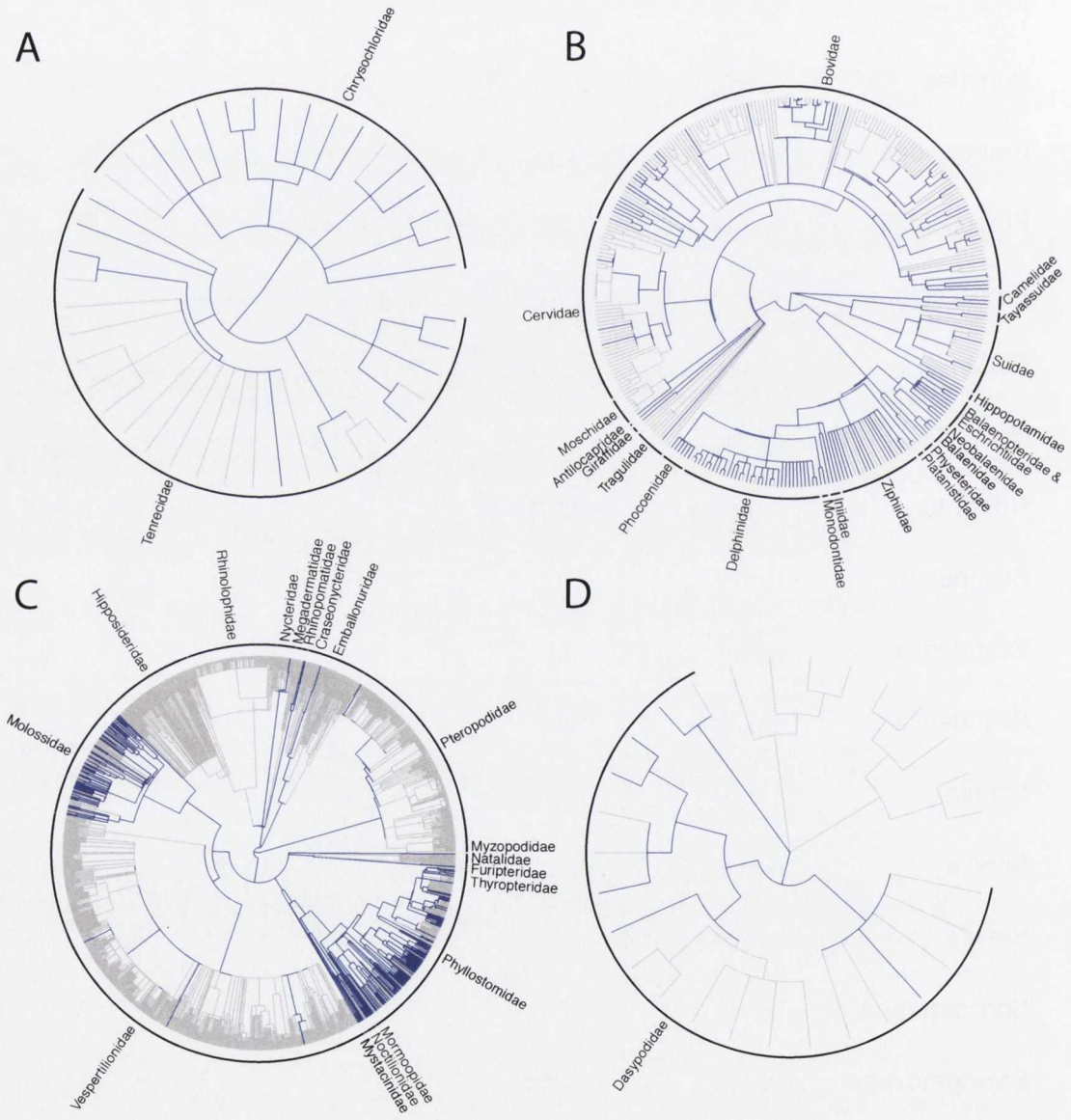



FIGURE B.1: Distribution of available morphological data across Afrosoricida (A), Cetartiodactyla (B), Chiroptera (C) and Cingulata (D). Edges are colored in grey when no morphological data are available or in blue when data are available.

