The Estimation of Phylogeny and Evolutionary Timescales Using Molecular and Morphological Data

By

Joseph Edward O'Reilly

School of Earth Sciences

The University of Bristol

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Doctor Of Philosophy in the Faculty of Science.

August 2017

Word Count: 34,710

"*Facts are meaningless. You can use facts to prove anything that's even remotely true.*"

Homer J. Simpson

Abstract

The estimation of accurate divergence times is key to the construction of a timescale of evolutionary events. Such a timescale is of vital importance in evolutionary biology as it is required for testing the majority of hypotheses regarding the process of evolution. The methods used to estimate such timescales have become increasingly sophisticated, with the most recent improvements allowing for the combined analysis of molecular data from extant taxa and morphological data from both extinct and extant taxa. This approach allows all available palaeontological data to inform divergence time estimates, something that was not possible before. Despite the obvious potential benefits of this total evidence approach there are many outstanding questions regarding the proper application of palaeontological data in this framework and the accuracy of the estimated divergence times. In this thesis I consider several such questions, concentrating on the key topics of fossil taxon calibration construction, the impact of differential preservation of fossil data on parameter estimates, effectively summarising posterior samples of fossil trees, and the general efficacy of probabilistic frameworks for morphological phylogenetic inference. I demonstrate the impact of these areas on the accuracy of methods incorporating fossil morphological data for divergence time analysis or phylogenetic reconstruction, highlighting both the challenges this framework presents when paleontological data is improperly incorporated but also the benefits a total-evidence approach affords when fossil data is properly integrated.

Dedications and Acknowledgements

First and foremost, I would like to thank my main supervisor, Phil Donoghue, who has provided guidance, insightful discussion and many academic opportunities for me over the course of my PhD, and with whom I am deeply grateful to have had the chance to work alongside. I would also like to thank my other supervisors, Davide Pisani and Ziheng Yang, for offering support and discussion. Additionally, I would also like to thank Mario dos Reis, who has been something of an unofficial supervisor to me, and who has happily answered many of my annoying statistical questions.

Thank you to the graduate students of The Bristol Palaeobiology research group for providing such a studious environment in which rigorous scientific enquiry is occasionally the priority. Specific thanks must go to, Mark Puttick, Joe Keating, Luke Parry, Al Tanner, James Fleming, Fiona Walker, Leanne Melbourne, and Richard Taylor for a mix of scientific and not-so-scientific discussion over the years. I would also like to thank Rachel Warnock, who set me up with some code in my first year that introduced me to programming and has since become absolutely indispensible.

I cannot thank my family enough. Thanks to my parents for showing pride in me, and being some of the handful of people who don't think that a PhD isn't "a real job". Thanks to Rose for moving to Bristol, seeing me on a monthly basis and providing free trips to the zoo; and thanks to Hugh "Jasper" O'Reilly for generally being a laugh.

Thanks must also go to the 2013 Earth Sciences PhD intake for providing lots of "social activities", particularly in first year. Thanks to Emma for supporting my initial decision to pursue a PhD. Thanks to Luke and Charlie for watching University Challenge with me every Monday. Thanks to Lou and Carmelisa for making me a more international person.

Finally, special thanks go to Hannah for unwavering support and care, and without whom I would have never achieved what I have so far. Thank you.

Dedicated to my Grandmother and the two close friends that I lost along the way.

Statement of Collaboration

Chapters 1,2,3,4,5, and 6 were collaborations with Philip C. J. Donoghue. Chapters 1 and 2 were collaborations with Mark Puttick, Luke Parry, Alastair R. Tanner, James E. Tarver, James Fleming, Davide Pisani, Lucy Holloway, and Jesus Lozano-Fernandez. Chapter 3 was produced in collaboration with Mario dos Reis. For all chapters I collected data, designed and performed analyses, interpreted results and led the writing.

Declaration

I declare that the work in this thesis was carried out in accordance with the Regulations of the
University of Bristol. The work is original, except where indicated by in text references, and no part has been submitted for any other academic award. Work done in collaboration with and assistance of others is indicated. Any views expressed are those of the author.

SIGNED: ..... DATE:.....

Table of Contents

Introduction, 12

The Importance of Morphology, 13

Reconstructing Phylogenetic Relationships With Morphological Data, 14

Divergence Time Estimation Methods, 15

Relaxing the Molecular Clock, 17

The Evolution of Inference Frameworks for Divergence Time Estimation, 27

Bayesian Calibration of The Clock, 19

Difficulties Encountered When Integrating Molecular and Morphological Data, 22

Overview of Chapters, 24

Chapter One - Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data

1.1 – Abstract, 28

1.2 – Introduction, 29

1.3 – Materials and Methods, 29

1.4 – Results, 31

1.5 – Discussion, 35

1.6 – Conclusions, 36

Chapter Two - Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data

2.1 – Abstract, 38

2.2 – Introduction, 39

2.3 – Materials and methods, 40

2.3.1 - Simulation of morphological matrices, 40

2.3.2 - Phylogenetic analysis, 40

2.3.3 - Empirical analyses, 41

2.4 – Results, 42

2.4.1 - Simulated data, 42

2.4.2 - Empirical phylogenies, 47

2.5 – Discussion, 50

2.5.1 - Simulations indicate that the Bayesian implementation of the Mk model outperforms all other methods and implementations, 50

2.5.2 - Analyses of empirical data bear out conclusions based on simulations, 52

2.5.3 - Implications for phylogenetic analysis of phenotypic data, 52
2.6 – Conclusions, 53

Chapter Three - Dating tips for divergence time estimation
3.1 – Abstract, 56
3.2 – Introduction, 57
3.3 – Models of morphological character evolution and the incompleteness of fossils, 59
3.4 – Dating tips and calibration strategies, 60
3.5 – Total Evidence Dating - less than the sum of its parts?, 68
3.6 – Conclusions, 70

Chapter Four - Tips and nodes are complimentary not competing approaches to the calibration of molecular clocks
4.1 – Abstract, 73
4.2 – Introduction, 74
4.3 – Materials and Methods, 75
4.4 – Results, 75
4.5 – Discussion, 79
4.6 – Conclusions, 80

Chapter Five - Isolating and Mitigating the Effects of Fossilization Processes on the Accuracy of Divergence Time Estimates
5.1 – Abstract, 83
5.2 – Introduction, 84
5.3 – Materials and Methods, 87
5.3.1 - Simulating the Loss of Soft Tissue Anatomical Data, 87
5.3.2 - Simulating the Effects of Post-Decay Biostratinomic Processes, 87
5.3.3 - Exploring the Effects of Character Matrix Dimensions, 88
5.3.4 - Divergence Time Estimation, 88
5.3.5 - Averaging Over Consensus Trees, 88
5.4 – Results, 89
5.4.1 - Loss of Data Resulting From the Decay of Soft Tissue Characters, 89
5.4.2 - Loss of Data Resulting from the Effects of Later Biostratinomic Processes, 91
5.5 – Discussion, 93
5.5.1 - The Impact of Data Loss Resulting From Decay of Soft Tissue Anatomy, 93
5.5.2 - The Impact of Data Loss Resulting From Post-Decay Biostratinomic Processes, 93

5.5.3 - Mitigating Against the Effects of Missing Data, 94
5.6 – Conclusion, 94

Chapter Six - Consensus trees and the estimation of topology and time from morphological data
6.1 – Abstract, 97
6.2 – Introduction, 98
6.3 – Materials and Methodology, 99
6.3.1 - Simulated Matrices, 100
6.3.2 - Empirical Matrices, 100
6.3.3 - Divergence Time Estimation, 100
6.3.4 - Consensus Tree Efficacy Tests, 100
6.4 – Results, 101
6.4.1 - Simulated Matrices, 102
6.4.2 - Empirical Matrices, 110
6.5 – Discussion, 112
6.6 – Conclusion, 113

Conclusions, 115
References, 118
Appendix, 127

List of Figures

Figure 1.1 – The relative performance of Bayesian and Parsimony Inference, 32
Figure 1.2 – The Mk model exhibits higher accuracy with lower precision than parsimony methods, 34
Figure 2.1 - Contour plots of Robinson-Foulds distance against phylogenetic resolution, 43
Figure 2.2 - Accuracy of nodes is higher for those closer to the tips in the asymmetrical Trees, 44
Figure 2.3 - Alternative phylogenetic reconstruction methods alter our understanding of evolution with empirical matrices, 48
Figure 2.4 - Alternative phylogenetic reconstruction methods produce generally congruent reconstructions of evolution with empirical matrices, 49
Figure 3.1 - Construction of a tip-calibration, 63
Figure 3.2 - A dated phylogeny of Hymenoptera produced using node-calibrations, 66
Figure 3.3 - Comparison between marginal posterior distributions on 9 node ages estimated with TED, and prior clade-age constraints employed for node-calibrated analysis of the same data, 67
Figure 4.1 - Time calibrated phylogenies of Hymenoptera based on: tip calibration; node calibration; and combined tip and node calibration, 77
Figure 4.2 - Infinite-sites plots for three alternative calibration approaches for both the posterior and prior distribution of times of 9 clades, 78
Figure 5.1 - The effect of different stages of the fossilization process on the distribution of missing data in morphological matrices, 86
Figure 5.2 - Mammalian time scaled phylogeny, 90
Figure 5.3 - Mammalian time scaled phylogeny, 92
Figure 6.1 - The number and proportion of erroneous bifurcations in maximum clade credibility and majority rule consensus summarisation of simulated data sets of different sizes plotted against the number of unique bifurcations sampled in each analysis, 107
Figure 6.2 - The number and proportion of correct and incorrect clades presented in consensus trees, 108
Figure 6.3 – Histogram showing that both MCC and MAP trees contain more incorrect nodes than correct ones, 109
Figure 6.4 - Maximum Clade Credibility and Majority Rule consensus trees for Hymenoptera ,111

List of Tables

Table 1.1 - The differences in median and the 95th percentile range of Robinson–Foulds values between the Mk and both parsimony models, 33
Table 2.1 - Bayesian approaches produce the most accurate trees for all character sets, 45
Table 2.2 - P Values from Spearman's rank correlation between the percentage of nodes being accurately reconstructed and their distance from the root, 46
Table 6.1 - Number of unique sampled bipartitions obtained from the posterior distribution for 100 replicate simulated datasets, 103
Table 6.2 - Absolute number of incorrect clades in maximum clade credibility (MCC), majority rule consensus (MRC), and maximum a posteriori (MAP) trees constructed for posterior distributions sampled from 100 replicate simulated data sets, 104
Table 6.3 - Features of valid but unrepresented clades in MCC trees used to summarise posterior samples of trees using different data types, 105
Table 6.4 - Features of valid but unrepresented clades in MAP trees constructed from posterior distributions obtained using different data types., 106

Introduction

The Importance of Morphology

The morphological characters that unite clades or define taxa provide a key source of insight into the evolutionary process that has given rise to the diversity of both past and present life. The manner in which characters have been acquired and lost over time allows for the construction of hypotheses regarding many enigmatic aspects of evolution, such as rapid diversifications (Beck and Lee, 2014, Lee et al., 2013) or the developmental origins of specific morphological structures (Donoghue and Keating, 2014). If we wish to test these hypotheses, a model of taxonomic classification describing the relationships and shared ancestry amongst taxa is required, and Phylogenetic trees provide the model against which such hypotheses can be tested. Phylogenies are constructed by inferring the shared ancestry of taxa through interpretation of the evolutionary history of homologous characters across a sample of taxa. This interpretative approach can be formalised in a statistical framework, in which trees are estimated using one many available optimality criteria. Once taxonomic relationships are estimated it is then possible to apply fossil occurrence data in combination with morphology to elucidate the absolute timescale over which the constituent clades of the tree appeared (Pyron, 2011, Ronquist et al., 2012a). This method allows for further tests, regarding the rate and timing of evolutionary phenomena supported by the phylogeny.

As the cost and time associated with obtaining molecular sequence data has decreased, it has broadly supplanted morphology as the standard data source for inferring phylogenies and evolutionary timescales. Despite its popularity, molecular data is almost exclusively unavailable for extinct taxa. As a result of this, many evolutionary hypotheses regarding extinct taxa are unable to be adequately tested with molecular phylogenies alone. In such cases the use of morphological data, either on its own or in tandem with molecular data, can enable these tests by estimating the phylogenetic relationships of fossil taxa. Furthermore, extinct taxa are part of the evolutionary diversification process, and accounting for their morphology and distribution through time will provide a more comprehensive insight into the history of life (Heath et al., 2014, Zhang et al., 2016, Gavryushkina et al., 2014). Recognition of the importance of fossil taxa for investigating evolutionary history has resulted in the recent development of a number of methods in which morphological data is applied to estimate both evolutionary timescales and phylogenetic relationships (Ronquist et al., 2012a, Pyron, 2011, Heath et al., 2014). These methods allow morphological data to inform evolutionary timescales and phylogenetic relationships, either alone or in combination with molecular data. This approach appropriates molecular clock methodology for morphological data, allowing the chronological distribution and phylogenetic relationships of fossil taxa to refine estimates of clade age. Such a revolutionary application of morphological data has rejuvenated interest in the assessment of

the accuracy of methods that use this data source for estimating phylogenetic relationships and evolutionary timescales.

Reconstructing Phylogenetic Relationships with Morphological Data

For divergence time analyses using morphological data, whether co-estimating times and topology or estimating ages on a fixed topology, the accuracy of divergence time estimates is contingent on an ability to accurately estimate phylogenetic relationships. When co-estimating times and topology an exclusively probabilistic approach to accommodating morphological data is required. When estimating ages on a fixed topology a number of methods can be applied to obtain the tree on which ages are subsequently estimated. The relative accuracy of these competing methods is therefore important to consider when selecting between different approaches to divergence time estimation.

The use of morphological data to establish phylogenetic relationships amongst taxa has traditionally relied on the use of the maximum parsimony optimality criterion to choose between competing hypotheses (Farris, 1970). The most parsimonious phylogeny is the one that requires the least number of character changes to explain. Therefore, this approach assumes that the simplest explanation of morphological evolution provides the best approximation of the true relationships between taxa.

Recently there has been renewed interest in applying probabilistic methods to estimate phylogenies from morphological data (Bapst et al., 2016, Wright and Hillis, 2014, Wright et al., 2016). In a probabilistic framework the process of morphological change is explicitly modelled to allow for estimation of the tree topology and evolutionary distances between taxa at the same time (Lewis, 2001). Application of a probability-based approach allows for the use of maximum likelihood as the optimality criterion when estimating model parameters. Alternatively, a Bayesian approach can be applied, in which the posterior distribution of the phylogenetic relationships between taxa is estimated.

Each of these available inference frameworks has unique strengths and weaknesses. Parsimony does not explicitly model the process by which changes in characters occur, it also struggles to accommodate homoplasy, and has been demonstrated to be statistically inconsistent (Felsenstein, 1978). Maximum Likelihood estimation of topology does not allow for uncertainty in tree structure and requires *post hoc* measures of clade support (Felsenstein, 1985), but it does employ a model of morphological evolution. A relatively small amount of information in morphological datasets means that Bayesian inference is likely to often result in a diffuse

posterior distribution of trees (Gelman et al., 2013), which will be difficult to reconcile in a single consensus tree. Despite this, Bayesian phylogenetic inference allows for an easily interpretable measure of tree support with the use of clade posterior probabilities. It is still unclear which methodological framework provides the most accurate estimates of phylogeny for this data source. Initial attempts to assess the relative efficacy of the different available methods supported the use of Bayesian inference (Wright and Hillis, 2014). Unfortunately, this result was built on a simulation framework that was potentially biased and resulted in data that contained higher levels of homoplasy than expected in empirical datasets (Sanderson and Donoghue, 1989, Sanderson, 1996). The bias in the simulated data strongly favoured Bayesian methods and it is therefore still unknown which method is the most accurate when analysing morphological data.

Divergence Time Estimation Methods

Establishing a timescale for the Tree of Life is a key goal for palaeontologists due to its necessity for testing hypotheses regarding the tempo and mode of evolutionary processes. The manner by which such timescales are estimated has become increasingly sophisticated, initial efforts simply attempted to use putative inter-specific relationships between fossil taxa and the age of the geological units from which they were sampled to obtain relatively crude timescales (Gidley, 1907, Stirton, 1940). The advent of widely available molecular sequence data and statistical molecular phylogenetic methods introduced a marked improvement in the way in which timescales could be estimated thanks to the models of molecular substitution that are required in this framework.

A key requirement for statistical phylogenetic inference is a model of the stochastic process of evolutionary change that occurs within individual morphological or molecular characters across a phylogenetic tree. A range of stochastic models have been developed that employ continuous time Markov chains to describe the process by which these changes accumulate. The specific random variables in each of these models allow for estimation of specific aspects of the molecular substitution process, such as the degree of transition transversion bias, but all of these models require the estimation of the underlying amount of evolutionary change, which is represented in the estimated branch length as a product of the substitution rate and the time period over which changes have accumulated. Explicitly modelling the substitution process in this way made it possible to estimate molecular phylogenies with estimated branch lengths representing evolutionary distance d, measured in units of expected number of molecular substitutions per site, for each individual lineage. The distance between taxa is a product of both

the rate at which substitutions accumulate r, and the time over which the distance has accumulated, t,

$$d = rt\,.$$

Unfortunately, these parameters are confounded and not individually identifiable, as molecular sequence data alone contains no information regarding absolute rates or times. Therefore, without information about both the timescale over which the phylogeny has developed, and the distribution of evolutionary rate across this envelope of time, branch lengths may only be expressed in terms of estimated evolutionary distance. Fortunately, the molecular clock hypothesis provides a method by which to separate the rates and times when a probabilistic framework is employed, by assuming that the rate of evolutionary change is constant (Zuckerkandl and Pauling, 1965). The molecular clock therefore assumes that there is a linear relationship between the number of molecular differences between any two extant taxa and the time since their most recent common ancestor. This relationship was demonstrated using the observed differences between proteins isolated from different taxa and the divergence time of the lineages leading to these taxa as supported by the fossil record. It was observed that different proteins accumulated molecular substitutions at different rates, but the linearity was expected to broadly hold across different loci and lineages (Zuckerkandl and Pauling, 1965). Therefore, using the strict molecular clock assumption that the rate of change is constant across all lineages, we can estimate the value of the constant r by dividing the estimated evolutionary distance by the time period over which it accumulated,

$$r = d/t.$$

The values of each t associated with the bifurcations in a tree can then be calculated, provided that we have some method of including external information about the time duration for the accumulation of any evolutionary distance represented in the tree. Information about r is therefore obtainable through calibration of the clock to absolute time by using paleontological evidence for the age of one or more of the divergence events in the tree. If we estimate the distance, measured in expected changes per site, leading to a divergence of known age we can then deduce the absolute rate, measured in expected changes per site per year, which can then be applied across the rest of the phylogeny to elucidate the ages of other divergences through the equation

$$t = d/r.$$

Palaeontological data allows for an external source of information regarding the timing of events within the tree that is to be dated. For a tree of *s* taxa, there are *s*-1 nodes for which palaeontological data is either able or unable to provide an absolute age. Therefore, *t* is a vector of *s*-1 node ages constructed from two further vectors, the calibrated node ages, t_c, and the uncalibrated node ages, t_{c-}, for which ages are to be estimated. As the rate of molecular change is dependent on the time over which evolutionary distances have accumulated, the uncalibrated node ages and the rate of change are treated as random variables and are estimated together, unless some prior information allows for the rate to be fixed to a certain value.

The application of external paleontological data in this manner unifies molecular clock methodology with more traditional stratigraphic methods for elucidating timescales. This approach provides a coherent framework in which divergence times can be objectively estimated using both extant molecular data and a selection of suitable fossil occurrence data.

Relaxing the Molecular Clock

Since the first proposal of the molecular clock hypothesis it has been widely recognised that assuming an invariable global substitution rate is often a gross oversimplification of the evolutionary process, and that in reality substitution rates will vary amongst distant lineages due to factors controlling the rate of mutation and substitution, such as variation in life history traits and DNA repair mechanism efficiency (Gu and Li, 1992, Ayala, 1997). Allowing for variable rates of evolutionary change across the tree potentially leads to more accurate divergence time estimates in trees with wide taxonomic scope where the factors constraining substitution rate are expected to vary wildly.

Different approaches to relaxing the clock are available, with the models that are most commonly applied broadly falling into two categories; auto-correlated clocks and uncorrelated clocks. Auto correlated clocks enforce a relatively smooth change of rates across lineages by drawing the rates on descendent lineages from a parametric distribution constrained by the rate on the ancestral lineage (Lepage et al., 2007, Kishino et al., 2001). This approach leads to a Brownian motion style diffusion of rates along branches. Uncorrelated clocks draw rates from a parametric distribution allowing the rates on each lineage to be independent of the rates on surrounding lineages, accommodating large changes in rate on proximal branches (Lepage et al., 2007, Drummond et al., 2006, Rannala and Yang, 2007).

The Evolution of Inference Frameworks for Divergence Time Estimation

The statistical frameworks in which divergence time estimation has been applied have evolved over time, with a succession of inference methods used to obtain evolutionary timescales when a stochastic model of character change is employed. Initially, maximum-likelihood was used to obtain estimates of divergence times and absolute evolutionary rates. In the maximum-likelihood estimation framework the probability of obtaining the sequence data, X, given some set of parameter values, Φ, is used as the optimality criterion. Therefore, the set of parameter values that maximise the likelihood-function, $P(X|\Phi)$, provide the best estimate of divergence times and evolutionary rates. Over time, further refinement of the divergence time model applied in the maximum-likelihood framework allowed for the inclusion of lineage specific rate variation methods that were similar to those discussed in the previous section, but properly accommodating the distribution of uncertainty of the fossil record in such analyses proved more difficult (Sanderson, 2002).

In the presence of some prior knowledge regarding the distribution of parameters in the phylogenetic model, a natural extension to maximum-likelihood estimation that can accommodate this information is Bayesian inference. In the case of divergence time estimation, the prior information regarding the distribution of some parameters concerns the ages of fossil taxa that can be unequivocally assigned to divergence events in the tree. The ability to explicitly account for uncertainty in prior information in this manner for a range of model parameters has resulted in Bayesian phylogenetic inference becoming a widely used method.

The widespread use of Bayesian inference has allowed for the application of more intricate evolutionary models and has become particularly beneficial for accurately incorporating palaeontological data when calibrating the clock model. The Bayesian phylogenetic framework involves obtaining a sample of the joint posterior distribution of the model parameters. For a node calibrated divergence time analysis, we are primarily concerned with obtaining the marginal distributions of uncalibrated node ages, t. The other parameters of this model are the topology τ, the clock rate μ, the parameters of the clock model ν, the parameters of the substitution model θ and the parameters of the tree prior T. Here, for the sake of the simplicity of notation these parameters are presented as a parameter vector $\Phi = [t, \tau, \mu, \nu, \theta, \mathrm{T}]$. The distribution of these parameters given the sequence data, X, can be obtained by applying Bayes theorem:

$$P(\Phi|X) = \frac{P(X|\Phi)\,P(\Phi)}{P(X)}.$$

The denominator on the right-hand side of this equation is the probability of observing the data given the model and is called the marginal likelihood. This value is non-trivial to calculate as it often involves solving a high dimensional integral (Lartillot and Philippe, 2006). The marginal likelihood is a normalising constant, therefore MCMC sampling can be used to obtain an accurate estimate of the posterior distribution of the parameters measured on a scale of relative probability. The likelihood of a tree and its set of divergence times is given by $P(\mathrm{X}|\Phi)$, and $P(\Phi)$ represents the joint prior distribution of the parameters to be estimated.

Bayesian Calibration of the Clock

Node Calibration

The Bayesian paradigm allowed for significant advances in the incorporation of uncertainty in the age of calibrated divergence events as parametric prior distributions (Yang and Rannala, 2006). In the divergence time estimation frameworks that preceded Bayesian implementations fossil calibrations were commonly applied as point estimates of the age of a known divergence event, often assigned to the age of the oldest unequivocal fossil species belonging to the subtending clade (Nei and Glazko, 2002, Hedges and Kumar, 2004). This method introduces considerable error, as the paleontological data supporting the age of any known divergence event will have a degree of associated uncertainty that will be poorly summarised as a single point in geological time. Furthermore, the assumption that the most recent common ancestor of a clade is of the same age as its oldest sampled fossil member is unreasonable as the incomplete fossil record is unlikely to preserve many older but valid clade members (Donoghue and Benton, 2007, dos Reis et al., 2016).

The Bayesian framework allows for the incorporation of the uncertainty associated with the age of a divergence event through the use of parametric probability distributions describing the *a priori* expectation of the distribution of this parameter (Yang and Rannala, 2006). The construction of such fossil calibrations usually involves identifying the oldest unequivocal member of the clade subtending the calibrated divergence event; the age of the geological unit this taxon was recovered from then provides the minimum age of the clade and the minimum bound of the parametric distribution describing the probability of the divergence time (Donoghue and Benton, 2007, Parham et al., 2012). The minimum bound of a calibration is relatively easy to construct as the calibrated divergence event can older than the oldest sampled unequivocal member of the subtending clade, therefore no positive probability should be assigned to ages younger than the age of this taxon and this bound is therefore usually "hard" (Donoghue and Benton, 2007, Parham et al., 2012). A method for objectively defining the maximum bound of the distribution is far less obvious and often relies on the absence of

evidence for older members of the calibrated clade in the fossil record. The difficulty in defining the maximum bound of calibrations often results in a "soft" distribution being employed in which there is a non-zero probability that the age of the divergence event is older than the empirical evidence for the maximum. Similarly, encapsulating *a priori* knowledge about the age of a clade through the parameters of a probability distribution is non-trivial, with the arbitrary distribution of probability between the minimum and maximum bounds directly influencing the posterior estimates of clade age (Parham et al., 2012, Warnock et al., 2015).

Another issue presented by node-calibration is the difference between the assigned prior probabilities on calibrated clades and the effective prior (Warnock et al., 2015, Heled and Drummond, 2012). In the node-calibration framework, the construction of the time prior involves combining the calibration information with the prior distribution of node ages defined by the tree prior. One of the initial implementations of Bayesian divergence time estimation on a fixed topology provides a simple example of why this is the case (Yang and Rannala, 2006). For an exclusively bifurcating fixed tree of s taxa provided for a divergence time analysis there are $s - 1$ nodes. These nodes can be split into three categories; the root t_r, the calibrated nodes t_c, for which fossil data can constrain the timing of divergence, and the uncalibrated nodes t_{c-} for which no chronological data is available but an age will be estimated. The distribution on the age of uncalibrated nodes is primarily constrained by the model of diversification applied to describe the prior probability of a tree; this is commonly a birth-death model and in this implementation this process is conditioned on the age of the root and number of taxa. The prior on the ages of all the nodes in the tree, apart from the root, is therefore the joint probability of these distributions

$$P(t) = P(t_c, t_{c-}).$$

To obtain this joint distribution, the density on node ages from the birth death process conditioned on the information for calibrated nodes, the root age, and number of taxa, is multiplied by the calibration information C for the calibrated nodes

$$P(t) = P_{BD}(t_{c-}|t_c, t_r, s)\, P(t_c|C).$$

The effect of multiplying these distributions together is that the effective prior on calibrated nodes will often be different to the assigned distribution $P(t_c|C)$. This results in the information provided by the fossil record often being incorrectly applied when estimating divergence times and therefore negatively influencing the accuracy of age estimates. Furthermore, there are

different methods by which the calibration information can be combined with the prior node age densities resulting from the underlying tree prior. The “calibration density” approach is a widely implemented time prior construction method that is required for computational purposes when topology and divergence times are co-estimated (Bouckaert et al., 2014, Heled and Drummond, 2012). In this approach, the calibration information is multiplied by the node age densities induced by the tree prior but without conditioning on the calibration information (Heled and Drummond, 2012). This approach is arguably improper, as it does not follow the rules of probability calculus (Heled and Drummond, 2012), and is therefore likely to lead to further discrepancies between the assigned node calibrations and the effective time prior.

Tip Calibration and Total Evidence Dating

Many recent developments in divergence time estimation methodology have been based around the application of a total-evidence approach (Pyron, 2011, Ronquist et al., 2012a). Total-evidence dating (TED) methods are a collection of approaches that deviate from the node calibration framework by allowing for divergence time estimation on non-ultrametric trees through the representation of both extant and extinct taxa in the tree. This is commonly achieved by analysing molecular data from extant taxa in tandem with morphological data from both extant and extinct taxa, although entirely morphological or entirely fossil analyses have also been performed. The molecular data is analysed with standard substitution models and morphological data is analysed with a simple probabilistic model of morphological change (Lewis, 2001). Importantly, the inclusion of fossil taxa allows for calibrations to be placed on fossil tips nested within the tree instead of on nodes.

With this modification of the calibration method the TED framework does not rely on the construction of accurate node calibrations, directly avoiding many of the issues presented by the node-calibration framework. For node-calibration methods the available paleontological data is reduced down to the age of the unequivocal oldest members of calibrated clades only; with TED any fossil with categorical morphological data can be included in analyses, allowing the full range of paleontological data to refine estimated timescales. The process of calibration construction is relatively straightforward in this framework and consists of simply defining the bounds on the age of each individual fossil taxon for which morphological data is available (O'Reilly et al., 2015). This method does not rely on absence of evidence for constructing maxima, and most calibrations can be safely applied with hard bounds as uniform distributions. As there are no arbitrary distributions placed on node ages the issues of combining calibration information with the tree prior that are encountered with node-calibration can also be avoided if a tree prior that accounts for sampling through time is used (Gavryushkina et al., 2014).

Furthermore, the requirement of the arguably improper "calibration density" construction technique allows for a more justified co-estimation of time and topology in the TED framework (Heled and Drummond, 2012).

The first TED analyses were reliant on tree priors that were often arguably inappropriate for non-ultrametric trees, as they did not explicitly account for the sampling of extinct taxa (Pyron, 2011, Ronquist et al., 2012a). The fossilised birth death (FBD) process promises to circumvent this issue by explicitly modelling the sampling of fossil taxa in the tree prior and inducing a more appropriate prior density on node ages (Gavryushkina et al., 2014, Zhang et al., 2016). The development of the FBD tree prior demonstrates that TED methods present a coherent framework for divergence time estimation, but that this framework is still in its infancy and that further improvements to this method are required to obtain the most accurate estimates of clade ages.

Difficulties Encountered When Integrating Molecular and Morphological Data

Despite the potential benefits of analysing palaeontological data in combination with molecular data in the TED framework, there are several issues with current TED implementations that may limit the accuracy of estimated timescales. Many of these issues are caused by the simplicity of the model of morphological evolution or by the application of assumptions that are justifiable for molecular clock analyses but that are questionable when analysing morphological data. For molecular data, an assumption is often made to simplify the calculation of likelihood that each site in an alignment evolves along the same underlying phylogeny but independent of all other characters (Felsenstein, 1981). With this assumption each site is considered to be an independent observation from the same underlying distribution (IID). The assumption of independence is perhaps debatable for molecular data as secondary and tertiary structure is likely to constrain the probability of substitutions at different sites, as are shared selective pressures. For morphological data the IID assumption is likely to be far less appropriate, particularly when character contingencies are present in datasets. A contingent character is one that is only applicable to a taxon if a particular state is present at another character, therefore, contingent characters must co-vary by definition. The presence of contingencies in morphological datasets presents an obvious issue for estimating the evolutionary distance between taxa in a probabilistic phylogenetic framework, and as such, contingencies are likely to influence estimates of clade age.

With the use of morphological data to refine divergence time estimates a relaxed "morphological clock" must be assumed, in addition to the relaxed molecular clock. While the

general molecular clock principle has been empirically demonstrated, a linear relationship between morphological change and divergence time is not immediately obvious (dos Reis et al., 2016). Furthermore, the way in which the relationship between the rate of molecular and morphological evolution should be modelled is not clear. There are two possible approaches to handling the relative molecular and morphological rates; separate clocks for molecular and morphological partitions, or a single underlying clock with rate multipliers describing the relative rate between partitions (Pyron, 2011). If the former approach is taken then molecular and morphological rate are unlinked on each lineage. If the latter approach is taken then molecular and morphological rate are contingent on one another on each lineage. This is achieved through the estimation of rate multiplier parameters that describe the relative rate difference between partitions, effectively meaning that molecular and morphological evolution co-vary. Therefore, on a given branch a relatively high rate in one partition must result in a relatively high rate in the other partition (Nylander et al., 2004). There is currently no consensus on how molecular and morphological rates co-vary (Omland, 1997, Bromham et al., 2002), and what the effects of different approaches to modelling this relationship are in this framework.

The construction of the prior density on divergence times has certainly improved in the TED framework, but the distribution of the effective prior may potentially allow for age estimates that are incongruent with the fossil record (O'Reilly and Donoghue, 2016). The effective prior distribution on divergence times in the TED framework is constrained entirely by the fossil tip calibrations and the tree prior. There are no arbitrary distributions on clade ages constraining this distribution to be congruent with fossil evidence for clade age. This is potentially problematic as the effective prior on node ages may conflict with empirical fossil occurrence evidence for the minimum constraints on clade ages. How often this phenomenon occurs is unknown as the effective time prior is difficult to obtain when the topology is estimated alongside the divergence times. Therefore, it is possible that divergence times estimated in the TED framework may violate empirical fossil evidence due to inappropriate prior densities on node ages.

Qualities of morphological data that are not present in molecular data may present issues for TED analyses. Missing data in morphological matrices is likely to be systematically distributed due to fossilisation biases, with certain character types being more likely to be degraded or absent (Sansom and Wills, 2013, Sansom et al., 2010)}. A systematic distribution of missing data may well influence estimates of divergence time as it has been shown that taphonomic bias can affect the accuracy of topology estimates (Sansom and Wills, 2013). The taphonomic process is a multifaceted one, and each stage is likely to introduce a characteristic distribution of missing data and a concomitant effect on the estimation of divergence times. Whether the

systematic distribution of missing data in fossil morphological matrices has an influence on age estimate accuracy will therefore require investigation. If it can be demonstrated that missing fossil morphological data do have an influence on age estimates, then characterising the relative influence of different fossilisation processes may provide an insight into which types of character are problematic for divergence time analyses.

Overview of Chapters

Over the course of this thesis I investigate the influence that several outstanding issues exert on the accuracy of phylogenetic relationships and divergence times estimated when morphological data is analysed in a probabilistic framework. In a number of cases methodological solutions are presented that may be implemented to improve the accuracy of divergence time estimates.

Chapter One

In the first chapter I address a fundamental question underlying many of the more sophisticated combined data source techniques such as TED; the suitability of a probabilistic framework for analysing categorical morphological data. TED methods often co-estimate topology and divergence times, and as such it is important to identify the ability of the commonly applied model of morphological evolution to accurately estimate phylogenetic relationships. I investigate the relative efficacy of competing methods of phylogenetic reconstruction, and test whether a probabilistic framework outperforms parsimony in terms of the accuracy of reconstructed trees. I utilise a simulation framework in which morphological data is simulated along a known topology and then analysed by competing reconstruction methods. To ensure that the analysed data reflects empirical data, the simulation framework is constructed such that the resulting data matches a range of homoplasy seen in empirical morphological matrices.

Chapter Two

In Chapter 2 I build on both the framework and findings of the first chapter to investigate the efficacy of competing phylogenetic reconstruction methods for morphological data when the true topology is either extremely symmetric or extremely asymmetric. Trees containing fossil taxa are often strongly asymmetric, and therefore the ability of different inference frameworks to accurately reconstruct such topologies is a pertinent question given the common use of co-estimation of topology and divergence time in TED methods. The simulation framework from Chapter 1 is altered so that simulated data matches an empirical distribution of homoplasy, as opposed to simply the range. I also use empirical morphological matrices to investigate whether

a number of published evolutionary hypotheses that are supported by parsimony analyses are also supported by probabilistic methods.

Chapter Three

Chapter 3 is presented as a thorough review of TED methodology and the expected benefits afforded by the integration of morphological data in such analyses. I also highlight several issues presented when attempting to analyse fossil morphology alongside molecular data, many of which are addressed in later chapters. In this chapter I also consider the correct manner in which calibrations should be constructed for TED analyses. Many of the initial TED analyses assumed that fossil taxon age was known without error; I demonstrate the correct method to objectively construct fossil tip calibrations and the effect that including this information has on divergence time estimates.

Chapter Four

A common observation from TED analyses is unexpectedly ancient divergence time estimates (Ronquist et al., 2016). This observation is potentially caused, at least partially, by the effective prior on divergence times placing exaggerated probability on ancient ages. In this Chapter I demonstrate a method of combined tip and node calibration that can constrain evolutionary timescales to be congruent with fossil occurrence data. Using an approximation of the effective prior on node ages, I investigate whether the effective tree prior in a tip-calibrated analysis may potentially place non-zero probability on clade ages that violate the fossil record. Through a combination of tip and node calibrations with the uniform tree prior I describe a method with which ages that are wholly compatible with fossil evidence can be estimated.

Chapter Five

The distribution of missing data in fossil morphological matrices is likely to be primarily constrained by biases in the fossilisation process. A systematic distribution of missing data may potentially lead to systematic biases in divergence time estimates in a tip-calibrated framework, and as such the effect of fossilisation biases on divergence times requires characterisation. In Chapter Five I investigate the relative influence of different stages of the fossilisation on the accuracy of divergence times. This is achieved through the use of a nearly complete Mammalian morphological data set and an empirically guided fossilisation simulation method.

Chapter Six

Summarising a posterior sample of trees presents a challenge for analyses using morphological data. This is due to the high level of variance expected in the posterior distribution due to a relatively small amount of phylogenetic information. A diffuse posterior sample will contain

few clades with high support and many clades with low support, meaning that summary methods that take a single sampled topology as the consensus may not provide a good representation of the posterior distribution of trees. The popular maximum clade credibility (MCC) consensus method takes such an approach and results in a fully resolved consensus topology (Heled and Bouckaert, 2013). In Chapter Six I explain how a diffuse posterior distribution presents issues for the accuracy of both topology and divergence time estimates presented on MCC trees. I then demonstrate these issues through the analysis of simulated datasets containing increasing quantities of phylogenetic information. I also demonstrate how the inclusion of empirical fossil morphological data influences the variance of a posterior sample of trees, and explain the effect that this has on the accuracy of MCC consensus trees constructed when empirical data is analysed.

Chapter 1

Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data

Joseph E. O'Reilly*, Mark N. Puttick*, Luke Parry, Alastair R. Tanner, James E. Tarver, James Fleming, Davide Pisani, Philip C. J. Donoghue

*These authors contributed equally to this study

This Chapter was published in *Biology Letters* on April 19th 2016
DOI: 10.1098/rsbl.2016.0081

1.1 - ABSTRACT

Different analytical methods can yield competing interpretations of evolutionary history and, currently, there is no definitive method for phylogenetic reconstruction using morphological data. Parsimony has been the primary method for analysing morphological data, but there has been a resurgence of interest in the likelihood-based Mk-model. Here we test the performance of the Bayesian implementation of the Mk-model relative to both equal and implied-weight implementations of parsimony. Using simulated morphological data, we demonstrate that the Mk-model outperforms equal-weights parsimony in terms of topological accuracy, and implied-weights performs the most poorly. However, the Mk-model produces phylogenies that have less resolution than parsimony methods. This difference in the accuracy and precision of parsimony and likelihood approaches to topology estimation needs to be considered when selecting a method for phylogeny reconstruction.

1.2 - Introduction

Morphology once provided the only means of inferring evolutionary trees, but it was effectively rendered obsolete by molecular sequence data and the development of sophisticated molecular evolutionary models for phylogenetic analysis (Scotland et al., 2003). However, with the recognition that fossil species are integral to correctly inferring patterns of character evolution, changes in diversity, as well as in establishing evolutionary timescales, morphological data is enjoying a phylogenetic renaissance (Lee and Palci, 2015), allowing fossil species to be assigned to their correct branches in the Tree of Life. Methods for phylogenetic analysis of morphological data remain underdeveloped and though likelihood models are available that may more accurately accommodate the vagaries of morphological datasets (Lewis, 2001), including high rates of heterogeneity and a preponderance of missing data (Wagner, 2012), parsimony remains the method of choice, principally perhaps as a consequence of tradition. Indeed, a recent simulation-based study by Wright and Hillis (2014) demonstrated that a Bayesian implementation of the Mk-model (Lewis, 2001) strongly outperforms parsimony, especially when rates of character change are high, or when relatively few characters are analysed. The conclusions drawn by Wright and Hillis (2014) were based on data effectively simulated using the Mk-model, potentially biasing the test in favour of the Mk-model. Furthermore, they did not consider whether the simulated data exhibited realistic levels of homoplasy, analysed unrealistically large simulated datasets, and evaluated only the relative performance of equal-weights parsimony when morphological data are now commonly analysed under implied-weights parsimony (Goloboff et al., 2008a).

In an attempt to evaluate the relative performance of likelihood and parsimony methods for the phylogenetic analysis of discrete character morphological data, we simulated datasets of 100, 350 and 1000 discrete morphological characters using a modified HKY85 model, discriminating datasets that failed to meet expected levels of homoplasy. We evaluated the relative performance of equal-weights parsimony, implied-weights parsimony, and model-based methods of phylogenetic analysis in terms of their ability to recover the tree used to simulate the data. We found that the Mk-model performs best in the analysis of all simulated datasets, largely because the Bayesian consensus trees are poorly resolved. Equal-weights parsimony exhibits lower levels of accuracy but this is combined with higher resolution. Implied-weights parsimony performed most poorly of all the methods considered.

1.3 - Materials and Methods

To simulate binary morphological data we used the HKY+$\Gamma_{continuous}$ model to generate nucleotide data with π = [A, C, G, T] = [0.2, 0.3, 0.2, 0.3] which we translated into purines (0)

and pyrimidines (1) – R/Y coding, resulting in $\pi = [R, Y] = [0.4, 0.6]$. Due to the symmetry of unordered binary characters, it is possible to recode each character such that the global distribution of character states matches the expected [0.5, 0.5] distribution of the Mk model, potentially improving the fit of the model to empirical data. This is not done in practice, so here we keep the stationary frequencies as that which they are simulated as. The recoded HKY-model possesses an uneven equilibrium distribution of state frequencies, resulting in structurally realistic morphological matrices while facilitating violation of assumptions of the Mk-model; thus, our data is not biased in favour of either method of phylogenetic inference. Initial tests were performed to determine values for the model parameters that produce binary data with empirically observed levels of homoplasy (Sanderson and Donoghue, 1989). Following (Wright and Hillis, 2014), data was simulated using the lissamphibian tree presented in (Pyron, 2011), yielding datasets of 100, 350 and 1000 characters; most real morphological datasets contain in the order of 100 characters, but we included 350 and 1000 character matrices to investigate the effect of scaling and for ease of comparison toWright and Hillis (2014). 100 unique underlying substitution rates were drawn from a U(0.1,10) distribution, and the branch lengths of the original lissamphibian tree were multiplied by these underlying rates, resulting in rates spanning two orders of magnitude. For each substitution rate 10 unique matrices were produced, modelling among-character rate heterogeneity as gamma distributed uniquely within each matrix.

Matrices were analysed with the Mk+Γ model using default priors in MrBayes 3.2 (Ronquist et al., 2012b), and both standard and implied-weights parsimony in TNT (Goloboff, 2000). The Mk-model is more suitable for our simulated data than the Mkv-model as we did not strip invariant sites from the final matrices. Majority-rule consensus trees were produced for each method. For implied-weights parsimony, we used a range of K values: 2, 3, 5, 10, 20 and 200. As the underlying substitution rate is varied, the per-matrix level of homoplasy may violate the empirically observed range; to produce the most empirically justified morphological matrices we implemented an empirically derived minimum consistency index (CI) cut-off of 0.26 (Sanderson and Donoghue, 1989) for each simulated dataset and repeated analyses for these treated matrices (Figure S1). This cut-off reduced the size of the datasets to 128 (100 characters), 149 (350 characters), and 126 (1000 characters) matrices. In depth description of the initial parameter value tests and further details of matrix generation are presented in the Supplementary Information.

The accuracy of topologies estimated by the different reconstruction techniques was assessed using the Robinson-Foulds distance (Robinson and Foulds, 1981) from the generator tree. We

also explored the relationship between resolution of output trees, measured by the number of nodes per tree.

1.4 - Results

The Mk-model achieved the highest levels of accuracy across all datasets. Median Robinson-Foulds distances were lower for the Mk-model compared to both equal-weights and implied-weights parsimony (table 1.1; figure 1.1), and for all approaches, accuracy of topology reconstruction increases with increasing dataset size. Furthermore, equal-weights parsimony out-performs implied-weights parsimony for all datasets and values of K, but this is less pronounced for the 1000 character dataset (table 1.1). For convenience, all further results for implied weights are for K = 2.

The same relative performance of the phylogenetic reconstruction methods is seen when considering only those datasets exhibiting realistic levels of homoplasy. The median Robinson-Foulds distance for the Mk-model is still lowest for each dataset, but the median and range of Robinson-Foulds distances for equal and implied-weights parsimony are closer to the distribution seen from the Mk-model (table 1.1; figure 1.1). Additionally, for a given dataset, there is a similar Robinson-Foulds distance regardless of the reconstruction method employed (supplementary figure S2). Unless otherwise stated, all subsequent results are from the subset of datasets exhibiting realistic levels of homoplasy.

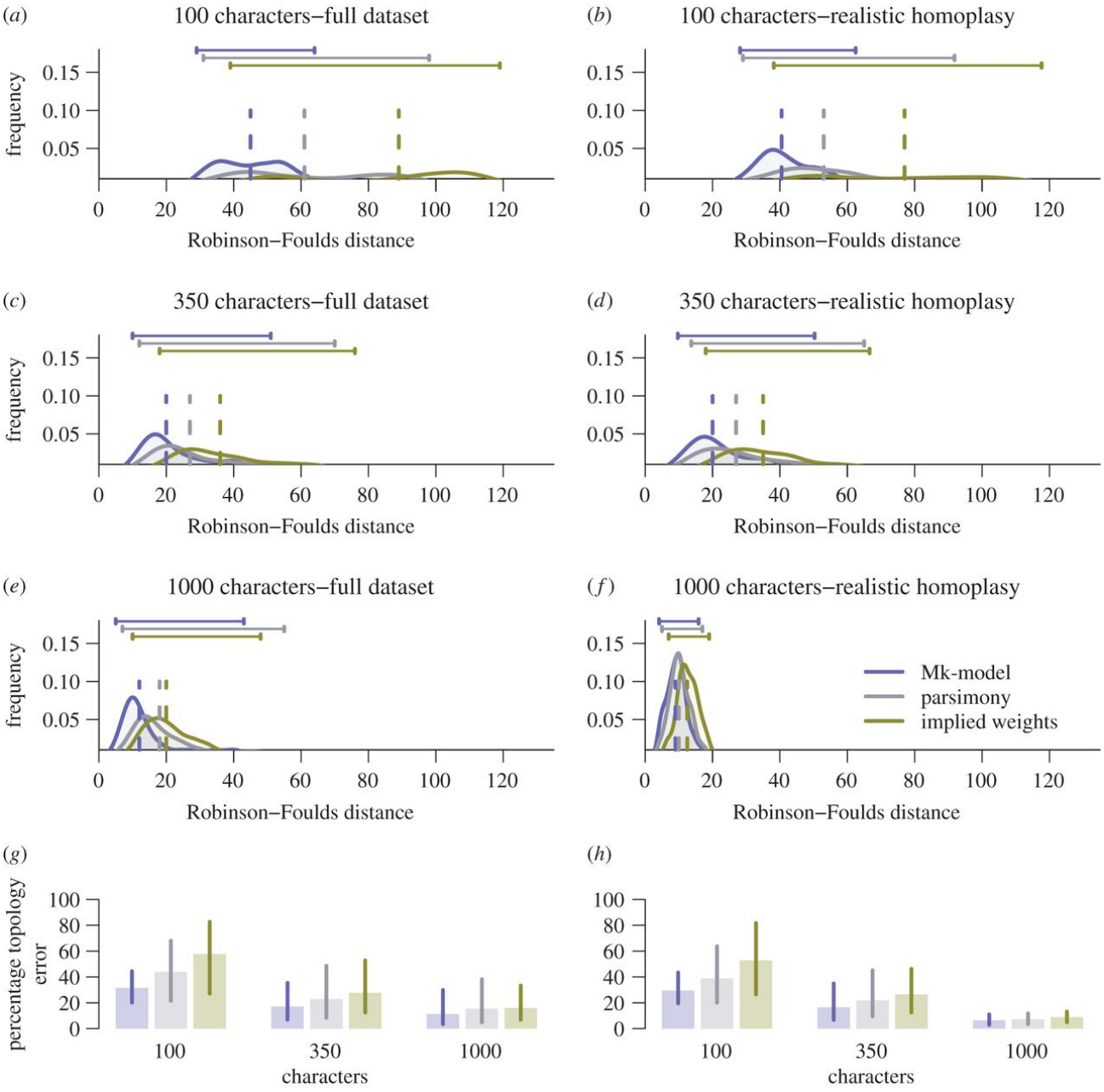


Figure 1.1 - Mk tree reconstructions (blue) outperform equal-weights parsimony (grey) and implied-weights parsimony (green) for 100, 350, and 1000 characters (a, c, e, g), and these differences remain with the subset of data with realistic homoplasy (b, d, f, h). Bars above the plots mark the 95th percentile range for each method, and dotted vertical lines show the median values. Percentage topology error (g, h) is the Robinson-Foulds value of the reconstructed tree compared to the worst possible value, as shown in (Wright and Hillis, 2014).

	100 characters	100 characters CI	350 characters	350 characters CI	1000 characters	1000 characters CI
mk	45 (29–64)	40.5 (28.2–62.5)	20 (10–51)	19.5 (10.2–57.3)	19.5 (10.2–57.3)	11 (5–27.8)
ew	61 (31–98)	53 (29–91.8)	27 (12–70)	28 (12–74.8)	28 (12–74.8)	16 (6.2–43.7)
iw k2	89 (39–119)	77 (38.2–117.7)	36 (18–76)	36 (17.2–81.3)	36 (17.2–81.3)	19.5 (10–35.7)
iw k3	76 (38–112)	69 (36.4–108)	32 (16–69)	34 (15.2–70)	34 (15.2–70)	18 (9.2–35.7)
iw k5	68 (36–104)	61 (32.2–102)	30 (14–66)	31.5 (15.2–68)	31.5 (15.2–68)	18 (9–34)
iw k10	63 (34–100)	55.5 (32–98)	28 (13–68)	30 (15.2–69.7)	30 (15.2–69.7)	16 (8–34)
iw k20	64 (34–100)	53 (33–97.8)	28 (14–68)	30 (13.2–71.7)	30 (13.2–71.7)	17 (8–39.3)
iw k200	65 (34–100)	55 (32.2–97.7)	28 (14–72)	30.5 (15–76)	30.5 (15–76)	18 (8–44)

Table 1.1: The differences in median and the 95th percentile range of Robinson–Foulds values between the Mk and both parsimony models are greater in the full dataset compared with the realistic homoplasy subsets. mk, Bayesian Mk model; ew, equal-weights parsimony; iw, implied weights parsimony and its attendant K values.

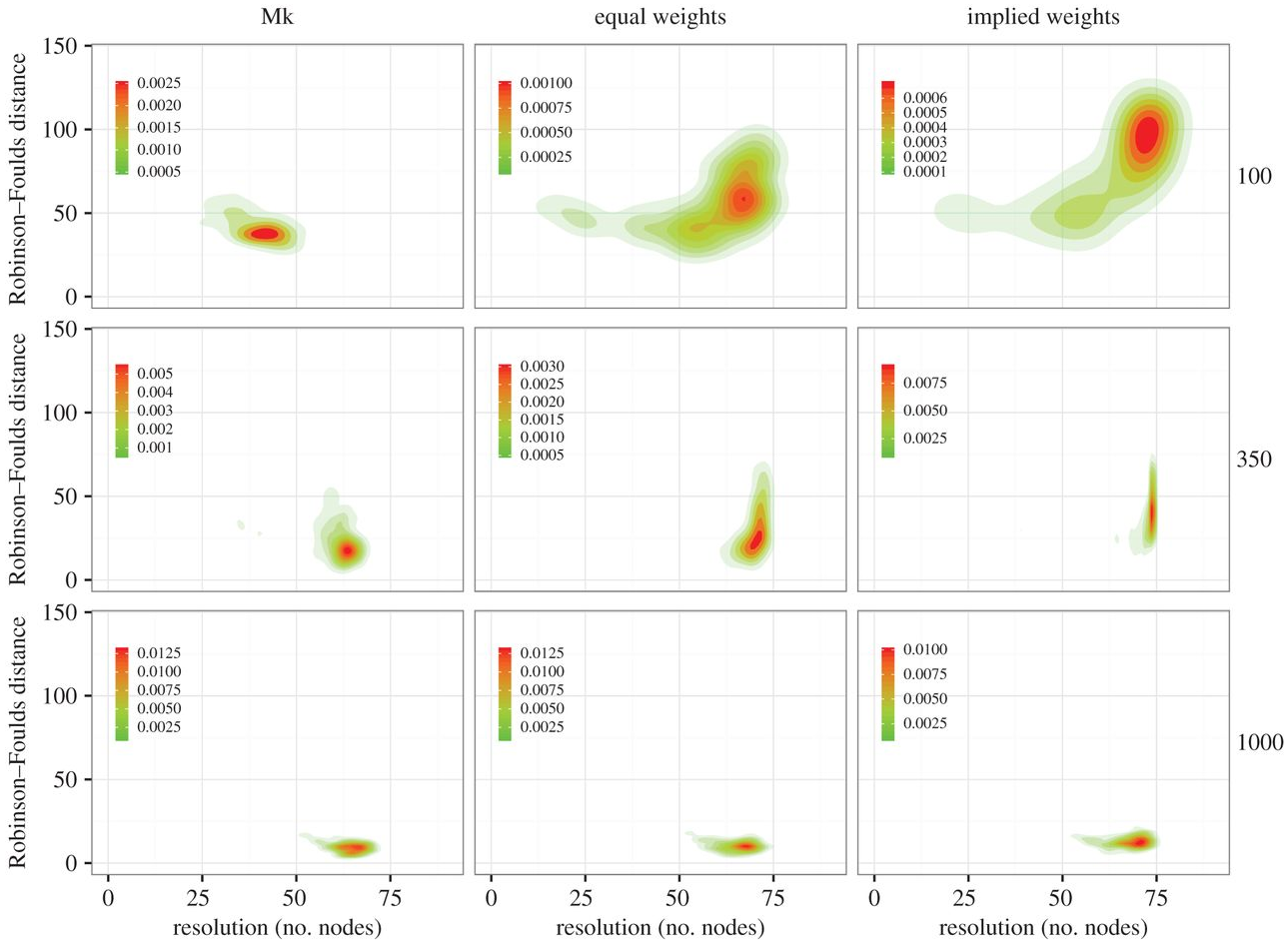


Figure 1.2 - The Mk model exhibits higher accuracy with lower precision than parsimony methods; these results are less clear as more characters are added. Contour plots of Robinson-Foulds distances plotted against the number of resolved nodes in each tree; the z-axis represents the density of the distribution of trees.

The higher accuracy (lower Robinson-Foulds values) of the Mk-model against other methods for 100 and 350 characters is due to trees being less resolved (figure 1.2). The density of Robinson-Foulds distance is lower for the Mk compared to equal weights, which itself is lower than implied weights, but both equal and implied weights achieve higher levels of precision (number of nodes reconstructed). These differences are negligible in the 1000 character datasets (figure 1.2).

There is a significant overlap in the set of nodes correctly recovered across methods, when mapped against the reference phylogeny (Figure 1.2, S3). In particular, for all methods there is a trend for nodes closer to the root to be more accurately estimated in small datasets, but this relationship decreases as the number of characters increases (Supplementary table S2, figures 1.2, S4, S5). The percentage of times a node from the reference tree was accurately reconstructed showed a strong correlation for 100 and 350 characters, but decreases with 1000 characters (table 1.2).

1.5 - Discussion

Only minor differences are seen in the accuracy of phylogenetic topology reconstruction between the Bayesian implementation of the Mk-model and parsimony methods. Our findings both support and contradict elements of the results of Wright and Hillis (2014) in that we can corroborate their observation, that the Mk-model outperforms equal-weights parsimony in accuracy, but the Mk-model achieves this at the expense of precision. Unexpectedly, implied-weights parsimony is less effective than either equal-weights parsimony or the Mk-model, in datasets with small numbers of characters. Implied-weights parsimony outperforms equal-weights parsimony only in the analyses of unrealistically large datasets. These results challenge the increasingly common view that implied-weighting better accommodates homoplasy than does equal-weights parsimony (Goloboff et al., 2008a), and this result is true for a range of K values (table 1.1).

In comparison to the other approaches, equal-weights parsimony analyses of the datasets exhibiting realistic levels of homoplasy and large number of characters, yield a set of trees with a longer tailed distribution of Robinson-Foulds distances. In large part, this reflects estimation of a small quantity of trees markedly different from the generating tree (figure 1.1). Inaccuracy in topological estimation is more prevalent towards the tips in all analyses, with the inclusion of more characters reducing the intensity of this phenomenon. For this effect to be completely removed it would require the analysis of well over 1000 empirically justifiable characters, a number that is rarely achieved for morphological data sets. The accuracy of node reconstruction is correlated significantly between all three techniques, demonstrating that most nodes in the tree that were difficult to resolve for one method were difficult to resolve for all. This phenomenon is observed across all character quantities, and suggests a general difficulty in accurately estimating topology given the same data.

Our results can be interpreted to advocate use of the Mk-model over parsimony methods in the analysis of discrete morphological data. Parsimony methods produce precision without the

accuracy achieved by the Mk-model and precision without accuracy is a poor basis for any science. We anticipate that the implementation of the Mk-model within a maximum-likelihood framework will exhibit levels of accuracy and precision more comparable to the parsimony methods, simply because it estimates a single, fully-resolved topology. Integration over parameters while producing an acceptable level of accuracy is a quality of Bayesian inference, and our Mk-model results are probably dependent on a Bayesian implementation. While comparative phylogenetic methods often require fully resolved trees, these may be accommodated through analyses utilising the posterior sample of trees estimated using the Mk-model. Therefore, the prior requirement of a fully-resolved tree need not unnecessarily lead to a preference for parsimony over the Mk-model.

In comparison to parsimony methods, the Mk-model has undergone little development since its conception (Klopfstein et al., 2015, Wright et al., 2016), while attempts to improve the performance of parsimony methods, like implied-weights parsimony (Lewis, 2001), have not led to increased accuracy (table 1.1). Thus, model-based phylogenetics can be expected to offer more opportunity for development, e.g. through relaxing the assumption of symmetrically distributed stationary distribution of character states (Klopfstein et al., 2015, Wright et al., 2016) and improvement in the accuracy of phylogeny estimation from discrete character data. We suggest, however, that more focus should be invested in assessing whether the data are sufficiently informative to discriminate between competing phylogenetic hypotheses.

1.6 - Conclusions

Phylogenies produced using likelihood models are more accurate than parsimony approaches, but have lower precision. Likelihood models offer greater scope for development in attempting to achieve greater accuracy but, in the interim, we suggest that phylogeneticists should consider the aims of their analyses when choosing the appropriate method.

Chapter 2

Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data

Mark N. Puttick*, J. E. O'Reilly*, Alastair R. Tanner, James F. Fleming, James Clark, Lucy Holloway, Jesus Lozano-Fernandez, Luke A. Parry, James E. Tarver, Davide Pisani, Philip C.J. Donoghue

*These authors contributed equally to this study

This Chapter was published in *Proceedings Of The Royal Society B* on 11th January 2017
DOI: 10.1098/rspb.2016.2290

2.1 - Abstract

Morphological data provides the only means of classifying the majority of life's history, but the choice between competing phylogenetic methods for the analysis of morphology is unclear. Traditionally, parsimony methods have been favoured but recent studies have shown that these approaches are less accurate than the Bayesian implementation of the Mk model. Here we expand on these findings in several ways: we assess the impact of tree shape and maximum-likelihood estimation using the Mk model, as well as analysing data composed of both binary and multistate characters. We find that all methods struggle to correctly resolve deep clades within asymmetric trees, and when analysing small character matrices. The Bayesian Mk model is the most accurate method for estimating topology, but with lower resolution than other methods. Equal weights parsimony is more accurate than implied weights parsimony, and Maximum Likelihood estimation using the Mk model is the least accurate method. We conclude that the Bayesian implementation of the Mk model should be the default method for phylogenetic estimation from phenotype datasets, and we explore the implications of our simulations in reanalysing several empirical morphological character matrices. A consequence of our finding is that high levels of resolution or the ability to classify species or groups with much confidence should not be expected when using small datasets. It is now necessary to depart from the traditional parsimony paradigms of constructing character matrices, towards datasets constructed explicitly for Bayesian methods.

2.2 - Introduction

The fossil record affords the only direct insight into evolutionary history of life on Earth, but the incomplete preservation and temporal distribution of fossils has long prompted biologists to seek alternative perspectives, such as molecular phylogenies of living species, eschewing palaeontological evidence altogether (Harvey et al., 1994). However, there is increasing acceptance that analyses of historical diversity cannot be made without phylogenies that incorporate fossil species (Rabosky, 2010, Losos et al., 2013)and calibrating molecular phylogenies to time cannot be achieved effectively without recourse to the fossil record (dos Reis et al., 2016). Integrating fossil and living species has become the grand challenge and there has been a modest proliferation of phylogenetic approaches to the analysis of phenotypic data. While conventional parsimony remains the most widely employed method, alternative parsimony (Goloboff et al., 2008b) and probabilistic (Lewis, 2001) models have been developed to better accommodate heterogeneity in the rate of evolution among characters and across phylogeny. Unfortunately, these competing methods invariably yield disparate phylogenetic hypotheses among which it is difficult to discriminate since the true tree is never known for empirical data.

A number of studies have attempted to establish the efficacy of competing phylogenetic methods using data simulated from known trees (Wright and Hillis, 2014, O'Reilly et al., 2016, Congreve and Lamsdell, 2016), finding that the probabilistic Mkv model outperforms parsimony methods, among which, conventional equal-weights parsimony performs best. However, these studies were potentially biased by their experimental design: (i) two of the studies employed a generating tree that was unresolved and, therefore, biased against parsimony methods which recover resolved trees; (ii) these studies did not discriminate between the impact of the probabilistic model and its implementation in a Bayesian framework; (iii) based on single empirical trees, the impact of tree symmetry, which is known to confound phylogeny estimation (Holton et al., 2014), was not explored; (iv) only binary characters were considered, whereas empirical datasets are commonly a mixture of binary and multistate characters. Therefore, we compare the performance of equal-weights parsimony (EW-Parsimony), implied-weights parsimony (IW-Parsimony), Maximum Likelihood and Bayesian implementations of the Mk model, based on datasets with different numbers of characters, comprised of binary and multistate characters and simulated on a fully balanced and a maximally-imbalanced phylogenetic tree. We find that Bayesian inference out-performs all other methods, while EW-Parsimony performs better than IW-Parsimony, and Maximum Likelihood performs worst of all. We apply these competing phylogenetic methods to empirical morphological datasets of

similar sizes to our simulated datasets and explore the efficacy of the ensuing phylogenetic hypotheses in light of the conclusions derived from of our simulation-based study.

2.3 - Materials and Methods

2.3.1 - Simulation of Morphological Matrices

We simulated data on two ultrametric 32-taxon generating trees at the extremes of tree symmetry: one fully asymmetrical and one fully symmetrical (see Supplementary Fig S2.1). The root height was set to 1, for the symmetric tree this results in all branch lengths being equal. For each tree we simulated matrices of three sizes: 100, 350 and 1000 characters. We generated matrices using the HKY + Γ Continuous model of molecular substitution, with $\kappa = 2$, the shape (set equal to rate) of the gamma distribution and underlying substitution rate for each replicate sampled from independent and identically distributed exponential distributions with a mean of one, and character state stationary frequencies fixed as $\pi = [A, C, G, T] = [0.2, 0.3, 0.2, 0.3]$. We used a fixed and uneven stationary distribution of nucleotide frequencies to ensure our simulation model did not collapse into the Mk model, as this would bias the analysis in favour of Mk model-based approaches. We simulated 1000 replicate matrices with unique substitution parameters for each tree and each character number, resulting in a total of 6000 matrices. We set two types of character within each matrix, binary and multistate, and we simulated a proportion of 55 binary:45 multistate characters, based on the mean ratio found in a survey of empirical morphological data matrices (Guillerme and Cooper, 2016). We established binary characters by converting data simulated under the HKY model to R/Y coding (i.e. 0/1), this results in the elimination of the transition transversion ratio and a new stationary distribution of state frequencies, $\pi = [R, Y] = [0.4, 0.6]$: morphological multistate characters were simulated by converting DNA bases to integers.

To ensure that our simulated data are realistic, we generated each set of 1000 unique replicate matrices such that the among-matrix distribution of homoplasy approximated the distribution of empirical homoplasy, characterised by the Consistency Index (CI), reported by Sanderson (1996). To approximate this distribution of homoplasy we placed the Sanderson and Donoghue data into quantised bins of CI spanning 0.05, between the empirical bounds of 0.26 and 1.0, and simulated matrices until we matched this expected density per bin (Supplementary Figure S2).

The code used to simulate these data is available in Electronic Supplementary Materials.

2.3.2 - Phylogenetic analysis

We analysed the simulated matrices with EW-Parsimony, IW-Parsimony (k=2), and the Mk model (Lewis, 2001) under both Maximum Likelihood and Bayesian implementations. EW-Parsimony and IW-Parsimony estimation of topology was performed in TNT (Goloboff et al., 2008b). We used the Mk plus gamma model for Maximum Likelihood estimation of topology in RAxML 7.2 (Stamatakis, 2014), and Bayesian estimation of topology in MrBayes 3.2 (Ronquist et al., 2012b). As the approximate likelihood calculation of RAxML may be distant from the true likelihood (Wright et al., 2015), we conducted a sensitivity test by re-analysing a subset of our data with the likelihood implementation of the MK model in IQ-tree (Nguyen et al., 2015); both methods gave effectively identical results, indicating results from the likelihood Mkv model are not software-specific.

The Mkv model is inappropriate due to the lack of acquisition bias in the simulated data. For both Maximum Likelihood and Bayesian analyses we applied the discretised gamma distribution model to account for between-character rate heterogeneity. For Bayesian analyses the posterior distribution was sampled one million times by four chains using the Metropolis-coupled Markov-chain Monte Carlo algorithm, two independent runs were performed for each replicate and the two resulting posterior samples were combined after qualitative assessment of convergence. For parity, we characterised the result of all phylogenetic methods as the majority-rule consensus of resultant tree samples. We did not employ bootstrap methods to measure support for parsimony and likelihood analyses because phenotypic data does not meet the assumption that phylogenetic signal is distributed randomly among characters.

We used the Robinson-Foulds metric (Robinson and Foulds, 1981) to compare the similarity of estimated topologies against their respective generating tree. We also noted the per-node resolution, and the variation of node accuracy across the topology.

2.3.3 - Empirical analyses

We analysed four published palaeontological phenotype character matrices that encompass a range of character numbers and a diverse sample of taxa from the Tree of Life (Hilton and Bateman, 2006, Luo et al., 2015, Nesbitt et al., 2013, Sutton et al., 2012). We resolved any ambiguities in character coding to their most derived state for each matrix to make analyses compatible across the different phylogenetic methods, facilitating comparison of results. We analysed each matrix by applying the same settings used to analyse our simulated matrices: EW-Parsimony, IW-Parsimony, as well as Bayesian and Maximum Likelihood implementations of the Mk model. Empirical morphological matrices are rarely constructed to contain invariant or parsimony uninformative characters. Therefore, the Mkv extension of the Mk model, which

uses conditional likelihood to correct for such acquisition biases, is more appropriate than the Mk model for analysis of these empirical data matrices(Lewis, 2001).

2.4 - Results

2.4.1 - Simulated data

Accuracy is higher for trees inferred from data simulated on a symmetrical topology compared to trees estimated from data simulated on the asymmetrical topology (compare figures 2.2). Bayesian consensus phylogenies are generally the least well resolved (Figure 1). All methods estimated topologies with greater accuracy as the number of analysed characters increased (figs 2.2; supplementary table S2.5-S2.7). All methods, apart from Maximum Likelihood, produced phylogenies with greater resolution with higher numbers of characters (figure 2.1).

For all implementations and dataset sizes, the Bayesian implementation of the Mk model achieves higher accuracy compared to other methods (table 2.1; figs 2.1-2.2). The two parsimony methods achieved the next highest levels of accuracy, EW-Parsimony achieving greater accuracy than IW-Parsimony. Maximum Likelihood was the least accurate method for topology reconstruction for both the symmetrical and asymmetrical phylogenies (table 2.1). The relative accuracy of these phylogenetic methods remains the same across all dataset sizes and the two simulation topologies (table 2.1; figs 2.1-2.2).

Nodes closer to the tips are significantly more accurately reconstructed in the asymmetrical phylogenies across all dataset sizes (table 2.2; figure 2.2; supplementary figure S2.8). In the symmetrical trees, there was no significant correlation between distance from the tips and the accuracy of node reconstruction, except in the Maximum Likelihood analysis of 100 characters (figure 2.2; table 2.2).

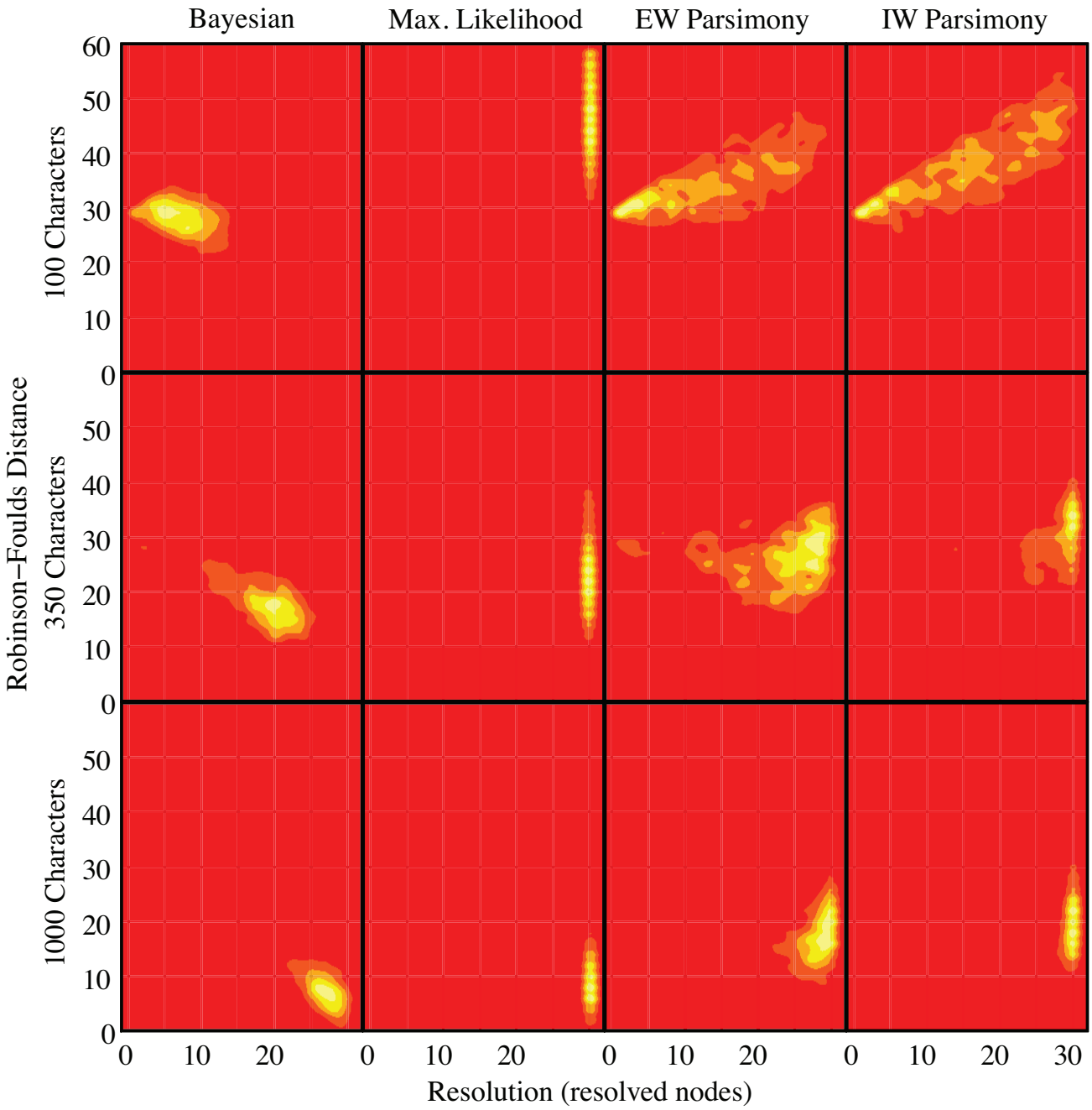


Figure 2.1 - Contour plots of Robinson-Foulds distance against phylogenetic resolution, indicating the higher accuracy of Bayesian implementations against all other methods with data generated on the asymmetrical phylogeny. The spectrum of red to yellow, reflect lower to higher density of trees. As the number of characters increases all methods converge on the correct phylogeny, although Bayesian phylogenies are generally the least resolved. The other methods achieve higher resolution but at a cost of lower accuracy. Data generated on the symmetrical phylogeny shows similar patterns but with much less variance and higher accuracy for all iterations; this lack of variance means point estimates cannot be shown as density estimates.

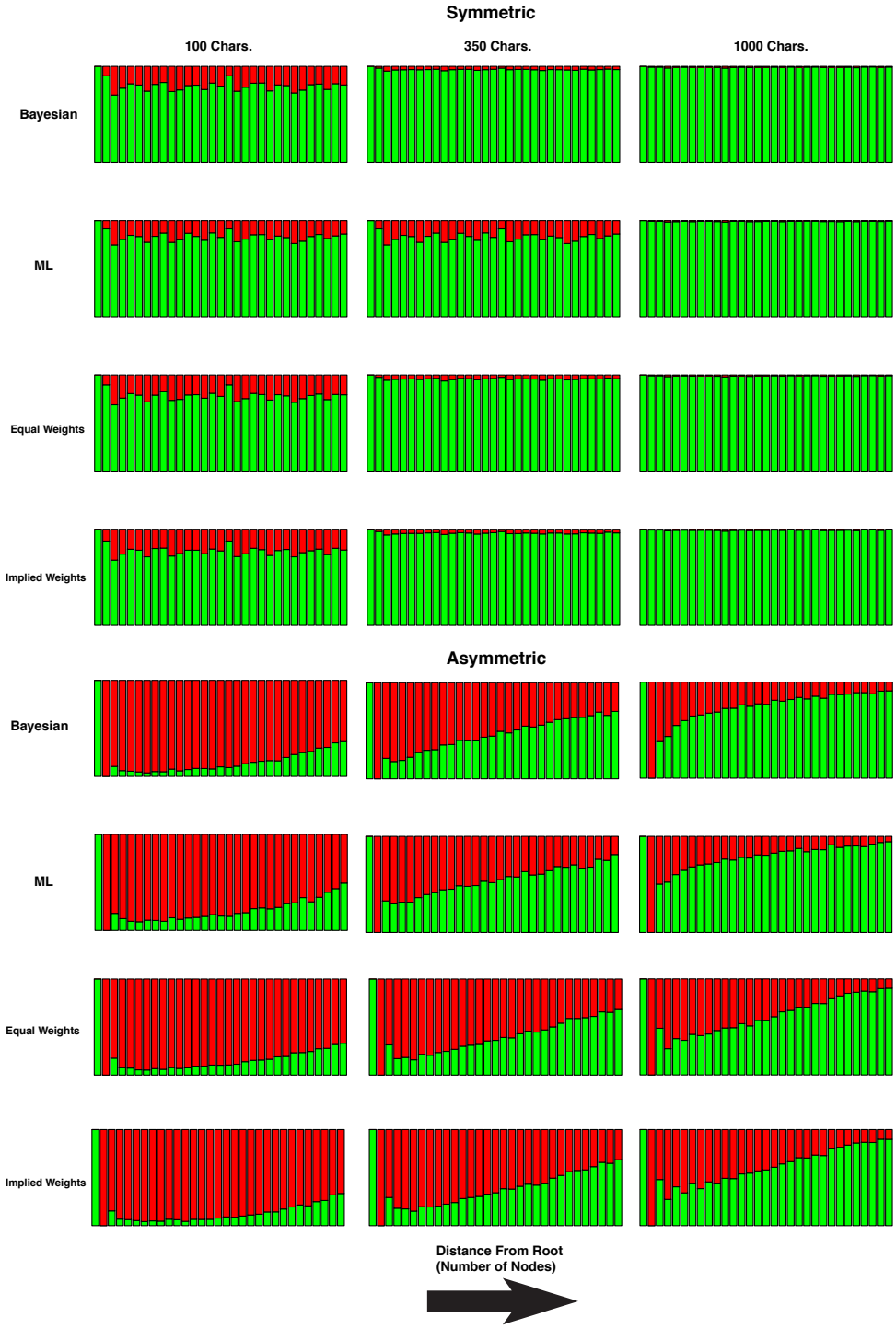


Figure 2.2 - Accuracy of reconstructed nodes is high across the symmetrical phylogeny, wheras accuracy increases with distance from the root in asymmetrical trees. The proportion of times a node was accurately reconstructed is shown in green, and the proportion it was inaccurately reconstructed is shown in red.

	equal weights parsimony	implied weights parsimony	maximum likelihood	Bayesian
asymmetrical generating phylogeny				
100	34.89 (22–56)	37.85 (22–56)	45.84 (20–58)	28.1 (18–39)
350	26.57 (11–51)	29.2 (12–51)	26.49 (6–58)	19.21 (7–35)
1000	17.82 (3–40)	19.16 (2–33)	11.94 (0–58)	9.34 (0–31)
symmetrical generating phylogeny				
100	8.08 (0–33)	9.29 (0–29)	10.1 (0–58)	7.51 (0–29)
350	1.33 (0–28)	1.43 (0–28)	1.8 (0–52)	1.2 (0–28)
1000	0.32 (0–26)	0.31 (0–26)	0.51 (0–52)	0.31 (0–26)

Table 2.1 - Bayesian approaches produce the most accurate trees for all character sets. Mean and range (in brackets) of Robinson–Foulds distances are lower for topologies estimated using Bayesian methods for both the symmetrical and asymmetrical generating tree. Maximum likelihood is the generally the most inaccurate method for the symmetrical generating tree, and implied weights parsimony performs worst for the asymmetrical generating tree.

.

	asymmetrical tree	symmetrical tree
MB 100	< 0.001	0.09919
Maximum Likelihood 100	< 0.001	0.027295
EW 100	< 0.001	0.106712
IW 100	< 0.001	0.092736
MB 350	< 0.001	0.638242
Maximum Likelihood 350	< 0.001	0.057809
EW 350	< 0.001	0.19683
IW 350	< 0.001	0.148108
MB 1000	< 0.001	0.256976
Maximum Likelihood 1000	< 0.001	0.085987
EW 1000	< 0.001	0.179186
IW 1000	< 0.001	0.287058

Table 2.2 - P Values from Spearman's rank correlation between the percentage of nodes being accurately reconstructed and their distance from the root. Nodes closer to the tips are significantly more likely to be accurately reconstructed in asymmetrical trees but this is not generally true for symmetrical phylogenies.

2.4.2 - Empirical phylogenies

Patterns of resolution achieved from the simulated datasets are similar for the empirical datasets. The Bayesian implementation of the Mk model estimates the least resolved phylogenies and Maximum Likelihood produces fully-resolved trees (full trees are shown supplementary figure S2.9-S2.15).

Kulindroplax, from the Sutton et al. dataset, is supported as a crown-mollusc based on Maximum Likelihood, EW-Parsimony, and IW-Parsimony (figure 2.3a-d). The results of the IW-Parsimony analysis are most similar to the original results (Sutton et al., 2012), with *Kulindroplax* resolved as a crown-aplacophoran; Maximum Likelihood analysis of the dataset resolved *Kulindroplax* as the stem-aplacophoran. The result of the Bayesian analysis of the dataset is largely unresolved, and *Kulindroplax* is not discriminated as a member of any clade within molluscs or even as a member of total-group Mollusca.

The anthophyte hypothesis (non-monophyletic gymnosperms sister to seed ferns plus angiosperms) recovered by Hilton and Bateman (2006) is supported by our EW-Parsimony and Maximum Likelihood analyses of their dataset which recovered a paraphyletic seed ferns plus Gnetophyta as sister to angiosperms (figure 2.3f, g); the results of Bayesian and IW-Parsimony analyses of the same dataset contradict the anthophyte hypothesis (figure 2.3e, h). The Bayesian analysis produced a non-monophyletic gymnosperms with the relationships between them and seed ferns unresolved with the exception of *Bennettitales* which resolved as a gnetophyte, and *Caytonia* as sister to the angiosperms.

Analyses of the Luo et al. (2015) dataset yielded congruent results with the original study, with the placement of *Haramiyavia* outside of crown Mammalia and multituberculates, although some haramiyids are resolved as crown mammals in the IW-Parsimony analysis (figure 2.4a-d).

Nyasasaurus is recovered as a member of Dinosauria in the Maximum Likelihood, EW-Parsimony, and IW-Parsimony analyses of the dataset from Nesbitt et al. (2013) (figure 2.4e-h). The Bayesian analysis recovers *Nyasasaurus* in a polytomy with the two major clades of dinosaurs, corroborating the conclusion of Nesbitt et al. (2013) that, given the data, its precise phylogenetic position is uncertain.

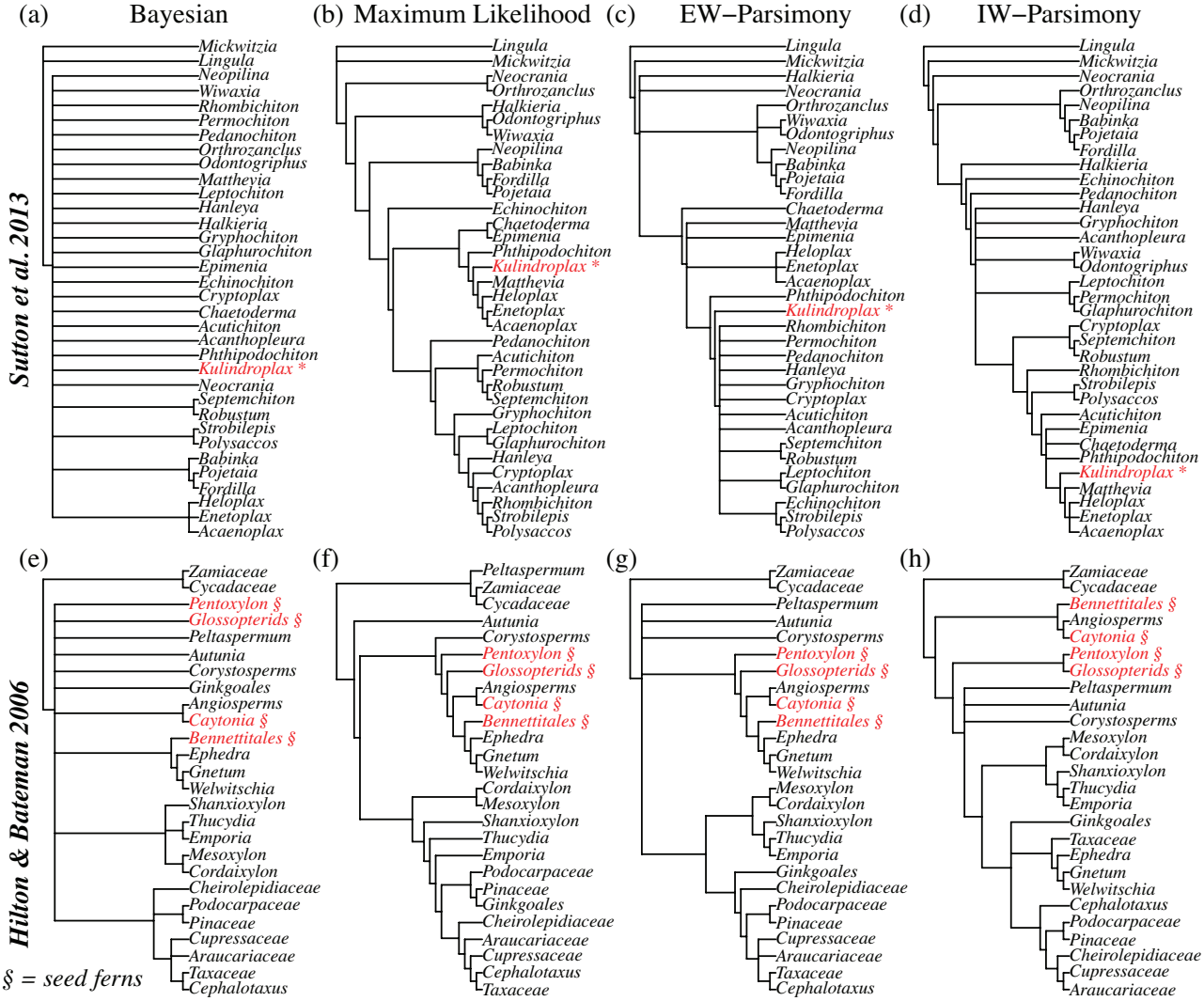


Figure 2.3 - Alternative phylogenetic reconstruction methods alter our understanding of evolution with empirical matrices. However, the relationship of fossil seed ferns from Hilton and Bateman (Hilton and Bateman, 2006) is changed according to implementation (a-d), although *Caytonia* remains as sister to angiosperms in all analyses. Alternative analyses change the taxonomic affinity of *Kulindroplax* from Sutton et al. (Sutton et al., 2012) (e-h).

