THE UNIVERSITY OF CHICAGO

NOVEL APPROACHES TO THE PARAMETRIC INFERENCE OF PHYLOGENIES AND DIVERSIFICATION RATES FROM FOSSIL DATA

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF THE GEOPHYSICAL SCIENCES

BY
DAVID ČERNÝ

CHICAGO, ILLINOIS
AUGUST 2024

To my family, without whom none of this would have been possible.

[A]lle Wissenschaft wäre überflüssig, wenn die Erscheinungsform und das Wesen der Dinge unmittelbar zusammenfielen[.]

K.M., 1864–1865.

La conoscenza umana avanza per rivoluzione. La conoscenza umana avanza per rivoluzioni sociali. Il resto è silenzio.

A.B. / P.C.I., 1960.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

A doctorate is not an individual achievement, but the outcome of a collective effort. With that in mind, I would like to express my profound gratitude to my advisor, Graham Slater: I could not have wished for a better one. This is no less true of the remaining members of my committee, Tracy Heath, Michael Foote, and Richard Ree, who have been incredibly supportive throughout the entire process of completing this dissertation. If I regret anything at all about asking them to take on this role, it is that I have perhaps not sought their guidance as often as I should have. I would like to thank my co-authors, especially Daniel Madzia, Rossy Natale, Ashley Simonoff, and Paul van Els, mostly for being extremely patient with me. Dan, of course, has earned his place in these acknowledgments for far more than that, since it was he who got me interested in paleobiology all those years ago when I was still blissfully unaware of what an MCMC sampler was. I want to further thank those collaborators with whom I had a chance to work on projects that, for various reasons, ultimately haven't resulted in a publication – at least not by the time I'm writing these lines. These include Tetsuto Miyashita, Orlando Schwery, and Juned Zariwala. A special mention must definitely go to Michael Alfaro, who not only made a computational biologist out of me in the first place, but who also first pointed out to me that the University of Chicago and the Slater Lab might be a great match for my research interests. I am deeply grateful to Joëlle Barido-Sottani and Walker Pett, two former members of the Heath Lab, who patiently taught me the basics of RevBayes and of software development in general. I might still not be particularly good at it, but what I've learned I consider extremely valuable. Speaking of RevBayes, I would be remiss not to mention the entire development team, especially Sebastian Höhna, Ben Redelings, April Wright, Rachel Warnock, Michael Landis, Andy Magee, and Jiansi Gao, with whom I have not only had many fascinating discussions over the years, but who have also made my research so much easier with their many contributions ranging from bug fixes to conceptual clarification.

I'd like to thank all of my friends, who have made my graduate school experience better in more ways than I can count – whether they be on the second floor of Hinds, elsewhere at UChicago, or completely free of any connections to that institution. I have really enjoyed the time we've spent together, from trivia games at The Pub to late-night conference discussions about overcoming nihilism. I would also like to express my gratitude to Graduate Students United (GSU) and, more broadly, to the United Electrical, Radio and Machine Workers of America (UE). It has been an honor to organize with them for better conditions in our own workplace, and to engage in solidarity efforts with other workers' struggles. I am very satisfied that after six years, I am leaving a university where graduate workers have a legally recognized union and a collective bargaining agreement. Finally, and most importantly of all, I would like to thank my family, who has always supported me unconditionally in my pursuit of higher education, no matter how impractical my choice of field may have seemed. I can only repeat myself: *bez vás bych tohle nikdy nemohl dokázat.*

ABSTRACT

Fossils offer a unique window into the history of life on Earth, but incorporating them into algorithmic approaches for the study of macroevolutionary patterns and processes has been fraught with difficulty. Morphological data are much more difficult to collect at scale than DNA sequences, limiting the size of phylogenetic trees that can be inferred for extinct organisms, and the methods used to analyze them have long lagged behind those that are routinely applied to molecular datasets in terms of sophistication. Here, I aim to alleviate these issues by developing several novel parametric approaches to the inference of fossil phylogenies and their use in downstream analyses of macroevolutionary rates. In Chapter 1, I present a new method called Bayesian Least-Squares Supertrees (BLeSS), which takes as its input a profile of previously inferred time-calibrated trees with partially overlapping leaf sets, and returns a posterior distribution of time-calibrated supertrees on the union of the leaf sets of the individual source trees. I provide a fast and flexible implementation of BLeSS in a pre-existing software environment, and validate it using a recently introduced simulation-based pipeline. Using both synthetic and empirical datasets, I identify the conditions under which BLeSS performs well, demonstrate its ability to handle supertrees of up to 1000 leaves, and provide guidelines for its use as well as suggestions for further extensions and performance boosts. In Chapter 2 (published as Černý and Simonoff, 2023), I adopt a protocol originally developed for phylogenetic analyses of genomic data, and repurpose it for the study of morphological characters to address a controversial problem in fossil phylogenetics: the earliest divergence among dinosaurs. I show that despite substantial differences in character coding, which were thought to be of critical importance by previous studies, neither of the two examined datasets significantly favors any of the three plausible hypotheses over the other two. Moreover, each hypothesis is supported by a nearly equal number of characters, including characters with high phylogenetic signal. This suggests the datasets in question suffer from a high degree of internal conflict, and are inadequate for resolving the problem