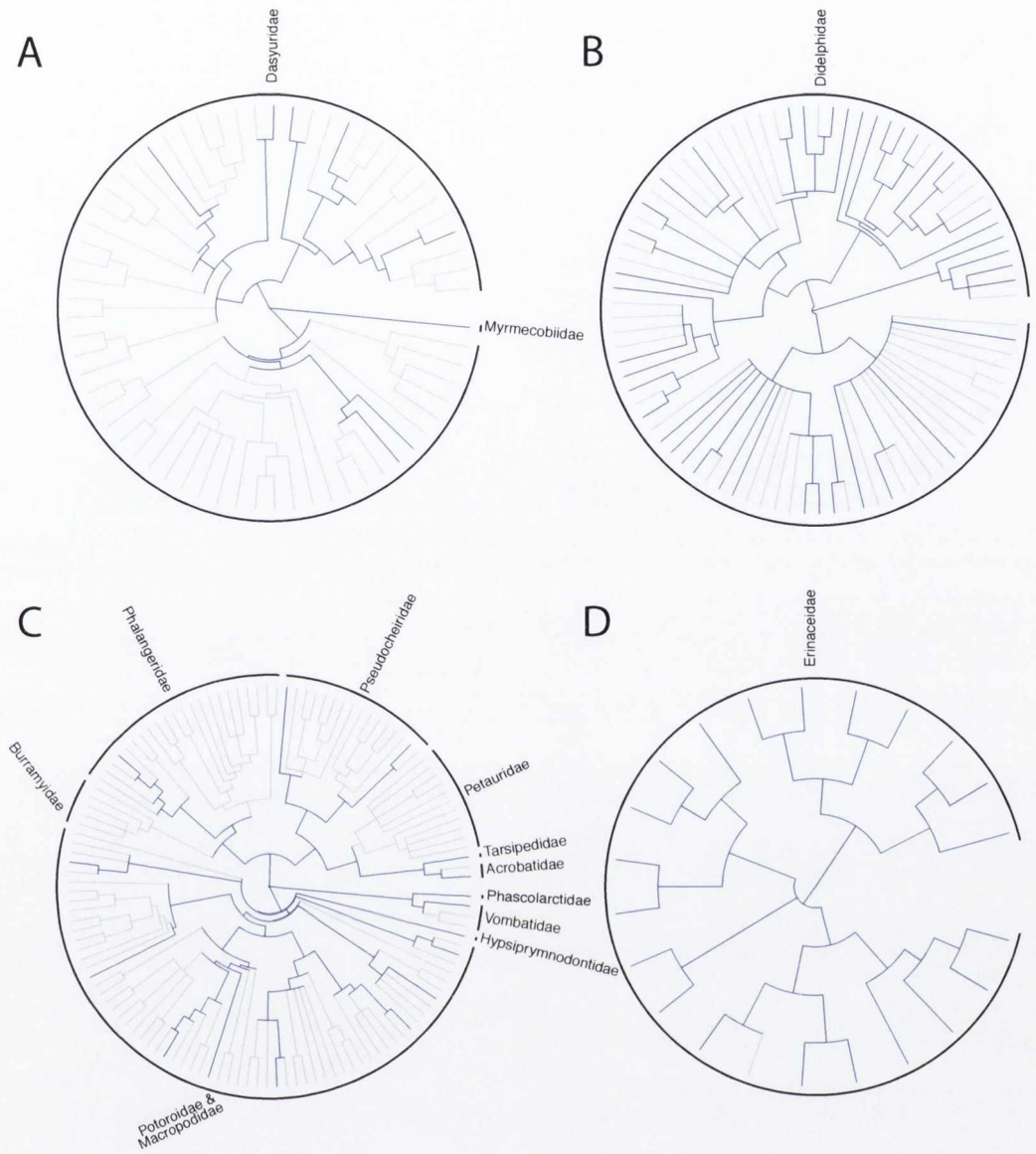



FIGURE B.2: Distribution of available morphological data across Dasyuromorphia (A), Didelphimorphia (B), Diprotodontia (C) and Erinaceomorpha (D). Edges are colored in grey when no morphological data are available or in blue when data are available.