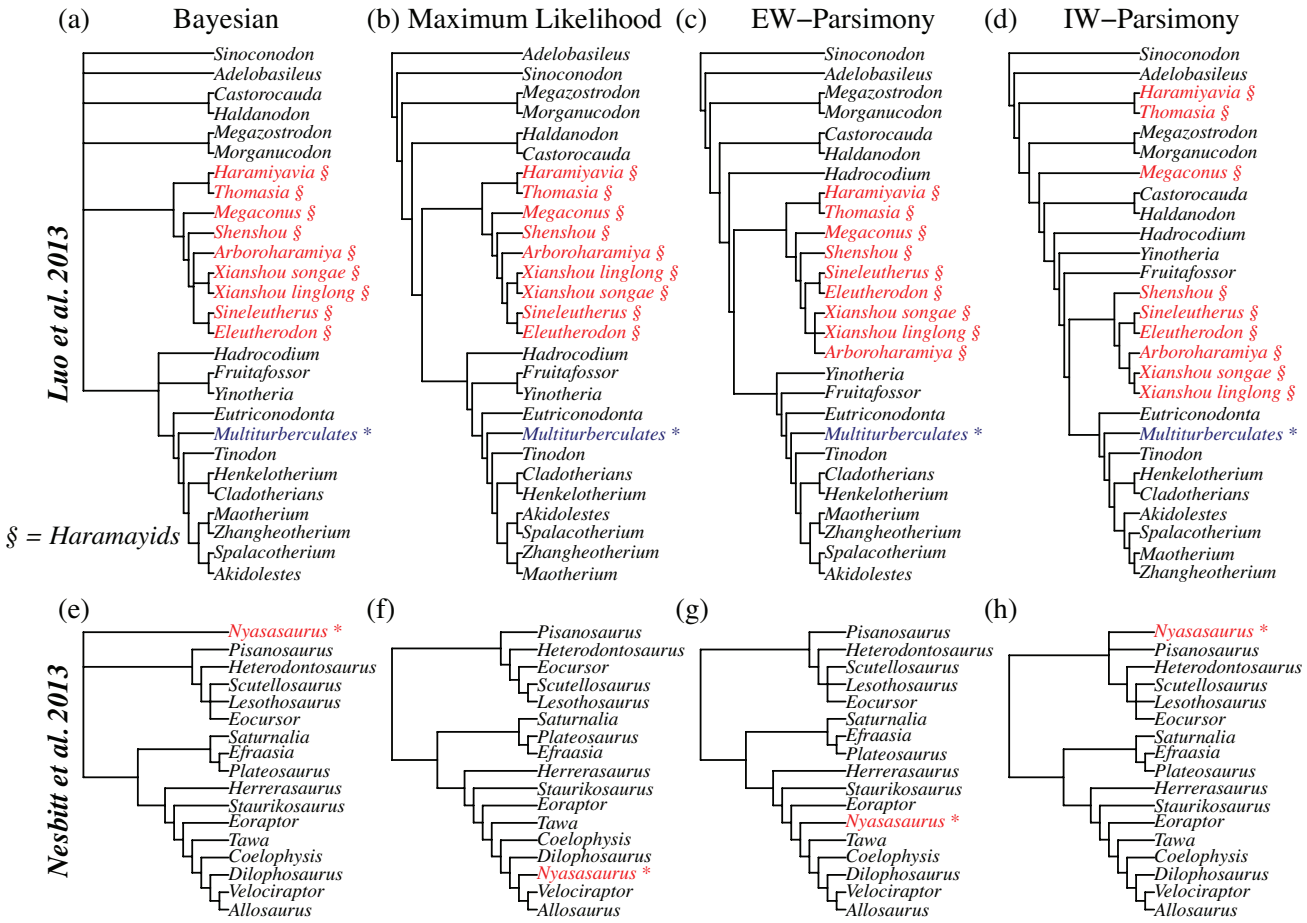


Figure 2.4 - Alternative phylogenetic reconstruction methods produce generally congruent reconstructions of evolution with empirical matrices. For Luo et al. (Luo et al., 2015), the relationship between the haramiyids and multituberculates is largely unchanged across analyses (a-d). IW-Parsimony (g) and Bayesian analyses place *Nyasasaurus* as close to the earliest dinosaur (e) and IW-Parsimony places it close to the earliest diverging taxa (g), but EW-Parsimony and Maximum Likelihood place the taxa as a derived member of Dinosauria (f,h).

2.5 - Discussion

2.5.1 - Simulations indicate that the Bayesian implementation of the Mk model outperforms all other methods and implementations

Previous simulation-based analyses that have attempted to evaluate the performance of likelihood and parsimony-based phylogenetic methods for analysing phenotypic data have found that the probabilistic model performs best (Wright and Hillis, 2014, O'Reilly et al., 2016). However, these studies were biased against parsimony because they employed an unresolved generating tree that is problematic since parsimony methods will attempt to recover a fully resolved tree from the simulated data yielding a non-zero RF distance from the generating tree, even if the two trees are effectively compatible. Furthermore, since previous simulation studies considered the Mk model only within a Bayesian framework, they did not distinguish between the impact of the probabilistic model of character evolution and the statistical framework in which it was implemented.

Our analyses control for these shortcomings of previous simulation studies and show consistently that the Bayesian implementation of the Mk model performs best. In line with previous simulations (O'Reilly et al., 2016), we found that EW-Parsimony performs better than IW-Parsimony. There is overlap between model performance shown by the distribution of Robinson-Foulds distances (table 2.1), but there is reason to have different degrees of confidence in the models; only the Bayesian implementation produces a relatively small distribution of tree performance compared to the large tails signifying worse performance in the two parsimony methods (table 2.1). We also found that the Bayesian implementation of the Mk model outperforms the Maximum Likelihood implementation, indicating that it is not merely the probabilistic transition model that outperforms parsimony methods, but the implementation of the Mk model within a Bayesian statistical framework. Indeed, the Maximum Likelihood implementation of the Mk model was the worst-performing method, worse even than IW-Parsimony. In part, the poor performance of the Maximum Likelihood-Mk method is because we did not capture phylogenetic uncertainty associated with this phylogenetic method. This is normally achieved in analyses of molecular datasets through bootstrapping methods, but these are inappropriate for the analysis of phenotypic data since the basic methodological assumption, that the phylogenetic signal is randomly distributed across sites (characters), is not true for morphological data.

However, irrespective of the phylogenetic method used, dataset size correlated positively with both phylogenetic accuracy and resolution, diminishing differences in the relative performance of the competing phylogenetic methods. All phylogenetic methods also performed best when

attempting to recover a symmetrical target tree; all methods found recovery of asymmetrical trees challenging and phylogenetic accuracy diminished from tip to root. The impact of tree topology is of particular concern since empirical phylogenetic trees are invariably asymmetric (Mooers and Heard, 1997), and trees of fossil species are infamous for their asymmetry (Shao and Sokal, 1990, Harcourt-Brown et al., 2001). However, there is a broad spectrum of tree symmetry, with fully symmetric and fully asymmetric trees representing end-members. Palaeontological trees with the dimensions used in our simulations are typically far from the fully asymmetric pectinate generating tree we employed (Ic = ~0.4 for 32 species) (Harcourt-Brown et al., 2001). Furthermore, the asymmetry of many palaeontological trees is often a representational artefact of attempting to summarise character evolution, or an analytic artefact of analysing the relationships among diverse clades based on representative species or higher taxa (Panchen, 1982). Thus, the challenge of recovering trees of extinct taxa may not be as great as a simplistic interpretation of our results might suggest.

2.5.2 - Analyses of empirical data bear out conclusions based on simulations

Maximum Likelihood, IW-Parsimony and EW-Parsimony methods of the simulated datasets commonly identify a single optimal tree, but the differences between the optimal trees derived from these methods provides no confidence that any one of the inferred topologies is accurate with reference to the placement of a taxon of interest. This view is corroborated by our reanalysis of empirical datasets which recovered poorly resolved trees using the Bayesian implementation of the Mk model, and in a number of instances, indicate that the conclusions drawn in the corresponding original studies are not supported by the data.

In an extreme example, our re-analyses of the dataset published by Sutton et al. (2012), which attempted to demonstrate a crown-aplacophoran mollusc affinity for *Kulindroplax*, yielded disparate hypotheses of affinity. EW-Parsimony and IW-Parsimony recovered the published result, while Maximum Likelihood recovered *Kulindroplax* as a stem-aplacophoran, and Bayesian could not discriminate *Kulindroplax* as a total-group mollusc (figure 2.4a). This poor resolution is unlikely to be a result of poor fossil evidence but, rather, the lack of discriminatory power in the small character matrix. Among the analyses of the dataset from Hilton and Bateman (2006) we recovered some of the principal competing topologies that have featured in debate over the affinity of seed plants in past decades. However, the Bayesian analysis of the dataset recovered a topology that is largely unresolved in terms of the relationships among key clades. This suggests that the available data are insufficient to discriminate among the competing hypotheses, and this long-standing debate is largely an artefact of the false resolution of parsimony methods.

Bayesian analyses need not overturn the results from previous analyses based on deterministic phylogenetic methods like EW-Parsimony, IW-Parsimony, and Maximum Likelihood. A phylogenetic position for haramiyids, outside crown-Mammalia, is corroborated by our Bayesian analysis of the dataset from Luo et al. (2015) - in contrast to the crown-Mammalia affinity recovered for some haramiyids through IW-Parsimony analysis of the same data (Figure 2.4d). Similarly, *Nyasasaurus* was posited as the earliest dinosaur, and this conclusion is supported by the Bayesian analyses (Figure 2.4e) although this is not supported by EW-Parsimony, IW-Parsimony, and Maximum Likelihood analyses (Figure 2.4f-h). However, the Bayesian analysis is more robust in expressing the phylogenetic ambiguity identified by the original authors (Nesbitt et al., 2013), as *Nyasasaurus* falls in a polytomy alongside the two major clades of dinosaurs.

Some of the differences between methods may simply reflect the dimensions of the dataset. The two datasets that cannot resolve relationships under Bayesian inference and exhibit significant topological discordance among phylogenetic methods (Hilton and Bateman, 2006, Sutton et al., 2012), are both comparatively small (34 taxa, 48 characters and 48 taxa, 82 characters, respectively). These both fall within the scope of simulated datasets that yield low resolution from the Bayesian method and, from other phylogenetic methods, high resolution but low accuracy (Figure 2.1). The two empirical datasets that yield trees with greater congruence from the different phylogenetic methods, are both larger: Luo (114 taxa, 497 characters) and Nesbitt (82 taxa, 413 characters). The size of these matrices is comparable with our simulation results in which we see marked increases in topological accuracy and agreement between methods (Figure 2.1 - between 350 and 1000 characters).

2.5.3 - Implications for phylogenetic analysis of phenotypic data

The results of our simulation studies indicate that the cadre of phylogenetic hypotheses generated from phenotypic data using parsimony methods require reassessment using the Bayesian implementation of the Mk model. It is likely that many evolutionary interpretations are contingent on precise but inaccurate phylogenetic hypotheses. In this undertaking it is important that the implications of our simulation studies are considered in the design of phylogenetic studies.

Firstly, phylogenies of fossils tend toward strong asymmetries (Harcourt-Brown et al., 2001) and, like all phylogenetic methods, Bayesian inference struggles with the recovery of deep nodes within asymmetric trees. Therefore, it is important that outgroups are sampled extensively, ensuring that contentious in-group relationships are closer to the tips, where

topological accuracy is highest. Further, in-group lineages should be sampled in a manner that does not accentuate tree asymmetry.

Secondly, phylogenetic accuracy and resolution correlates positively with the relative dimensions of the dataset. Accordingly, phylogenetic resolution or certainty should not be expected from cladistic analyses of small morphological datasets (i.e, those around 100 characters or fewer), particularly if they include fossils. There are finite limits to the number of available phylogenetically-informative characters (Scotland et al., 2003) and, for well-studied clades, it may be perceived that these phylogenetically informative characters have already been found. However, it is important to note that the concept of phylogenetic informativeness is different within a likelihood versus a parsimony framework: in parsimony characters that undergo few changes are prized in favour of homoplastic characters. Under the likelihood model, branch length, informed by the number of character changes, contributes to topology estimation. Thus, traditionally 'bad' phylogenetic characters (those exhibiting homoplasy) may find utility in expanding the dimensions of phenotypic character matrices as long as homoplasy falls within the limits that the model can accommodate. In a Bayesian framework this can be tested using posterior predictive tests of model adequacy (e.g. (Tarver et al., 2016)).

Finally, we may need to alter our expectations to anticipate less well-resolved but more accurate phylogenetic hypotheses, which will both constrain and guide research. Greater resolution may be found by generating matrices suited to likelihood- rather than parsimony-based phylogenetic methods. However, we must also come to terms with the prospect that for some groups of organisms, or their fossil remains, there may be insufficient data. As such, their evolutionary relationships might not therefore be resolvable using morphological data alone and, if they are fossils, their evolutionary significance may never be realised. Nevertheless, resolving phylogenies is not the end game for evolutionary biology. Incompletely resolved trees can still be used as a basis for investigating interesting macroevolutionary questions, and methods exist for incorporating tree uncertainty in phylogenetic comparative methods (e.g (Healy et al., 2014)).

2.6 - Conclusions

A growing consensus shows that the Bayesian Mk model is the most accurate method of phylogenetic reconstruction, and here we show that this remains true across dramatically different tree shapes, when analysing datasets composed of both multistate and binary characters, and when compared to Maximum Likelihood estimation using the Mk model. We would recommend that Bayesian implementations of the Mk model should become the default

method for phylogenetic analyses of cladistic morphological datasets, and we should expect low levels of resolution with small datasets. As parsimony methods appear to be less effective than probabilistic approaches, it may be necessary to alter data collection practices by moving away from choosing a selection of characters that undergo few changes, and moving towards scoring all possible characters from the available taxa irrespective of their expected homoplasy.

Chapter 3

Dating tips for divergence time estimation

Joseph O'Reilly, Mario dos Reis, and Philip C. J. Donoghue

This Chapter was published in *Trends in Genetics* Volume 31, Issue 11
DOI: 10.1016/j.tig.2015.08.001

3.1 - Abstract - The molecular clock is the only viable means of establishing an accurate timescale for Life on Earth yet it remains reliant on a capricious fossil record for calibration. 'Tip-dating' promises a conceptual advance, integrating fossil species among their living relatives using molecular and morphological datasets and evolutionary models. Fossil species of known age establish calibration directly and their phylogenetic uncertainty is accommodated through the coestimation of time and topology. However, challenges remain including: a dearth of effective models of morphological evolution, rate correlation, the non-random nature of missing characters in fossil data and, most importantly, accommodating uncertainty in fossil age. We show uncertainty in fossil-dating propagates to divergence time estimates, yielding estimates that are older and less precise than those based on traditional node calibration. Ultimately, node and tip calibrations are not mutually incompatible and may be integrated to achieve more accurate and precise evolutionary timescales

3.2 - Introduction

Establishing an evolutionary timescale for Life on Earth has long been a fundamental goal of evolutionary biology, providing the framework for inferring modes and rates of molecular and phenotypic evolution, as well as a means of associating intrinsic evolutionary change to extrinsic causal factors. This endeavour was originally the domain of palaeontologists, but it is now widely accepted that fossil data alone are insufficient because of the incompleteness of the fossil record (Benton and Donoghue, 2007). Molecular clock dating methodology can be used to establish an evolutionary timescale by calculating the molecular distance between species, and estimating absolute molecular evolutionary rates based on the oldest fossil evidence for the antiquity of the living lineages (Zuckerkandl and Pauling, 1965). This powerful combination of molecular and palaeontological data sees through the gaps in the fossil record, providing the only viable means of establishing an accurate evolutionary timescale.

Molecular clock methodology has been developed to accommodate tree-wide substitution rate heterogeneity (Sanderson, 2002, Thorne et al., 1998, Rannala and Yang, 2007, Drummond et al., 2006) and precision has increased with the availability of genome-scale datasets (i.e. an effectively infinite amount of sequence data) (dos Reis et al., 2012). However, further increases in accuracy and precision may only be possible with a concomitant increase in the precision of calibrations (Yang and Rannala, 2006, Dos Reis and Yang, 2013, Zhu et al., 2015, Rannala and Yang, 2007). Hence, recent years have witnessed attempts to constrain the uncertainties associated with fossil-based calibrations, including phylogenetic position, age interpretation, and the degree to which calibrating fossils approximate the true time of divergence for the nodes that they calibrate (Donoghue and Benton, 2007, Parham et al., 2012, Benton and Donoghue, 2007). Controversially, this requires not just the oldest fossil records of extant clades on which minimum age constraints are established, but also interprets the absence of older fossils attributable to the clade to establish maximum age constraints (Donoghue and Benton, 2007, Parham et al., 2012). Or else simple mathematical functions are employed to express, probabilistically, a visceral perception of the degree to which fossil minima reflect the time of lineage divergence (Ho and Phillips, 2009, Donoghue and Benton, 2007). Or fossil occurrence data can be modelled statistically, with or without reference to a phylogeny, to determine the extent of the temporal gap between the age of a clade and its oldest fossils (Marshall, 1994, Wilkinson et al., 2011, Heath et al., 2014). Attempts to constrain uncertainty with fossil calibrations must be welcomed, but they have hardly led to increased precision in divergence time estimation, not least since node calibrations require complex and often ad hoc interpretations of fossil and phylogenetic evidence to establish probabilistic calibrations, which are viewed by some as a grossly over-interpreted yet inadequate solution to a complex problem (Heads, 2012).

The recent introduction of fossil tip calibration (Pyron, 2011, Ronquist et al., 2012a), also known as 'tip-dating' or 'Total Evidence Dating' has, therefore, enjoyed an enthusiastic welcome. This method requires both molecular sequence and morphological character datasets that are analysed using molecular and morphological models of evolution, but its chief innovation is that it allows fossil species to be incorporated into divergence time analyses on a par with their living relatives. This calibration methodology is analogous to the manner in which ancient DNA or archived viral sequences of known age are employed to infer rates of evolution among extant species or strains (Drummond et al., 2003). In this case, fossils of known age calibrate the rate of evolution based on their phylogenetic position, branch length, and an inferred rate of evolution. Phylogenetic topology may be estimated independently or co-estimated with the divergence time analysis and the rate of evolution maybe based on independent or correlated rates of morphological and molecular evolution.

Thus, tip-calibration obviates many of the controversies associated with node-calibration. First, fossil species inform the evolutionary rate without recourse to ad hoc assumptions about the degree to which these species approximating the age of a living clade. Second, since time and topology can be co-estimated, it becomes possible to include older, temporally more-informative fossils that could not be used for node-calibration because their phylogenetic position is uncertain. Third, since calibrations no longer serve as prior estimates of clade age, tip-calibrations can be drawn from any and all fossil species, removing restrictions of the amount paleontological data that can be included in divergence time studies. Finally, tip calibrations summarise the age of a single species only, avoiding the over-interpretation of negative evidence in establishing maximum constraints.

Tip-calibration was originally introduced based on empirical divergence time analyses of insects (Ronquist et al., 2012a) and amphibians (Pyron, 2011), and it has since been applied to mammals (Slater, 2013, Schrago et al., 2013, Tseng et al., 2014, Slater, 2015, Dembo et al., 2015, Marx and Fordyce, 2015), teleost fishes (Near et al., 2014, Dornburg et al., 2015, Alexandrou et al., 2013, Arcila et al., 2015), arachnid spiders (Wood et al., 2013, Sharma and Giribet, 2014), flies (Winterton and Ware, 2015), and plants (Larson-Johnson, 2016). The approach has been extended to analyses of entirely extinct clades, relying exclusively on morphological data (Lee et al., 2014). While tip-calibration was initially advocated on the basis that it was less sensitive to root time prior densities and yielded more precise divergence time estimates in comparison to node-calibration (Ronquist et al., 2012a), subsequent studies have shown the reverse to be true (Arcila et al., 2015, Wood et al., 2013). Furthermore, tip-calibration has proven consistently to yield older age estimates than traditional node-calibration

(Arcila et al., 2015, Tseng et al., 2014, Sharma and Giribet, 2014, Wood et al., 2013, Schrago et al., 2013, Ronquist et al., 2012a, Slater, 2013, Slater, 2015). Thus, while it is clear that in incorporating all data pertinent to divergence time estimation, tip-calibration is the most promising approach for establishing accurate and precise evolutionary timescales, at present it appears to be less accurate than conventional node calibration methods. Below we consider the factors biasing current methods employing tip-calibration and suggest ways in which they can be developed to obtain more accurate divergence time estimates.

3.3 - Models of morphological character evolution and the incompleteness of fossils

While there are a number of nested models of molecular substitution, morphological models have not enjoyed much development, with only a handful proposed to date and even fewer actually implemented in popular software packages (Alekseyenko et al., 2008, Bouckaert et al., 2014, Swofford, 1998, Felsenstein, 1989, Ronquist et al., 2012b, Stamatakis, 2014). The Mk model of discrete character evolution has been utilised in all published tip-calibrated analyses to date (Lewis, 2001). The Mk model is a *k* states generalisation of the JC69 model of molecular substitution and, inevitably, it possesses many simplifying assumptions that may not hold true for morphology (Jukes and Cantor, 1969). Independent evolution of sites and equal equilibrium frequencies are two factors that are particularly difficult to justify for morphological evolution. Alternative models utilising continuous characters (Felsenstein, 1973) or the threshold model (Felsenstein, 2012, Felsenstein, 2005) are appealing alternatives but they have yet to be implemented.

The inherently incomplete nature of fossil phenotypic data, in comparison to living species, is undoubtedly a challenge to tip-calibrated divergence time analyses. The impact of missing sequence data on Bayesian phylogenetic topology estimation has been investigated, with the majority of studies indicating that it is unlikely to have a strong negative impact (Wiens and Moen, 2008, Wiens and Morrill, 2011, Wiens and Tiu, 2012, Lemmon et al., 2009, Simmons, 2011) except where there is a comparatively small number (not proportion) of non-missing sites {Wiens, 2008 #48. This is clearly a problem for topology estimation based on phenotype where datasets are generally very small in comparison to molecular sequence alignments. This issue is exacerbated by the decidedly non-random nature of missing phenotype data in fossil species (Sansom and Wills, 2013, Sansom et al., 2010). Fossil data are invariably biased towards the preservation of phenotypic characters that are manifest in or as mineralised skeletal structures. Even where soft tissue characters are exceptionally preserved, some groups exhibit a phenomenon coined “stem-ward slippage” in which features are lost to decay in reverse phylogenetic order making their fossils appear artefactually to belong to more primitive

evolutionary grades (Sansom et al., 2010, Sansom and Wills, 2013). While the impact of these factors on topology estimation has been considered, it has not been investigated explicitly in the context of time and rate estimation (Sansom and Wills, 2013).

For tip-calibrated divergence time analyses, the likely impact is two-fold: calibrating fossil species will be assigned to erroneously early-branching positions with the phylogeny, and the branch lengths will be underestimated, both due to their lack of shared-derived and autapomorphic soft tissue characters, missing artefactually as a consequence of non-random decay patterns. Both of these phenomena will influence rate estimates and, therefore, divergence time estimates. To minimise the negative influence of missing data, sub-sampling approaches have been proposed, allowing the use of only the most completely coded taxa or characters. While it has been argued that such approaches have minimal impact on topology and age estimation (Ronquist et al., 2012a, Pyron, 2011), this is unlikely to hold true for non-random missing data. Alternatively, a model of fossilisation could be employed that accounts for the directed loss of characters during preservation, but modelling this process may be entirely unrealistic given that fossilization potential varies with environment and taxonomic group.

3.4 - Dating tips and calibration strategies

Almost all total-evidence dating studies conducted so far have employed point age-estimates for the fossil species used as tip-calibrations, assuming implicitly that the age of the fossil is known without error. This has been done on the sometimes explicit justification that the errors associated with the dating of fossils are negligible (Ronquist et al., 2012a, Sharma and Giribet, 2014). This approach is reminiscent of the point age estimates for node calibrations, employed when divergence time estimation was in its infancy, and none of the lessons learned from the development of node-calibration strategies (Donoghue and Benton, 2007, Reisz and Muller, 2004, Parham et al., 2012) have been transferred to studies that employ fossil tip-calibration. It is well established that the age of a fossil can rarely, if ever, be known without error and this uncertainty must be accommodated regardless of whether the fossil is used in the construction of a node or tip-calibration. The age of any fossil occurrence can be constrained only to within an envelope of minimum-maximum bounds, the span of which varies depending on the attendant evidential context. Node-calibrations are based principally on the earliest secure fossil record of a clade and, thus, it is necessary to determine only the minimum age interpretation of the calibrating fossil (Reisz and Muller, 2004, Benton and Donoghue, 2007). At the least, the age of a tip-calibrating fossil requires establishing both its minimum and maximum age interpretations. For both the minimum and maximum age interpretations, this invariably entails a tortuous daisy-chain of litho-, bio-, chemo-, cyclo, and /or magneto- stratigraphic correlations

between the site of the fossil occurrence and another in which a geochronological absolute date has been established, at each step taking the minimum or maximum relative age interpretation, as appropriate, leading to iteratively increasing age uncertainty; see Box 2 (3.7) for a worked example. It is likely that in many instances, this uncertainty will exceed that associated with local node-calibrations, though tip calibrations may prove more palatable since they rely on fewer assumptions.

Borrowing from practice in establishing node-calibrations, the age uncertainty associated with a fossil species can be modelled as a uniform distribution if there is equal probability of the age of the fossil, per unit time, between minimum-maximum age interpretations. Or else, the variety of parametric distributions already implemented for node calibrations may be redeployed in instances where there is justification for focussing uncertainty closer to the minimum, maximum or mid range between age bounds.

To demonstrate the method by which a robust fossil tip calibration can be constructed we can use one of the hymenopteran species from (Ronquist et al., 2012a) (Figure 3.1). *Palaeathalia laiangensis* was recovered from the Laiyang Formation in Liaoning, China, which can be divided into four members, the third of which has yielded most fossils. Although the Laiyang Formation contains no directly dateable elements, correlation with the base of the Yixian Formation, also of China, allows the use of radiometric dates for the base of this formation to inform the age of the Laiyang Formation. Similarly, the unit overlying the Laiyang Formation, the Houkuang Formation, contains dateable elements, allowing an age for the base of this formation to constrain the age of the top of the Laiyang Formation. As we consider the age of the fossil species *P. laiangensis* to lie within the chronological interval between the top and base of the unit of its provenance, and without further information to constrain the limits and distribution of probability, we can use the ages of these limits to determine the bounds of our calibration. Correlation with the Yixian Formation can be made based on numerous palynological and faunal similarities, mostly with the lowermost member of the Yixian Formation, the Lujiutun Bed. While these sources may not individually be considered conclusive, numerous biostratigraphic similarities strongly support this correlation (Zhang.J. and Rasnitsyn.A., 2006, Hu.C. et al., 2001, Chen.P. et al., 2005, Chen.P. et al., 2006, Zhou et al., 2003). Radiometric dates of 128.4 ± 0.2 Ma have been acquired from the base of the Lujitan Bed, which can be used to determine the age of the base of the Laiyang Formation on the basis of the correlation between these units (Zhoue.Z., 2006, Wang.S. et al., 2001, Zhou et al., 2003).

The Laiyang Formation is succeeded by the Qingshan Group, of which the Houkuang Formation is the lowermost member. As the Laiyang Formation can be no younger than the

overlying unit an age for the base of the Houkuang Formation can provide a minimum age for the Laiyang Formation. U-Pb dating of zircons from the base of the Houkuang Formation has yielded dates of 106 Ma ± 2 Myr, which can be used to constrain the minimum age of the Laiyang Formation (Ling.W. et al., 2007). As no dates are available to further constrain the limits of this formation, and without any further information regarding the manner in which the probability of the age of *P. laiangensis* should be distributed, a uniform distribution spanning the full range of uncertainty in radiometric dates across the interval (128.6 – 104 Ma). This tip age can be contrasted with that utilised by Ronquist et al. (Ronquist et al., 2012a) of a fixed age of 140 Ma, which falls significantly outside the bounds of this calibration.

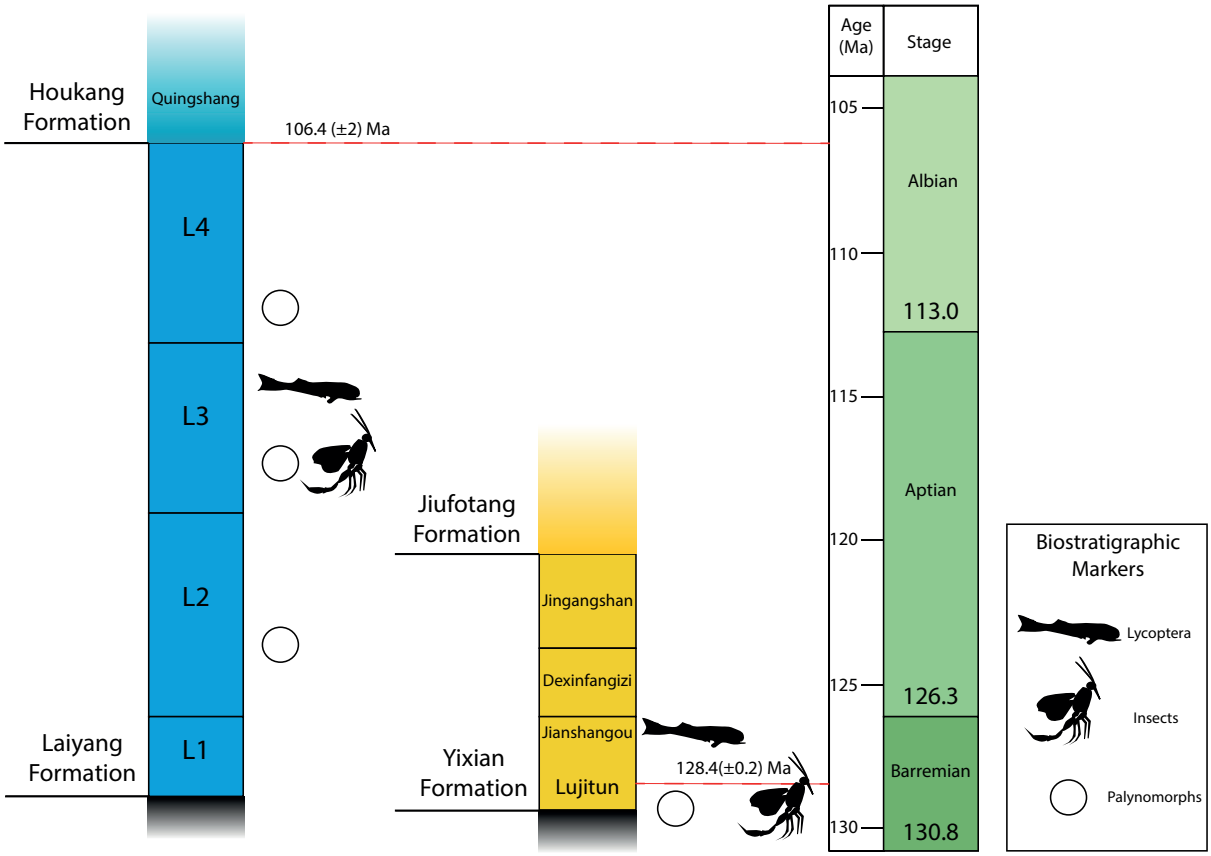


Figure 3.1 - Construction of a tip-calibration for P. laiangensis based on stratigraphic correlation between the unit of provenance, The Laiyang Formation, and the Yixian Formation of China.

Tip-calibrations present further peculiarities that should also be considered in attempting to integrate uncertainty associated with their age. For example, many fossil species employed in the node-calibration of divergence time analyses are not single occurrences but, rather, occur through a stratigraphic age range. This is of little relevance to node-calibration used to establish a clade age minimum, however, in establishing a tip-calibration, this is much more germane. Given that, by definition, such species will exhibit little or no morphological variation, it seems appropriate that this age range should be incorporated into the age uncertainty associated with the fossil. Ultimately, it may prove useful to integrate this information, in the form of effective stasis in the set of traits analysed, into the inference of rate variation across the tree.

Since tip-calibration and total evidence have been presented as a means of achieving greater precision in divergence time estimation (Ronquist et al., 2012a), it is pertinent to consider whether this can be sustained while integrating the uncertainty associated with the age of fossil tips. To this end, we reanalysed the dataset from the seminal total-evidence study (Ronquist et al., 2012a), in which tip-calibrations were utilised to estimate divergence times for Hymenoptera. Ronquist and colleagues were focussed on the theoretical and practical introduction of the method and they made no account of the uncertainty associated with the fossils used in tip-calibration. We reproduced the calibrations for each fossil tip, accommodating uncertainty in the age of each fossil species using probabilistic distributions (see Box 2 (3.7). for an example of this process). In contrast to previous assertions, that the uncertainties associated with tip ages would be negligible (Ronquist et al., 2012a, Sharma and Giribet, 2014), our attempts to capture a realistic estimate of the associated uncertainty results in tip-calibrations that span tens of millions of years – in contrast to the errorless estimates of age estimates used by the original authors (see Appendix for calibration details). These results can be contrasted with similar analyses performed on a much shallower timescale, in which the inclusion of tip age uncertainty resulted in a negligible impact on divergence time parameter estimates (Molak et al., 2013). To determine the performance of node-versus tip-calibration, we also constructed node-calibrations following established best practise (Parham et al., 2012) (see Appendix for calibration details). On average, recalibrated node priors were 23 Myr wider than the original calibrations. In both tip- and node-calibrations, uncertainty was modelled as a uniform distribution. Analyses were performed in MrBayes 3.2.2 (Ronquist et al., 2012b) in broadly the same manner as the original article. Precision was measured as the width of the 95% confidence interval (CI) for posterior estimates of node age for fourteen key in-group clades that could be resolved.

Our analyses show that when fossil age uncertainty is properly accounted for, tip-calibrated analyses do not necessarily yield divergence time estimates that are more precise than those

derived using node-calibration. Furthermore, for 27% of fossil taxa, the 95% HPD posterior estimates of fossil tip age did not encompass the original fixed tip-calibration, demonstrating the importance of appropriate prior construction. Divergence time estimates based on node-calibration are the most precise in all but four of the component clades (Figure 3.2). In line with almost all previous total-evidence studies, tip-calibration yields clade ages that are older, in general, than like-for-like estimates based on node-calibration; the only exceptions being divergences outside Hymenoptera. These deeper divergence times are most prominent in Vespina, where it appears that relaxing the constraint on the age of *Mesorussus*, (which was assigned to Vespina in both our analysis and the original analysis (Ronquist et al., 2012a)) from 94 Ma to 93.7-140.3 Ma leads to the older age estimates.

While we were able to repeat the results of the original analysis using the original calibrations, we were unable to reproduce the topological resolution and/or monophyly of Xyelidae, Pamphilioidea, and the placement of fossil taxa *Palaeathalia, Cleistogaster,* and *Prosyntexis* when employing our revised tip-calibrations. Since the only variable between our analyses is the method of calibration construction, it appears that the more realistic age-uncertainty associated with the fossils in our revised tip-calibrations has impacted on topology estimation as part of the co-estimation of topology and time. Thus, by implication, accommodating the realistic age uncertainty associated with fossil tip-calibrations also impacts rate and clade age estimates indirectly by contributing to topology estimation.

Claims of the superiority of tip-calibration over node-calibration appear unfounded when fossil age uncertainty is accommodated equally. Furthermore, it is not entirely clear that node calibrations are redundant in tip-calibration studies since, logically, they can still serve their purpose of constraining node age estimates and rate variation. One way to assess whether they are still useful in this role is in comparing traditional node calibrations and the posterior node-age estimates based on analyses employing tip-calibrations. We did this for the nine nodes for which we have constructed calibrations. The results (Figure 3.3) show that while all of the node age estimates derived from tip-calibration are old relative to the node calibrations, four fall fully outside these node age constraints. It could be argued that this demonstrates the fallacy of fossil-based maximum age constraint, however, two of the node age estimates include age ranges that are younger than the minimum age constraints based on the empirical palaeontologial evidence. Evidently, there remains a role for node age constraints, even in tip calibration divergence time analyses.

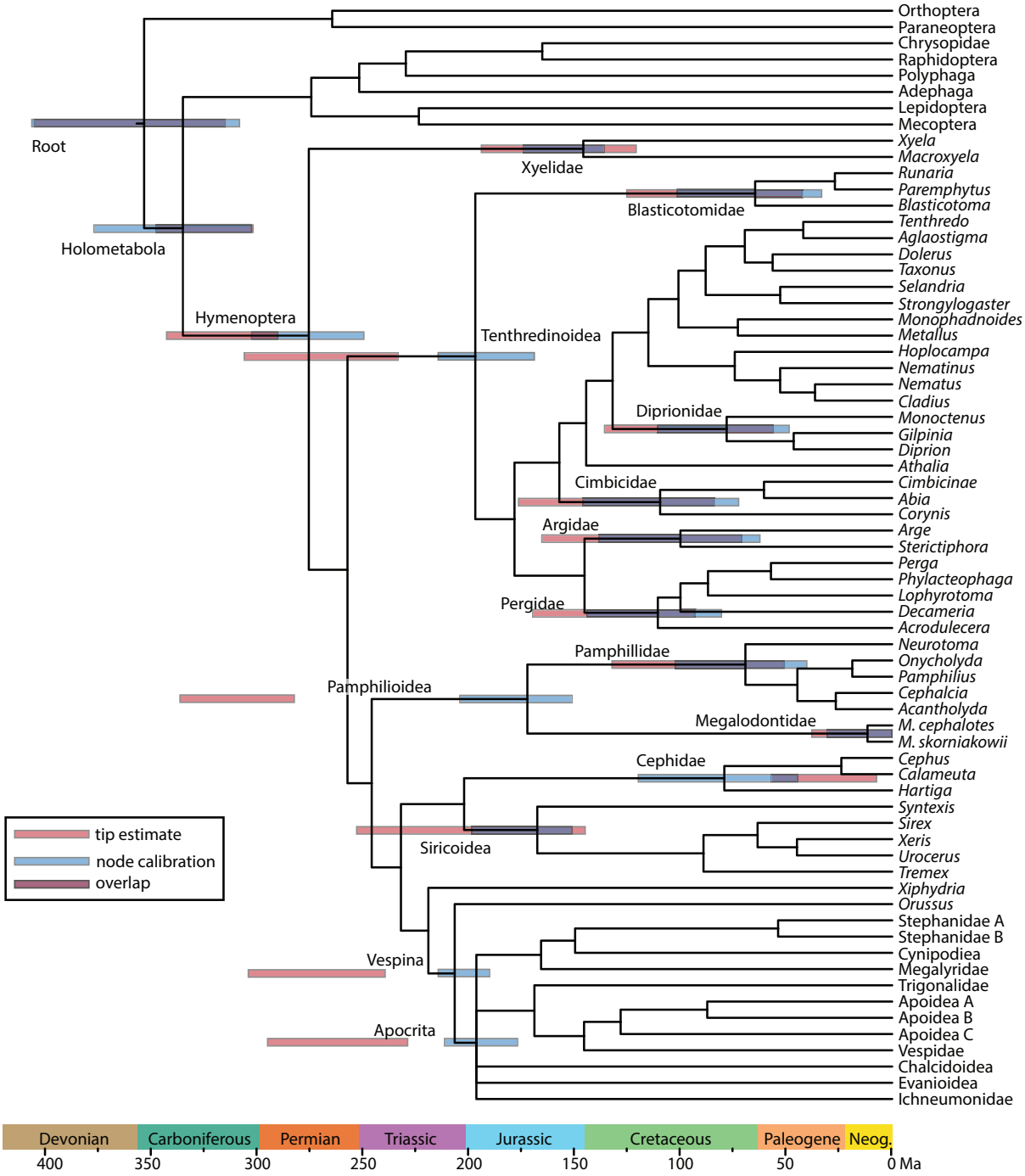


Figure 3.2 - A dated phylogeny of Hymenoptera produced using node-calibrations. Node bars represent 95% highest posterior density (HPD) for node ages estimated with either node-calibration or total-evidence dating (blue and red respectively). The dotted lines join HPD bars to the node for which they represent age estimate confidence and do not represent an extension of the confidence interval.

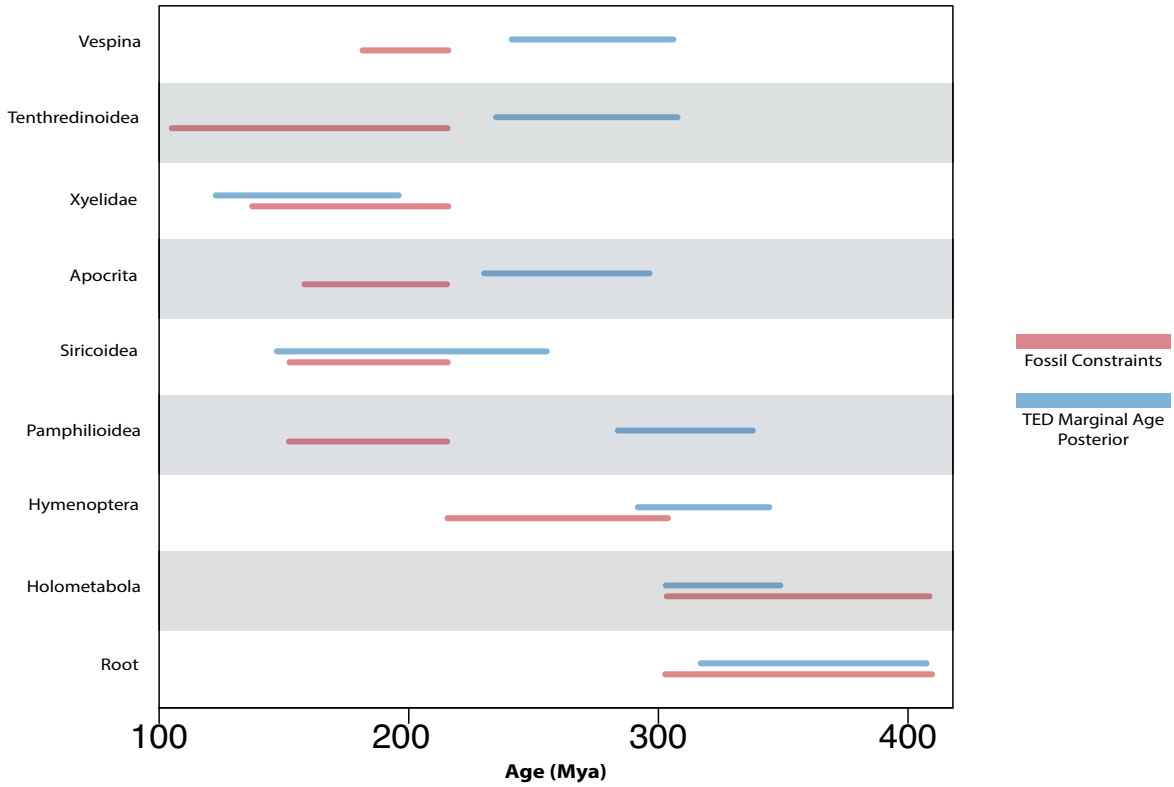


Figure 3.3 - Comparison between marginal posterior distributions on 9 node ages estimated with TED (blue), and prior clade-age constraints employed for node-calibrated analysis of the same data (red). The calibrations for node-calibrated analysis encapsulate the fossil evidence for the possible age of each clade. A lack of overlap at any node implies that there is discordance between the TED induced prior on that node and the fossil record. Discordance between these two distributions demonstrates that TED may lead to empirically unsupportable clade age estimates.

3.5 - Total Evidence Dating - less than the sum of its parts?

While total evidence dating has been presented as an alternative approach to conventional node-calibrated molecular clocks, this is a false dichotomy. Total-Evidence Dating is a particular combination of approaches that are neither inextricably linked, nor mutually exclusive from node-calibrated molecular clock analysis. These include: (i) the relaxed morphological clock, (ii) tip-calibration, and (iii) co-estimation of time and topology. In practise, these methods can and have been deployed in isolation in augmenting conventional molecular clock analyses. For example, Schrago et al. (2013) and colleagues' divergence time study of New World primates followed a two-step protocol, using the posterior age estimates from a conventional molecular clock analysis of living species as time-priors on node ages in a morphological clock analysis including both living and fossil species. At the least, this approach obviates the problematic assumption that molecular and morphological data co-vary, following a single rate model. Lee et al. (2014) co-estimated time and topology using dated tips and a morphological clock, eschewing molecular data altogether, in their analysis of body size evolution through the dinosaur-bird evolutionary transition. This approach will surely be adopted widely as palaeontologists seek to obtain clade ages, rather than minimum ages, for entirely extinct clades. However, this enthusiasm may be short lived given that tip-calibration approaches have consistently yielded older clade age estimates than conventional molecular clock studies – against which, palaeontologists have a long tradition of objecting violently (Donoghue and Smith, 2003). Combining ancient DNA and morphological data is another possibility afforded by tip-calibration, as has been applied to studying Pantherhine phylogeny (Tseng et al., 2014). This combination of ancient morphology and DNA may facilitate more accurate estimates of evolutionary rate.

While there has been enthusiasm in the application of the total evidence approach, not least since it provides a platform for the integration of so many disparate sources of uncertainty, it is arguable that in so doing this approach serves as a black box that disengages the user from the assumptions underpinning the analysis, many of which are very difficult to justify. One of the most problematic, potentially, is the co-estimation of time and topology, which, as we have demonstrated, allows fossil ages to constrain their phylogenetic position and, therefore, impact the estimation of rates and dates. This follows the common sense expectation that the age of a fossil species must reflect their phylogenetic position. Indeed, phylogeny estimation integrating the relative stratigraphic age of fossil species has a long tradition in palaeontology, but it has been much debated (Smith, 2000, Alroy, 2002, Wagner, 2002, Fisher et al., 2002, Sumrall and Brochu, 2003) and generally abandoned in favour of phylogenetics based on phenotype, perhaps refined by stratigraphy, except in groups with exceptionally rich fossil records that are

rarely if ever the focus of divergence time studies (Wickstrom and Donoghue, 2005). Though there is a broad correlation between clade age and phylogenetic branching order (Benton et al., 2000) this relationship breaks down as fossil taxon sampling decreases (Fortey and Jefferies, 1982). It is complicated further by secular biases in the rock record which serve to telescope temporally distinct fossil species originations and extinctions (Holland, 2000) and in the differential preservation of fossil groups and the environments in which they lived (Behrensmeyer et al., 2000). Thus, there appears little justification for the co-estimation of time and topology where fossil ages contribute to their phylogenetic position. We strongly advocate the prior analysis of topology before divergence time estimation. It is unfortunate that this approach precludes the integration of phylogenetic uncertainty into divergence time estimation, but resolving phylogenetic uncertainty using tip age is not viable using current methods.

The majority of TED analyses model branch rates as linked across morphological and molecular partitions (i.e. the application of rate multipliers to describe inter-partition rate heterogeneity (Ho and Lanfear, 2010, Yang, 1996, Nylander et al., 2004)). The suitability of this assumption for partitioned molecular data alone has been investigated, and partition-specific clocks developed for when this assumption is not met (Ho and Lanfear, 2010, Duchêne and Ho, 2014). However, the effect of morphological and molecular partition-specific clocks has barely been considered (Thornhill et al., 2012, Pyron, 2011, Ho and Lanfear, 2010), and most studies employ a single, partition-linked clock despite the fact that a strong co-varying relationship between molecular and morphological rates has never been demonstrated (Bromham et al., 2002, Seligmann, 2010, Davies and Savolainen, 2006). Morphological rate heterogeneity has long been considered likely to significantly dwarf its molecular counterpart, suggesting that the assumption of phenotypic and molecular rate correlation is unjustified (Kimura, 1983, Haldane, 1949). Molecular rates are interpreted as genome-wide measures of the number of substitutions accumulated per time unit, while morphological rates reflect only those aspects of the genome that specify the phenotypic traits analysed, further diminishing any expectation of covariance between molecular and morphological evolutionary rates (Bromham et al., 2002, Gillespie, 1991). In this light, it is perhaps unsurprising that unlinked partition-specific clocks have been found to be better-fitting than a single linked clock for mixed data analyses (Lee et al., 2013).

While node and tip-based calibration have been presented as competing approaches, they are not mutually exclusive. Indeed, some temporal constraints on clade age are better suited to being implemented as node-calibrations. This is particularly true of biogeographic calibrations where, based on the modern and ancient biogeographic distributions of evolutionary lineages, it is acceptable to assume that a dateable vicariance event, such as continental fragmentation, is

causal to lineage divergence. Similarly, some fossil-evidence is better reflected as node-age calibrations, rather than through including component fossil species as tip-calibrations. Node and tip-calibrations have already been employed together to calibrate interior nodes of the out-group, while allowing for an unconstrained in-group topology, or as part of a highly constrained topology in which fossil taxa are assigned to predetermined clades (Ronquist et al., 2012a, Beck and Lee, 2014). However, this must be extended to allow node-calibrations throughout the tree. This approach requires a fixed topology (or at least backbone constraints compatible with calibrated nodes) and, thus, precludes the possibility to co-estimating time and topology but, as we have argued, this may not be a material loss. Node calibrations may serve to mitigate against the propensity for tip-calibration-based studies to yield unacceptably ancient divergence dates, since it places additional constraints on the age of internal nodes of the tree, providing local checks on branch length and rate variation.

Finally, it is likely that the mismatch between divergence time estimates based on node and tip-calibration strategies is based at least in part in the shortcomings of the Mk model in explaining the phenotypic data commonly used in tip-calibration studies. The Mk model fails to account for expected characteristics of cladistic data, including the covariation of characters that are biologically linked, and logically linked through character design. Doubtless, the excitement surrounding the combined use of morphological and molecular data for divergence time analysis will lead to the development of this and other models of evolution. However, it may also be appropriate to consider different approaches to characterising phenotype, such as through the kinds of continuous variable characters obtained through morphometry of features such as skull suture patterns, tooth shape, or the dimensions of limb bones. The stochastic variation of such data is more similar to the variation seen in molecular sequence alignments and, as such, may be more readily modelled and better suited to combined data divergence time analysis.

3.6 – Conclusions

The advances inherent in Total Evidence Dating provide an excellent platform for the further development of methods for divergence time analysis. However, many aspects of the principal evolutionary model for phenotypic data currently employed are violated by the evolutionary process it attempts to encapsulate. The extent of these problems is so great that divergence time estimates derived using tip-calibration cannot enjoy the same confidence as conventional node-calibrated molecular clock studies. However, with the development of evolutionary models, protocols for dating fossil species and dealing with missing data, Total Evidence Dating encompasses a variety of powerful tools, the combination of which can be chosen to best test the hypothesis at hand. It also provides a viable framework for the best and greatest use of

palaeontological data that may serve as a nexus of the unification of palaeontological and molecular approaches to establishing evolutionary timescales.

Chapter 4

Tips and nodes are complementary not competing approaches to the calibration of molecular clocks

Joseph E. O'Reilly and Philip C. J. Donoghue

This Chapter was published in *Biology Letters* on 19th April 2016
DOI: 10.1098/rsbl.2015.0975

An invited contribution to the special feature 'Putting fossils in trees: combining morphology, time and molecules to estimate phylogenies and divergence times'

4.1 - Abstract - Molecular clock methodology provides the best means of establishing evolutionary timescales, the accuracy and precision of which remains reliant on calibration, traditionally based on fossil constraints on clade (node) ages. Tip-calibration has been developed to obviate undesirable aspects of node calibration, including the need for maximum age constraints that are invariably very difficult to justify. Instead, tip-calibration incorporates fossil species as dated tips alongside living relatives, potentially improving the accuracy and precision of divergence time estimates. We demonstrate that tip-calibration yields node calibrations that violate fossil evidence, contributing to unjustifiably young and ancient age estimates, less precise and (presumably) accurate than conventional node calibration. However, we go on to show that node and tip calibrations are complementary, producing meaningful age estimates, with node minima enforcing realistic ages and fossil tips interacting with node calibrations to objectively define maximum age constraints on clade ages. Together, tip and node calibration may yield evolutionary timescales that are better justified, more precise and accurate than either calibration strategy can achieve alone.

4.2 - Introduction

The molecular clock has displaced the fossil record as the primary means of establishing an evolutionary timescale, however, the accuracy and precision of divergence time estimates and their fossil calibrations remain inextricably linked (Rannala and Yang, 2007). Traditionally, divergence time estimation has achieved calibration based on geological (usually palaeontological) constraints on clade (node) ages. This approach has been developed to the extent that further improvements in accuracy and precision are limited by the inherent uncertainty in fossil evidence. Indeed, it is this uncertainty that has called into question the approach of node calibration, particularly what some see as the over interpretation of palaeontological data to establish maximum constraints on clade ages, and the difficulty in objectively representing prior evidence of node age as a probability distribution (Ronquist et al., 2012a). Furthermore, node age constraints invariably differ from those specified as a consequence of their integration into the joint time prior on node ages (Warnock et al., 2012). These concerns have led to the replacement of node calibrations with tip calibrations in which fossil species of a known age are integrated directly into divergence time analyses, supplementing sequence data from living species with morphological data from living and fossil species (Ronquist et al., 2012a, Pyron, 2011). However, there has been little effort to demonstrate the effect of different approaches to calibration and, indeed, to determine whether the effective prior on node ages resulting from tip calibration is compatible with the fossil evidence usually employed in node calibration. This is of particular interest given growing concern that tip-calibration consistently yields unrealistically ancient divergence time estimates (O'Reilly et al., 2015).

Hence, we sought to compare the efficacy of tip and node calibration by determining the compatibility of the resulting effective prior on node ages resulting from tip calibration and fossil-based node age constraints. This is readily sampled in node and tip calibrated analyses when the time-prior is conditioned on a fully constrained topology upon which ages are estimated. However, it is challenging where topology and time are co-estimated. Here we show that, in such circumstances, an approximation of the time prior can be obtained by conditioning on the consensus tree derived from a posterior sample of trees. Using an empirical dataset, we show that effective node age priors derived from tip calibration are often incompatible with fossil evidence, violating either minimum or maximum node age constraints. We argue that this contributes to the unrealistically ancient divergence time estimates produced by tip-calibration. These artefacts are diminished by combining tip and node calibrations, where node calibrations ensure that divergence time estimates never violate fossil-based minima and tip calibrations effectively establish node age maxima.

4.3 - Materials and Methods

We compared the effective node age priors and posteriors for tip and node calibration using a previously published hymenopteran dataset of molecular and morphological characters (Ronquist et al., 2012a). The original study assumed errorless tip-ages for fossil species. We employed revised ages for these species, integrating associated uncertainty, and derived node age constraints in order to compare effective priors on node ages to the palaeontological evidence (O'Reilly et al., 2015). Uncertainty in fossil taxon age was represented with uniform distributions whereas node calibrations were assigned offset exponential distributions, as in (Ronquist et al., 2012a). Unbounded distributions allow maxima to be defined by interaction between node and tip calibrations.

To obtain an approximation of the time prior we sampled from the prior while conditioning on the consensus of a sample from the posterior distribution of trees obtained from a standard tip-calibrated Total Evidence Dating (TED) analysis. We then constrained the topology to the consensus tree and sampled from the prior conditioned on this tree, providing a meaningful approximation of the effective time prior in a topologically unconstrained tip-calibrated analysis (see Appendix for details).

To evaluate the influence of tip calibrations, we compared effective priors and posterior estimates of node ages from tip-calibrated analysis to (i) the raw palaeontological constraints on node ages, and to the effective priors and posterior estimates of node ages derived from (ii) a node-calibrated analysis, and (iii) an analysis that implemented both tip- and node-calibrations. In the latter, fossil taxa were assigned to clades identified in the standard tip-calibrated analysis; where possible, the clades are assigned node calibrations. Minima on node-calibrated clades are defined by fossil evidence and maxima are established based on interaction between node and tip calibrations. We obtained a posterior sample of trees using the consensus tree produced from this sample to sample the effective time prior. Several fossil taxa and node calibrations could not be included in this analysis because of limitations of MrBayes (see Appendix for detail).

4.4 - Results

Our tip-calibrated consensus topology (Figure 4.1a) differs from (Ronquist et al., 2012a) in the placement of fossil Xyelidae, which could not be resolved in our analysis. *Spathoxyela* and *Mesoxyela* form a polytomy with extant Xyelidae because they are alternately assigned to crown or total-group Xyelidae in the tree sample; in the original analysis all fossil Xyelidae were resolved to the stem in the consensus tree. Following (Ronquist et al., 2012a), *Eoxyela*, the fossil defining the node calibration for Xyelidae, is resolved outside of crown Xyelidae. A number of fossil taxa, including *Palaeathalia* and *Cleistogaster*, were placed with higher

resolution in our recalibrated analysis than in the original. Like (Ronquist et al., 2012a), we were unable to recover unequivocal monophyly of Pamphilioidea.

The effective priors on node ages resulting from tip-calibration alone (excepting the two deepest nodes) consistently extend beyond the maximum palaeontological constraints on node ages, and include more ancient ages than the effective priors on node ages in the node-calibrated analysis. In two clades (Xyelidae and Siricoidea), tip calibration produces effective priors extending to the near recent. The effective time priors on these clades plus Pamphilioidea, extend beyond the minimum palaeontological constraints on the ages of these crown clades, and encompass younger ages than the effective priors on node ages in the node-calibrated analysis. In all instances, these differences propagate to the posterior estimates of clade ages. The anticipated linear relationship between node age and HPD width holds only for the node-calibrated analysis (figure 4.2). The results of the tip-calibrated analysis exhibit an inverse relationship, with uncertainty decreasing with proximity to the root.

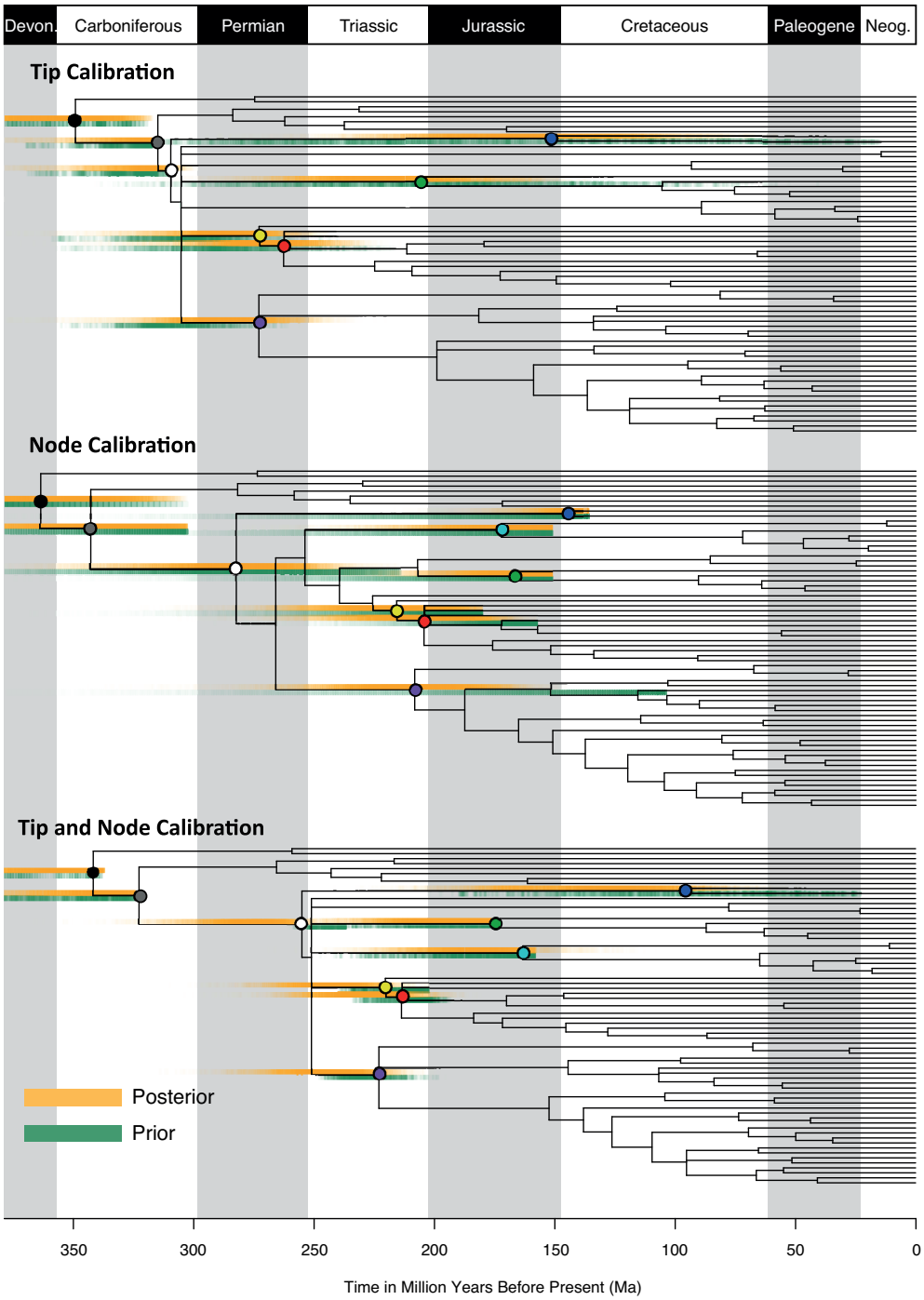
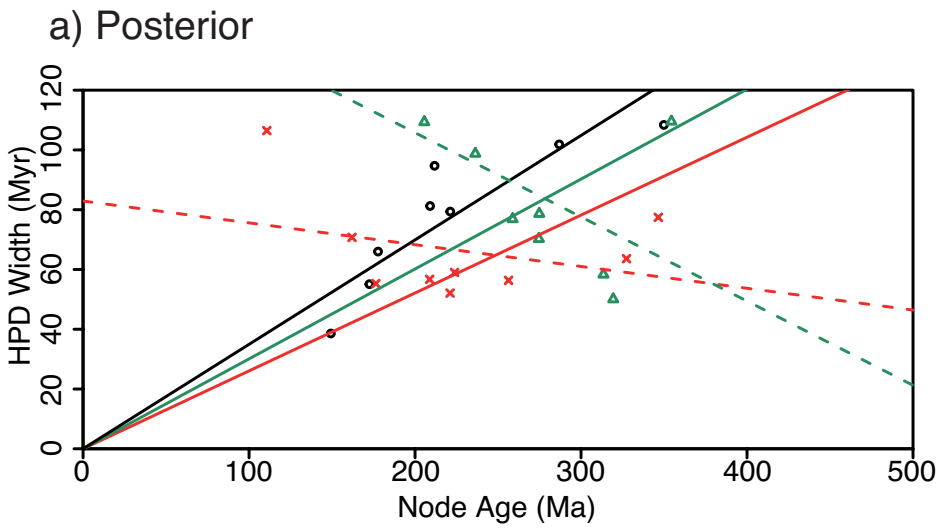
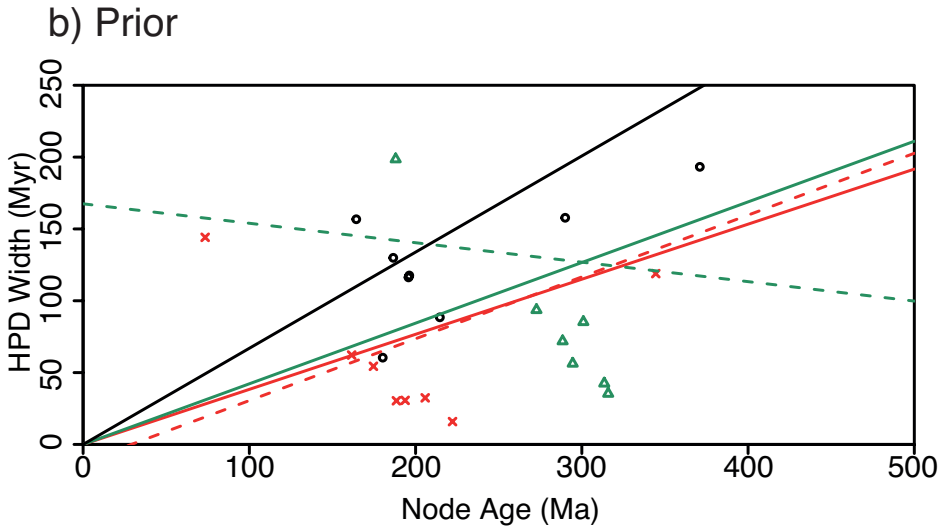


Figure 4.1 - Time calibrated phylogenies of Hymenoptera based on: (a) tip calibration; (b) node calibration; and (c) combined tip and node calibration. (a) and (c) are presented with fossil taxa removed, complete topologies are presented in the supplementary materials. Graduated bars represent the prior and posterior distribution of clade age, with colour density correlated with probability. Polytomies reflect topological uncertainty in the tree sample and are not indicative of simultaneous divergence. Coloured nodes indicate the position of the 9 clades of interest across the three topologies. Black (Neoptera), grey (Holometabola), white (Hymenoptera), yellow (Vespina), red (Apocrita), purple (Tenthredinoidea), blue (Xyelidae), turquoise (Pamphilioidea), green (Siricoidea).





c) R^2 values for linear models

	Node Calibration$_0$	Tip Calibration$_0$	Tip Calibration	Combined Calibration$_0$	Combined Calibration
Posterior	0.987	0.801	0.858	0.817	0.105
Prior	0.924	0.574	0.009	0.703	0.334

Figure 4.2 - Infinite-sites plots (1) for three alternative calibration approaches for both the posterior (a) and prior (b) distribution of times of 9 clades. Solid lines represent the fitted linear model for each independent set of node ages when forced through the origin, as in (1). Dotted lines represent the linear model for non node-calibrated analyses when not forced through the origin, demonstrating the lack of a linear decrease in clade age confidence interval width.

When tip- and node-calibrations are combined (figure 4.1c), the effective priors on node ages encompass dates younger than the minimum palaeontological constraints on the ages of crown Pamphilioidea and crown Xyelidae; in all other clades the effective priors and posterior age estimates fall fully within their palaeontological node age constraints. In all but the two deepest nodes the mean of posterior estimates of clade age are consistently and significantly younger than their counterparts when only tip calibrations are implemented. The distributions of posterior estimates of clade age are also more precise than their tip-calibrated counterparts in all but two most basal clades.

4.5 - Discussion

It has been accepted generally that, because user specified node-age priors are truncated in construction of the joint time prior, the effective prior should be assessed to determine whether it is consistent with the palaeontological constraints (Warnock et al., 2012). Our results indicate that this approach should be extended to tip-calibration. Tip calibrations consistently yielded older effective priors on node ages and older divergence time estimates. This occurs principally because of an absence of constraints on the ages of internal nodes within the tree, normally provided by node calibrations, allowing uncertainty to propagate from the tips, constrained only by the prior on the root age, skewing the distribution of prior probability toward ancient ages. We cannot conclude that these estimates are inaccurate merely because they are incompatible with palaeontological maximum age constraints. However, the effective priors derived from tip-calibration of some node ages are younger than their palaeontological minimum age constraints, which is unreasonable. This occurs because some crown clades (Xyelidae, Pamphilioidea) in the tree sample are often resolved without fossil members and so their minimum ages are bounded only by the Recent.

The node-calibrated analysis is compatible with the palaeontological constraints on clade ages because they are implemented as node calibrations. However, the combined node and tip-calibrated analyses yielded younger effective priors and posteriors than exclusively tip or node-calibrated analyses, while also conforming to the palaeontological minimum constraints. This is clear in the case of Siricoidea where no fossil member of the crown clade is represented but the zero-time constraint on the age of this clade in the tip-calibrated analysis is supplemented by a node age constraint in the combined tip- and node-calibrated analysis. The divergence time estimates derived from combined calibration are consistently younger - a consequence of the tip calibrations which act to truncate the broad priors of the node calibrations, extending from their hard minimum age constraints. This serves to draw the effective prior probability closer to the minima in the joint time prior, which propagates to the posterior divergence time estimates. In

effect, the tip- and node-calibrations interact to operationally establish maxima for the node calibrations.

It is reasonable to question whether tip- and node-calibrations should be implemented together and, certainly, the same data should not be represented in both calibration methods. However, there is no logical inconsistency between these approaches and some fossil data is better represented as a tip- or a node-calibration. While it has been argued that tip-calibration facilitates the inclusion of all fossil species in divergence time analyses (Ronquist et al., 2012a, Pyron, 2011), some fossil taxa are too incomplete to be effective tip-calibrations, but may be no less definitive in circumscribing the minimum age of a clade (e.g. the minimum age of angiosperms and echinoderms are constrained by tricolpate pollen and fragments of stereom, respectively).

A casualty of the implementation of node calibrations in MrBayes is the ability to perform coestimation of time and topology, a particular advantage of the tip-calibration approach (Ronquist et al., 2012a). However, fossil taxa are not commonly well-resolved through coestimation, a consequence of the paucity of morphological data and the non-random distribution of missing data for fossil species (O'Reilly et al., 2015). These challenges may be overcome simply by introducing a backbone of partial topological constraints, facilitating coestimation, but within the qualified phylogenetic uncertainty that is associated with most fossil species. Only BEAST is currently capable of fully accommodating this approach to combined calibration (Bouckaert et al., 2014). In our combined tip- and node-calibrated analysis we were forced to exclude any fossil species whose age overlapped or extended beyond the node calibration for the clade to which it was assigned. This limitation occurs because MrBayes unnecessarily considers ages for fossil species that can be older than their assigned clade, yielding a negative clock-rate and, therefore, an error when calculating the proposal ratio. Analyses employing the Fossilised Birth-Death Model (FBD; (Gavryushkina et al., 2014)) integrate fossil occurrences as data in coestimating time and topology, constraining node ages and, as such, they do not exhibit node age inflation seen in TED analyses that don't employ FBD. While we employ a total-evidence approach in our example here, combining node and tip-calibrations is also applicable to matrices consisting solely of fossil taxa and morphological characters.

4.6 - Conclusions

Nodes and tips are complementary, not competing approaches to the calibration of molecular clock analyses. Ancient age estimates have become synonymous with tip-calibrated analyses. The construction of the time prior itself is likely to be a causal factor. Our approach to

approximating the effective time prior in tip-calibrated analyses shows that when they are implemented alone, tip-calibrations can yield divergence time estimates that violate empirical fossil evidence or place exaggerated probability on overly ancient age estimates. Combining node- and tip-calibrations obviates these effects with the hard minima of node calibrations constraining the uncertainty associated with tip calibrations that, in turn, serve to objectively define the maxima of node age constraints. This approach is appealing because of the positive complementary interaction between the two classes of calibration, but also because it makes the best use of palaeontological data in the construction of evolutionary timescales.

Chapter 5

Isolating and Mitigating the Effects of Fossilization Processes on the Accuracy of Divergence Time Estimates

Joseph E. O'Reilly and Philip C. J. Donoghue

Authors' contributions - Both authors designed the study, interpreted the results and contributed to writing the manuscript; J.E.O'R. carried out the analysis and led the writing.

5.1 - Abstract - The importance of palaeontological data in divergence time estimation has increased with the introduction of Bayesian total-evidence dating methods which utilise fossil taxa directly for calibration, facilitated by the co-analysis of morphological and molecular data. Fossil taxa are invariably incompletely known as a consequence of taphonomic processes, resulting in the decidedly non-random distribution of missing data. The impact on age estimate accuracy exerted by this non-random distribution of missing data is unknown. To constrain the impact of biases in the taphonomic process on total evidence dating analyses, we compared clade ages estimated from a very complete morphological matrix to ages estimated from the same matrix permuted to simulate the progressive loss of anatomical information resulting from biostratinomic processes. We demonstrate that systematically distributed missing data negatively influences clade age estimates, but that successive stages within the biostratinomic process introduce greater levels of error in age estimation. We suggest that, in the absence of models that can explicitly account for the taphonomic process, morphological datasets should be constructed to minimise the impact of taphonomy on divergence time estimation.

5.2 - Introduction

Evolutionary timescales are essential to effect tests on the coevolution of Earth and Life and are necessary when testing many hypotheses rooted in historical biology. Molecular clock methodology has effectively displaced direct interpretation of the geologic and fossil record in this endeavour. Nevertheless, fossil data remain integral to divergence time estimation, in calibrating the rate of molecular evolution to time. This has traditionally been achieved indirectly through node calibration, using fossil and geologic evidence to constrain probabilistically the minimum and maximum age of clades (Parham et al., 2012). Tip-calibration overcomes challenges associated with the indirect interpretation of fossils to inform node-calibrations, allowing fossil taxa to inform molecular clock analyses directly, including them en par with their living relatives through the inclusion of a morphological dataset and model of evolution. This approach is particularly attractive since it allows all fossil species to be included in divergence time analyses, not just those informing the age of extant clades. Through co-estimation of time and topology in 'Total Evidence Dating' (TED; (Ronquist et al., 2012a)), the phylogenetic uncertainty of fossil species can be controlled for. However, fossil taxa are invariably incompletely preserved and the extent to which the taphonomic biases constraining the distribution of missing data impact molecular clock analyses has not yet been explored (O'Reilly et al., 2015).

The impact of missing data on the accuracy of topology estimates has been studied extensively, but almost all such studies have assumed that missing data is randomly distributed (Sansom and Wills, 2013). This is unlikely even for living species because of research biases resulting from the study of organ systems. It is also not expected in fossil taxa where biases in the processes of decay and preservation lead to progressive, but non-random, loss of biological information.

Long before the chemical processes of preservation and diagenesis occur, a prospective fossil may experience physical biostratinomic effects. These include decay, post-mortem collapse, disarticulation, fragmentation, erosion, and differential transportation (Behrensmeyer and Kidwell, 1985). The nature and degree of biological information loss is dependent on whether the effects of these processes can be diminished, such as through early burial and mineralization, and this is invariably dependent on the nature of the depositional environment (Behrensmeyer and Kidwell, 1985). At the most general level, soft tissue anatomy quickly rots away; only biomineralised anatomical elements are routinely fossilized, but even these have differential preservation potential, with lighter and more fragile skeletal elements most likely to be transported away, fragmented and eroded (Patoumathis, 1994). Hence, missing data are systematically distributed across fossil taxa, associated with specific classes of anatomical characters (Fig. 5.1). The non-random distribution of missing data has been shown to have a

biasing effect on topology estimation within a parsimony framework (Sansom and Wills, 2013), but its impact on tip-calibration and TED is unknown.

Here we explore the impact of non-random missing data on the accuracy and precision of divergence time analyses that employ tip-calibration. Traditionally, the effect of missing data has been investigated through simulation analyses in which the true tree is known and missing data is introduced randomly. However, designing simulations that approach the disparate fossilization process that vary both with intrinsic biology and the nature of the post mortem environment, is challenging. Instead we use a large, complete, empirical morphological matrix as the basis for our analyses in which we simulate the progressive loss of biological information that results from decay and biostratinomic processes, informed by empirical evidence for the proportional loss of anatomical characters. The ages estimated from the complete, unfossilized, dataset are used as a benchmark against which the accuracy and precision of divergence times estimated using datasets with artificial fossilization are measured.

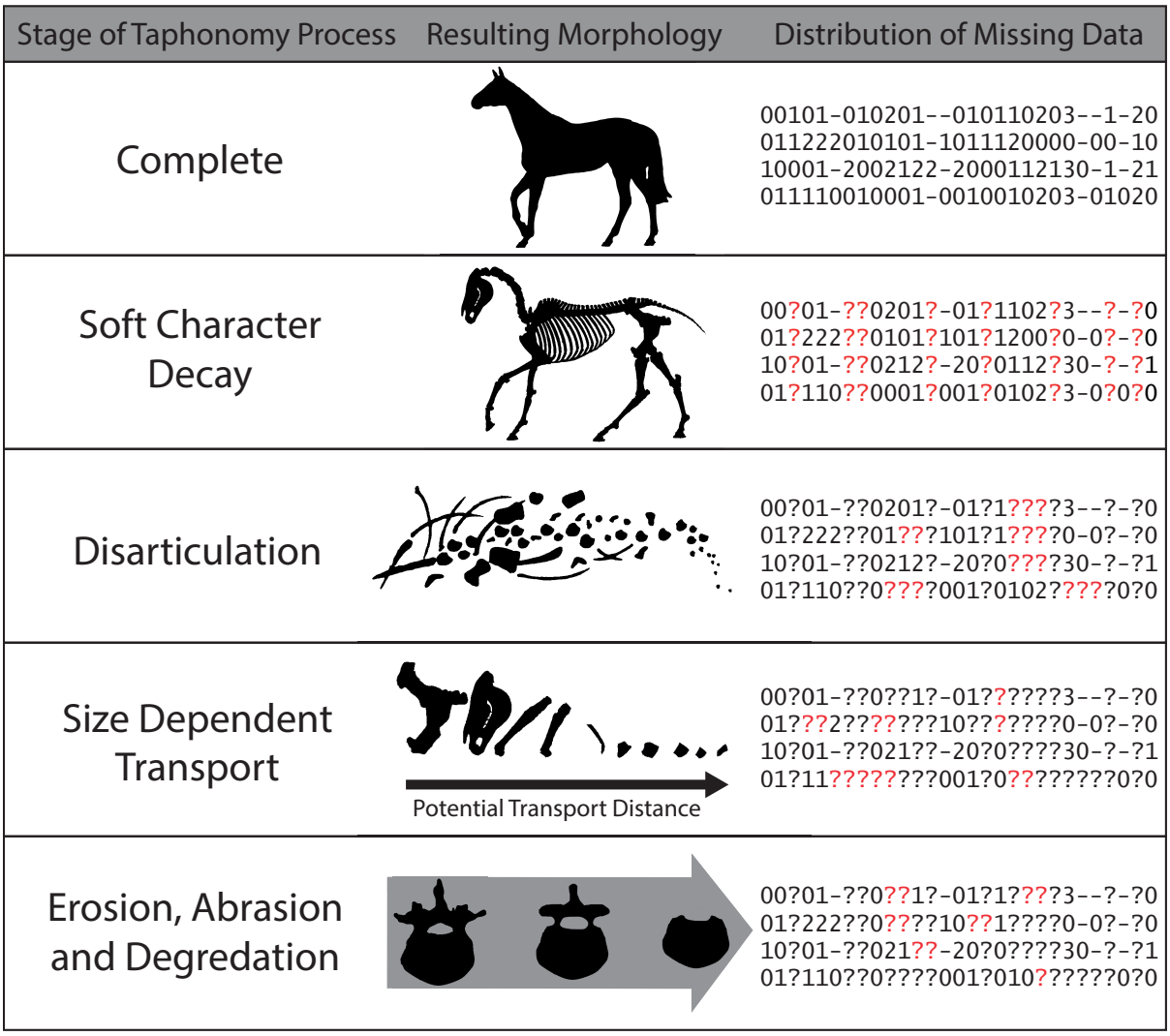


Figure 5.1 – The effect of different stages of the fossilization process on the distribution of missing data (characters in red font) in morphological matrices. Starting with a complete matrix before the influence of any stage of the fossilization process, the loss of soft characters through decay introduces a character-wise distribution of missing data as these characters are unlikely to preserve for any fossil taxon. Physical biostratinomic processes introduce further missing data in a taxon-wise manner as a number of characters from morphological structures are lost simultaneously. Disarticulation and size dependent transport leads to taxon specific loss of large numbers of characters associated with lost morphological structures. Erosion and abrasion lead to further loss of fine grain characters that are worn away before deposition.

5.3 - Materials and Methods

Given our goal, to determine the influence of non-random missing morphological data in molecular clock analyses, we required a dataset with a sufficiently large quantity of characters, such that branch lengths could be estimated with accuracy before and after missing data is introduced through artificial taphonomic biases. We used the 4541 morphological character dataset from (O'Leary et al., 2013), which encompasses the diversity of placental mammals, as the basis for our experiments.

To prepare the data to simulate the effects of the fossilization process we subsampled the data matrix, removing fossil taxa that are not members of crown-Mammalia, converted polymorphisms and ambiguities to missing data, and removed characters and taxa that have more than 25% missing data (inapplicable characters were considered as scored). This resulted in a dataset comprising 2454 characters and 66 taxa. Using this approximately complete dataset, we adopted two approaches to simulating the effects of fossilization: (i) decay of soft tissues, and (ii) later biostratinomic processes of disarticulation, transport, abrasion, erosion, fragmentation, etc).

5.3.1 - Simulating the Loss of Soft Tissue Anatomical Data

19% of the 4541 characters in the original (O'Leary et al., 2013) matrix are unlikely to be preserved through routine modes of fossilization. Thus, we selected at random 19% of the characters in the 2454 character matrix and changed their codings to missing data for all fossil taxa. This process was repeated to obtain 30 distinct matrices.

5.3.2 - Simulating the Effects of Post-Decay Biostratinomic Processes

Aslan and Behrensmeyer (1996) identified categories of mammalian osteological characters that are likely to be preserved in a fluvial deposit (the environment in which the majority of terrestrial vertebrates are preserved), with a probability of recovery reported for each category. The O'Leary dataset was reduced to only those characters relevant to the categories identified by Aslan and Behrensmeyer (1996), resulting in a dataset of 2454 characters. Using a bespoke script, we randomly sampled a selection of extinct taxa and converted codings to missing data within each class of characters, such that the amount of missing data was equal to the profile of the taphonomic filter. This approach introduced ~30% missing data into the matrix. To investigate the influence of increasingly extreme levels of missing data, we multiplied the quantity of missing data in each character category by 1, 1.4, 1.8, and 2. With a scaling factor of 2, we derived a matrix with 95% missing data for fossil taxa. A survey of ~250 Mammalian morphological matrices (Guillerme and Cooper, 2016) showed an approximately uniform distribution of missing data from 0 to 99%, with a mean of 37%. The process was repeated to

produce 30 matrices for each scaling factor in which a different random subsample of taxa was selected within each character class.

5.3.3 - Exploring the Effects of Character Matrix Dimensions

Though our experimental dataset has less than half the number of characters of the original (O'Leary et al., 2013), the remaining 2454 characters comprise an unusually large empirical phenotype dataset. Thus, we also explored the effects of taphonomic biases in datasets that approximate the dimensions of the majority of phylogenetic datasets, which are typically composed of characters measured in the low hundreds. To achieve this, we produced 5 smaller matrices in which each character class is reduced to a random subsample of 10% from the original matrix, resulting in 245 characters. For each of the 5 reductions, 30 replicates were performed. This approach was applied to explore the impact of smaller character sets in combination with simulations of the effects of soft tissue decay and later biostratinomic processes.

5.3.4 - Divergence Time Estimation

Divergence time analyses based on these matrices were performed using MrBayes 3.2 (Ronquist et al., 2012b). In all analyses, we employed a fixed topology compatible with the results of the combined molecular and morphological analysis of (O'Leary et al., 2013). We were not interested in the correctness of the fixed topology but, rather, use it to isolate the effects of the introduction of non-random missing data, without the confounding effect of topology estimation. Node calibrations were taken from (Benton et al., 2015) and assigned offset exponential distributions; tip calibrations were assigned uniform distributions, or point estimates, based on the ranges or single point ages reported in (O'Leary et al., 2013). A diffuse prior on the morphological rate was set as $N(0.001, 0.01)$ and is based on the prior for morphological rate of (Beck and Lee, 2014). The IGR clock model (Lepage et al., 2007) was used and assigned a variance parameter prior of $exp(10)$. Morphological data was analysed with the Mkv model (Lewis, 2001) and the uniform tree prior was applied. The mcmc approximation of the posterior distribution was performed for 40,000,000 generations, sampling every 4000^{th} over 4 runs of 4 chains. Convergence was assumed when effective sample sizes of greater than 200 were observed and with qualitative assessment of the stationarity of the chain in Tracer (Rambaut et al., 2014). Consensus trees were constructed after a conservative burn-in of 25%, after which sampling was deemed to be from the posterior distribution.