they were designed to address. In Chapter 3 (published as Černý et al., 2022), I use empirical and synthetic data to evaluate the performance of two recently introduced Bayesian methods for the inference of macroevolutionary rates from fossil phylogenies and occurrence records. I show that their ability to tease apart shifts in the rates of speciation, extinction, and fossil sampling is limited, possibly because of the paucity of data available to vertebrate paleontologists. In particular, the two approaches are unable to conclusively answer the question of whether non-avian dinosaurs entered a period of decline prior to their extinction at the K–Pg boundary. I outline best practices for informed use of model-based diversification rate estimation in extinct clades, and suggest how it can be profitably combined with more traditional paleobiological approaches.

INTRODUCTION

Over the last forty years, the field of phylogenetics has experienced a rapid growth and development, driven in large part by technological advances in DNA sequencing, which made it possible to collect and analyze enormous quantities of previously inaccessible data, and by the advent of modern comparative methods (Felsenstein, 1985), which made phylogenies essential to fields as diverse as ecology, behavior, and physiology. Paleontology has long been isolated from these developments, despite the fact that many questions that are routinely investigated in a phylogenetic context today were originally posed in paleontological settings (Simpson, 1944; Raup et al., 1973; Raup, 1985; Foote, 1996). In part, this isolation may have been due to the near-universal reliance of paleontologists on simple nonparametric phylogenetic methods, whose use tends to be justified in logical and philosophical rather than statistical terms. Moreover, these methods do not produce the time-calibrated phylogenies that represent the key input for comparative methods. However, the opposite problems also occurs, and existing comparative methods are occasionally implemented in ways that presuppose extant-only trees (Slater, 2014), thus posing an additional barrier to the integration of fossil and phylogenetic data in a single coherent statistical framework.

This situation started to change with the introduction of the first substitution model for discrete morphological characters by Lewis (2001), which opened up the possibility of parametric inference of fossil phylogenies. This development was accelerated by the advent of Bayesian tip-dating, which unifies extinct and extant lineages in a principled manner (Ronquist et al., 2012a; Gavryushkina et al., 2014; Heath et al., 2014; Wright et al., 2022). The resulting time-calibrated phylogenies, which accommodate uncertainty in both the shape of the tree (topology) and the ages of its nodes, have in turned allowed researchers to address a variety of macroevolutionary questions (Slater, 2015; Wisniewski et al., 2022), particularly those related to diversification rates (Mitchell et al., 2018; Crouch, 2020; Lloyd and Slater, 2021). Despite these advances, however, the development of parametric approaches to fossil-

based phylogenetics and diversification rate estimation still lags behind the rich variety of techniques that are available for the analysis of extant taxa. Here, I attempt to take several steps toward mitigating this problem.

In Chapter 1, I address one of the key problems with the application of comparative methods and diversification rate estimation to extinct lineages: the lack of sufficiently large time-scaled phylogenies. This, in turn, stems from the fact that morphological datasets are much more difficult to combine into "supermatrices" than molecular sequence alignments. I develop a novel Bayesian method, BLeSS, which circumvents this problem by using a supertree approach instead. Unlike most existing supertree techniques, BLeSS can accommodate both branch lengths (assumed to represent branch durations in units of calendar time) and topological uncertainty. While my demonstration of the method focuses on ultrametric trees, little prevents its extension to phylogenies with non-contemporaneous tips.