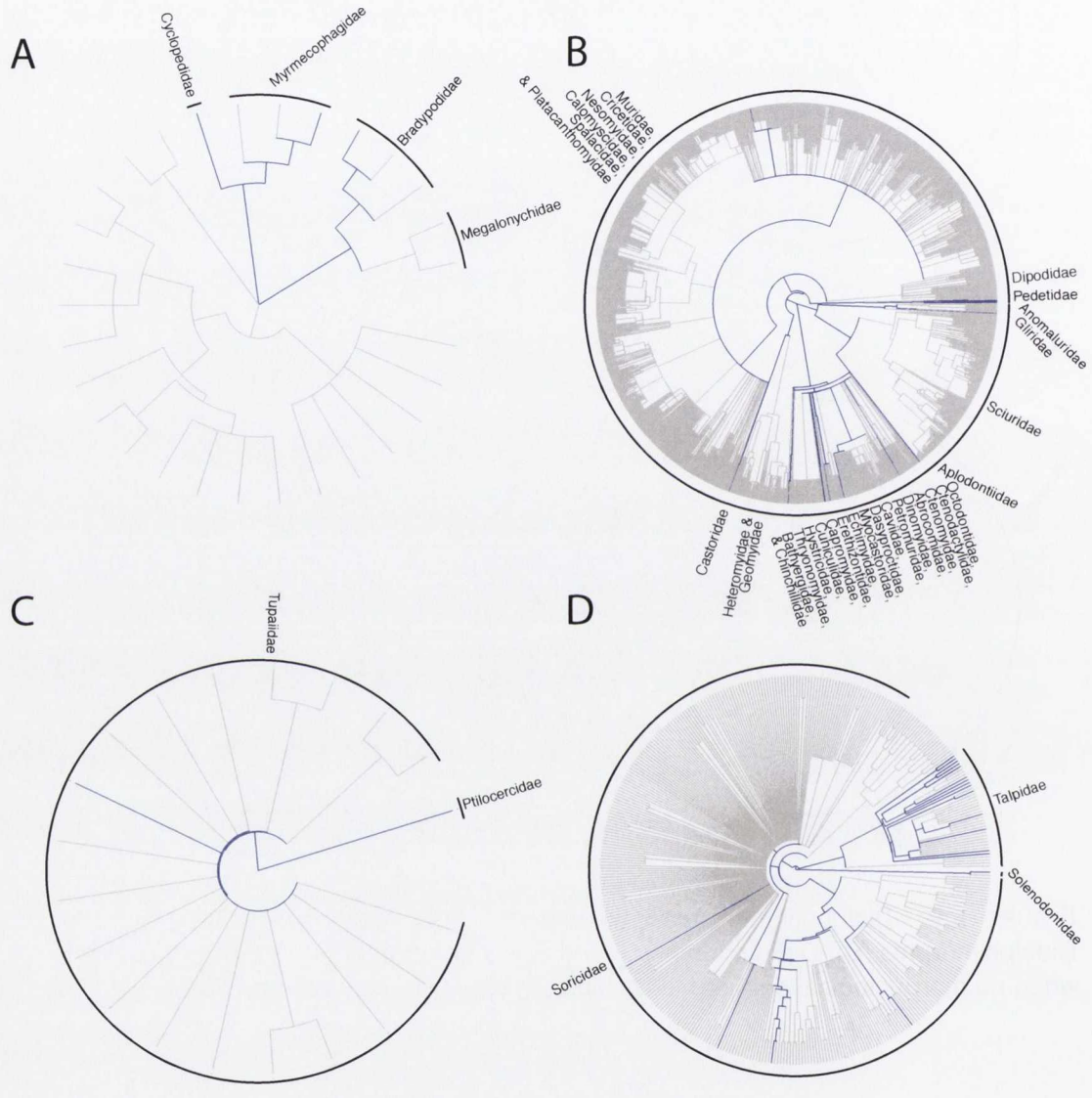



FIGURE B.3: Distribution of available morphological data across Pilosa (A), Rodentia (B), Scandentia (C) and Soricomorpha (D). Edges are colored in grey when no morphological data are available or in blue when data are available.

APPENDIX C

SUPPLEMENTARY INFORMATION TO CHAPTER 4

# Mammalian morphological diversity does not increase in response to the Cretaceous-Paleogene mass extinction and the extinction of the (non-avian) dinosaurs

## C.1 RAREFIED RESULTS FOR EACH DATASET

This section contains the results of the rarefied analyses of disparity through time for Mammaliaformes and Eutheria. We ran the rarefaction analysis to test whether disparity might be higher in subsamples with more phylogenetic elements simply because there are more taxa represented (see Chapter 4, section 4.2.6 for details). For each dataset, we reduced the number of tips and nodes to the strict minimum (i.e. three tips or nodes in each subsample) and to the minimum number of Mammaliaformes in the subsamples from 170 Ma to the present (eight tips or nodes in each subsample; as in Fig. 4.4). Note that these analyses do not change our results, see Tables 4.1 and 4.2 in Chapter 4 for more details.