5.3.5 - Averaging Over Consensus Trees

For each analysis, 30 replicate majority-rule consensus trees are produced. We take the 30 consensus trees and for each constituent node we obtain the mean age estimate over the 30

replicates, producing a single tree that encompasses the results from 30 replicate analyses. These averaged trees can then be compared to a consensus tree estimated from the untreated matrix that their respective fossilised matrices were produced from. We also consider aspects of the distribution of the 30 individual replicate trees before averaging over their structures. We did not consider replicates with negligible interior branch lengths, as the subtending node is collapsed into a polytomy.

5.4 - Results

5.4.1 - Loss of Data Resulting From the Decay of Soft Tissue Characters

Full matrix. — Differences in node age estimates between the full matrix and treated matrix replicates are small, with the range of differences in node age estimates across all 30 replicates extending from -4.75 to +3.47 Myr. We use the mean absolute node age error, obtained by averaging over the individual replicate consensus trees, to compare the differences between estimated node ages for untreated and treated matrices. The mean absolute error over 30 replicates was 0.58 Myr. Node age estimate error appears to be normally distributed around -0.13, meaning that there is no obvious directionality in the error in age estimates (Fig. 5.2b). When there is error in a node age estimate it is just as likely to be underestimated as overestimated. Similarly, there appears to be no pattern to the distribution of node age error across the tree, with no obvious association between error and the proximity of fossil taxa (Fig. 5.2a).

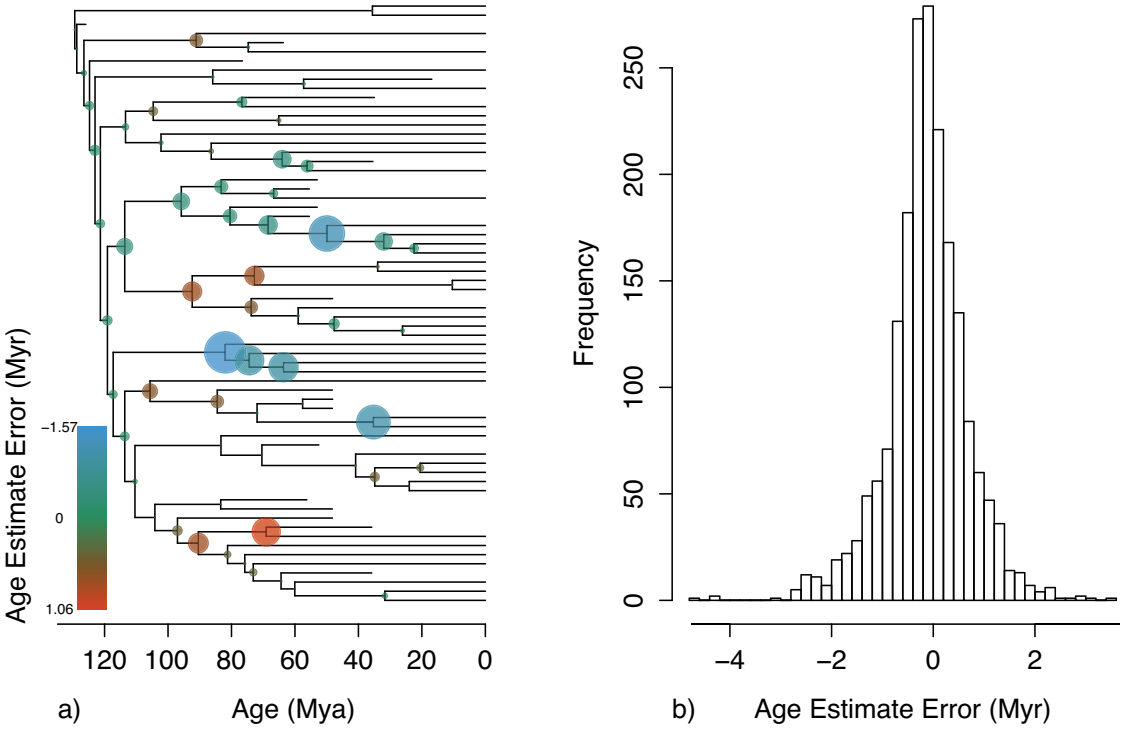


Figure 5.2 - (a) Mammalian time scaled phylogeny with node colour indicating the difference between mean estimated ages obtained from the full 2454 character untreated matrix and average ages obtained from 30 replicate treated matrices when the distribution of missing data approximates the distribution expected when soft characters are lost due to degradation and decay. The size of the node label is proportional to the magnitude of the error associated with that node age estimate. (b) Histogram of node age estimate error across all 30 replicate analyses.

Reduced matrices. — Over the 30 replicates of each of the 5 separate reductions of the full matrix, estimated node age differences between the untreated and treated matrices range from -9.98 to +19.33 Myr. Although, the majority of the replicates for each of the five reductions produced node age error ranges comparable to those seen in the full matrix (Fig. S5.1a-e). This variability reflects the heterogenous distribution of phylogenetic information in morphological matrices. The mean absolute error over all 30 replicates for 5 reductions is 1.08 Myr, which is around twice the mean absolute error obtained from artificial soft character fossilisation of the full matrix. Error is normally distributed, in this case around a mean of 0.17 Myr (Fig. S5.1f). The distribution of node age error across the tree is random as there appears to be no association between the proximity of fossil taxa and age estimate error. These results are broadly the same as those obtained from a full matrix in which only soft characters have been lost.

5.4.2 - Loss of Data Resulting from the Effects of Later Biostratinomic Processes

Full matrix. — As the scale factor is increased, the range of node age estimate differences between the treated and untreated matrices also increases, reaching a maximum of -22.80 to +14.68 Myr. With the smallest scale factor of 1, the node age error ranges from -19.92 to +6.97 Myr across all 30 replicates; this interval is greater than those seen when soft character treatment was applied to the full matrix. The mean absolute error also increases with scale factor (1.5, 3.1, 3.3, 3.4 Myr), demonstrating that this source of data loss introduces an increase in age estimate error. These measures of error are between 3 and 6 times greater than those obtained from the analysis of artificial soft character fossilisation of the full matrix. Across all scale factors, node age error is broadly normally distributed (Figs. S5.3f, S5.4f, S5.5f.), demonstrating a general lack of directionality in the error in age estimates. Similarly, there appears to be no relationship between proximity of fossil taxa and branch length estimate error (Figs. 3, S5.3a-e, S5.4a-e, S5.5a-e.).

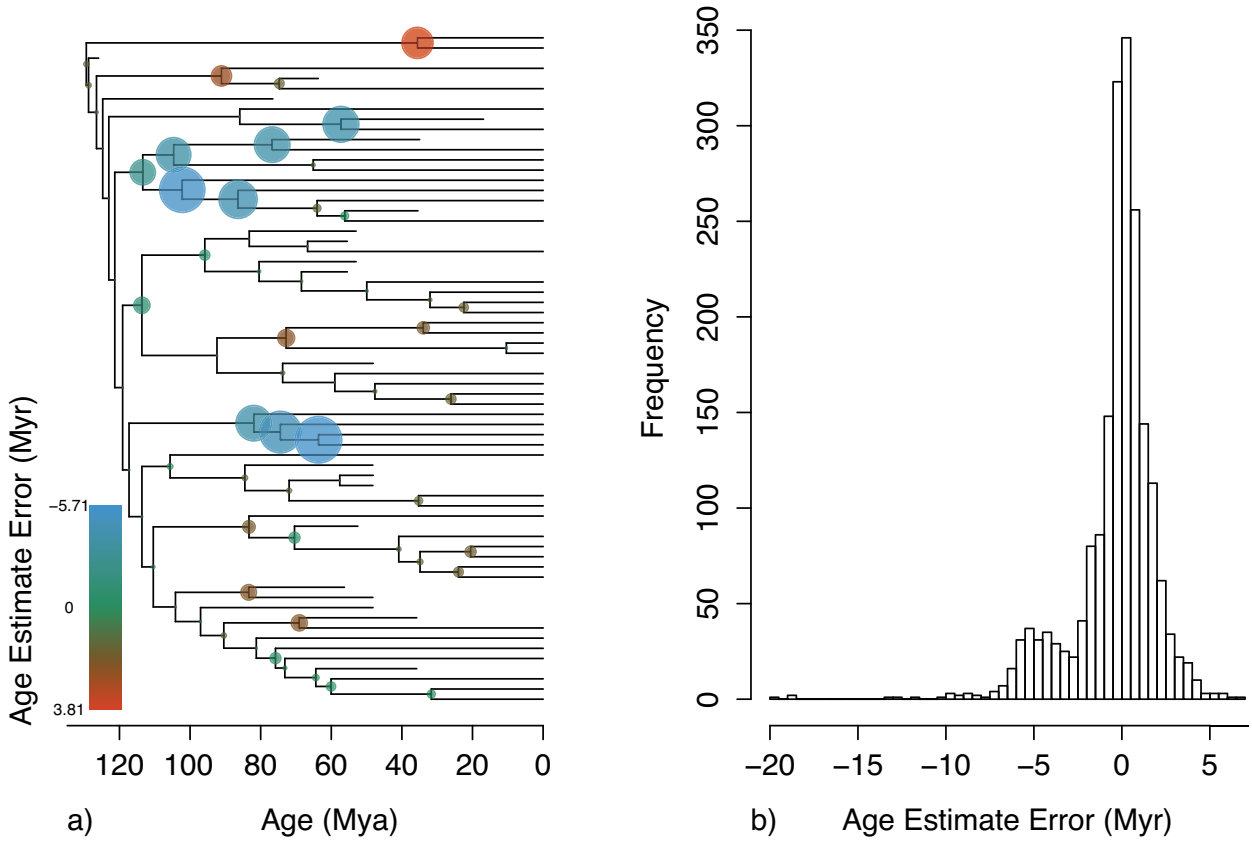


Figure 5.3 - (a) Mammalian time scaled phylogeny with node colour indicating the difference between mean estimated ages obtained from the full 2454 character untreated matrix and average ages obtained from 30 replicate treated matrices when the distribution of missing data approximates the distribution expected when characters are lost in blocks due to physical biostratinomic processes (scale factor =1). The size of the node label is proportional to the magnitude of the error associated with that node age estimate. (b) Histogram of node age estimate error across all 30 replicate analyses.

Reduced matrices. — Over 5 separate subsamples of the full matrix, the range of estimated node age differences between the untreated and treated matrices broadly increase as the scale factor is increased, with a maximum range of -25.94 to +42.50 Myr at a scale factor of 2, which is a smaller interval than obtained when the full matrix was analysed. With the smallest scale factor of 1, the node age estimate error ranges from -21.32 to +29.90 Myr, which is a greater interval than that obtained from the corresponding analyses of the full matrix. The mean absolute error increases with scale factor up to a scale factor of 2, where it decreases slightly (2.99, 3.19, 3.90, 3.20 Myr). This is possibly due to the disposal of replicates due to the presence of polytomies in consensus trees. Error is normally distributed around means of (0.93, 0.73, 2.02, 0.73) for the 4 scale factors, and across all analyses the distribution of error appears to have no relationship to the proximity of fossil taxa (Figs. S5.2f, S5.6f, S5.7f, S5.8f).

5.5 - Discussion

Complex taphonomic and biostratinomic processes control the distribution of missing fossil morphological data. Our results show the relative influence of two key stages in these processes on the accuracy of clade age estimation: (i) the loss of soft tissue characters due to decay and degradation, and (ii) the combined effect of physical biostratinomic processes such as disarticulation and transport.

5.5.1 - The Impact of Data Loss Resulting From Decay of Soft Tissue Anatomy

It is evident from our results that the effect of soft tissue character loss on divergence time estimation is negligible, with error in branch length estimates of approximately +/- 1 Myr irrespective of the size of the analysed matrix. This resilience to missing data is possibly due to the fact that extant taxa remain scored for these characters, and this influences the interpretation of missing data in fossil taxa in the posterior distribution. The ability of observed data to influence the interpretation of unobserved data is obviously reliant on the quality, distribution, and proportion of observed data relative to missing data. The distribution of fossil taxa in the analysed matrix is fairly even, allowing interspersed extant taxa to inform parameter estimates across the tree. An uneven distribution of fossil taxa is likely to amplify the deleterious effects of missing soft character data.

5.5.2 - The Impact of Data Loss Resulting From Post-Decay Biostratinomic Processes

Missing data resulting from physical biostratinomic processes introduces a marked reduction in divergence time estimate accuracy. One simple reason for this is that a greater proportion of data are lost because of post-decay biostratinomic processes, as opposed to loss through the decay of soft tissue characters alone. The levels of missing data in our treated matrices are based on empirical observations (Guillerme and Cooper, 2016), with soft character loss

affecting proportionally far fewer characters than other taphonomic processes, but the distinct difference in accuracy appears to be more than would be expected given the difference in the number of scored characters. Another factor that is likely to amplify the effect of blocks of missing fossil data is the non-independence of morphological characters. While relationships between characters are not explicitly accounted for in this Bayesian framework, the loss of an entire suite of characters may still exert a strong influence over the accuracy of divergence time estimates. Our results suggest that extant taxa can compensate for random distribution of characters consisting entirely of missing data in extinct taxa, but the systematic loss of a number of related characters cannot be accommodated as readily. The tip-calibration framework now facilitates divergence time analyses consisting entirely of fossil taxa; for such analyses the beneficial effects of well scored extant taxa will be missing and therefore the negative influence of missing data introduced by physical biostratinomic processes is likely to be exacerbated.

Both the process of soft-character decay and the effects of biostratinomic processes will vary greatly between different clades due to differences in morphology and depositional environment. For example, there are several phyla with a poor fossil record due to a paucity of hard parts, these taxa will possess a higher ratio of soft to hard parts and are likely to be more susceptible to errors caused by a loss of soft characters. Similarly, taxa that possess many easily transportable characters are likely to experience an increase in the effect of biostratinomic processes. Therefore, the effects of taphonomic biases presented here cannot be directly applied to other clades as the simulation process is explicitly built on a study of the taphonomic process in Mammals. For this reason, the effects of the taphonomic process on the accuracy of estimated divergence times in other clades requires direct investigation.

5.5.3 - Mitigating Against the Effects of Missing Data

Mitigating against the impact of these biases in the taphonomic process could be achieved by modifying the phylogenetic model to explicitly account for taphonomic processes, or by altering the qualities of the data exposed to the model. However, our results suggest that the simplest solution may be to subsample datasets to minimise the number of characters that are coded for only a subset of fossil taxa. Eliminating characters may strongly limit the statistical power of the remaining phenotypic data**,** but such an approach could be used to minimise the number of characters thought to possess a distribution of missing data constrained primarily by physical biostratinomic processes. The positive influence on divergence time estimate accuracy introduced by this subsampling method has been demonstrated in our analyses, with the case in which the distribution of simulated missing data is constrained exclusively by soft tissue decay being analogous to an empirical dataset subsampled to minimise physical biostratinomic

effects. The alternative to subsampling, development of a model of phylogenetic inference that accounts for taphonomic and biostratinomic processes, is decidedly non-trivial due to the disparity of fossilization pathways that vary with intrinsic biology and the extrinsic environment of fossilization.

5.6 - Conclusion

Missing data in morphological matrices is distributed systematically, with certain character types more likely to be missing than others in fossil taxa. Using a large paleontological dataset and empirically derived distributions of simulated missing data, we have demonstrated the relative influence of missing morphological data distributed according to different biostratinomic and taphonomic processes. The degradation and decay of soft characters appears to introduce little error into age estimates, whereas the loss of characters due to physical biostratinomic processes is likely to significantly impact estimated clade ages. Mitigation against this effect may be achievable by subsampling matrices such that the distribution of missing data introduced by physical biostratinomic processes is minimised.

Chapter 6

The Efficacy of Consensus Tree Methods for Summarising Phylogenetic Relationships from a Posterior Sample of Trees Estimated with Morphological Data

Joseph E. O'Reilly and Philip C. J. Donoghue

Authors' contributions - Both authors designed the study, interpreted the results and contributed to writing the manuscript; J.E.O'R. carried out the analysis and led the writing.

6.1 Abstract - Consensus trees are required to summarise trees obtained through MCMC sampling of a posterior distribution, providing an overview of the distribution of estimated parameters such as topology, branch lengths and divergence times. Numerous consensus tree construction methods are available, each presenting a different interpretation of the tree sample. The rise of morphological clock and sampled-ancestor methods of divergence time estimation, in which times and topology are co-estimated, has increased the popularity of the maximum clade credibility (MCC) consensus tree method. The MCC method assumes that the sampled, fully resolved topology with the highest clade credibility contains an adequate summary of the most probable clades, with parameter estimates from compatible sampled trees used to obtain the marginal distributions of parameters such as clade ages and branch lengths. Using both simulated and empirical data, we demonstrate that MCC trees, and trees constructed using the similar maximum *a posteriori* (MAP) method, often include poorly supported and incorrect clades when summarising diffuse posterior samples of trees. We demonstrate that the paucity of information in morphological datasets contributes to the inability of MCC and MAP trees to present an accurate summary of the posterior distribution. Conversely, majority-rule consensus (MRC) trees report a lower proportion of incorrect nodes when summarising the same posterior samples of trees. Thus, we advocate the use of MRC trees, in place of MCC or MAP trees, in attempts to summarise the results of Bayesian phylogenetic analyses of morphological data.

6.2 - Introduction

Recently developed tip-calibration methods facilitate the inclusion of fossil species and morphological data alongside living species and molecular data for phylogenetic and divergence time estimation, making use of all available relevant data. Indeed, these methods have ignited interest among palaeontologists in estimating true evolutionary timescales, even for entirely extinct clades for which only morphological data are available (Bapst et al., 2016, Matzke and Wright, 2016). Thus, these methods have the potential to revolutionise our understanding of evolutionary history and unite the hitherto disparate disciplines of palaeontological and molecular phylogenetics. However, interpreting the results of such analyses is complicated because morphological data are expected to often yield a very diffuse posterior sample of topologically disparate trees that are difficult to reconcile meaningfully in a single consensus tree.

Several methods are available to summarise the results from Bayesian posterior tree samples. A straightforward approach to representing the posterior distribution of trees is to choose a single tree from the sample of trees that can be considered optimal by maximising some criterion of support. One such approach is to use the single sampled topology with the greatest posterior probability, the maximum *a posteriori* tree (MAP. As the posterior probability of a tree is the joint probability of the model parameters as well as the tree topology (both conditioned on the data), it is possible that MAP trees will have strongly supported non-topology parameter estimates on an incorrect topology. Another sampled tree method, Maximum Clade Credibility (MCC) is less susceptible to this source of error as it considers the distribution of clade support in the posterior sample of trees. The MCC method has become one of the most popular consensus methods for summarising tree samples obtained in tip-calibrated analyses in which morphological data are analysed (Pyron, 2011, Wood et al., 2013 , Dornburg et al., 2015, Dembo et al., 2015, Dembo et al., 2016, Herrera and Davalos, 2016, Matzke and Wright, 2016). This popularity is perhaps because it yields consensus trees that are highly resolved unlike more conventional methods, such as Majority Rule Consensus (MRC), or alternatively because it is the default method in the popular TreeAnnotator consensus tree construction software package. The MCC method identifies the single tree in the posterior sample with the largest sum (or alternatively, product) of posterior probabilities across its constituent bifurcations (Heled and Bouckaert, 2013). Like the MAP tree, this tree will not explicitly account for topological uncertainty in its structure, as each sampled tree in the posterior sample is invariably fully resolved. As with the MAP tree, the fully resolved nature of the MCC tree is appealing, but it may be a poor summary of the posterior distribution of trees, as it has the potential to include clades with low posterior probabilities that are poorly supported by the data. The inclusion of such clades in MCC trees is likely to be caused by morphological data which often yield a

diffuse sample of topologically disparate trees from the posterior distribution, principally due to a relative lack of phylogenetic information distributed across a matrix consisting of few characters (Gelman et al., 2013, Steel, 2013). Thus, MCC trees based on morphological data, have the potential to over-represent clades with low posterior probabilities which are, by definition, poorly supported by the data and, therefore, likely to be spurious.

The MRC tree method offers an alternative approach to summarisation by sacrificing potentially unjustifiable precision for topological accuracy. MRC trees present divergence times on a set of well-supported (posterior probability > 0.5) bifurcations, or soft polytomies, in the presence of uncertainty. Such a conservative approach to presenting topological uncertainty may be desirable, particularly in a Bayesian framework in which obtaining the marginal posterior distribution of model parameters results in the explicit estimation of their uncertainty.

For morphological clock analyses, the accuracy of the consensus tree topology upon which clade ages are presented is integral to the accuracy of the reported timescale. This is because the marginal distributions of clade ages are constructed from only the trees in the sample that are compatible with the consensus topology (Heled and Bouckaert, 2013). Reporting ages for spurious clades is obviously problematic and will have a significant impact on interpretations of evolutionary history. The consensus tree used to summarise the posterior distribution must, therefore, minimise incorrect clades while also maximising the inclusion of correct clades. Here, using simulated datasets containing variable levels of phylogenetic information, we demonstrate that the increased variance of a finite posterior sample of trees obtained from morphological data is often poorly summarised by MCC and MAP trees which often include incorrect clades. We also show that MRC trees outperform MCC and MAP trees in summarising diffuse posterior distributions, presenting a more conservative summary of topology, including fewer incorrect clades. Finally, by analysing several empirical data matrices that are expected to carry varying levels of observed information about the same set of divergences, we demonstrate that MCC and MAP trees are likely to be inappropriate when summarising posterior samples of trees obtained from empirical morphological data.

6.3 Materials and Methodology

We simulated matrices that exhibit varying levels of phylogenetic information, performed co-estimation of divergence times and topology on these matrices, and then constructed MCC, MAP, and MRC trees from samples of the posterior distribution. To simulate matrices with varying levels of observed information, we exploited the relationship between the quantity of independent and identically distributed (*i.i.d)* data drawn from the underlying process in question and the variance of the posterior distribution around the true value of parameter

estimates; this relationship is commonly termed consistency (Gelman et al., 2013). We can therefore assume that small matrices simulated with the standard morphological model and analysed with that same model will produce a more diffuse posterior distribution than larger matrices, and that larger matrices will contain more information about the distribution of parameters in the model.

6.3.1 - Simulated Matrices

All simulations were performed on an arbitrary 36-tip time scaled phylogeny containing 4 fossil taxa. The simulations used the Mk model of morphological evolution (Lewis, 2001), with 100 replicate matrices of either 100, 1000, or 10000 binary characters produced using this model.

6.3.2 - Empirical Matrices

We analysed 3 empirical matrices spanning common data sources for divergence time estimation. These matrices were obtained by splitting the total-evidence matrix from Ronquist et al. (2012a) into its constituent elements. The first empirical matrix consisted of the morphological characters from this analysis only, the second matrix consisted of the molecular characters only, with all fossil taxa removed; the final matrix was a total-evidence combination of both molecular and morphological characters for extinct and extant taxa.

6.3.3 - Divergence time estimation

For analyses of our simulated datasets, a posterior sample of trees was obtained using MrBayes 3.2 (Ronquist et al., 2012b). For simulated matrices, errorless point calibrations were applied to non-contemporaneous tips and a root calibration was applied as a gamma distribution with mean = 1 and standard deviation = 0.1. A strict clock was employed with a prior on the rate of $G(1,2)$. The Mkv model (Lewis, 2001) was used to analyse the simulated data. Four chains of Metropolis-coupled MCMC sampling were performed for one million generations. For empirical datasets, we performed analyses as in O'Reilly et al. (2015). Consensus trees were constructed for each posterior sample of trees after a 25% burnin, MCC trees were constructed in TreeAnnotator, MAP trees were taken from the MrBayes output files, and MRC trees were constructed by the `sumt` function in MrBayes.

6.3.4 - Consensus Tree Efficacy Tests

For both simulated and empirical matrices, we performed several tests of the ability of the MCC, MAP, or MRC tree to summarise the posterior distribution of sampled trees. Using a custom R script, we identified all bipartitions present in each individual tree in the posterior sample (post burn-in) for each replicate and then obtained the posterior probability for each of

these bifurcations. We use the number of unique bipartitions sampled from posterior distribution in each analysis to approximate the variance of the posterior distribution itself.

Effective MCMC sampling of the posterior distribution requires a Markov-chain with a stationary distribution of the posterior distribution, in addition to a finite number of samples that is large enough to accurately approximate the distribution. When MCMC sampling of the posterior distribution is performed, estimated parameter values are sampled with a frequency proportional to their posterior probability. Therefore, if the posterior distribution of a discrete parameter is highly concentrated around the true value of the estimated parameter, a small number of discrete parameter values close to the true value will be sampled regularly from a finite sample of the posterior distribution, with values further from the true value sampled infrequently. We therefore consider the number of unique sampled bifurcations as an acceptable approximation of the variance of the posterior distribution of topologies.

For simulated matrices, we obtained the number of clades in the MCC, MAP, or MRC tree that are not found in the generating tree, and are therefore incorrect. This was achieved by comparing the constituent taxa of each clade in the generating tree with the constituent taxa of each clade, whether defined by a bifurcation or a soft polytomy, in the consensus trees. If any of the clades in the consensus tree consisted of taxa that did not form a clade in the generating tree, they were considered incorrect. MCC and MAP trees will possess a constant number of constituent clades, whereas MRC trees may possess soft polytomies and will therefore present a variable number of clades. Therefore, we also calculated the proportion of nodes that are incorrect in MRC, MAP, or MCC trees, correcting for the resolution of the consensus tree. For each replicate we also subtracted the number of incorrect nodes from the number correct nodes for each consensus tree to obtain a score for the overall accuracy of the presented topologies.

For empirical matrices we also consider the degree of underrepresentation of clades with relatively high support across consensus tree construction methods. We use three different criteria to determine whether a sampled clade should be represented in the MCC or MAP tree but is missing from either, i.e. if it meets at least one of these criteria it is considered valid but unrepresented. The three separate criteria for a clade to be considered valid but unrepresented are: (i) possession of a posterior probability greater than that of most poorly supported clade in the MCC or MAP consensus tree; (ii) possession of a posterior probability greater than the mean posterior probability of clades in the MCC or MAP tree; or (iii) possession of a posterior probability greater than 0.5.

6.4 - Results

6.4.1 - Simulated Matrices

As the number of simulated characters was increased the posterior sample of trees became more concentrated, as reflected in the decreasing number of unique sampled bifurcations (Table 6.1). For all consensus tree methods, the mean and range of both the number and percentage of incorrect nodes decreased as the posterior distribution became less diffuse (Table6. 2). When the posterior sample of trees was at its most diffuse, with 100 characters, MRC consensus trees included far fewer incorrect nodes than MCC or MAP trees, whether expressed in absolute terms or as a proportion of the total number of clades in the consensus tree. MAP trees often contained more incorrect nodes than MCC trees (Table 6.2; Figs. 6.1,6.2,6.3). With 100 characters, MRC trees were never fully resolved, with trees containing a mean of 22 resolved nodes out of a possible 35 (63%), with a range of 16 to 29 (46% - 83%). With 100 characters, MCC trees possessed the most correct nodes in absolute terms (on average), with MRC and MAP trees possessing a similar number of correct clades to one another (Fig. 6.2). Conversely, MCC and MAP trees also included more incorrect nodes than MRC trees in both absolute and proportional terms (Table 6.2; Fig. 6.2). When the number of correct nodes is expressed as a proportion of the total number of resolved nodes in the consensus tree, MRC trees greatly outperformed MCC and MAP trees (Table 6.2; Fig. 6.2). When the number of incorrect nodes is subtracted from the number of correct nodes presented in each consensus tree, the MRC tree often exhibited a better total level of accuracy than MCC trees, which in turn often exhibited a higher level of accuracy than MAP trees (Fig. 6.2). Both MCC and MAP trees occasionally produced topologies with more incorrect nodes than correct nodes, as can be seen in the frequency of replicates with a value below than 0 in Figure 6.2.

The differences between the performance of MRC, MAP, and MCC trees diminished as the number of analysed characters increased and the posterior sample became less diffuse, with differences between the three methods becoming indistinguishable when 10000 characters were analysed (Fig. 6.1; Table 6.2.).

Table 6.1. Number of unique sampled bipartitions obtained from the posterior distribution for 100 replicate simulated datasets.

	Number of unique sampled bipartitions	
Number of characters	Mean	Range
100	3882.0	1010-9726
1000	208.7	177-246
10000	176.9	142-220

Table 6.2. Absolute number of incorrect clades in maximum clade credibility (MCC), majority rule consensus (MRC), and maximum *a posteriori* (MAP) trees constructed for posterior distributions sampled from 100 replicate simulated data sets. The percentage of nodes that are incorrect for each consensus tree method are presented in parentheses.

	MCC		MRC		MAP	
Num. of characters	Mean (%)	Range (%)	Mean (%)	Range (%)	Mean (%)	Range (%)
100	10.5 (30.1)	5-20 (14.3-57.1)	2.7 (11.9)	0-7 (0-35)	14.8 (42.3)	5-22 (14.3-62.9)
1000	1.69 (4.8)	0-5 (0-14.3)	1.03 (3)	0-4 (0-12.1)	1.72 (4.9)	0-5 (0-14.3)
10000	0.03 (0.1)	0-1 (0-2.9)	0.02 (0.1)	0-1 (0-2.9)	0.03 (0.1)	0-1 (0-2.9)

Table 6.3. Features of valid but unrepresented clades in MCC trees constructed from posterior distributions obtained using different data types.

Data Source	Number of Sampled Clades	Number of clades missing from MCC tree with pp > minimum pp of clades in MCC tree	Max pp of clades missing from MCC tree with pp > minimum pp of clades in MCC tree	Number of MRC Clades Not Presented in MCC Tree	Number of clades missing from MCC tree with pp > mean pp of clades in MCC tree
Molecular	256	13	0.55	1	0
Morphology	94660	41623	0.84	8	11
Total Evidence	29610	1372	0.78	3	1

2. Posterior probability (pp)

Table 6.4. Features of valid but unrepresented clades in MAP trees constructed from posterior distributions obtained using different data types.

Data Source	Number of clades missing from MAP tree with pp > minimum pp of clades in MAP tree	Max pp of clades missing from MAP tree with pp > minimum pp of clades in MAP tree	Number of MRC Clades Not Presented in MAP Tree	Number of clades missing from MAP tree with pp > mean pp of clades in MAP tree
Molecular	31	0.90	5	1
Morphology	41629	0.95	13	28
Total Evidence	5803	0.65	4	2

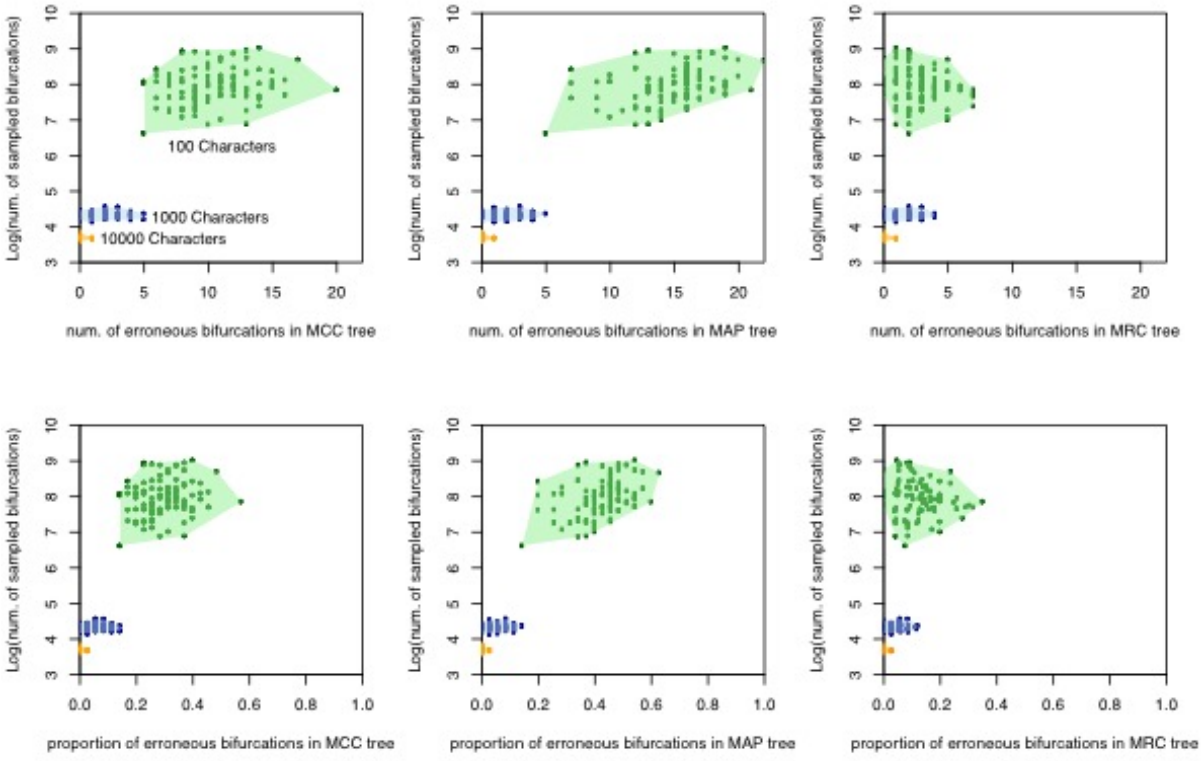


Figure 6.1. – The number and proportion of incorrect bifurcations in maximum clade credibility (MCC), maximum *a posteriori* (MAP), and majority rule consensus (MRC) summarisation of simulated data sets of different sizes plotted against the number of unique bifurcations sampled in each analysis. MRC trees present fewer incorrect nodes in both absolute and proportional terms, with MCC presenting fewer incorrect nodes than MAP

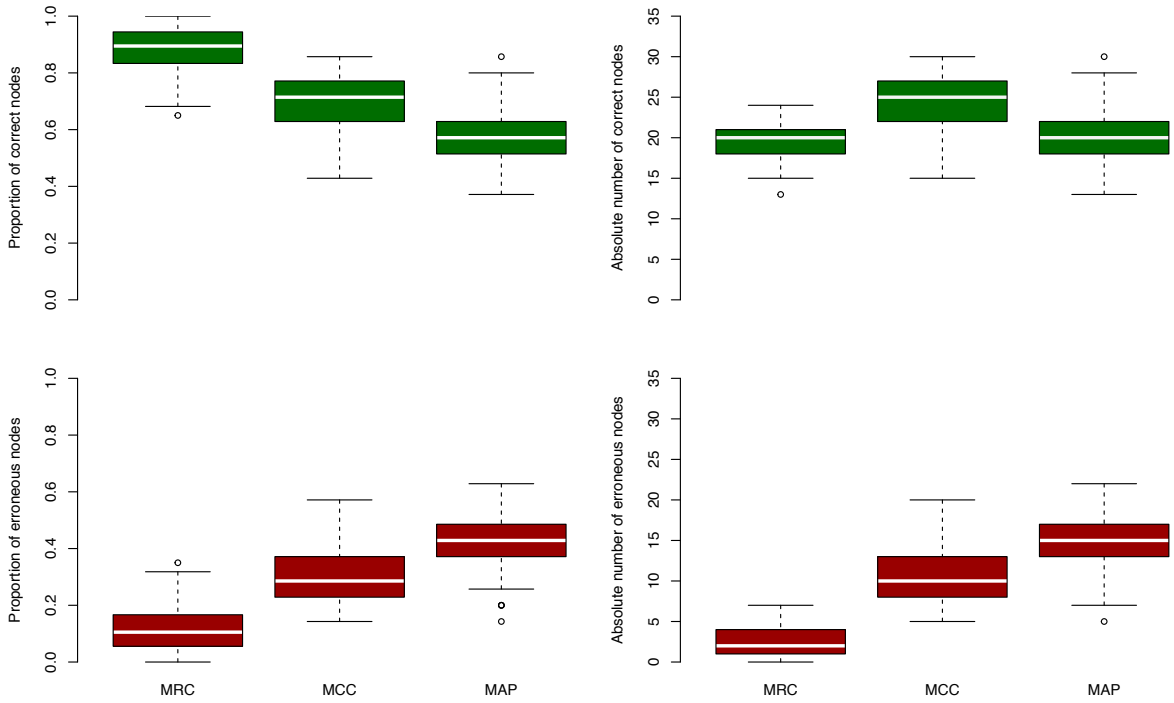


Figure 6.2. – The number and proportion of correct and incorrect clades presented in consensus trees constructed from 100 replicate matrices of 100 simulated characters. maximum clade credibility (MCC) trees often present more correct nodes than majority rule consensus (MRC) or maximum *a posteriori* (MAP) trees in absolute terms, but they also present many more incorrect nodes than MRC trees in absolute terms. MRC trees present proportionally more correct nodes and fewer incorrect nodes than either MAP or MCC trees.

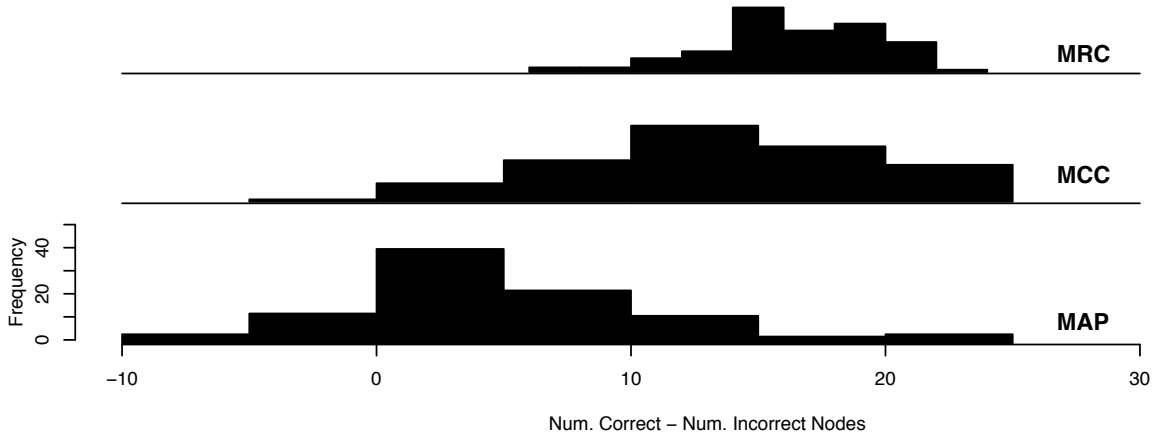


Figure 6.3. – The number of incorrect nodes found within each consensus subtracted from the number of correct nodes found tree within the same consensus trees constructed from 100 replicate matrices of 100 morphological characters. In several cases both MCC and MAP trees contain more incorrect nodes than correct ones.

6.4.2 - Empirical Matrices

The posterior sample of trees obtained using molecular data was the least diffuse of all three data types (Table 6.3). The addition of morphological data in the total-evidence analysis dramatically increased the number of sampled clades and, therefore, the diffusion of the posterior sample. It should be noted, however, that this analysis involves more taxa and will therefore naturally allow for more unique sampled bipartitions. Analysis of morphological data on its own produced the most diffuse posterior sample, as inferred from the number of unique sampled clades.

The number of valid but unrepresented clades derived from each set of posterior samples, increased as the diffusion was inferred to increase, with molecular data resulting in MCC and MAP trees with few overlooked clades and the smallest posterior probabilities for unrepresented clades (Table 6.3; Table 6.4). The addition of morphological data, or the analysis of morphological data alone, greatly increased the number of clades unrepresented in MCC and MAP trees.

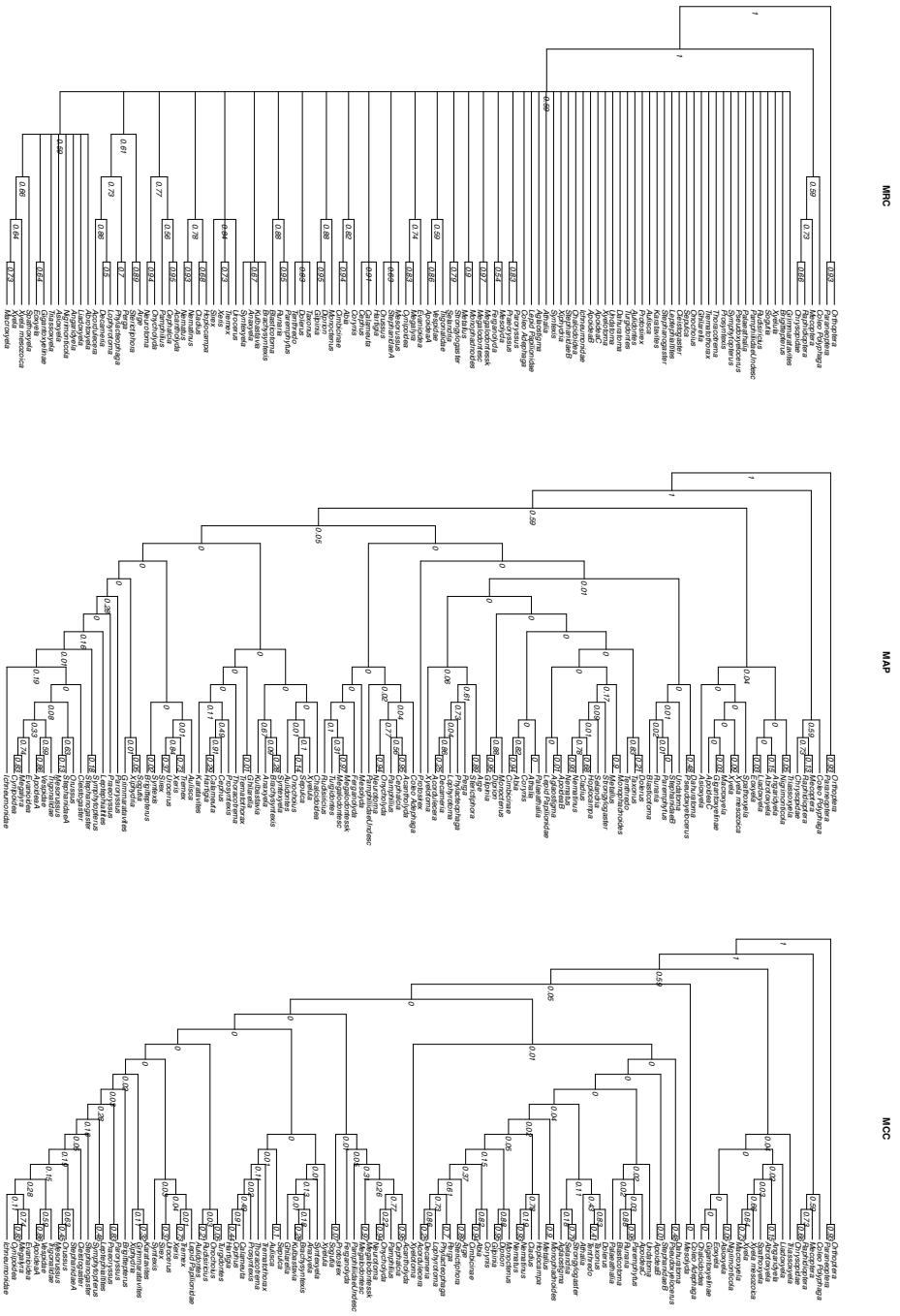


Figure 6.4. –Majority rule (MRC), maximum *a posteriori* (MAP), and maximum clade credibility (MCC) consensus trees for Hymenoptera constructed from a posterior sample of trees obtained when only morphological data was analysed. The trees are presented as cladograms with arbitrary branch lengths as summary of the posterior distribution of branch lengths results in numerous branches with negative length for the maximum clade credibility tree. The posterior probability of each clade is presented as a node label. In the MCC and MAP trees many clades express very low support and most are not consistent with molecular phylogenies or conventional understanding of Hymenopteran evolution. For the MRC tree few clades can be presented with support > 0.5, but those that are present are broadly congruent with established Hymenopteran phylogenetic relationships.

6.5 - Discussion

Using simulated data, we have demonstrated that when summarising a diffuse posterior sample of trees, MAP and MCC consensus trees often include many correct clades, but at the cost of the inclusion of large numbers of incorrect clades. Conversely, MRC trees present fewer clades in total, but those that are presented are very rarely incorrect. With 100 characters, there is a reduced level of information regarding the distribution of the parameters of the model. In this case, the MRC tree often outperforms the MCC and MAP tree at minimising the inclusion of both the absolute number and percentage of incorrect clades (Fig. 6.1; Fig. 6.3). With larger numbers of simulated characters there is increased information about the estimated model parameters and, therefore, the posterior distribution is expected to become more concentrated around the true value of the model parameters (Gelman et al., 2013). In these cases, MCC, MAP, and MRC methods obtain comparable levels of accuracy, as measured by the proportion of incorrect clades that are presented by these consensus tree methods (Fig. 6.1). These results suggest that when analysing matrices with large numbers of informative characters, the use of either the MCC, MAP or MRC method is justifiable, but when there is a paucity of information in the observed data, such as when morphological data are analysed, a more considered choice needs to be made. This choice must be based on whether the inclusion of an incorrect clade is considered worse than the omission of a correct clade, with the MRC tree likely minimizing the inclusion of incorrect clades, the MCC tree more likely to maximise the inclusion of accurate clades, and the MAP tree including fewer correct nodes than the MCC tree and more incorrect clades. The trade-off between type I and type II errors when choosing consensus tree construction methods has previously been explored in a decision theory framework, with the MRC tree considered the optimal topology if the inclusion of an incorrect clade is considered worse than the exclusion of a correct clade (Holder et al., 2008). The results we present here are congruent with this, and suggest that the optimal consensus tree for reporting parameter estimates from morphological datasets is likely to be the MRC tree.

Categorical morphological data can inform parameter estimates alone, or in tandem with molecular data, in a total-evidence approach (Pyron, 2011, Ronquist et al., 2012a). By using the number of unique sampled bipartitions as a proxy for the variance of the posterior distribution, we have shown that empirical analyses that utilise morphological data in both exclusive and total-evidence approaches, are likely to possess a markedly more diffuse posterior distribution than if molecular data from the same group is analysed alone, and are therefore less suitable for summarisation by MCC or MAP trees. Indeed, the MCC tree constructed from the posterior distribution obtained from morphological data alone contains few clades consistent with other published hymenopteran phylogenies (Fig. 6.4) (Klopfstein et al., 2015 , Zhang et al., 2016).

Conversely, the MRC tree contains few resolved clades, but those that are presented are broadly congruent with established phylogenetic relationships.

The inability of the MCC and MAP trees to recover a correct topology when summarising a diffuse posterior distribution is intimately linked to their inability to represent topological uncertainty in the posterior distribution with soft polytomies. A finite sample of trees from a highly concentrated posterior distribution will consist of only a small number of clades that have been visited frequently by the MCMC algorithm. In such circumstances, the probability of the set of best-supported clades appearing simultaneously in any sampled tree and, therefore the MCC or MAP tree, is high. Conversely, when the tree sample is diffuse the probability of the set of best-supported clades appearing simultaneously in any single sampled tree is reduced; increasing the chances that the MCC or MAP tree includes poorly supported and incorrect clades. This phenomenon is likely to be exacerbated as the number of taxa included in an analysis increases as this will also increase the number of possible bifurcations, reducing the chance of the best set of bifurcations appearing simultaneously. Alternatively, MRC trees represent frequently sampled bifurcations exclusively, collapsing topological uncertainty due to low support into soft polytomies, and naturally reducing the inclusion of spurious and infrequently sampled, poorly supported clades.

The relative efficacy of consensus tree construction methods has a large influence over the presentation and interpretation of divergence time estimates in molecular, morphological or total-evidence clock analyses. To obtain the marginal distributions on node ages, the tree sample is examined for clades that are compatible with the consensus tree and a vector of node ages is constructed for each clade in the consensus tree from these compatible sampled trees (Heled and Bouckaert, 2013). Therefore, it is likely that for morphological clock or total evidence dating analyses estimates of clade age are more likely to be presented for incorrect or poorly supported clades in MCC or MAP trees than in MRC trees. As MCC and MAP trees are more likely to present a spurious evolutionary timescale, the MRC tree can be considered the optimal topology upon which divergence time estimates obtained using morphological data should be reported.

6.6 - Conclusion

Using simulated data, we have demonstrated that MCC and MAP consensus trees often include more incorrect nodes than MRC trees when attempting to summarise particularly diffuse posterior distributions. With empirical data, we have shown that when morphological data are added to analyses, MCC and MAP methods have a propensity to include poorly supported and likely incorrect clades. These results suggest that MCC or MAP trees may be unsuitable for use

with phylogenetic methods that attempt to integrate morphological data, especially those in which parameter estimates are summarised on a consensus tree, such as divergence time estimation analyses.

Conclusions

The integration of morphological and molecular data in a probabilistic phylogenetic framework promises to refine estimates of both phylogenetic relationships and evolutionary timescales. Despite the obvious benefits offered by this framework there are several areas in which the manner in which morphological data is incorporated must be improved if we are to obtain parameter estimates of the greatest accuracy. Throughout this thesis several factors influencing both the accuracy of phylogenetic estimates obtained through the analysis of morphological data and of divergence times estimated in a total-evidence framework have been considered. In Chapter One the relative efficacy of the parsimony and Bayesian methods for phylogenetic reconstruction using morphological data are identified. This is achieved through the analysis of morphological data simulated to match the range of empirical homoplasy. Parsimony methods are shown to be more precise than Bayesian approaches to topology estimation, but this precision is achieved at the expense of accuracy. Conversely, Bayesian estimates of topology benefit from the explicit estimation of uncertainty, which can then be accounted for in a majority rule consensus tree.

The framework applied in Chapter One is built upon in Chapter Two to investigate the efficacy of phylogenetic reconstruction methods when the analysed data is simulated along trees with different levels of symmetry. Bayesian methods are again shown to outperform parsimony methods in terms of accuracy. It is also demonstrated that maximum-likelihood estimation of topology from morphological data suffers from the same reduced accuracy for higher resolution pay-off of parsimony methods. Using this result it was possible to demonstrate that a number of evolutionary hypotheses that are supported solely by parsimony analysis cannot be supported by Bayesian analysis. These results highlight the possibility that our understanding of many aspects of evolutionary history may well be inaccurate when built upon parsimony inference alone. In both Chapter One and Two the accuracy of estimated topologies are presented with no consideration of measures of support for the constituent nodes of these trees. Future investigation of the level of support assigned to incorrect clades estimated in different inference frameworks will hopefully allow for a more detailed understanding of the relative accuracy of phylogenetic reconstruction methods.

In Chapter Three an overview of the TED framework is provided, in addition to the highlighting of several issues presented by this framework that require investigation if we are to obtain the most accurate estimates of clade ages. It is also demonstrated that incorporation of the uncertainty in the age of fossil taxon calibrations influences age estimates. A method by

which fossil calibrations should be constructed to properly encapsulate the uncertainty surrounding the age of each fossil taxon using a process derived from the standard protocol for constructing node-calibrations is outlined. The question of what a tip calibration actually represents is still unclear though. Unlike node calibrations or sequentially sampled viral tip calibrations, where a calibration is constructed for a specific event, the identification of the event occurring at a fossil tip is unclear. If we consider a taxon with a long history of morphological stasis, i.e. the taxon exhibits the same morphology across many geological units, it is unclear if we should construct a calibration for the first appearance of this taxon or for the full extent of its appearance in the fossil record. If we wish to obtain accurate estimates of morphological evolutionary rate it seems that the presence of morphological stasis should be accounted for in analyses, particularly as we can consider the presence of stasis as observed data. The influence of different approaches to the calibration of tips in such cases requires further investigation if we are to improve rate estimates.

Unexpectedly ancient divergence times are often estimated in a TED framework (Ronquist et al., 2016). The construction of the time prior is likely to be a driving factor in such cases, with non-zero probability being placed on divergence times that are in direct conflict with fossil occurrence data. In Chapter Four a viable method is demonstrated for constraining the time prior in TED analyses such that the prior on node ages is congruent with empirical fossil evidence. This approach is achieved through a mixture of topology constraints, the uniform tree prior, and a combination of tip and node calibrations. A novel method for approximating the prior distribution on node ages in tip-calibrated analyses was required to achieve this. Since the publication of the work contained in this chapter there have been a number of developments regarding alternative tree priors, most notably the fossilised birth death process (Heath et al., 2014, Zhang et al., 2016). The fossilised birth death process provides a well-defined framework for constructing the time prior on node ages, explicitly accounting for the process of fossil sampling. Despite the obvious potential of this model, the congruency of the effective time prior induced by the fossilised birth death process and empirical fossil evidence has yet to be thoroughly addressed.

Missing data for fossil taxa in morphological matrices is likely to be systematically distributed due to biases in the fossilisation process (Sansom et al., 2010, Sansom and Wills, 2013). In Chapter Five it is demonstrated that the characteristic distributions of missing data introduced by different stages of the fossilisation process have different effects on the accuracy of divergence times estimated using morphological data. Through the use of an empirically guided simulation framework and a particularly complete morphological matrix it is demonstrated that the loss of soft characters due to decay introduces a relatively small amount of error into

divergence time estimates. Conversely, biostratinomic process such as transport and size sorting will introduce large amounts of error into age estimates. In this Chapter it is suggested that, in the absence of a meaningful model of the fossilisation process, that subsampling morphological matrices to minimise the influence of biostratinomic processes is likely to be the best way to minimise age estimate error caused by fossilisation biases. The framework employed to reach these findings assumed that the topology on which divergence times were estimated was known without error. Many divergence time methods now attempt to co-estimate divergence times and topology, and the effect of taphonomic biases on topology estimates is still unknown when morphological data is analysed in a probabilistic framework. Given the increasing use of probabilistic methods for estimating phylogenies with morphological data, an assessment of the influence of fossilisation bias on the accuracy of topology in this framework is now also required.

MCC trees have become a popular choice for summarising the posterior sample of trees obtained in TED analyses. In Chapter Six the unsuitability of maximum clade credibility (MCC) trees for Bayesian analyses of morphological data is demonstrated through the use of both simulated and empirical data. The fact that MCC consensus trees often present numerous erroneous nodes, or present poorly supported clades at the expense of well-supported ones, when attempting to summarise a diffuse posterior sample of trees is highlighted. In this chapter it is explained that the method by which divergence times are summarised on MCC trees can lead to poorly sampled clade ages and the presentation of divergence times for clades with poor support at the expense of better supported clades.

To summarise, TED methods present a framework in which the full breadth of palaeontological data is able to inform our understanding of evolutionary history. The combined analysis of extant and extinct taxa allows for insight into the evolutionary process, of which they are both part. Despite the obvious benefits of this approach, initial applications of TED did not account for many aspects of morphological evolution that are likely to influence the accuracy of age estimates. As demonstrated in this thesis, consideration of the peculiarities of morphological and fossil data can further improve the TED framework and refine evolutionary timescales.

References

Alekseyenko, A. V., Lee, C. J. & Suchard, M. A. 2008. Wagner and Dollo: A Stochastic Duet by Composing Two Parsimonious Solos. *Syst Biol,* 57, 772-84.

Alexandrou, M. A., Swartz, B. A., Matzke, N. J. & Oakley, T. H. 2013. Genome Duplication and Multiple Evolutionary Origins of Complex Migratory Behavior in Salmonidae. *Molecular Phylogenetics and Evolution,* 69, 514-523.

Alroy, J. 2002. Stratigraphy in Phylogeny Reconstruction - Reply to Smith (2000). *Journal of Paleontology,* 76, 587-589.

Arcila, D., Pyron, R. A., Tyler, J. C., Orti, G. & Betancur-R, R. 2015. An Evaluation of Fossil Tip-Dating Versus Node-Age Calibrations in Tetraodontiform Fishes (Teleostei: Percomorphaceae). *Molecular Phylogenetics and Evolution,* 82, 131-145.

Aslan, A. & Behrensmeyer, A. K. 1996. Taphonomy and Time Resolution of Bone Assemblages in a Contemporary Fluvial System: The East Fork River, Wyoming. *Palaios,* 11, 411-421.

Ayala, F. J. 1997. Vagaries of the Molecular Clock. *Proceedings of the National Academy of Sciences of the United States of America,* 94, 7776-7783.

Bapst, D. W., Wright, A. M., Matzke, N. J. & Lloyd, G. T. 2016. Topology, Divergence Dates, and Macroevolutionary Inferences Vary between Different Tip-Dating Approaches Applied to Fossil Theropods (Dinosauria). *Biology Letters,* 12.

Beck, R. M. D. & Lee, M. S. Y. 2014. Ancient Dates or Accelerated Rates? Morphological Clocks and the Antiquity of Placental Mammals. *Proceedings of the Royal Society B-Biological Sciences,* 281, 10.

Behrensmeyer, A. K. & Kidwell, S. M. 1985. Taphonomy Contributions to Paleobiology. *Paleobiology,* 11, 105-119.

Behrensmeyer, A. K., Kidwell, S. M. & Gastaldo, R. A. 2000. Taphonomy and Paleobiology. *Paleobiology,* 26, 103-147.

Benton, M. J. & Donoghue, P. C. 2007. Paleontological Evidence to Date the Tree of Life. *Mol Biol Evol,* 24, 26-53.

Benton, M. J., Donoghue, P. C. J., Asher, R. J., Friedman, M., Near, T. J. & Vinther, J. 2015. Constraints on the Timescale of Animal Evolutionary History. *Palaeontologia Electronica,* 18, 107.

Benton, M. J., Wills, M. A. & Hitchin, R. 2000. Quality of the Fossil Record through Time. *Nature,* 403, 534-537.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C. H., Xie, D., Suchard, M. A., Rambaut, A. & Drummond, A. J. 2014. Beast 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput Biol,* 10, e1003537.

Bromham, L., Woolfit, M., Lee, M. S. & Rambaut, A. 2002. Testing the Relationship between Morphological and Molecular Rates of Change Along Phylogenies. *Evolution,* 56, 1921-30.

Chen.P., Li.J., Matsukawa.M. & Al, E. 2006. Geological Ages of Track Bearing Formations in China. *Cretaceous Research* 27, 22-32.

Chen.P., Wang.Q., Zhang.H. & Al., E. 2005. Jianshangou Bed of the Yixian Formation in West Liaoning, China. *Science in China Series D; Earth Sciences* 48, 298-312.

Congreve, C. R. & Lamsdell, J. C. 2016. Implied Weighting and Its Utility in Palaeontological Datasets: A Study Using Modelled Phylogenetic Matrices. *Palaeontology,* 59, 447-462.

Davies, T. J. & Savolainen, V. 2006. Neutral Theory, Phylogenies, and the Relationship between Phenotypic Change and Evolutionary Rates. *Evolution,* 60, 476-83.

Dembo, M., Matzke, N. J., Mooers, A. O. & Collard, M. 2015. Bayesian Analysis of a Morphological Supermatrix Sheds Light on Controversial Fossil Hominin Relationships. *Proceedings of the Royal Society B-Biological Sciences,* 282, 133-141.

Dembo, M., Radovcic, D., Garvin, H. M., Laird, M. E., Schroeder, L., Scott, J. E., Brophy, J., Ackermann, R. R., Musiba, C. M., De Ruiter, D. J., Mooers, A. O. & Collard, M. 2016. The Evolutionary Relationships and Age of Homo Naledi: An Assessment Using Dated Bayesian Phylogenetic Methods. *Journal of Human Evolution,* 97, 17-26.

Donoghue, P. C. & Benton, M. J. 2007. Rocks and Clocks: Calibrating the Tree of Life Using Fossils and Molecules. *Trends Ecol Evol,* 22, 424-31.

Donoghue, P. C. J. & Keating, J. N. 2014. Early Vertebrate Evolution. *Palaeontology,* 57, 879-893.

Donoghue, P. C. J. & Smith, M. P. 2003. *Telling the Evolutionary Time: Molecular Clocks and the Fossil Record*, CRC Press.

Dornburg, A., Friedman, M. & Near, T. J. 2015. Phylogenetic Analysis of Molecular and Morphological Data Highlights Uncertainty in the Relationships of Fossil and Living Species of Elopomorpha (Actinopterygii: Teleostei). *Molecular Phylogenetics and Evolution,* 89, 205-218.

Dos Reis, M., Donoghue, P. C. J. & Yang, Z. H. 2016. Bayesian Molecular Clock Dating of Species Divergences in the Genomics Era. *Nature Reviews Genetics,* 17, 71-80.

Dos Reis, M., Inoue, J., Hasegawa, M., Asher, R. J., Donoghue, P. C. J. & Yang, Z. 2012. Phylogenomic Datasets Provide Both Precision and Accuracy in Estimating the Timescale of Placental Mammal Phylogeny. *Proceedings of the Royal Society B: Biological Sciences,* 279, 3491-3500.

Dos Reis, M. & Yang, Z. H. 2013. The Unbearable Uncertainty of Bayesian Divergence Time Estimation. *Journal of Systematics and Evolution,* 51, 30-43.

Drummond, A. J., Ho, S. Y., Phillips, M. J. & Rambaut, A. 2006. Relaxed Phylogenetics and Dating with Confidence. *PLoS Biol,* 4, e88.

Drummond, A. J., Pybus, O. G., Rambaut, A., Forsberg, R. & Rodrigo, A. G. 2003. Measurably Evolving Populations. *Trends in Ecology & Evolution,* 18, 481-488.

Duchêne, S. & Ho, S. Y. 2014. Using Multiple Relaxed-Clock Models to Estimate Evolutionary Timescales from DNA Sequence Data. *Mol Phylogenet Evol,* 77, 65-70.

Farris, J. S. 1970. Methods for Computing Wagner Trees. *Systematic Zoology,* 19, 83-&.

Felsenstein, J. 1973. Maximum-Likelihood Estimation of Evolutionary Trees from Continuous Characters. *American Journal of Human Genetics,* 25, 471-492.

Felsenstein, J. 1978. Cases in Which Parsimony or Compatibility Methods Will Be Positively Misleading. *Systematic Zoology,* 27, 401-410.

Felsenstein, J. 1981. Evolutionary Trees from DNA-Sequences - a Maximum-Likelihood Approach. *Journal of Molecular Evolution,* 17, 368-376.

Felsenstein, J. 1985. Confidence-Limits on Phylogenies - an Approach Using the Bootstrap. *Evolution,* 39, 783-791.

Felsenstein, J. 1989. Phylip - Phylogeny Inference Package (Version 3.2). *Cladistics,* 5, 164-166.

Felsenstein, J. 2005. Using the Quantitative Genetic Threshold Model for Inferences between and within Species. *Philosophical Transactions of the Royal Society B-Biological Sciences,* 360, 1427-1434.

Felsenstein, J. 2012. A Comparative Method for Both Discrete and Continuous Characters Using the Threshold Model. *American Naturalist,* 179, 145-156.

Fisher, D. C., Foote, M., Fox, D. L. & Leighton, L. R. 2002. Stratigraphy in Phylogeny Reconstruction - Comment on Smith (2000). *Journal of Paleontology,* 76, 585-586.

Fortey, R. A. & Jefferies, R. P. S. 1982. Fossils and Phylogeny –a Compromise Approach. *In:* JOYSEY, K. A. & FRIDAY, A. E. (eds.) *Problems of Phylogenetic Reconstruction.*

Systematics Association Special Volume 21. Academic Press.

Gavryushkina, A., Welch, D., Stadler, T. & Drummond, A. J. 2014. Bayesian Inference of Sampled Ancestor Trees for Epidemiology and Fossil Calibration. *Plos Computational Biology,* 10.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. 2013. *Bayesian Data Analysis*, Chapman & Hall.

Gidley, J. W. 1907. Revision of the Miocene and Pliocene Equidae of North America. *Bulletin of the American Museum of Natural History,* 23, 865-934.

Gillespie, J. H. 1991. *The Causes of Molecular Evolution,* New York ; Oxford, Oxford University Press.

Goloboff, P. A., Carpenter, J. M., Arias, J. S. & Esquivel, D. R. M. 2008a. Weighting against Homoplasy Improves Phylogenetic Analysis of Morphological Data Sets. *Cladistics,* 24, 758-773.

Goloboff, P. A., Farris, J. S. & Nixon, K. C. 2008b. Tnt, a Free Program for Phylogenetic Analysis. *Cladistics,* 24, 774-786.

Goloboff, P. A., Farris, S., Nixon, K. 2000. Tnt (Tree Analysis Using New Technology).

Gu, X. & Li, W. H. 1992. Higher Rates of Amino Acid Substitution in Rodents Than in Humans. *Molecular Phylogenetics and Evolution,* 1, 211-214.

Guillerme, T. & Cooper, N. 2016. Assessment of Available Anatomical Characters for Linking Living Mammals to Fossil Taxa in Phylogenetic Analyses. *Biology Letters,* 12.

Haldane, J. B. S. 1949. Suggestions as to Quantitative Measurement of Rates of Evolution. *Evolution,* 3, 51-56.

Harcourt-Brown, K. G., Pearson, P. N. & Wilkinson, M. 2001. The Imbalance of Paleontological Trees. *Paleobiology,* 27, 188-204.

Harvey, P. H., May, R. M. & Nee, S. 1994. Phylogenies without Fossils. *Evolution,* 48, 523-529.

Heads, M. 2012. Bayesian Transmogrification of Clade Divergence Dates: A Critique. *Journal of Biogeography,* 39, 1749-1756.

Healy, K., Guillerme, T., Finlay, S., Kane, A., Kelly, S. B. A., Mcclean, D., Kelly, D. J., Donohue, I., Jackson, A. L. & Cooper, N. 2014. Ecology and Mode-of-Life Explain Lifespan Variation in Birds and Mammals. *Proceedings of the Royal Society B-Biological Sciences,* 281.

Heath, T. A., Huelsenbeck, J. P. & Stadler, T. 2014. The Fossilized Birth-Death Process for Coherent Calibration of Divergence-Time Estimates. *Proc Natl Acad Sci U S A,* 111, E2957-66.

Hedges, S. B. & Kumar, S. 2004. Precision of Molecular Time Estimates. *Trends in Genetics,* 20, 242-247.

Heled, J. & Bouckaert, R. R. 2013. Looking for Trees in the Forest: Summary Tree from Posterior Samples. *Bmc Evolutionary Biology,* 13.

Heled, J. & Drummond, A. J. 2012. Calibrated Tree Priors for Relaxed Phylogenetics and Divergence Time Estimation. *Systematic Biology,* 61, 138-149.

Herrera, J. P. & Davalos, L. M. 2016. Phylogeny and Divergence Times of Lemurs Inferred with Recent and Ancient Fossils in the Tree. *Systematic Biology,* 65, 772-791.

Hilton, J. & Bateman, R. M. 2006. Pteridosperms Are the Backbone of Seed-Plant Phylogeny. *Journal of the Torrey Botanical Society,* 133, 119-168.

Ho, S. Y. & Lanfear, R. 2010. Improved Characterisation of among-Lineage Rate Variation in Cetacean Mitogenomes Using Codon-Partitioned Relaxed Clocks. *Mitochondrial DNA,* 21, 138-46.

Ho, S. Y. & Phillips, M. J. 2009. Accounting for Calibration Uncertainty in Phylogenetic Estimation of Evolutionary Divergence Times. *Syst Biol,* 58, 367-80.

Holder, M. T., Sukumaran, J. & Lewis, P. O. 2008. A Justification for Reporting the Majority-Rule Consensus Tree in Bayesian Phylogenetics. *Systematic Biology,* 57, 814-821.

Holland, S. M. 2000. The Quality of the Fossil Record: A Sequence Stratigraphic Perspective. *Paleobiology,* 26, 148-168.

Holton, T. A., Wilkinson, M. & Pisani, D. 2014. The Shape of Modern Tree Reconstruction Methods. *Systematic Biology,* 63, 436-441.

Hu.C., Cheng.Z. & Pang.W. 2001. *Shantungosaurus Giganteus*.

Jukes, T. H. & Cantor, C. R. 1969. *Evolution of Protein Molecules.* New York: Academic Press.

Kimura, M. 1983. *The Neutral Theory of Molecular Evolution,* Cambridge, Cambridge University Press.

Kishino, H., Thorne, J. L. & Bruno, W. J. 2001. Performance of a Divergence Time Estimation Method under a Probabilistic Model of Rate Evolution. *Mol Biol Evol,* 18, 352-61.

Klopfstein, S., Vilhelmsen, L. & Ronquist, F. 2015. A Nonstationary Markov Model Detects Directional Evolution in Hymenopteran Morphology. *Systematic Biology,* 64, 1089-1103.

Larson-Johnson, K. 2016. Phylogenetic Investigation of the Complex Evolutionary History of Dispersal Mode and Diversification Rates across Living and Fossil Fagales. *New Phytologist,* 209, 418-435.

Lartillot, N. & Philippe, H. 2006. Computing Bayes Factors Using Thermodynamic Integration. *Systematic Biology,* 55, 195-207.

Lee, M. S. Y., Cau, A., Naish, D. & Dyke, G. J. 2014. Morphological Clocks in Paleontology, and a Mid-Cretaceous Origin of Crown Aves. *Systematic Biology,* 63, 442-449.

Lee, M. S. Y. & Palci, A. 2015. Morphological Phylogenetics in the Genomic Age. *Current Biology,* 25, R922-R929.

Lee, M. S. Y., Soubrier, J. & Edgecombe, G. D. 2013. Rates of Phenotypic and Genomic Evolution During the Cambrian Explosion. *Current Biology,* 23, 1889-1895.

Lemmon, A. R., Brown, J. M., Stanger-Hall, K. & Lemmon, E. M. 2009. The Effect of Ambiguous Data on Phylogenetic Estimates Obtained by Maximum Likelihood and Bayesian Inference. *Syst Biol,* 58, 130-45.

Lepage, T., Bryant, D., Philippe, H. & Lartillot, N. 2007. A General Comparison of Relaxed Molecular Clock Models. *Mol Biol Evol,* 24, 2669-80.

Lewis, P. O. 2001. A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data. *Syst Biol,* 50, 913-25.

Ling.W., Xie.X., Liu.X. & Cheng.J. 2007. Zircon U-Pb Dating on the Mesozoic Volcanic Suite from the Qingshan Group Stratotype Section in Eastern Shandong Province and Its Tectonic Significance. *Science in China Series D; Earth Sciences,* 50, 813-824.

Losos, J. B., Arnold, S. J., Bejerano, G., Brodie, E. D., Hibbett, D., Hoekstra, H. E., Mindell, D. P., Monteiro, A., Moritz, C., Orr, H. A., Petrov, D. A., Renner, S. S., Ricklefs, R. E., Soltis, P. S. & Turner, T. L. 2013. Evolutionary Biology for the 21st Century. *Plos Biology,* 11.

Luo, Z. X., Gatesy, S. M., Jenkins, F. A., Amaral, W. W. & Shubin, N. H. 2015. Mandibular and Dental Characteristics of Late Triassic Mammaliaform Haramiyavia and Their Ramifications for Basal Mammal Evolution. *Proceedings of the National Academy of Sciences of the United States of America,* 112, E7101-E7109.

Marshall, C. R. 1994. Confidence-Intervals on Stratigraphic Ranges - Partial Relaxation of the Assumption of Randomly Distributed Fossil Horizons. *Paleobiology,* 20, 459-469.

Marx, F. G. & Fordyce, R. E. 2015. Baleen Boom and Bust: A Synthesis of Mysticete Phylogeny, Diversity and Disparity. *Royal Society Open Science,* 2.

Matzke, N. J. & Wright, A. 2016. Inferring Node Dates from Tip Dates in Fossil Canidae: The Importance of Tree Priors. *Biology Letters,* 12, 4.

Molak, M., Lorenzen, E. D., Shapiro, B. & Ho, S. Y. W. 2013. Phylogenetic Estimation of Timescales Using Ancient DNA: The Effects of Temporal Sampling Scheme and Uncertainty in Sample Ages. *Molecular Biology and Evolution,* 30, 253-262.

Mooers, A. O. & Heard, S. B. 1997. Inferring Evolutionary Process from Phylogenetic Tree Shape. *Quarterly Review of Biology,* 72, 31-54.

Near, T. J., Dornburg, A. & Friedman, M. 2014. Phylogenetic Relationships and Timing of Diversification in Gonorynchiform Fishes Inferred Using Nuclear Gene DNA Sequences (Teleostei: Ostariophysi). *Molecular Phylogenetics and Evolution,* 80, 297-307.

Nei, M. & Glazko, G. V. 2002. Estimation of Divergence Times for a Few Mammalian and Several Primate Species. *Journal of Heredity,* 93, 157-164.

Nesbitt, S. J., Barrett, P. M., Werning, S., Sidor, C. A. & Charig, A. J. 2013. The Oldest Dinosaur? A Middle Triassic Dinosauriform from Tanzania. *Biology Letters,* 9.

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. 2015. Iq-Tree: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution,* 32, 268-274.

Nylander, J. A., Ronquist, F., Huelsenbeck, J. P. & Nieves-Aldrey, J. L. 2004. Bayesian Phylogenetic Analysis of Combined Data. *Syst Biol,* 53, 47-67.

O'Leary, M. A., Bloch, J. I., Flynn, J. J., Gaudin, T. J., Giallombardo, A., Giannini, N. P., Goldberg, S. L., Kraatz, B. P., Luo, Z. X., Meng, J., Ni, X. J., Novacek, M. J., Perini, F. A., Randall, Z. S., Rougier, G. W., Sargis, E. J., Silcox, M. T., Simmons, N. B., Spaulding, M., Velazco, P. M., Weksler, M., Wible, J. R. & Cirranello, A. L. 2013. The Placental Mammal Ancestor and the Post-K-Pg Radiation of Placentals. *Science,* 339, 662-667.

O'Reilly, J. E. & Donoghue, P. C. J. 2016. Tips and Nodes Are Complementary Not Competing Approaches to the Calibration of Molecular Clocks. *Biology Letters,* 12.

O'Reilly, J. E., Dos Reis, M. & Donoghue, P. C. J. 2015. Dating Tips for Divergence-Time Estimation. *Trends in Genetics,* 31, 637-650.

O'Reilly, J. E., Puttick, M. N., Parry, L., Tanner, A. R., Tarver, J. E., Fleming, J., Pisani, D. & Donoghue, P. C. J. 2016. Bayesian Methods Outperform Parsimony but at the Expense of Precision in the Estimation of Phylogeny from Discrete Morphological Data. *Biology Letters,* 12.

Omland, K. E. 1997. Correlated Rates of Molecular and Morphological Evolution. *Evolution,* 51, 1381-1393.

Panchen, A. L. 1982. The Use of Parsimony in Testing Phylogenetic Hypotheses. *Zoological Journal of the Linnean Society,* 74, 305-328.

Parham, J. F., Donoghue, P. C., Bell, C. J., Calway, T. D., Head, J. J., Holroyd, P. A., Inoue, J. G., Irmis, R. B., Joyce, W. G., Ksepka, D. T., Patané, J. S., Smith, N. D., Tarver, J. E., Van Tuinen, M., Yang, Z., Angielczyk, K. D., Greenwood, J. M., Hipsley, C. A., Jacobs, L., Makovicky, P. J., Müller, J., Smith, K. T., Theodor, J. M., Warnock, R. C. & Benton, M. J. 2012. Best Practices for Justifying Fossil Calibrations. *Syst Biol,* 61, 346-59.

Patoumathis, M. 1994. Paleobiology - a Synthesis - Briggs,Deg, Crowther,Pr. *Anthropologie,* 98, 502-502.

Pyron, R. A. 2011. Divergence Time Estimation Using Fossils as Terminal Taxa and the Origins of Lissamphibia. *Systematic Biology,* 60, 466-481.

Rabosky, D. L. 2010. Extinction Rates Should Not Be Estimated from Molecular Phylogenies. *Evolution,* 64, 1816-1824.

Rambaut, A., Suchard, M., Xie, D. & Drummond, A. 2014. Tracer V1.6.

Rannala, B. & Yang, Z. 2007. Inferring Speciation Times under an Episodic Molecular Clock. *Syst Biol,* 56, 453-66.

Reisz, R. R. & Muller, J. 2004. Molecular Timescales and the Fossil Record: A Paleontological Perspective. *Trends in Genetics,* 20, 237-241.

Robinson, D. F. & Foulds, L. R. 1981. Comparison of Phylogenetic Trees. *Mathematical Biosciences,* 53, 131-147.

Ronquist, F., Klopfstein, S., Vilhelmsen, L., Schulmeister, S., Murray, D. L. & Rasnitsyn, A. P. 2012a. A Total-Evidence Approach to Dating with Fossils, Applied to the Early Radiation of the Hymenoptera. *Systematic Biology,* 61, 973-999.

Ronquist, F., Lartillot, N. & Phillips, M. J. 2016. Closing the Gap between Rocks and Clocks Using Total-Evidence Dating. *Philosophical Transactions of the Royal Society B-Biological Sciences,* 371.

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A. & Huelsenbeck, J. P. 2012b. Mrbayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice across a Large Model Space. *Syst Biol,* 61, 539-42.

Sanderson, M. J. 2002. Estimating Absolute Rates of Molecular Evolution and Divergence Times: A Penalized Likelihood Approach. *Mol Biol Evol,* 19, 101-9.

Sanderson, M. J. & Donoghue, M. J. 1989. Patterns of Variation in Levels of Homoplasy. *Evolution,* 43, 1781-1795.

Sanderson, M. J., Donoghue, M. 1996. The Relationship between Homoplasy and the Confidence in a Phylogenetic Tree. *Homoplasy: The Recurrence of Similarity in Evolution.* Academic Press.

Sansom, R. S., Gabbott, S. E. & Purnell, M. A. 2010. Non-Random Decay of Chordate Characters Causes Bias in Fossil Interpretation. *Nature,* 463, 797-800.

Sansom, R. S. & Wills, M. A. 2013. Fossilization Causes Organisms to Appear Erroneously Primitive by Distorting Evolutionary Trees. *Scientific Reports,* 3, 5.

Schrago, C. G., Mello, B. & Soares, A. E. 2013. Combining Fossil and Molecular Data to Date the Diversification of New World Primates. *J Evol Biol,* 26, 2438-46.

Scotland, R. W., Olmstead, R. G. & Bennett, J. R. 2003. Phylogeny Reconstruction: The Role of Morphology. *Systematic Biology,* 52, 539-548.

Seligmann, H. 2010. Positive Correlations between Molecular and Morphological Rates of Evolution. *J Theor Biol,* 264, 799-807.

Shao, K. T. & Sokal, R. R. 1990. Tree Balance. *Systematic Zoology,* 39, 266-276.

Sharma, P. P. & Giribet, G. 2014. A Revised Dated Phylogeny of the Arachnid Order Opiliones. *Frontiers in Genetics,* 5.

Simmons, M. 2011. Misleading Results of Likelihood-Based Phylogenetic Analyses in the Presence of Missing Data. *Cladistics,* 28, 208-222.

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. 2015. Iterative Adaptive Radiations of Fossil Canids Show No Evidence for Diversity-Dependent Trait Evolution. *Proceedings of the National Academy of Sciences of the United States of America,* 112, 4897-4902.

Smith, A. B. 2000. Stratigraphy in Phylogeny Reconstruction. *Journal of Paleontology,* 74, 763-766.

Stamatakis, A. 2014. Raxml Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics,* 30, 1312-1313.

Steel, M. 2013. Consistency of Bayesian Inference of Resolved Phylogenetic Trees. *Journal of Theoretical Biology,* 336, 246-249.

Stirton, R. A. 1940. *Phylogeny of North American Equidae*, University of California press.

Sumrall, C. D. & Brochu, C. A. 2003. Resolution, Sampling, Higher Taxa and Assumptions in Stratocladistic Analysis. *Journal of Paleontology,* 77, 189-194.

Sutton, M. D., Briggs, D. E. G., Siveter, D. J. & Sigwart, J. D. 2012. A Silurian Armoured Aplacophoran and Implications for Molluscan Phylogeny. *Nature,* 490, 94-97.

Swofford, D. 1998. *Paup*. Phylogenetic Analysis Using Parsimony(*and Other Methods). Version 4.*, Sinauer Associates.

Tarver, J. E., Dos Reis, M., Mirarab, S., Moran, R. J., Parker, S., O'reilly, J. E., King, B. L., O'connell, M. J., Asher, R. J., Warnow, T., Peterson, K. J., Donoghue, P. C. J. & Pisani, D. 2016. The Interrelationships of Placental Mammals and the Limits of Phylogenetic Inference. *Genome Biology and Evolution,* 8, 330-344.

Thorne, J. L., Kishino, H. & Painter, I. S. 1998. Estimating the Rate of Evolution of the Rate of Molecular Evolution. *Mol Biol Evol,* 15, 1647-57.

Thornhill, A. H., Popple, L. W., Carter, R. J., Ho, S. Y. & Crisp, M. D. 2012. Are Pollen Fossils Useful for Calibrating Relaxed Molecular Clock Dating of Phylogenies? A Comparative Study Using Myrtaceae. *Mol Phylogenet Evol,* 63, 15-27.

Tseng, Z. J., Wang, X. M., Slater, G. J., Takeuchi, G. T., Li, Q., Liu, J. & Xie, G. P. 2014. Himalayan Fossils of the Oldest Known Pantherine Establish Ancient Origin of Big Cats. *Proceedings of the Royal Society B-Biological Sciences,* 281.

Wagner, P. J. 2002. Testing Phylogenetic Hypotheses with Stratigraphy and Morphology - a Comment on Smith (2000). *Journal of Paleontology,* 76, 590-593.

Wagner, P. J. 2012. Modelling Rate Distributions Using Character Compatibility: Implications for Morphological Evolution among Fossil Invertebrates. *Biology Letters,* 8, 143-146.

Wang.S., Hu.H., Li.P. & Wang.Y. 2001. Further Discussion on Geologic Age of Sihetun Vertebrate Assemblage in Western Liaoning China: Evidence from Ar-Ar Dating. *Petrelog. Sinica* 17, 663-668.

Warnock, R. C., Yang, Z. & Donoghue, P. C. 2012. Exploring Uncertainty in the Calibration of the Molecular Clock. *Biol Lett,* 8, 156-9.

Warnock, R. C. M., Parham, J. F., Joyce, W. G., Lyson, T. R. & Donoghue, P. C. J. 2015. Calibration Uncertainty in Molecular Dating Analyses: There Is No Substitute for the Prior Evaluation of Time Priors. *Proceedings of the Royal Society B-Biological Sciences,* 282.

Wickstrom, L. M. & Donoghue, P. C. J. 2005. Cladograms, Phylogenies and the Veracity of the Conodont Fossil Record. *Conodont Biology and Phylogeny: Interpreting the Fossil Record*, 185-218.

Wiens, J. & Moen, D. 2008. Missing Data and the Accuracy of Bayesian Phylogenetics *Journal of Systematics and Evolution,* 46, 307-314.

Wiens, J. J. & Morrill, M. C. 2011. Missing Data in Phylogenetic Analysis: Reconciling Results from Simulations and Empirical Data. *Syst Biol,* 60, 719-31.

Wiens, J. J. & Tiu, J. 2012. Highly Incomplete Taxa Can Rescue Phylogenetic Analyses from the Negative Impacts of Limited Taxon Sampling. *PLoS One,* 7, e42925.

Wilkinson, R. D., Steiper, M. E., Soligo, C., Martin, R. D., Yang, Z. H. & Tavare, S. 2011. Dating Primate Divergences through an Integrated Analysis of Palaeontological and Molecular Data. *Systematic Biology,* 60, 16-31.

Winterton, S. L. & Ware, J. L. 2015. Phylogeny, Divergence Times and Biogeography of Window Flies (Scenopinidae) and the Therevoid Clade (Diptera: Asiloidea). *Systematic Entomology,* 40, 491-519.

Wood, H. M., Matzke, N. J., Gillespie, R. G. & Griswold, C. E. 2013. Treating Fossils as Terminal Taxa in Divergence Time Estimation Reveals Ancient Vicariance Patterns in the Palpimanoid Spiders. *Systematic Biology,* 62, 264-284.

Wright, A. M. & Hillis, D. M. 2014. Bayesian Analysis Using a Simple Likelihood Model Outperforms Parsimony for Estimation of Phylogeny from Discrete Morphological Data. *Plos One,* 9.

Wright, A. M., Lloyd, G. T. & Hillis, D. M. 2016. Modeling Character Change Heterogeneity in Phylogenetic Analyses of Morphology through the Use of Priors. *Systematic Biology,* 65, 602-611.

Wright, A. M., Lyons, K. M., Brandley, M. C. & Hillis, D. M. 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. Maximum-Likelihood Models for Combined Analyses of Multiple Sequence Data. *J Mol Evol,* 42, 587-96.

Yang, Z. & Rannala, B. 2006. Bayesian Estimation of Species Divergence Times under a Molecular Clock Using Multiple Fossil Calibrations with Soft Bounds. *Mol Biol Evol,* 23, 212-26.

Zhang, C., Stadler, T., Klopfstein, S., Heath, T. A. & Ronquist, F. 2016. Total-Evidence Dating under the Fossilized Birth-Death Process. *Systematic Biology,* 65, 228-249.