Chapter 2 (Černý and Simonoff, 2023) represents a novel application of methods that were originally introduced to investigate the distribution of support for competing topologies in phylogenomic data to paleontological character matrices. When used to evaluate alternative hypotheses about higher-level dinosaur phylogeny, these likelihood ratio-based approaches reveal a large amount of conflict among the individual characters included in the examined datasets, and the resulting inability to decide in favor of one topology over another. Notably, these conclusions are not affected by attempts to fix errors in character coding, which have been considered to be of great importance in the surrounding debate. The uncertainty may have deeper causes, and its resolution may require changes to both data collection practices (subjective character delimitation) and phylogenetic methodology.

Chapter 3 (Černý et al., 2022) investigates the performance of two recently developed frameworks for model-based estimation of macroevolutionary rates, PyRate (Silvestro et al., 2014b) and Fossil BAMM (Mitchell et al., 2018). The two methods are shown to yield conflicting results when applied to the controversial question of a potential diversity decline of

non-avian dinosaurs prior to their eventual extinction at the Cretaceous–Paleogene boundary. Moreover, a set of simulations designed to mimic paleobiologically realistic conditions reveals a variety of problems with both approaches, including some that may be indicative of underlying identifiability problems (e.g., correlated shifts in the rates of speciation and extinction). We provide practical guidelines for interpreting the results yielded by the two methods in light of their limitations, and point out potential causes of the observed problems.

CHAPTER 1

BAYESIAN LEAST-SQUARES SUPERTREES (BLESS): FLEXIBLE INFERENCE OF LARGE TIME-CALIBRATED PHYLOGENIES

1.1 Abstract

Time-calibrated phylogenies are key to macroevolutionary hypothesis testing and parameter inference, but their estimation is difficult when the number of tips is large. Despite its attractive properties, the joint Bayesian inference of topology and divergence times remains computationally prohibitive for large supermatrices. Historically, supertrees represented a popular alternative to supermatrix-based phylogenetic methods, but most of the existing supertree techniques do not accommodate branch lengths or topological uncertainty, rendering them unfit to supply input for modern comparative methods. Here, we present Bayesian Least-Squares Supertrees (BLeSS), a new approach that takes a profile of time trees with partially overlapping leaf sets as its input, and returns the joint posterior distribution of supertree topologies and divergence times as its output. Building upon the earlier exponential error model and average consensus techniques, BLeSS transforms the profile into path-length distance matrices, computes their arithmetic average, and uses MCMC to sample time-calibrated supertrees according to their least-squares fit to the average distance matrix. We provide a fast, flexible, and validated implementation of BLeSS in the program RevBayes, and test its performance using a comprehensive set of simulations. We show that the method performs well across a wide range of conditions, including variation in missing data treatment and the steepness of the error function. Finally, we apply BLeSS to an empirical dataset comprising 33 time trees for 260 species of carnivorans, illustrating its ability to recover well-supported clades and plausible node ages, and discuss how the method can best be used in practice, outlining possible extensions and performance boosts.

1.2 Introduction

Time trees, or phylogenetic trees with branch lengths scaled to calendar time, are widely used to investigate questions ranging from trait evolution (Ree, 2005; Slater and Harmon, 2013; Hopkins and Smith, 2015; Slater, 2015; Caetano and Harmon, 2018) to diversification rate estimation (Nee et al., 1994; Rabosky, 2014; Maliet et al., 2019; Barido-Sottani et al., 2020) and historical biogeography (Ree and Smith, 2008; Landis et al., 2013; Meseguer et al., 2014). Over the last two decades, joint Bayesian inference of tree topology and divergence times (Drummond et al., 2006; Ronquist et al., 2012a) has become the preferred method for producing time trees thanks to its ability to synthesize multiple types of evidence and accommodate the different sources of uncertainty that are associated with them (dos Reis et al., 2016). Unfortunately, due to the intrinsic difficulty of sampling from complex posterior distributions, Bayesian co-estimation of topology and node ages remains computationally unfeasible for datasets comprising thousands or tens of thousands of tips, which are increasingly used to shed light on the patterns and processes underlying major radiations across the tree of life (Jetz et al., 2012; Rabosky et al., 2018; Upham et al., 2019; Janssens et al., 2020).