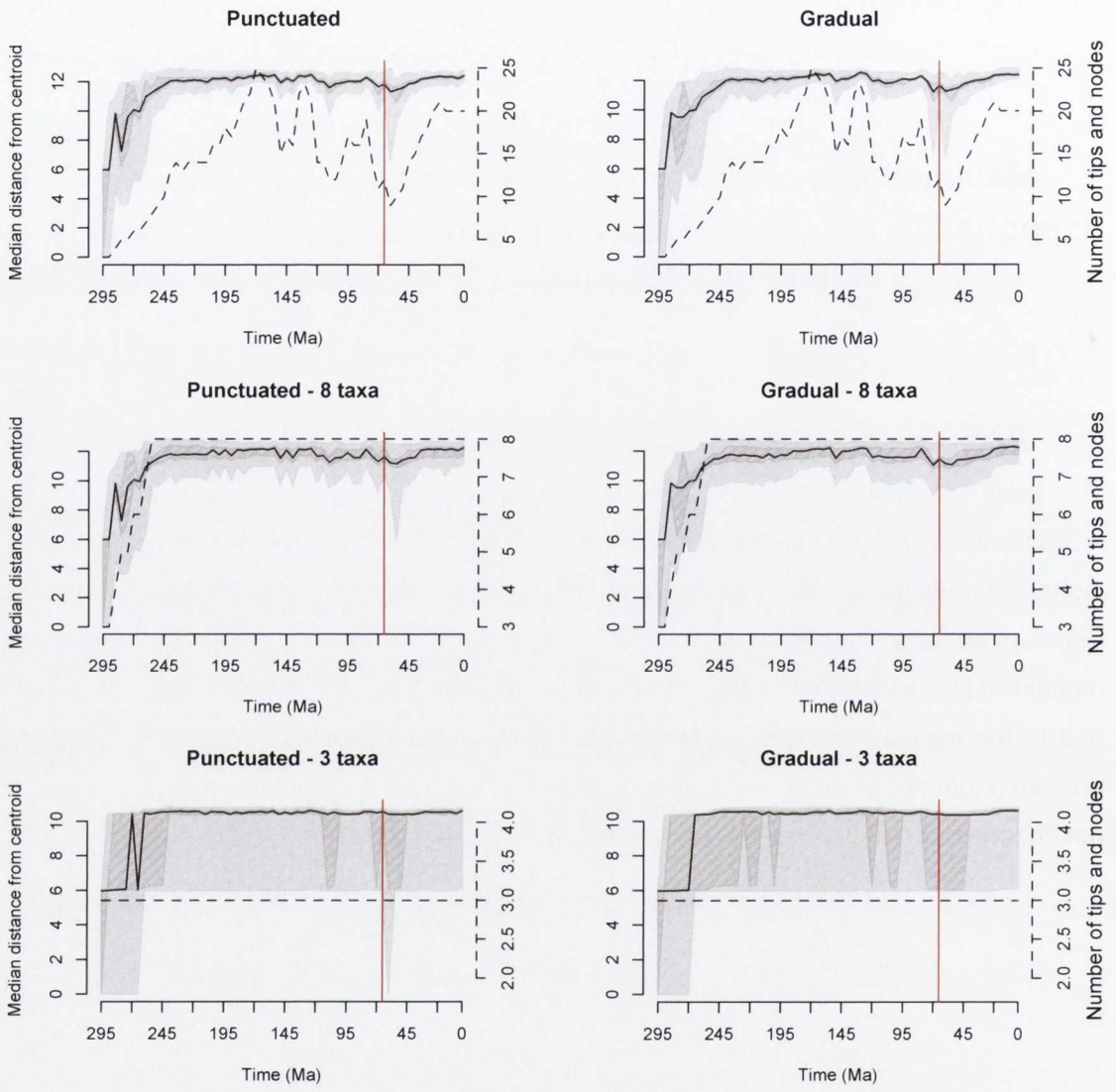



FIGURE C.1: **Observed and rarefied variation of disparity through time among Mammaliaformes with a punctuated or gradual evolution model.** The x axis represents time in millions of years before the present (Ma). The y axis represents disparity, measured as the median distance between the centroid of the ordinated space and the tips/nodes in each time subsample. The solid black lines show the mean disparity estimated from 1000 bootstrapped pseudoreplicates and confidence intervals (CI) are represented by the grey polygons (50% CI in dark grey and 95% CI in light grey). The dashed line and the right hand axis represents the number of tips/nodes in each time slice. The red vertical line indicates the Cretaceous-Paleogene (K-Pg) boundary (66 Ma). Note that scale bars differ among panels.