Zhang.J. & Rasnitsyn.A. 2006. New Extinct Taxa of Pelecinidae Sensu Lato (Hymenoptera:Proctotrupidea) in the Laiyang Formation, Shadong, China. *Cretaceous Research* 27, 684-688.

Zhou, Z., Barrett, P. M. & Hilton, J. 2003. An Exceptionally Preserved Lower Cretaceous Ecosystem. *Nature,* 421, 807-14.

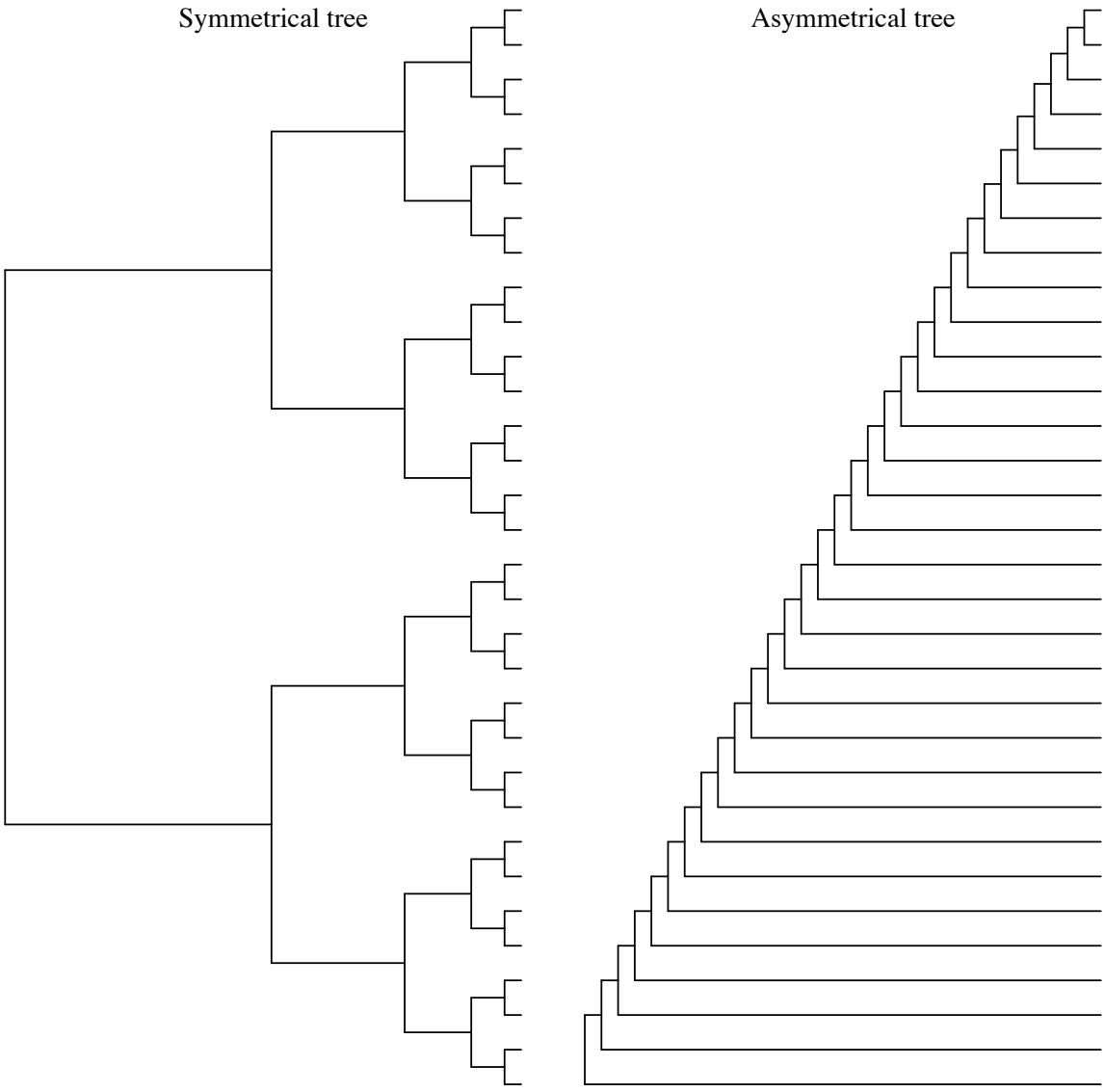
Zhoue.Z. 2006. Evolutionary Radiation of the Jehol Biota: Chronological and Ecological Perspectives. *Geological Journal* 41, 377-393.

Zhu, T., Dos Reis, M. & Yang, Z. 2015. Characterization of the Uncertainty of Divergence Time Estimation under Relaxed Molecular Clock Models Using Multiple Loci. *Syst Biol,* 64, 267-80.

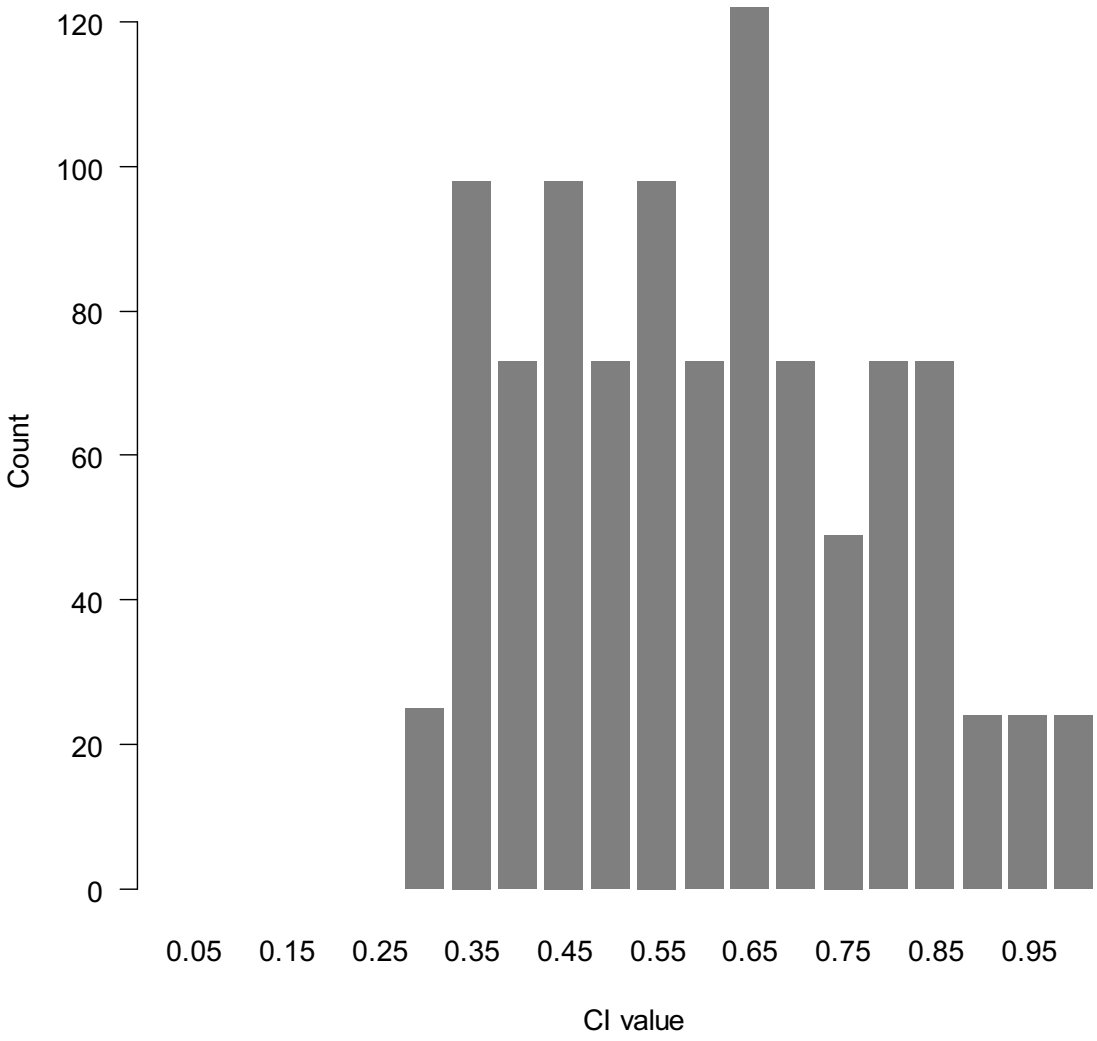
Zuckerkandl, E. & Pauling, L. 1965. Molecules as Documents of Evolutionary History. *J Theor Biol,* 8, 357-66.

Appendix

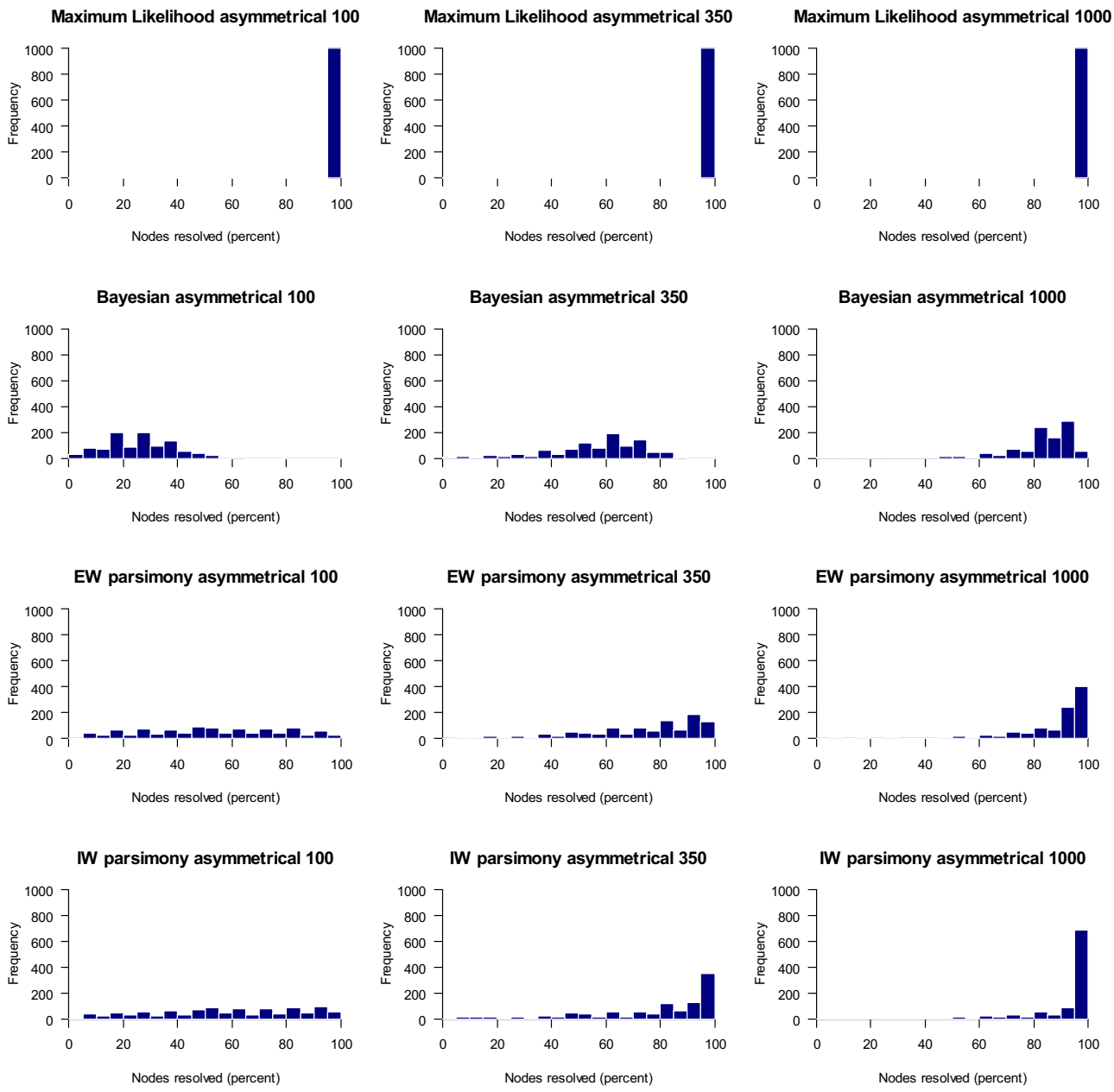
Chapter Two Supplementary Material



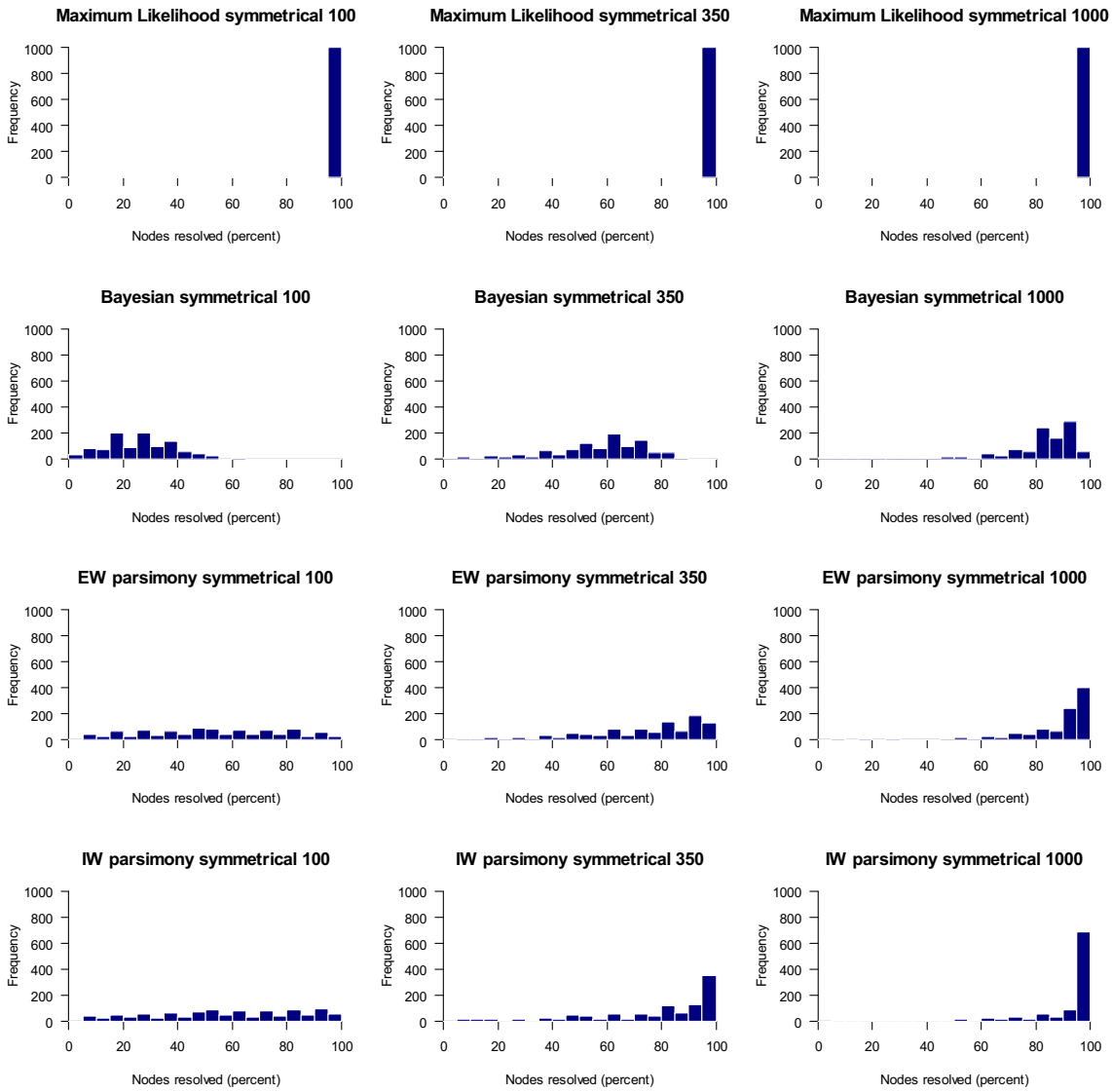
Supplementary Figure S2.1. The two phylogenies used for simulations



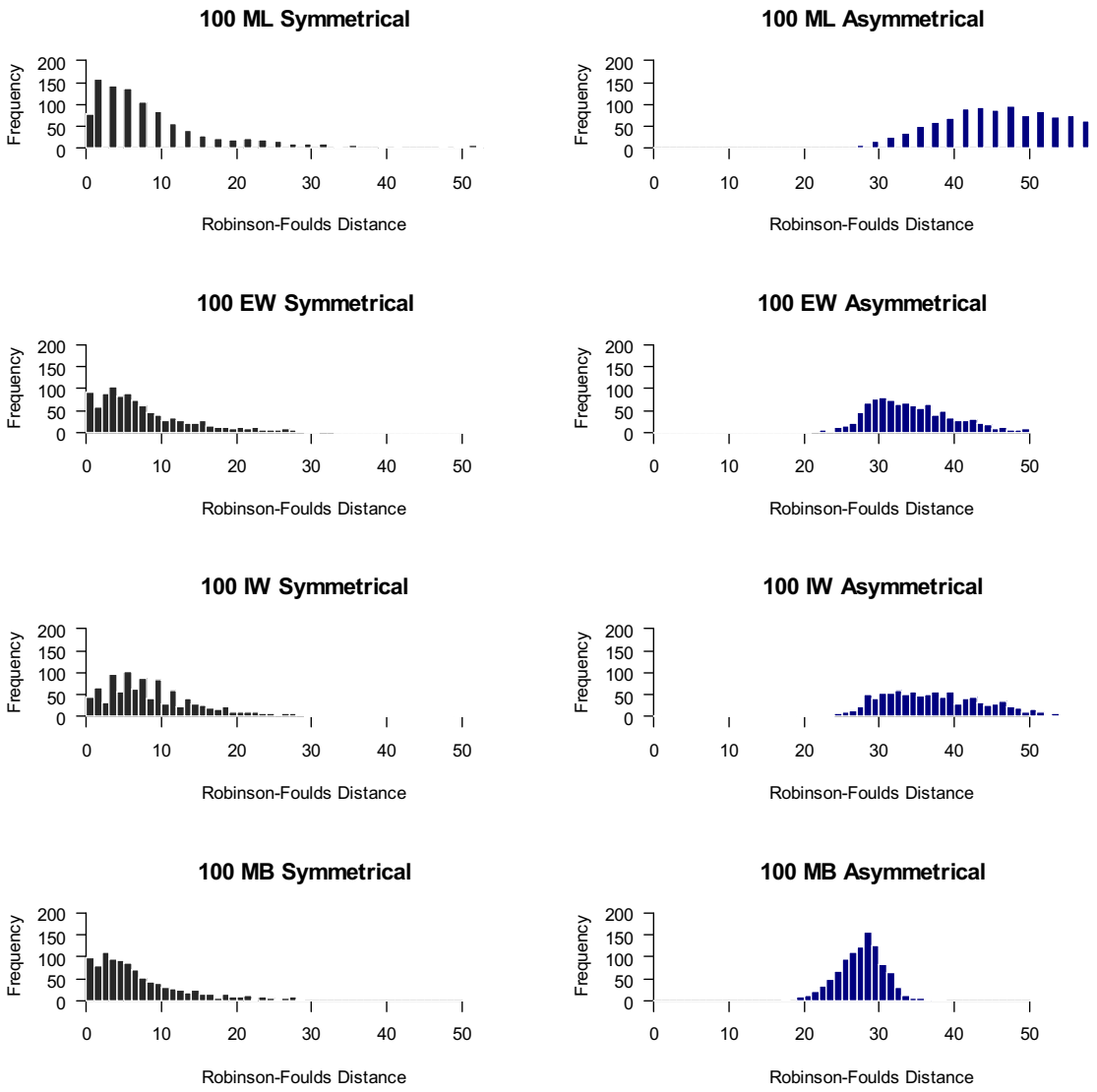
Supplementary Figure S2.2. Target distribution of CI values for datasets produced from the simulation phylogenies.



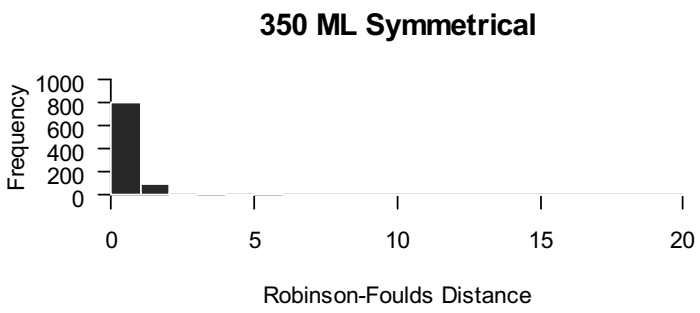
Supplementary Figure S2.3. The percentage of nodes resolved from asymmetrical generated trees for all methods and dataset sizes.

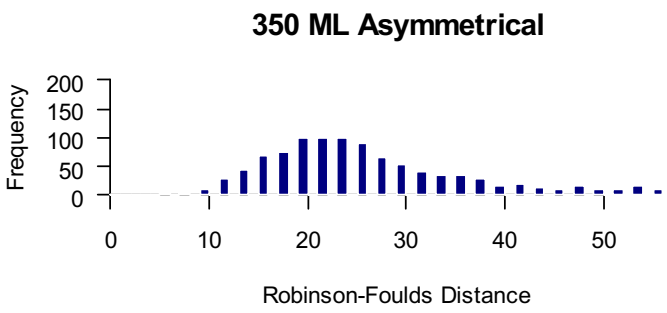


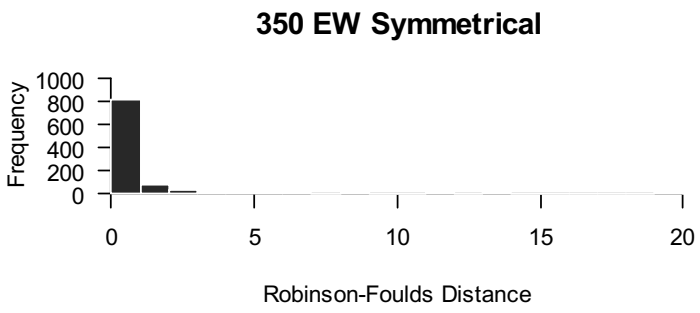
Supplementary Figure S2.4. The percentage of nodes resolved from symmetrical generated trees for all methods and dataset sizes.

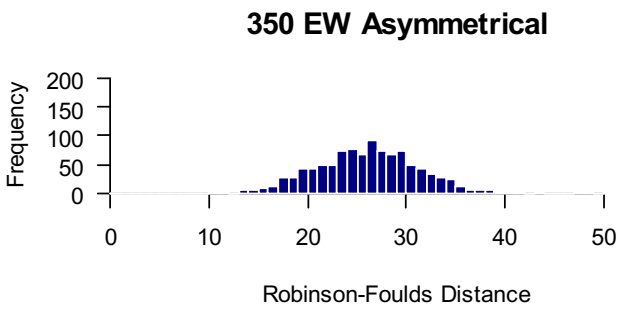


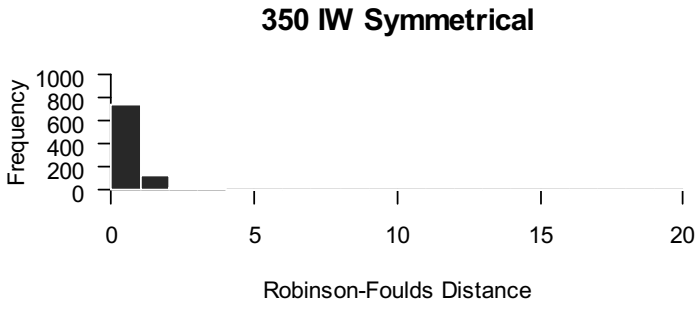
Supplementary Figure S2.5. Robinson-Foulds scores and for all phylogenies based on the 100 character simulated datasets.

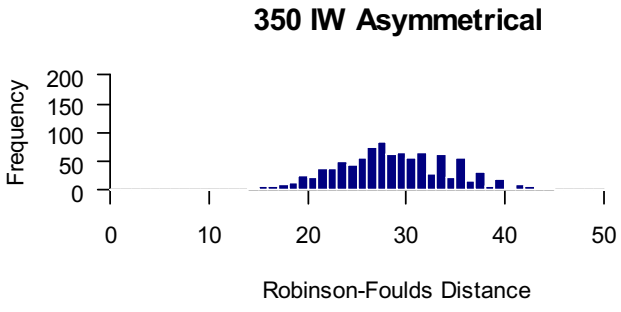


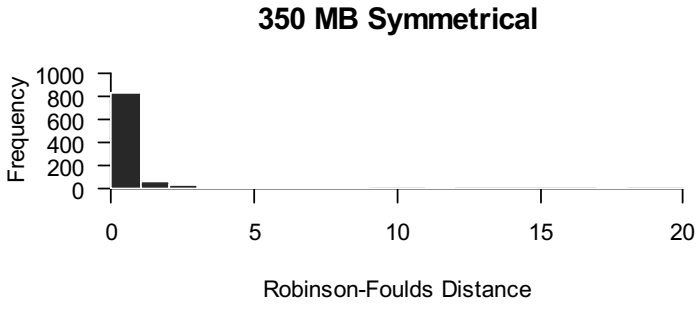


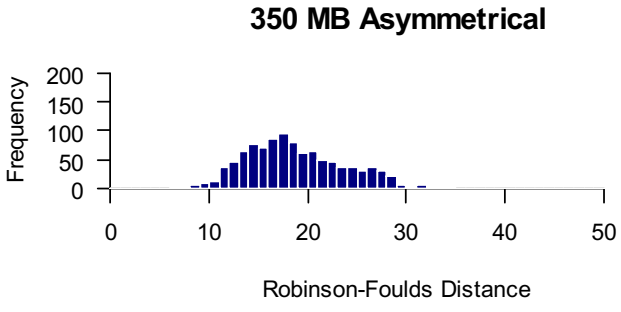




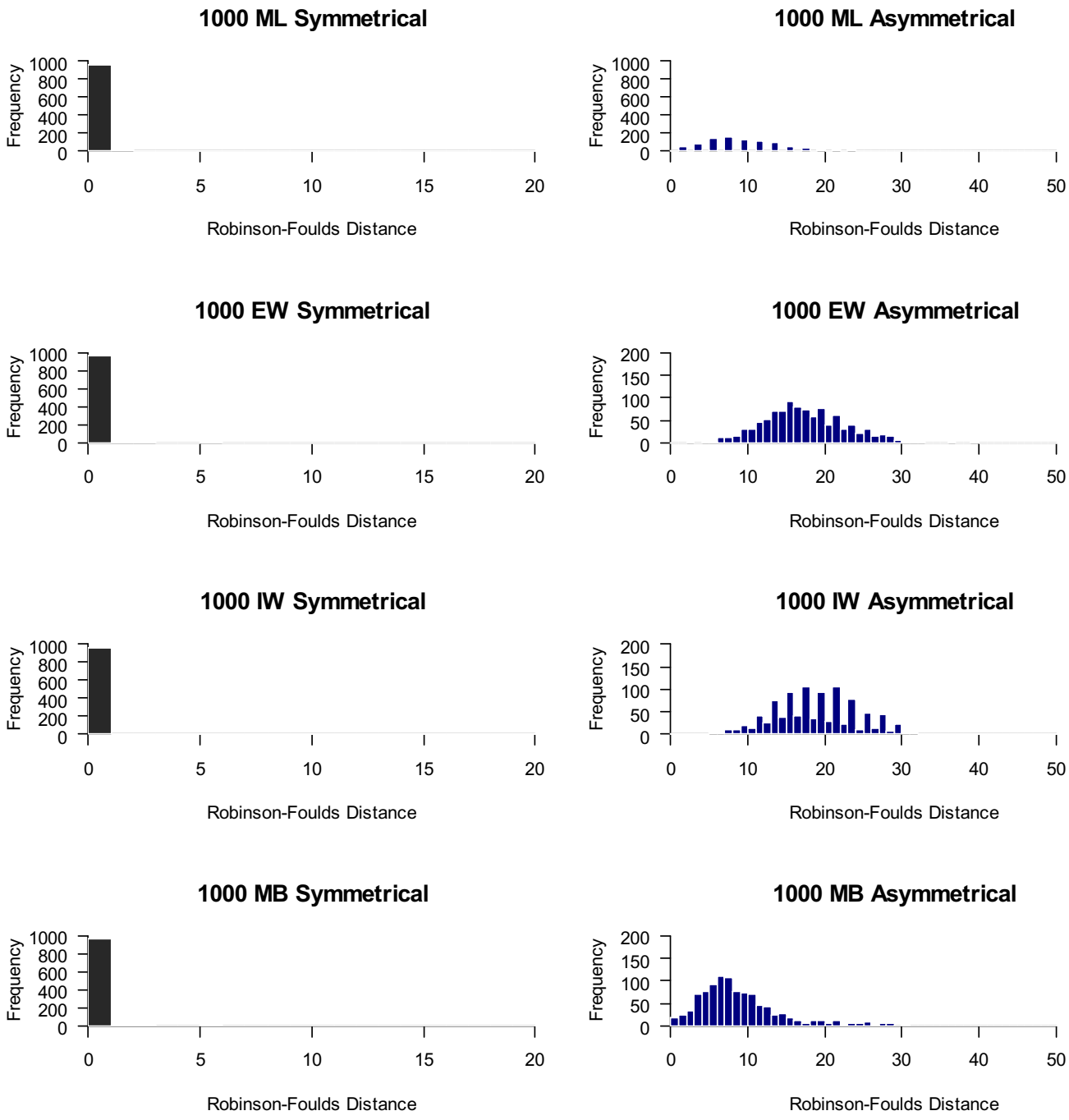




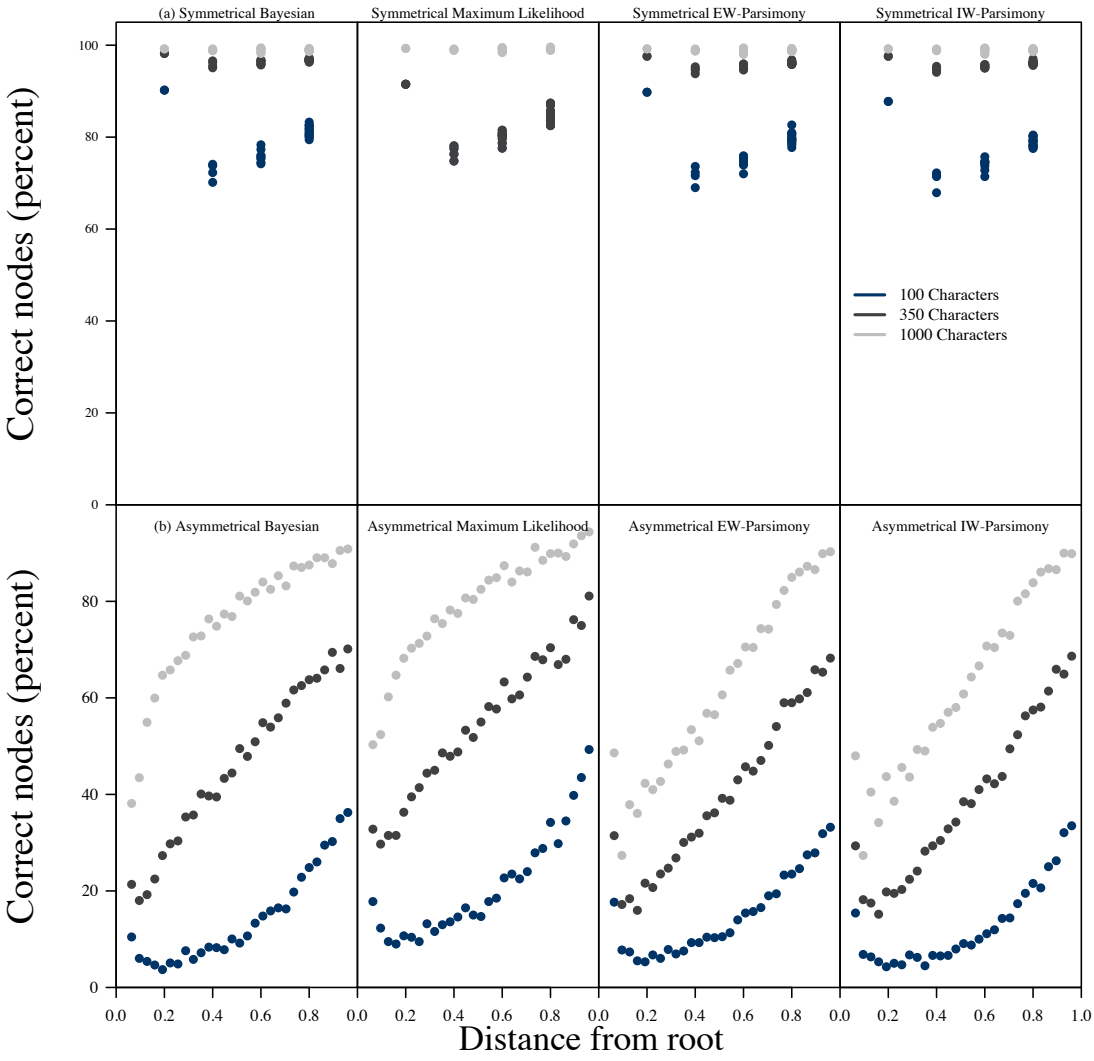




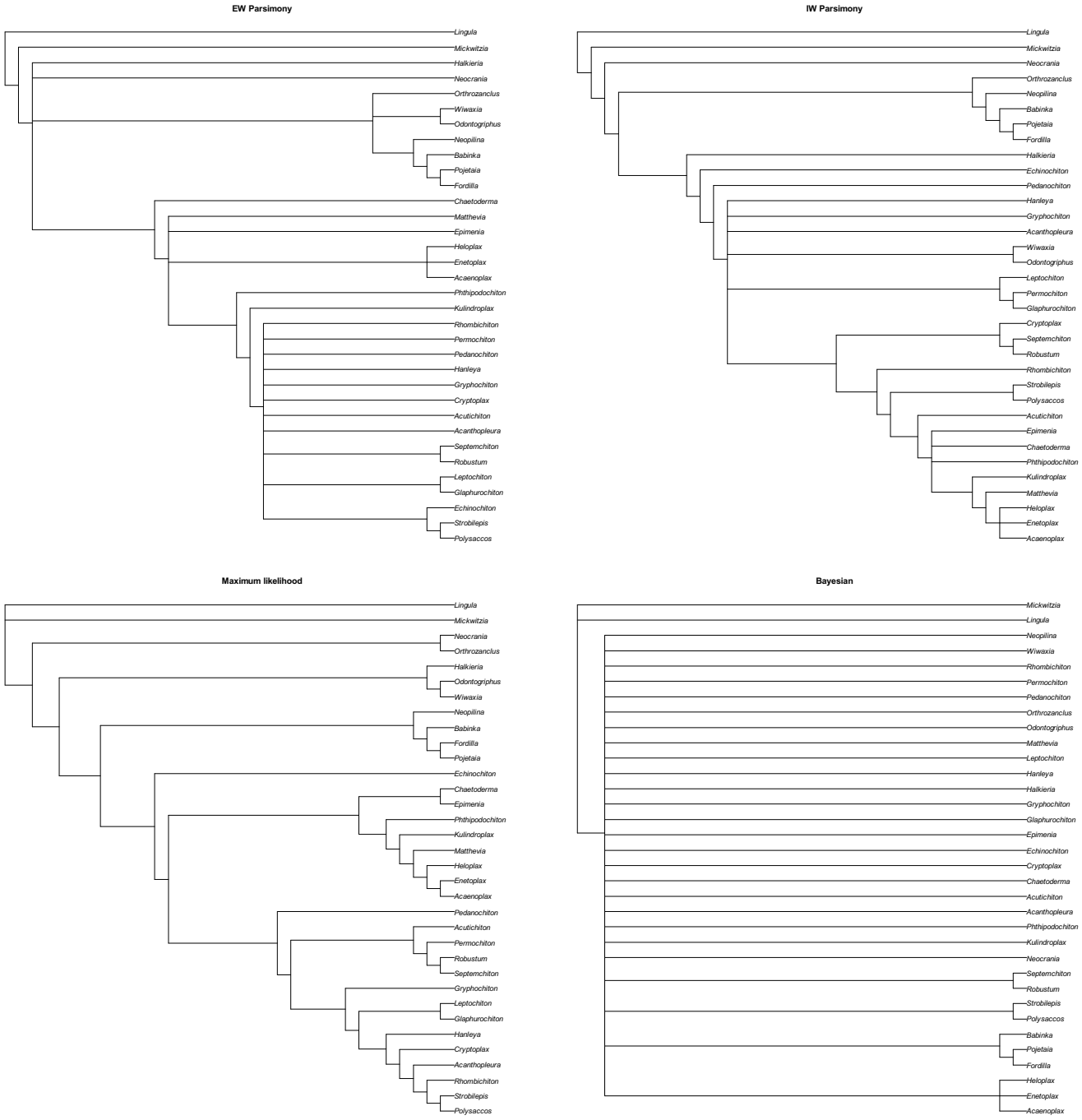
Supplementary Figure S2.6. Robinson-Foulds scores and for all phylogenies for the 350 character simulated datasets.



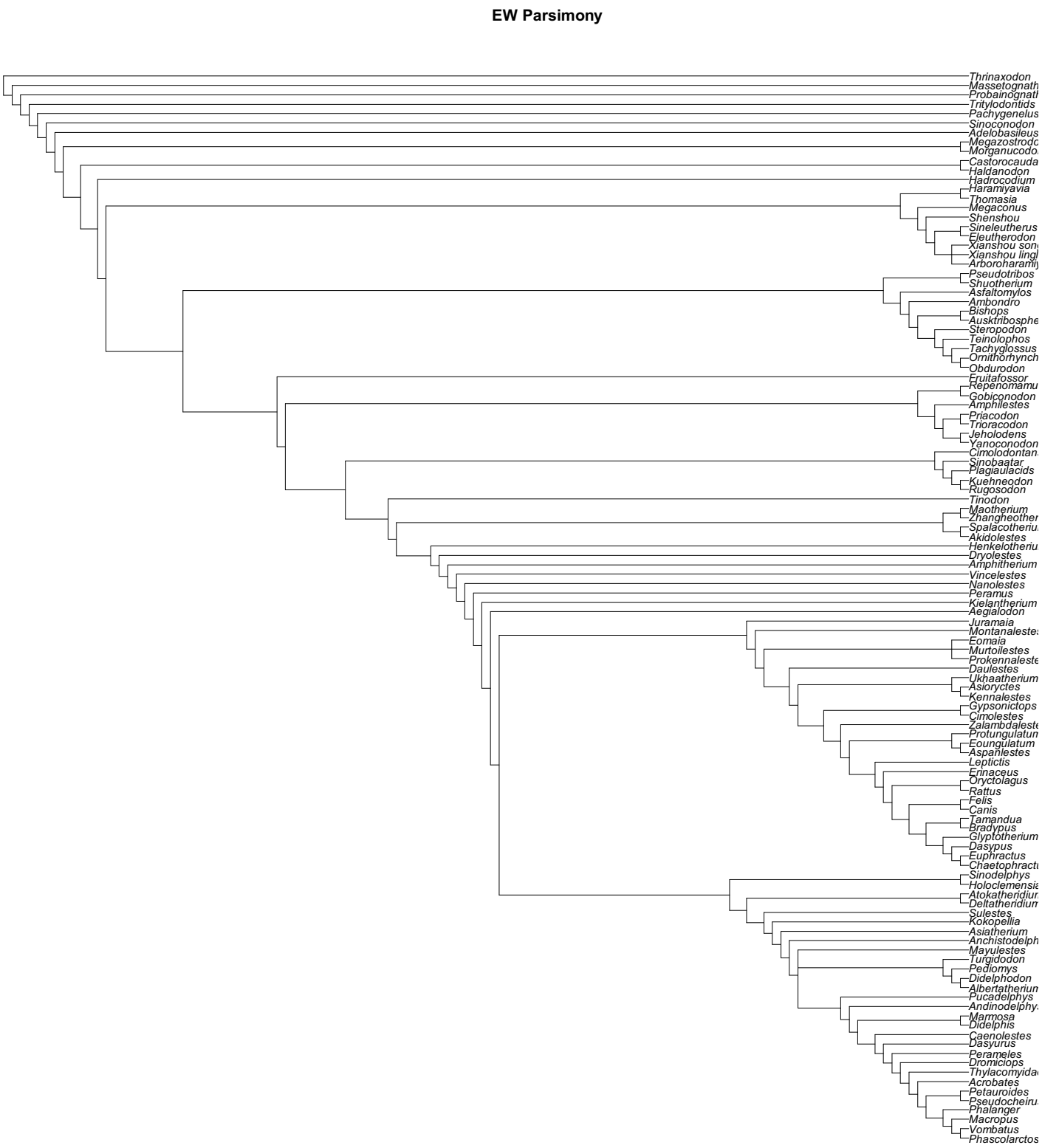
Supplementary Figure S2.7. Robinson-Foulds scores and for all phylogenies for the 1000 character simulated datasets.



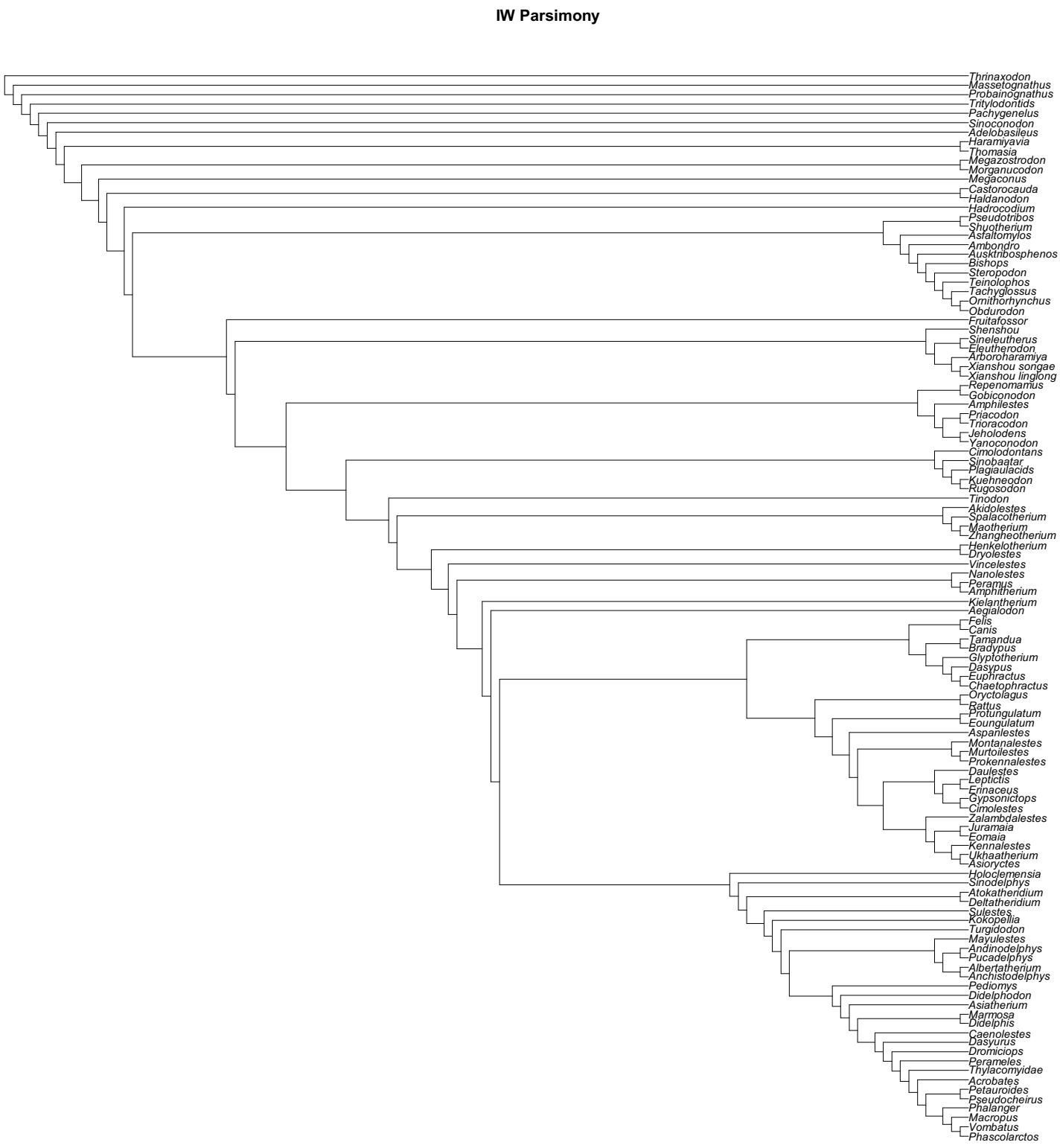
Supplementary Figure S2.8. The accuracy of reconstructions of node against the distance of each node from the root of the phylogeny. For the symmetrical simulation trees (a) there is generally high performance of all nodes. However, the asymmetrical phylogenies show a positive relationship as nodes further from the root are reconstructed more accurately (b).



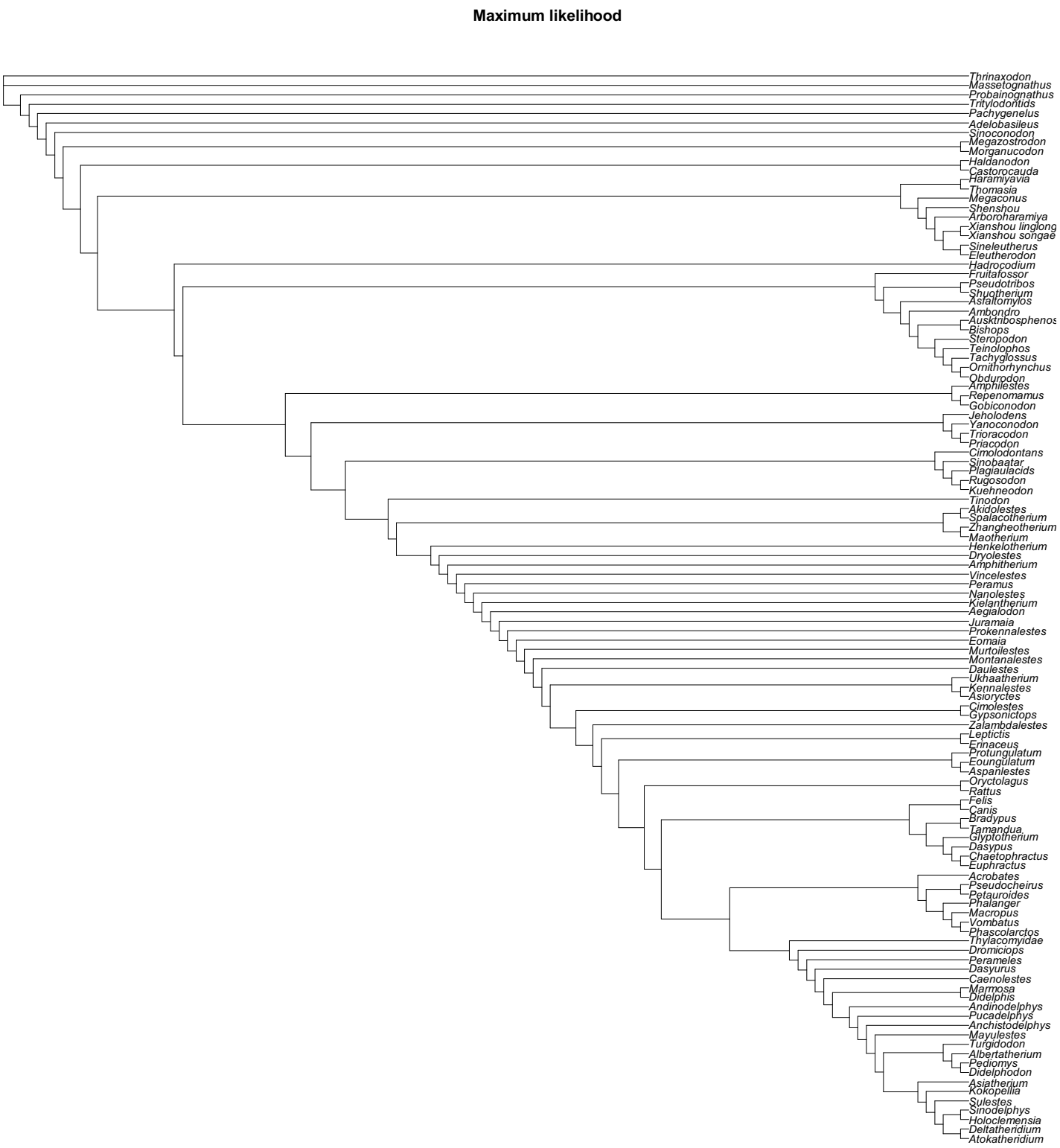
Supplementary Figure S2.9. Full phylogenies of Sutton et al. (2015) using equal weights parsimony, implied weights parsimony, maximum likelihood and Bayesian.



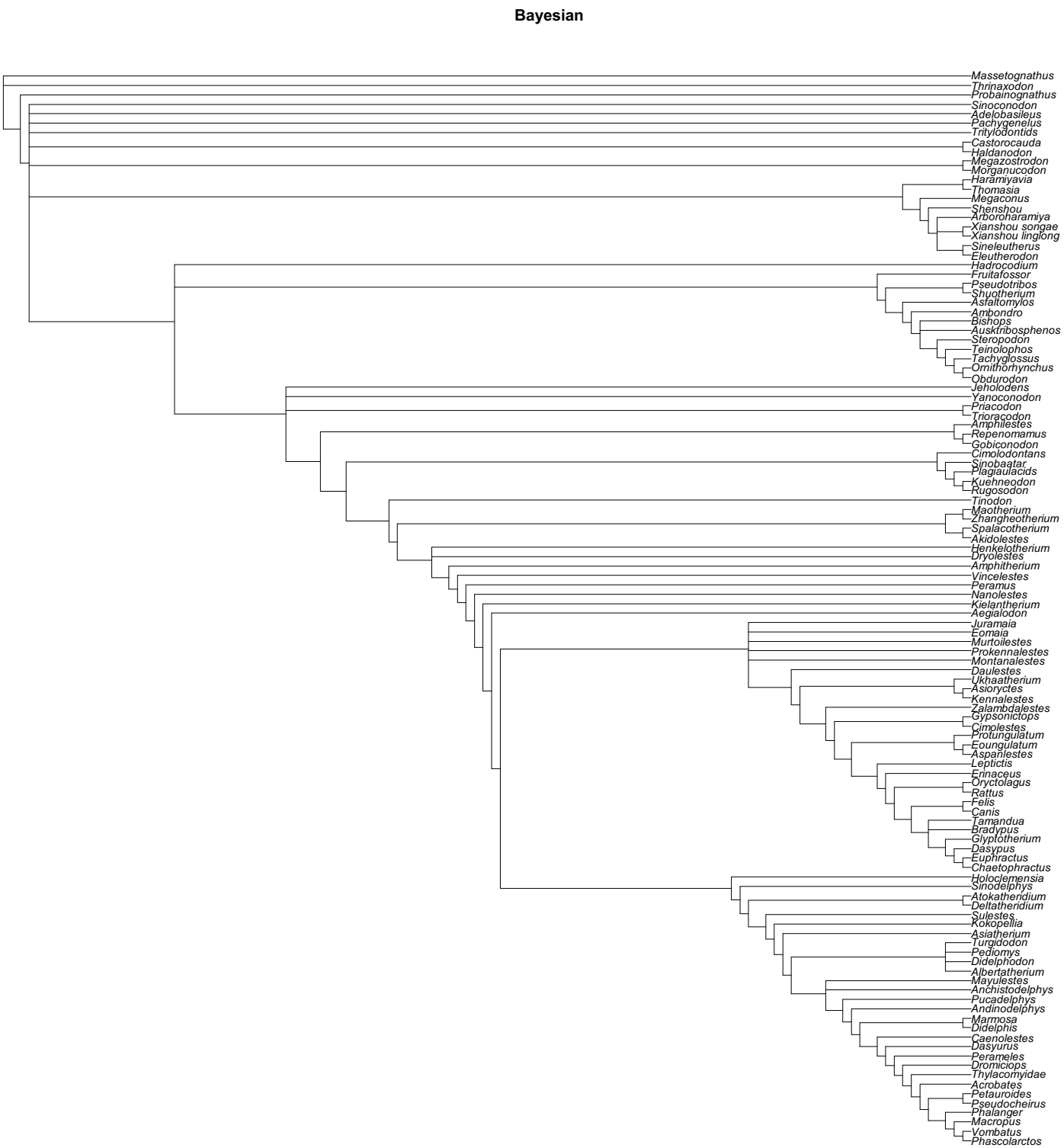
Supplementary Figure S2.10. Full phylogeny of Luo et al. (2015) using equal weights parsimony.



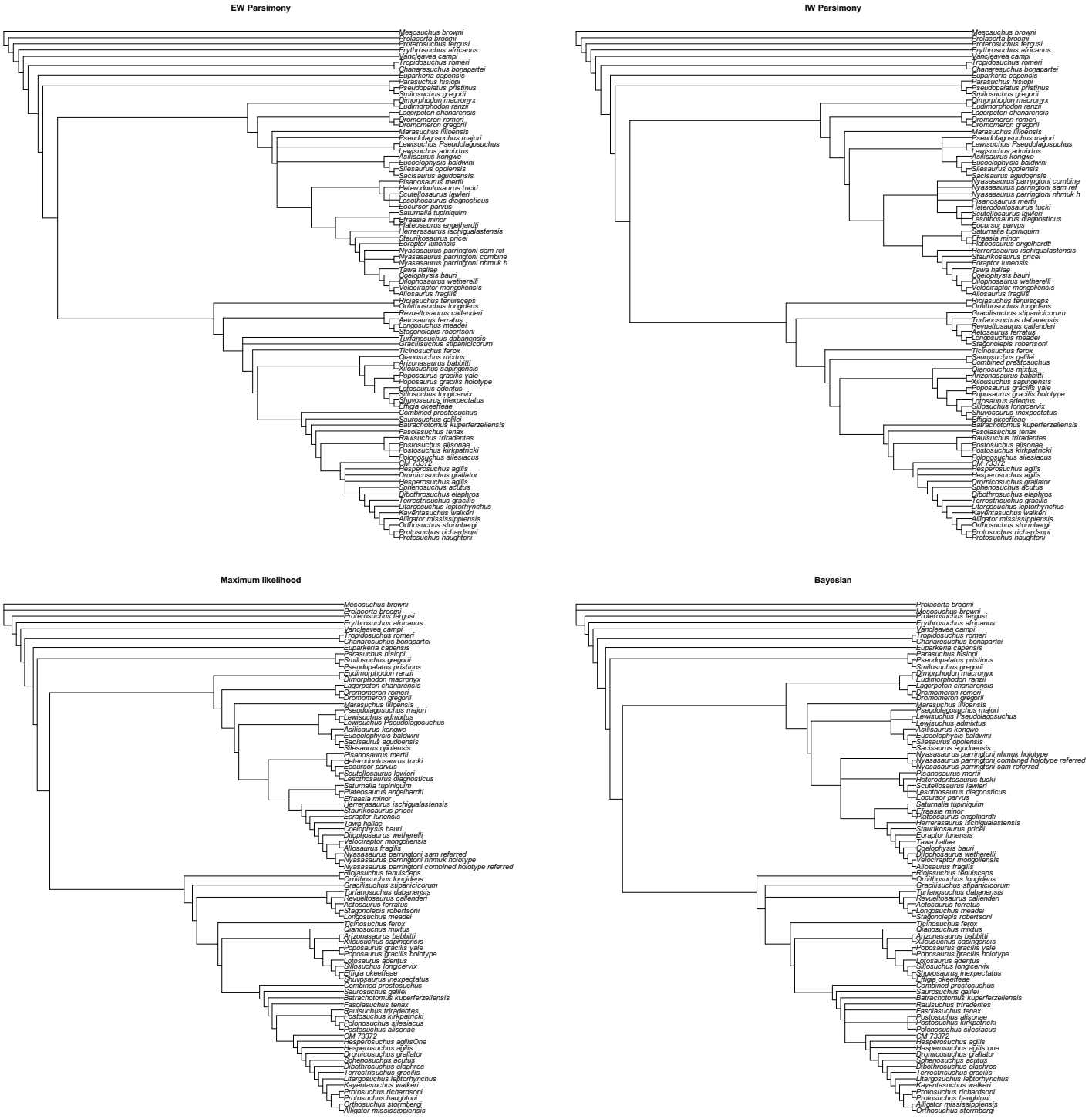
Supplementary Figure S2.11. Full phylogeny of Luo et al. (2015) using implied weights parsimony.



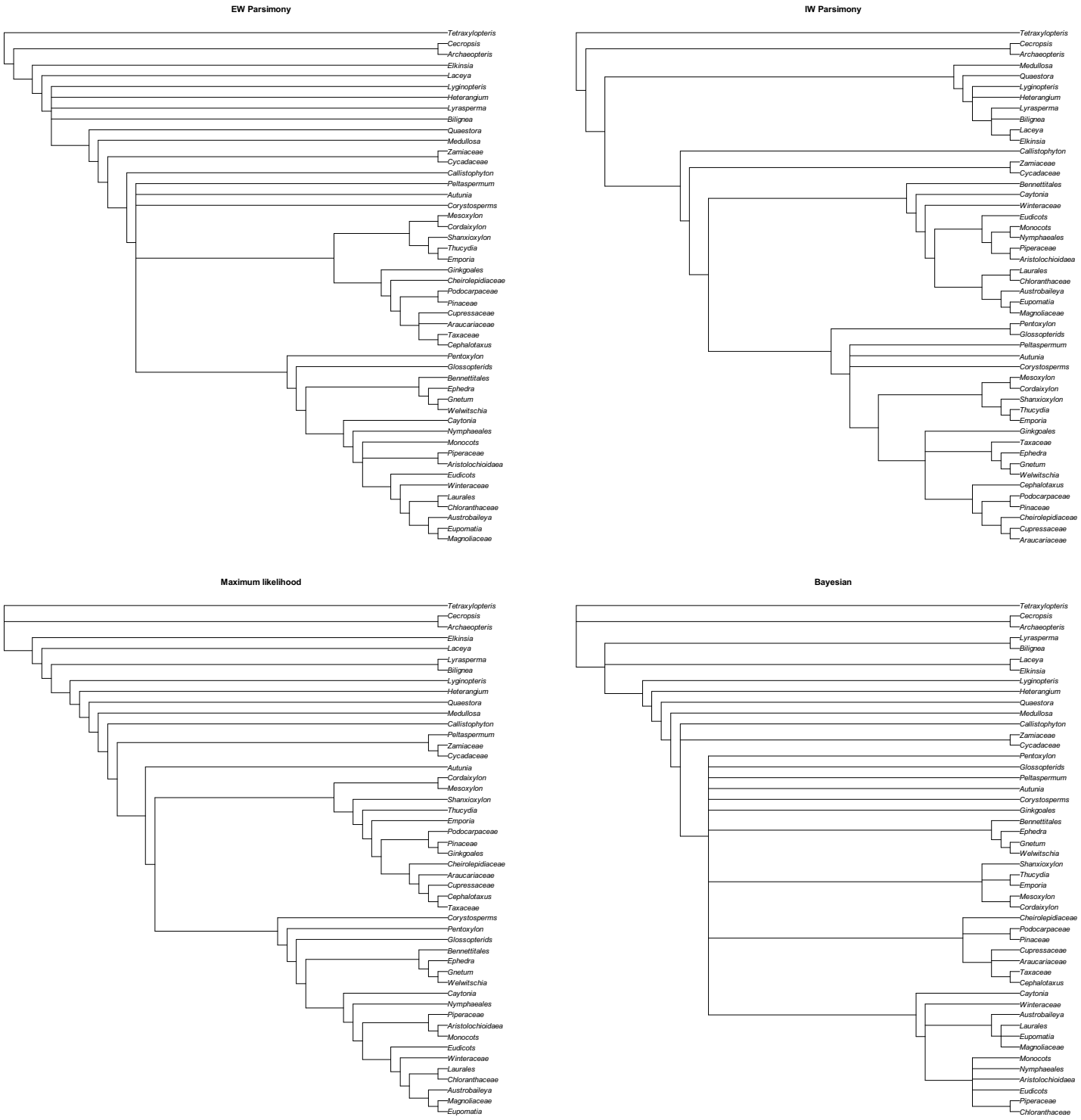
Supplementary Figure S2.12. Full phylogeny of Luo et al. (2015) using maximum likelihood.



Supplementary Figure S2.13. Full phylogeny of Luo et al. (2015) using Bayesian implementation in MrBayes.



Supplementary Figure S2.14. Full phylogenies of Nesbitt et al. (2013) using equal weights parsimony, implied weights parsimony, maximum likelihood and Bayesian.



Supplementary Figure S2.15. Full phylogenies of Hilton and Bateman (2006) using equal weights parsimony, implied weights parsimony, maximum likelihood and Bayesian.

Chapter Three Supplementary Material

Node Calibrations

Node A – Neoptera

The calibration of Node A is constructed using the oldest possible first appearance of Insects and the latest possible first appearance of Neoptera. Currently *Rhyniognatha*, found in the Rhynie Formation of Scotland, is considered to be the oldest fossil evidence of Insects, and *Ctenopilus elongatus*, from the Commentry Basin, Allier, France, is considered to be the oldest Neopteran fossil species currently known; differing from the choice of Ronquist (Ronquist et al., 2012b), who considered *Katerinka* to be representative of the earliest appearance of Neoptera.

The shale and sandstone deposits of the Rhynie Formation contain spores that were used by Rice et al. (Rice et al., 1995) to acquire a date for this Formation, which built on previous palynologicaly derived ages for the Formation. Richardson (Richardson.J., 1967) retrieved spores of the genera *Retusotriletes*, *Apiculiretusispora* and *Emphanisporites*, which could be attributed to the Devonian, from the Chert deposits of Rhynie, this allowed for the comparison of Rhynie to the Devonian Ousdale deposits of England due to similar palynological assemblages. Rice et al. (Rice et al., 1995) were able to further refine this date to the Pragian due to the presence of species including: *Ambitisporites sp;*, *Apiculatisporites sp. cf. A. microconus, Apiculiretusispora arenorugosa, A. brandtii, Calamospora spp., Cirratriradites sp., Cyclogranisporites sp., Emphanisporites micrornatus, E. neglectus, E. rotatus, E. zavallatus, Retusotriletes maculatus,* and *R. rotundu*, all of which are found in Pragian deposits (Richardson.J. and McGregor.D., 1986). Wellman et al. (Wellman et al., 2006) consider the spore assemblage of the Rhynie to be comparable to the polygonalis–emsiensis Spore Assemblage Biozone (Richardson.J. and McGregor.D., 1986) and the PoW Oppel Zone (Streel.M. et al., 1987). This relationship allows for an early Pragian–earliest Emsian age to be assigned to this Formation, adding further support for a Pragian age for the base of this Formation. Despite this, Wellman et al. (2006) point out that the base of this Formation can be considered early Pragian, but not earliest Pragian based on the age of these related assemblages, as the PoW Oppel Zone and polygonalis–emsiensis do not reach the base of the Pragian (Gradstein.F. et al., 2012).

Rice et al. (Rice et al., 1995) also utilised Ar^{40}-Ar^{39} radiometric dating to confirm an Early Devonian (Pragian-Emsian) age (396 ± 12 Ma) for the Rhynie Formation. This radiometric date can be utilised as a maximum constraint on the age of the Rhynie Formation, providing a date of 408 Ma which conforms to the early, but not earliest, Pragian age suggested by the Rhynie spore assemblage.

C.elongatus is found in the Stephanian B-C deposits of the Commentry Basin, France. The top of the Stephanian C is equivalent to the base of the Pavlovoposadian from the Russian Platform (Gradstein.F. et al., 2012). The Pavlovoposadian has been radiometrically dated as 301.29 ± 0.07 Ma (Schmitz and Davydov, 2012); allowing a minimum constraint on the first appearance of Neoptera to be placed at 301.22 Ma.

Min – 301.22 Ma

Max- 408 Ma

Node B – Insect Gall (Oldest Holometabola)

The oldest evidence for Holometabola are instances of Holometabolan larval gall from the Mattoon Formation of Illinois, The United States of America (Labandeira and Phillips, 1996). As there is no radiometric dating for this Formation it is necessary to utilise biostratigraphic sources to form correlations between units. A number of Conodont species have been recovered from the Mattoon Formation (Martin and Merrill, 1976) and provide a robust biostratigraphic marker for this calibration.

The Little Vermillion member of the upper Mattoon Formation contains numerous Conodont species, two of which are present in the Conodont zonation of the Carboniferous outlined by Gradstein et al. (Gradstein.F. et al., 2012). These species are *Streptognathodus cancellosus* and *Streptognathodus simulator* from the Kasimovian and Gzhelian respectively. The presence of these two species in the Mattoon Formation can be used to infer an age for this Formation of 304.83 Ma ± 0.36 (base *of S. cancellosus* zone) to 303.1 Ma ± 0.36 (top of *S. simulator* zone).

Min – 302.74 Ma

Max - 408

Node C – Hymenoptera

The Hymenopteran node calibration is based on the first appearance of Hymenoptera, assumed to be *Triassoxyela* and *Asioxyela* from the Madygen Formation of Kyrgyzstan.

The floral assemblage of The Madygen Formation of Kyrgystan, located to the south of the Fergana Valley (Shcherbakov.D., 2008), can be correlated with the *Scytophyllum* flora of the upper Keuper lithographic unit on the basis of the presence of key plant fossils of *Scytophyllum* and *Neocalamites* in the Madygen Formation, indicative of the Ladinian-Carnian *Scytophyllum* flora (Dobruskina.I.A., 1995, I.A., 1994). Dobruskina (Dobruskina.I.A., 1995, I.A., 1994) proposed that the Madygen Flora was most similar to the middle Triassic Floras of Eurasia as no Early/Late Triassic floral assemblage contained enough common taxonomic groups to support a correlation. The most similar flora to that of Madygen are: *Priuralye*, *Nikolayevka* and *Garazhovka* (Donetsk Basin) and *Bogoslovsk*, all of which are Ladinian to Carnian in age (Dobruskina.I.A., 1995, I.A., 1994).

Correlation with the Priuraly flora is based on the appearance of the following groups in both locations: *Filicophyta*; *Chiropteris*; *Lepidopteris*; *Scytophyllum*; *Vittaephylum*; *Glossophyllum*. Correlation to the Nikolayevka and Garazhovka flora of the Donetsk Basin is based on the shared appearance of: *Neocalamites; Chiropteris; Lepidopteris; Scytophyllum; Vittaephyllum; Glossophyllum* (Dobruskina.I.A., 1995). Correlation with the Carnian Svalbard flora is based on the shared appearance of the Glossophyllacea Family (Dobruskina.I.A., 1995).

The Scytophyllum Flora is correlated with the Cortaderitian Stage of Gondwana due to similarities in floral assemblages, particularly the abundance of *Scytophyllum* (Morel et al., 2003). The Cortaderitian Stage is divided into 3 Biozones; a $^{40}Ar/^{39}Ar$ radiometric date for the middle biozone of the Cortaderitian Stage of 228.5 ± 0.3 Ma was measured by Rogers et al. (Rogers et al., 1993), supporting the Ladinian – Carnian age for the *Scytophyllum* flora and the

Madygen Formation. Further support for the Ladinian - Carnian age of the Madygen Formation can also be derived from the Gondwanian floral stages; the Puesto Viejo Formation, part of the Barrealian Stage underlying the Cortaderitian Stage (and therefore the *Scytophyllum* Flora), has been radiometrically dated to 232± 4 Ma (Valencio.D. et al., 1975), this would suggest that the Cortaderitian Stage can be no older than 236 Ma, and therefore the *Scytophyllum* flora and Madygen Formation can be no older than this age either.

A minimum constraint on the age of the Madygen Formation can be inferred from the strong support for a Ladinian – Carnian age for this Formation, allowing for the end of the Carnian (216.5 ± 2 Ma) to be utilised as the minimum age for the Madygen Formation.

The top of the unit in which the first evidence for Holometabola has been found provides the maximum constraint for this calibration. This is the Mattoon Formation of The United States of America, the minimum age of which has been outlined in the construction of the calibration for Holometabola (Node B).

Min – 214.5 Ma

Max – 302.74 Ma

Node D – Xyelidae

The oldest representative of Xyelidae in this analysis is considered to be *Eoxyela*, found at Novospasskoye, Ichetuy Formation in Transbaikalia, Siberia, Russia. As the first appearance of Hymenoptera, provided by fossils from the Madygen Formation, is used to determine the oldest possible age for this node we are only concerned in deriving a minimum age for the Novospasskoye Formation.

The Ichetuy Formation is thought to be of an Early-Middle Jurassic age on the basis of the biostratigraphic composition of the Formation (Skoblo.V. and Lyamina.N., 1965, Rasnitsyn.A. and Quicke.D., 2002, Ponomarenko.A., 1993), although radiometric dating of volcanogenic material suggests a younger age than this (Metelkin.D. et al., 2007). The basalt covers at the top of the Ichetuy Formation have been dated to 145 ± 4 Ma through the use of K-Ar dating (Ivanov.V. et al., 1995), this date can be utilised as a minimum constraint on the age of the Ichetuy Formation as the date is measured from material overlying the formation. Other dates measured for the volcanogenic material of the Ichetuy Formation are 158 ± 8 (Rb-Sr wr; (Gordienko et al., 1997)), 150 ± 5 (K-Ar; (Ivanov.V. et al., 1995)), 158 ± 4 (Rb-Sr wr; (Shadaev.M. et al., 1992)), 156 ± 4 – 146 ±3 (K-Ar; (Ivanov.V. et al., 1995)), 150 ± 4 – 140 ±4 (K-Ar; (Ivanov.V. et al., 1995)), and 159.1 ±2.7 (Rb-Sr wr; (Andryushchenko et al., 2010)). These dates range from 162 Ma to 136 Ma, suggesting that the Ichetuy Formation is of Callovian – Berriasian age (Gradstein.F. et al., 2012). The latest of these radiometric dates is 136 Ma and can be utilised as a minimum constraint for this calibration.

Min - 136

Max - 214.5

Node E – Pamphilioidea

The calibration of Node E is based on the first appearance of two fossil species, *Aulidontes mandibulatus* and *Pamphiliidae undescribed*. Ronquitst et al. (Ronquist et al., 2012a) considered the Formations that these two species are found in (Karatau, Karabastau locality, Kazakhstan and Daohugou, China, respectively) to be of the same age and therefore treated them as equally likely candidates for the first appearance of *Pamphilioide*a. Karatau and Daohugou are widely considered to be comparable in age, yet when all sources in the literature are taken into account it can be shown that whilst they are of a comparable age, with both Formations starting at the same time, the Karatau Beds have a more recent age attributed to their upper members.

The Karatau locality consists of a group of deposits situated in the Jambul Province, Kazakhstan. The most notable sites are Aulie (also known as Mikhailovka), Karabastau, and Uspenovka (formerly Galkino), located within the Kulbastau Mountain Range. The floral composition of the Karatau mountain range is well documented and specific floral assemblages have been identified (Doludenko and Orlovskaya, 1976). The Karabastausky floral assemblage was initially identified at the Karabastau site and the comparison of floral assemblages at Galinko allowed this site to be assigned to the Karabastausky assemblage (Doludenko and Orlovskaya, 1976). One of the characteristics of the Karabastausky flora is an abundance of *Classopolis* pollen (95-100%)(Doludenko and Orlovskaya, 1976). Vakhrameev (Vakhrameev.V., 1991) analysed the fluctuations in *Classopolis* abundance across Eastern Europe and Asia and compared them with major Geological events; this analysis showed that *Classopolis* in Kazakhstan, Middle Asia, Ukraine and Crimea only reached abundances of +95% during the Oxfordian and Kimmeridgian, before decreasing rapidly during the Late Kimerridgian – Tithonian. The Karabastausky Assemblage is positioned above the Borolsaisky Assemblage but it is unknown what length of time separates these two assemblages (Doludenko and Orlovskaya, 1976); Despite this, the Karabaustsky Assemblage must be no older than the Borolosaisky, so the age of the top of this assemblage can still constrain the age of the base of the Karabustsky Assemblage. The upper parts of the Borolosaisky Assemblage are considered to be of a Lower to Middle Callovian age as they contain around 50% *Classopolis*. Doludenko and Orlovskaya (Doludenko and Orlovskaya, 1976) and Sakulina (Sakulina.G.V., 1971) has shown that this level of abundance is indicative of an Early Middle Callovian age, whereas higher abundances are indicative of Upper Callovian- Tithonian ages; The Borolosaisky Assemblage reaches a peak *Classopolis* abundance of 50% before dropping back down to 10%, supporting an Early – Middle Callovian age, meaning that a the age of the base of the Callovian (166.1 ± 1.2 Ma) (Gradstein.F. et al., 2012) can be assigned to the base of the Karabaustsky Assemblage. If we consider the 95% abundance of *Classopolis* in the Karabastau Assemblage as indicative of a pre Late Kimmeridgian reduction in *Classopolis* abundance then we can assign the age of the base of the Tithonian (152.1 ±0.9 Ma) (Gradstein.F. et al., 2012) as the age of the top of the Karabastausky Assemblage.

Min – 151.2

Max – 214.5

Node F – Siricoidea

The earliest representatives of Siricoidea in this analysis are considered to be *Aulisca, Anaxyela, Syntexyela, Kulbastavia* and *Brachysyntexis*, all of which are from the Karatau Locality, Kazakhstan, the minimum age of which will provide the minimum constraint on the age of this particular calibration. The maximum age of Holometabola, provided by the age of the Madygen Formation, will be used as the maximum constraint; therefore we are only concerned with deriving the minimum age of the Karatau Locality.

The Karatau locality consists of a group of deposits situated in the Jambul Province, Kazakhstan. The most notable sites are Aulie (also known as Mikhailovka), Karabastau, and Uspenovka (formerly Galkino), located within the Kulbastau Mountain Range. The floral composition of the Karatau mountain range is well documented and specific floral assemblages have been identified (Doludenko and Orlovskaya, 1976). The Karabastausky floral assemblage was initially identified at the Karabastau site and the comparison of floral assemblages at Galinko allowed this site to be assigned to the Karabastausky assemblage (Dorludenko and Orlovskaya 1976). One of the characteristics of the Karabastausky flora is an abundance of Classopolis pollen (95-100%)(Doludenko and Orlovskaya, 1976). Vakhrameev (Vakhrameev.V., 1991) analysed the fluctuations in *Classopolis* abundance across Eastern Europe and Asia and compared them with major Geological events; this analysis showed that *Classopolis* in Kazakhstan, Middle Asia, Ukraine and Crimea only reached abundances of +95% during the Oxfordian and Kimmeridgian, before decreasing rapidly during the Late Kimerridgian – Tithonian. The Karabastausky Assemblage is positioned above the Borolsaisky Assemblage but it is unknown what length of time separates these two assemblages (Doludenko and Orlovskaya, 1976); Despite this, the Karabaustsky Assemblage must be no older than the Borolosaisky, so the age of the top of this assemblage can still constrain the age of the base of the Karabustsky Assemblage. The upper parts of the Borolosaisky Assemblage are considered to be of a Lower to Middle Callovian age as they contain around 50% *Classopolis*. Doludenko and Orlovskaya. (Doludenko and Orlovskaya, 1976) and Sakulina (Sakulina.G.V., 1971)has shown that this level of abundance is indicative of an Early Middle Callovian age, whereas higher abundances are indicative of Upper Callovian- Tithonian ages; The Borolosaisky Assemblage reaches a peak *Classopolis* abundance of 50% before dropping back down to 10%, supporting an Early – Middle Callovian age, meaning that a the age of the base of the Callovian (166.1 ± 1.2 Ma) (Gradstein.F. et al., 2012) can be assigned to the base of the Karabaustsky Assemblage. If we consider the 95% abundance of *Classopolis* in the Karabastau Assemblage

as indicative of a pre Late Kimmeridgian reduction in *Classopolis* abundance then we can assign the age of the base of the Tithonian (152.1 ±0.9 Ma) (Gradstein.F. et al., 2012) as the age of the top of the Karabastausky Assemblage.

Min – 151.2

Max - 214.5

Node G – Vespina

The oldest representative of Vespina used by Ronquist et al. (Ronquist et al., 2012a) is the species *Brigittepteris brauckmanni*, found in the Dobbertin Locality of Germany. The age of this locality can inform the minimum constraint on the calibration of the Vespina node with the earliest possible appearance of Hymenoptera, seen in the Madygen Formation, providing the maximum constraint. As we are only concerned with the latest possible appearance of Vespina we only need to determine a minimum age for the Dobbertin locality.

The Dobbertin locality of Mecklenburg-Vorpommern, Northern Germany, is widely considered to be lower Toarcian in age. Insect finds from this locality are assigned to the *Harpoceras falciferum* ammonoid zone (Vrsansky and Ansorge, 2007, Krzeminski.W., 1995, Ansorge.J., 1993b). The presence of a biostratigraphic marker as strong as the ammonite *H.falciferum* means that a strongly supported age can be inferred for this locality.

The *falciferum* ammonoid zone is the second earliest of the Toarcian. The end of the *falciferum* ammonoid zone, which has been established to be 182.0 +3.3/-1.8 Ma (Palfy et al., 2002), can be utilized as a minimum constraint on the age of the Dobbertin locality.

Min – 180.2

Max - 214.5

Node H - Apocrita

The First appearance of Apocrita can be inferred from the age of the fossil species belonging to Mesoserphidae. This group contains the oldest representatives of Proctotupoidea, whose divergence from Chalcidoidea represents the formation of the crown group Apocrita (Warnock et al., 2012)**.** As the maximum age of this node is provided by the first appearance of Holometabola, from the Madygen Formation, we are only concerned with the youngest possible age for the appearance of this species and therefore the youngest possible age for this Formation.

Early members of Mesoserphidae are found in the Daohugou Beds in China (Rasnitsyn and Zhang, 2004). The Daohugou Beds have been radiometrically dated, allowing for a robust minimum for the Apocrita calibration. $^{40}Ar/^{39}Ar$ radiometric dating provides the most accurate representation of the minimum age of the Jiulongshau Formation of the Daohugou Beds. Chang et al. (Chang et al., 2009) used this method on two tuffs to date the very bottom of the Lanqui Formation, which lies just above the Haifanggou Formation. The Lanqui and Haiffangou formations are situated in Liaoning Province, the same formation is referred to as the Jiulongshan Formation in Hebei province and therefore Haiffangou is equivalent to Jiulongshan. This relationship means that the oldest possible age for the Lanqui Formation can be used as a minimum constraint on the age of the Jiulongshan Formation. The youngest tuff measured by Chang et al. (Chang et al., 2009) was dated as 158 Ma ± 0.6; this gives a minimum constraint of 157.4 Ma for the Jiulongshan Formation.

Min – 157.4

Max – 214.5

Node I – Tenthredinoidea

The calibration for the Tenthredinoidea clade can be derived from the age of the formation in which the oldest fossil of this species, *Palaeathalia laiangensis*, is found in. *P.laiangensis* is found in the Laiyang Formation of China; as the maximum constraint for this calibration is provided by the first appearance of Holometabola, from the Madygen Formation, we are only concerned with the youngest possible first appearance of *P.laiangensis* and therefore the minimum possible age for the top of the Laiyang Formation.

An estimate of the minimum age of the Laiyang Formation can be derived from the oldest known age of the formation know to be positioned above it stratigraphically. The Qingshan Group is positioned above the Laiyang Formation (Ling et al., 2007) and therefore a date for the bottom of the Qingshan Formation can be utilised as a minimum constraint on the age of the Laiyang Formation. The use of Zircon U-Pb dating on a number of samples from the lowest part of the Houkuang Formation, part of the Qingshan Group were measured as 106 Ma ± 2 (Ling et al., 2007); this therefore places a minimum constraint on the Laiyang Formation of 104 Ma.

Min – 104

Max – 214.5

Fossil Tip Calibrations

Madygen Formation

The floral assemblage of the Madygen Formation of Kyrgystan, located to the south of the Fergana Valley (Shcherbakov.D., 2008), has been correlated with the Scytophyllum flora of the Upper Keuper lithographic unit on the presence of *Scytophyllum* and *Neocalamites* remains in the Madygen Formation (Dobruskina.I.A., 1995, I.A., 1994). The Scytophyllum flora ranges in age from the start of the Ladinian to the end of the Carnian (Dobruskina.I., 1993). Dobruskina (Dobruskina.I.A., 1995, I.A., 1994) proposed that the Madygen Flora was most similar to the Middle Triassic floras of Eurasia as no Early/Late Triassic floral assemblages contained enough common taxa to support a correlation. The most similar flora to that of Madygen are the Priuralye, Nikolayevka and Garazhovka (Donetsk Basin) and Bogoslovsk, all of which are

Ladinian to Carnian in age (Dobruskina.I.A., 1995, I.A., 1994).

Correlation with the Priuraly flora has been based on the presence of remains of *Filicophyta, Chiropteris, Lepidopteris, Scytophyllum, Vittaephylum*, and *Glossophyllum* in both locations. Correlation to the Nikolayevka and Garazhovka flora of the Donetsk Basin is based on the shared presence of *Neocalamites, Chiropteris, Lepidopteris, Scytophyllum, Vittaephyllum*, and *Glossophyllum* (Dobruskina.I.A., 1995). Correlation with the Carnian Svalbard flora is based on the shared presence of remains attributable to *Glossophyllacea* (Dobruskina.I.A., 1995).

The Scytophyllum Flora has been attributed to the Cortaderitian Stage of Gondwana due to similarities in floral assemblages, particularly the abundance of *Scytophyllum* (Morel et al., 2003). The Cortaderitian Stage is divided into 3 biozones, the middle one of which was dated 228.5 Ma ± 0.3 Myr by Rogers et al. (Rogers et al., 1993), supporting the Ladinian–Carnian age for the Scytophyllum flora and, therefore, the Madygen Formation. A maximum constraint on the age of the Madygen Formation can also be derived from Gondwanan floral stages. Specifically, the Puesto Viejo Formation, part of the Barrealian Stage underlying the Cortaderitian Stage (and therefore the Scytophyllum Flora), has been dated radiometrically to 232 Ma ± 4 Myr (Valencio.D. et al., 1975) . This would suggest that the Cortaderitian Stage, Scytophyllum flora and Madygen Formation can be no older than 236 Ma.

Given the evidence for a Ladinian – Carnian age for the Madygen Formation, a minimum constraint on the age of Hymenoptera can be established on the minimum age interpretation for the Carnian-Norian Boundary, 228.4 Ma ± 2 Myr (Gradstein.F. et al., 2012; though there remains uncertainty over the definition of this boundary), thus, 226.4 Ma.

Minimum – 226.4 Ma

Maximum – 236 Ma

Ronquist et al. 2012 – 235 Ma

Turga Formation

The age of the Turga Formation of Transbaikalia, Siberia, is not known with any great degree of accuracy given an absence of reliable stratigraphic markers (Friis.E. et al., 2011) . Thus, any

calibration based on fossil taxa from this deposit must rely heavily upon the weak biostratigraphic correlations available to better-dated formations located elsewhere. The Turga Formation has been estimated loosely as early Cretaceous but there is the possibility of a late Jurassic age for the lower part of the Formation (Vakhrameev.V., 1991).

Rasnitsyn and Quicke (Rasnitsyn.A. and Quicke.D., 2002) quote unpublished radiometric dates of 134 Ma ± 2 Myr and 131 Ma ± 5 Myr for the Turga Formation using Kr-Ar and Rb-Sr methods respectively. However, these have not been substantiated and so they must be discounted. Otherwise, the Turga Formation has been correlated with the Baissa Formation on the basis of a similar faunal (*Ephemeropsis* abundance (Zherekhin.V., 1978)) and floral (shared presence of *Asteropollis*; (Godefroit.P., 2012); (Vakhrameev.V. and Kotova.I., 1977)) assemblage. Thus, the age of the Baissa Formation has been estimated loosely as ranging from Late Jurassic (145 Ma ± 4 Myr; (Kopylov.D., 2010)) to Barremian (130 Ma ± 1.5 Myr to 125 Ma ± 1 Myr), though the evidence substantiating this is weak.