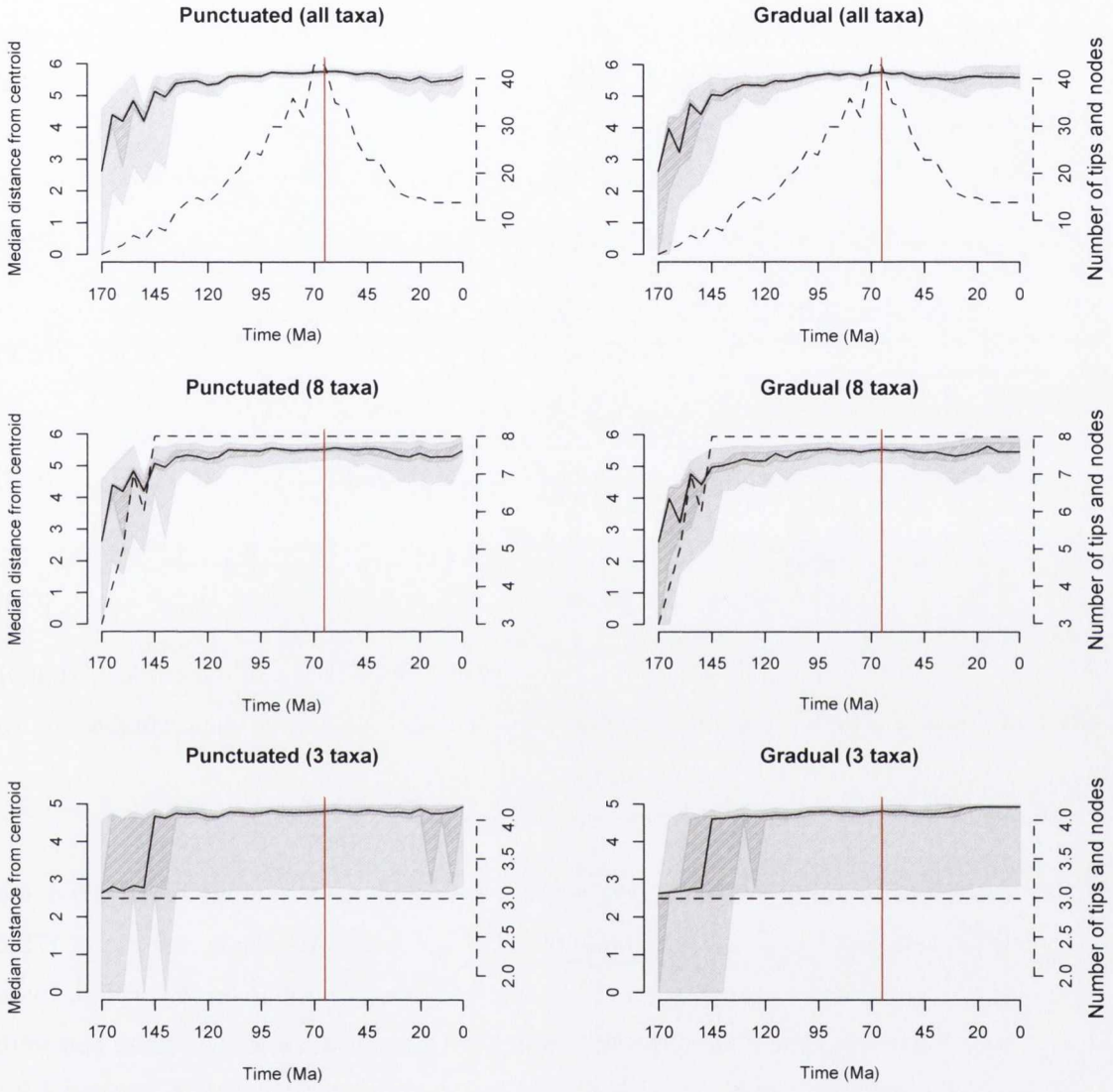



FIGURE C.2: **Observed and rarefied variation of disparity through time among Eutheria with a punctuated or gradual evolution model.** The x axis represents time in millions of years before the present (Ma). The y axis represents disparity, measured as the median distance between the centroid of the ordinated space and the tips/nodes in each time subsample. The solid black lines show the mean disparity estimated from 1000 bootstrapped pseudoreplicates and confidence intervals (CI) are represented by the grey polygons (50% CI in dark grey and 95% CI in light grey). The dashed line and the right hand axis represents the number of tips/nodes in each time slice. The red vertical line indicates the Cretaceous-Paleogene (K-Pg) boundary (66 Ma). Note that scale bars differ among panels.

## C.2 COMPARISON OF DIFFERENT DISPARITY METRICS AND TIME SAMPLING METHODS

This section (Fig. C.3, C.4, C.5 and C.6) contains the results of the variation of disparity through time analyses for Mammaliaformes and Eutheria using all the disparity metrics and methods for sampling disparity through time. The different disparity metrics are the median distance from centroid (see Chapter 4, section 4.2.4), and the sum and products of ranges and variances of the cladisto-space dimensions (Wills et al., 1994). The sum and products of ranges and variances are calculated as follows:

$$\text{Sum of ranges} = \sum (max(\mathbf{v}_n) - min(\mathbf{v}_n)) \tag{C.1}$$

$$\text{Sum of variances} = \sum (\sigma^2(\mathbf{v}_n)) \tag{C.2}$$

$$\text{Product of ranges} = \prod (max(\mathbf{v}_n) - min(\mathbf{v}_n)) \tag{C.3}$$

$$\text{Product of variances} = \prod (\sigma^2(\mathbf{v}_n)) \tag{C.4}$$

where $\mathbf{v}_n$ is any of the $n$ eigenvectors (i.e. any of the $n$ dimensions of the cladisto-space), $max$ and $min$ are respectively the maximum and minimum values of each eigenvector $\mathbf{v}_n$, and $\sigma^2$ is the variance of each eigenvector $\mathbf{v}_n$.

The different time sampling methods are as follows:

1. **Intervals (tips only)**. We selected every tip present at every geological stage (i.e. the smaller stratigraphic units) from the early Middle Jurassic (Bajocian, starting at 170.3 Ma) to the present. We collapsed together every stage containing fewer than three tips so that every time interval contained at least three tips. Note that some tips were present in multiple stages due to their occurrence data (see Chapter 4, section 4.2.1 for details).

2. **Intervals (tips and nodes)**. We selected tips and nodes present at every stage from the early Middle Jurassic (Bajocian, starting at 170.3 Ma) to the present. We collapsed together every stage containing fewer than three elements (tips and/or nodes).

3. **Slices (punctuated)**. These are the results presented in the main text where the cladisto-space is sampled every 5 Ma and the mode of evolution between each sub-sample is assumed to be punctuated (randomly selecting either data from the descendant or the ancestor when slicing through a branch; see Chapter 4, section 4.2.5 for details).

4. **Slices (punctuated: ACCTRAN)**. Similar to the slices (punctuated) method but data are always selected from the descendant (see Chapter 4, section 4.2.5 for details).

5. **Slices (punctuated: DELTRAN)**. Similar to the slices (punctuated) method but data are always selected from the ancestor (see Chapter 4, section 4.2.5 for details).

6. **Slices (gradual)**. These are the results presented in the main text where the cladistospace is sampled every 5 Ma and the mode of evolution is assumed to be gradual (data is selected from the descendant or the ancestor based on branch length; see Chapter 4, section 4.2.5 for details).