The Baissa Formation has in turn correlated with the Purbeck Formation of England (Rasnitsyn.A. and Quicke.D., 2002, Zherekhin.V., 1978, Zherikhin.V. et al., 1998) on the basis of the presence of the hymenopteran subfamily Bassinae in both deposits. The giant Mayfly *Ephemeropsis*, (Godefroit.P., 2012, Zherikhin.V. et al., 1998) *Trematorax baissensis* and three other members of the *Trematorax* genus are common to both deposits (Rasnitsyn et al., 1998, Rasnitsyn.A., 1988), and have been exploited in establishing a biostratigraphic correlation. The Tithonian-Berriasian (140.2 Ma ± 3 Myr) boundary is thought to lie to the base of the Purbeck Formation (Allen and Wimbledon, 1991), although magnetostratigraphy has suggested that the true location of this boundary may lay between the Purbeck Formation and ostracod-rich freestone that is positioned below (Ogg.J. et al., 1994). However, Hymenopterea are poor biostratigraphic markers and correlations based upon them are unlikely to have fidelity over such vast paleogeographic distances.

A more reliable biostratigraphic correlation can be drawn from the presence of the early angiosperm *Asteropollis* in the Baissa and Turga formations (Vakhrameev.V. and Kotova.I., 1977), which has a well-characterised global distribution (Friis et al., 2005). *Asteropollis asteroides* is a broadly defined species of *Asteropollis* and, as a result, its age range may be no better constrained than that of the genus. The oldest instances of *Asteropollis* pollen (Martinez et al., 2013) occur in Portuguese coastal sections, most notably associated with a female flower likely related to the extant genus *Hedyosmum*, which possesses pollen extremely similar to that

of *Asteropollis* (Friis et al., 1999). *Asteropollis* has also been found in a number of contemporaneous (considered so due to high biostratigraphic similarities; (Friis et al., 1999)) formations in Portugal and is dated to the Barremian or Aptian based on the biostratigraphic observations of Friis et al. (Friis et al., 1999).

A more precise assessment of the age of these formations can be derived from palynological observations made by Heimhofer et al. (Heimhofer.U. et al., 2007) of a number of chronologically diagnostic dinoflagellate species in deposits from the Lusitanian basin (Cresmina section, to which the floral sites investigated by Friis et al. (Friis et al., 1999) are attributed). The first occurence of the dinoflagellate species *Cerbia tabulata* is at the base of the Cresmina section, *C. tabulata* is indicative of the Early-Late Barremian boundary (Stover.L.E. et al., 1996), and is usually found just below this point in time, suggesting a mid-Barremian age for the base of this Formation. Thus, a maximum age constraint on the first appearance of *Asteropollis* can be established on the base of the Barremian, 130.8 Ma ± 0.5 Myr (Gradstein.F. et al., 2012), thus, 131.3 Ma.

Asteropollis does not appear in the fossil record after the Early Campanian, with the latest instance observed in sections in Antarctica (Martinez et al., 2013, Dettmann and Thomson, 1987). This last occurrence of *Asteropollis* co-occurs with the Ammonite species *Submortoniceras chicoense* which is indicative of the Lower Campanian (Haggart, 1984) and the dinoflagellate *Xenikoon australis* (Dettmann and Thomson, 1987), which is indicative of the *X.australis* biozone, which is dated to the Campanian (Helby.R. et al., 1987). Thus, a minimum age constraint on the last appearance of *Asteropollis* pollen can be established from the age of the end of the Campanian, 72.1Ma ± 0.2 (Gradstein.F. et al., 2012), thus, 71.9 Ma.

Minimum – 71.9 Ma

Maximum – 131.3 Ma

Ronquist et al. 2012– 130 Ma

Baissa / Zaza Formation

The Zaza Formation of Baissa, Transbaikalia, Siberia, can be correlated with the Turga Formation, also of Transbaikalia, based on the shared presence of key components of each formations respective floral assemblage. The most notable similarity between these floral

assemblages is the shared presence of *Asteropollis asteroides*, *Dicotylophyllum pusilum, Baisa hirsuita, Podozamites, Schizolepis, Pseudolarix, Phoenicopsis, Czekanowskia rigida* and *Sphenobaiera* (Godefroit.P., 2012, Vakhrameev.V. and Kotova.I., 1977, Vakhrameev.V., 1991, Krassilov, 1986). The age of the Turga flora and Formation has been discussed previously and is based on the chronological distribution of *Asteropollis* type pollen, but correlation with the Yixian Formation of China is also supported strongly (Godefroit.P., 2012), allowing for refinement of the *Asteropollis*-derived ages. Correlation between Turga and Yixian is based on similarities in the floral assemblages of these two formations, with the shared presence of the species *Baisa hirsuita, Botrychites reheensis, Neozamites verchojanensis, Pityolepis pseudotsugaoides, Brachyphyllum longispicum, Scarbugia hilii, Ephedrites chenii, Carpolithus multiseminalis, Carpolithus pachythelis, Schizolepis, Baiera, Coniopteris, Ginkoites, Pityocladus , Pityospermum* and *Elatocladus* (Godefroit.P., 2012, Chen.P. et al., 2005, Krassilov, 1986).

The overlap in the floral assemblages is due to the fact that all of these formations are members of the Jehol Biota and therefore are closely related in composition and age (Godefroit.P., 2012). To derive a maximum and minimum constraint for the Zaza Formation of Transbaikalia it is necessary to incorporate evidence of the age of all correlated formations. The shared presence of *Asteropollis asteroides* in Turga and Zaza allows for the use of this palynomorphs chronological range to determine a maximum and minimum age for Zaza. *Asteropollis* first appears in the fossil record in coastal Portugal and is dated to roughly 127.8 Ma (Friis et al., 1999, Martinez et al., 2013), the last appearance of *Asteropollis* is in Antarctica (Dettmann and Thomson, 1987) and is dated to the end-Campanian at the latest 72.1Ma ± 0.2 (Gradstein.F. et al., 2012); in-depth discussion on the subject of the chronological range of *Asteropollis* is presented in the calibration justification for the Turga Formation.

Incorporation of data related to the chronological range of the Yixian Formation measured with radiometric methods allows for the refinement of the age suggested by the palynological composition of Zaza. A brief outline of the dating of the Yixian Formation is presented here; a more in-depth discussion is presented in the calibration justification for the Laiyang Formation. The base of the Yixian Formation, the Lujitan Bed, has been dated through the use of the $^{40}Ar/^{39}Ar$ radiometric method, providing a maximum constraint on the age of this Formation of 128.6 Ma (Zhoue.Z., 2006, Wang.S. et al., 2001, Zhou et al., 2003). This age is consistent with the first appearance of *Asteropollis* (Martinez et al., 2013, Friis et al., 1999).

The Yixian Formation is correlated with the Laiyang Formation of Lianoning, China through the components of both insect and floral assemblages (see Laiyang Formation calibration for details). Therefore, the age of the base of the Formation which overlies the Laiyang Formation, the Houkang Formation of the Qingshan Group, can be utilized as a minimum constraint on the age of the Yixian Formation and therefore the Zaza Formation. U-Pb zircon dating from the Houkang Formation has yielded an age of 106 ± 2 Ma (Ling et al., 2007), providing a minimum constraint of 104 Ma for the Zaza Formation.

Minimum – 104 Ma

Maximum –128.6 Ma

Ronquist et al 2012 – 140 Ma

Daohugou Bed / Jiulongshan Formation

The Daohugou bed has produced a rich selection of fossil insects (Rasnitsyn and Zhang, 2004), plants and vertebrates (Wang.X. et al., 2005) and there have been numerous attempts to date this stratum and its surrounding strata .. The formation that the Daohugou bed belongs to has been the subject of considerable debate. It is currently thought that the Daohugou Bed belongs to the Middle Jurassic Jiulongshan Formation (Gao and Ren, 2006), but other researchers believe that it belongs to the Tiaojishan Formation (Liu et al., 2006) or the Early Cretaceous Yixian Formation (Wang.X. et al., 2005). Complex stratigraphy caused by possible overturning of the sequence (He.H.l. et al., 2004) and instances of unconformities (Gao and Ren, 2006) make it difficult to constrain from among these possibilities.

Shen et al. (Shen.Y. et al., 2003) noted that the conchostracans found in the Daohugou Bed belong to the Bajocian-Bathonian Euestheria ziliujingensis fauna and that the conchostracan species *E. luanpingensis* is found in both the Daohugou Bed and the Jiulongshan Formation, suggesting that the Daouhugou Bed belongs to the Jiulongshan Formation. However, the conchostracan species found in Daohugou are notably different from species of the Eosestheria fauna of the Yixian Formation suggesting that Daohugou does not belong to this Formation (Shen.Y. et al., 2003). The Bajocian - Bathonian age range for the Eustheria ziliujingensis fauna is concordant with radiometric dates established for Daohugou (Liu et al., 2006)(discussed below).

Ren et al. (Ren.D. et al., 2002) demonstrated similarities between the Jiulongshan Formation and Dahougou Bed insect assemblages and noted that, based on biostratigraphic inference, the age of the Daohugou Bed was not Early Cretaceous and could not, therefore, be assigned to the Yixian Formation. The presence of *Ephemeropsis* in the Daohugou Bed was used previously to support an Early Cretaceous age, but it was shown that this was actually misidentified *Mesoneta* (Ren.D. et al., 2002), which is known from the Bathonian in Mongolia (Sinitshenkova.N., 1985).

The Daohugou Bed contains none of the most indicative early Cretaceous hymenopterans, suggesting that it should not be assigned to the Early Cretaceous Yixian Formation, as proposed by Wang et al. (Wang.X. et al., 2005). Instead, Rasnitsyn and Zhang (Rasnitsyn and Zhang, 2004) argue that the Daohugou Bed should be assigned a Middle to Late Jurassic age due to an overlap in hymenopteran assemblages with the Karatau locality. The genera Xyelidae, Siricidae, Xyelydidae, Anaxyelidae, Mesoserphidae, Megalyridae, Praeaulacidae are found in both locations and are also among the most abundant (Rasnitsyn and Zhang, 2004).

SHRIMP U-Pb Zircon dating on a number of samples from the Daohugou Biota and the strata lying both above and below showed that samples positioned above the famous fossil salamander bearing layers at Reshuitang (which are overlain by the bottom of the Daohugou Bed) could be dated to 164 Ma ± 4 Myr, and the youngest possible date for strata overlaying the Daouhugou Bed was observed in the Xiaoxigou-Xiaoliangqian section, at the bottom of the layer overlying the Daohugou Bed which has been dated to 152 Ma ± 2.3 Myr (Liu et al., 2006).

Alternately, Chang et al. (Chang et al., 2009) used $^{40}Ar/^{39}Ar$ dating on two tuffs to date the very bottom of the Lanqui Formation, which lies just above the Haifanggou Formation. The Lanqui and Haiffangou formations are situated in Liaoning Province, the same formation is referred to as the Jiulongshan Formation in Hebei province and therefore Haiffangou is equivalent to Jiulongshan. This relationship means that the oldest possible age for the Lanqui Formation can be used as a minimum constraint on the age of the Jiulongshan Formation. The youngest tuff measured by Chang et al. (Chang et al., 2009) was dated to 158 Ma ± 0.6 Myr, yielding a minimum constraint of 157.4 Ma for the age of the Jiulongshan Formation.

Minimum – 157.4 Ma

Maximum – 168 Ma

Ronquist et al. 2012 – 161 Ma

Karatau Locality (Kulbastau/Galkino Provenance)

The Karatau Formation consists of a group of deposits located in Jambul Province, Kazakhstan. The most notable sites are Aulie (also known as Mikhailovka), Karabastau, and Uspenovka (formerly Galkino), located within the Kulbastau Mountain Range. The floral composition of the strata comprising the Karatau Mountain Range is well documented and specific floral assemblages have been identified (Doludenko and Orlovskaya, 1976). The Karabastausky floral assemblage was identified initially at the Karabastau site and the comparison of floral assemblages at Galinko allowed this site also to be assigned to the Karabastausky assemblage (Doludenko and Orlovskaya, 1976). One of the characteristics of the Karabastausky flora is an abundance (95-100% of the floral assemblage) of *Classopollis* pollen (Doludenko and Orlovskaya, 1976). Vakhrameev (Vakhrameev.V., 1991) analysed the fluctuations in *Classopollis* abundance across Eastern Europe and Asia and compared them with major geological events. This analysis showed that *Classopollis* in Kazakhstan, Middle Asia, Ukraine and Crimea only reached abundances of +95% during the Oxfordian and Kimmeridgian, before decreasing rapidly during the Late Kimerridgian – Tithonian. The Karabastausky Assemblage is positioned above the Borolsaisky Assemblage but it is unknown what length of time separates these two assemblages (Doludenko and Orlovskaya, 1976). Despite this, the Karabaustsky Assemblage must be no older than the Borolosaisky and so the minimum age of this assemblage can still be used to constrain the age of the base of the Karabustsky Assemblage. The upper parts of the Borolosaisky Assemblage are considered to be of a Lower to Middle Callovian age as they contain around 50% *Classopollis*. Doludenko and Orlovskaya (Doludenko and Orlovskaya, 1976) and Sakulina (Sakulina.G.V., 1971) has shown that this level of abundance is indicative of an Early Middle Callovian age, whereas higher abundances are indicative of Upper Callovian- Tithonian ages. The Borolosaisky Assemblage reaches a peak *Classopollis* abundance of 50% before dropping back down to 10%, supporting an Early – Middle Callovian age.

Thus, the minimum age of the Karabaustsky Assemblage can be established on 95% abundance of *Classipollis* at the top of the Karatau Formation, which must predate the pre-Late Kimmeridgian reduction in *Classopollis* abundance, which can be dated arbitrarily but

objectively on the base of the Tithonian, viz. 152.1 Ma ± 0.9 Myr (Gradstein.F. et al., 2012) and, thus, 151.2 Ma

A correlation, supported by numerous biostratigraphic similarities, between the Karabastau locality and the Daohugou bed in China has been proposed (Rasnitsyn and Zhang, 2004, Zhang, 2010, Zhang, 2011). Kovalevisargid flies described by Zhang (Zhang, 2011) were noted as having strong similarities to kovalevisargid flies observed in Daohugou deposits. The rarity of these flies coupled with their lack of diversity in the fossil record substantiates the correlation between these localities (Zhang, 2011). Furthermore, *Pterosagus* found in the Daohugou Biota has similar wing venation to *Nagotomukha karabas* from the Karabastau locality (Zhang, 2010). Representatives of Archisargidae, such as *Archirhagio, Archisargus, Mesosolva* and *Calosargus*, have been retrieved from both the Karabastau and Daohugou localities (Zhang, 2010). Further evidence for a correlation between Karatau and Daohugou is the presence of *Protoscelinae* leaf beetles, known only only from these two localities (Zhang, 2005b). It has been suggested that *Protoscelinae* existed only for a relatively short time and over a small geographic range (Zhang, 2005a), providing further evidence for a correlation between Daohugou and Karatau.

The correlation between these two sites allows the radiometrically derived age for the strata occuring below the Daohugou Biota to be utilised to infer maximum dates on the Karatau Formation and to further refine the age range suggested by Vakhrameev (Vakhrameev.V., 1970). Liu et al. (Liu et al., 2006) used SHRIMP U-Pb zircon dating on a number of samples from the Daohugou Biota and the strata lying both above and below. They showed that samples positioned above the salamander bearing layers at Reshuitang (which are overlain by the Daohugou Biota) could be dated to 164 ± 4 Ma. This date would suggest a maximum age of 168 Ma for this Formation.

Minimum – 151.2 Ma

Maximum – 168 Ma

Ronquist et al. 2012 -161 Ma

Laiyang Formation

On the basis of the shared content of their respective fossil insect assemblages, the Laiyang

Formation of Liaoning, China, can be correlated with the Yixian Formation, also of China and the Zaza Formation, Transbaikalia (Zhang and Rasnitsyn, 2006, Zhang and Rasnitsyn, 2004). Nine species of Pelecinidae wasps are found in both the Laiyang, Zaza, and Yixian formations (*Iscopinus baissicus, Sinopelecinus delicatus, S. epigaeus, S. magicus, S. viriosus, Eopelecinus vicinus, E. shanyuanensis, E. similaris*, and *Scorpiopelecinus versatilis*; (Zhang and Rasnitsyn, 2006)). Zhang and Rasnitsyn (Zhang and Rasnitsyn, 2006) consider that any difference in the assemblages at Yixian and Laiyang are due to taphonomic processes as opposed to chronological differences.

The Yixian and Laiyang formations can also be correlated by their shared floral assembalges; both formations contain members of the genera *Brachyphyllum, Cupressinocladus* and *Schizolepis* (Hu.C. et al., 2001, Chen et al., 2005). The Yixian and Laiyang formations both contain *Classopollis parvus* and *Solenites murrayama*, in addition to members of the genera *Cedripites* and *Cicatricosisporites* (Chen et al., 2005, Hu.C. et al., 2001). These floral and palynological remains are found in the lower beds of the Yixian Formation, the Jianshangou beds (Chen et al., 2005), providing support for the correlation of the base of the Yixian Formation with the base of the Laiyang Formation. Choncostracans attributed to *Yanjiestheria* are found in both the Yixian (Lujiatun Bed) and Laiyang formations (Chen et al., 2006, Chen et al., 2005). The ostracod genus *Cypridea* is found in both the Yixian and Laiyang formations (Chen et al., 2006, Chen et al., 2005), as is the bivalve species *Sphaerium anderssoni* and members of the gastropod genus *Probaicalia* (Chen et al., 2006, Chen et al., 2005). The fish *Lycoptera sinensis* is found in both the Laiyang Formation (Chen et al., 2006) and also in the Jianshangou Bed (Chen et al., 2005, Zhou et al., 2003). This species provides a comparatively strong correlation and its presence at the base of the Yixian Formation demonstrates the chronological relationship between the Yixian and Laiyang formations, allowing the Lujiatun Bed (which is the lowermost part of the Jianshangou Bed (Chen et al., 2005)) to be used to derive a date for the base of the Laiyang Formation. The wide range of biostratigraphic sources available to correlate the Yixian and Laiyang formations includes insect, floral, palynological and vertebrate assemblages. While the utility of any single biostratigraphic marker could be called into question, the number and range of sources available for this correlation provides overwhelming support.

The use of $^{40}Ar/^{39}Ar$ radiometric dating on the bottom of the Lujiatun Bed at the very base of the Yixian Formation yields a date of 128.4 ± 0.2 Ma (Zhoue.Z., 2006, Wang et al., 2001, Zhou et al., 2003). This date can be used to give a maximum constraint on the Laiyang Formation of

128.6 Ma. Other attempts to date the base of the Yixian Formation the $^{40}Ar/^{39}Ar$ system (Lo.C. et al., 1999) have yielded dates around 20 Myr older than those estimated by Wang et al. (Wang et al., 2001). These older dates have been considered unreliable as the samples may have contained trapped Argon that may have distorted results (Swisher et al., 2002, Zhou et al., 2003).

The minimum age for the Laiyang Formation is best established from the perspective of the overlying Qingshan Group, the lowest unit within which is the Houkuang Formation, U-Pb dating of zircons within which has yielded a date of 106 Ma ± 2 Myr (Ling et al., 2007). Thus, we establish a minimum constraint on the age of the Laiyang Formation at 104 Ma.

Minimum – 104 Ma

Maximum – 128.6 Ma

Ronquist et al. 2012 – 140 Ma

Bon Tsagan / Khurulit Rock Unit

The Bon Tsagan locality of Central Mongolia can be divided into a number of formations: the Undur-Ukhin Formation is the lowermost unit and is comparable to the Tsagen Tsab Formation of Eastern Mongolia, the middle unit is the Anda-Khuduk Formation and is comparable with the Shin Khuduk Formation in Eastern Mongolia, the uppermost unit is the Khulsyn-Gol Formation. The Khurilt rock unit is assigned to the Anda Khuduk Formation (Krassilov, 1982).

Correlation with the Zaza and Yixian Formations of Transbaikalia and China, respectively, is possible based on shared flora. Shared elements of the Yixian and Bon-Tsagan flora are *Schizolepis, Pseudolarix, Baiera, Sphenobaiera, Phoenicopsis, Ginkgoites, Pityocladus, Pityospermum, Brachyphyllum, Erenia stenoptera* and *Leptostrobus* (Krassilov, 1982, Cao et al., 1998, Chen et al., 2005). Krassilov (Krassilov, 1982) assigned the Anda Khuduk and Shin Khuduk formations to the *Baierella hastate* phytostratigraphic unit, which was thought to be Aptian in age. The palynological assemblage at Shin Khuduk is dominated by *Pinaceae* pollen (Krassilov, 1982); *Pinaceae* plants are also found in the Yixian Formation, (Shang et al., 2001) further supporting a correlation between these two localities. Shared flora between Yixian and Bon Tsagan are distributed between all three subformations of the Bon Tsagan unit (and their

related formations in Eastern Mongolia) (Krassilov, 1982), allowing the age of the Yixian Formation to inform the age of the strata at the Bon Tsagan locality. With respect to shared components of the insect assemblages of these formations, the Caloblattinidae genus *Nuurcala* is of particular note. *Nuurcala obesa*, found in the Yixian Formation is considered closely related to *Nuurcala popovi* (Wang and Ren, 2013) found in the Anda Khuduk of the Khurilt Unit at Bon Tsagan (Vrsansky, 2003), further suggesting correlation between these locations.

The use of $^{40}Ar/^{39}Ar$ radiometric dating on the bottom of the Lujiatun Bed at the very base of the Yixian Formation yields a date of 128.4 ± 0.2 Ma (Zhoue.Z., 2006, Wang et al., 2001, Zhou et al., 2003). This date can be used to provide a maximum constraint on the Bon Tsagan locality of 128.6 Ma. Other attempts to date the base of the Yixian Formation with the $^{40}Ar/^{39}Ar$ method have yielded dates around 20 Ma older (Lo.C. et al., 1999) than those measured by Wang et al. (Wang et al., 2001), though these older dates are considered to be unreliable, as the samples measured may have contained trapped Argon, which can distort results (Swisher et al., 2002, Zhou et al., 2003).

The Yixian Formation is succeeded by the Jiufontang Formation (He.H.l. et al., 2004), and, therefore, the oldest possible date for the Jiufontang Formation can be utilised as a minimum constraint on the age of the Yixian Formation and therefore the Bon Tsagan Locality. The use of $^{40}Ar/^{39}Ar$ dating on a number of samples from the Jiufontang Formation allowed an age of 120.3 ± 0.7 Ma to be assigned to volcanic tuffs present in the Formation (He.H.l. et al., 2004). Whilst these tuffs do not exist at the very bottom of the Jiufontang Formation, they may still be used to derive a minimum constraint on the Yixian Formation as it must be no younger than 119.6 Ma, which is concordant with the Aptian age for the flora of the Baierella hastate unit.

In the petrified Suihent Forrest of South Eastern Mongolia, the lower Tsagen Tsaab Formation (also called Tsagen Tsaav), whose comparison to Undur-Ukhin has previously been discussed, is dated to 156 ± 0.76 Ma (Late Jurassic) through the dating of volcanic tuffs using the $^{40}Ar/^{39}Ar$ method (Keller and Hendrix, 1997) . This estimate pushes the constraint on the age of the base of the Bon Tsagan unit back to 156.76 Ma, although the vast majority of the Formation is much younger than this and likely Cretaceous in age (Rothwell et al., 2012).

Minimum - 119.6 Ma

Maximum - 156.76 Ma

Ronquist et al. 2012 – 121 Ma

Dolgan Formation / Agapa

The Dolgan (also referred to as Dolganskaya Formation) of the Nizhnyaya Agapa river locality in Northwest Siberia has yielded numerous biostratigraphic markers that allow the age of this Formation to be identified as early Upper Cretaceous. The Dorozhkov Member overlies the Dolgan Member; both of these deposits contain *Inoceramuae* species that strongly support the Cenomanian – Turonian boundary as the maximum constraint on the age of the Dolgan Member. *Inoceramus pictus* is found in the Dolgan Member, and the first appearance of *I. pictus* corresponds with the top of the *Acanthoceras jukesbrowni* zone (Rohstoffe), which suggests a middle to upper Cenomanian age (Caldwell.M.W. and Cooper.J., 1999). The Dorozhkov Member contains the lower Turonian species *Inoceramus labiatus*. The boundary between the *I.pictus* and *I.labiatus* zones is recognised as the Cenomanian – Turonian boundary in Siberia (Sahagian.D. et al., 1994, Zakharov.V. et al., 2002, Birkelund et al., 1984), which lends strong support to the use of the Cenomanian – Turonian boundary as the minimum constraint on the age of the Dolgan member. *I.labiatus* is a member of the subgenus Mytiloides (Zakharov.V. et al., 2002), which is a common find after the first appearance of the lower Turonian ammonite species *Watinoceras devonense* (Bengston.P., 1996, Zakharov.V. et al., 2002) adding further support to the use of the Cenomanian – Turonian boundary. The Cenomanian – Turonian boundary is dated to 93.9 Ma ± 0.2 (Gradstein.F. et al., 2012).

The palynological evidence to place the Dolgan Member in the Cenomanian stage is also strong. The species *Balmeisporites glenelgensis* is found in both the very base of the Dolgan Member and also the Upper Cretaceous deposits of Victoria, Australia

(Cookson.I. and Dettmann.M, 1958)(Lebedev and Zverev. 2003). *Balmeisporites glenelgensis* is also found in the middle to late Cenomanian deposits of the Peace River of North Western Alberta, Canada (Hu et al., 2008) and the Sargeant Bluff and Stone Park lignite of Iowa and Nebraska (Hu et al., 2008). *Balmeisporites glenelgensis* is found in the Raritan Formation of New Jersey, the age of this Formation is disputed but the majority of researchers consider it to be Cenomanian in age due to the presence of numerous biostratigraphic markers indicative of a lower late Cretaceous age (Kimyai.A., 1966); notably *Arcellites*, which is found in the Cenomanian deposits of Disko Island, Greenland (Miner.E., 1935) and Grill Coal, Iowa (Schemel, 1950).

The presence of the microspore *Schizosporis sp.* (Lebedev and Zverev. 2003) can also be used to infer the age of the Dolgan Formation as *Schizosporis* has been found in boreholes from Cenomian deposits in Australia (Cookson.I. and Dettmann.M, 1959). Similarly, the presence of moss spores *Stereisporites (spp.)* (Lebedev and Zverev. 2003) further support a Cenomanian age for the Dolgan Formation. *Stereisporites* occurs alongside the dinocyst species *Epelidosphaeridia spinosa* in 4 boreholes in the Bohemian Cretaceous Basin (Cech.S. et al., 2005). *E. spinosa* is considered to be Cenomanian in age, and it has been found in deposits of early, middle and late Cenomanian ages (Cech.S. et al., 2005). *E. spinosa* was found in the lower and middle Cenomanian of Northern Europe (*Mantelliceras dixoni* and *Acanthoceras rhotomagense* zones respectively) (Paul et al., 1994, Cech.S. et al., 2005, Robazynski.F. and Caron.M, 1988).

The presence of *Taxodiaceaepollenites hiatus* pollen at Dolgan (Lebedev and Zverev. 2003) suggests a much older age than The Cenomanian due to the appearance of this species in the *Dicheiropollis-Classopollis-Cicatricosisporites* assemblage of the Suowa of the Qinghai-Xizang Plateau of China (Li and Batten, 2004). The presence in this assemblage of *Dicheiropolli*s suggests an Early Cretaceous age (Hochuli, 1981). Li and Batten (Li and Batten, 2004) suggested a Valanginian – Berremian age for this assemblage based on the presence of *Dicheiropollis, Cicatricosisporites, Concavissimisporites, Impardecispora, Lygodioisporites,* and *Pilosisporites*.

By considering the zonation of palynological assembalges of the Pacific coast of Russia, a similar pre-Cenomanian age can be suggested for the Dolgan Formation based on the presence of *Taxodiaceaepollenites hiatus*. *Taxodiaceaepollenites hiatus* is assigned to the *Gleicheniidites carinatus* - *Pilosisporites echinaceus* palynozone (Markevitch, 1994) which is considered to be Valanginian in age, confirmed by similarities to Valanginian floras which have their date confirmed by the presence of fossil molluscs (Markevitch, 1994). This support for a Valanginian age for elements of the palynological assemblage of Dolgan allow the use of the Beriasian – Valanginian boundary as a maximum constraint on the age of the Dolgan Formation, which has been dates to 139.4 Ma ± 0.7 Myr.

Minimum – 93.7 Ma

Maximum – 140.3 Ma

Ronquist et al. 2012 – 94 Ma

Novospasskoye/Ichetuy

The Novospasskoye (also referred to as Novospasskoe) locality of the volcanogenic Ichetuy (Ichetui) Formation is located outside the town of Novospasskoye in the Tugny (Tugni) Depression, Transbaikalia, Russia (Kotov, 2007). It has proven difficult to determine an exact age for this Formation, as demonstrated by the disparity in age estimates formed from interpretations of the insect assemblage and estimates derived from the radiometric dating of the volcanic sediment (Metelkin.D. et al., 2007). The accuracy of any age based purely on the structure of the insect assemblage must be treated with caution, as it has been demonstrated that the assemblage contains chronological anomalies, including the find of a probable early Cretaceous Coptoclavid beetle *Bolbonectes* alongside the Jurassic beetle species *Stygeonectes jurassicus* (Ponomarenko.A., 1993). Therefore, a less subjective method of dating this formation is preferred; the volcanogenic nature of this sediment allows for the use of radiometric dates acquired from dating volcanic events in this locality. The Ichetuy Formation is thought to be of an Early-Middle Jurassic age on the basis of the biostratigraphic composition of the Formation (Ponomarenko.A., 1993, Skoblo.V. and Lyamina.N., 1965, Rasnitsyn.A. and Quicke.D., 2002), although radiometric dating of volcanogenic material suggests a younger age than this (Metelkin.D. et al., 2007). A K-Ar date of 145 Ma ± 4 Myr derived from a basalt overlying the Ichetuy Formation (Ivanov.V. et al., 1995) provides for an effective minimum constraint on its age, thus 141 Ma.

The oldest date obtained from the measurement of volcanogenic material at the Ichetuy Formation is 162 ± 6 Ma (Donskaya et al., 2013, Shadaev.M. et al., 1992) and was measured using the Rb-Sr whole rock method. This pushes the oldest possible age for the Ichetuy Formation back to 168 Ma.

Minimum – 141 Ma

Maximum – 168 Ma

Ronquist et al. 2012 – 176 Ma

Ola Formation

The Ola Formation of the Arman' and Ola rivers interfluve, North Eastern Russia, can be correlated with the Barykov Formation of the Amaam Lagoon, North Eastern Russia, on the

basis of shared floral assemblages. Both of these formations contain the angiosperm species *Macclintockia beringiana*, (Moiseeva, 2011) which is considered to be indicative of the Barykovsk floral assemblage (Vakhrameev.V., 1991, Herman, 2007). This floral assemblage is dated based on the presence of fossil bivalves in the Barykova Formation at Ugolnaya Harbour, which is assigned to the Barykovsk floral assemblage. At Ugolnaya bay the Barykova Formation can be split into four sections, the uppermost of which is terrestrial with the three underlying sections considered to be marine (Moiseeva, 2011). The date of the uppermost marine section can be utilised as a maximum constraint on the age of the Barykov Formation, as all preceding deposits are marine and the hymenopteran species of Ronquist et al. (Ronquist et al., 2012a) are lacustrine. A date for the uppermost marine section can be inferred from the presence of the bivalve species *Inoceramus Patootensis* (Moiseeva, 2011, Vakhrameev.V., 1991), which is considered to be indicative of middle Santonian to base Campanian age (Jones.D. and Gryc.G., 1960) and so we can derive a maximum age on the base of the Santonian 86.3 Ma ± 0.5 Myr (Gradstein.F. et al., 2012). The terrestrial section of the Barykov Formation can be further divided into three more sections consisting of two coal containing beds positioned above and below a bed lacking coal deposits; the contents of the coal containing beds are made of elements of the Barykov floral assemblage, and are the location of fossilised *Macclintockia* (Moiseeva, 2011).

The Koryak Formation overlies the Barykov Formation (Moiseeva, 2011), and therefore its oldest possible age can be utilised as a constraint on the minimum age of the Barykov Formation. The middle section of the Koryak Formation contains the bivalve species *Inoceramus pilvoensis* and *Patagiosites alaskensis*, which suggest an base Maastrichtian age 72.1 Ma ± 0.2 Myr (Gradstein.F. et al., 2012) for the middle of the Koryak Formation (Moiseeva, 2008), which can be utilised as a minimum constraint on the age of the Barykov Formation and therefore the Ola Formation.

Minimum – 71.9 Ma

Maximum – 86.8 Ma

Ronquist et al. 2012 – 140 Ma

Unda / Glushkovo Formation

The age of the Glushkovo Formation is considered to be either Late Jurassic or Early Cretaceous on the basis of the composition of its insect assemblage (Ponomarenko.A., 1993, Rasnitsyn.A. and Quicke.D., 2002). The stoneflies found within the Glushkovo Formation would suggest an Early Cretaceous age due to the presence of species, including *Dimoula dimi* (Sinitshenkova.N, 2005), not found in Jurassic strata. Furthermore, numerous modern sandfly species not found in Jurassic deposits are found in Glushkovo, supporting a Cretaceous age for this Formation (Sinitshenkova.N, 2005).

Ignatov et al. (Ignatov.M. et al., 2011) correlate the Glushkovo Formation with the Baigul locality of Transbaikalia on the basis of the shared presence of the insect species *Proameletus caudatus*, members of the taxonomic group *Isophlebiidae*, and the *Equisetaceae* plant species *Equisetum undense*, in addition to the crustacean species *Prolepidurus schewija* and *Chirocephalus rasnitsyni*. The strata at the Baigul locality can, in turn, be correlated with the Ulugey (Ulugei) Formation of Mongolia due to the shared presence of the moss genera *Bryokhutuliinia* (Newton.A. and Tangney.R., 2007), which is one of only five known taxonomic groups of Jurassic mosses (Ignatov.M. et al., 2011). There is a paucity of literature regarding the precise age of the Ulugey Formation, with most estimates of its age listed as Late Jurassic – Early Cretaceous (Menier et al., 2004, Legalov, 2010a, Legalov, 2010b, Gratshev and Aeranob, 2009). *Pityospermum sp.* was also found in Baigul (Ignatov.M. et al., 2011) but the lack of a species level classification of this specimen makes it a poor quality biostratigraphic marker as the genus *Pityospermum* is known from the Permian to the Late Cretaceous (Spicer et al., 2002, Henan., 1989).

Members of the coleopteran Genus *Gobicar* are found at Khutuliyn-Khira locality of the Ulugey Formation and also in the Montsec Fauna of Spain (Legalov, 2010a), allowing for a correlation between these sites for the purpose of this calibration. The Montsec locality has been dated as late Berriasian at the earliest and Aptian at the very latest (Ansorge.J., 1993a, Peybernes.B. and Oertli.H., 1972). The presence of the freshwater Ostracods (*Cypridea .sp*) in Montsec suggests a Berriasian – Valenginain age as similar *Cypridea* is also known from the Berriasian – Valenginain boundary of Aquitaine, France (Peybernes.B. and Oertli.H., 1972, Oertli.H., 1963, El Albani et al., 2004). This age range supports the fossil insect evidence that suggests an early Cretaceous age for Glushkovo. Despite this, Peybernès and Oertli (Peybernes.B. and Oertli.H., 1972) offer no solid reason to assume that the *Cypridea* present in Montsec is of the same species as that found in Aquitaine. If the uncertainty in the classification of the species of *Cypridea* found at Montsec is taken into account then this find can only be

confidently considered indicative of an age within the full temporal distribution of *Cypridea*. This ostracod genus has a constant presence from the Middle Jurassic to Late Cretaceous and is seen from the Bathonian (Pozo.J., 1971) until the Maastrichtian (Langston, 1975), allowing for the end of the Masstrichtian (66 ± 0.05 Ma (Gradstein.F. et al., 2012)) to be utilized as a minimum constraint on the age of this Formation . If the lack of insect fossils from the Jurassic in Glushkovo is assumed to indicate that this Formation is of a Cretaceous age only then the maximum constraint on the age of this Formation can be placed at the base of the Cretaceous 145.0 Ma ± 0.8 Myr (Gradstein.F. et al., 2012), as it is known that *Cypridea* extends well into the Jurassic.

Minimum – 65.9 Ma

Maximum – 145.8 Ma

Ronquist et al. 2012 – 146 Ma

Sogul Formation / Sagul

The Sogul Formation is located in the Osh Region of Kyrgyzstan, a handful of sites belonging to this Formation with similar geology are referred to as the Say-Saigul (also known as Sai-Sagul, Shurab 3, or Svodovoe Ruslo) locality.

The age of this locality has not firmly been established due to the endemic nature of the faunal assemblage present (Yang et al., 2012) and a relative lack of literature related to the floral assemblage; despite this, the insect assemblage has been used to infer a late Early Jurassic to early Middle Jurassic age for Sagul (Sukacheva and Rasnitsyn, 2004). The presence of the wood morphogenus *Xenoxylon* can be used to infer a rough age range for the Sagul locality. The species *X. barberi, X. hopeiense, X. latiporosum*, and *X. suljuctense* are all found in the Sogul Formation (Philippe and Thevenard, 1996). *X. barberi* is also found in the Aalenian–Bajocian (Thies 1989) of Germany and Toarcian of France (*serpentinium* zone, equivalent to *falciferum* zone (180.2 – 183 Ma, discussed in Toarcian calibrations)), *X. hopeiense* is found in the Middle Jurassic (Aalenian–Bajocian (Nosova, 2013)) of the Angren Formation of Uzbekistan. *X. latiprosum* covers a large chronological range, from the Late Triassic to the Early Cretaceous, but a number of examples of this species have been found in Early and Middle Jurassic deposits, *X. suljuctense* is only known from the Sogul Formation (Philippe and

Thevenard, 1996).

The presence of the species *Shurabia angustata* at Sagul can be used to make a loose correlation with the Ust'-Balei locality of Transbaikalia, where this species is also found, and allows the age of this Formation to influence the calibration of fossils found in Sagul.

The floral assemblage of Ust'-Balei can be correlated with the Tolyinsk suite found to the West of Lake Baikal due to the shared presence of *Sphenobaiera longifolia* (Vakhrameev.V., 1991, Zhou, 2010). There is no consensus on the age of this suite as it is thought to be either Bathonian – Callovian or Lower Oxfordian in age; the suite can be no younger than Upper Oxfordian though, as the top of this suite contains Upper Oxfordian marine sediments (Vakhrameev.V., 1991). If we assume that this suite must exist within this range of ages we can derive a Bathonian to Oxfordian – Kimmeridgian age for Sagul based purely on the floral assemblage.

If both fossil insect and fossil floral evidence is considered then the Sogul Formation can be assumed to have occurred during the period from the base of the Toarcian (which is better dated than the very slightly younger *falciferum* zone (c. 182.7 Ma ± 0.7 Myr), to the end of the Oxfordian (157.3 Ma ± 1.0) (Gradstein.F. et al., 2004).

Minimum – 156.3 Ma

Maximum – 183.4 Ma

Ronquist et al. 2012 – 176 Ma

Bascharage

The Bascharage locality of Luxembourg is though to be Toarcian in age as fossil insects retrieved from these deposits co-occur with the ammonite species *Harpoceras falciferum* (Szwedo, 2011). *H. falciferum* is a strong biostratigraphic marker and can be utilised to infer a strongly supported age for these deposits.

The *falciferum* Boreal ammonoid zone is the second earliest of the Toarcian, time-equivalent to the Tethyan *serpentinum* ammonoid zone, the base of which has been dated to 181.7 Ma

(Gradstein.F. et al., 2012). Both the *falciferum* and *serpentinum* zones are succeeded by the *bifrons* ammonoid zone, the base of which has been dated to 180.36 (Gradstein.F. et al., 2012), both of which have attendant errors of 0.7 Myr. This affords an age range of 182.4-179.66 Ma for the strata the Bascharage Locality.

Minimum – 179.66 Ma

Maximum – 182.4 Ma

Ronquist et al. 2012 – 180 Ma

Dobbertin

The Dobbertin locality of Mecklenburg-Vorpommern, Northern Germany, is widely considered to be lower of lower Toarcian age. Insect finds from this locality are assigned to the *Harpoceras falciferum* ammonoid zone (Vrsansky and Ansorge, 2007, Ansorge.J., 1993b, Krzeminski.W., 1995). The presence of a biostratigraphic marker as strong as the ammonite *H.falciferum* means that a strongly supported age can be inferred for this locality.

The *falciferum* Boreal ammonoid zone is the second earliest of the Toarcian, time-equivalent to the Tethyan *serpentinum* ammonoid zone, the base of which has been dated to 181.7 Ma (Gradstein.F. et al., 2012). Both the *falciferum* and *serpentinum* zones are succeeded by the *bifrons* ammonoid zone, the base of which has been dated to 180.36 (Gradstein.F. et al., 2012), both of which have attendant errors of 0.7 Myr. This affords an age range of 182.4-179.66 Ma for the strata the Bascharage Locality.

Minimum – 179.66 Ma

Maximum – 182.4 Ma

Ronquist et al. 2012 – 180 Ma

Grimmen

The Grimmen Deposit located in Germany is thought to be Toarcian in age. Strong support for this age can be drawn from the presence of Ammonites, indicative of particular zones, within this locality. Notably, *Harpoceras falciferum* is found in the insect bearing layers of this deposit (Ansorge.J., 1999). The presence of this strong biostratigraphic marker means that a well-

supported age range for this deposit can be derived.

The *falciferum* Boreal ammonoid zone is the second earliest of the Toarcian, time-equivalent to the Tethyan *serpentinum* ammonoid zone, the base of which has been dated to 181.7 Ma (Gradstein.F. et al., 2012). Both the *falciferum* and *serpentinum* zones are succeeded by the *bifrons* ammonoid zone, the base of which has been dated to 180.36 (Gradstein.F. et al., 2012), both of which have attendant errors of 0.7 Myr. This affords an age range of 182.4-179.66 Ma for the strata the Bascharage Locality.

Minimum – 179.66 Ma

Maximum – 182.4 Ma

Ronquist et al. 2012 – 180 Ma

Allen, P. & Wimbledon, W. A. 1991. Correlation of Nw European Purbeck Wealden (Nonmarine Lower Cretaceous) as Seen from the English Type-Areas. *Cretaceous Research,* 12, 511-526.

Andryushchenko, S. V., Vorontsov, A. A., Yarmolyuk, V. V. & Sandimirov, I. V. 2010. Evolution of Jurassic-Cretaceous Magmatism in the Khambin Volcanotectonic Complex (Western Transbaikalia). *Russian Geology and Geophysics,* 51, 734-749.

Ansorge.J. 1993a. Bemerkenswerte Lebensspuren Und ?Cretosphex Catalunicus N. Sp. (Insecta; Hymenoptera) Aus Den Unterkretazischen Plattenkalken Der Sierra Del Montsec (Provinz Lérida, Ne - Spanien). *Neues Jahrbuch für Geologie und Paläontologie Abhhandlungen* 190, 19-35.

Ansorge.J. 1993b. Parabittacus Analis Handlirsch 1939 Und Parabittacus Lingula Bode 1953, Neorthophlebiiden (Insecta: Mecoptera) Aus Dem Oberen Lias Von Deutschland. *Palaeontologische Zeitschrift,* 67, 293-298.

Ansorge.J. 1999. Aenne Liasina Gen. Et Sp. N. - the Most Primitive Non Biting Midge (Diptera: Chironomidae: Aenneinae Subfam. N.) - from the Lower Jurassic of Germany. *Polskie Pismo Entomologiczne* 68.

Bengston.P. 1996. The Turonian Stage and Substage Boundaries *In:* RAWSON.P. (ed.) *Proceedings of the Second International Symposium on Cretaceous Stage Boundaries, Brussels, 8-16 September 1995.*

Birkelund, T., Hancock, J. M., Hart, M. B., Rawson, P. F., Remane, J., Robaszynski, F., Schmid, F. & Surlyk, F. 1984. Cretaceous Stage Boundaries Proposals. *Bulletin of the Geological Society of Denmark,* 33, 3-20.

Caldwell.M.W. & Cooper.J. 1999. Redescription, Palaebiogeography,and Palaeoecology of Coniasaurus Crassidens Coniasaurus Crassidens Owen, 1850 Owen, 1850(Squamata) from the English Chalk (Cretaceous; Cenomanian) *Zoological Journal of the Linnean Society,* 127, 423-452.

Cao, Z. Y., Wu, S. Q., Zhang, P. & Li, J. R. 1998. Discovery of Fossil Monocotyledons from Yixian Formation Western Liaoning. *Chinese Science Bulletin,* 43, 230-+.

Cech.S., Hradecka.L., Svobodova.M. & Svabenicka.L. 2005. Cenomanian and Cenomanian-Turonian Boundary in the Southern Part of the Bohemian Cretaceous Basin, Czech Republic. *Bulletin of Geosciences* 80, 321-354.

Chang, S.-C., Zhang, H., Renne, P. R. & Fang, Y. 2009. High-Precision Ar-40/Ar-39 Age Constraints on the Basal Lanqi Formation and Its Implications for the Origin of Angiosperm Plants. *Earth and Planetary Science Letters,* 279, 212-221.

Chen, P. J., Li, J. J., Matsukawa, M., Zhang, H. C., Wang, Q. F. & Lockley, M. G. 2006. Geological Ages of Dinosaur-Track-Bearing Formations in China. *Cretaceous Research,* 27, 22-32.

Chen, P. J., Wang, Q. F., Zhang, H. C., Cao, M. Z., Li, W. B., Wu, S. Q. & Shen, Y. B. 2005. Jianshangou Bed of the Yixian Formation in West Liaoning, China. *Science in China Series D-Earth Sciences,* 48, 298-312.

Chen.P., Wang.Q., Zhang.H. & Al., E. 2005. Jianshangou Bed of the Yixian Formation in West Liaoning, China. *Science in China Series D; Earth Sciences* 48, 298-312.

Cookson.I. & Dettmann.M 1958. "Megaspores" and a Closely Associated Microspore from the Australian Region. *Micropaelontology,* 4, 39-49.

Cookson.I. & Dettmann.M 1959. On Schizosporis, a New Form Genus from Australian Cretaceous Deposits. *Micropaleontology,* 5, 213-216.

Dettmann, M. E. & Thomson, M. R. A. 1987. Cretaceous Palynomorphs from the James-Ross Island Area, Antarctica - a Pilot-Study. *British Antarctic Survey Bulletin*, 13-59.

Dobruskina.I. 1993. Relationships of Floral and Faunal Evolution During the Transition from the Paleozoic to the Mesozoic. *In:* LUCAS.S.G. & MORALES.M. (eds.) *The Nonmarine Triassic.*

Dobruskina.I.A. 1995. Madygen, Triassic Lagerstätte Number One, before and after Sharov. *New Mexico Museum of Natural History Bulletins,* 5, 1-95.

Doludenko, M. P. & Orlovskaya, E. R. 1976. Jurassic Floras of the Karatau Range Southern Kazakhstan Ussr. *Palaeontology (Oxford),* 19, 627-640.

Donskaya, T. V., Gladkochub, D. P., Mazukabzov, A. M. & Ivanov, A. V. 2013. Late Paleozoic-Mesozoic Subduction-Related Magmatism at the Southern Margin of the Siberian Continent and the 150 Million-Year History of the Mongol-Okhotsk Ocean. *Journal of Asian Earth Sciences,* 62, 79-97.

El Albani, A., Fursich, F. T., Colin, J. P., Meunier, A., Hochuli, P., Martin-Closas, C., Mazin, J. M. & Billon-Bruyat, J. P. 2004. Palaeoenvironmental Reconstruction of the Basal Cretaceous Vertebrate Bearing Beds in the Northern Part of the Aquitaine Basin (Sw France): Sedimentological and Geochemical Evidence. *Facies,* 50, 195-215.

Friis, E. M., Pedersen, K. R. & Crane, P. R. 1999. Early Angiosperm Diversification: The Diversity of Pollen Associated with Angiosperm Reproductive Structures in Early Cretaceous Floras from Portugal. *Annals of the Missouri Botanical Garden,* 86, 259-296.

Friis, E. M., Pedersen, K. R. & Crane, P. R. 2005. When Earth Started Blooming: Insights from the Fossil Record. *Current Opinion in Plant Biology,* 8, 5-12.

Friis.E., Crane.P & Pedersen.K 2011. *Early Flowers and Angiosperm Evolution*.

Gao, K. Q. & Ren, D. 2006. Radiometric Dating of Ignimbrite from Inner Mongolia Provides No Indication of a Post-Middle Jurassic Age for the Daohugou Beds. *Acta Geologica Sinica-English Edition,* 80, 42-45.

Godefroit.P. 2012. *Bernissart Dinosaurs and Early Cretaceous Terrestrial Ecosystems*.

Gordienko, I. V., Klimuk, V. S., Ivanov, V. G. & Posokhov, V. F. 1997. New Data on Composition and Age of Bimodal Volcanic Series of the Tugnui Riftogenic Depression, Trans-Baikalial Region. *Doklady Earth Sciences* 353, 273-276.

Gradstein.F., Ogg.J., Schmitz.M. & Ogg.G. 2012. *The Geologic Timescale 2012*.

Gradstein.F., Ogg.J. & Smith.A. 2004. *A Geologic Timescale 2004*.

Gratshev, V. G. & Aeranob, A. A. 2009. New Taxa of the Family Nemonychidae (Coleoptera) from Jurassic and Early Cretaceous. *Evraziatskii entomologicheskii Zhurnal,* 8, 411-416.

Haggart, J. W. 1984. Upper Cretaceous (Santonian-Campanian) Ammonite and Inoceramid Biostratigraphy of the Chico Formation, California. *Cretaceous Research,* 5, 225-241.

He.H.L., Wang.X.L., Zhoue.Z., Zhu.R.X., Jin.F., Wang.F., Ding.X. & Boven.A. 2004. 40ar/39ar Dating of Ignimbrite from Inner Mongolia, Northeastern China, Indicates a Post-Middle Jurassic Age for the Overlying Daohugou Bed. *Geophysical Research Letters,* 31.

Heimhofer.U., Hochuli.P., Burla.S. & Weissert.H. 2007. New Records of Early Cretaceous Angiosperm Pollen from Portugese Coastal Deposits: Implications for the Timing of the Early Angiosperm Radiation. *Review of Paleobotany and Palynology,* 144, 39-76.

Helby.R., Margan.R. & Partridge.A. 1987. A Palynological Zonation of the Australian Mesozoic. *Memoirs of the Association of Autralasian Paleontologists,* 4.

Henan., B. O. G. a. M. R. O. 1989. Regional Geology of Henan Province. People's Republic of China Ministry of Geology and Mineral Resources. *Geological Memoirs* 1, 1-772.

Herman, A. B. 2007. Comparative Paleofloristics of the Albian-Early Paleocene in the Anadyr-Koryak and North Alaska Subregions, Part 1: The Anadyr-Koryak Subregion. *Stratigraphy and Geological Correlation,* 15, 321-332.

Hochuli, P. A. 1981. North Gondwanan Floral Elements in Lower to Middle Cretaceous Sediments of the Southern Alps (Southern Switzerland, Northern Italy). *Review of Palaeobotany and Palynology,* 35, 337-358.

Hu, S., Jarzen, D. M. & Dilcher, D. L. 2008. New Species of Angiosperm Pollen from the Dakota Formation (Cenomanian, Upper Cretaceous) of Minnesota, USA. *Palynology,* 32, 17-26.

Hu.C., Cheng.Z. & Pang.W. 2001. *Shantungosaurus Giganteus*.

I.A., D. 1994. *Triassic Floras of Eurasia (Schriftenreihe Der Erdwissenschaftlichen Kommission)*, Springer.

Ignatov.M., Karasev.E. & Sinitsa.S. 2011. Upper Jurassic Mosses from Baigul (Transbaikalia, South Siberia). *Arctoa,* 20, 43-64.

Ivanov.V., Yramoluk.V. & Smirnov.V. 1995. New Data on the Age of Volcanism Evidence in West-Zabaikalian Late Mesozoic-Cenozoic Volcanic Domain. *Doklady Akademii Nauk,* 345, 648-652.

Jones.D. & Gryc.G. 1960. Upper Cretaceous Pelecypods of the Genus Inoceramus from Northern Alaska. *Shorter Contributions to General Geology*, 344-e.

Keller, A. M. & Hendrix, M. S. 1997. Paleoclimatologic Analysis of a Late Jurassic Petrified Forest, Southeastern Mongolia. *Palaios,* 12, 282-291.

Kimyai.A. 1966. New Plant Microfossils from the Raritan Formation (Cretaceous) in New Jersey. *Micropaelontology,* 12, 461-476.

Kopylov.D. 2010. Ichneumonids of the Subfamily Tancychorinae (Insecta: Hymenoptera: Ichneumonidae) from the Lower Cretaceous of Transbaikalia and Mongolia. *Paleontological Journal* 44, 180-187.

Kotov, A. A. 2007. Jurassic Cladocera (Crustacea, Branchiopoda) with a Description of an Extinct Mesozoic Order. *Journal of Natural History,* 41, 13-37.

Krassilov, V. 1982. Early Cretaceous Flora of Mongolia. *Palaeontographica Abteilung B Palaeophytologie,* 181, 1-43.

Krassilov, V. A. 1986. New Floral Structure from the Lower Cretaceous of Lake Baikal Area. *Review of Palaeobotany and Palynology,* 47, 9-16.

Krzeminski.W., A. J. A. 1995. Revision of Mesorhyphus Handlirsch, Eoplecia Handlirsch and Heterorhyphus Bode (Diptera: Anisopodomorpha, Bibionomorpha) from the Upper Liassic of Germany. *Palaeontologische Zeitschrift* 69, 167-172.

Labandeira, C. C. & Phillips, T. L. 1996. A Carboniferous Insect Gall: Insight into Early Ecologic History of the Holometabola. *Proceedings of the National Academy of Sciences of the United States of America,* 93, 8470-8474.

Langston, W. 1975. Ceratopsian Dinosaurs and Associated Lower-Vertebrates from St-Mary River Formation (Maestrichtian) at Scabby Butte, Southern Alberta. *Canadian Journal of Earth Sciences,* 12, 1576-1608.

Legalov, A. A. 2010a. Checklist of Mesozoic Curculionoidea (Coleoptera) with Description of New Taxa. *Baltic Journal of Coleopterology,* 10, 71-101.

Legalov, A. A. 2010b. Phylogeny of the Family Nemonychidae (Coleoptera) with Descriptions of New Taxa. *Evraziatskii entomologicheskii Zhurnal,* 9, 457-473.

Li, H. G. & Batten, D. J. 2004. Early Cretaceous Palynofloras from the Tanggula Mountains of the Northern Qinghai-Xizang (Tibet) Plateau, China. *Cretaceous Research,* 25, 531-542.

Ling, W., Xie, X., Liu, X. & Cheng, J. 2007. Zircon U-Pb Dating on the Mesozoic Volcanic Suite from the Qingshan Group Stratotype Section in Eastern Shandong Province and Its Tectonic Significance. *Science in China Series D-Earth Sciences,* 50, 813-824.

Liu, Y., Liu, Y., Ji, S. A. & Yang, Z. 2006. U-Pb Zircon Age for the Daohugou Biota at Ningcheng of Inner Mongolia and Comments on Related Issues. *Chinese Science Bulletin,* 51, 2634-2644.

Lo.C., Chen.P.J., Tsou.T.Y., Sun.S.S. & Lee.C.Y. 1999. 40ar-39ar Laser Single-Grain and K-Ar Dating of the Yixian Formation, Ne China. *Paleoworld,* 11, 329-340.

Markevitch, V. S. 1994. Palynological Zonation of the Continental Cretaceous and Lower Tertiary of Eastern Russia. *Cretaceous Research,* 15, 165-177.

Martin, M. D. & Merrill, G. K. 1976. Environmental Control of Conodont Distribution in the Bond and Mattoon Formations (Pennsylvanian, Missourian), Northern Illinois. *In:* R., B. C. (ed.) *Conodont Paleoecology.*

Martinez, C., Madrinan, S., Zavada, M. & Jaramillo, C. A. 2013. Tracing the Fossil Pollen Record of Hedyosmum (Chloranthaceae), an Old Lineage with Recent Neotropical Diversification. *Grana,* 52, 161-180.

Menier, J. J., Nel, A., Waller, A. & De Ploeg, G. 2004. A New Fossil Ichneumon Wasp from the Lowermost Eocene Amber of Paris Basin (France), with a Checklist of Fossil Ichneumonoidea S.L. (Insecta: Hymenoptera: Ichneumonidae: Metopiinae). *Geologica Acta,* 2, 83-94.

Metelkin.D., Gordienko.I. & Klimuk.V. 2007. Paleomagnetism of Upper Jurassic Basalts Fom Transbaikalia: New Data on the Time of Closure of the Mongol-Okhotsk Ocean and Mesozoic Intraplate Tectonics of Central Asia. *Russian Geology and Geophysics* 48, 825-834.

Miner.E. 1935. Paleobotanical Examinations of Cretaceous and Tertiary Coals. *The American Midland Naturalist Journal,* 16, 585-625.

Moiseeva, M. G. 2008. New Angiosperms from the Maastrichtian of the Amaam Lagoon Area (Northeastern Russia). *Paleontological Journal,* 42, 313-327.

Moiseeva, M. G. 2011. New Species of the Genus Macclintockia (Angiosperms) from the Campanian of the Ugol'naya Bay (Northeastern Russia). *Paleontological Journal,* 45, 207-223.

Morel, E. M., Artabe, A. E. & Spalletti, L. A. 2003. Triassic Floras of Argentina: Biostratigraphy, Floristic Events and Comparison with Other Areas of Gondwana and Laurasia. *Alcheringa,* 27, 231-243.

Newton.A. & Tangney.R. 2007. *Pleurocarpous Mosses: systematics and Evolution*, CRC Press.

Nosova, N. 2013. Revision of the Genus Grenana Samylina from the Middle Jurassic of Angren, Uzbekistan. *Review of Palaeobotany and Palynology,* 197, 226-252.

Oertli.H. 1963. *Faunes D'ostracodes Du Mésozoïque De France: Mesozoic Ostracod Faunas of France*.

Ogg.J., Hasenyager.R. & Wimbledon.W. 1994. Jurassic-Cretaceous Boundary:Portland-Purbeck Magnetostratigraphy and Possible Correlation to the Tethyan Faunal Realm. *Geobios,* 27, 519-527.

Palfy, J., Smith, P. L. & Mortensen, J. K. 2002. Dating the End-Triassic and Early Jurassic Mass Extinctions, Correlative Large Igneous Provinces, and Isotopic Events. *Catastrophic Events and Mass Extinctions: Impacts and Beyond,* 356, 523-532.

Paul, C. R. C., Mitchell, S. F., Marshall, J. D., Leary, P. N., Gale, A. S., Duane, A. M. & Ditchfield, P. W. 1994. Palaeoceanographic Events in the Middle Cenomanian of Northwest Europe. *Cretaceous Research,* 15, 707-738.

Peybernes.B. & Oertli.H. 1972. La Série De Passage Du Jurassique Au Crétacé Dans Le Bassin Sud-Pyrénén (Espagne). *Compte Rendus de l'Académie des Sciences, Paris* 274, 3348–3351.

Philippe, M. & Thevenard, F. 1996. Distribution and Palaeoecology of the Mesozoic Wood Genus Xenoxylon: Palaeoclimatological Implications for the Jurassic of Western Europe. *Review of Palaeobotany and Palynology,* 91, 353-370.

Ponomarenko.A. 1993. Two New Species of Mesozoic Dysticoid Beetles from Asia. *Paleontological Journal,* 27, 182-191.

Pozo.J., R. D. 1971. Algunas Observaciones Sobre El Jurasico De Alava, Burgos Y Santander. *Cuadernos Geología Ibérica* 2, 491-508.

Rasnitsyn, A. P., Jarzembowski, E. A. & Ross, A. J. 1998. Wasps (Insecta: Vespida =Hymenoptera) from the Purbeck and Wealden (Lower Cretaceous) of Southern England and Their Biostratigraphical and Paleoenvironmental Significance *Cretaceous Research,* 19, 329-391.

Rasnitsyn, A. P. & Zhang, H. C. 2004. Composition and Age of the Daohugou Hymenopteran (Insecta, Hymenoptera = Vespida) Assemblage from Inner Mongolia, China. *Palaeontology,* 47, 1507-1517.

Rasnitsyn.A. 1988. An Outline of Evolution of the Hymenopterous Insects (Order Vespida). *Oriental Insects,* 22, 115-145.

Rasnitsyn.A. & Quicke.D. 2002. *History of Insects*, Kluwer Academic Publishers.

Ren.D., Gao.K., Ziguang.G., Shuan.J., Jingjing.T. & Zhuo.S. 2002. Stratigraphic Division of the Jurassic in the Daohugou Area, Nincheng, Inner Mongolia. *Geological Bulletin of China* 21, 584-591.

Rice, C. M., Ashcroft, W. A., Batten, D. J., Boyce, A. J., Caulfield, J. B. D., Fallick, A. E., Hole, M. J., Jones, E., Pearson, M. J., Rogers, G., Saxton, J. M., Stuart, F. M., Trewin,

N. H. & Turner, G. 1995. A Devonian Auriferous Hot-Spring, System, Rhynie, Scotland. *Journal of the Geological Society,* 152, 229-250.

Richardson.J. 1967. Some British Lower Devonian Spore Assemblages and Their Stratigraphic Significance. *Review of Paleobotany and Palynology,* 1, 111-129.

Richardson.J. & Mcgregor.D. 1986. Silurian and Devonian Spore Zones of the Old Red Sandstone Continent and Adjacent Regions. *Bulletin of The Geological Survey of Canada,* 364, 1-79.

Robazynski.F. & Caron.M 1988. Foraminifčres Planctoniques Du Crétacé Commentaire De La Zonation Europe – Méditerranée. *Bulletin de la Société géologique de France* 166, 681-692.

Rogers, R. R., Swisher, C. C., Sereno, P. C., Monetta, A. M., Forster, C. A. & Martinez, R. N. 1993. The Ischigualasto Tetrapod Assemblage (Late Triassic, Argentina) and Ar-40/Ar-39 Dating of Dinosaur Origins. *Science,* 260, 794-797.

Rohstoffe, B. F. G. U. *Brochterbeck-Formation* [Online]. Available: http://www.bgr.de/app/litholex/bilder/2008013.pdf. [Accessed December 11th 2013].

Ronquist, F., Klopfstein, S., Vilhelmsen, L., Schulmeister, S., Murray, D. L. & Rasnitsyn, A. P. 2012a. A Total-Evidence Approach to Dating with Fossils, Applied to the Early Radiation of the Hymenoptera. *Systematic Biology,* 61, 973-999.

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A. & Huelsenbeck, J. P. 2012b. Mrbayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice across a Large Model Space. *Syst Biol,* 61, 539-42.

Rothwell, G. W., Mapes, G., Stockey, R. A. & Hilton, J. 2012. The Seed Cone Eathiestrobus Gen. Nov.: Fossil Evidence for a Jurassic Origin of Pinaceae. *American Journal of Botany,* 99, 708-720.

Sahagian.D., Beisel.A. & Zakharov.V. 1994. Sequence Stratigraphy Enhancement of Biostratigraphic Correlation with Application to the Upper Cretaceous of Northern Siberia: A Potential Tool for Petroleum Exploration. *International Geology Reviews* 36, 359-372.

Sakulina.G.V. 1971. Classopollis Pollen in the Upper Jurassic Deposits of Southern Kazakhstan. *Palynology of Kazakhstan. Problems of the Geology of the Weathering Crust.*

Schemel, M. P. 1950. Cretaceous Plant Microfossils from Iowa. *American Journal of Botany,* 37, 750-754.

Schmitz, M. D. & Davydov, V. I. 2012. Quantitative Radiometric and Biostratigraphic Calibration of the Pennsylvanian Early Permian (Cisuralian) Time Scale and Pan-

Euramerican Chronostratigraphic Correlation. *Geological Society of America Bulletin,* 124, 549-577.

Shadaev.M., Poskhov.V. & Drubetskoi.E. 1992. New Data on Age of the Ichetui Suite in Western Transbaikalia: Rb-Sr and K-Ar Data. *Geologica I Geofizika* 33, 41-44.

Shang, H., Cui, J.-Z. & Li, C.-S. 2001. Pityostrobus Yixianensis Sp. Nov., a Pinaceous Cone from the Lower Cretaceous of North-East China. *Botanical Journal of the Linnean Society,* 136, 427-437.

Shcherbakov.D. 2008. Madygen, Triassic Lagerstätte Number One, before and after Sharov. *Alavesia,* 2, 113-124.

Shen.Y., Chen.P. & Huang.D. 2003. Age of the Fossil Conchostracans from Daouhugou of Nincheng, Inner Mongolia. *Journal of Stratigraphy* 27, 311-313.

Sinitshenkova.N 2005. The Oldest Known Record of an Imago of Nemouridae (Insecta: Perlida = Plecoptera) in the Late Mesozoic of Eastern Transbaikalia *Paleontological Journal* 39, 39-41.

Sinitshenkova.N. 1985. Jurskie Podenki (Ephemerida-Ephemeroptera) Yuzhnoy Sibiri I Zapadnoy Mongolii. *In:* RASNITSYN.A (ed.) *Jurskie Nasekomye Sibiri I Mongolii*

Skoblo.V. & Lyamina.N. 1965. On the Tugnai Type Locality. *Materials on Geology and Mineral Fossils of the Buryat Autonomous Republic.*

Spicer, R. A., Ahlberg, A., Hermana, A. B., Kelley, S. P., Raikevich, M. I. & Rees, P. M. 2002. Palaeoenvironment and Ecology of the Middle Cretaceous Grebenka Flora of Northeastern Asia. *Palaeogeography Palaeoclimatology Palaeoecology,* 184, 65-105.

Stover.L.E., Briknhius.H., Damassa.S., Verteuill.L., D., Helby.R.J., Monteil.E., Partdridge.A. & Powell.A. 1996. Mesozoic–Tertiary Dinoflagellates, Acritarchs and Prasinophytes. *In:* JANSONIUS.J. & MCGREGOR.D (eds.) *Palynology: Principles and Applications, Volume Ii.*

Streel.M., Higgs.K., Loboziak.S., Riegel.W. & Steemans.P. 1987. Spore Stratigraphy and Correlation with Faunas and Floras in the Type Marine Devonian of the Ardenne-Rhenish Regions. *Review of Paleobotany and Palynology,* 50, 211-229.

Sukacheva, I. D. & Rasnitsyn, A. P. 2004. Jurassic Insects (Insecta) from the Sai-Sagul Locality (Kyrgyzstan, Southern Fergana). *Paleontological Journal,* 38, 182-186.

Swisher, C. C., Wang, X. L., Zhou, Z. H., Wang, Y. Q., Jin, F., Zhang, J. Y., Xu, X., Zhang, F. C. & Wang, Y. 2002. Further Support for a Cretaceous Age for the Feathered-Dinosaur Beds of Liaoning,China: New Ar-40/Ar-39 Dating of the Yixian and Tuchengzi Formations. *Chinese Science Bulletin,* 47, 135-138.

Szwedo, J. 2011. The Coleorrhyncha (Insecta: Hemiptera) of the European Jurassic, with a Description of a New Genus from the Toarcian of Luxembourg. *Volumina Jurassica,* 9, 3-19.

Vakhrameev.V. 1970. Range and Paleoecology of Mesozoic Conifers, the Cheirolepidiaceae. *Paleontological Journal* 4, 12-25.

Vakhrameev.V. 1991. *Jurassic and Cretaceous Floras and Climates of the Earth*

Vakhrameev.V. & Kotova.I. 1977. Ancient Angiosperms and Accompanying Plants from the Lower Cretaceous of Transbaikalia. *Paleontological Journal* 4, 487-495.

Valencio.D., Mendia.J. & Vilas.J. 1975. Paleomagnetism and K/Ar Ages of Triassic Igneous Rocks from the Ischigualasto-Ischichuca Basin and Puesto Viejo Formation, Argentina. *Earth and Planetary Science Letters,* 26, 319-330.

Vrsansky, P. 2003. Unique Assemblage of Dictyoptera (Insecta - Blattaria, Mantodea, Isoptera) from the Lower Cretaceous of Bon Tsagaan Nuur in Mongolia. *Entomological Problems,* 33, 119-151.

Vrsansky, P. & Ansorge, J. 2007. Lower Jurassic Cockroaches (Insecta: Blattaria) from Germany and England. *African Invertebrates,* 48, 103-126.

Wang, C. & Ren, D. 2013. Nuurcala Obesa Sp N. (Blattida, Caloblattinidae) from the Lower Cretaceous Yixian Formation in Liaoning Province, China. *Zookeys*, 35-46.

Wang, S. S., Hu, H. G., Li, P. X. & Wang, Y. Q. 2001. Further Discussion on the Geologic Age of Sihetun Vertebrate Assemblage in Western Liaoning, China: Evidence from Ar-Ar Dating. *Acta Petrologica Sinica,* 17, 663-668.

Wang.S., Hu.H., Li.P. & Wang.Y. 2001. Further Discussion on Geologic Age of Sihetun Vertebrate Assemblage in Western Liaoning China: Evidence from Ar-Ar Dating. *Petrelog. Sinica* 17, 663-668.

Wang.X., Zhoue.Z., He.H., Jin.F., Wang.Y., Zhang.J., Wang.Y., Xu.X. & Zhang.F. 2005. Stratigraphy and Age of the Dahougou Bed in Nincheng, Inner Mongolia. *Chinese Science Bulletin* 50, 2369-2376.

Warnock, R. C., Yang, Z. & Donoghue, P. C. 2012. Exploring Uncertainty in the Calibration of the Molecular Clock. *Biol Lett,* 8, 156-9.

Wellman, C. H., Kerp, H. & Hass, H. 2006. Spores of the Rhynie Chert Plant Aglaophyton (Rhynia) Major (Kidston and Lang) D.S Edwards, 1986. *Review of Palaeobotany and Palynology,* 142, 229-250.

Yang, Q., Makarkin, V. N., Winterton, S. L., Khramov, A. V. & Ren, D. 2012. A Remarkable New Family of Jurassic Insects (Neuroptera) with Primitive Wing Venation and Its Phylogenetic Position in Neuropterida. *Plos One,* 7.

Zakharov.V., Lebedeva.N. & Khometovsky.O. 2002. Upper Cretaceous Inoceramid and Dinoflaggelate Cyst Biostratigraphy of the Northern Siberia. *In:* MICHALÍK.J. (ed.) *Tethyan/Boreal Cretaceous Correlation: Mediterranean and Boreal Cretaceous Paleobiogeographic Areas in Central and Eastern Europe.*

Zhang, J. 2011. Three Distinct but Rare Kovalevisargid Flies from the Jurassic Daohugou Biota, China (Insecta, Diptera, Brachycera, Kovalevisargidae). *Palaeontology,* 54, 163-170.

Zhang, J. & Rasnitsyn, A. P. 2006. New Extinct Taxa of Pelecinidae Sensu Lato (Hymenoptera : Proctotrupoidea) in the Laiyang Formation, Shandong, China. *Cretaceous Research,* 27, 684-688.

Zhang, J. F. 2005a. Eight New Species of the Genus Eopelecinus (Hymenoptera : Proctotrupoidea : Pelecinidae) from the Laiyang Formation, Shandong Province, China. *Paleontological Journal,* 39, 417-427.

Zhang, J. F. 2005b. The First Find of Chrysomelids (Insecta : Coleopetra : Chrysomeloidea) from Callovian-Oxfordian Daohugou Biota of China. *Geobios,* 38, 865-871.

Zhang, J. F. 2010. Records of Bizarre Jurassic Brachycerans in the Daohugou Biota, China (Diptera, Brachycera, Archisargidae and Rhagionemestriidae). *Palaeontology,* 53, 307-317.

Zhang, J. F. & Rasnitsyn, A. P. 2004. Minute Members of Baissinae (Insecta : Hymenoptera : Gasteruptiidae) from the Upper Mesozoic of China and Limits of the Genus Manlaya Rasnitsyn, 1980. *Cretaceous Research,* 25, 797-805.

Zherekhin.V. 1978. Development and Changes of the Cretaceous and Cenozoic Faunal Assemblages (Tracheata and Chelicerata). *Transactions of the Paleontological Institute, Academy of Sciences USSR* 165, 1–198.

Zherikhin.V., Mostovski.M., Vrsansky.P., Blagoderov.V. & Lukashevich.E. 1998. The Unique Lower Cretaceous Locality Baissa and Other Contemporaneous Fossil Insect Sites in North and West Transbaikalia. *Proceedings of the First International Palaeoentomological Conference, Moscow*

Zhou, Z., Barrett, P. M. & Hilton, J. 2003. An Exceptionally Preserved Lower Cretaceous Ecosystem. *Nature,* 421, 807-14.

Zhou, Z.-Y. 2010. An Overview of Fossil Ginkgoales (Vol 18, Pg 1, 2009). *Palaeoworld,* 19, 206-206.

Zhoue.Z. 2006. Evolutionary Radiation of the Jehol Biota: Chronological and Ecological Perspectives. *Geological Journal* 41, 377-393.

Chapter Four Supplementary Material

Supplementary Methods - O'Reilly and Donoghue (2016): Tips and nodes are complimentary not competing approaches to the calibration of molecular clocks. *Biology Letters*

Conditional Prior Justification

For a matrix consisting of *s* taxa we can obtain the total number of potential bifurcating, rooted, topologies using

$$\tau_{\text{total}} = \frac{(2s-3)!}{2^{(s-2)}(s-2)!}$$ (Felsenstein, 1978),

and for any purely bifurcating tree there is a vector *t* of times on the *s*-1 nodes

$$t = [t_1,\ldots,t_{s\text{-}1}]$$

To obtain the effective joint time prior on all possible topologies we must marginalise over topology. For rooted, strictly bifurcating trees this is

$$\mathrm{P}(t) = \sum_{\mathrm{i}=1}^{\tau_{\text{total}}} \mathrm{P}(t|\,\tau_{\mathrm{i}})$$

For a tree of 113 taxa, as we have here, this would involve the unfeasible assessment of the time prior for $\frac{223!}{2^{111}\times 111!}$ bifurcating trees, with an even greater number if allowing for multifurcations. We consider the time prior conditioned on the majority rule consensus tree topology $\mathrm{P}(t|\tau_{\text{cons}})$ a suitable approximation of the time prior as this topology is composed of the most commonly sampled clades, and therefore reflects the distribution of the most commonly sampled time prior.

Divergence Time Estimation - Settings Shared Across All Analyses

Three approaches to divergence time estimation were taken: Node calibration; tip calibration; combined tip and node calibration. For each of these approaches to calibration we obtained a

sample of trees from the posterior and a sample of trees from the prior distribution, or in the case of tip-calibration, the "conditional" prior distribution. For all analyses, whether tip, node or combined tip and node calibrated, we utilised the same clock model (IGR), partitioning scheme, partition specific substitution models and prior distributions on parameters associated with these as were used in (Ronquist et al., 2012a). The uniform tree prior was also employed for all analyses, including node-calibrated analyses, as in the original article. The only priors changed from (Ronquist et al., 2012a) were the node and tip calibrations, which were swapped for those in (O'Reilly et al., 2015), with offset exponential distributions assigned to the node calibrations instead of uniform. Offset exponential distributions possess no hard maximum, allowing interaction with other priors to induce maxima on the node calibrations. For all analyses we performed 4 separate runs in MrBayes to assess convergence (Ronquist et al., 2012b). These 4 separate runs were combined into a single set of sampled trees; convergence was considered to have been achieved with ESS scores >200, split frequency values of <0.05, and through qualitative assessment of the marginal distributions of parameters.

Node Calibration

A posterior sample of trees was obtained under node calibration by running MrBayes for 20,000,000 generations, sampling every 2000th. A sample of trees from the prior distribution was obtained by sampling for the same number of generations at the same frequency, but with the mcmc usedata = NO setting in MrBayes.

Tip Calibration

Tip calibrated analyses ran for 50,000,000 generations, sampling every 5000th when sampling from the posterior, and 20,000,000 sampling every 200th when sampling from the prior conditioned on the consensus topology.

Tip + Node Calibration

Taxa, both fossil and extant, were assigned to clades based on their placement in the consensus tree produced from the posterior sample of tip-calibrated trees. Topological constraints were then designed to enforce monophyly of these clades, facilitating the application of node calibrations to each clade that was calibrated in the exclusively node-calibrated analyses. As tip-calibration is a relatively new approach to divergence time estimation many aspects of standard phylogenetic algorithms are yet to properly accommodate tip calibration. One such area is the proposal distribution for clade age when there is an overlap between a tip-calibration and the node calibration for the clade to which it is assigned. MrBayes appears to allow for the proposal of tip ages that are older than the clade to which they are assigned, leading to an error when assessing if the new, impossible, state should be accepted in the chain. To deal with this we

have discarded any fossil that could conflict with its subtending node calibration, in the case of Tenthredinoidea and Apocrita we have not calibrated the clade, as this would require the disposal of a large number of fossil taxa. By enforcing monophyly of clades, but not any more fine scale topological constraints, we allow for the placement of all fossil taxa assigned to a given clade as stem group members, this allows for violation of the minimum bound for our node calibrations. This phenomenon is prominent in Pamphilioidea where a small number of trees in the sample possess all fossils in this clade as stem group members, effectively removing the minimum age constraint for the crown group. This issue highlights the need for non-exclusive constraints that also allow for calibration in MrBayes.

Xyelidae was not assigned a node calibration due to issues highlighted by the original authors with the placement of the fossil used to define the node calibration as a stem and not crown group member. Similarly, in our tip-calibrated analyses *Eoxyela* was often placed outside of crown Xyelidae and further demonstrating its unsuitability as a node calibration for crown Xyelidae.

Tip + Node calibrated analyses ran for 50,000,000 generations, sampling every 5000th when sampling from the posterior, and 20,000,000 sampling every 2000th when sampling from the prior conditioned on the consensus topology.

Node Calibration Prior Retrieval

To obtain the marginal distributions on the ages of the calibrated nodes we used an R script that calculated the age of each clade in each of the trees sampled from the prior (4). Clades were identified on the basis of their extant members, diagnosed from the topology retrieved from the consensus tree produced from the sample of trees from the node calibrated posterior distribution.

Marginal distributions for the ages of the nine calibrated clades employed by the original authors were retrieved using R. For any given node, the marginal distribution of that node age can be approximated from the posterior sample of trees, as the age of each clade is a parameter in the phylogenetic model. Clades were identified on the basis of their extant members, determined from the consensus tree produced from the sample of trees from the node calibrated posterior distribution, in each sampled tree the node that defines the most recent common ancestor of this set of taxa is identified and its age is stored. The distribution of these ages is the marginal distribution of the age of the clade.

Tip Calibration and Tip+Node Calibration Conditional Prior Retrieval

To obtain the conditional prior for tip-calibrated analyses we fix the topology to that of the consensus tree constructed from the posterior sample of tip-calibrated trees. An R script was written that defines all internal nodes within a given tree (including multifurcations) as topological constraints for MrBayes. The prior is then sampled from conditional on this fixed topology; the sample of trees from the prior are obtained by setting usedata = NO in MrBayes.

Marginal distributions for the ages of the nine calibrated clades employed by the original authors were retrieved using R. For any given node, the marginal distribution of that node age can be approximated from the posterior sample of trees, as the age of each clade is a parameter in the phylogenetic model. Clades were identified on the basis of their extant members, determined from the topology retrieved from a previous node-calibrated analysis of the data (Ronquist et al., 2012a) and then in each sampled tree the node that defines the most recent common ancestor of this set of taxa is identified and its age is stored. The distribution of these ages is the marginal distribution of the age of the clade.

Felsenstein, J. 1978. Number of Evolutionary Trees. *Systematic Zoology,* 27**,** 27-33.

O'reilly, J. E., Dos Reis, M. & Donoghue, P. C. J. 2015. Dating Tips for Divergence-Time Estimation. *Trends in Genetics,* 31**,** 637-650.

Ronquist, F., Klopfstein, S., Vilhelmsen, L., Schulmeister, S., Murray, D. L. & Rasnitsyn, A. P. 2012a. A Total-Evidence Approach to Dating with Fossils, Applied to the Early Radiation of the Hymenoptera. *Systematic Biology,* 61**,** 973-999.

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A. & Huelsenbeck, J. P. 2012b. Mrbayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice across a Large Model Space. *Syst Biol,* 61**,** 539-42.

Chapter Five Supplementary Material

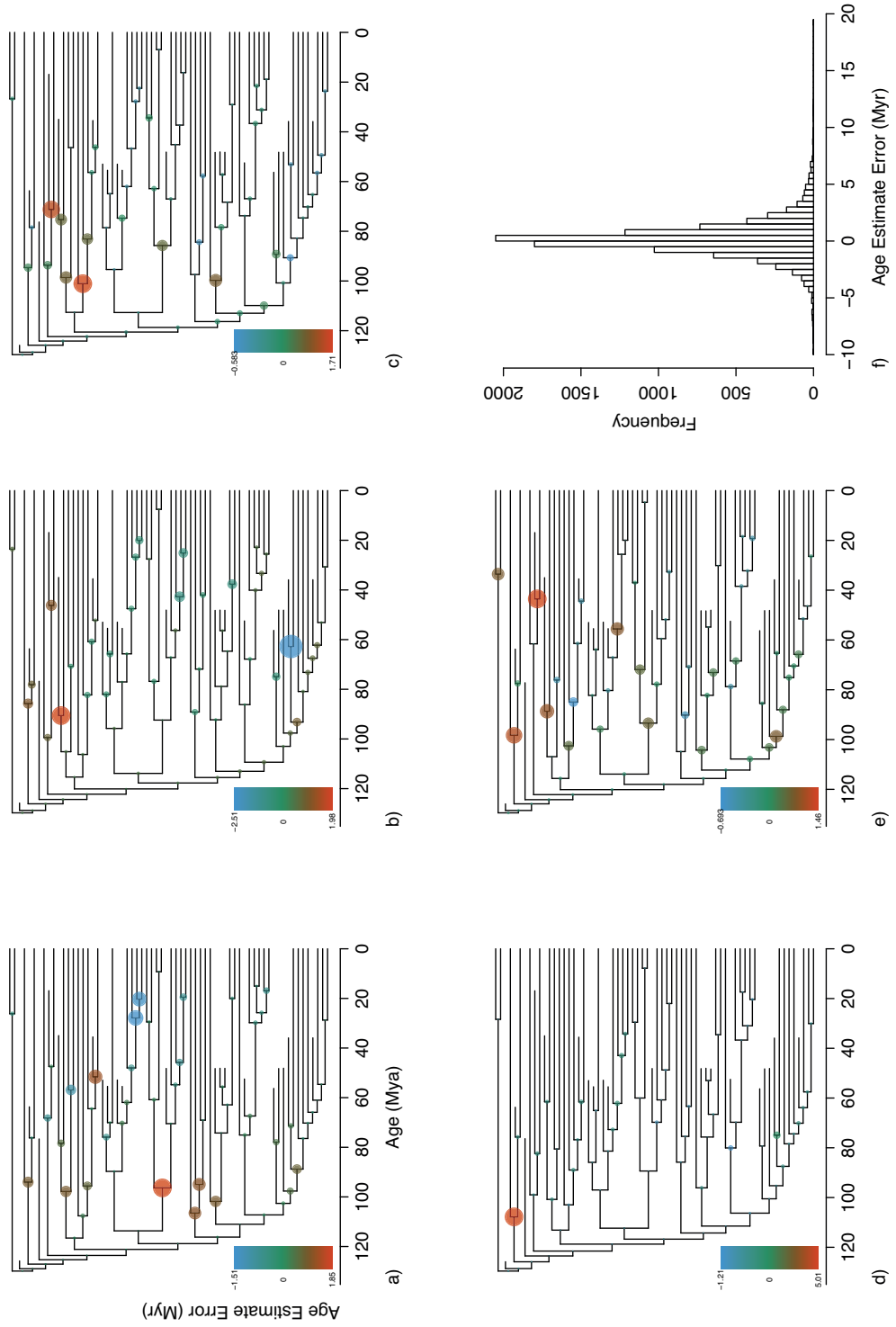


Figure S5.1 - (a-e) Mammalian time scaled phylogenies with node colour indicating the difference between mean estimated ages obtained from each respective reduced 245 character untreated matrix and average ages obtained from 30 replicate treated matrices when the distribution of missing data approximates that expected when soft characters are lost due to degradation and decay. (f) Histogram of node age estimate error across all 30 replicate analyses for all 5 independent matrix reductions.

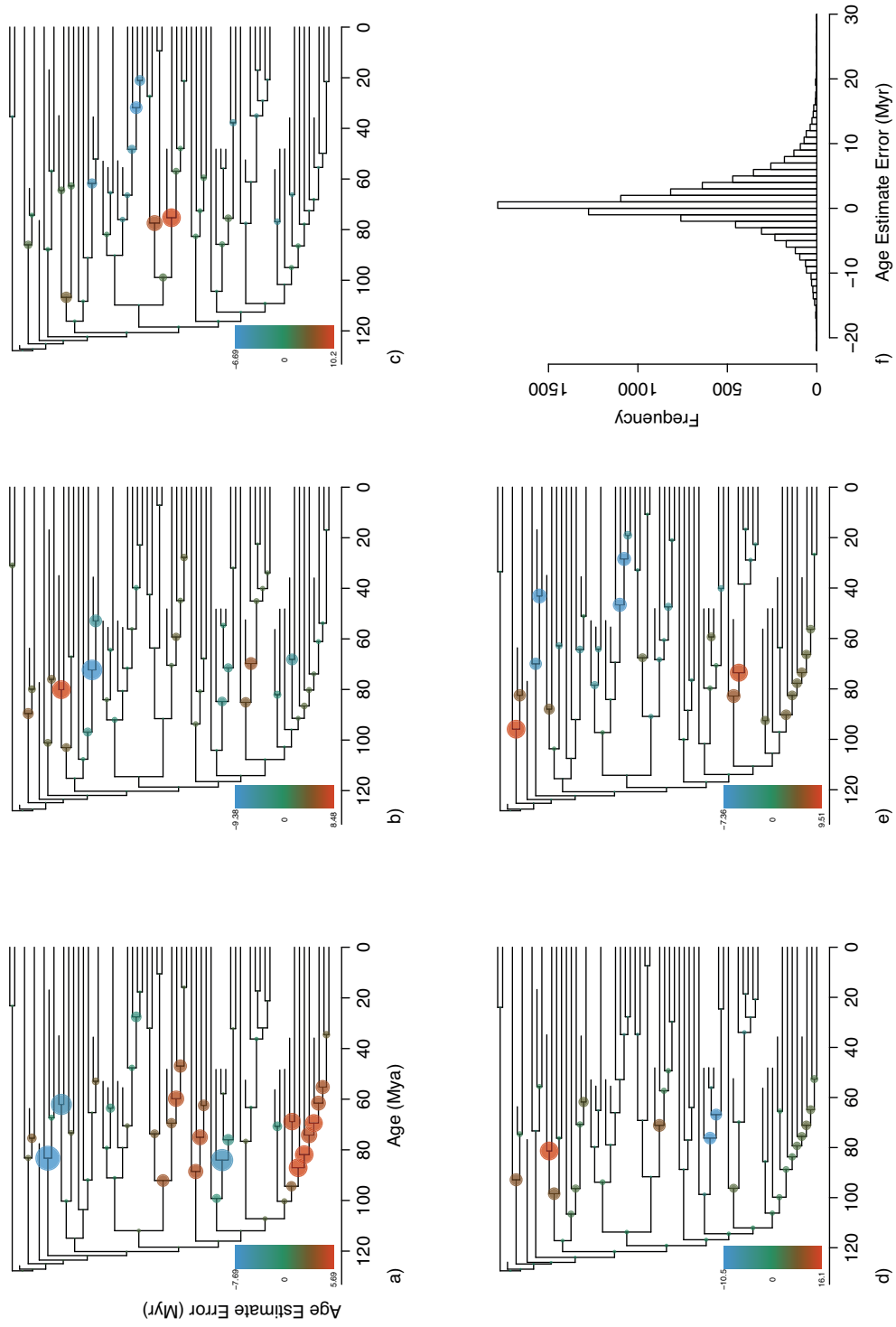


Figure S5.2 - (a-e) Mammalian time scaled phylogenies with node colour indicating the difference between mean estimated ages obtained from each respective reduced 245 character untreated matrix and average ages obtained from 30 replicate treated matrices when the distribution of missing data approximates that expected when characters are lost in blocks due to physical biostratinomic processes (scale factor =1). (f) Histogram of node age estimate error across all 30 replicate analyses for all 5 independent matrix reductions.

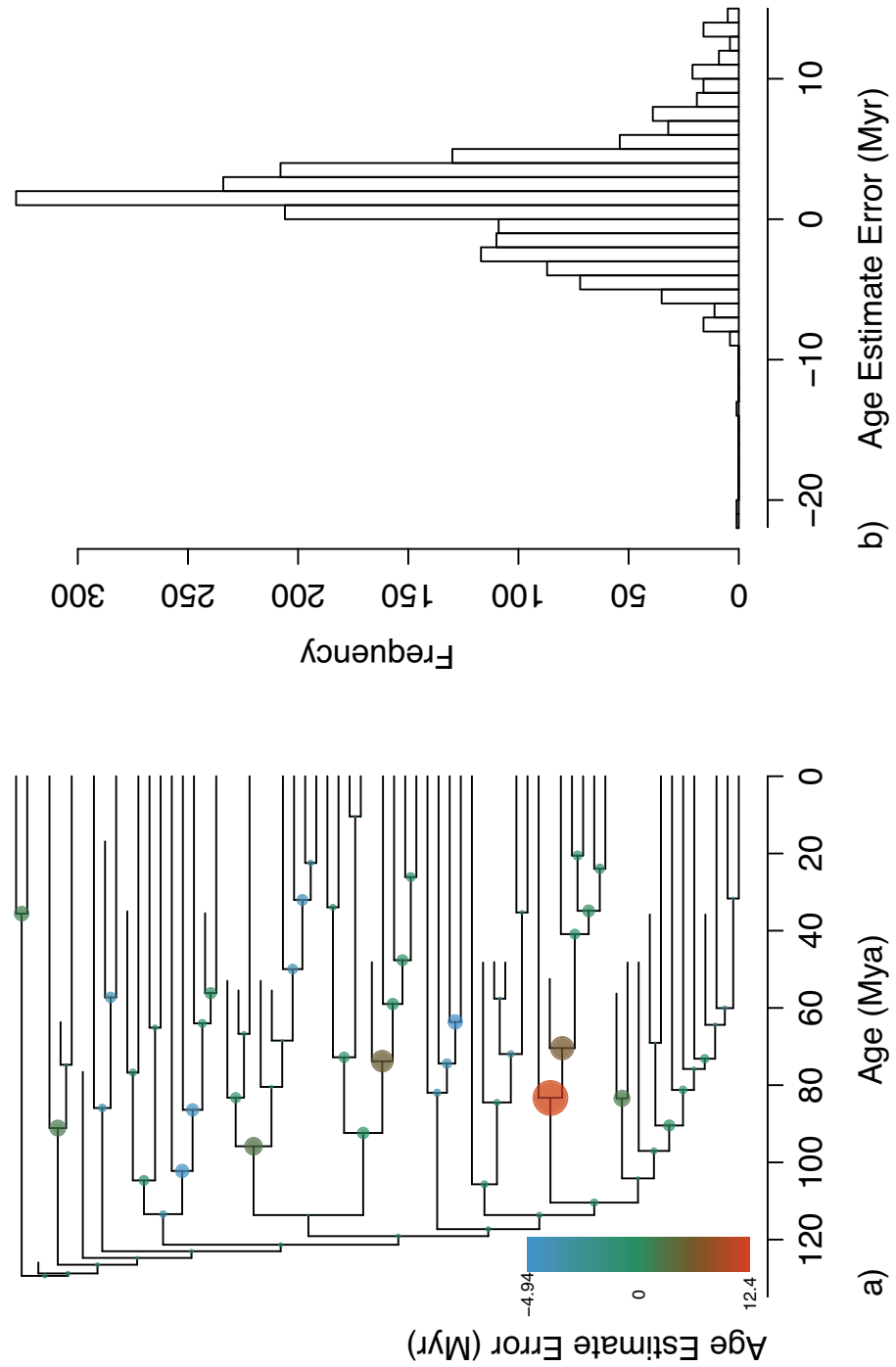


Figure S5.3 - (a) Mammalian time scaled phylogeny with node colour indicating the difference between mean estimated ages obtained from the full 2454 character untreated matrix and average ages obtained from 30 replicate treated matrices when the distribution of missing data approximates the distribution expected when characters are lost in blocks due to physical biostratinomic processes (scale factor = 1.4). (b) Histogram of node age estimate error across all 30 replicate analyses.

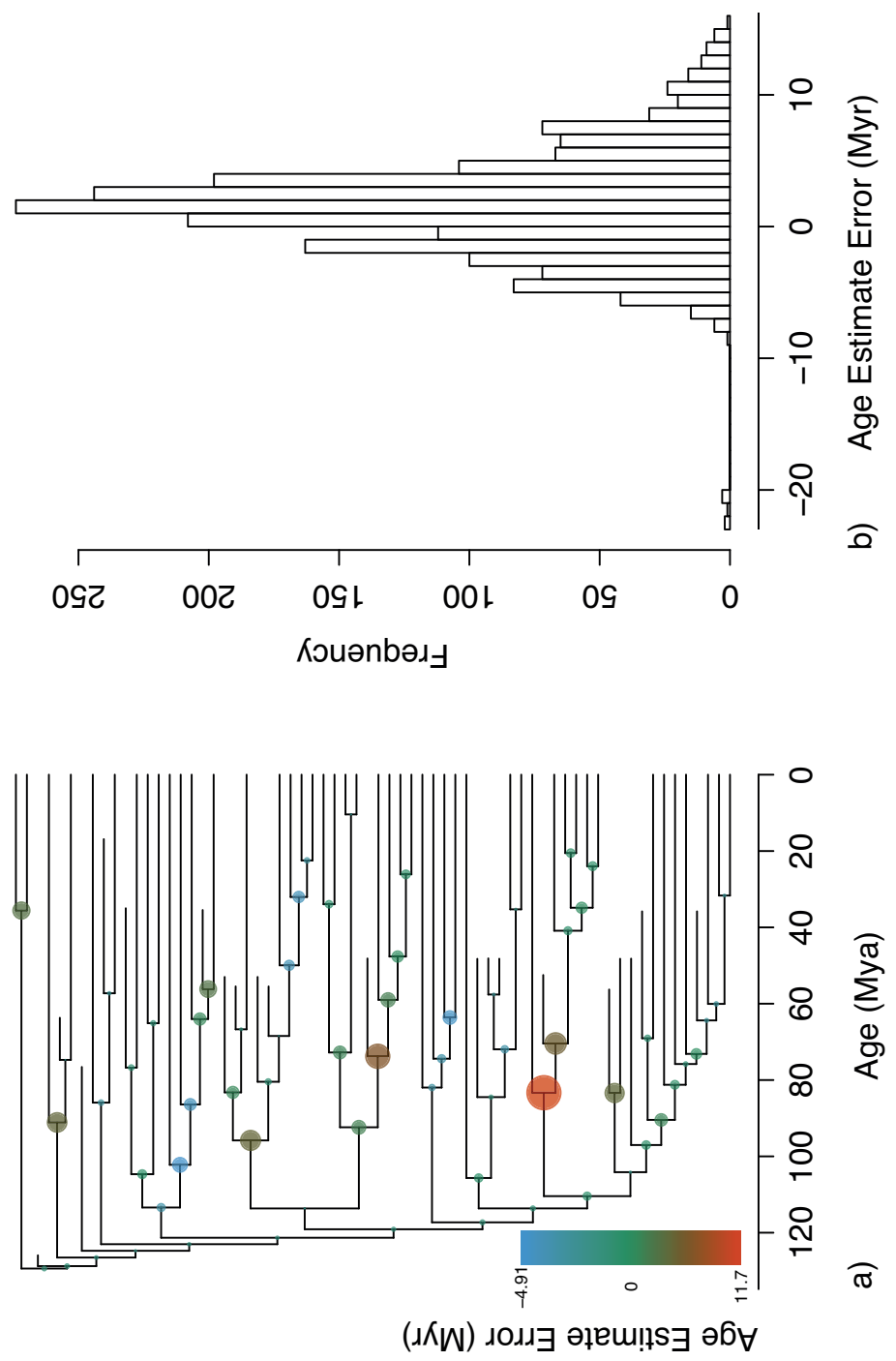


Figure S5.4 - (a) Mammalian time scaled phylogeny with node colour indicating the difference between mean estimated ages obtained from the full 2454 character untreated matrix and average ages obtained from 30 replicate treated matrices when the distribution of missing data approximates the distribution expected when characters are lost in blocks due to physical biostratinomic processes (scale factor = 1.8). (b) Histogram of node age estimate error across all 30 replicate analyses.

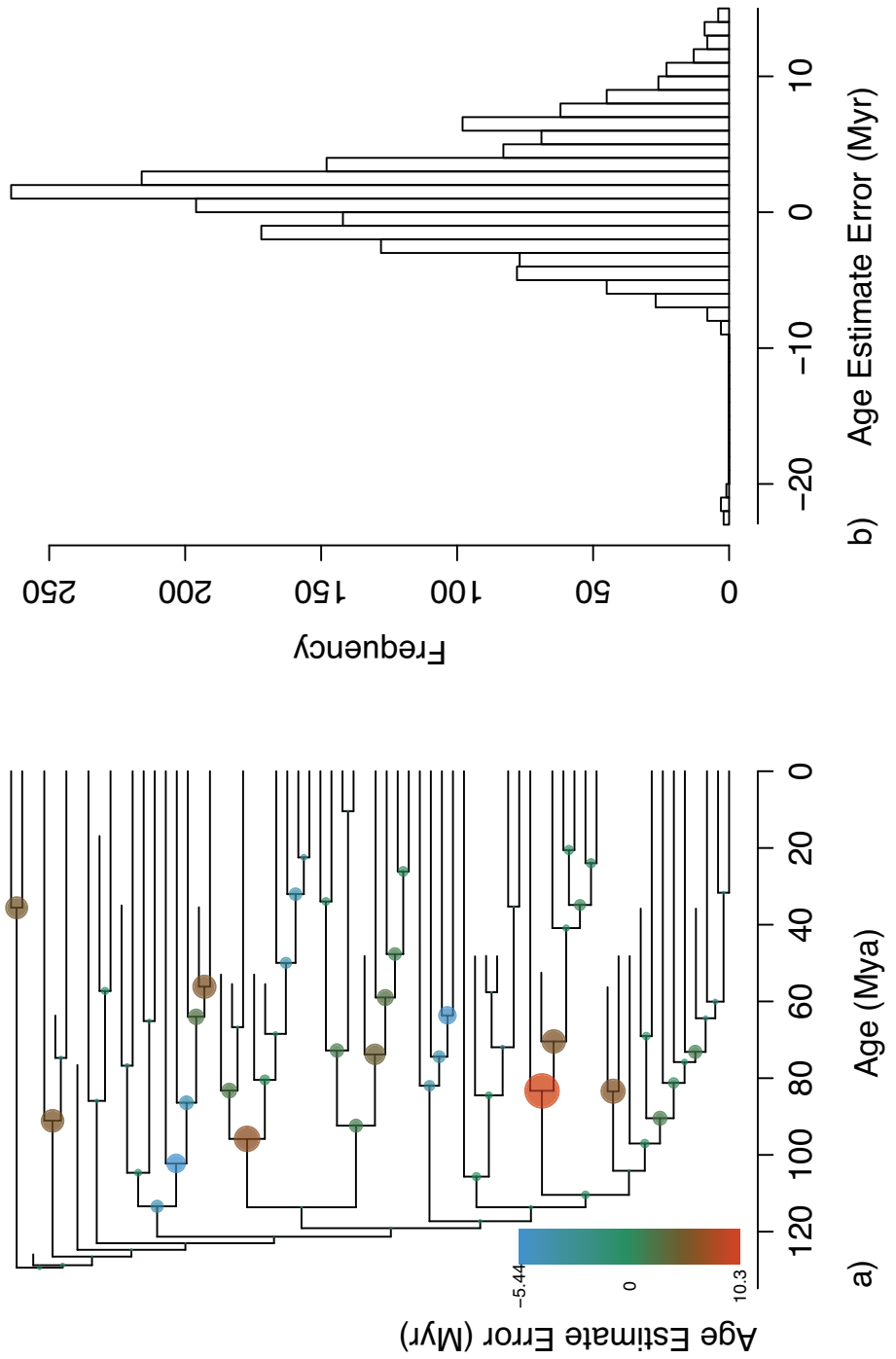


Figure S5.5 - (a) Mammalian time scaled phylogeny with node colour indicating the difference between mean estimated ages obtained from the full 2454 character untreated matrix and average ages obtained from 30 replicate treated matrices when the distribution of missing data approximates the distribution expected when characters are lost in blocks due to physical biostratinomic processes (scale factor = 2). (b) Histogram of node age estimate error across all 30 replicate analyses.

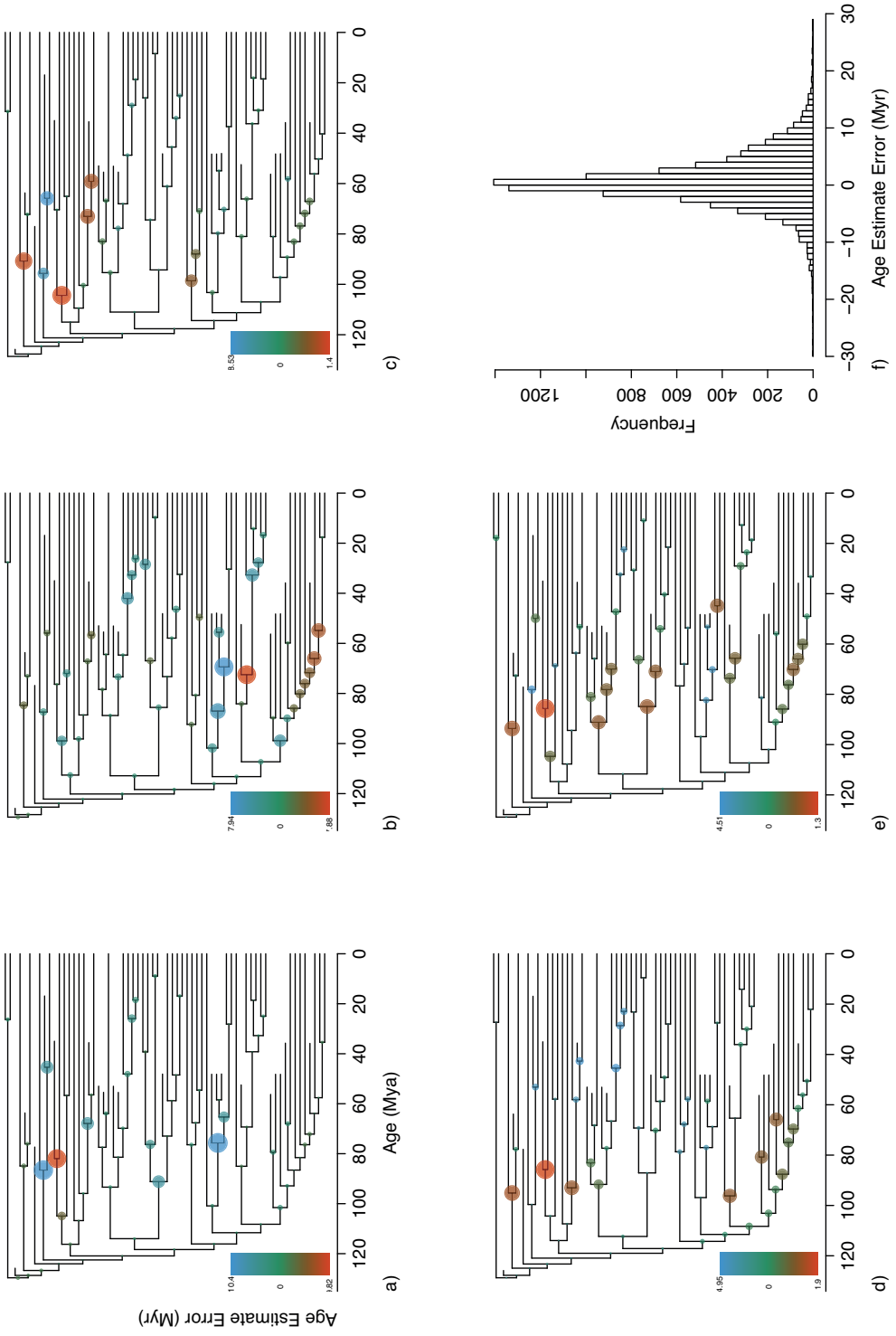


Figure S5.6 – (a-e) Mammalian time scaled phylogenies with node colour indicating the difference between mean estimated ages obtained from each respective reduced 245 character untreated matrix and average ages obtained from 30 replicate treated matrices when the distribution of missing data approximates that expected when soft characters are lost due to physical biostratinomic processes (scale factor = 1.4). (f) Histogram of node age estimate error across all 30 replicate analyses for all 5 independent matrix reductions.

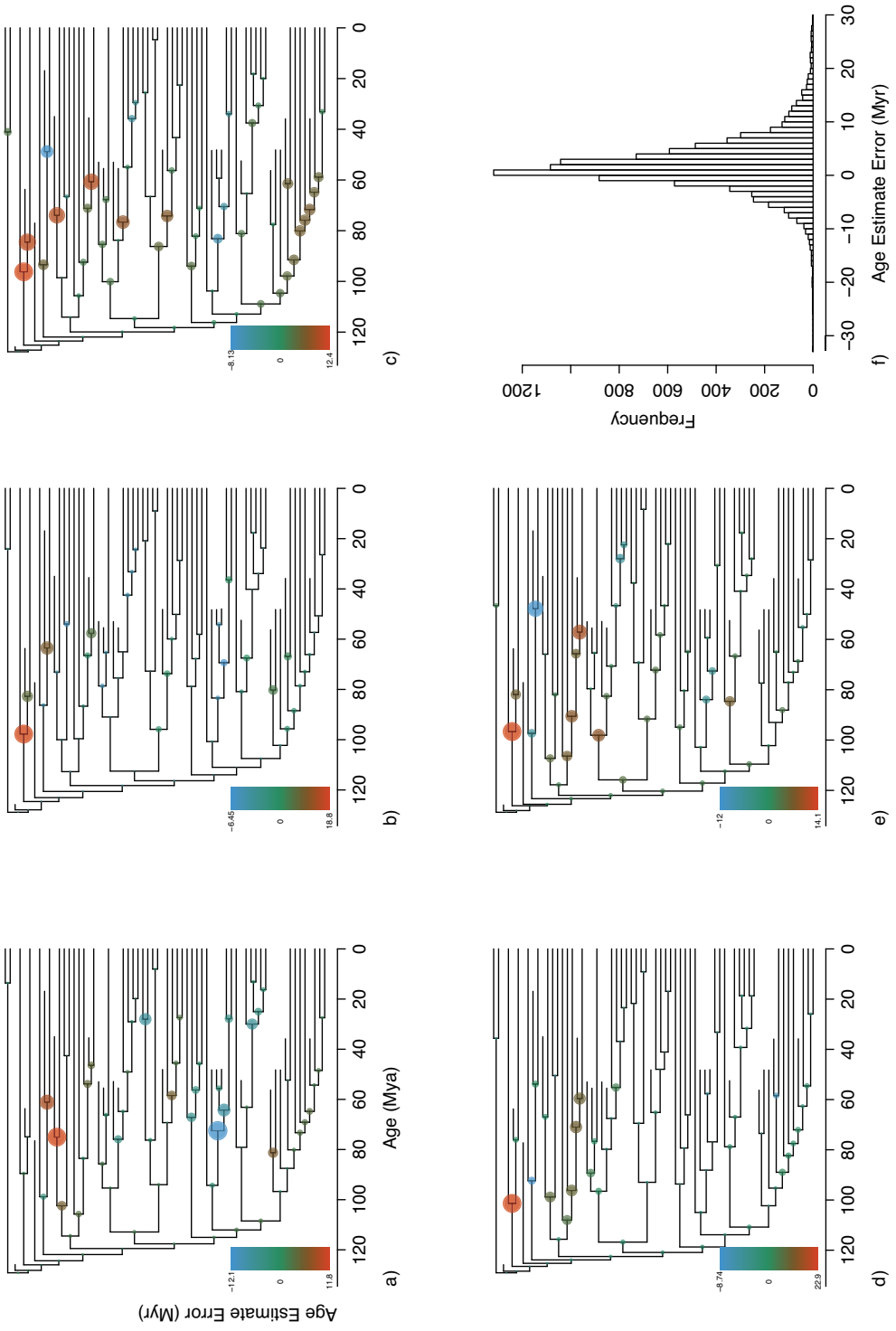


Figure S5.7 - (a-e) Mammalian time scaled phylogenies with node colour indicating the difference between mean estimated ages obtained from each respective reduced 245 character untreated matrix and average ages obtained from 30 replicate treated matrices when the distribution of missing data approximates that expected when soft characters are lost due to physical biostratinomic processes (scale factor = 1.8). (f) Histogram of node age estimate error across all 30 replicate analyses for all 5 independent matrix reductions.

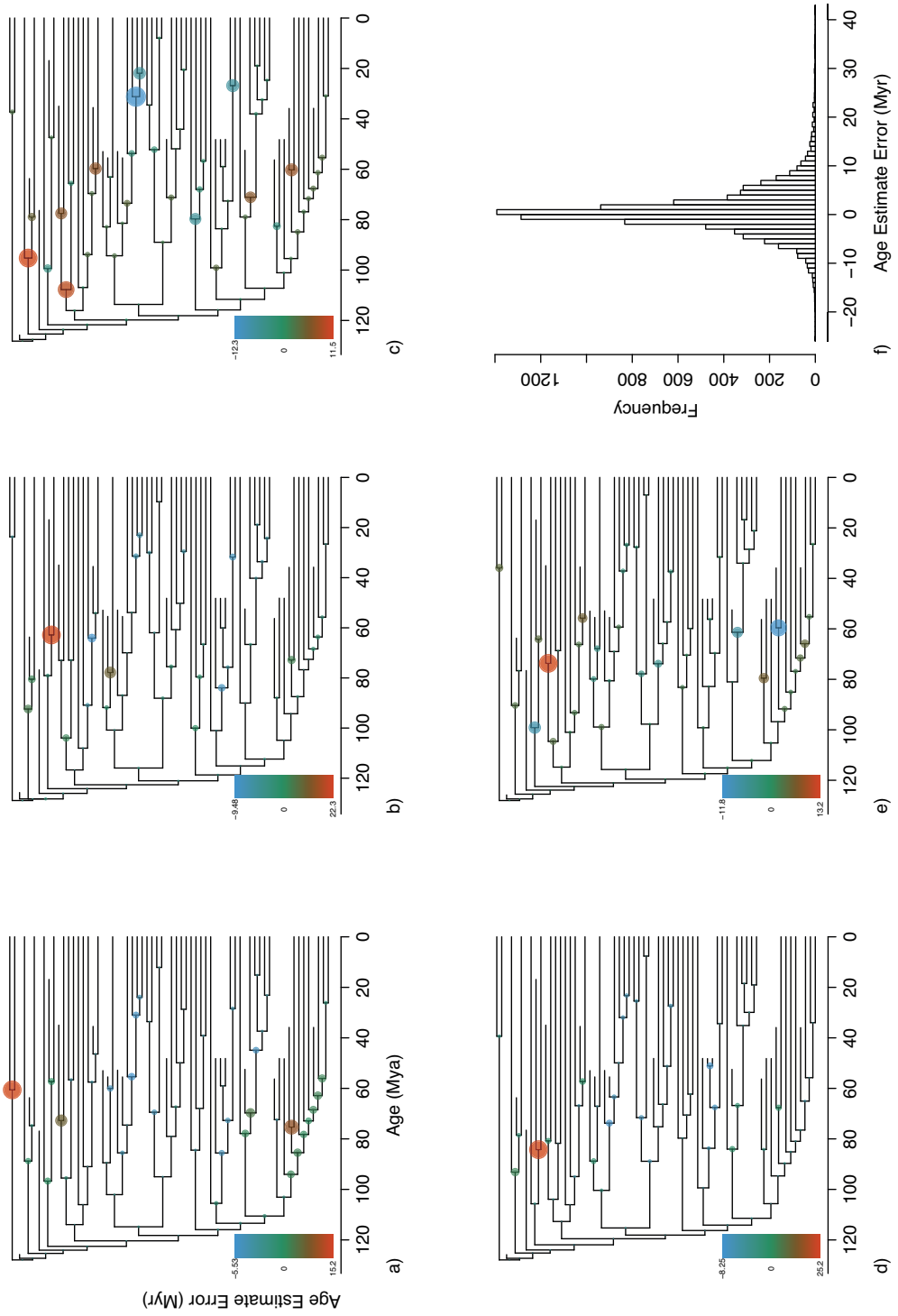


Figure S5.8 - (a-e) Mammalian time scaled phylogenies with node colour indicating the difference between mean estimated ages obtained from each respective reduced 245 character untreated matrix and average ages obtained from 30 replicate treated matrices when the distribution of missing data approximates that expected when soft characters are lost due to physical biostratinomic processes (scale factor = 2). (f) Histogram of node age estimate error across all 30 replicate analyses for all 5 independent matrix reductions.

Chapter Six Supplementary Material

# of characters	Max probability of unrepresented clades	Number of unrepresented clades with probability > 0.5	% of replicates with an MRC clade not presented in MCC tree	% of replicates with a missing clade in MCC tree with pp > mean pp of clades in MCC tree
100	0.74	32	0%	4%
1000	0.45	0	0%	0%
10000	NA	NA	NA	NA

Supplementary Table S6.1. Features of unrepresented clades in MCC trees constructed from posterior samples obtained using 100 replicated simulated datasets. For 10000 characters no valid but unrepresented clades were recovered in the posterior sample.

Published Articles

CellPress

Review

Dating Tips for Divergence-Time Estimation

Joseph E. O'Reilly,[1] Mario dos Reis,[2,3] and Philip C.J. Donoghue[1,*]

The molecular clock is the only viable means of establishing an accurate timescale for Life on Earth, but it remains reliant on a capricious fossil record for calibration. 'Tip-dating' promises a conceptual advance, integrating fossil species among their living relatives using molecular/morphological datasets and evolutionary models. Fossil species of known age establish calibration directly, and their phylogenetic uncertainty is accommodated through the co-estimation of time and topology. However, challenges remain, including a dearth of effective models of morphological evolution, rate correlation, the non-random nature of missing characters in fossil data, and, most importantly, accommodating uncertainty in fossil age. We show uncertainty in fossil-dating propagates to divergence-time estimates, yielding estimates that are older and less precise than those based on traditional node calibration. Ultimately, node and tip calibrations are not mutually incompatible and may be integrated to achieve more accurate and precise evolutionary timescales.

Trends

Total evidence dating constitutes a significant advance in divergence-time estimation. It overcomes problems with calibration by including fossil species on a par with their living relatives, using molecular sequence data from living species supplemented by morphological data from both living and fossil species.

The method relies on the controversial hypothesis of a morphological clock and suffers from the lack of development of realistic models of morphological evolution.

Most studies have failed to accommodate fossil age uncertainty. We present a protocol for characterizing and implementing this uncertainty, and demonstrate its impact on divergence-time estimation.

We argue that total evidence dating encompasses a suite of methods that can be used in bespoke combinations chosen to best suit the nature of specific divergence-time estimation studies.

Establishing an evolutionary timescale for Life on Earth has long been a fundamental goal of evolutionary biology, providing the framework for inferring modes and rates of molecular and phenotypic evolution, as well as a means of associating intrinsic evolutionary change to extrinsic causal factors. This endeavor was originally the domain of paleontologists, but it is now widely accepted that fossil data alone are insufficient because of the incompleteness of the fossil record [1]. Molecular clock dating methodology can be used to establish an evolutionary timescale by calculating the molecular distance between species, and by estimating absolute molecular evolutionary rates based on the oldest fossil evidence for the antiquity of the living lineages [2]. This powerful combination of molecular and paleontological data sees through the gaps in the fossil record, providing the only viable means of establishing an accurate evolutionary timescale.

Molecular clock methodology has been developed to accommodate tree-wide substitution rate heterogeneity [3–6], and precision has increased with the availability of genome-scale datasets (i.e., an effectively infinite amount of sequence data) [7]. However, further increases in accuracy and precision may only be possible with a concomitant increase in the precision of calibrations [5,8–10]. Hence, recent years have witnessed attempts to constrain the uncertainties associated with fossil-based calibrations, including phylogenetic position, age interpretation, and the degree to which calibrating fossils approximate the true time of divergence for the nodes that they calibrate [1,11,12]. Controversially, this requires not only the oldest fossil records of extant clades on which minimum age constraints are established, but also interprets the absence of older fossils attributable to the clade to establish maximum age constraints [1,11]. Alternatively, simple mathematical functions are employed to express, probabilistically, a visceral perception of the degree to which fossil minima reflect the time of lineage divergence [1,13]. Fossil occurrence data can also be modeled statistically, with or without reference to a phylogeny, to determine the extent of the temporal gap between the age of a clade and its oldest fossils

[1]School of Earth Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol BS8 1TQ, UK
[2]Department of Genetics, Evolution and Environment, University College London, London, WC1E 6BT, UK
[3]Present address: School of Biological and Chemical Sciences, Queen Mary, University of London, London, E1 4NS, UK

*Correspondence: phil.donoghue@bristol.ac.uk (P.C.J. Donoghue).

Box 1. Node Calibration

The development of TED has been shaped by a desire to overcome perceived shortcomings in node-calibration, the traditional means by which molecular clock analyses have been calibrated to absolute time. Node calibrations are established based on the oldest evidence for the existence of a clade and, most commonly, this is evidenced by the oldest fossil record of the clade. Thus, node calibrations require a prior phylogenetic hypothesis. This establishes a minimum age for the clades, but this must be complemented by a maximum age constraint. Deriving a maximum bound is more difficult to justify because it must, by necessity, rely on negative evidence. There are many methods for establishing maxima, including birth–death models [16] and statistical analysis of the stratigraphic distribution of fossils [81]. However, most commonly, maxima are established using taphonomic controls from the existence of outgroup taxa to interpret evidence of absence of ingroup taxa [82]. It is also necessary to establish the prior probability of the time of divergence between (and, using soft bounds [8], beyond) the minimum and maximum age constraints. The resulting probability density functions for each node calibration are ultimately combined with a stochastic branching model to derive effective priors on non-calibrated nodes in the tree, facilitating divergence-time estimates for all nodes.

Node calibrations have been considered unsatisfactory because they require a prior phylogenetic hypothesis and they fail to integrate uncertainty in the phylogenetic affinity of the calibrating fossils. This is problematic because the earliest fossil occurrences are often fragmentary, and therefore of uncertain affinity, and they are therefore ignored in favor of younger, better-known, and therefore phylogenetically-secure species. However, this leads to less-certain and less-informative calibrations – and dismisses an effectively infinite amount of other rate-informative fossil evidence. Some consider maximum age constraints based on fossil evidence or, rather, its absence, as unjustifiable, and establishing the nature of a probability density function spanning minimum and maximum constraints has little justification beyond gut-feeling. Unfortunately, arbitrary choices between competing parameters have an almost overwhelming impact on divergence-time estimates [83,84]. Finally, the node calibrations specified by users are invariably transformed in the establishment of the joint time prior, to the extent that they sometimes bear little relation to the original fossil evidence [7,83–85].

[14–16]. Attempts to constrain uncertainty with fossil calibrations must be welcomed, but they have not led to significantly increased precision in divergence-time estimation, not least because node calibrations require complex and often *ad hoc* interpretations of fossil and phylogenetic evidence to establish probabilistic calibrations, which are viewed by some as a grossly over-interpreted yet inadequate solution to a complex problem [17] (Box 1).

The recent introduction of fossil tip calibration [18,19], also known as 'tip-dating' or 'total evidence dating' (TED) has, therefore, enjoyed an enthusiastic welcome. This method requires both molecular sequence and morphological character datasets that are analyzed using molecular and morphological models of evolution, but its chief innovation is that it allows fossil species to be incorporated into divergence-time analyses on a par with their living relatives. This calibration methodology is analogous to the manner in which ancient DNA or archived viral sequences of known age are employed to infer rates of evolution among extant species or strains [20]. In this case, fossils of known age calibrate the rate of evolution based on their phylogenetic position, branch length, and an inferred rate of evolution. Phylogenetic topology may be estimated independently or co-estimated with the divergence-time analysis, and the rate of evolution maybe based on independent or correlated rates of morphological and molecular evolution.

Thus, tip-calibration obviates many of the controversies associated with node-calibration. First, fossil species inform the evolutionary rate without recourse to *ad hoc* assumptions about the degree to which these species approximate the age of a living clade. Second, because time and topology can be co-estimated, it becomes possible to include older, temporally more-informative fossils that could not be used for node-calibration because their phylogenetic position is uncertain. Third, because calibrations no longer serve as prior estimates of clade age, tip-calibrations can be drawn from any and all fossil species, removing restrictions on the amount paleontological data that can be included in divergence-time studies. Finally, tip calibrations summarize the age of a single species only, avoiding the over-interpretation of negative evidence in establishing maximum constraints.

Tip-calibration was originally introduced based on empirical divergence-time analyses of insects [19] and amphibians [18], and it has since been applied to mammals [21–26], teleost fishes [27–31],

arachnid spiders [32,33], flies [34], and plants [35]. The approach has been extended to analyses of entirely extinct clades, relying exclusively on morphological data [36]. While tip-calibration was initially advocated on the basis that it was less sensitive to root time prior densities, and yielded more precise divergence-time estimates in comparison to node-calibration [19], subsequent studies have shown the reverse to be true [30,32]. Furthermore, tip-calibration has proven consistently to yield older age-estimates than traditional node-calibration [19,21–24,30,32,33]. Thus, while it is clear that in incorporating all data pertinent to divergence-time estimation, and tip-calibration is the most promising approach for establishing accurate and precise evolutionary timescales, at present it appears to be less accurate than conventional node calibration methods. Below we consider the factors biasing current methods employing tip-calibration, and suggest ways in which they can be developed to obtain more accurate divergence-time estimates.

Models of Morphological Character Evolution and the Incompleteness of Fossils

While there are several nested models of molecular substitution, morphological models have not enjoyed much development, with only a handful proposed to date and even fewer actually implemented in popular software packages [37–42]. The Mk model of discrete character evolution has been utilized in all published tip-calibrated analyses to date [43]. The Mk model is a *k* states generalization of the JC69 model of molecular substitution and, inevitably, it possesses many simplifying assumptions that may not hold true for morphology [44]. Independent evolution of sites and equal equilibrium frequencies are two factors that are particularly difficult to justify for morphological evolution. Alternative models utilizing continuous characters [45] or the threshold model [46,47] are appealing alternatives, but they have yet to be implemented.

The inherently incomplete nature of fossil phenotypic data, in comparison to living species, is undoubtedly a challenge to tip-calibrated divergence-time analyses. The impact of missing sequence data on Bayesian phylogenetic topology estimation has been investigated, with the majority of studies indicating that it is unlikely to have a strong negative impact [48–52], except where there is a comparatively small number (not proportion) of non-missing sites [49]. This is clearly a problem for topology estimation based on phenotype where datasets are generally very small in comparison to molecular sequence alignments. This issue is exacerbated by the decidedly non-random nature of missing phenotype data in fossil species [53,54]. Fossil data are invariably biased towards the preservation of phenotypic characters that are manifest in, or as, mineralized skeletal structures. Even where soft tissue characters are exceptionally well preserved, some groups exhibit a phenomenon coined ‘stem-ward slippage’ in which features are lost to decay in reverse phylogenetic order, making their fossils appear artefactually to belong to more primitive evolutionary grades [53,54]. While the impact of these factors on topology estimation has been considered, it has not been investigated explicitly in the context of time and rate estimation [53].

For tip-calibrated divergence-time analyses, the likely impact is twofold: calibrating fossil species will be assigned to erroneously early-branching positions within the phylogeny, and the branch lengths will be underestimated, both owing to their lack of shared-derived and autapomorphic soft-tissue characters, missing artefactually as a consequence of non-random decay patterns. Both these phenomena will influence rate estimates and, therefore, divergence-time estimates. To minimize the negative influence of missing data, sub-sampling approaches have been proposed, allowing the use of only the most completely coded taxa or characters. While it has been argued that such approaches have minimal impact on topology and age estimation [18,19], this is unlikely to hold true for non-random missing data. Alternatively, a model of fossilization could be employed that accounts for the directed loss of characters during

preservation, but modeling this process may be entirely unrealistic given that fossilization potential varies with environment and taxonomic group.

Dating Tips and Calibration Strategies

Almost all TED studies conducted so far have employed point age estimates for the fossil species used as tip-calibrations, assuming implicitly that the age of the fossil is known without error. This has been done on the sometimes explicit justification that the errors associated with the dating of fossils are negligible [19,33]. This approach is reminiscent of the point age estimates for node calibrations, employed when divergence-time estimation was in its infancy, and none of the lessons learned from the development of node-calibration strategies [1,11,13] have been transferred to studies that employ fossil tip-calibration. It is well established that the age of a fossil can rarely, if ever, be known without error, and this uncertainty must be accommodated regardless of whether the fossil is used in the construction of a node or tip-calibration. The age of any fossil occurrence can be constrained only to within an envelope of minimum–maximum bounds, the span of which varies depending on the attendant evidential context. Node-calibrations are based principally on the earliest secure fossil record of a clade (Box 1), and it is thus necessary to determine only the minimum age interpretation of the calibrating fossil [13,55]. At the least, the age of a tip-calibrating fossil requires establishing both its minimum and maximum age interpretations. For both the minimum and maximum age interpretations, this invariably entails a tortuous daisy-chain of litho-, bio-, chemo-, cyclo-, and/or magneto–stratigraphic correlations between the site of the fossil occurrence and another in which a geochronological absolute date has been established, at each step taking the minimum or maximum relative age interpretation, as appropriate, leading to iteratively increasing age uncertainty (Box 2 gives a worked example). It is likely that, in many instances, this uncertainty will exceed that associated with local node-calibrations, although tip calibrations may prove more palatable because they rely on fewer assumptions.

Borrowing from practice in establishing node-calibrations, the age uncertainty associated with a fossil species can be modeled as a uniform distribution if there is equal probability of the age of the fossil, per unit time, between minimum–maximum age interpretations. Alternatively, the variety of parametric distributions already implemented for node calibrations may be redeployed in instances where there is justification for focusing uncertainty closer to the minimum, maximum, or mid-range between age bounds. The range of available distributions and instances in which they may be deployed, are discussed in Box 3.

Tip-calibrations present further peculiarities that should also be considered in attempting to integrate uncertainty associated with their age. For example, many fossil species employed in the node-calibration of divergence-time analyses are not single occurrences but, rather, occur through a stratigraphic age range. This is of little relevance to node-calibration used to establish a clade age minimum; however, in establishing a tip-calibration this is much more germane. Given that, by definition, such species will exhibit little or no morphological variation, it seems appropriate that this age range should be incorporated into the age uncertainty associated with the fossil (Box 4 expands upon this idea). Ultimately, it may prove useful to integrate this information, in the form of effective stasis in the set of traits analyzed, into the inference of rate variation across the tree.

Because tip-calibration and TED have been presented as a means to achieve greater precision in divergence-time estimation [19], it is pertinent to consider whether this can be sustained while integrating the uncertainty associated with the age of fossil tips. To this end, we reanalyzed the dataset from the seminal TED study [19], in which tip-calibrations were utilized to estimate divergence times for Hymenoptera. Ronquist and colleagues focused on the theoretical and practical introduction of the method, and they did not take account of the uncertainty associated

Box 2. The Construction of a Tip Calibration

Palaeathalia laiangensis was recovered from the Laiyang Formation in Liaoning, China, which can be divided into four members, the third of which has yielded most fossils. Although the Laiyang Formation contains no directly dateable elements, correlation with the base of the Yixian Formation, also of China, allows the use of radiometric dates for the base of this formation to inform the age of the Laiyang Formation. Similarly, the unit overlying the Laiyang Formation, the Houkuang Formation, contains dateable elements, allowing an age for the base of this formation to constrain the age of the top of the Laiyang Formation. Because we consider the age of the fossil species *P. laiangensis* to lie within the chronological interval between the top and base of the unit of its provenance, and without further information to constrain the limits and distribution of probability, we can use the ages of these limits to determine the bounds of our calibration. Correlation with the Yixian Formation can be made based on numerous palynological and faunal similarities, mostly with the lowermost member of the Yixian Formation, the Lujiutun Bed. While these sources may not individually be considered conclusive, numerous biostratigraphic similarities strongly support this correlation [86–90]. Radiometric dates of 128.4 Ma ± 0.2 Myr have been acquired from the base of the Lujitan Bed, which can be used to determine the age of the base of the Laiyang Formation on the basis of the correlation between these units [90–92] (Figure I).

The Laiyang Formation is succeeded by the Qingshan Group, of which the Houkuang Formation is the lowermost member. Because the Laiyang Formation can be no younger than the overlying unit, an age for the base of the Houkuang Formation can provide a minimum age for the Laiyang Formation. U–Pb dating of zircons from the base of the Houkuang Formation has yielded dates of 106 Ma ± 2 Myr, which can be used to constrain the minimum age of the Laiyang Formation [93]. Because no dates are available to further constrain the limits of this formation, and without any further information regarding the manner in which the probability of the age of *P. laiangensis* should be distributed, a uniform distribution spanning the full range of uncertainty in radiometric dates across the interval (128.6–104 Ma). This tip age can be contrasted with that utilized by Ronquist *et al.* [19] of a fixed age of 140 Ma, which falls significantly outside the bounds of this calibration.

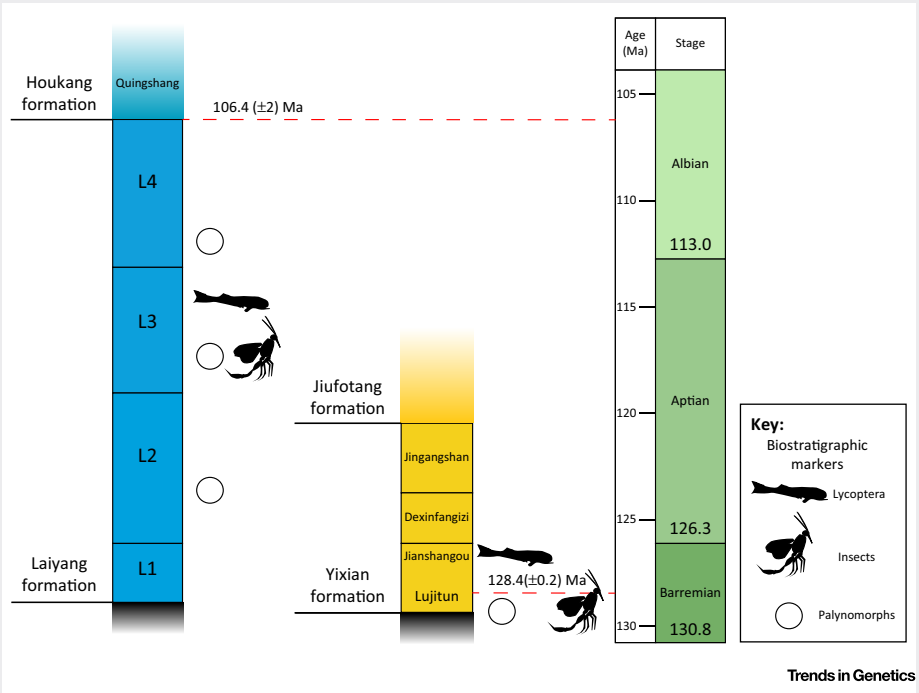


Figure I. Construction of a Tip-Calibration for *P. laiangensis* Based on Stratigraphic Correlation Between the Unit of Provenance, The Laiyang Formation, and the Yixian Formation of China.

with the fossils used in tip-calibration. We reproduced the calibrations for each fossil tip, accommodating uncertainty in the age of each fossil species using probabilistic distributions (Box 2 gives an example of this process). In contrast to previous assertions, that the uncertainties associated with tip ages would be negligible [19,33], our attempts to capture a realistic estimate of the associated uncertainty results in tip-calibrations that span tens of millions of years – in contrast to the errorless estimates of age estimates used by the original authors. To determine

the performance of node- versus tip-calibration, we also constructed node-calibrations following established best practice [11] (see the supplemental information online). On average, recalibrated node priors were 23 Myr wider than the original calibrations. In both tip- and node-calibrations, uncertainty was modeled as a uniform distribution. Analyses were performed in MrBayes 3.2.2 [41] in broadly the same manner as the original article (see the supplemental information online for details). Precision was measured as the width of the 95% confidence interval (CI) for posterior estimates of node age for 14 key in-group clades that could be resolved.

Our analyses show that when fossil age uncertainty is properly accounted for, tip-calibrated analyses do not necessarily yield divergence-time estimates that are more precise than those derived using node-calibration. Furthermore, for 27% of fossil taxa, the 95% highest posterior density (HPD) estimates of fossil tip age did not encompass the original fixed tip-calibration, demonstrating the importance of appropriate prior construction. Divergence-time estimates based on node-calibration are the most precise in all but four of the component clades (Figure 1). In line with almost all previous TED studies, tip-calibration yields clade ages that are older, in general, than like-for-like estimates based on node-calibration, the only exceptions being divergences outside Hymenoptera. These deeper divergence times are most prominent in

Box 3. Density Distributions for Fossil Tip-Calibration

The wide range and flexibility of probability distributions has allowed the accurate incorporation of uncertainty into fossil calibrations. Unfortunately, encapsulating prior knowledge of fossil age as a density distribution is not a straightforward task, and the application of density distributions with arbitrarily assigned parameters can have profound effects on age estimates [84]. Although computational methods exist for the integration of fossil stratigraphic range and geochronological age data [94], they are rarely implemented in evolutionary studies, and in their place it is important that the construction of density distributions is justified explicitly. For tip-calibration, several distributions are applicable, depending on the context in which uncertainty manifests itself. Six distributions are presented here using the calibration of the Hymenopteran fossil *Eoxyela* (minimum = 141 Ma, maximum = 168 Ma) as an example (Figure I).

(i) Exponential Distribution (Figure I,i). Exponential distributions introduce diminishing probability over time. These calibrations are particularly useful when the weight of evidence suggests that the true age of a tip is close to the minimum bound but a much more ancient age cannot be ruled out. The rate parameter determines how far back the distribution extends to (λ), with its reciprocal being equal to the mean. Here two parameterizations reflect separate assumptions of how ancient the clade may be.

(ii) Gamma Distribution (Figure I,ii). The gamma distribution has two parameters, shape (α) and rate (β), and is relatively flexible compared to other available distributions. For example, when $\alpha < 1$, the distribution is L-shaped with the mode at zero and with a long tail. When $\alpha = 1$ the distribution reduces to the exponential distribution. Finally, when $\alpha > 1$ the distribution has a mode away from zero. As the value of α increases, the gamma converges to the normal distribution with mean of α/β and variance α/β^2. An offset is required to express the distribution relative to the minimum age; here an offset of 141 Ma is used.

(iii) Normal Distribution (Figure I,iii). The normal distribution has seen limited use for node calibrations, but it may prove more useful in a tip-dating context. Normal distributions place equal diminishing probability (determined by the variance σ^2) either side of the mean (μ), and may be useful when a species is known from the middle of a unit only. Here the upper and lower bounds of the species chronological distribution are set at 2 standard deviations from the mean, allowing for age estimates that violate the bounds (Figure I,iii).

(iv) Point Calibrations (Figure I,iv). Assume that the provided age is absolutely correct, disregarding any meaningful interpretation of the fossil record; therefore, erroneously inflated confidence in posterior age estimates is introduced owing to increased specificity in the prior distribution [95]. Here the assumed tip age is at the mid-point of the chronological distribution of the taxon.

(v) Uniform Distributions (Figure I,v). Uniform distributions place equal probability across the interval (*a,b*). This distribution is applicable when a fossil is known from a single unit in which dates can be derived for the base and top, but no additional constraints on the distribution of age can be demonstrated.

(vi) Lognormal Distribution (Figure I,vi). Lognormal distributions allow the assignment of diminishing probability that the first appearance of a species is actually described by the age of the fossil specimen itself. The distribution has two parameters, the log-mean (μ) and log-standard deviation (σ). σ determines the shape of the distribution; when it is close to zero, the distribution is symmetrical, and when it is large, the distribution becomes very asymmetric with a long tail and with the mode of the distribution moving towards zero.

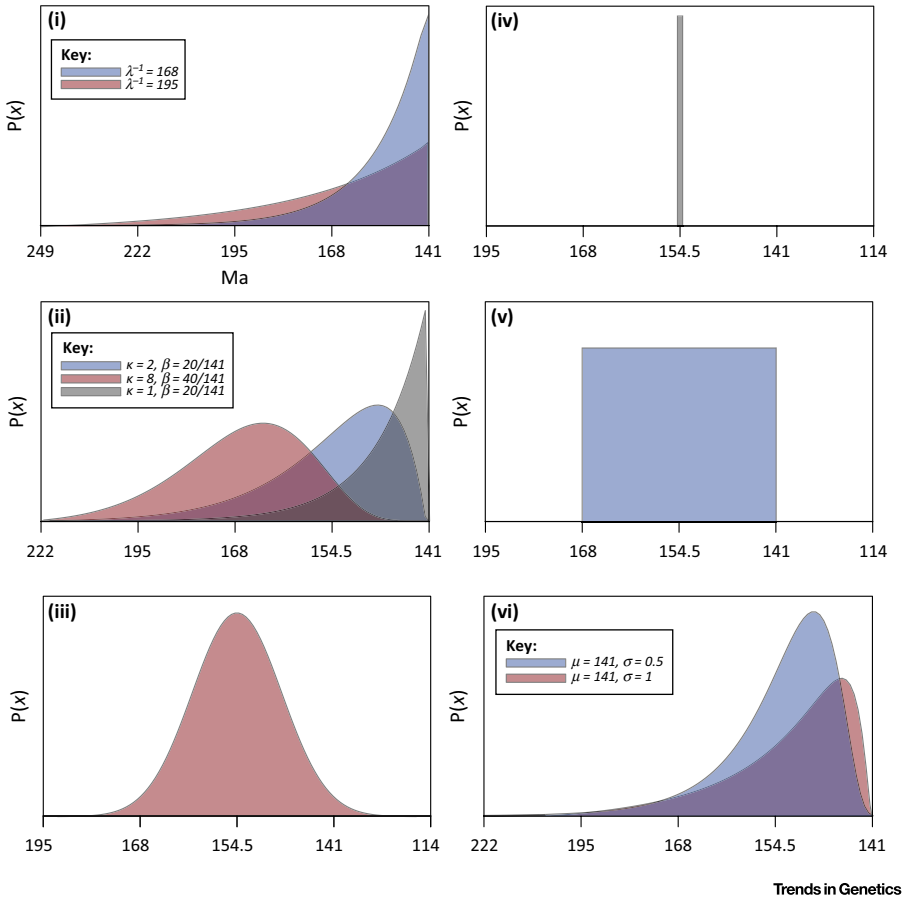


Figure I. Six Alternative Probability Density Functions Commonly Used to Encapsulate Prior Knowledge of the Chronological Distribution of a Fossil Tip. Here the calibration of the fossil taxon *Eoxyela* is used to demonstrate the characteristics of the different distributions.

Vespina, where it appears that relaxing the constraint on the age of *Mesorussus* (which was assigned to Vespina in both our analysis and the original analysis [19]) from 94 Ma to 93.7–140.3 Ma leads to the older age estimates.

While we were able to repeat the results of the original analysis using the original calibrations, we were unable to reproduce the topological resolution and/or monophyly of Xyelidae, Pamphilioidea, and the placement of fossil taxa *Palaeathalia, Cleistogaster*, and *Prosyntexis* when employing our revised tip-calibrations. Because the only variable between our analyses is the method of calibration construction, it appears that the more realistic age-uncertainty associated with the fossils in our revised tip-calibrations has impacted on topology estimation as part of the co-estimation of topology and time. Thus, by implication, accommodating the realistic age uncertainty associated with fossil tip-calibrations also impacts indirectly on rate and clade age estimates by contributing to topology estimation.

Claims of the superiority of tip-calibration over node-calibration appear unfounded when fossil age uncertainty is accommodated equally. Furthermore, it is not entirely clear that node calibrations are redundant in tip-calibration studies because, logically, they can still serve their purpose of constraining node age estimates and rate variation. One way to assess whether they

Box 4. Tip Calibrations and Apparent Morphological Stasis

The exact definition of what a fossil tip represents has not yet been defined explicitly because it is currently not clear whether calibrations should be constructed based on the age of an individual fossil, or to reflect the minimum age of the fossil species to which it is assigned, or the total known temporal range of that species. For a species with only one known fossil the situation is simple: the tip represents the evolutionary path to the first appearance of the suite of characters it possesses, and it is therefore justifiable to assign a calibration based on the provenance of that individual fossil. It is less clear how a fossil species known from several temporal intervals should be represented in terms of the tip-age. For example, consider the scenario outlined in Figure I. A fossil species (†) with a chronological distribution of 10 Myr is recovered from two serial units (A and B), each of 5 Myr in length, with no overlap. The suite of characters at the start of the first deposit and at the end of the last deposit is the same; there is effective morphological stasis. In this scenario, morphological and molecular rates are certainly unlinked because, despite the perceived evolutionary stasis, there will be molecular evolutionary change. The choice of calibration bounds in this situation can readily lead to the over- or under-estimation of rates on surrounding branches, by compressing or stretching the length of the branch subtending the fossil species. If the tip age is constrained based on the limits of the oldest occurrence, apparent morphological stasis is not accommodated; constraining tip age based on the combined time span of both temporal occurrences is likely to inference of lower rates on other branches [96]. An alternative calibration strategy might be to assign point estimates based on the statistically derived 95% CI for the lower limit of the true stratigraphic range of such fossil species, ignoring the protracted stasis but explicitly calibrating the origin of the suite of fossilized characters [97]. Is this morphological stasis a derived state that should not be extrapolated across the tree, or it is inherited from earlier members of the lineage and should therefore be used to inform rates elsewhere? Morphological stasis is hypothesized to be driven largely by either stabilizing selection [98] or developmental constraints [99], but a consensus as to which is the controlling factor has still to be reached [100]. If the latter obtains, it is likely that calibrations need to incorporate stasis as it is an inherited trait.

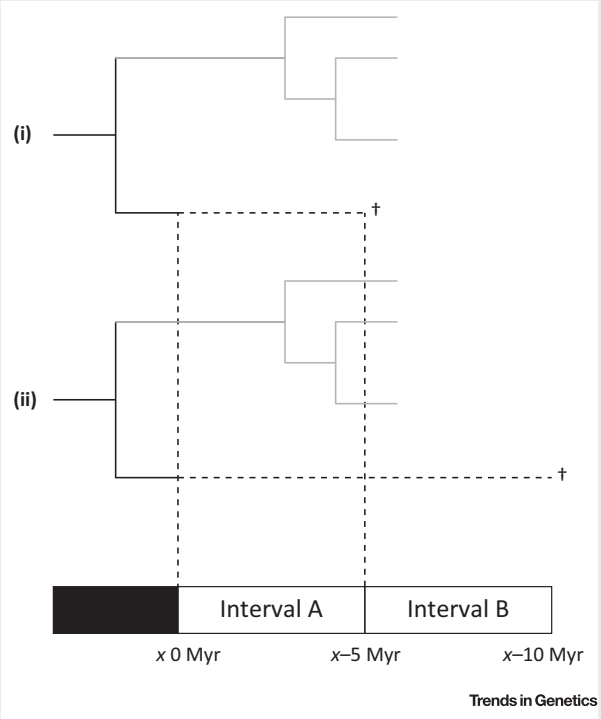


Figure I. Potential Branch Lengths (Dotted Lines) When Fossil Tip Age Is Calibrated Based on Different Stratigraphic Limits when a Fossil Taxon (†) Is Recovered from Multiple Units (A and B). Calibrations constructed from the full stratigraphic range of the fossil taxon will incorporate stasis into the model, but may induce lower rates on nearby branches (ii). Calibrations constructed from the first appearance on the fossil taxon ignore the protracted stasis, and may induce inflated rates on surrounding branches (i).

are still useful in this role is in comparing traditional node calibrations and the posterior node-age estimates based on analyses employing tip-calibrations. We did this for the nine nodes for which we have constructed calibrations. The results (Figure 2) show that, while all of the node age estimates derived from tip-calibration are old relative to the node calibrations, four fall fully outside these node age constraints. It could be argued that this demonstrates the fallacy of

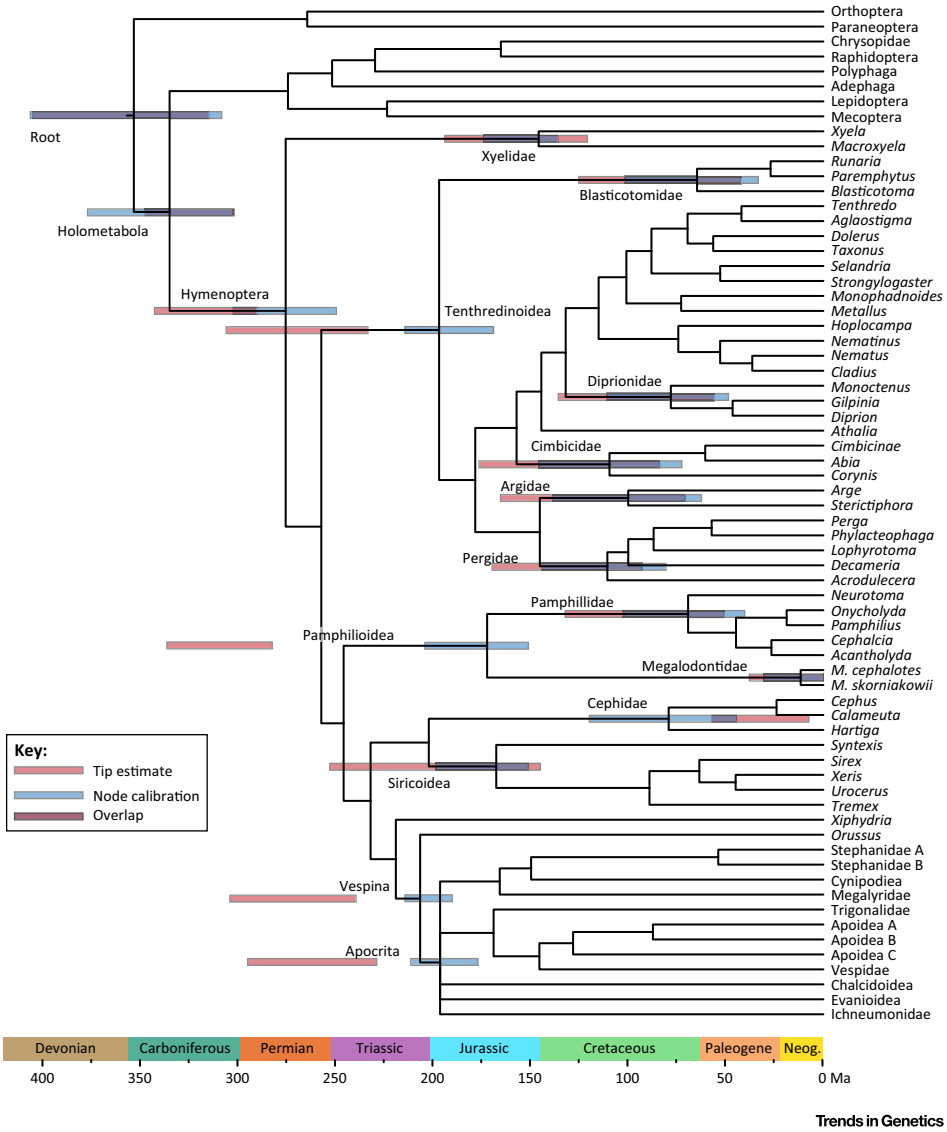


Figure 1. A Dated Phylogeny of Hymenoptera Produced Using Node-Calibrations. Node bars represent 95% highest posterior density (HPD) for node ages estimated with either node-calibration or total evidence dating (blue and red respectively). The dotted lines join HPD bars to the node for which they represent age estimate confidence, and do not represent an extension of the confidence interval.

fossil-based maximum age constraint, however, two of the node age estimates include age ranges that are younger than the minimum age constraints based on the empirical paleontological evidence. Evidently, there remains a role for node age constraints, even in tip calibration divergence-time analyses.

TED – Less Than the Sum of Its Parts?

While TED has been presented as an alternative approach to conventional node-calibrated molecular clocks, this is a false dichotomy. TED is a specific combination of approaches that are neither inextricably linked, nor mutually exclusive from node-calibrated molecular clock analysis. These include: (i) the relaxed morphological clock, (ii) tip-calibration, and (iii) co-estimation of time and topology. In practice, these methods can and have been deployed in isolation in augmenting

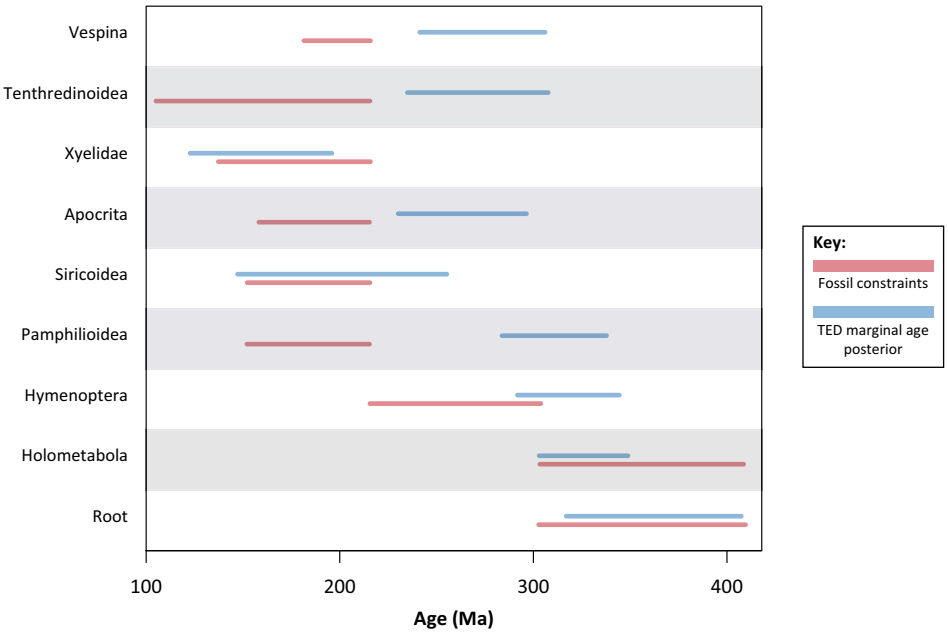


Figure 2. Comparison Between Marginal Posterior Distributions on Nine Node Ages Estimated with Total Evidence Dating (TED, Blue), and Prior Clade-Age Constraints Employed for Node-Calibrated Analysis of the Same Data (Red). The calibrations for node-calibrated analysis encapsulate the fossil evidence for the possible age of each clade. A lack of overlap at any node implies that there is discordance between the TED effective prior on that node and the fossil record. Discordance between these two distributions demonstrates that TED may lead to empirically unsupportable clade age estimates.

conventional molecular clock analyses. For example, the divergence-time study of Schrago and colleagues [22] of New World primates followed a two-step protocol, using the posterior age estimates from a conventional molecular clock analysis of living species as time-priors on node ages in a morphological clock analysis including both living and fossil species. At the least, this approach obviates the problematic assumption that molecular and morphological data co-vary, following a single rate model. Lee *et al.* [56] co-estimated time and topology using dated tips and a morphological clock, eschewing molecular data altogether, in their analysis of body size evolution through the dinosaur–bird evolutionary transition. This approach will surely be adopted widely as paleontologists seek to obtain clade ages, rather than minimum ages, for entirely extinct clades. However, this enthusiasm may be short-lived given that tip-calibration approaches have consistently yielded older clade age estimates than conventional molecular clock studies – against which paleontologists have a long tradition of objecting violently [57]. Combining ancient DNA and morphological data is another possibility afforded by tip-calibration, as has been applied to studying Pantherhine phylogeny [23]. This combination of ancient morphology and DNA may facilitate more accurate estimates of evolutionary rate.

While there has been enthusiasm in the application of the TED approach, not least because it provides a platform for the integration of many disparate sources of uncertainty, it is arguable that in so doing this approach serves as a black box that disengages the user from the assumptions underpinning the analysis, many of which are very difficult to justify (see Outsanding Questions). One of the most problematic, potentially, is the co-estimation of time and topology, which, as we have demonstrated, allows fossil ages to constrain their phylogenetic position and, therefore, impact on the estimation of rates and dates. This follows the common-sense

expectation that the age of a fossil species must reflect their phylogenetic position. Indeed, phylogeny estimation integrating the relative stratigraphic age of fossil species has a long tradition in paleontology, but it has been much debated [58–62] and generally abandoned in favor of phylogenetics based on phenotype, perhaps refined by stratigraphy, except in groups with exceptionally rich fossil records that are rarely if ever the focus of divergence-time studies [63]. Although there is a broad correlation between clade age and phylogenetic branching order [64], this relationship breaks down as fossil taxon sampling decreases [65]. It is complicated further by secular biases in the rock record, which serve to telescope temporally-distinct fossil species originations and extinctions [66], and in the differential preservation of fossil groups and the environments in which they lived [67]. Thus, there appears little justification for the co-estimation of time and topology where fossil ages contribute to their phylogenetic position. We strongly advocate the prior analysis of topology before divergence-time estimation. It is unfortunate that this approach precludes the integration of phylogenetic uncertainty into divergence-time estimation, but resolving phylogenetic uncertainty using tip age does not appear viable using current methods.

The majority of TED analyses model branch rates as linked across morphological and molecular partitions (i.e., the application of rate multipliers to describe inter-partition rate heterogeneity [68–70]). The suitability of this assumption for partitioned molecular data alone has been investigated, and partition-specific clocks have been developed for when this assumption is not met [68,71]. However, the effect of morphological and molecular partition-specific clocks has barely been considered [18,68,72], and most studies employ a single, partition-linked clock despite the fact that a strong covarying relationship between molecular and morphological rates has never been demonstrated [73–75]. Morphological rate heterogeneity has long been considered likely to significantly dwarf its molecular counterpart, suggesting that the assumption of phenotypic and molecular rate correlation is unjustified [76,77]. Molecular rates are interpreted as genome-wide measures of the number of substitutions accumulated per time unit, while morphological rates reflect only those aspects of the genome that specify the phenotypic traits analyzed, further diminishing any expectation of covariance between molecular and morphological evolutionary rates [73,78]. In this light, it is perhaps unsurprising that unlinked partition-specific clocks have been found to be better-fitting than a single linked clock for mixed data analyses [79].

While node and tip-based calibration have been presented as competing approaches, they are not mutually exclusive. Indeed, some temporal constraints on clade age are better suited to being implemented as node-calibrations. This is particularly true of biogeographic calibrations where, based on the modern and ancient biogeographic distributions of evolutionary lineages, it is acceptable to assume that a dateable vicariance event, such as continental fragmentation, is causal to lineage divergence. Similarly, some fossil evidence is better reflected as node-age calibrations rather than through including component fossil species as tip-calibrations. Node and tip-calibrations have already been employed together to calibrate interior nodes of the out-group, while allowing an unconstrained in-group topology, or as part of a highly constrained topology in which fossil taxa are assigned to predetermined clades [19,80]. However, this must be extended to allow node-calibrations throughout the tree. This approach requires a fixed topology (or at least backbone constraints compatible with calibrated nodes) and, thus, precludes the possibility to co-estimating time and topology, but, as we have argued, this may not be a material loss. Winterton and Ware [34] have shown that combining node and tip calibrations in this way yields younger estimates than node (or, presumably, tip) calibrations alone. Node calibrations may serve to mitigate against the propensity for tip-calibration-based studies to yield unacceptably-ancient divergence dates because it places additional constraints on the age of internal nodes of the tree, providing local checks on branch length and rate variation.

Finally, it is likely that the mismatch between divergence-time estimates based on node and tip-calibration strategies is based at least in part in the shortcomings of the Mk model in explaining the phenotypic data commonly used in tip-calibration studies. The Mk model fails to account for expected characteristics of cladistic data, including the covariation of characters that are biologically linked and logically linked through character design. Doubtless, the excitement surrounding the combined use of morphological and molecular data for divergence-time analysis will lead to the development of this and other models of evolution. However, it may also be appropriate to consider different approaches to characterizing phenotype, such as through the types of continuous variable characters obtained through morphometry of features such as skull suture patterns, tooth shape, or the dimensions of limb bones. The stochastic variation of such data is more similar to the variation seen in molecular sequence alignments and, as such, may be more readily modeled and better suited to combined data divergence-time analysis.

Concluding Remarks

The advances inherent in TED provide an excellent platform for the further development of methods for divergence-time analysis. However, many aspects of the principal evolutionary model for phenotypic data currently employed are violated by the evolutionary process it attempts to encapsulate. The extent of these problems is so great that divergence-time estimates derived using tip-calibration cannot enjoy the same confidence as conventional node-calibrated molecular clock studies. However, with the development of evolutionary models, protocols for dating fossil species and dealing with missing data, TED encompasses a variety of powerful tools, the combination of which can be chosen to best test the hypothesis at hand. It also provides a viable framework for the best and greatest use of paleontological data that may serve as a nexus of the unification of paleontological and molecular approaches to establishing evolutionary timescales.

Outstanding Questions

How adequate is the Mk model of morphological evolution for estimating divergence times? There has been little development of this model in the past 15 years. Its suitability for morphology-based divergence-time estimation remains unclear.

What is the best method for modeling the relationship between molecular and morphological evolutionary rate? Many analyses model these rates as correlated variables, but it is unclear how well this approach encapsulates their true relationship.

How congruent with the fossil record are tip-calibration node-age priors? Exploring the induced time prior is a non-trivial task for TED analyses owing to the co-estimation of time and topology. Without knowledge of the time prior it is not possible to determine whether zero probability is being assigned to age estimates that violate minima derived from the empirical evidence contained within the fossil record.

Are morphological data best characterized as categorical or continuous variable data for the purposes of divergence-time estimation?

Supplemental Information

Supplemental information associated with this article can be found online at http://dx.doi.org/10.1016/j.tig.2015.08.001.

References

1. Donoghue, P.C.J. and Benton, M.J. (2007) Rocks and clocks: calibrating the Tree of Life using fossils and molecules. *Trends Ecol. Evol.* 22, 424–431
2. Zuckerkandl, E. and Pauling, L. (1965) Molecules as documents of evolutionary history. *J. Theor. Biol.* 8, 357–366
3. Sanderson, M.J. (2002) Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Mol. Biol. Evol.* 19, 101–109
4. Thorne, J.L. *et al.* (1998) Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* 15, 1647–1657
5. Rannala, B. and Yang, Z. (2007) Inferring speciation times under an episodic molecular clock. *Syst. Biol.* 56, 453–466
6. Drummond, A.J. *et al.* (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 4, e88
7. dos Reis, M. *et al.* (2012) Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proc. R. Soc. B: Biol. Sci.* 279, 3491–3500
8. Yang, Z. and Rannala, B. (2006) Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol. Biol. Evol.* 23, 212–226
9. Dos Reis, M. and Yang, Z. (2013) The unbearable uncertainty of Bayesian divergence time estimation. *J. Syst. Evol.* 51, 30–43
10. Zhu, T. *et al.* (2015) Characterization of the uncertainty of divergence time estimation under relaxed molecular clock models using multiple loci. *Syst. Biol.* 64, 267–280
11. Parham, J.F. *et al.* (2012) Best practices for justifying fossil calibrations. *Syst. Biol.* 61, 346–359
12. Benton, M.J. and Donoghue, P.C. (2007) Paleontological evidence to date the tree of life. *Mol. Biol. Evol.* 24, 26–53
13. Ho, S.Y. and Phillips, M.J. (2009) Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. *Syst. Biol.* 58, 367–380
14. Marshall, C.R. (1994) Confidence-intervals on stratigraphic ranges – partial relaxation of the assumption of randomly distributed fossil horizons. *Paleobiology* 20, 459–469
15. Wilkinson, R.D. *et al.* (2011) Dating primate divergences through an integrated analysis of palaeontological and molecular data. *Syst. Biol.* 60, 16–31
16. Heath, T.A. *et al.* (2014) The fossilized birth-death process for coherent calibration of divergence-time estimates. *Proc. Natl. Acad. Sci. U.S.A.* 111, E2957–E2966
17. Heads, M. (2012) Bayesian transmogrification of clade divergence dates: a critique. *J. Biogeogr.* 39, 1749–1756
18. Pyron, R.A. (2011) Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. *Syst. Biol.* 60, 466–481
19. Ronquist, F. *et al.* (2012) A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. *Syst. Biol.* 61, 973–999
20. Drummond, A.J. *et al.* (2003) Measurably evolving populations. *Trends Ecol. Evol.* 18, 481–488
21. Slater, G.J. (2013) Phylogenetic evidence for a shift in the mode of mammalian body size evolution at the Cretaceous–Palaeogene boundary. *Methods Ecol. Evol.* 4, 734–744

22. Schrago, C.G. *et al.* (2013) Combining fossil and molecular data to date the diversification of New World Primates. *J. Evol. Biol.* 26, 2438–2446
23. Tseng, Z.J. *et al.* (2014) Himalayan fossils of the oldest known pantherine establish ancient origin of big cats. *Proc. R. Soc. B: Biol. Sci.* 281, 20132686
24. Slater, G.J. (2015) Iterative adaptive radiations of fossil canids show no evidence for diversity-dependent trait evolution. *Proc. Natl. Acad. Sci. U.S.A.* 201403666
25. Dembo, M. *et al.* (2015) Bayesian analysis of a morphological supermatrix sheds light on controversial fossil hominin relationships. *Proc. Natl. Acad. Sci. U.S.A.* 282, 20150943
26. Marx, F.G. and Fordyce, R.E. (2015) Baleen boom and bust: a synthesis of mysticete phylogeny, diversity and disparity. *R. Soc. Open Sci.* 2, 140434
27. Near, T.J. *et al.* (2014) Phylogenetic relationships and timing of diversification in gonorynchiform fishes inferred using nuclear gene DNA sequences (Teleostei: Ostariophysi). *Mol. Phylogenet. Evol.* 80, 297–307
28. Dornburg, A. *et al.* (2015) The impact of shifts in marine biodiversity hotspots on patterns of range evolution: evidence from the Holocentridae (squirrelfishes and soldierfishes). *Evolution* 69, 146–161
29. Alexandrou, M.A. *et al.* (2013) Genome duplication and multiple evolutionary origins of complex migratory behavior in Salmonidae. *Mol. Phylogenet. Evol.* 69, 514–523
30. Arcila, D. *et al.* (2015) An evaluation of fossil tip-dating versus node-age calibrations in tetraodontiform fishes (Teleostei: Percomorphaceae). *Mol. Phylogenet. Evol.* 82, 131–145
31. Dornburg, A. *et al.* (2015) Phylogenetic analysis of molecular and morphological data highlights uncertainty in the relationships of fossil and living species of Elopomorpha (Actinopterygii: Teleostei). *Mol. Phylogenet. Evol.* 89, 205–218
32. Wood, H.M. *et al.* (2013) Treating fossils as terminal taxa in divergence time estimation reveals ancient vicariance patterns in the palpimanoid spiders. *Syst. Biol.* 62, 264–284
33. Sharma, P.P. and Giribet, G. (2014) A revised dated phylogeny of the arachnid order Opiliones. *Front. Genet.* 5, 255
34. Winterton, S.L. and Ware, J.L. (2015) Phylogeny, divergence times and biogeography of window flies (Scenopinidae) and the therevoid clade (Diptera: Asiloidea). *Syst. Entomol.* 40, 491–519
35. Larson-Johnson, K. (2015) Phylogenetic investigation of the complex evolutionary history of dispersal mode and diversification rates across living and fossil Fagales. *New phytol.* Published online July 21, 2015. http://dx.doi.org/10.1111/nph.13570
36. Lee, M.S.Y. *et al.* (2014) Morphological clocks in paleontology, and a Mid-Cretaceous origin of crown aves. *Syst. Biol.* 63, 442–449
37. Alekseyenko, A.V. *et al.* (2008) Wagner and Dollo: a stochastic duet by composing two parsimonious solos. *Syst. Biol.* 57, 772–784
38. Bouckaert, R. *et al.* (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* 10, e1003537
39. Swofford, D.L. (1998) *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods) Version 4,* Sinauer Associates
40. Felsenstein, J. (1989) PHYLIP – Phylogeny Inference Package (Version 3.2). *Cladistics* 5, 164–166
41. Ronquist, F. *et al.* (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542
42. Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313
43. Lewis, P.O. (2001) A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* 50, 913–925
44. Jukes, T.H. and Cantor, C.R. (1969) Evolution of protein molecules. In *Mammalian Protein Metabolism* (Munro, H.N., ed.), pp. 21–132, Academic Press
45. Felsenst, J. (1973) Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am. J. Hum. Genet.* 25, 471–492
46. Felsenstein, J. (2012) A comparative method for both discrete and continuous characters using the threshold model. *Am. Nat.* 179, 145–156
47. Felsenstein, J. (2005) Using the quantitative genetic threshold model for inferences between and within species. *Philos. Trans. R. Soc. B: Biol. Sci.* 360, 1427–1434
48. Wiens, J.J. and Morrill, M.C. (2011) Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. *Syst. Biol.* 60, 719–731
49. Wiens, J. and Moen, D. (2008) Missing data and the accuracy of Bayesian phylogenetics. *J. Syst. Evol.* 46, 307–314
50. Wiens, J.J. and Tiu, J. (2012) Highly incomplete taxa can rescue phylogenetic analyses from the negative impacts of limited taxon sampling. *PLoS ONE* 7, e42925
51. Lemmon, A.R. *et al.* (2009) The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. *Syst. Biol.* 58, 130–145
52. Simmons, M.P. (2011) Misleading results of likelihood-based phylogenetic analyses in the presence of missing data. *Cladistics* 28, 208–222
53. Sansom, R.S. and Wills, M.A. (2013) Fossilization causes organisms to appear erroneously primitive by distorting evolutionary trees. *Sci. Rep.* 3, 5
54. Sansom, R.S. *et al.* (2010) Non-random decay of chordate characters causes bias in fossil interpretation. *Nature* 463, 797–800
55. Reisz, R.R. and Muller, J. (2004) Molecular timescales and the fossil record: a paleontological perspective. *Trends Genet.* 20, 237–241
56. Lee, M.S.Y. *et al.* (2014) Sustained miniaturization and anatomical innovation in the dinosaurian ancestors of birds. *Science* 345, 562–566
57. Donoghue, P.C.J. and Smith, M.P., eds (2003) *Telling the Evolutionary Time: Molecular Clocks and the Fossil Record*, CRC Press
58. Smith, A.B. (2000) Stratigraphy in phylogeny reconstruction. *J. Paleontol.* 74, 763–766
59. Alroy, J. (2002) Stratigraphy in phylogeny reconstruction – reply to Smith (2000). *J. Paleontol.* 76, 587–589
60. Wagner, P.J. (2002) Testing phylogenetic hypotheses with stratigraphy and morphology – a comment on Smith (2000). *J. Paleontol.* 76, 590–593
61. Fisher, D.C. *et al.* (2002) Stratigraphy in phylogeny reconstruction – comment on Smith (2000). *J. Paleontol.* 76, 585–586
62. Sumrall, C.A. and Brochu, C.A. (2003) Resolution, sampling, higher taxa and assumptions in stratocladistic analysis. *J. Paleontol.* 77, 189–194
63. Wickström, L.M. and Donoghue, P.C.J. (2005) Cladograms, phylogenies and the veracity of the conodont fossil record. *Special Papers Palaeontol.* 73, 185–218
64. Benton, M.J. *et al.* (2000) Quality of the fossil record through time. *Nature* 403, 534–537
65. Fortey, R.A. and Jefferies, R.P.S. (1982) Fossils and phylogeny – a compromise approach. In *Problems of Phylogenetic Reconstruction. Systematics Association Special Volume 21* (Joysey, K.A. and Friday, A.E., eds), pp. 197–234, Academic Press
66. Holland, S.M. (2000) The quality of the fossil record: a sequence stratigraphic perspective. *Paleobiology* 26 (Suppl.), 148–168
67. Behrensmeyer, A.K. *et al.* (2000) Taphonomy and paleobiology. *Paleobiology* 26, 103–147
68. Ho, S.Y. and Lanfear, R. (2010) Improved characterisation of among-lineage rate variation in cetacean mitogenomes using codon-partitioned relaxed clocks. *Mitochondrial DNA* 21, 138–146
69. Yang, Z. (1996) Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.* 42, 587–596
70. Nylander, J.A. *et al.* (2004) Bayesian phylogenetic analysis of combined data. *Syst. Biol.* 53, 47–67

71. Duchene, S. and Ho, S.Y. (2014) Using multiple relaxed-clock models to estimate evolutionary timescales from DNA sequence data. *Mol. Phylogenet. Evol.* 77, 65–70
72. Thornhill, A.H. *et al.* (2012) Are pollen fossils useful for calibrating relaxed molecular clock dating of phylogenies? A comparative study using Myrtaceae. *Mol. Phylogenet. Evol.* 63, 15–27
73. Bromham, L. *et al.* (2002) Testing the relationship between morphological and molecular rates of change along phylogenies. *Evolution* 56, 1921–1930
74. Seligmann, H. (2010) Positive correlations between molecular and morphological rates of evolution. *J. Theor. Biol.* 264, 799–807
75. Davies, T.J. and Savolainen, V. (2006) Neutral theory, phylogenies, and the relationship between phenotypic change and evolutionary rates. *Evolution* 60, 476–483
76. Kimura, M. (1983) *The Neutral Theory of Molecular Evolution,* Cambridge University Press
77. Haldane, J.B.S. (1949) Suggestions as to quantitative measurement of rates of evolution. *Evolution* 3, 51–56
78. Gillespie, J.H. (1991) *The Causes of Molecular Evolution,* Oxford University Press
79. Lee, M.S.Y. *et al.* (2013) Rates of phenotypic and genomic evolution during the cambrian explosion. *Curr. Biol.* 23, 1889–1895
80. Beck, R.M.D. and Lee, M.S.Y. (2014) Ancient dates or accelerated rates? Morphological clocks and the antiquity of placental mammals. *Proc. R. Soc. B: Biol. Sci.* 281, 10
81. Marshall, C.R. (1990) Confidence-intervals on stratigraphic ranges. *Paleobiology* 16, 1–10
82. Benton, M.J. *et al.* (2009) Calibrating and constraining molecular clocks. In *The Timetree of Life* (Hedges, S.B. and Kumar, S., eds), pp. 35–86, Oxford University Press
83. Warnock, R.C.M. *et al.* (2015) Calibration uncertainty in molecular dating analyses: there is no substitute for the prior evaluation of time priors. *Proc. R. Soc. B: Biol. Sci.* 282
84. Warnock, R.C. *et al.* (2012) Exploring uncertainty in the calibration of the molecular clock. *Biol. Lett.* 8, 156–159
85. Inoue, J. *et al.* (2010) The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. *Syst. Biol.* 59, 74–89
86. Zhang, J. and Rasnitsyn, A. (2006) New extinct taxa of Pelecinidae sensu lato (Hymenoptera:Proctotrupidea) in the Laiyang Formation, Shadong, China. *Cretaceous Res.* 27, 684–688
87. Hu, C. *et al.* (2001) *Shantungosaurus giganteus,* Beijing Geological Publishing House (in Chinese)
88. Chen, P. *et al.* (2005) Jianshangou Bed of the Yixian Formation in West Liaoning, China. *Sci. China Series D: Earth Sci.* 48, 298–312
89. Chen, P. *et al.* (2006) Geological ages of track bearing formations in China. *Cretaceous Res.* 27, 22–32
90. Zhou, Z. *et al.* (2003) An exceptionally preserved Lower Cretaceous ecosystem. *Nature* 421, 807–814
91. Zhoue, Z. (2006) Evolutionary radiation of the Jehol Biota: chronological and ecological perspectives. *Geological J.* 41, 377–393
92. Wang, S. *et al.* (2001) Further discussion on geologic age of Sihetun vertebrate assemblage in western Liaoning China: evidence from Ar–Ar dating. *Petrelog. Sinica* 17, 663–668
93. Ling, W. *et al.* (2007) Zircon U–Pb dating on the Mesozoic volcanic suite from the Qingshan Group stratotype section in eastern Shandong Province and its tectonic significance. *Sci. China Series D: Earth Sci.* 50, 813–824
94. Sadler, P.M. (2004) Quantitative biostratigraphy – achieving finer resolution in global correlation. *Annu. Rev. Earth Planet. Sci.* 32, 187–213
95. Graur, D. and Martin, W. (2004) Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *Trends Genet.* 20, 80–86
96. Ho, S.Y. (2009) An examination of phylogenetic models of substitution rate variation among lineages. *Biol. Lett.* 5, 421–424
97. Marshall, C.R. (1997) Confidence intervals on stratigraphic ranges with nonrandom distributions of fossil horizons. *Paleobiology* 23, 165–173
98. Butlin, R. *et al.* (2012) What do we need to know about speciation? *Trends Ecol. Evol.* 27, 27–39
99. Maynard Smith, J. *et al.* (1985) Developmental constraints and evolution. *Q. Rev. Biol.* 60, 265–287
100. Davis, C.C. *et al.* (2014) Long-term morphological stasis maintained by a plant-pollinator mutualism. *Proc. Natl. Acad. Sci. U.S. A.* 111, 5914–5919

BIOLOGY LETTERS

rsbl.royalsocietypublishing.org

Research

Cite this article: O'Reilly JE, Puttick MN, Parry L, Tanner AR, Tarver JE, Fleming J, Pisani D, Donoghue PCJ. 2016 Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. *Biol. Lett.* **12**: 20160081.
http://dx.doi.org/10.1098/rsbl.2016.0081

Received: 28 January 2016
Accepted: 21 March 2016

Subject Areas:
palaeontology, taxonomy and systematics, evolution

Keywords:
parsimony, Bayesian, likelihood, phylogenetics, morphology

Authors for correspondence:
Davide Pisani
e-mail: davide.pisani@bristol.ac.uk
Philip C. J. Donoghue
e-mail: phil.donoghue@bristol.ac.uk

†These authors contributed equally to this study.