We also rarefied both dataset for all the metrics and all the methods using the minimum of three taxa for the interval methods, and eight taxa for the slices methods.

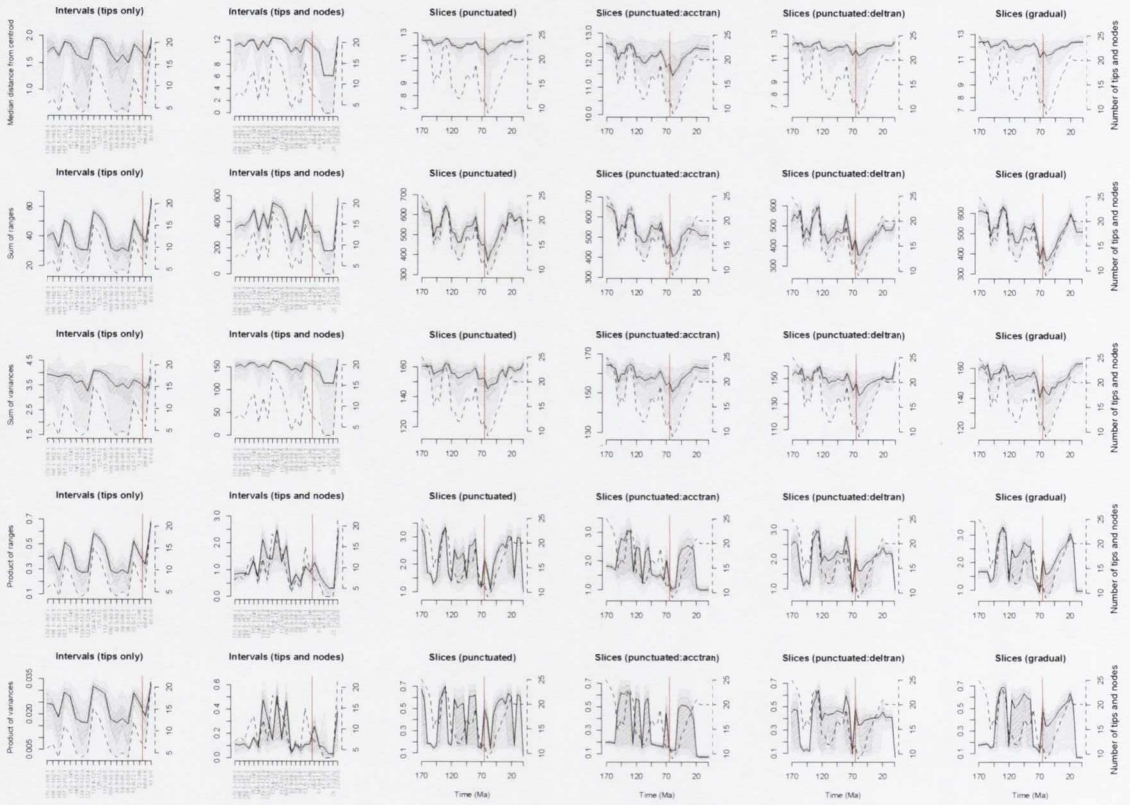



FIGURE C.3: **Variations in disparity through time among Mammaliaformes with different disparity measurements and methods for sampling disparity through time.** The x axis represents time in millions of years before the present (Ma). The y axis represents disparity, measured either as the median distance between the centroid of the ordinated space and the tips/nodes in each time subsample, or the sum and product of ranges and variances of each axis of the cladisto-space. The solid black lines show the mean disparity estimated from 1000 bootstrapped pseudoreplicates and confidence intervals (CI) are represented by the grey polygons (50% CI in dark grey and 95% CI in light grey). The dashed line and the right hand axis represents the number of tips/nodes in each time slice. The red vertical line indicates the Cretaceous-Paleogene (K-Pg) boundary (66 Ma). Note that scale bars differ among panels.