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsbl.2016.0081 or via http://rsbl.royalsocietypublishing.org.

THE ROYAL SOCIETY PUBLISHING

Palaeontology

Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data

Joseph E. O'Reilly[1,†], Mark N. Puttick[1,†], Luke Parry[1], Alastair R. Tanner[1,2], James E. Tarver[1], James Fleming[1], Davide Pisani[1,2] and Philip C. J. Donoghue[1]

[1]School of Earth Sciences, and [2]School of Biological Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol BS8 1TQ, UK

JEO, 0000-0001-9775-253X; DP, 0000-0003-0949-6682; PCJD, 0000-0003-3116-7463

Different analytical methods can yield competing interpretations of evolutionary history and, currently, there is no definitive method for phylogenetic reconstruction using morphological data. Parsimony has been the primary method for analysing morphological data, but there has been a resurgence of interest in the likelihood-based Mk-model. Here, we test the performance of the Bayesian implementation of the Mk-model relative to both equal and implied-weight implementations of parsimony. Using simulated morphological data, we demonstrate that the Mk-model outperforms equal-weights parsimony in terms of topological accuracy, and implied-weights performs the most poorly. However, the Mk-model produces phylogenies that have less resolution than parsimony methods. This difference in the accuracy and precision of parsimony and Bayesian approaches to topology estimation needs to be considered when selecting a method for phylogeny reconstruction.

1. Introduction

Morphology once provided the only means of inferring evolutionary trees, but it was effectively rendered obsolete by molecular sequence data and the development of sophisticated molecular evolutionary models for phylogenetic analysis [1]. However, with the recognition that fossil species are integral to correctly inferring patterns of character evolution and changes in diversity, as well as in establishing evolutionary timescales, morphological data are enjoying a phylogenetic renaissance [2], allowing fossil species to be assigned to their correct branches in the Tree of Life. Methods for phylogenetic analysis of morphological data remain underdeveloped and though likelihood models are available that may more accurately accommodate the vagaries of morphological datasets [3], including high rates of heterogeneity and a preponderance of missing data [4], parsimony remains the method of choice, principally perhaps as a consequence of tradition. Indeed, a recent simulation-based study by Wright & Hillis [5] demonstrated that a Bayesian implementation of Lewis's Mk-model [3] strongly outperforms parsimony, especially when rates of character change are high, or when relatively few characters are analysed. The conclusions drawn by Wright & Hillis [5] were based on data effectively simulated using the Mk-model, potentially biasing the test in favour of the Mk-model. Furthermore, they did not consider whether the simulated data exhibited realistic levels of homoplasy, analysed unrealistically large simulated datasets, and evaluated only the relative performance of

Table 1. The differences in median and the 95th percentile range of Robinson–Foulds values between the Mk and both parsimony models are greater in the full dataset compared with the realistic homoplasy subsets. mk, Bayesian Mk model; ew, equal-weights parsimony; iw, implied weights parsimony and its attendant *K* values.

	100 characters	100 characters CI	350 characters	350 characters CI	1000 characters	1000 characters CI
mk	45 (29–64)	40.5 (28.2–62.5)	20 (10–51)	19.5 (10.2–57.3)	19.5 (10.2–57.3)	11 (5–27.8)
ew	61 (31–98)	53 (29–91.8)	27 (12–70)	28 (12–74.8)	28 (12–74.8)	16 (6.2–43.7)
iw k2	89 (39–119)	77 (38.2–117.7)	36 (18–76)	36 (17.2–81.3)	36 (17.2–81.3)	19.5 (10–35.7)
iw k3	76 (38–112)	69 (36.4–108)	32 (16–69)	34 (15.2–70)	34 (15.2–70)	18 (9.2–35.7)
iw k5	68 (36–104)	61 (32.2–102)	30 (14–66)	31.5 (15.2–68)	31.5 (15.2–68)	18 (9–34)
iw k10	63 (34–100)	55.5 (32–98)	28 (13–68)	30 (15.2–69.7)	30 (15.2–69.7)	16 (8–34)
iw k20	64 (34–100)	53 (33–97.8)	28 (14–68)	30 (13.2–71.7)	30 (13.2–71.7)	17 (8–39.3)
iw k200	65 (34–100)	55 (32.2–97.7)	28 (14–72)	30.5 (15–76)	30.5 (15–76)	18 (8–44)

equal-weights parsimony when morphological data are now commonly analysed under implied-weights parsimony [6].

In an attempt to evaluate the relative performance of likelihood and parsimony methods for the phylogenetic analysis of discrete character morphological data, we simulated datasets of 100, 350 and 1000 discrete morphological characters using a modified HKY85 model, discriminating datasets that failed to meet expected levels of homoplasy. We evaluated the relative performance of equal-weights parsimony, implied-weights parsimony and model-based methods of phylogenetic analysis in terms of their ability to recover the tree used to simulate the data. We found that the Mk-model performs best in the analysis of all simulated datasets, largely because the Bayesian consensus trees are poorly resolved. Equal-weights parsimony exhibits lower levels of accuracy but this is combined with higher resolution. Implied-weights parsimony performed most poorly of all the methods considered.

2. Material and methods

To simulate binary morphological data, we used the HKY + $\Gamma_{\text{continuous}}$ model to generate nucleotide data which we translated into purines (0) and pyrimidines (1)—R/Y coding. The recoded HKY-model possesses an uneven equilibrium distribution of state frequencies, resulting in structurally realistic morphological matrices while facilitating violation of assumptions of the Mk-model; thus, our data are not biased in favour of either method of phylogenetic inference. Initial tests were performed to determine values for the model parameters which produce binary data with empirically observed levels of homoplasy [7]. Following [5], data were simulated using the lissamphibian tree presented in [8], yielding datasets of 100, 350 and 1000 characters; most real morphological datasets contain in the order of 100 characters, but we included 350 and 1000 character matrices to investigate the effect of scaling and for ease of comparison to [5]. In total, 100 unique underlying substitution rates were drawn from a U(0.1,10) distribution, facilitating rates spanning two orders of magnitude. For each substitution rate, 10 unique matrices were produced, modelling among-character rate heterogeneity as gamma distributed uniquely within each matrix.

Matrices were analysed with the Mk + Γ model using default priors in MrBayes v. 3.2 [9], and both standard and implied-weights parsimony in TNT [10]. The Mk-model is more suitable for our simulated data than the Mkv-model as we did not strip invariant sites from the final matrices. Majority-rule consensus trees were produced for each method. For implied-weights parsimony, we used a range of *K*-values: 2, 3, 5, 10, 20 and 200. As the underlying substitution rate is varied, the per-matrix level of homoplasy may violate the empirically observed range; to produce the most empirically justified morphological matrices, we implemented an empirically derived minimum consistency index (CI) cut-off of 0.26 [7] for each simulated dataset and repeated analyses for these treated matrices (electronic supplementary material, figure S1). This cut-off reduced the size of the datasets to 128 (100 characters), 149 (350 characters) and 126 (1000 characters) matrices. In-depth description of the initial parameter value tests and further details of matrix generation are presented in the electronic supplementary material.

The accuracy of topologies estimated by the different reconstruction techniques was assessed using the Robinson–Foulds distance [11] from the generator tree. We also explored the relationship between resolution of output trees, measured by the number of nodes per tree.

3. Results

The Mk-model achieved the highest levels of accuracy across all datasets. Median Robinson–Foulds distances are lower for the Mk-model compared with both equal-weights and implied-weights parsimony (table 1 and figure 1), and for all approaches, accuracy of topology reconstruction increases with increasing dataset size. Furthermore, equal-weights parsimony out-performs implied-weights parsimony for all datasets and values of *K*, but this is less pronounced for the 1000 character dataset (table 1). For convenience, all further results for implied weights are for $K = 2$.

The same relative performance of the phylogenetic reconstruction methods is seen when considering only those datasets exhibiting realistic levels of homoplasy. The median Robinson–Foulds distance for the Mk-model is still lowest for each dataset, but the median and range of Robinson–Foulds distances for equal and implied-weights parsimony are closer to the distribution seen from the Mk-model (table 1 and figure 1). Additionally, for a given dataset, there is a similar Robinson–Foulds distance regardless of the reconstruction method employed (electronic supplementary material, figure S2). Unless otherwise stated, all subsequent results are from the subset of datasets exhibiting realistic levels of homoplasy.

The higher accuracy (lower Robinson–Foulds values) of the Mk-model against other methods for 100 and 350 characters is due to trees being less resolved (figure 2). The density of Robinson–Foulds distance is lower for the Mk compared with equal weights, which itself is lower than implied weights, but both equal and implied weights achieve higher levels of

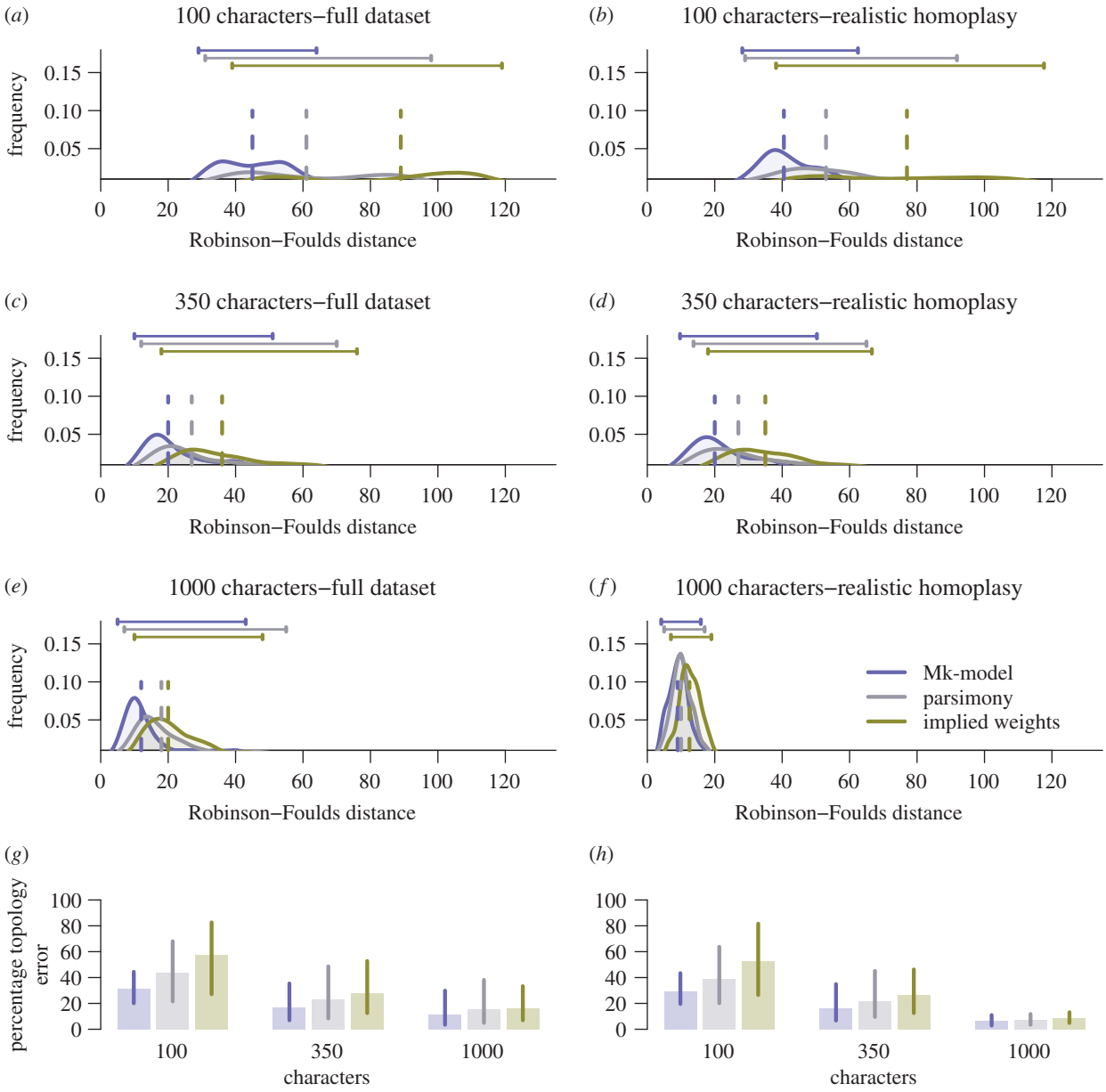


Figure 1. Mk tree reconstructions (blue) outperform equal-weights parsimony (grey) and implied-weights parsimony (green) for 100, 350 and 1000 characters (*a*,*c*,*e*,*g*), and these differences remain in the subset of the simulated data matrices that exhibit realistic levels of homoplasy (*b*,*d*,*f*,*h*). Bars above the plots mark the 95th percentile range for each method, and dashed vertical lines show the median values. Percentage topology error (*g*,*h*) is the Robinson–Foulds value of the reconstructed tree compared with the worst possible value, as shown in [5].

precision (number of nodes reconstructed). These differences are negligible in the 1000 character datasets (figure 2).

There is a significant overlap in the set of nodes correctly recovered across methods, when mapped against the reference phylogeny (figure 2; electronic supplementary material, figure S3). In particular, for all methods there is a trend for nodes closer to the root to be more accurately estimated in small datasets, but this relationship decreases as the number of characters increases (electronic supplementary material, table S2 and figures S2, S4, S5). The percentage of times a node from the reference tree was accurately reconstructed showed a strong correlation for 100 and 350 characters, but decreases with 1000 characters (electronic supplementary material, table S2).

4. Discussion

Only minor differences are seen in the accuracy of phylogenetic topology reconstruction between the Bayesian implementation of the Mk-model and parsimony methods. Our findings both support and contradict elements of the results of Wright & Hillis [5] in that we can corroborate their observation, that the Mk-model outperforms equal-weights parsimony in accuracy, but the Mk-model achieves this at the expense of precision. Unexpectedly, implied-weights parsimony is less effective than either equal-weights parsimony or the Mk-model, in datasets with small numbers of characters. Implied-weights parsimony outperforms equal-weights parsimony only in the analyses of unrealistically large datasets. These results challenge the increasingly common view that implied-weighting better accommodates homoplasy than does equal-weights parsimony [6], and this result is true for a range of *K*-values (table 1).

In comparison with the other approaches, equal-weights parsimony analyses of the datasets exhibiting realistic levels of homoplasy and large number of characters yield a set of trees with a longer tailed distribution of Robinson–Foulds distances. In large part, this reflects estimation of a small quantity of trees markedly different from the generating tree (figure 1).

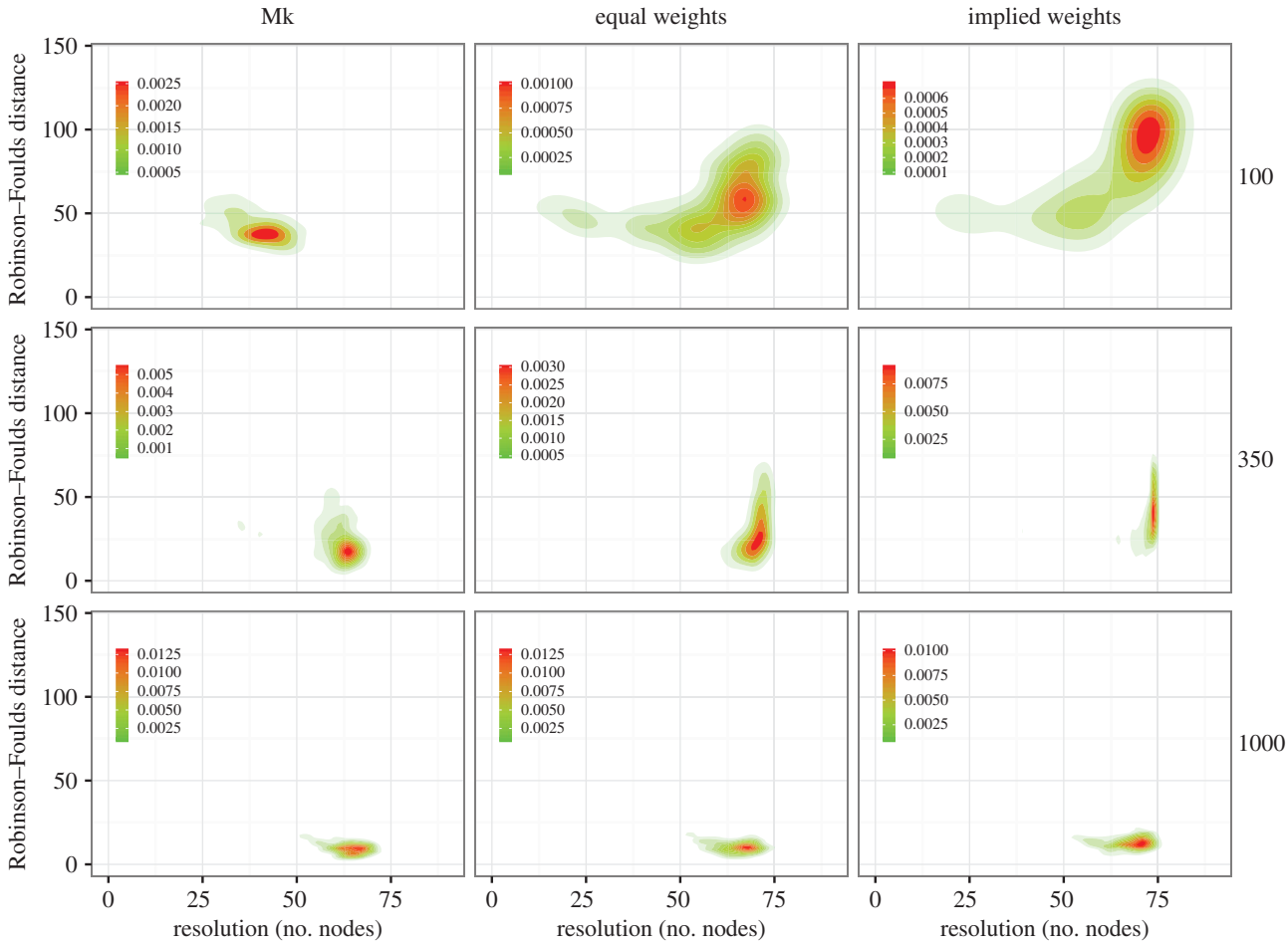


Figure 2. The Mk model exhibits higher accuracy with lower precision than parsimony methods; these results are less clear as more characters are added. Contour plots of Robinson–Foulds distances against the number of resolved nodes in each tree; the contours represent the density of the distribution of trees.

Inaccuracy in topological estimation is more prevalent towards the tips in all analyses, with the inclusion of more characters reducing the intensity of this phenomenon. For this effect to be completely removed, it would require the analysis of well over 1000 empirically justifiable characters, a number that is rarely achieved for morphological datasets. The accuracy of node reconstruction is correlated significantly between all three techniques, demonstrating that most nodes in the tree that were difficult to resolve for one method were difficult to resolve for all. This phenomenon is observed across all character quantities and suggests a general difficulty in accurately estimating topology given the same data.

Our results can be interpreted to advocate use of the Mk-model over parsimony methods in the analysis of discrete morphological data. Parsimony methods produce precision without the accuracy achieved by the Mk-model and precision without accuracy is a poor basis for any science. We anticipate that the implementation of the Mk-model within a maximum-likelihood framework will exhibit levels of accuracy and precision more comparable to the parsimony methods, simply because it estimates a single, fully resolved topology. Integration over parameters while producing an acceptable level of accuracy is a quality of Bayesian inference, and our Mk-model results are probably dependent on a Bayesian implementation. While comparative phylogenetic methods often require fully resolved trees, these may be accommodated through analyses using the posterior sample of trees estimated using the Mk-model. Therefore, the prior requirement of a fully resolved tree need not necessarily lead to a preference for parsimony over the Mk-model.

In comparison to parsimony methods, the Mk-model has undergone little development since its conception [12,13], while attempts to improve the performance of parsimony methods, like implied-weights parsimony [3], have not led to increased accuracy (table 1). Thus, model-based phylogenetics can be expected to offer more opportunity for development, e.g. through relaxing the assumption of symmetrically distributed stationary distribution of character states [12,13] and improvement in the accuracy of phylogeny estimation from discrete character data. We suggest, however, that more focus should be invested in assessing whether the data are sufficiently informative to discriminate between competing phylogenetic hypotheses.

5. Conclusion

Phylogenies produced using likelihood models are more accurate than parsimony approaches, but have lower precision. Likelihood models offer greater scope for development in attempting to achieve greater accuracy but, in the interim, we suggest that phylogeneticists should consider the aims of their analyses when choosing the appropriate method.

Data accessibility. Data used in this manuscript are archived at http://dx.doi.org/10.5061/dryad.10qf3 and as the electronic supplementary material.

Authors' contributions. All authors made substantial contributions to (i) conception and design, or acquisition of data, or analysis and interpretation of data and (ii) drafting the article or revising it critically for important intellectual content; gave final approval of the version to be published, and agreed to be accountable for all aspects

of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Competing interests. We have no competing interests.

Funding. This research was funded by NERC (NE/L501554/1 to J.E.O'R. and L.P.; NE/K500823/1 to M.N.P.; NE/L002434/1 to J.F.; NE/N003438/1 to P.C.J.D.), BBSRC (BB/N000919/1 to P.C.J.D.), the University of Bristol (STaR scholarship to A.R.T.), Royal Society Wolfson Research Merit Award (P.C.J.D.) and the John Templeton Foundation (43915 to D.P.).

Acknowledgements. We thank the other members of the Bristol Palaeobiology research group for discussion, as well as April Wright and Peter Wagner for the constructive reviews that helped shape the published version of our study.

References

1. Scotland RW, Olmstead RG, Bennett JR. 2003 Phylogeny reconstruction: the role of morphology. *Syst. Biol.* **52**, 539–548. (doi:10.1080/10635150390223613)
2. Lee MSY, Palci A. 2015 Morphological phylogenetics in the genomic age. *Curr. Biol.* **25**, R922–R929. (doi:10.1016/j.cub.2015.07.009)
3. Lewis PO. 2001 A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* **50**, 913–925. (doi:10.1080/106351501753462876)
4. Wagner PJ. 2012 Modelling rate distributions using character compatibility: implications for morphological evolution among fossil invertebrates. *Biol. Lett.* **8**, 143–146. (doi:10.1098/rsbl.2011.0523)
5. Wright AM, Hillis DM. 2014 Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. *PLoS ONE* **9**, e109210. (doi:10.1371/journal.pone.0109210)
6. Goloboff PA, Carpenter JM, Arias JS, Rafael D, Esquivel M. 2008 Weighting against homoplasy improves phylogenetic analysis of morphological data sets. *Cladistics* **24**, 758–773. (doi:10.1111/j.1096-0031.2008.00209.x)
7. Sanderson MJ, Donoghue MJ. 1989 Patterns of variation in levels of homoplasy. *Evolution* **43**, 1781–1795. (doi:10.2307/2409392)
8. Pyron RA. 2011 Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. *Syst. Biol.* **60**, 466–481. (doi:10.1093/sysbio/syr047)
9. Ronquist F *et al.* 2012 MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542. (doi:10.1093/sysbio/sys029)
10. Goloboff P, Farris S, Nixon K. 2000 *TNT* (*Tree analysis using New Technology*). Tucumán, Argentina: published by the authors.
11. Robinson DR, Foulds LR. 1981 Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147. (doi:10.1016/0025-5564(81)90043-2)
12. Klopfstein S, Vilhelmsen L, Ronquist F. 2015 A nonstationary Markov model detects directional evolution in hymenopteran morphology. *Syst. Biol.* **64**, 1089–1103. (doi:10.1093/sysbio/syv052)
13. Wright AM, Lloyd GT, Hillis D. 2015 Modeling character change heterogeneity in phylogenetic analyses of morphology through the use of priors. *Syst. Biol.* syv122. (doi:10.1093/sysbio/syv122)

BIOLOGY LETTERS

rsbl.royalsocietypublishing.org

Research

Cite this article: O'Reilly JE, Donoghue PCJ. 2016 Tips and nodes are complementary not competing approaches to the calibration of molecular clocks. *Biol. Lett.* **12**: 20150975. http://dx.doi.org/10.1098/rsbl.2015.0975

Received: 20 November 2015
Accepted: 24 March 2016

Subject Areas:
palaeontology, taxonomy and systematics, evolution

Keywords:
molecular clock, calibration, tip, node, Hymenoptera

Authors for correspondence:
Joseph E. O'Reilly
e-mail: joe.oreilly@bristol.ac.uk
Philip C. J. Donoghue
e-mail: phil.donoghue@bristol.ac.uk

An invited contribution to the special feature 'Putting fossils in trees: combining morphology, time and molecules to estimate phylogenies and divergence times'.

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsbl.2015.0975 or via http://rsbl.royalsocietypublishing.org.

THE ROYAL SOCIETY PUBLISHING

Palaeontology

Tips and nodes are complementary not competing approaches to the calibration of molecular clocks

Joseph E. O'Reilly and Philip C. J. Donoghue

School of Earth Sciences, University of Bristol, Life Sciences Building, Bristol BS8 1TQ, UK

JEO, 0000-0001-9775-253X; PCJD, 0000-0003-3116-7463

Molecular clock methodology provides the best means of establishing evolutionary timescales, the accuracy and precision of which remain reliant on calibration, traditionally based on fossil constraints on clade (node) ages. Tip calibration has been developed to obviate undesirable aspects of node calibration, including the need for maximum age constraints that are invariably very difficult to justify. Instead, tip calibration incorporates fossil species as dated tips alongside living relatives, potentially improving the accuracy and precision of divergence time estimates. We demonstrate that tip calibration yields node calibrations that violate fossil evidence, contributing to unjustifiably young and ancient age estimates, less precise and (presumably) accurate than conventional node calibration. However, we go on to show that node and tip calibrations are complementary, producing meaningful age estimates, with node minima enforcing realistic ages and fossil tips interacting with node calibrations to objectively define maximum age constraints on clade ages. Together, tip and node calibrations may yield evolutionary timescales that are better justified, more precise and accurate than either calibration strategy can achieve alone.

1. Introduction

The molecular clock has displaced the fossil record as the primary means of establishing an evolutionary timescale; however, the accuracy and precision of divergence time estimates and their fossil calibrations remain inextricably linked [1]. Traditionally, divergence time estimation has achieved calibration based on geological (usually palaeontological) constraints on clade (node) ages. This approach has been developed to the extent that further improvements in accuracy and precision are limited by the inherent uncertainty in fossil evidence. Indeed, it is this uncertainty that has called into question the approach of node calibration, particularly what some see as the over-interpretation of palaeontological data to establish maximum constraints on clade ages, and the difficulty in objectively representing prior evidence of node age as a probability distribution [2]. Furthermore, node age constraints invariably differ from those specified as a consequence of their integration into the joint time prior on node ages [3]. These concerns have led to the replacement of node calibrations with tip calibrations in which fossil species of a known age are integrated directly into divergence time analyses, supplementing sequence data from living species with morphological data from living and fossil species [2,4]. However, there has been little effort to demonstrate the effect of different approaches to calibration and, indeed, to determine whether the effective prior on node ages resulting from tip calibration is compatible with the fossil evidence usually employed in node calibration. This is of particular

interest given growing concern that tip calibration consistently yields unrealistically ancient divergence time estimates [5].

Hence, we sought to compare the efficacy of tip and node calibrations by determining the compatibility of the resulting effective prior on node ages resulting from tip calibration and fossil-based node age constraints. This is readily sampled in node- and tip-calibrated analyses when the time prior is conditioned on a fully constrained topology upon which ages are estimated. However, it is challenging where topology and time are coestimated. Here, we show that, in such circumstances, an approximation of the time prior can be obtained by conditioning on the consensus tree derived from a posterior sample of trees. Using an empirical dataset, we show that effective node age priors derived from tip calibration are often incompatible with fossil evidence, violating either minimum or maximum node age constraints. We argue that this contributes to the unrealistically ancient divergence time estimates produced by tip calibration. These artefacts are diminished by combining tip and node calibrations, where node calibrations ensure that divergence time estimates never violate fossil-based minima and tip calibrations effectively establish node age maxima.

2. Material and methods

We compared the effective node age priors and posteriors for tip and node calibrations using a previously published hymenopteran dataset of molecular and morphological characters [2]. The original study assumed errorless tip-ages for fossil species. We employed revised ages for these species, integrating associated uncertainty and derived node age constraints in order to compare effective priors on node ages to the palaeontological evidence [5]. Uncertainty in fossil taxon age was represented with uniform distributions, whereas node calibrations were assigned offset exponential distributions, as in [2]. Unbounded distributions allow maxima to be defined by interaction between node and tip calibrations.

To obtain an approximation of the time prior, we sampled from the prior while conditioning on the consensus of a sample from the posterior distribution of trees obtained from a standard tip-calibrated total evidence dating (TED) analysis. We then constrained the topology to the consensus tree and sampled from the prior conditioned on this tree, providing a meaningful approximation of the effective time prior in a topologically unconstrained tip-calibrated analysis (electronic supplementary material methods).

To evaluate the influence of tip calibrations, we compared effective priors and posterior estimates of node ages from tip-calibrated analysis to the raw palaeontological constraints on node ages, and to the effective priors and posterior estimates of node ages derived from (i) a node-calibrated analysis and (ii) an analysis that implemented both tip and node calibrations. In the latter, fossil taxa were assigned to clades identified in the standard tip-calibrated analysis; where possible, the clades were assigned node calibrations. Minima on node-calibrated clades are defined by fossil evidence and maxima are established based on interaction between node and tip calibrations. We obtained a posterior sample of trees using the consensus tree produced from this sample to sample from the effective time prior. Several fossil taxa and node calibrations could not be included in this analysis because of limitations of MrBayes (see the electronic supplementary material for detail).

3. Results

Our tip-calibrated consensus topology (figure 1*a*) differs from [2] in the placement of fossil Xyelidae, which could not be resolved in our analysis. *Spathoxyela* and *Mesoxyela* form a polytomy with extant Xyelidae, because they are alternately assigned to crown or total-group Xyelidae in the tree sample; in the original analysis, all fossil Xyelidae were resolved to the stem in the consensus tree. Following [2], *Eoxyela*, the fossil defining the node calibration for Xyelidae, is resolved outside of crown Xyelidae. A number of fossil taxa, including *Palaeathalia* and *Cleistogaster*, were placed with higher resolution in our recalibrated analysis than in the original. Similar to [2], we were unable to recover unequivocal monophyly of Pamphilioidea.

The effective priors on node ages resulting from tip calibration alone (excepting the two deepest nodes) consistently extend beyond the maximum palaeontological constraints on node ages, and include more ancient ages than the effective priors on node ages in the node-calibrated analysis. In two clades (Xyelidae and Siricoidea), tip calibration produces effective priors extending to the near Recent. The effective time priors on these clades plus Pamphilioidea extend beyond the minimum palaeontological constraints on the ages of these crown clades, and encompass younger ages than the effective priors on node ages in the node-calibrated analysis (figure 1*b*). In all instances, these differences propagate to the posterior estimates of clade ages. The anticipated linear relationship between node age and highest posterior density (HPD) width holds only for the node-calibrated analysis (figure 2). The results of the tip-calibrated analysis exhibit an inverse relationship, with uncertainty decreasing with proximity to the root.

When tip and node calibrations are combined (figure 1*c*), the effective priors on node ages encompass dates younger than the minimum palaeontological constraints on the ages of crown Pamphilioidea and crown Xyelidae; in all other clades, the effective priors and posterior age estimates fall fully within their palaeontological node age constraints. In all but the two deepest nodes the means of posterior estimates of clade age are consistently and significantly younger than their counterparts when only tip calibrations are implemented. The distributions of posterior estimates of clade age are also more precise than their tip-calibrated counterparts in all but the two most basal clades.

4. Discussion

It has been accepted generally that, because user-specified node age priors are truncated in construction of the joint time prior, the effective prior should be assessed to determine whether it is consistent with the palaeontological constraints [3]. Our results indicate that this approach should be extended to tip calibration. Tip calibrations consistently yielded older effective priors on node ages and older divergence time estimates. This occurs principally because of an absence of constraints on the ages of internal nodes within the tree, normally provided by node calibrations, allowing uncertainty to propagate from the tips, constrained only by the prior on the root age, skewing the distribution of prior probability towards ancient ages. We cannot conclude that these estimates are inaccurate merely because they are incompatible with palaeontological maximum age constraints. However, the effective priors derived from tip calibration of some node ages are younger than their palaeontological minimum age constraints, which is unreasonable. This occurs because some crown clades (Xyelidae, Pamphilioidea) in the tree sample are often resolved without fossil members and so their minimum ages are bounded only by the Recent.

The node-calibrated analysis is compatible with the palaeontological constraints on clade ages, because they are

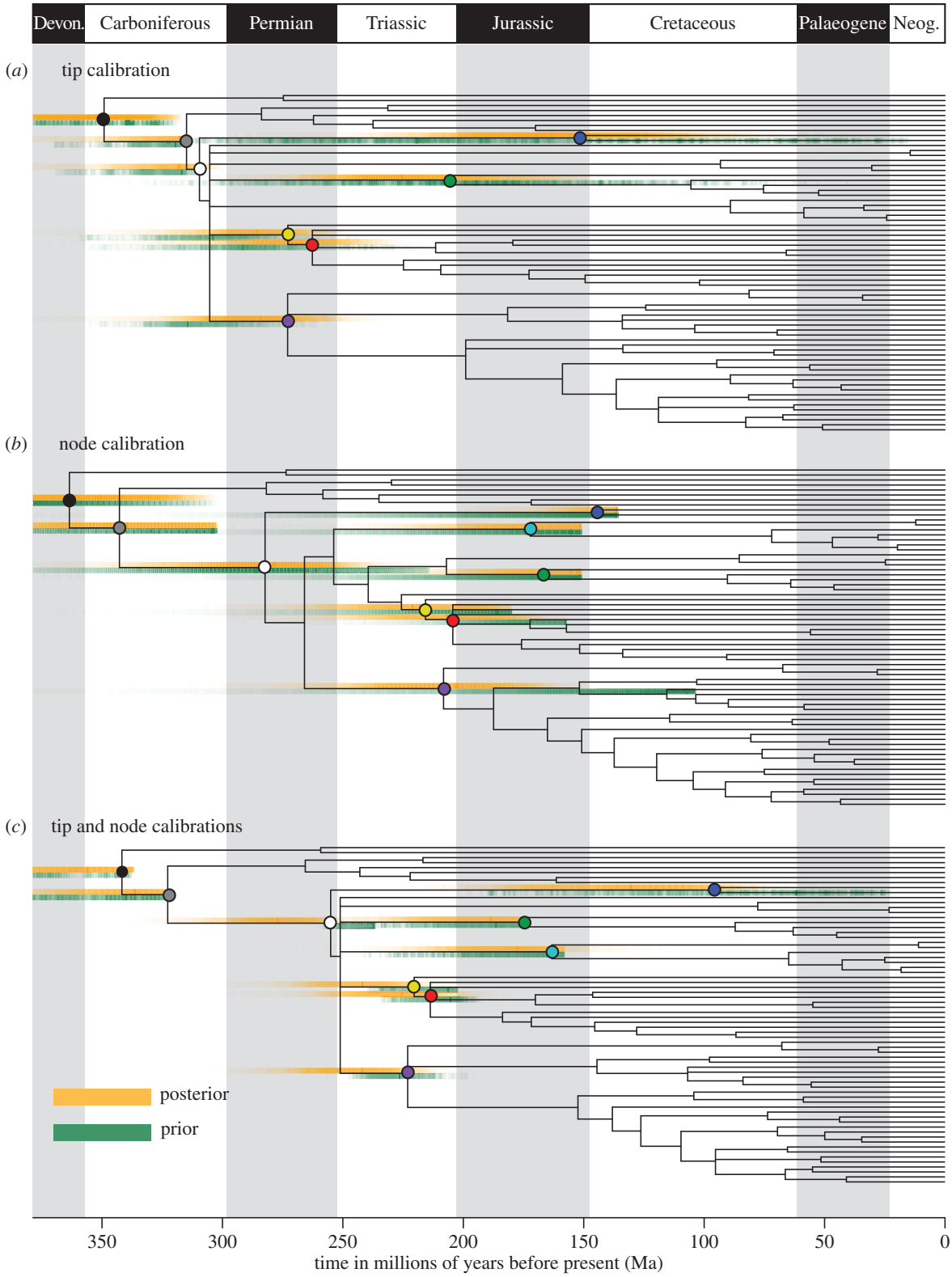
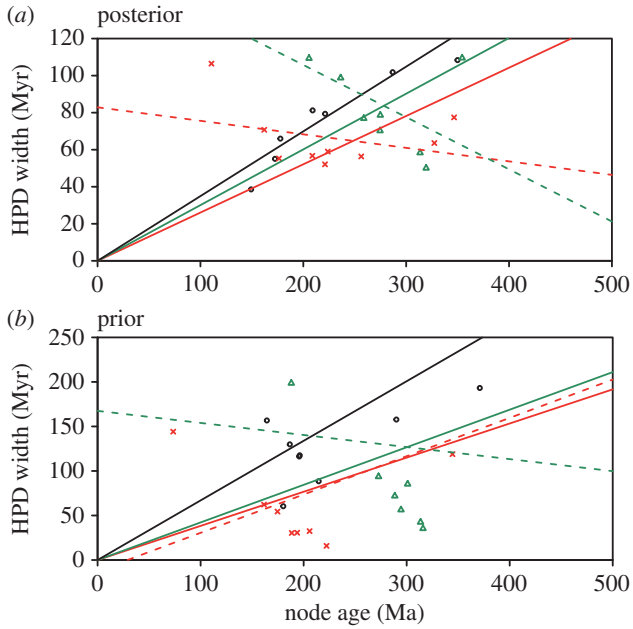


Figure 1. Time-calibrated phylogenies of Hymenoptera based on: (*a*) tip calibration, (*b*) node calibration and (*c*) combined tip and node calibrations. Panels (*a*,*c*) are presented with fossil taxa removed, complete topologies are presented in the electronic supplementary material. Graduated bars represent the prior and posterior distribution of clade age, with colour density correlated with probability. Polytomies reflect topological uncertainty in the tree sample and are not indicative of simultaneous divergence. Coloured nodes indicate the position of the nine clades of interest across the three topologies. Black (Neoptera), grey (Holometabola), white (Hymenoptera), yellow (Vespina), red (Apocrita), purple (Tenthredinoidea), blue (Xyelidae), turquoise (Pamphilioidea) and green (Siricoidea).

implemented as node calibrations. However, the combined node and tip-calibrated analyses yielded younger effective priors and posteriors than exclusively tip- or node-calibrated analyses, while also conforming to the palaeontological minimum constraints. This is clear in the case of Siricoidea, where no fossil member of the crown clade is represented but the zero-time constraint on the age of this clade in the tip-calibrated analysis is supplemented by a node age constraint in the combined tip- and node-calibrated analyses. The divergence time estimates derived from combined calibration are consistently younger—a consequence of the tip calibrations which act to truncate the broad priors of the node calibrations, extending from their hard minimum age constraints. This serves to draw the effective prior probability



(c) R^2 values for linear models

	node calibration$_O$	tip calibration$_O$	tip calibration	combined calibration$_O$	combined calibration
posterior	0.987	0.801	0.858	0.817	0.105
prior	0.924	0.574	0.009	0.703	0.334

Figure 2. Infinite-sites plots [1] for three alternative calibration approaches for both the posterior (*a*) and prior (*b*) distribution of times of nine clades for which node calibrations could be applied. Solid lines represent the fitted linear model for each independent set of node ages when forced through the origin, as in [1]. Dotted lines represent the linear model for non-node-calibrated analyses when not forced through the origin, demonstrating the lack of a linear decrease in clade age confidence interval width. The fit of linear models is presented in (*c*). Models forced through the origin are indicated with a subscript O.

closer to the minima in the joint time prior, which propagates to the posterior divergence time estimates. In effect, the tip and node calibrations interact to operationally establish maxima for the node calibrations.

It is reasonable to question whether tip and node calibrations should be implemented together and, certainly, the same data should not be represented in both calibration methods. However, there is no logical inconsistency between these approaches, and some fossil data are better represented as a tip calibration or as a node calibration. While it has been argued that tip calibration facilitates the inclusion of all fossil species in divergence time analyses [2,4], some fossil taxa are too incomplete to be effective tip calibrations, but may be no less definitive in circumscribing the minimum age of a clade (e.g. the minimum ages of angiosperms and echinoderms are constrained by tricolpate pollen and fragments of stereom, respectively).

A casualty of the implementation of node calibrations in MrBayes is the ability to perform coestimation of time and topology, a particular advantage of the tip calibration approach [2]. However, fossil taxa are not commonly well-resolved through coestimation, a consequence of the paucity of morphological data and the non-random distribution of missing data for fossil species [5]. These challenges may be overcome simply by introducing a backbone of partial topological constraints, facilitating coestimation, but within the qualified phylogenetic uncertainty that is associated with most fossil species. Only BEAST is currently capable of fully accommodating this approach to combined calibration [6]. In our combined tip- and node-calibrated analyses, we were forced to exclude any fossil species whose age overlapped or extended beyond the node calibration for the clade to which it was assigned. This limitation occurs because MrBayes unnecessarily considers ages for fossil species that can be older than their assigned clade, yielding a negative clock-rate and, therefore, an error when calculating the proposal ratio. Analyses employing the fossilized birth–death (FBD) model [7] integrate fossil occurrences as data in coestimating time and topology, constraining node ages and, as such, they do not exhibit node age inflation seen in TED analyses that do not employ FBD. While we employ a total evidence approach in our example, combining node and tip calibrations is also applicable to matrices consisting solely of fossil taxa and only morphological characters.

5. Conclusion

Nodes and tips are complementary, not competing, approaches to the calibration of molecular clock analyses. Ancient age estimates have become synonymous with tip-calibrated analyses. The construction of the time prior itself is likely to be a causal factor. Our approach to approximating the effective time prior in tip-calibrated analyses shows that when they are implemented alone, tip calibrations can yield divergence time estimates that violate empirical fossil evidence or place exaggerated probability on overly ancient age estimates. Combining node and tip calibrations obviates these effects with the hard minima of node calibrations constraining the uncertainty associated with tip calibrations that, in turn, serve to objectively define the maxima of node age constraints. This approach is appealing because of the positive complementary interaction between the two classes of calibration, but also because it makes the best use of palaeontological data in the construction of evolutionary timescales.

Data accessibility. The dataset supporting this article is available at: http://dx.doi.org/10.5061/dryad.2q3k2.

Authors' contributions. Both authors designed the study, interpreted the results and contributed to writing the manuscript; J.E.O'R. carried out the analysis and led the writing. Both authors approved the manuscript and are accountable for the work herein.

Competing interests. We declare we have no competing interests.

Funding. This work is supported by NERC (NE/L501554/1 to J.E.O'R.), BBSRC (BB/J009709/1, BB/N000919/1 to P.C.J.D.) and the Royal Society (Wolfson Merit Award to P.C.J.D.).

Acknowledgements. Mario dos Reis for discussion.

References

1. Rannala B, Yang Z. 2007 Inferring speciation times under an episodic molecular clock. *Syst. Biol.* **56**, 453–466. (doi:10.1080/10635150701420643)

2. Ronquist F, Klopfstein S, Vilhelmsen L, Schulmeister S, Murray DL, Rasnitsyn AP. 2012

A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. *Syst. Biol.* **61**, 973–999. (doi:10.1093/sysbio/sys058)

3. Warnock RC, Yang Z, Donoghue PCJ. 2012 Exploring uncertainty in the calibration of the molecular clock. *Biol. Lett.* **8**, 156–159. (doi:10.1098/rsbl.2011.0710)
4. Pyron RA. 2011 Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. *Syst. Biol.* **60**, 466–481. (doi:10.1093/sysbio/syr047)
5. O'Reilly J, dos Reis M, Donoghue PCJ. 2015 Dating tips for divergence time estimation. *Trends Genet.* **31**, 637–650. (doi:10.1016/j.tig.2015.08.001)
6. Drummond AJ, Rambaut A. 2007 BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214. (doi:10.1186/1471-2148-7-214)
7. Gavryushkina A, Welch D, Stadler T, Drummond AJ. 2014 Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLoS Comput. Biol.* **10**, e1003919. (doi:10.1371/journal.pcbi.1003919)

The Interrelationships of Placental Mammals and the Limits of Phylogenetic Inference

James E. Tarver[1,2,*], Mario dos Reis[3,4], Siavash Mirarab[5,6], Raymond J. Moran[7], Sean Parker[2], Joseph E. O'Reilly[2], Benjamin L. King[8], Mary J. O'Connell[7], Robert J. Asher[9], Tandy Warnow[5,6,10], Kevin J. Peterson[11], Philip C.J. Donoghue[2], and Davide Pisani[2,12,*]

[1]Department of Biology, The National University of Ireland, Maynooth, Ireland

[2]School of Earth Sciences, University of Bristol, United Kingdom

[3]Department of Genetics, Evolution and Environment, University College London, United Kingdom

[4]School of Biological and Chemical Sciences, Queen Mary University of London, United Kingdom

[5]Department of Computer Science, University of Texas at Austin

[6]Department of Electrical and Computer Engineering, University of California, San Diego

[7]Computational and Molecular Evolutionary Biology Group, School of Biology, Faculty of Life Sciences, University of Leeds

[8]Mount Desert Island Biological Laboratory, Salisbury Cove, Maine

[9]Museum of Zoology, University of Cambridge, United Kingdom

[10]Departments of Bioengineering and Computer Science, University of Illinois at Urbana-Champaign

[11]Department of Biological Sciences, Dartmouth College, Hanover, New Hampshire

[12]School of Biological Sciences, University of Bristol, United Kingdom

*Corresponding author: james.tarver@bristol.ac.uk; davide.pisani@bristol.ac.uk

Accepted: December 23, 2015

Abstract

Placental mammals comprise three principal clades: Afrotheria (e.g., elephants and tenrecs), Xenarthra (e.g., armadillos and sloths), and Boreoeutheria (all other placental mammals), the relationships among which are the subject of controversy and a touchstone for debate on the limits of phylogenetic inference. Previous analyses have found support for all three hypotheses, leading some to conclude that this phylogenetic problem might be impossible to resolve due to the compounded effects of incomplete lineage sorting (ILS) and a rapid radiation. Here we show, using a genome scale nucleotide data set, microRNAs, and the reanalysis of the three largest previously published amino acid data sets, that the root of Placentalia lies between Atlantogenata and Boreoeutheria. Although we found evidence for ILS in early placental evolution, we are able to reject previous conclusions that the placental root is a hard polytomy that cannot be resolved. Reanalyses of previous data sets recover Atlantogenata + Boreoeutheria and show that contradictory results are a consequence of poorly fitting evolutionary models; instead, when the evolutionary process is better-modeled, all data sets converge on Atlantogenata. Our Bayesian molecular clock analysis estimates that marsupials diverged from placentals 157–170 Ma, crown Placentalia diverged 86–100 Ma, and crown Atlantogenata diverged 84–97 Ma. Our results are compatible with placental diversification being driven by dispersal rather than vicariance mechanisms, postdating early phases in the protracted opening of the Atlantic Ocean.

Key words: placental, phylogeny, mammalian, genome, microRNA, palaeontology.

Introduction

The quest for the root of placental mammal phylogeny has achieved the status of an iconic controversy (Teeling and Hedges 2013), with three principal competing hypotheses that resolve either 1) Xenarthra (e.g., armadillos and sloths; Kriegs et al. 2006; Churakov et al. 2009; O'Leary et al. 2013), 2) Afrotheria (e.g., elephants and tenrecs; Murphy et al. 2001; Asher 2007; Nishihara et al. 2007; Hallstrom and Janke 2010; McCormack et al. 2012; Romiguier et al. 2013), or 3) Atlantogenata (i.e., Xenarthra plus Afrotheria; Murphy et al.

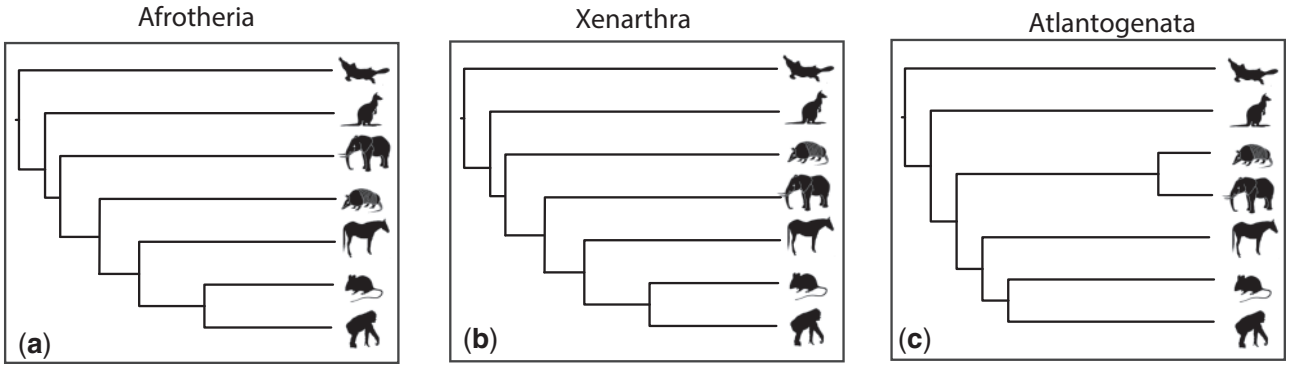


Fig. 1. The three principal competing hypotheses for the higher-level relationships among placental mammals, with either (*a*) Afrotheria, (*b*) Xenarthra, or (*c*) Atlantogenata being the sister taxon to all other placentals.

2007; Wildman et al. 2007; Prasad et al. 2008; Meredith et al. 2011; Song et al. 2012; Morgan et al. 2013) as the sister to all other placentals (fig. 1). Previous analyses have found support for all three hypotheses, leading some to conclude that this phylogenetic problem is impossible to resolve (Churakov et al. 2009; Nishihara et al. 2009; Hallstrom and Janke 2010). This has been considered a consequence of incomplete lineage sorting (ILS; Churakov et al. 2009; Nishihara et al. 2009; Hallstrom and Janke 2010; Guo et al. 2012), reflected in large scale gene tree heterogeneity, a result of the apparent rapidity of successive vicariance-driven divergence events associated with the fragmentation of the Pangaean and Gondwanan supercontinents (Murphy et al. 2001; Wildman et al. 2007; Nishihara et al. 2009). Thus, if placental mammals evolved extremely rapidly, then the root of the placental radiation may be theoretically unresolvable, as it was not strictly bifurcating (Nishihara et al. 2009; Hallstrom and Janke 2010) in the first instance. However, it is possible that phylogenetic resolution has been precluded by practical constraints, which include the availability of adequate models of molecular evolution (Morgan et al. 2013), compositional biases, and/or long branch attraction (Romiguier et al. 2013), and computational limitations on the scale of molecular sequence data sets with limited gene and/or taxon sampling (Morgan et al. 2013). Resolution among these three competing hypotheses is essential to understand the evolutionary origin and diversification of placentals, the most phenotypically diverse group of vertebrates, occupying terrestrial, aerial, and aquatic ecological niches, with body sizes spanning several orders of magnitude (Wilson and Reeder 2005) and which were accompanied by both large scale genomic (e.g., transposable elements, Lynch et al. 2015; conserved noncoding RNAs, Mikkelsen et al. 2007) and morphological (e.g., the placenta; Carter and Mess 2007) innovation.

In an attempt to resolve this phylogenetic controversy, we undertook analyses of two genome-scale data sets representing both coding and noncoding regions of the genome: a 21.4 million nucleotide superalignment of 14,631 genes from 36 taxa, and a 16,050 nucleotide superalignment of 239 pre-miRNAs from 39 taxa. In addition, we reanalyzed the data from three recent analyses that obtained results incongruent with those from our protein coding and nonprotein coding data sets (Hallstrom and Janke 2010; O'Leary et al. 2013; Romiguier et al. 2013), and tested the extent to which morphological data can inform mammal phylogenetics using the 4,541 character data set of (O'Leary et al. 2013).

Materials and Methods

Phylogenetic Analyses

Model Testing

We performed phylogenetic analyses of two nucleotide data sets and three amino acid data sets. The nucleotide data sets were a superalignment of 14,631 protein-coding genes and 36 taxa (totaling 32,116,455), and a superalignment of pre-miRNA sequences comprising 16,050 sites and 42 taxa. The three amino acid data sets were the 11,365 amino acid data set of O'Leary et al. (2013), the AT-rich amino acid data set of Romiguier et al. (2013), and the amino acid data set of Hallstrom and Janke (2010). For all considered data sets Posterior Predictive Analysis (PPA) of biochemical specificity was performed to investigate whether standard, compositionally site-homogeneous, models (e.g., general time reversible [GTR] and Whelan and Goldman [WAG]) provided an adequate fit to the data or whether a more complex (compositionally site–heterogeneous) model (e.g., CAT–GTR; Lartillot and Philippe 2004; Lartillot et al. 2007) was necessary to adequately fit the data. For the nucleotide and microRNA (miRNA) data sets two models were tested, the GTR+G model and CAT–GTR+G. For the amino acid data sets PPA was used to compare the model used in the original studies (Jones, Taylor, and Thorton [JTT]+G [O'Leary et al. 2013]; LG+G [Romiguier et al. 2013]; and WAG+G [Hallstrom and Janke 2010]), against the CAT–GTR+G model. PPA was performed using the serial version of Phylobayes 3.3f (following suggestions from Nicolas Lartillot) using data sets that were

Table 1
Total Size of All Five Data Sets Analyzed and the Percentage of Missing Data in Each

Data Set	Total Sites	% Missing
miRNAs	674,100	22
Nucleotide	770,794,920	39
Hallström and Janke (2010)	7,116,417	21
O'Leary et al. (2013)	522,790	8
Romiguier et al. (2013)	1,065,012	52

subsampled to include a set of approximately 5,000 randomly selected characters. The final number of characters is variable (but comparable) across the different data sets, because of the subsampling strategy we used. However, this is not important as models are compared on the same data sets and not across data sets.

General

Total data set size and percentage of missing data is record in table 1. All Bayesian analyses were performed using the CAT–GTR+G model and implemented with the MPI version of the software Phylobayes (Phylobayes MPI 1.5a; Lartillot et al. 2013). For all Phylobayes analyses two chains were run. Burn-in varied and all chains were run until convergence (which was tested using the BPCOMP software, which is part of the Phylobayes suite). Following the Phylobayes manual, chains were considered to have converged on the same solution when the Maxdiff (maximal difference between observed bipartitions) dropped below 0.2. Maximum Likelihood analyses were performed using RAxML (Stamatakis 2006; Stamatakis et al. 2008) under a GTR+G model, and the bootstrap (100 replicates) was used to estimate support.

Nucleotide

The genome alignment of dos Reis et al. (2012), comprising 36 taxa and 14,631 protein-coding genes was used. Codon sequences were aligned using PRANK with no guide tree to minimize bias associated with any guide tree, although we note that alternate alignment software will generate alternate alignments and subsequent analyses should examine whether such alignments affect our results. The first and second codon positions of all genes were concatenated into a single partition (21,410,970 nt). Because of computational limitations, the full data set could only be analyzed using maximum likelihood. We investigated whether the results of our maximum likelihood GTR+G analyses were supported also under CAT–GTR+G, but because a CAT–GTR+G analysis of the entire superalignment is unfeasible, we removed all the constant and parsimony uninformative sites prior to the analysis. We recognize that this is not ideal, as it can introduce biases and this analysis can consequently be considered to have only an exploratory nature. Initial CAT–GTR+G analyses included all the taxa but did not converge. Inspection of the two chains showed that the horse and tree shrew were unstable within Boreoeutheria. As these taxa are irrelevant to investigate the relationships at the root of the placental tree (Boreoeutheria was monophyletic in both chain and with a posterior probability of 1), we repeated analyses excluding these two taxa. This analysis converged on the same topology within 150 generations (with a Maximal Difference between observed bipartitions dropping to zero).

After having run our phylogenetic analyses we investigated whether the data could significantly discriminate between alternative hypotheses of placental relationships. As CAT–GTR+G and GTR+G supported the same tree for the nucleotide data set these analyses were only explicitly performed under maximum likelihood using the GTR+G model. To do so, the three competing hypotheses were fixed and compared using the approximately unbiased (AU)-Test. Sitewise likelihood values were obtained (under each considered hypothesis of placental relationships) using BASEML (Yang 2007), and CONSEL (Shimodaira and Hasegawa 2001) was used to calculate the AU test. Because of computational limitations AU tests was only performed using the superalignment, and not on the 14,631 individual gene alignments constituting our superalignment. For the gene-by-gene analyses a reduced data set of 11,169 genes was used so that every gene had at least one non-placental outgroup, a Xenarthran, Atlantogenatan, and Boreoeutherian present in the alignment so that the tree could not only be rooted but was also informative as to the relationships between these key clades. For each gene we then estimated the likelihood of each considered tree and performed two different analyses. First, we calculated how many genes supported each alternative hypothesis without considering whether the differences in likelihood between compared trees have been significant. This identified the number of genes for which each considered topology is optimal. Subsequently the Akaike Information Criterion (AIC) test was used to determine whether the genes supporting each specific tree topology, supported that topolgy significantly better than the other tree topologies.

Incomplete Lineage Sorting

The reduced data set of 11,169 genes from the gene-by-gene analyses (see above) was used to define the set of unbinned gene trees. We also used a statistical binning pipeline (Mirarab, Bayzid, et al. 2014) with support threshold set to 50% to create 2,513 bins of genes (1,373 bins with four genes, 1,139 bins with five genes, and one bin with six genes) and estimated a supergene tree for each bin. ASTRAL version 4.7.6 was run on both sets of inputs: the 11,169 unbinned gene trees, and the 2,513 supergene trees, weighting each supergene tree by size of the corresponding bin (weighted statistical binning; Mirarab, Reaz, et al. 2014; Bayzid et al. 2015). To test for the number of gene trees that supported each hypothesis with support

above 50% or 75% threshold, we first restricted each gene tree to branches that have support above the chosen threshold. We then compared each collapsed gene tree against three unresolved trees that represented the three hypotheses. A gene tree can either reject all three hypotheses (i.e., when Xenarthra, Afrotheria, Boreoeutheria, or the branch uniting the three outgroups are rejected), or be indecisive (i.e., be compatible with all three hypotheses; this happens when in the collapsed gene tree, the relationship between Xenarthra, Afrotheria, Boreoeutheria is unresolved), or can support one of the three hypotheses. Thus, five outcomes are possible, and we note the percentage of times each outcome is observed. We also note the percentage of gene trees that support each of the three hypotheses out of those that support just one hypothesis. This produces three estimated probabilities, one for each hypothesis, and we can convert these probabilities to coalescent unit branch lengths by calculating $-\ln(3/2^*(1-p))$ where p is the probability of a hypothesis (Degnan and Rosenberg 2009). For example, for unbinned gene trees, out of 3,495 genes that exclusively supported one the three hypotheses with at least 50% BS, 48.4% of them supported Atlantogenata, which puts the branch length in coalescent units at $-\ln(3/2^*(1-0.484)) = 0.257$. Using 75% threshold with unbinned gene trees results in a length of 0.415, and using supergene trees with 50% and 75% threshold result in lengths of 0.135 and 0.192, respectively.

microRNA

Small RNA libraries were generated from whole juvenile specimens of Armadillo (*Dasypus novemcinctus*), Rabbit (*Oryctolagus cuniculus*), and Guniea Pig (*Cavia porcellus*) using the Illumina Tru-seq small RNA prep kits. In brief, this process involves taking 1 μg of total RNA and adding 5′- and 3′-adapters, which were then reverse transcribed, barcoded, and amplified using polymerase chain reaction. The sample was run out on a Novex 6% TBE Page gel using electrophoresis allowing size fractionation of the sample. The relevant size fraction will be excised and eluted overnight to increase total product. The eluate will be precipitated using EtOH, glycogen, and sodium acetate for 24 h before being resuspended and submitted for sequencing on a GAIIx sequencer at the University if Bristol Transcriptomics Facility. Total read counts were approximately 22M for Armadillo, approximately 13.5M for Guinea Pig, and approximately 21M for Rabbit, and the data processed using in-house algorithms. These read data were used to verify the mature and star reads and hence the end of the pre sequence, which was used for the pre-mir alignments and have been deposited in miRBase. In addition, BLAST searches were conducted for an additional 42 taxa to identify additional miRNA loci. Orthology for each individual miRNA was checked using both distance and, when possible, syntenic analysis. Each individual pre-miRNA from the 42 taxa analyzed was concatenated into the tetrapod superalignment of Field et al. (2014) and analyzed as a standard superalignment (Tarver et al. 2013; Field et al. 2014; Kenny et al. 2015) comprising 15,590 sites and 42 taxa, using the GTR+G model.

Reanalyses

Several recent studies addressed the relationships among the placental mammals finding contradictory results (Hallstrom and Janke 2010; O'Leary et al. 2013; Romiguier et al. 2013). A feature characterizing these studies is the heterogeneity in the choice of the model used for phylogenetic analyses, and the fact that in all cases the substitution model used to analyzed the data was selected in either a subjective way or from a subset of models that did not include well-performing (parameter rich) site-heterogeneous models. Following the results of our PPA (see above), which showed that the models used in the original studies did not fit the data adequately, the three data sets associated with these studies (the 11,365 amino acid data set of O'Leary et al. [2013], the AT-rich amino acid data set of Romiguier et al. [2013], and the amino acid data set of Hallstrom and Janke (2010)) were reanalyzed under the site-heterogeneous CAT–GTR+G model.

Morphological Data Analysis

O'Leary et al. (2013) recently presented a 4,541 character morphological data set. We tested whether this morphological data set could distinguish between the three alternative hypotheses of placental relationships. As in the case of the nucleotide data set the AU-Test was used (implemented in CONSEL), with character-wise likelihood values estimated in RaXML under the MKv model.

Molecular Clock Analysis

The 21m nucletotide alignment was used for the molecular clock analysis. This alignment has previously been used (dos Reis et al. 2012), however, the discovery of new fossil material, as well as revised stratigraphy and phylogenetic placement of taxa means that 20 of the 23 calibration points shared between studies had to be revised (table 2). The previously unpublished calibration on node 37 is justified below following best practice guidelines (Parham et al. 2012).

Calibration on Node 37—Mammalia

Fossil Taxon and Specimen: *Haramiyavia clemmenseni* (Museum of Comparative Zoology MCZ 7/G95) from the Tait Bjerg Beds, Ørsted Dal Member of the Fleming Fjord Formation with an age corresponding to the Late Triassic (?Norian-Rhaetic) (Jenkins et al. 1997).

Phylogenetic Justification: Prior to the discovery of *Haramiyavia clemmenseni*, haramiyids were known from two genera. However, the taxonomic status of these genera was uncertain, and while *H. clemmenseni* exhibited highly specialized dentition it also retained features of the jaw and

Table 2
All 23 Fossil Calibrations Used in This Study

	Node	Minimum Soft Bound	Maximum Soft Bound	References
37	Mammalia—Root	201.1[a]	252.23	Herein—see below
38	Theria	156.3[b]	169.6[c]	Benton et al. (2015)
39	Marsupialia	47.6[d]	131.3[c]	Benton et al. (2015)
40	Placentalia	—	164.6[b]	Benton et al. (2015)
42	Xenarthra	47.6[c]	—	Benton et al. (2015)
43	Afrotheria	56.0[b]	—	Benton et al. (2015)
47	Eulipotyphla	61.6[a]	—	Benton et al. (2015)
49	Chiroptera	45.0[a]	58.9	Phillips (2015)
51	Carnivora	37.3[c]	66.0[c]	Benton et al. (2015)
52	Euungulata	62.5	—	dos Reis et al. (2012)
53	Artiodactyla	—	66.0[c]	Benton et al. (2015)
55	Dolphin/Cow	52.4	—	dos Reis et al. (2012)
56	Euarchontoglires	61.6[a],[c]	—	Benton et al. (2015)
59	Lagomorpha	47.6[c]	66.0[c]	Benton et al. (2015)
60	Rodentia	56.0[c]	66.0[c]	Benton et al. (2015)
61	Guinea Pig/Rat	47.6[c]	59.2[c]	Benton et al. (2015)
63	Muridae	10.4	14.0	dos Reis et al. (2012)
64	Primates	56.0[c]	—	Benton et al. (2015)
65	Strepsirrhini	33.9[c]	56.0[c]	Benton et al. (2015)
67	Anthropoidea	33.9[c]	—	Benton et al. (2015)
68	Catarrhini	24.44[a]	33.9[c]	Benton et al. (2015)
69	Hominidae	11.6[c]	—	Benton et al. (2015)
71	Hominini	6.5[c]	10.0	Benton et al. (2015)

Note.—There are 12 joint (maximum and minimum), two maximum and nine minimum bounds with all maximum and minimum bounds being 'soft'. Although many of the same nodes are calibrated as in dos Reis et al. (2012), only three of the calibrations are retained with all of the others being revised due to:

[a]Change to a different but existing fossil.

[b]Discovery of a new fossil.

[c]Revision of timescale.

[d]Revision of phylogeny.

post-dentary apparatus that indicated a position among stem mammals, cladistically more basal than crown Mammalia, i.e., the clade encompassing monotremes and therians (Jenkins et al. 1997; Zhou et al. 2013). Some recent phylogenetic studies (Zheng et al. 2013; Bi et al. 2014; Krause et al. 2014) have placed Haramiyavia as sister taxon to multituberculates, which are closer to therians than to monotremes and thereby within crown Mammalia. In contrast, other studies argue that the anatomical similarities between haramiyids and multituberculates are convergent (Jenkins et al. 1997; Zhou et al. 2013). We tentatively use Triassic haramiyids as a minimum calibration for Mammalia but are keen to see future, more thorough phylogenetic tests of haramiyid affinities.

Minimum Age: 201.1 Ma

Soft Maximum Age: 252.23 Ma

Age Justification: At present *Haramiyavia clemmenseni* is the oldest known haramiyid from the Tait Bjerg Beds, Ørsted Dal Member of the Fleming Fjord Formation with an age corresponding to the Late Triassic (?Norian-Rhaetic). This stage (Rhaetic) has a minimum bound of 201.3 Ma ± 0.2 Myr (Gradstein et al. 2012) and so the soft minima is set at 201.1 Ma.

Broader Justification: *Hadrocodium* and Docodonta (Luo et al. 2002; Meng et al. 2011) are the closest relatives to crown mammals. *Hadrocodium* is from the early Jurassic of Yunnan Province, China (Sinemurian; Luo et al. 2001), and the oldest docodonts are from the Bathonian of Europe, both of which are younger than *Haramiyavia.* More distantly related taxa such as Morganucodontidae, *Sinoconodon*, and *Adelobasileus*, are known from the late Triassic and early Jurassic and are contemporaneous with *Haramiyavia*, implying substantial ghost lineages in many of these taxa, as such a broad prior is used, setting the soft maxima at the PT extinction, dated at the base of the Induan, 252.17 Ma ± 0.06 Myr (Gradstein et al. 2012) and so the soft maxima is set at 252.23 Ma.

The molecular clock analysis was conducted with MCMCTREE v. 4 .8 a (Yang 2007), using the approximate likelihood method (dos Reis and Yang 2011; Thorne et al. 1998) by calculating the maximum-likelihood estimates of the branch lengths, the gradient vector and Hessian matrix, using BASEML, under the HKY+G4 model (Hasegawa et al. 1985; Yang 1994). We then used the Markov chain Monte Carlo algorithm to estimate divergence times on the constrained tree topology with two separate runs being performed. The

auto-correlated rates model (Thorne et al. 1998; Rannala and Yang 2007) was used to specify the prior of rates, and we followed (dos Reis et al. 2012) for other parameters, that is; the time unit was 100 Myr; a diffuse gamma prior G(1, 1) was used for the overall substitution rate; a rate drift parameter σ^2 was assigned G(1, 1); and the parameters of the birth–death process with species sampling were fixed at $\lambda = \mu = 1$ and $\rho = 0$. The alignment was analyzed as a single partition and we conducted 2,000,000 iterations, sampling every 200 a burn-in of 25%, and with both runs being concatenated post burn-in, after thinning down to 10,000 samples per run, to provide the final posterior values.

Results

Concatenated 21m Nucleotide Phylogenomic Alignment

A fully resolved phylogeny with 100% support for both a sister group relationship between Afrotheria and Xenarthra (Atlantogenata) and between Atlantogenata and Boreoeutheria (fig. 2, left; supplementary fig. S1, Supplementary Material online) was recovered in the analysis of the 21.4 million nucleotide alignment (first and second nucleotide positions) using a single GTR+G model. Of the 35 internal nodes, 32 were recovered with 100% support. Further analyses were performed using the compositionally site–heterogeneous CAT–GTR+G model, which accommodates among-site amino acid (and nucleotide) compositional heterogeneity. This analysis recovered the same topology with all nodes exhibiting 100% support (fig. 2, left; supplementary fig. S2, Supplementary Material online). Unambiguous statistical support for Atlantogenata was confirmed using the AU test, which assesses the level of support for each topology through a site-by-site analysis of the entire data set. The results of this analysis rejected basal positions for both Afrotheria and Xenarthra ($P \leq 0.01$) in favor of Atlantogenata ($P \geq 0.99$) (table 3).

Despite strong support for Atlantogenata, we decided to investigate the level of support for each of the three topologies from the individual genes. We therefore removed all of those genes that were unique to individual lineages, that is, Euarchontoglires, Laurasiatheria, Primates etc., or that were not represented by at least one member of Xenarthra, Atlantogenata, Boreoeutheria, and a nonplacental outgroup. This was done so that each individual gene had the potential as to be informative to the placental root, resulting in a reduced data set of 11,169 genes. The number of individual gene trees recovering alternative topologies (albeit not necessarily with high support) is comparable: Atlantogenata (~33.88%), Afrotheria (~33.84%), Xenarthra (~29.9%), and indecisive (~2.3%) (see table 3). These results could be interpreted to support the prevailing view that the early phylogenetic history of placental mammals was such a rapid radiation that it was not strictly bifurcating. However, 99.4% of the genes fail to discriminate among the competing hypotheses with statistical significance as measured by the AIC test, leaving only 0.2% of genes supporting Atlantogenata, 0.12% supporting Afrotheria, and 0.22% supporting Xenarthra (table 3). Thus, the distribution of support for competing topologies largely reflects the weak phylogenetic signal present in any single gene alignment, rather than suggesting a hard polytomy or very high levels of ILS.

Coalescent-Based Species Tree Estimation

It is known that concatenation analyses, such as those performed here, can be statistically inconsistent or even positively misleading in the presence of sufficient levels of ILS (Roch and Steel 2015). Thus, we further tested the robustness of our phylogeny through the use of ASTRAL-2 (Mirarab and Warnow 2015), a coalescent-based species tree estimation method that is robust to the presence of ILS (Mirarab, Reaz, et al. 2014). We also explored the use of weighted statistical binning (Mirarab, Bayzid, et al. 2014; Bayzid et al. 2015), a technique designed to improve species tree estimation when gene trees have poor resolution. Thus, we used ASTRAL with and without weighted statistical binning, applied to the same 11,169 genes used in the gene-by-gene analysis described earlier.

In both cases a fully resolved tree with 100% support for Atlantogenata (fig. 2, left; supplementary fig. S3, Supplementary Material online) was returned, supporting the concatenation analysis. After restricting analyses to the set of gene trees with high bootstrap support (50% or 75%) for one of the considered hypotheses, support for Atlantogenata was strengthened (supplementary fig. S4, Supplementary Material online). For example, 48% of the unbinned genes and 42% of the binned supergenes that met the 50% bootstrap support threshold supported Atlantogenata, with almost equal numbers of genes supporting an Afrotheran (26% or 30%) or Xenarthran outgroup (26% or 28%). When the level of bootstrap support necessary for the gene trees to be included in the analyses was increased to 75%, the preference for Atlantogenata further increased to 56% of the unbinned genes, and 45% of the binned supergenes, with corresponding decreases in the levels of support for Afrotheria or Xenarthra (fig. 3). This suggests that some (and perhaps much) of the incongruence observed across the gene trees is the result of stochastic errors in gene tree estimation, not ILS. When restricted to gene tree branches that have bootstrap support above 50% or 75%, the branch length for the Atlantogenata group is between 0.14 and 0.42 coalescent units (depending on the threshold and/or the type of gene trees used; see table 4 and Materials and Methods). Critically, the highest levels of support and longest branch lengths in terms of coalescent units for Atlantogenata are returned when we analyze the data using unbinned gene trees. Our estimated coalescent unit branch lengths point to a short branch, but not an extremely short branch that would

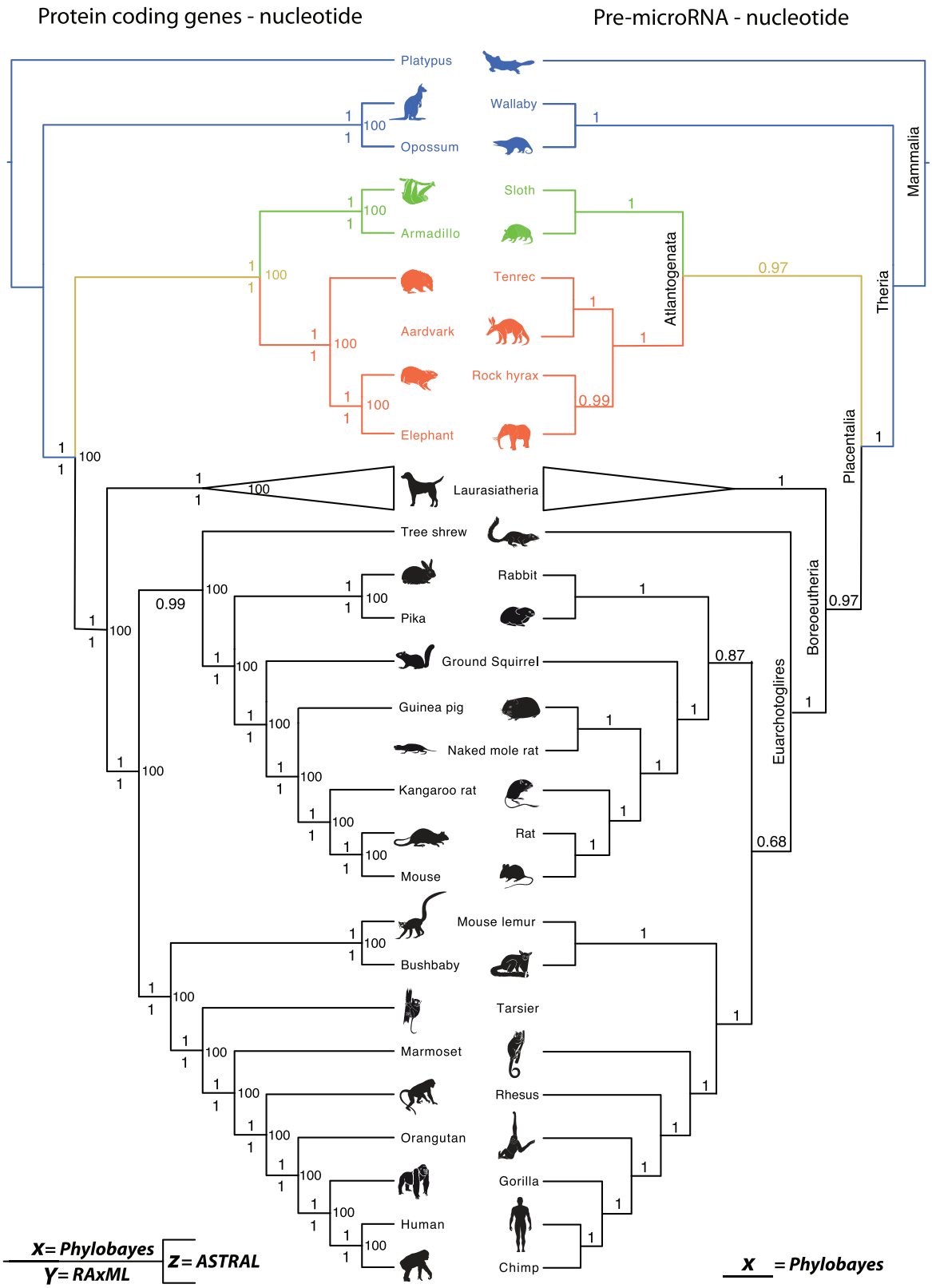


Fig. 2. Results from four of the phylogenetic analyses with each one providing support for Atlantogenata as the sister taxon to all other eutherians. (*a*) The 21.4 million whole-genome nucleotide alignment analyzed using Phylobayes (CAT–GTR+G), RAxML and ASTRAL with support values for almost all nodes being either 1 or 100. (*b*) The single concatenated nucleotide alignment for the pre-mir sequences analyzed under GTR+G in Phylobayes. Laurasiatheria is shown collapsed as the interrelationships among the constituent taxa vary between data sets.

Table 3

Results from the Likelihood Tests of the 21.4m Nucleotide Data Set

Topology	GTR+G4, 1st+2nd Sites, 1 Partition			GTR+G4, 1st +2nd+3rd Sites, 1 Gene Per Partition			AIC Test of Significance
	lnL	Delta lnL	AU test	lnL	Delta lnL	%	%
Atlantogenata	**−115121891**	**0**	***P* ≤ 0.99**	**−196918173**	**0**	**33.9**	**0.2**
Afrotheria	−115123016	1125	*P* ≤ 0.01	−196918837	664	33.84	0.12
Xenarthra	−115123409	1518	*P* ≤ 0.01	−196919286	1113	29.90	0.22
Indecisive						2.30	99.46

Note.—The total log likelihoods for the single partition (1st and 2nd sites) were calculated using BASEML under a GTR+G4 model, with the AU test being conducted on these log likelihoods, and showing unequivocal statistical support for Atlantogenata. Additional log likelihoods were then calculated for each individual gene from a reduced dataset of 11,169 genes (see methods) using a GTR+G4 model with all sites included; given the size of this dataset it is computationally impossible to conduct the AU test (as above) although it is clear that Atlantogenata is the most highly supported topology based on the Delta lnL values. Intriguingly, this topology was not supported by a majority of the genes with approximately 30–33% of genes supporting each alternate topology. However, results of the the AIC test of significance show that 99.46% of genes were unable to distinguish between the three competing hypotheses, while the distribution of support for competing topologies reflects the weak phylogenetic signal present in single gene alignments.

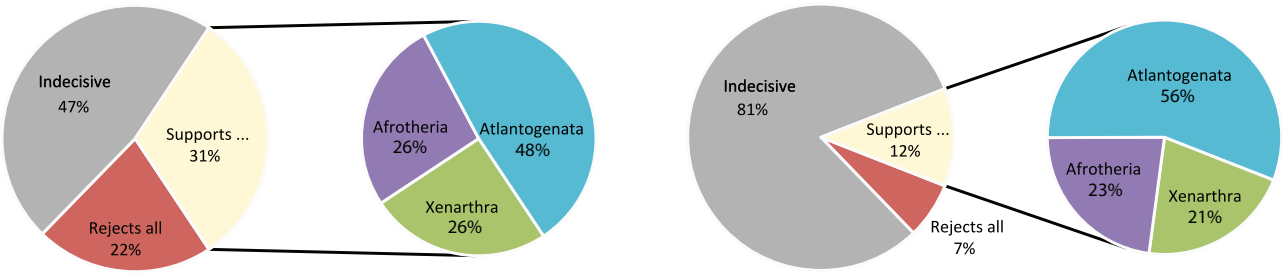


Fig. 3 Results from the discordance analysis of the unbinned gene trees with a threshold bootstrap support value of 50% ("left") and 75% ("right"). These results clearly show that Atlantogenata is the preferred topology, and that much of the incongruence observed across gene trees is due to stochastic errors and not ILS.