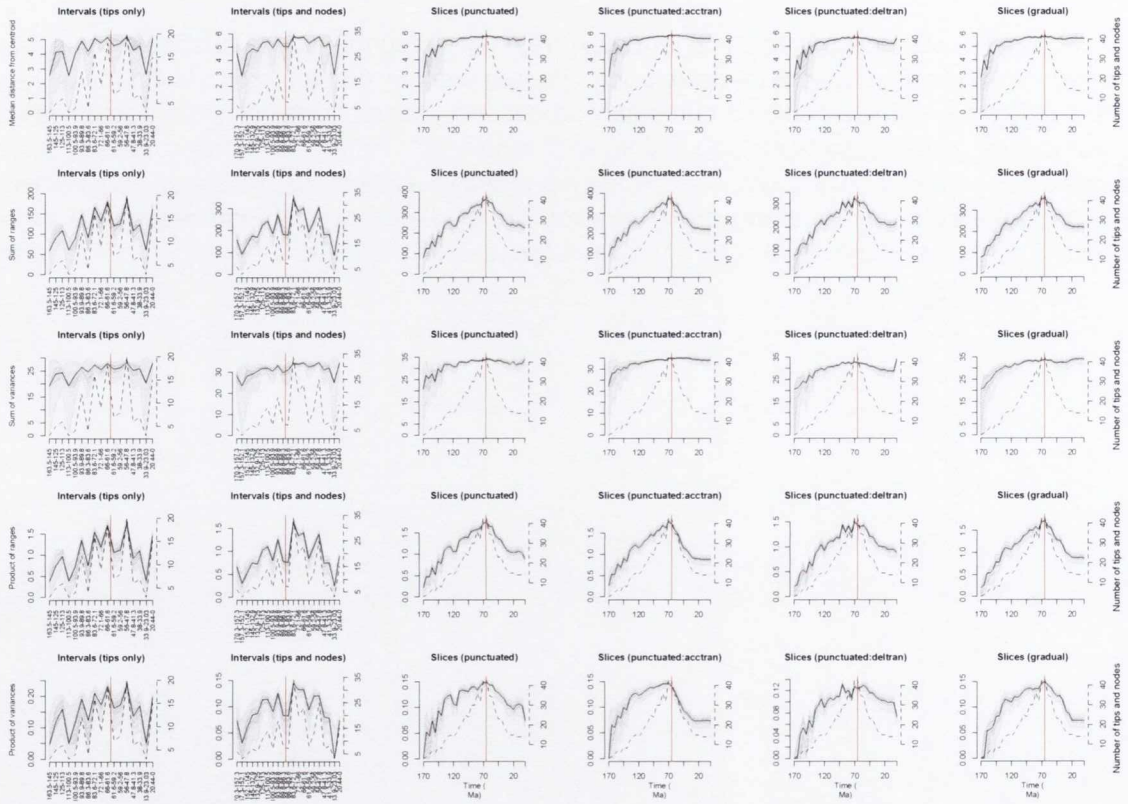



FIGURE C.4: **Variations in disparity through time among Eutheria with different disparity measurements and methods for sampling disparity through time.** The x axis represents time in millions of years before the present (Ma). The y axis represents disparity, measured as either the median distance between the centroid of the ordinated space and the tips/nodes in each time subsample, or the sum and product of ranges and variances of each axis of the cladisto-space. The solid black lines show the mean disparity estimated from 1000 bootstrapped pseudoreplicates and confidence intervals (CI) are represented by the grey polygons (50% CI in dark grey and 95% CI in light grey). The dashed line and the right hand axis represents the number of tips/nodes in each time slice. The red vertical line indicates the Cretaceous-Paleogene (K-Pg) boundary (66 Ma). Note that scale bars differ among panels.

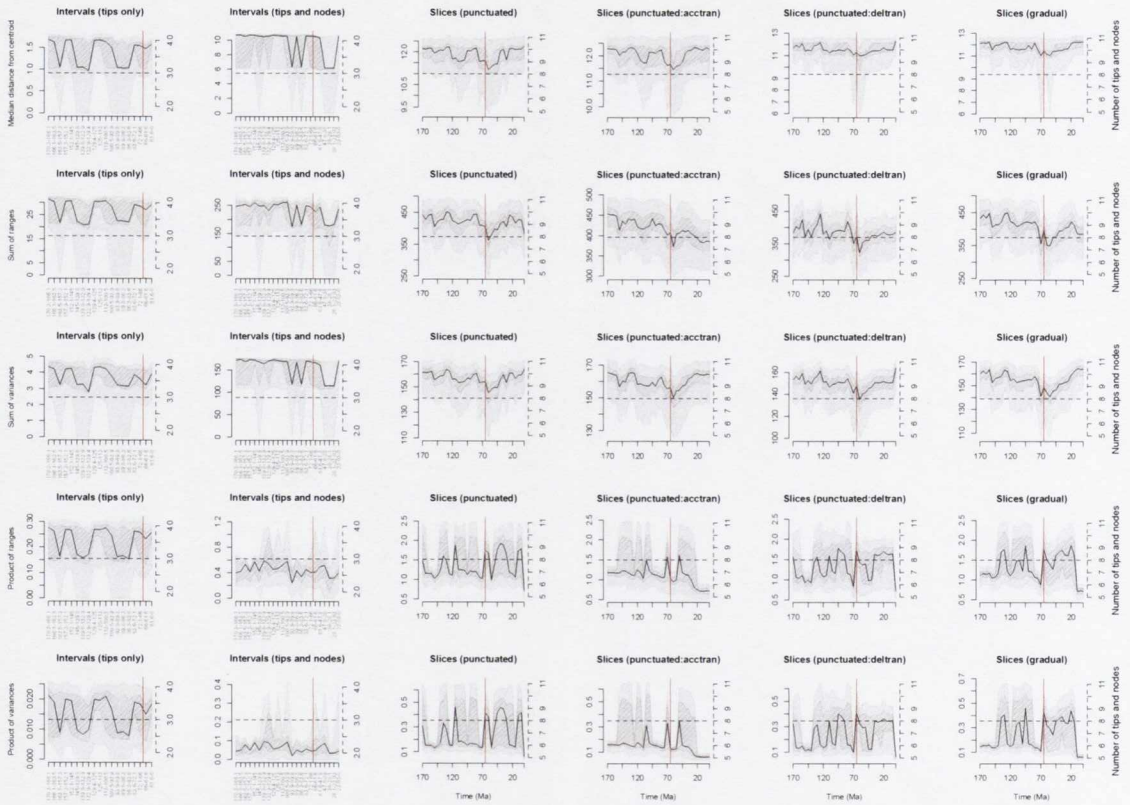



FIGURE C.5: **Variations in disparity through time among Mammaliaformes with different disparity measurements and methods for sampling disparity through time (rarefied with three taxa for the intervals methods and eight taxa for the slices methods).** The x axis represents time in millions of years before the present (Ma). The y axis represents disparity, measured as either the median distance between the centroid of the ordinated space and the tips/nodes in each time subsample, or the sum and product of ranges and variances of each axis of the cladisto-space. The solid black lines show the mean disparity estimated from 1000 bootstrapped pseudoreplicates and confidence intervals (CI) are represented by the grey polygons (50% CI in dark grey and 95% CI in light grey). The dashed line and the right hand axis represents the number of tips/nodes in each time slice. The red vertical line indicates the Cretaceous-Paleogene (K-Pg) boundary (66 Ma). Note that scale bars differ among panels.

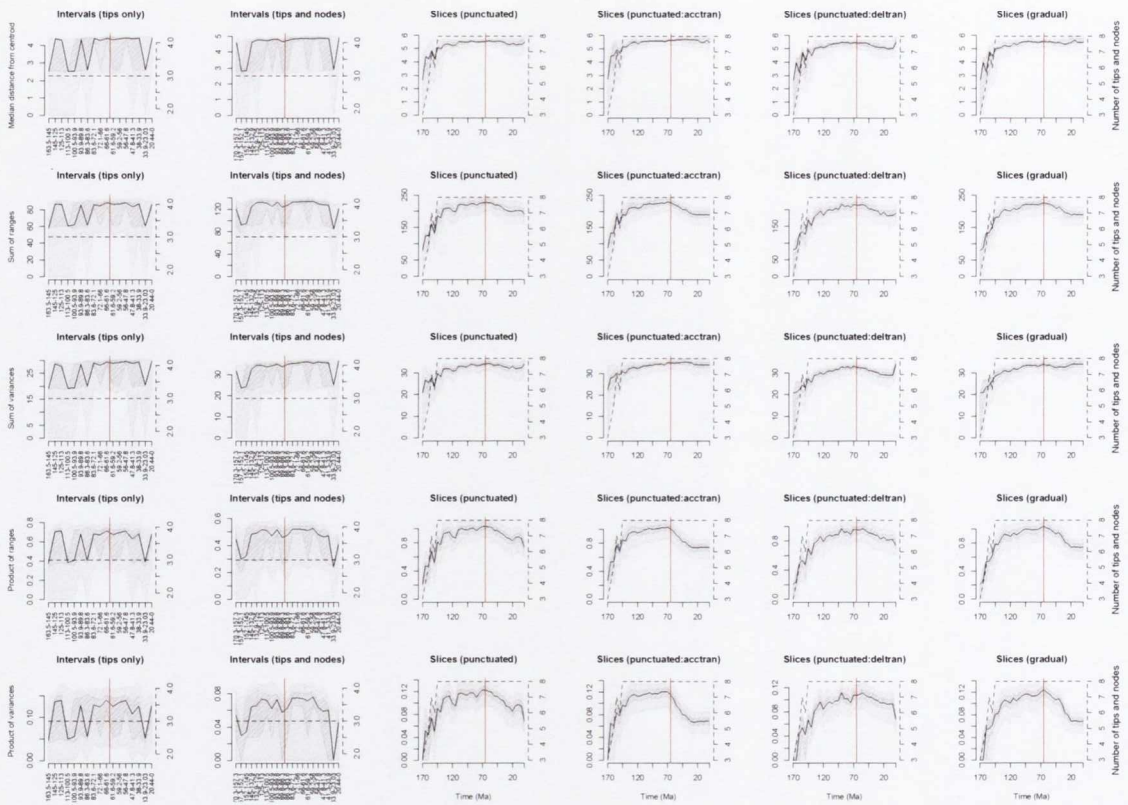



FIGURE C.6: **Variations in disparity through time among Eutheria with different disparity measurements and methods for sampling disparity through time (rarefied with three taxa for the intervals methods and eight taxa for the slices methods).** The x axis represents time in millions of years before the present (Ma). The y axis represents disparity, measured as either the median distance between the centroid of the ordinated space and the tips/nodes in each time subsample, or the sum and product of ranges and variances of each axis of the cladisto-space. The solid black lines show the mean disparity estimated from 1000 bootstrapped pseudoreplicates and confidence intervals (CI) are represented by the grey polygons (50% CI in dark grey and 95% CI in light grey). The dashed line and the right hand axis represents the number of tips/nodes in each time slice. The red vertical line indicates the Cretaceous-Paleogene (K-Pg) boundary (66 Ma). Note that scale bars differ among panels.