Table 4

Shows the Number of Genes, Either Binned or Unbinned Which Support One of Five Outcomes

	Binned (50%)	Unbinned (50%)	Binned (75%)	Unbinned (75%)
Reject all three hypotheses	2994	2418	1673	763
Indecisive	1788	5256	5293	9079
Xenarthra	1751	876	1113	281
Afrotheria	1969	926	1199	303
Atlantogenata	2667	1693	1891	743
Sum of three hypotheses	6387	3495	4203	1327
% Supporting Atlantogenata	0.417566933	0.484406295	0.449916726	0.55990957
Length in coalescent units	0.135075898	0.256971108	0.192220497	0.415309943

Note.—A gene tree can either reject all three hypotheses (i.e., when Xenarthra, Afrotheria, Boreoeutheria, or the branch uniting the three outgroups are rejected), or be indecisive (i.e., be compatible with all three hypotheses; this happens when in the collapsed gene tree, the relationship between Xenarthra, Afrotheria, Boreoeutheria is unresolved), or can support one the three hypotheses. The number of genes that support Atlantogenata is divided by the total number of gene trees that support one of the three hypotheses giving a percentage which can then be used to calculate branch lengths in coalescent units following Degnan and Rosenberg (2009), see Materials and Methods.

violate the hypotheses of a strictly bifurcating tree. These results are largely congruent with concatenation analyses, and suggest that the amount of discordance due to ILS is not sufficient to mislead the concatenation analysis. Thus, although the two analyses are based on data sets of different sizes (11k and 14k genes, respectively), both types of analysis—coalescent-based and concatenation—are highly congruent, and both provide high support for Atlantogenata.

Pre-miRNA Superalignment

In addition to protein-coding genes, we also assembled a concatenated superalignment of 239 noncoding RNA miRNAs consisting of 16,050 nt, which was analyzed under the GTR+G model (see table 5). This miRNA data set provides a second independent molecular data set, that of noncoding RNA genes, to complement protein-coding gene analyses,

Table 5

Posterior Predictive Analyses Conducted to Assess the Fit of the Model to the Data

	O'Leary et al. (2013)		Hallström and Janke (2010)		Romiguier et al. (2013)		Nucleotide		miRNAs	
	JTT+G	CAT–GTR+G	WAG2000–G+I	CAT–GTR+G	LG+G	CAT–GTR+G	GTR+G	CAT–GTR+G	GTR+G	CAT–GTR+G
Observed Diversity	3.1336	3.1336	1.8485	1.8485	2.1998	2.1998	3.1998	3.1998	1.3715	1.3715
Posterior Predictive	3.5652	3.1694	2.0711	1.8597	2.3331	2.2086	3.2733	3.2038	1.3297	1.4800
PP Value	0	0.12	0	0.2381	0	0.4367	0	0.3333	0.9588	0.0557

Note.—For each of the three previously published data sets, the models used in the original studies, JTT+G, WAG2000+G and LG+G' did not adequately fit the data. In comparison the CAT–GTR+G model, which we used in the reanalyses was an adequate fit to the data. For our nucleotide and miRNAs data sets the CAT–GTR+G model was compared with a GTR+G model, for the nucleotide analysis CAT+GTR+G was found to be the best fitting model, while for the miRNAs data set it was the GTR+G model, in both instances the better fitting model was used.

and these data can be analyzed using the same model based approaches. Such an approach has been shown previously to be suitable in resolving interspecies relationships among reptiles (Field et al. 2014), primates (Kenny et al. 2015), nematodes (Kenny et al. 2015), and drosophilids (Kenny et al. 2015). Our pre-miRNA superalignment recovered a fully resolved tree with an Atlantogenata outgroup exhibiting a posterior probability of 0.97 (fig. 2 right; supplementary fig. S5, Supplementary Material online), in agreement with the protein-coding gene analyses. We again used an AU test to investigate site-by-site support for the three topologies on the entire data set with the results significantly rejecting Afrotheria with $P = 0.028$ (Xenarthra $P = 0.250$; Atlantogenata $P = 0.795$), once more providing support against a hard polytomy.

Reanalysis of Three Previously Published Data Sets

Given the consistent support in our two data sets for Atlantogenata, we explored why some previous data sets did not find support for this rooting. Amino acid data sets have yielded support for Afrotheria (Hallstrom and Janke 2010) and Xenarthra (O'Leary et al. 2013), and analysis of an AT-rich amino acid data set supported Afrotheria (Romiguier et al. 2013). We focused on model selection and using PPA we showed that the models used in the original studies (WAG2000+G, JTT+G, LG+G, respectively) did not adequately fit the data (see table 5). In contrast, for each of these three data sets, the compositionally site-heterogeneous CAT–GTR+G model was found to be a satisfactory fit to the data. Reanalysis of all three data sets using the CAT–GTR+G model found variable support for Atlantogenata (supplementary figs. S6–S8, Supplementary Material online), and not for the relationships reported in the original studies, undermining their conclusions. Support values for an Atlantogenata root vary considerably between the three reanalyses with values of 1 (Hallstrom and Janke 2010; supplementary fig. S6, Supplementary Material online), 0.79 (O'Leary et al. 2013; supplementary fig. S7, Supplementary Material online), and 0.5 (Romiguier et al. 2013; supplementary fig. S8, Supplementary Material online). While a support value of 50% is uninformative the original paper had a bootstrap support of 100% for Afrotheria. Thus, although this reanalysis does not have high support the use of a better fitting model fundamentally overturned the previous hypothesis, which was itself very highly supported. Likewise, the results of O'Leary et al. (2013), which previously supported Xenarthra, were overturned to support Atlantogenata. Furthermore, these two data sets with the lowest levels of support either contained low numbers of loci (27 nuclear genes) as in O'Leary et al. (2013) or sampled a nonrandom selection of genes, focusing on AT-rich genes as in Romiguier et al. (2013), such approaches are likely to exacerbate phylogenetic artefacts through both compositional and long branch attraction.

In addition to their amino acid data set, O'Leary et al. (2013) also used a 4,541 character morphological datamatrix. When this matrix was analyzed using the AU test in RAxML (with a constraint tree to make Afrotheria, Xenarthra, and Atlantogenata monophyletic) the morphological data set was unable to distinguish between the three competing hypotheses (Afrotheria $P = 0.288$, Xenarthra $P = 0.212$, and Atlantogenata $P = 0.363$). Thus, when analyzed in isolation, the morphological data are indecisive concerning the earliest diverging lineage of placental mammals.

Timing of Placental Radiation

We estimate the mean divergence times for crown Theria as 164 Ma (CI = 157–170 Ma), crown Placentalia as 93 Ma (CI = 86–100 Ma), and crown Atlantogenata as 90 Ma (CI = 84–97 Ma) (fig. 4 and table 6). These dates are considerably younger than some studies (Springer et al. 2003; Bininda-Emonds et al. 2007), older than others (O'Leary et al. 2013), and congruent with others still (Hallstrom and Janke 2010; Meredith et al. 2011; dos Reis et al. 2012, 2014). As expected, our revised calibrations, older than those employed by dos Reis et al. (2012, 2014), have the effect of making the posterior ages slightly older (Placentalia and Atlantogenata increase in mean age by 3.1 and 2.5 Myr, respectively), while the 95% CI broadens from 88.3–91.6 to 86.5–99.9 Ma in placentals, and 85.9–89.1 to 83.7–96.5 Ma in Atlantogenata. This broadening in the 95% CI reflects the use of a single data partition, in comparison to dos Reis et al. (2012) in which 20 partitions were used. We estimate

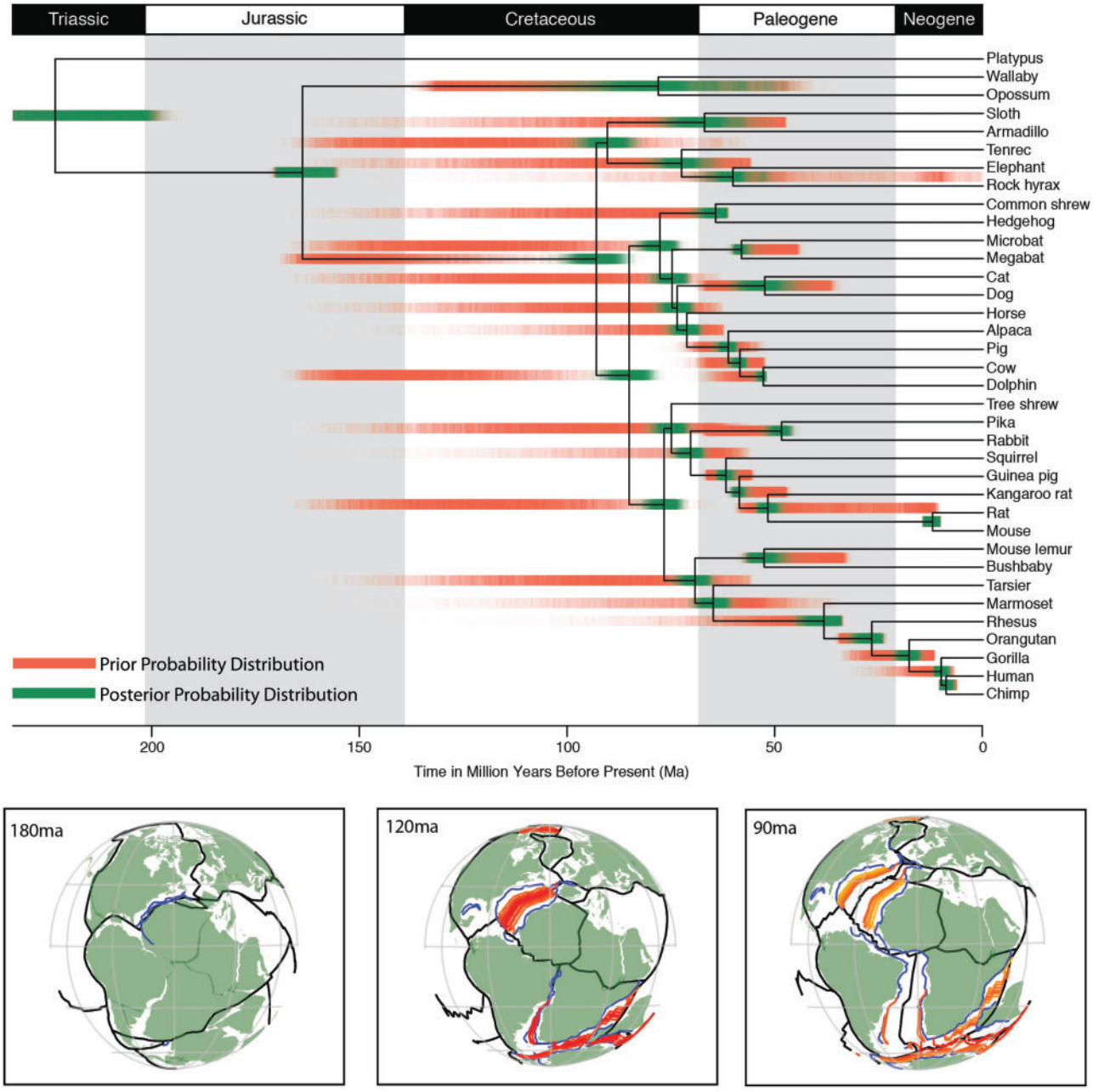


Fig. 4. Results from the molecular clock analysis showing the divergence times for placental lineages with all posterior probabilities shown in "green" and overlaid on the joint prior shown in "red," with both shaded to show values of highest likelihood (see table 6 for the 95% HPD values). Current biogeographic reconstructions for the breakup of Pangea at 180, 120, and 90 Ma, from "left to right," respectively, with hotter colors ("red") indicating faster rates of sea floor formation than colder colors ("blue") based on Seton et al. (2012) and downloadable from http://www.earthbyte.org/Resources/global_plate_model_ESR12.html. Both the Northern and Southern hemisphere continents have separated by 90 Ma, highlighting the role of dispersal, rather than vicariance, for the biogeographic distribution of crown placentals as the breakup of Pangaea predates current molecular clock estimates for the divergence of crown placentals.

diversification of placental orders overlapping the K-Pg mass extinction event at 66 Ma, with all placental orders diversifying between 76 and 51 Ma.

We advocate the use of providing details of not only the combined posterior, but also the marginal prior, which is an analysis run without the sequence data so that the effect of all calibrations can be assessed, as the marginal prior for any node can differ from the original fossil calibration (Warnock et al. 2015). Here we observe that the marginal prior closely approximates (<2 Myr) the calibration points at

Table 6
Prior and Posterior Divergence Times for All Nodes in the Mammal Tree

Node		Marginal Prior[a]			Posterior[b]		
		Mean	95% HPD		Mean	95% HPD	
			Lower	Upper		Lower	Upper
37	Mammalia—Root	226.58	200.99	252.03	223.75	200.47	251.31
38	Theria	163.55	156.45	169.68	163.92	156.67	169.79
39	Marsupialia	95.24	49.77	132.30	78.28	49.11	104.26
40	Placentalia	144.55	115.18	166.80	92.96	86.43	99.91
41	Atlantogenata	119.81	78.70	160.81	90.32	83.73	96.54
42	Xenarthra	80.50	47.46	125.59	67.08	56.62	76.83
43	Afrotheria	90.58	55.92	134.03	72.54	64.85	79.20
44	Paenungulata	49.22	0.12	91.45	60.21	51.45	67.79
45	Boreotheria	135.49	104.09	163.06	85.07	79.93	90.42
46	Laurasiatheria	121.96	84.99	157.18	77.74	73.75	81.96
47	Eulipotyphla	92.65	61.31	132.29	64.35	61.49	67.63
48	Scrotifera	107.66	70.35	146.41	74.82	71.17	78.50
49	Chiroptera	52.07	44.98	58.88	58.23	55.71	60.28
50	Carnivora/Euungulata	95.27	64.18	131.20	73.65	70.18	77.18
51	Carnivora	52.25	37.43	66.15	52.61	45.22	59.70
52	Euungulata	81.43	62.35	110.69	71.35	68.16	74.58
53	Artiodactyla	63.37	55.61	69.78	61.40	59.38	63.40
54	Pig/Cow	59.83	53.00	66.50	58.62	56.92	60.23
55	Dolphin/Cow	56.60	52.21	63.53	52.98	52.10	54.26
56	Euarchontoglires	119.78	84.01	155.50	76.69	72.63	80.60
57	Tree Shrew/Glires	96.61	60.66	136.15	74.93	71.26	78.77
58	Glires	78.48	56.43	115.21	70.34	67.22	73.53
59	Lagomorpha	56.43	47.29	65.59	48.57	46.13	51.65
60	Rodentia	60.89	55.93	65.84	61.97	60.07	63.78
61	Guinea Pig/Rat	53.92	47.79	59.20	58.76	56.91	60.28
62	Kangarro Rat/Rat	36.64	11.75	56.26	51.91	49.14	54.13
63	Muridae	12.18	10.40	13.98	12.21	10.46	13.98
64	Primates	93.18	57.25	130.31	69.27	65.64	72.96
65	Strepsirrhini	44.92	34.04	55.71	52.77	47.91	56.69
66	Haplorrhini	68.64	39.642	108.87	64.96	61.18	68.72
67	Anthropoidea	50.22	33.85	77.91	38.35	33.95	42.77
68	Catarrhini	29.49	24.53	33.97	26.84	24.11	30.32
69	Hominidae	19.04	11.59	28.17	17.85	15.40	20.52
70	Homininae	13.06	6.85	21.03	10.12	8.53	11.45
71	Hominini	8.26	6.51	10.01	8.94	7.51	10.10

[a]The marginal prior was constructed for each node using either the fossil calibrations or from a birth–death process if no calibration was available.
[b]Posterior time estimates for each node based upon the calibrations and the 14k gene data set.

the majority of nodes, with only four exceptions. The only substantial deviation (>4 Myr) is with the soft minima on node 56 (Euarchontoglires). Such results show that the priors were performing as expected based upon the initial fossil calibrations.

Discussion

Thus, far from an intractable phylogenetic problem, it is evident that conflicting placental phylogenies have been a consequence of the use of poorly fitting evolutionary models. Evidently, there was some gene tree heterogeneity caused by ILS during placental diversification. However, we can reject the view (Churakov et al. 2009; Hallstrom and Janke 2010) that this was so rampant as to obscure the fundamental relationships among placental mammals. Instead, our results demonstrate that the primary evidence on which such ideas are based, that is, an equal number of genes supporting mutually exclusive topologies, is the consequence of weak signal in single gene alignments rather than the result of ILS alone. As articulated elsewhere (e.g., Gatesy and Baker 2005; Thompson et al. 2012; Pattinson et al. 2015), isolated, historical signal becomes stronger when individual partitions (such as gene alignments) are combined. Thus, we reject the view

that the root of the placental mammal tree is an unresolvable polytomy, concluding instead that it is correctly resolved as a fundamental divergence between Atlantogenata and Boreoeutheria.

We do not doubt that evidence of ILS reflects the fact that the initial diversification of placentals was rapid as is observed in our molecular clock analysis, the results of which are comparable to those reported elsewhere (dos Reis et al. 2012, 2014; Meredith et al. 2011). Although the discovery of several recent fossils has led to the calibrations being revised substantially with the minimum ages for the root (Mammalia) and Theria being pushed back 38.2 and 32.3 Myr, respectively, whereas the maxima age for Placentalia was pushed back by 33.1 Myr. Yet, such revisions had only minor changes in the estimated mean age of diversification for Placentalia (+3.1 Myr), Atlantogenata (+2.5 Myr), and Boreoeutheria (+2.6 Myr), however, larger changes were observed for the Mammalia (+38.9 Myr), and Theria (−11.5 Myr) in comparison to the results of dos Reis et al. (2012). These dates support dispersal, rather than vicariance, as the underlying mechanism in placental mammal biogeography as they postdate not only the fragmentation of Pangaea, but also the later splitting of Gondwana due to the opening of the Atlantic Ocean (Seton et al. 2012).

Previous studies (Hedges and Maxson 1996; Wildman et al. 2007; Nishihara et al. 2009) have suggested a clear pattern of biogeographic diversification for placentals into four principle lineages (Afrotheria, Xenarthra, Laurasiatheria, and Euarchontoglires) caused by drift-vicariance, which followed the continental breakup of Pangaea into the northern continent of Laurasia (Laurasiatheria + Euarchontoglires) and a Southern Gondwanan continent (Afrotheria and Xenarthra) in the Jurassic (201.3–145 Ma). This was followed by the later breakup of Gondwana into South America (Xenarthra) and Africa (Afrotheria) due to the opening of the Atlantic during the Cretaceous approximately 110 Ma (Smith et al. 1994; Hay et al. 1999; Milani and Thomaz Filho 2000). Recent analyses of global plate tectonics suggests these dates for the complete breakup of Gondwana into S. America and Africa are too old and that this separation was fully complete by 100 Ma (Torsvik et al. 2009; Seton et al. 2012). However, these dates not only predate our mean divergence time for the divergence of Afrotheria from Xenarthra by approximately 10 Myr, but they also lie outside of the 95% HPD (83.73–96.54 Ma), suggesting dispersal by a group of stem Xenarthrans across the Atlantic. While dispersal across the proto Atlantic Ocean may seem unpalatable, the scale of the Atlantic ocean barrier in the Late Cretaceous (fig. 4) was far less significant than that between Africa and Madagascar which has, nevertheless, witnessed multiple post-Mesozoic dispersal events of placentals, including tenrecs, rodents, primates, and carnivores (Yoder and Nowak 2006). Oceanic dispersal of rodents and primates across the South Atlantic during the Eocene (when the overwater distance between Africa and S. America was wider compared with the Cretaceous) is also uncontroversial (Bond et al. 2015).

With the resolution of the evolutionary relationships among Afrotheria, Boreoutheria and Xenarthra, attention must now turn to resolving the problematic relationships within Laurasiatheria and to understanding of the role of dispersal in effecting placental diversification. The results of both our RAxML and ASTRAL analyses as well as the reanalyses of Hallstrom and Janke (2010) and O'Leary et al. (2013) place the tree shrew as sister taxa to Glires, and the horse in an Euungulata clade, and suggests that classical groupings such as Euarchonta and Ferungulata are not supported. Although such results have been presented before (Meredith et al. 2011) this is an area of significant conflict between previously published studies (Kriegs et al. 2006; Murphy et al. 2007; Nishihara et al. 2009; Hallstrom and Janke 2010; Meredith et al. 2011; McCormack et al. 2012; Nery et al. 2012; Song et al. 2012; Morgan et al. 2013; O'Leary et al. 2013; Romiguier et al. 2013). It is these two rogue taxa (tree shrew and horse) which are the cause of alternate tree topologies, and it is no surprise that the same two taxa were the ones that needed to be removed from our phylobayes analysis as they prevented the runs from converging. In future increased taxonomic sampling of additional perissodactyl lineages, that is, Equidae (donkeys, and zebras), Rhinocerotidae (rhinos), and Tapiridae (tapirs) as well as Scandentia lineages, that is, *Anathana* (Madras treeshrew), *Dendrogale* (Bornean smooth-tailed treeshrew), and *Ptilocercus* (Pen-tailed treeshrew), will lead to increased confidence in the phylogenetic placement of these lineages. While a better understanding for the role of dispersal through not only the late Mesozoic but also the Paleogene (or early Cenozoic) can be achieved through a more precise understanding of the geography including sea-level changes, and not merely the tectonics and biogeography through this interval. In addition the inclusion of fossils within analyses of their living relatives needs to become more widespread, allowing not only greater precision in divergence time estimation through the use of tip dating in molecular clock analyses (Ronquist et al. 2012), but also to better understand the pattern of character acquisition (Patterson 1981), and changes in diversity, either to identify diversification rate shifts (Tarver and Donoghue 2011; Wagner and Estabrook 2014) or broader patterns of biological diversity (Wagner 2000; Tarver et al. 2011; Losos et al. 2013).

The results of our study suggest that other seemingly intractable phylogenetic debates, such as the position of ctenophores, chaetognaths, Acoelomorpha, and the relationships among lophotrochozoans (Dunn et al. 2014), may be solvable by combining genome-scale data sets with realistic models of molecular evolution and rigorous coalescent-based species tree estimation methods.

Supplementary Material

Supplementary figures S1–S8 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by Irish Research Council EMPOWER Postdoctoral Fellowship (J.E.T., D.P.); Marie Curie actions of EU FP7 Fellowship (J.E.T., P.C.J.D.); BBSRC Standard Grant BB/J009709/1 (M.R., P.C.J.D.); Irish Research Council Postgraduate Scholarship GOIPG/2014/306 (R.J.M.); NERC studentship NE/L501554/1 (J.E.O'R.); Royal Society Wolfson Merit Award (P.C.J.D.); Leverhulme Trust Research Fellowship (P.C.J.D.); U.S. National Institutes of Health grants GM104318 and GM103423 (B.L.K.); the National Aeronautic and Space Agency (K.J.P); the U.S. National Science Foundation 1461364 (T.W.); and the Howard Hughes Medical Institute Graduate Fellowship (S.M.).

Note Added in Proof

Whilst this paper was in review following revision, Luo et al. (2015) published a phylogeny which placed haramiyids as stem mammals. In the molecular clock study here haramiyids were considered crown mammals and were tentatively used as a calibration point on the root (node 37). We therefore reran the molecular clock analysis using the much younger Mammalian calibration found in Benton et al. (2015) which places a soft minima at 164.9 Ma and a soft maxima at 201.5 Ma. As expected the root age of Mammalia changes by 40 Myr to become younger, whilst the age of Theria changes by 550,000 years, Marsupialia by 140,000 years and all other nodes by less than 100,000 years. The results of this reanalysis had no material effect on the conclusions of this study.

Literature Cited

Asher RJ. 2007. A web-database of mammalian morphology and a reanalysis of placental phylogeny. BMC Evol Biol. 7:108.

Bayzid S, Mirarab S, Boussau B, Warnow T. 2015. Weighted Statistical Binning: enabling statistically consistent genome-scale phylogenetic analyses. PLoS One 10:e0129183.

Benton MJ, et al. 2015. Constraints on the timescale of animal evolutionary history. Palaeontol Electron. 18:1–106.

Bi S, Wang Y, Guan J, Sheng X, Meng J. 2014. Three new Jurassic euharamiyidan species reinforce early divergence of mammals. Nature 514:579–584.

Bininda-Emonds ORP, et al. 2007. The delayed rise of present-day mammals. Nature 446:507–512.

Bond M, et al. 2015. Eocene primates of South America and the African origins of New World monkeys. Nature 520:538–541.

Carter AM, Mess A. 2007. Evolution of the placenta in eutherian mammals. Placenta 28:259–262.

Churakov G, et al. 2009. Mosaic retroposon insertion patterns in placental mammals. Genome Res. 19:868–875.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol. 24:332–340.

dos Reis M, Donoghue PCJ, Yang Z. 2014. Neither phylogenomic nor palaeontological data support a Palaeogene origin of placental mammals. Biol Lett. 10:20131003.

dos Reis M, et al. 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. Proc Biol Sci. 279:3491–3500.

dos Reis M, Yang Z. 2011. Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. Mol Biol Evol. 28:2161–2172.

Dunn CW, Giribet G, Edgecombe GD, Hejnol A. 2014. Animal phylogeny and its evolutionary implications. Ann Rev Ecol Evol Syst. 45:371–395.

Field DJ, et al. 2014. Toward consilience in reptile phylogeny: miRNAs support an archosaur, not lepidosaur, affinity for turtles. Evol Dev. 16:189–196.

Gatesy J, Baker RH. 2005. Hidden likelihood support in genomic data: can forty-five wrongs make a right? Syst Biol 54:483–492.

Gradstein FM, Ogg JG, Schmitz M, Ogg G. 2012. The geologic time scale 2012. Amsterdam: Elsevier.

Guo C-Q, et al. 2012. *Riccardiothallus devonicus* gen. et sp. nov., the earliest simple thalloid liverwort from the Lower Devonian of Yunnan, China. Rev Palaeobot Palynol. 176–177:35–40.

Hallstrom BM, Janke A. 2010. Mammalian evolution may not be strictly bifurcating. Mol Biol Evol. 27:2804–2816.

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

Hay WW, et al. 1999. Alternative global Cretaceous paleogeography. In: Barrera E, Johnson C, editors. Evolution of the Cretaceous ocean/climate system. Boulder (CO): The Geological Society of America. p. 1–48.

Hedges SB, Maxson LR. 1996. Molecules and morphology in amniote phylogeny. Mol Phylogenet Evol. 6:312–314.

Jenkins FA, Gatesy SM, Shubin NH, Amaral WW. 1997. Haramiyids and Triassic mammalian evolution. Nature 385:715–718.

Kenny NJ, et al. 2015. The phylogenetic utility and functional constraint of microRNA flanking sequences. Proc Biol Sci. 282:20142983

Krause DW, et al. 2014. First cranial remains of a gondwanatherian mammal reveal remarkable mosaicism. Nature 515:512–517.

Kriegs JO, et al. 2006. Retroposed elements as archives for the evolutionary history of placental mammals. PLoS Biol. 4:537–544.

Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol. 7:S4.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol. 21:1095–1109.

Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI. Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst Biol. 62:611–615.

Losos JB, et al. 2013. Evolutionary biology for the 21st century. PLoS Biol. 11:e1001466.

Luo ZX, Crompton AW, Sun AL. 2001. A new mammaliaform from the early Jurassic and evolution of mammalian characteristics. Science 292:1535–1540.

Luo ZX, Gatesy SM, Jenkins FA, Amaral WW, Shubin NH 2015. Mandibular and dental characteristics of Late Triassic mammaliaform Haramiyavia and their ramifications for basal mammal evolution. Proc Natl Acad Sci USA. 112(51):E7101–E7109.

Luo Z-X, Kielan-Jaworowska Z, Cifelli RL. 2002. In quest for a phylogeny of Mesozoic mammals. Acta Palaeontol Pol. 47:1–78.
Lynch VJ, et al. 2015. Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. Cell Rep. 10:551–561.
McCormack JE, et al. 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. Genome Res. 22:746–754.
Meng J, Wang Y, Li C. 2011. Transitional mammalian middle ear from a new Cretaceous Jehol eutriconodont. Nature 472:181–185.
Meredith RW, et al. 2011. Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. Science 334:521–524.
Mikkelsen TS, et al. 2007. Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. Nature 447:167–177.
Milani EJ, Thomaz Filho A. 2000. In: Cordani UG, Milani EJ, Thomaz Filho A, Campos DA, editors. Tectonic evolution of South America. Rio de Janiero: International Geological Congress. p. 389–449.
Mirarab S, Bayzid MS, Boussau B, Warnow T. 2014. Statistical binning enables an accurate coalescent-based estimation of the avian tree. Science 346:1250463.
Mirarab S, Reaz R, et al. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30:i541–i548.
Mirarab S, Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31(12):i44–i52.
Morgan CC, et al. 2013. Heterogeneous models place the root of the placental mammal phylogeny. Mol Biol Evol. 30:2145–2156.
Murphy WJ, et al. 2001. Molecular phylogenetics and the origins of placental mammals. Nature 409:614–618.
Murphy WJ, Pringle TH, Crider TA, Springer MS, Miller W. 2007. Using genomic data to unravel the root of the placental mammal phylogeny. Genome Res. 17:413–421.
Nery MF, González DJ, Hoffmann FG, Opazo JC. 2012. Resolution of the laurasiatherian phylogeny: evidence from genomic data. Mol Phylogenet Evol. 64:685–689.
Nishihara H, Maruyama S, Okada N. 2009. Retroposon analysis and recent geological data suggest near-simultaneous divergence of the three superorders of mammals. Proc Natl Acad Sci U S A. 106:5235–5240.
Nishihara H, Okada N, Hasegawa M. 2007. Rooting the eutherian tree: the power and pitfalls of phylogenomics. Genome Biol. 8:R199.
O'Leary MA, et al. 2013. The placental mammal ancestor and the Post-K-Pg radiation of placentals. Science 339:662–667.
Parham JF, et al. 2012. Best practices for justifying fossil calibrations. Syst Biol. 61:346–359.
Patterson C. 1981. Significance of fossils in determining evolutionary relationships. Ann Rev Ecol Syst. 12:195–223.
Pattinson DJ, Thompson RS, Piotrowski AK, Asher RJ. 2015. Phylogeny, paleontology, and primates: do incomplete fossils bias the tree of life? Syst Biol 64:169–186.
Phillips MJ 2015. Four mammal fossil calibrations: balancing competing palaeontological and molecular considerations. Palaeontol Electron. 18: 1–16.
Prasad AB, Allard MW, Green ED, Nisc Comparat Sequencing P. 2008. Confirming the phylogeny of mammals by use of large comparative sequence data sets. Mol Biol Evol. 25:1795–1808.
Rannala B, Yang Z. 2007. Inferring speciation times under an episodic molecular clock. Syst Biol. 56:453–466.
Roch S, Steel M. 2015. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. Theor Popul Biol. 100C:56–62.
Romiguier J, Ranwez V, Delsuc F, Galtier N, Douzery EJP. 2013. Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. Mol Biol Evol. 30:2134–2144.
Ronquist F, et al. 2012. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. Syst Biol. 61:973–999.
Seton M, et al. 2012. Global continental and ocean basin reconstructions since 200 Ma. Earth-Sci Rev. 113:212–270.
Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246–1247.
Smith AG, Smith DG, Funnell BM. 1994. Atlas of Cenozoic and Mesozoic coastlines. Cambridge: Cambridge University Press.
Song S, Liu L, Edwards SV, Wu S. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc Natl Acad Sci U S A. 109: 14942–14947.
Springer MS, Murphy WJ, Eizirik E, Brien SJO. 2003. Placental mammal diversification and the Cretaceous-Tertiary boundary. Proc Natl Acad Sci U S A. 100:1056–1061.
Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.
Stamatakis A, Hoover P, Rougemont J. 2008. A Rapid Bootstrap Algorithm for the RAxML Web Servers. Syst Biol. 57:758–771.
Tarver JE, Donoghue PCJ. 2011. The trouble with topology: phylogenies without fossils provide a revisionist perspective of evolutionary history in topological analyses of diversity. Syst Biol. 60:700–712.
Tarver JE, Donoghue PCJ, Benton MJ. 2011. Is evolutionary history repeatedly rewritten in light of new fossil discoveries? Proc R Soc B Biol Sci. 278:599–604.
Tarver JE, et al. 2013. miRNAs: small genes with big potential in metazoan phylogenetics. Mol Biol Evol. 30:2369–2382.
Teeling EC, Hedges SB. 2013. Making the impossible possible: rooting the tree of placental mammals. Mol Biol Evol. 30:1999–2000.
Thompson RS, BÄRmann EV, Asher RJ. 2012. The interpretation of hidden support in combined data phylogenetics Die Interprätation von "hidden support" in phylogenetischen Analysen mit kombinierten Datensätzen. J Zool Syst Evol Res. 50:251–263.
Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol Biol Evol. 15:1647–1657.
Torsvik TH, Rousse S, Labails C, Smethurst MA. 2009. A new scheme for the opening of the South Atlantic Ocean and the dissection of an Aptian salt basin. Geophys J Int. 177:1315–1333.
Wagner PJ. 2000. The quality of the fossil record and the accuracy of phylogenetic inferences about sampling and diversity. Syst Biol. 49:65–86.
Wagner PJ, Estabrook GF. 2014. Trait-based diversification shifts reflect differential extinction among fossil taxa. Proc Natl Acad Sci U S A. 111:16419–16424.
Warnock RC, Parham JF, Joyce WG, Lyson TR, Donoghue PC. 2015. Calibration uncertainty in molecular dating analyses: there is no substitute for the prior evaluation of time priors. Proc Biol Sci. 282:20141013.
Wildman DE, et al. 2007. Genomics, biogeography, and the diversification of placental mammals. Proc Natl Acad Sci U S A. 104:14395–14400.
Wilson DE, Reeder DM. 2005. Mammal species of the world: a taxonomic and geographic reference. Baltimore: John Hopkins University Press.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.
Yang ZH. 1994. Maximum-likelihood phylogenetic estimation from DNA-sequences with variable rates over sites—approximate methods. J Mol Evol. 39:306–314.

Yoder AD, Nowak MD. 2006. Has vicariance or dispersal been the predominant biogeographic force in Madagascar? Only time will tell. Ann Rev Ecol Evol Syst. 37:405–431.

Zheng X, Bi S, Wang X, Meng J. 2013. A new arboreal haramiyid shows the diversity of crown mammals in the Jurassic period. Nature 500:199–202.

Zhou CF, Wu S, Martin T, Luo ZX. 2013. A Jurassic mammaliaform and the earliest mammalian evolutionary adaptations. Nature 500:163–167.

Associate editor: Gunter Wagner

PROCEEDINGS B

rspb.royalsocietypublishing.org

Research

Cite this article: Puttick MN *et al*. 2017 Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data. *Proc. R. Soc. B* **284**: 20162290.
http://dx.doi.org/10.1098/rspb.2016.2290

Received: 19 October 2016
Accepted: 5 December 2016

Subject Category:
Palaeobiology

Subject Areas:
evolution, palaeontology, taxonomy and systematics

Keywords:
phylogeny, Bayesian, parsimony, cladistics, morphology, palaeontology

Authors for correspondence:
Davide Pisani
e-mail: davide.pisani@bristol.ac.uk
Philip C. J. Donoghue
e-mail: phil.donoghue@bristol.ac.uk

†These authors contributed equally to this study.

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3653186.

THE ROYAL SOCIETY PUBLISHING

Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data

Mark N. Puttick[1,3,†], Joseph E. O'Reilly[1,†], Alastair R. Tanner[2], James F. Fleming[1], James Clark[1], Lucy Holloway[1], Jesus Lozano-Fernandez[1,2], Luke A. Parry[1], James E. Tarver[1], Davide Pisani[1,2] and Philip C. J. Donoghue[1]

[1]School of Earth Sciences, and [2]School of Biological Sciences, University of Bristol, Life Sciences Building, 24 Tyndall Avenue, Bristol BS8 1TQ, UK
[3]Department of Life Sciences, Natural History Museum, Cromwell Road, London SW7 5BD, UK

MNP, 0000-0002-1011-3442; JEO, 0000-0001-9775-253X; JC, 0000-0003-2896-1631; LH, 0000-0003-1603-2296; JL-F, 0000-0003-3597-1221; LAP, 0000-0002-3910-0346; PCJD, 0000-0003-3116-7463

Morphological data provide the only means of classifying the majority of life's history, but the choice between competing phylogenetic methods for the analysis of morphology is unclear. Traditionally, parsimony methods have been favoured but recent studies have shown that these approaches are less accurate than the Bayesian implementation of the Mk model. Here we expand on these findings in several ways: we assess the impact of tree shape and maximum-likelihood estimation using the Mk model, as well as analysing data composed of both binary and multistate characters. We find that all methods struggle to correctly resolve deep clades within asymmetric trees, and when analysing small character matrices. The Bayesian Mk model is the most accurate method for estimating topology, but with lower resolution than other methods. Equal weights parsimony is more accurate than implied weights parsimony, and maximum-likelihood estimation using the Mk model is the least accurate method. We conclude that the Bayesian implementation of the Mk model should be the default method for phylogenetic estimation from phenotype datasets, and we explore the implications of our simulations in reanalysing several empirical morphological character matrices. A consequence of our finding is that high levels of resolution or the ability to classify species or groups with much confidence should not be expected when using small datasets. It is now necessary to depart from the traditional parsimony paradigms of constructing character matrices, towards datasets constructed explicitly for Bayesian methods.

1. Introduction

The fossil record affords the only direct insight into evolutionary history of life on the Earth, but the incomplete preservation and temporal distribution of fossils has long prompted biologists to seek alternative perspectives, such as molecular phylogenies of living species, eschewing palaeontological evidence altogether [1]. However, there is increasing acceptance that analyses of historical diversity cannot be made without phylogenies that incorporate fossil species [2,3] and calibrating molecular phylogenies to time cannot be achieved effectively without recourse to the fossil record [4]. Integrating fossil and living species has become the grand challenge and there has been a modest proliferation of phylogenetic approaches to the analysis of phenotypic data. While conventional parsimony remains the most widely employed method, alternative parsimony [5] and probabilistic [6] models have been developed to better accommodate heterogeneity in

the rate of evolution among characters and across phylogeny. Unfortunately, these competing methods invariably yield disparate phylogenetic hypotheses among which it is difficult to discriminate as the true tree is never known for empirical data.

A number of studies have attempted to establish the efficacy of competing phylogenetic methods using data simulated from known trees [7–9], finding that the probabilistic Mkv model outperforms parsimony methods, among which, conventional equal-weights parsimony (EW-Parsimony) performs best. However, these studies were potentially biased by their experimental design: (i) two of the studies employed a generating tree that was unresolved and, therefore, biased against parsimony methods which recover resolved trees; (ii) these studies did not discriminate between the impact of the probabilistic model and its implementation in a Bayesian framework; (iii) based on single empirical trees, the impact of tree symmetry, which is known to confound phylogeny estimation [10], was not explored; and (iv) only binary characters were considered, whereas empirical datasets are commonly a mixture of binary and multistate characters. Therefore, we compare the performance of EW-Parsimony, implied-weights parsimony (IW-Parsimony), maximum-likelihood and Bayesian implementations of the Mk model, based on datasets with different numbers of characters, comprising binary and multistate characters and simulated on a fully balanced and a maximally imbalanced phylogenetic tree. We find that Bayesian inference outperforms all other methods, while EW-Parsimony performs better than IW-Parsimony, and maximum likelihood performs worst of all. We apply these competing phylogenetic methods to empirical morphological datasets of similar sizes to our simulated datasets and explore the efficacy of the ensuing phylogenetic hypotheses in the light of the conclusions derived from our simulation-based study.

2. Material and methods

(a) Simulation of morphological matrices

We simulated data on two 32-taxon generating trees at the extremes of tree symmetry: one fully asymmetrical and one fully symmetrical (see electronic supplementary material, figure S1). For each tree, we simulated matrices of three sizes: 100, 350 and 1000 characters. We generated matrices using the HKY + Γ Continuous model of molecular substitution, with $\kappa = 2$, the shape (set equal to rate) of the gamma distribution and underlying substitution rate for each replicate sampled from independent and identically distributed exponential distributions with a mean of 1, and character state stationary frequencies fixed as $\pi = [0.2,0.2,0.3,0.3]$. We used a fixed and uneven stationary distribution of nucleotide frequencies to ensure our simulation model did not collapse into the Mk model, as this would bias the analysis in favour of Mk model-based approaches. We simulated 1000 replicate matrices with unique substitution parameters for each tree and each character number, resulting in a total of 6000 matrices. We set two types of character within each matrix, binary and multistate, and we simulated a proportion of 55 binary : 45 multistate characters, based on the mean ratio found in a survey of empirical morphological data matrices [11]. We established binary characters by converting data simulated under the HKY model to R/Y coding (i.e. 0/1): morphological multistate characters were simulated by converting DNA bases to integers.

To ensure that our simulated data are realistic, we generated each set of 1000 unique replicate matrices such that the among-matrix distribution of homoplasy approximated the distribution of empirical homoplasy, characterized by the consistency index (CI), reported by Sanderson & Donoghue [12]. To approximate this distribution of homoplasy, we placed the Sanderson and Donoghue data into quantized bins of CI spanning 0.05, between the empirical bounds of 0.26 and 1.0, and simulated matrices until we matched this expected density per bin (electronic supplementary material, figure S2).

The code used to simulate these data is available in the electronic supplementary material.

(b) Phylogenetic analysis

We analysed the simulated matrices with EW-Parsimony, IW-Parsimony ($k = 2$) and the Mk model [6] under both maximum-likelihood and Bayesian implementations. EW-Parsimony and IW-Parsimony estimation of topology was performed in TNT [13]. We used the Mk + Γ model for maximum-likelihood estimation of topology in RAxML v. 7.2 [14], and Bayesian estimation of topology in MrBayes v. 3.2 [15]. As the approximate likelihood calculation of RAxML may be distant from the true likelihood [16], we conducted a sensitivity test by re-analysing a subset of our data with the likelihood implementation of the Mk model in IQ-tree [17]; both methods gave effectively identical results, indicating results from the likelihood Mkv model are not software specific.

The Mkv model is inappropriate due to the lack of acquisition bias in the simulated data. For maximum-likelihood and Bayesian analyses, we applied the discretized gamma distribution model to account for between-character rate heterogeneity. For Bayesian analyses, the posterior distribution was sampled 1 million times by four chains using the Metropolis-coupled Markov-chain Monte Carlo algorithm with every 100th sample stored, resulting in 10 000 samples; two independent runs were performed for each replicate and the two resulting posterior samples were combined after qualitative assessment of convergence. For parity, we characterized the result of all phylogenetic methods as the majority-rule consensus of resultant tree samples. We did not employ bootstrap methods to measure support for parsimony and likelihood analyses because phenotypic data does not meet the assumption that phylogenetic signal is distributed randomly among characters.

We used the Robinson–Foulds metric [18] to compare the similarity of estimated topologies against their respective generating tree. We also noted the per-node resolution, and the variation of node accuracy across the topology.

(c) Empirical analyses

We analysed four published palaeontological phenotype character matrices that encompass a range of character numbers and a diverse sample of taxa from the Tree of Life [19–22]. We resolved any ambiguities in character coding to their most derived state for each matrix to make analyses compatible across the different phylogenetic methods, facilitating comparison of results. We analysed each matrix by applying the same settings used to analyse our simulated matrices: EW-Parsimony, IW-Parsimony, as well as Bayesian and maximum-likelihood implementations of the Mk model. Empirical morphological matrices are rarely constructed to contain invariant or parsimony uninformative characters. Therefore, the Mkv extension of the Mk model, which uses conditional likelihood to correct for such acquisition biases, is more appropriate than the Mk model for analysis of these empirical data matrices [6].

3. Results

(a) Simulated data

Accuracy is higher for trees inferred from data simulated on a symmetrical topology compared with trees

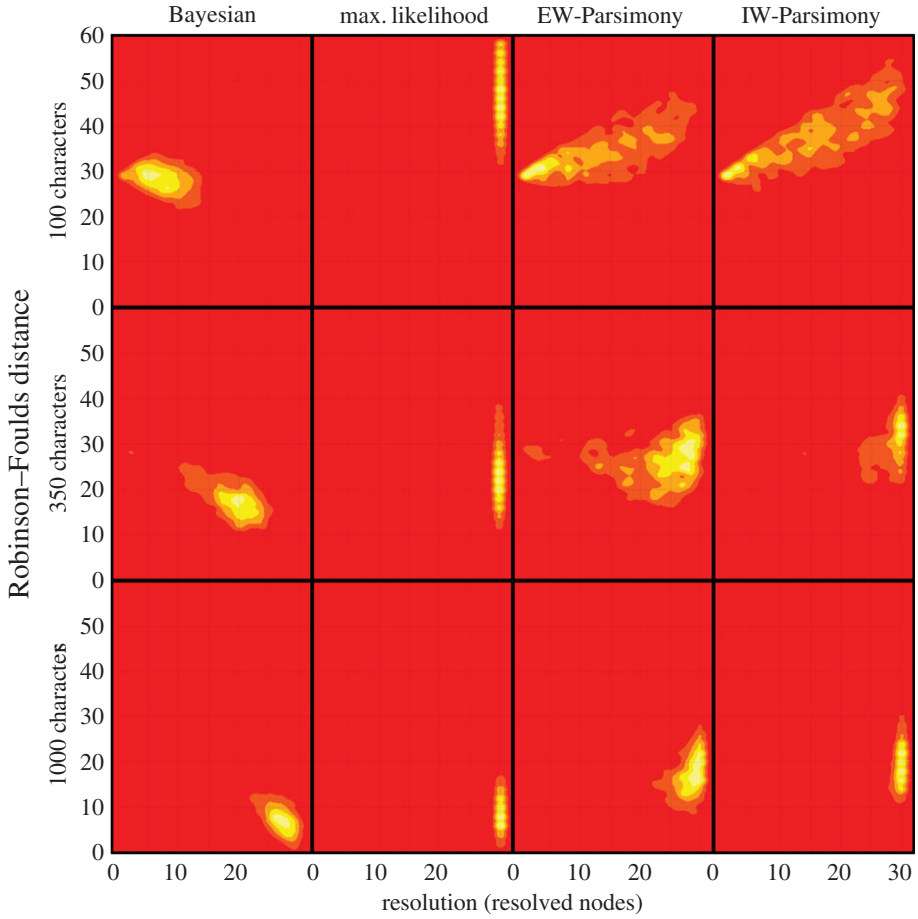


Figure 1. Contour plots of Robinson–Foulds distance against phylogenetic resolution, indicating the higher accuracy of Bayesian implementations against all other methods with data generated on the asymmetrical phylogeny. The spectrum of red to yellow, reflect lower to higher density of trees. As the number of characters increases all methods converge on the correct phylogeny, although Bayesian phylogenies are generally the least resolved. The other methods achieve higher resolution but at a cost of lower accuracy. Data generated on the symmetrical phylogeny shows similar patterns but with much less variance and higher accuracy for all iterations; this lack of variance means point estimates cannot be shown as density estimates. (Online version in colour.)

estimated from data simulated on the asymmetrical topology (cf. figures 2 and 3). Bayesian consensus phylogenies are generally the least well-resolved (figure 1). All methods estimated topologies with greater accuracy as the number of analysed characters increased (figures 2 and 3; electronic supplementary material, table S5–S7). All methods, apart from maximum likelihood, produced phylogenies with greater resolution with higher numbers of characters (figure 1).

For all implementations and dataset sizes, the Bayesian implementation of the Mk model achieves higher accuracy compared with other methods (table 1; figures 1–3). The two parsimony methods achieved the next highest levels of accuracy, EW-Parsimony achieving greater accuracy than IW-Parsimony. Maximum likelihood was the least accurate method for topology reconstruction for both the symmetrical and asymmetrical phylogenies (table 1). The relative accuracy of these phylogenetic methods remains the same across all dataset sizes and the two simulation topologies (table 1; figures 1–3).

Nodes closer to the tips are significantly more accurately reconstructed in the asymmetrical phylogenies across all dataset sizes (table 2 and figure 2; electronic supplementary material, figure S8). In the symmetrical trees, there was no significant correlation between distance from the tips and the accuracy of node reconstruction, except in the maximum-likelihood analysis of 100 characters (figure 2 and table 2).

(b) Empirical phylogenies

Patterns of resolution achieved from the simulated datasets are similar for the empirical datasets. The Bayesian implementation of the Mk model estimates the least resolved phylogenies and maximum likelihood produces fully resolved trees (full trees are shown electronic supplementary material, figure S9–S15).

Kulindroplax, from the Sutton *et al.* [22] dataset, is supported as a crown-mollusc based on maximum likelihood, EW-Parsimony and IW-Parsimony (figure 4*a–d*). The results of the IW-Parsimony analysis are most similar to the original results [22], with *Kulindroplax* resolved as a crown-aplacophoran; maximum-likelihood analysis of the dataset resolved *Kulindroplax* as the stem-aplacophoran. The result of the Bayesian analysis of the dataset is largely unresolved, and *Kulindroplax* is not discriminated as a member of any clade within molluscs or even as a member of total-group Mollusca.

The anthophyte hypothesis (non-monophyletic gymnosperms sister to seed ferns plus angiosperms) recovered by Hilton & Bateman [19] is supported by our EW-Parsimony and maximum-likelihood analyses of their dataset which recovered a paraphyletic seed ferns plus Gnetophyta as sister to angiosperms (figure 4*f,g*); the results of Bayesian and IW-Parsimony analyses of the same dataset contradict the anthophyte hypothesis (figure 4*e,h*). The Bayesian analysis produced a non-monophyletic gymnosperms with the relationships between them and seed ferns unresolved with the exception of

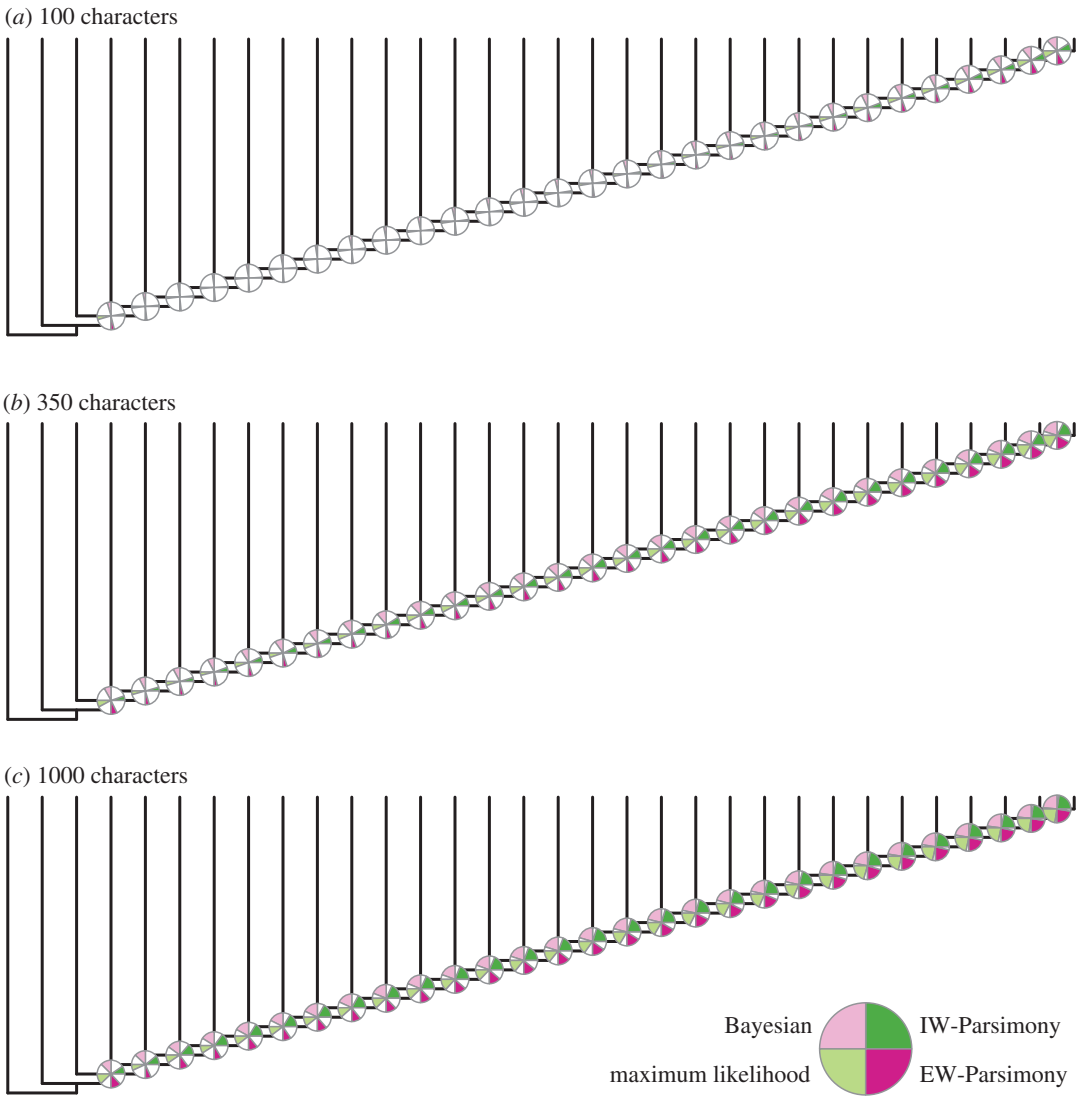


Figure 2. Accuracy of nodes is higher for those closer to the tips in the asymmetrical trees. The percentage of times a node was accurately reconstructed is shown as a proportion of a quarter of a circle in anticlockwise order for Bayesian, maximum likelihood, EW-Parsimony and IW-Parsimony at each node. Accuracy of reconstructions is significantly lower in the 100 character dataset (*a*), and increases in the 350 character (*b*) and 1000 character datasets (*c*). (Online version in colour.)

Bennettitales which resolved as a gnetophyte, and *Caytonia* as sister to the angiosperms.

Analyses of the Luo *et al.* [20] dataset yielded congruent results with the original study, with the placement of *Haramiyavia* outside of crown-Mammalia and multituberculates, although some haramiyids are resolved as crown mammals in the IW-Parsimony analysis (figure 5*a–d*).

Nyasasaurus is recovered as a member of Dinosauria in the maximum likelihood, EW-Parsimony and IW-Parsimony analyses of the dataset from Nesbitt *et al.* [21] (figure 5*e–h*). The Bayesian analysis recovers *Nyasasaurus* in a polytomy with the two major clades of dinosaurs, corroborating the conclusion of Nesbitt *et al.* [21] that, given the data, its precise phylogenetic position is uncertain.

4. Discussion

(a) Simulations indicate that the Bayesian implementation of the Mk model outperforms all other methods and implementations

Previous simulation-based analyses that have attempted to evaluate the performance of likelihood and parsimony-based phylogenetic methods for analysing phenotypic data have found that the probabilistic model performs best [7,8]. However, these studies were biased against parsimony because they employed an unresolved generating tree that is problematic as parsimony methods will attempt to recover a fully resolved tree from the simulated data yielding a non-zero RF distance from the generating tree, even if the two trees are effectively compatible. Furthermore, since previous simulation studies considered the Mk model only within a Bayesian framework, they did not distinguish between the impact of the probabilistic model of character evolution and the statistical framework in which it was implemented.

Our analyses control for these shortcomings of previous simulation studies and show consistently that the Bayesian implementation of the Mk model performs best. In line with previous simulations [8], we found that EW-Parsimony performs better than IW-Parsimony. There is overlap between model performance shown by the distribution of Robinson–Foulds distances (table 1), but there is reason to have different degrees of confidence in the models; only the Bayesian implementation produces a relatively small distribution of tree performance compared with the large tails signifying worse performance in the two parsimony methods (table 1). We also found that the Bayesian implementation of the Mk model outperforms the

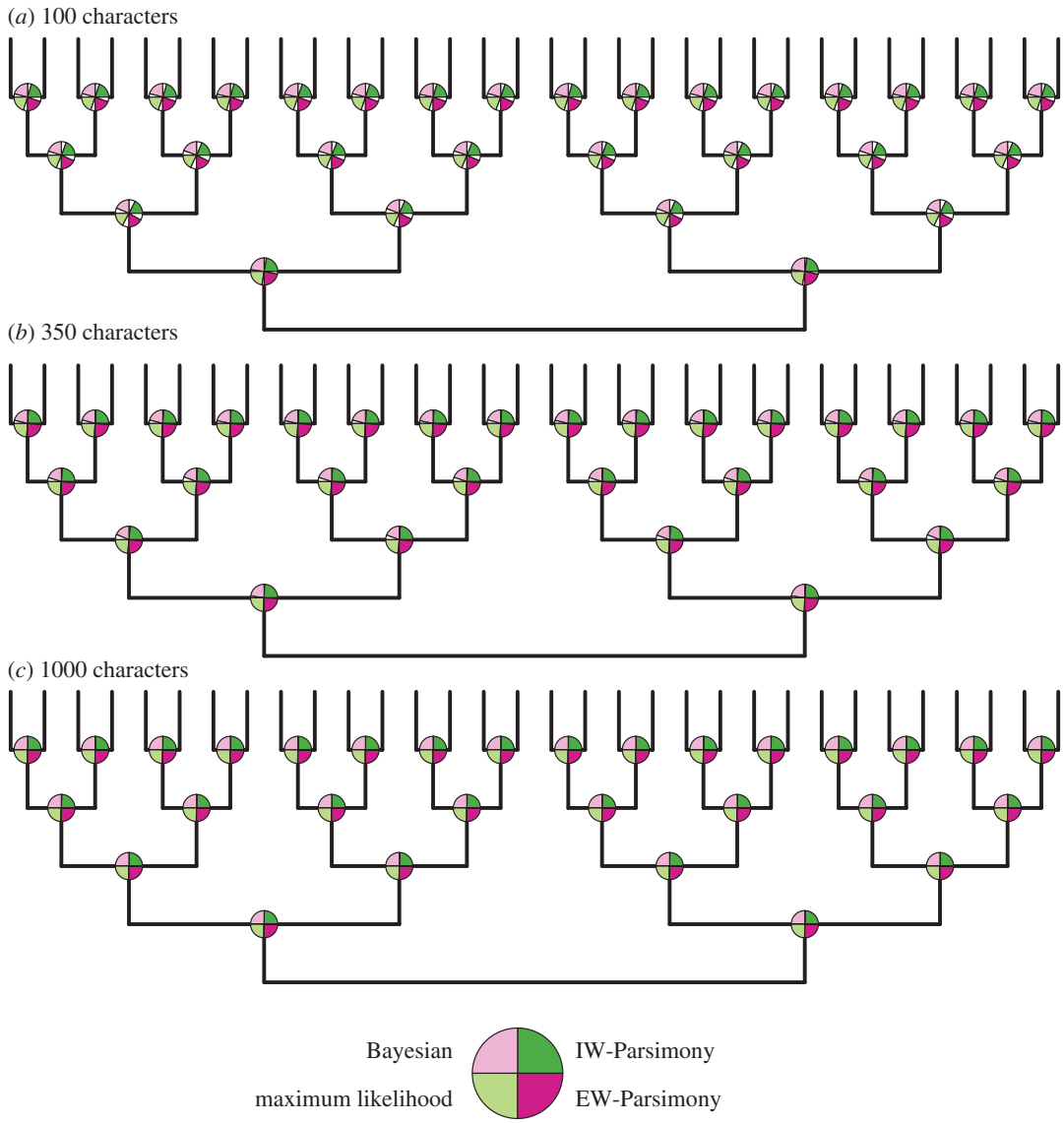


Figure 3. Accuracy of nodes is high for all nodes in the symmetrical phylogeny. The percentage of times a node was accurately reconstructed is shown as a proportion of a quarter of a circle in anticlockwise order for Bayesian, maximum likelihood, EW-Parsimony and IW-Parsimony at each node. Accuracy of reconstructions is high in each dataset size, but there is a non-significant increase in accuracy as dataset size increases (*a*–*c*). (Online version in colour.)

Table 1. Bayesian approaches produce the most accurate trees for all character sets. Mean and range (in brackets) of Robinson–Foulds distances are lower for topologies estimated using Bayesian methods for both the symmetrical and asymmetrical generating tree. Maximum likelihood is the generally the most inaccurate method for the symmetrical generating tree, and implied weights parsimony performs worst for the asymmetrical generating tree.

	equal weights parsimony	implied weights parsimony	maximum likelihood	Bayesian
asymmetrical generating phylogeny				
100	34.89 (22–56)	37.85 (22–56)	45.84 (20–58)	28.1 (18–39)
350	26.57 (11–51)	29.2 (12–51)	26.49 (6–58)	19.21 (7–35)
1000	17.82 (3–40)	19.16 (2–33)	11.94 (0–58)	9.34 (0–31)
symmetrical generating phylogeny				
100	8.08 (0–33)	9.29 (0–29)	10.1 (0–58)	7.51 (0–29)
350	1.33 (0–28)	1.43 (0–28)	1.8 (0–52)	1.2 (0–28)
1000	0.32 (0–26)	0.31 (0–26)	0.51 (0–52)	0.31 (0–26)

maximum-likelihood implementation, indicating that it is not merely the probabilistic transition model that outperforms parsimony methods, but the implementation of the Mk model within a Bayesian statistical framework. Indeed, the maximum-likelihood implementation of the Mk model was the worst-performing method, worse even than IW-Parsimony. In part, the poor performance of the maximum-likelihood-Mk method is because we did not capture phylogenetic uncertainty

Table 2. *p*-Values from Spearman's rank correlation between the percentage of nodes being accurately reconstructed and their distance from the root. Nodes closer to the tips are significantly more likely to be accurately reconstructed in asymmetrical trees but this is not generally true for symmetrical phylogenies.

	asymmetrical tree	symmetrical tree
MB 100	<0.001	0.09919
maximum likelihood 100	<0.001	0.027295
EW 100	<0.001	0.106712
IW 100	<0.001	0.092736
MB 350	<0.001	0.638242
maximum likelihood 350	<0.001	0.057809
EW 350	<0.001	0.19683
IW 350	<0.001	0.148108
MB 1000	<0.001	0.256976
maximum likelihood 1000	<0.001	0.085987
EW 1000	<0.001	0.179186
IW 1000	<0.001	0.287058

associated with this phylogenetic method. This is normally achieved in analyses of molecular datasets through bootstrapping methods, but these are inappropriate for the analysis of phenotypic data as the basic methodological assumption, that the phylogenetic signal is randomly distributed across sites (characters), is not true for morphological data.

However, irrespective of the phylogenetic method used, dataset size correlated positively with both phylogenetic accuracy and resolution, diminishing differences in the relative performance of the competing phylogenetic methods. All phylogenetic methods also performed best when attempting to recover a symmetrical target tree; all methods found recovery of asymmetrical trees challenging and phylogenetic accuracy diminished from tip to root. The impact of tree topology is of particular concern since empirical phylogenetic trees are invariably asymmetric [23], and trees of fossil species are infamous for their asymmetry [24,25]. However, there is a broad spectrum of tree symmetry, with fully symmetric and fully asymmetric trees representing end-members. Palaeontological trees with the dimensions used in our simulations are typically far from the fully asymmetric pectinate-generating tree we employed (Ic = ~0.4 for 32 species) [25]. Furthermore, the asymmetry of many palaeontological trees is often a representational artefact of attempting to summarize character evolution, or an analytic artefact of analysing the relationships among diverse clades based on representative species or higher taxa [26]. Thus, the challenge of recovering trees of extinct taxa may not be as great as a simplistic interpretation of our results might suggest.

(b) Analyses of empirical data bear out conclusions based on simulations

Maximum-likelihood, IW-Parsimony and EW-Parsimony methods of the simulated datasets commonly identify a single optimal tree, but the differences between the optimal trees derived from these methods provides no confidence that any one of the inferred topologies is accurate with reference to the placement of a taxon of interest. This view is corroborated by our reanalysis of empirical datasets which recovered poorly resolved trees using the Bayesian implementation of the Mk model, and in a number of instances, indicate that the conclusions drawn in the corresponding original studies are not supported by the data.

In an extreme example, our re-analyses of the dataset published by Sutton *et al.* [22], which attempted to demonstrate a crown-aplacophoran mollusc affinity for *Kulindroplax*, yielded disparate hypotheses of affinity. EW-Parsimony and IW-Parsimony recovered the published result, while maximum likelihood recovered *Kulindroplax* as a stem-aplacophoran, and Bayesian could not discriminate *Kulindroplax* as a total-group mollusc (figure 4*a*). This poor resolution is unlikely to be a result of poor fossil evidence but, rather, the lack of discriminatory power in the small character matrix. Among the analyses of the dataset from Hilton & Bateman [19], we recovered some of the principal competing topologies that have featured in debate over the affinity of seed plants in past decades. However, the Bayesian analysis of the dataset recovered a topology that is largely unresolved in terms of the relationships among key clades. This suggests that the available data are insufficient to discriminate among the competing hypotheses, and this long-standing debate is largely an artefact of the false resolution of parsimony methods.

Bayesian analyses need not overturn the results from previous analyses based on deterministic phylogenetic methods like EW-Parsimony, IW-Parsimony and maximum likelihood. A phylogenetic position for haramiyids, outside crown-Mammalia, is corroborated by our Bayesian analysis of the dataset from Luo *et al.* [20]—in contrast with the crown-Mammalia affinity recovered for some haramiyids through IW-Parsimony analysis of the same data (figure 5*d*). Similarly, *Nyasasaurus* was posited as the earliest dinosaur, and this conclusion is supported by the Bayesian analyses (figure 5*e*) although this is not supported by EW-Parsimony, IW-Parsimony and maximum-likelihood analyses (figure 5*f–h*). However, the Bayesian analysis is more robust in expressing the phylogenetic ambiguity identified by the original authors [19], as *Nyasasaurus* falls in a polytomy alongside the two major clades of dinosaurs.

Some of the differences between methods may simply reflect the dimensions of the dataset. The two datasets that cannot resolve relationships under Bayesian inference and exhibit significant topological discordance among phylogenetic methods [19,22] are both comparatively small (34 taxa, 48 characters and 48 taxa, 82 characters). These both fall within the scope of simulated datasets that yield low resolution from the Bayesian method and, from other phylogenetic methods, high resolution but low accuracy (figure 1). The two empirical datasets that yield trees with greater congruence from the different phylogenetic methods, are both larger: Luo (114 taxa, 497 characters) and Nesbitt (82 taxa, 413 characters). The size of these matrices is comparable with our simulation results in which we see marked increases in topological accuracy and agreement between methods (figure 1, between 350 and 1000 characters).

(c) Implications for phylogenetic analysis of phenotypic data

The results of our simulation studies indicate that the cadre of phylogenetic hypotheses generated from phenotypic data

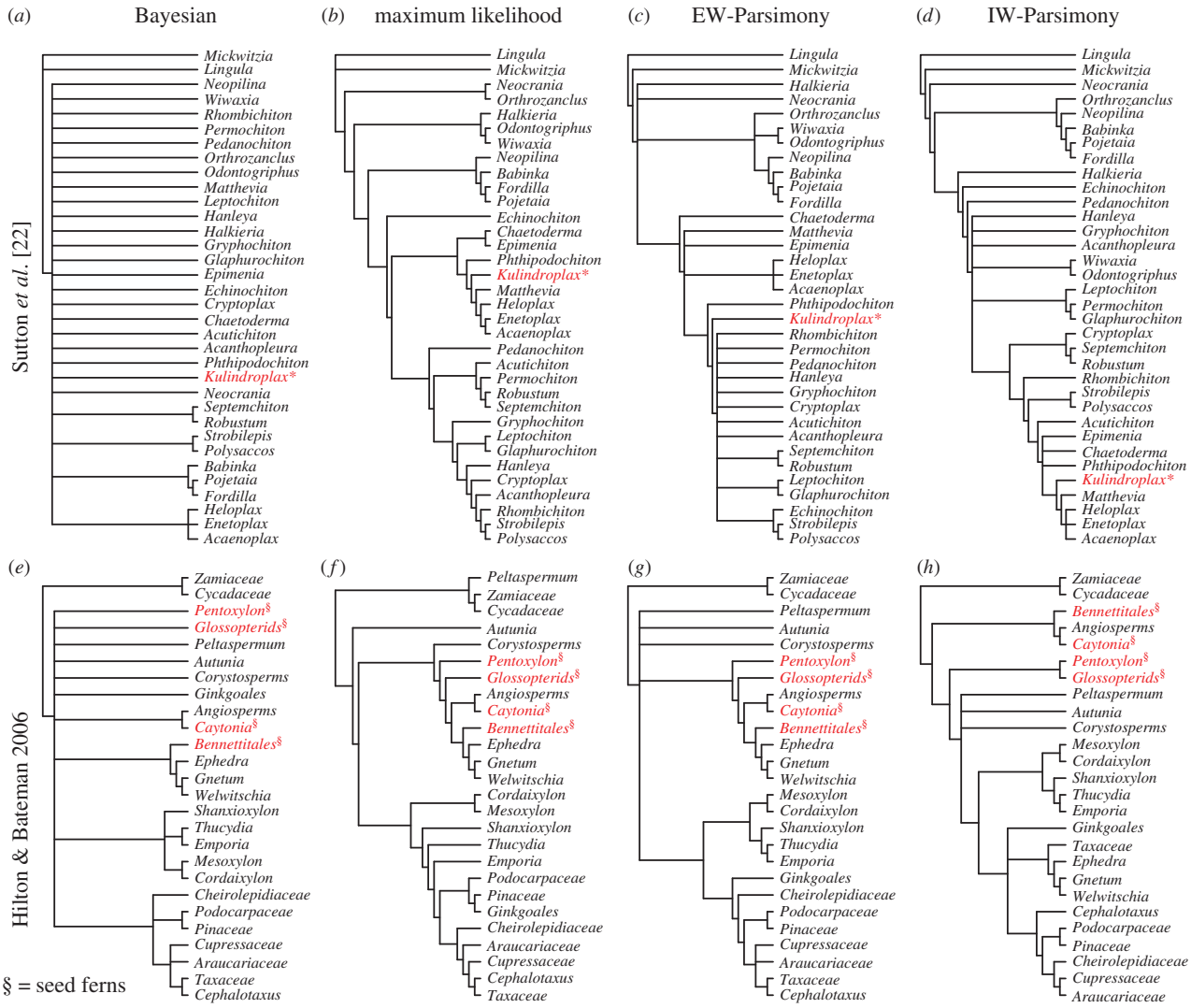


Figure 4. Alternative phylogenetic reconstruction methods alter our understanding of evolution with empirical matrices. However, the relationship of fossil seed ferns from Hilton & Bateman [19] is changed according to implementation (*a–d*), although *Caytonia* remains as sister to angiosperms in all analyses. Alternative analyses change the taxonomic affinity of *Kulindroplax* from Sutton *et al.* [22] (*e–h*). (Online version in colour.)

using parsimony methods require reassessment using the Bayesian implementation of the Mk model. It is likely that many evolutionary interpretations are contingent on precise but inaccurate phylogenetic hypotheses. In this undertaking, it is important that the implications of our simulation studies are considered in the design of phylogenetic studies.

Firstly, phylogenies of fossils tend towards strong asymmetries [25] and, like all phylogenetic methods, Bayesian inference struggles with the recovery of deep nodes within asymmetric trees. Therefore, it is important that outgroups are sampled extensively, ensuring that contentious in-group relationships are closer to the tips, where topological accuracy is highest. Further, in-group lineages should be sampled in a manner that does not accentuate tree asymmetry.

Secondly, phylogenetic accuracy and resolution correlates positively with the relative dimensions of the dataset. Accordingly, phylogenetic resolution or certainty should not be expected from cladistic analyses of small morphological datasets (i.e. those around 100 characters or fewer), particularly if they include fossils. There are finite limits to the number of available phylogenetically informative characters [27] and, for well-studied clades, it may be perceived that these phylogenetically informative characters have already been found. However, it is important to note that the concept of phylogenetic informativeness is different within a likelihood versus a parsimony framework: in parsimony characters that undergo few changes are prized in favour of homoplastic characters. Under the likelihood model, branch length, informed by the number of character changes, contributes to topology estimation. Thus, traditionally 'bad' phylogenetic characters (those exhibiting homoplasy) may find utility in expanding the dimensions of phenotypic character matrices as long as homoplasy falls within the limits that the model can accommodate. In a Bayesian framework, this can be tested using posterior predictive tests of model adequacy (e.g. [28]).

Finally, we may need to alter our expectations to anticipate less well-resolved but more accurate phylogenetic hypotheses, which will both constrain and guide research. Greater resolution may be found by generating matrices suited to likelihood- rather than parsimony-based phylogenetic methods. However, we must also come to terms with the prospect that for some groups of organisms, or their fossil remains, there may be insufficient data. As such, their evolutionary relationships might not therefore be resolvable using morphological data alone and, if they are fossils, their evolutionary significance may never be realized. Nevertheless, resolving phylogenies is not the end game for evolutionary biology.

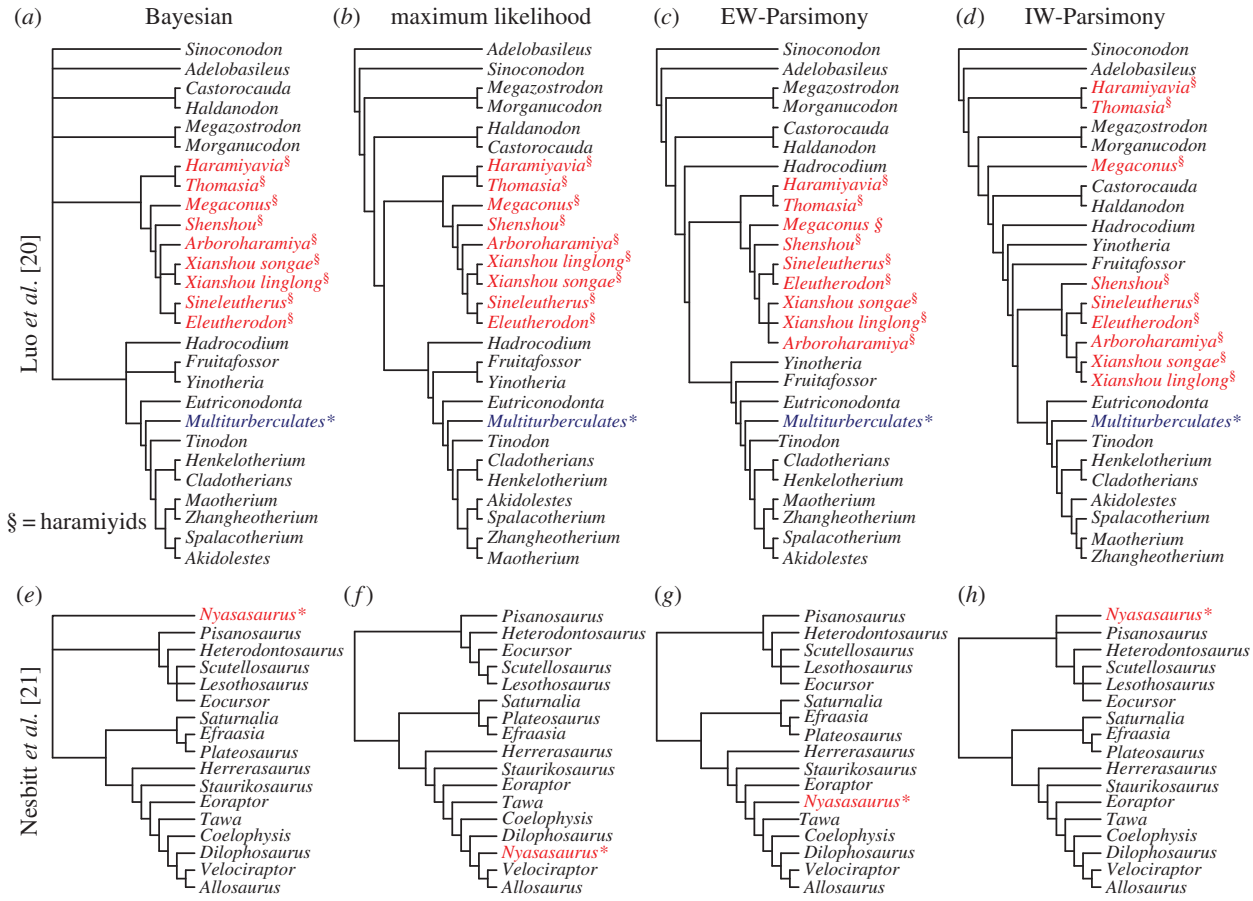


Figure 5. Alternative phylogenetic reconstruction methods produce generally congruent reconstructions of evolution with empirical matrices. For Luo *et al*. [20], the relationship between the haramiyids and multituberculates is largely unchanged across analyses (*a–d*). IW-Parsimony (*g*) and Bayesian analyses place *Nyasasaurus* as close to the earliest dinosaur (*e*) and IW-Parsimony places it close to the earliest diverging taxa (*g*), but EW-Parsimony and maximum likelihood place the taxa as a derived member of Dinosauria (*f*,*h*). (Online version in colour.)

Incompletely resolved trees can still be used as a basis for investigating interesting macroevolutionary questions, and methods exist for incorporating tree uncertainty in phylogenetic comparative methods (e.g. [29]).

5. Conclusion

A growing consensus shows that the Bayesian Mk model is the most accurate method of phylogenetic reconstruction, and here we show that this remains true across dramatically different tree shapes, when analysing datasets composed of both multistate and binary characters, and when compared with maximum-likelihood estimation using the Mk model. We recommend that Bayesian implementations of the Mk model should become the default method for phylogenetic analyses of cladistic morphological datasets, and we should expect low levels of resolution with small datasets. As parsimony methods appear to be less effective than probabilistic approaches, it may be necessary to alter data collection practices by moving away from choosing a selection of characters that undergo few changes, and moving towards scoring all possible characters from the available taxa irrespective of their expected homoplasy.

Data accessibility. Supplementary figures and the code used to simulate the data used in this publication can be accessed in the electronic supplementary material.

Authors' contributions. All authors contributed to the design of the study; M.N.P. and J.E.O.R. led the analyses; interpretation of results and writing was led by M.N.P., J.E.O.R., P.C.J.D. and D.P., though all authors contributed to the interpretation of results and the writing of the manuscript.

Competing interests. We have no competing interests.

Funding. This research was funded by NERC (NE/L501554/1 to J.E.O.R. and L.A.P.; NE/K500823/1 to M.N.P.; NE/L002434/1 to J.F.; NE/N003438/1 to P.C.J.D.), BBSRC (BB/N000919/1 to P.C.J.D.), the University of Bristol (STaR scholarship to A.R.T.), Royal Society Wolfson Research Merit Award (P.C.J.D.) and the John Templeton Foundation (43915 to D.P. and L.H.).

Acknowledgements. We thank the other members of the Bristol Palaeobiology research group for discussion; Rob Asher (Cambridge) and Thomas Guillerme for comments on the draft manuscript. We also thank April Wright and an anonymous reviewer for their help in improving the manuscript.

References

1. Harvey P, May R, Nee S. 1994 Phylogenies without fossils. *Evolution*. **48**, 523–529. (doi:10.2307/2410466)
2. Rabosky DL. 2010 Extinction rates should not be estimated from molecular phylogenies. *Evolution* **64**, 1816–1824. (doi:10.1111/j.1558-5646.2009.00926.x)

3. Losos JB *et al.* 2013 Evolutionary biology for the 21st century. *PLoS Biol*. **11**, e1001466. (doi:10.1371/journal.pbio.1001466)
4. dos Reis M, Donoghue PCJ, Yang Z. 2016 Bayesian molecular clock dating of species divergences in the genomics era. *Nat. Rev. Genet*. **17**, 1–10. (doi:10.1038/nrg.2015.8)
5. Goloboff PA, Carpenter JM, Arias JS, Miranda-Esquivel DR. 2008 Weighting against homoplasy improves phylogenetic analysis of morphological data sets. *Cladistics* **24**, 758–773. (doi:10.1111/j.1096-0031.2008.00209.x)
6. Lewis PO. 2001 A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol*. **50**, 913–925. (doi:10.1080/106351501753462876)
7. Wright AM, Hillis DM. 2014 Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. *PLoS ONE* **9**, e109210. (doi:10.1371/journal.pone.0109210)
8. O'Reilly JE, Puttick MN, Parry L, Tanner AR, Tarver JE, Fleming J, Pisani D, Donoghue PCJ. 2016 Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data. *Biol. Lett*. **12**, 20160081. (doi:10.1098/rsbl.2016.0081)
9. Congreve CR, Lamsdell JC. 2016 Implied weighting and its utility in palaeontological datasets: a study using modelled phylogenetic matrices. *Palaeontology* **59**, 447–462. (doi:10.1111/pala.12236)
10. Holton TA, Wilkinson M, Pisani D. 2014 The shape of modern tree reconstruction methods. *Syst. Biol*. **63**, 436–441. (doi:10.1093/sysbio/syt103)
11. Guillerme T, Cooper N. 2016 Effects of missing data on topological inference using a total evidence approach. *Mol. Phylogenet. Evol*. **94**, 146–158. (doi:10.1016/j.ympev.2015.08.023)
12. Sanderson MJ, Donoghue M. 1996 *The relationship between homoplasy and the confidence in a phylogenetic tree*. San Diego, CA: Academic Press.
13. Goloboff PA, Farris S, Nixon K. 2008 TNT, a free programm for phylogenetic analysis. *Cladistics* **24**, 774–786. (doi:10.1111/j.1096-0031.2008.00217.x)
14. Stamatakis A. 2014 RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313. (doi:10.1093/bioinformatics/btu033)
15. Ronquist F *et al.* 2012 Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol*. **61**, 539–542. (doi:10.1093/sysbio/sys029)
16. Wright AM, Lyons KM, Brandley MC, Hillis DM. 2015 Which came first: the lizard or the egg? Robustness in phylogenetic reconstruction of ancestral states. *J. Exp. Zool. B. Mol. Dev. Evol*. **324**, 504–516. (doi:10.1002/jez.b.22642)
17. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015 IQ-TREE: a fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. *Mol. Biol. Evol*. **32**, 268–274. (doi:10.1093/molbev/msu300)
18. Robinson DF, Foulds LR. 1981 Comparison of phylogenetic trees. *Math. Biosci*. **53**, 131–147. (doi:10.1016/0025-5564(81)90043-2)
19. Hilton J, Bateman RM. 2006 Pteridosperms are the backbone of seed-plant phylogeny. *J. Torrey Bot. Soc*. **133**, 119–168. (doi:10.3159/1095-5674)
20. Luo ZX, Gatesy SM, Jenkins FA, Amaral WW, Shubin NH. 2015 Mandibular and dental characteristics of Late Triassic mammaliaform *Haramiyavia* and their ramifications for basal mammal evolution. *Proc. Natl Acad. Sci. USA* **112**, E7101–E7109. (doi:10.1073/pnas.1519387112)
21. Nesbitt SJ, Barrett PM, Werning S, Sidor CA, Charig AJ. 2013 The oldest dinosaur? A Middle Triassic dinosauriform from Tanzania. *Biol. Lett*. **9**, 20120949. (doi:10.1098/rsbl.2012.0949)
22. Sutton MD, Briggs DEG, Siveter DJ, Siveter DJ, Sigwart JD. 2012 A Silurian armoured aplacophoran and implications for molluscan phylogeny. *Nature* **490**, 94–97. (doi:10.1038/nature11328)
23. Mooers AO, Heard SB. 1997 Inferring evolutionary process from phylogenetics tree shape. *Q. Rev. Biol*. **72**, 31–54. (doi:10.1086/419657)
24. Shao KT, Sokal RR. 1990 Tree balance. *Syst. Zool*. **39**, 266–276. (doi:10.1007/s13398-014-0173-7.2)
25. Harcourt-Brown K, Pearson P, Wilkinson M. 2001 The imbalance of paleontological trees. *Paleobiology* **27**, 188–204. (doi:10.1666/0094-8373(2001)027<0188:TIOPT>2.0.CO;2)
26. Panchen A. 1982 The use of parsimony in testing phylogenetic hypotheses. *Zool. J. Linn. Soc*. **74**, 305–328. (doi:10.1111/j.1096-3642.1982.tb01154.x)
27. Scotland RW, Olmstead RG, Bennett JR. 2003 Phylogeny reconstruction: the role of morphology. *Syst. Biol*. **52**, 539–548. (doi:10.1080/10635150390223613)
28. Tarver JE *et al.* 2016 The interrelationships of placental mammals and the limits of phylogenetic inference. *Genome Biol. Evol*. **8**, 330–334. (doi:10.1093/gbe/evv261)
29. Healy K *et al.* 2014 Ecology and mode-of-life explain lifespan variation in birds and mammals. *Proc R. Soc. B* **281**, 20140298. (doi:10.1098/rspb.2014.0298)