Two alternatives have emerged to the intractable fully Bayesian inference of very large time trees. One is the assembly of a comprehensive molecular dataset, or "supermatrix" (Delsuc et al., 2005; de Queiroz and Gatesy, 2007), followed by a time-free analysis employing a less computationally demanding method, such as a heuristic maximum likelihood search. The resulting phylogram is then time-scaled using nonparametric or semiparametric rate smoothing (Rabosky et al., 2018; Janssens et al., 2020). The other approach involves dividing the clade of interest into a number of non-overlapping subclades, each of which is sufficiently small for fully Bayesian divergence time estimation to remain feasible. The full-sized tree is then assembled by attaching the taxonomically comprehensive subclade phylogenies to a sparsely sampled backbone in which every subclade is represented by a single pair of tips. A pseudoposterior distribution of topologies and divergence times can be obtained by repeating

the attachment step on multiple samples from the subclade and backbone posteriors (Jetz et al., 2012; Tonini et al., 2016; Upham et al., 2019). Though occasionally contrasted with more traditional "supertree" methods (Upham et al., 2019, 2021), this "backbone-and-patch" approach can in fact be viewed as their special case, since the final tree is in both cases assembled from previously inferred source trees rather than estimated directly from character data (Sanderson et al., 1998; Bininda-Emonds et al., 2002; Wilkinson et al., 2005).

Despite their widespread use, both the *post hoc* time-scaling of phylograms and the backbone-and-patch approach present important shortcomings. The former method lacks some of the key properties of fully Bayesian time tree estimation, including the estimation of divergence times jointly with (rather than conditional on) tree topology, and the ability to accommodate various sources of uncertainty, such as variation in branch length estimates and fossil calibration error (Smith and O'Meara, 2012). The latter framework retains the advantages of fully Bayesian inference at the level of source trees, but fails to propagate uncertainty throughout the tree assembly process as a result of the limited overlap between the backbone and the patch trees. Moreover, the delimitation of patch clades requires potentially controversial monophyly assumptions (Lloyd and Slater, 2021), and the pseudoposterior produced by the method is not a statistically valid approximation of the full posterior, since the likelihoods of the backbone and patch trees are treated as independent despite being partly derived from the same set of sequences (Álvarez-Carretero et al., 2022).

The shortcomings of the backbone-and-patch framework can be overcome in a general supertree setting; however, existing supertree methods – including the widely used matrix representation approaches (Baum and Ragan, 2004; Ross and Rodrigo, 2004; Nguyen et al., 2012) – suffer from important drawbacks of their own. First, their ability to compete with alternative approaches to the inference of very large phylogenies is hampered by a lack of conceptual clarity on whether supertrees should be viewed merely as summaries of the information contained in the source trees, or estimates in their own right (Steel and Ro-

drigo, 2008). This has implications for whether supertree methods should attempt to resolve topological conflict, or merely represent it as a lack of resolution (the "voting" vs. "veto" approaches; Ranwez et al., 2007), and whether it is acceptable for the supertree to contain bipartitions not present in any of its source trees (Cotton et al., 2006; Wilkinson et al., 2007). Second, most supertrees do not come with branch lengths, which makes them of limited use as input for phylogenetic comparative methods. Branch lengths in units of expected substitutions per site or calendar time can be added to a supertree *post hoc* using sequence data (Binet et al., 2016; Kimball et al., 2019) or stratigraphic age ranges associated with the tips (Sakamoto et al., 2016), but few supertree methods are capable of estimating topology and branch lengths simultaneously, without relying on an increasingly tenuous chaining together of inferences (though see Willson, 2004; Lapointe and Levasseur, 2004; and Criscuolo et al., 2006 for exceptions). Third, most existing supertree methods yield point estimates without any measure of uncertainty. While extensions of the nonparametric bootstrap to the supertree setting have been proposed (Burleigh et al., 2006), they suffer from a variety of problems, including the impossibility of ensuring that boostrap pseudoreplicates will have the same leaf set as the original supertree (Bininda-Emonds, 2014). Bayesian posterior probabilities represent an attractive, easily interpretable alternative to bootstrap support values (Alfaro and Holder, 2006), but their use in supertree estimation remains uncommon (Ronquist et al., 2004; Akanni et al., 2015; Karcher et al., 2021).

Here, we present a new approach to supertree inference, the Bayesian Least-Squares Supertrees (BLeSS), which takes a profile of time trees with overlapping leaf sets as its input, and returns a posterior distribution of time-calibrated supertrees as its output. BLeSS can be used in a manner similar to other supertree methods, with each dataset providing a single source tree representing a point estimate of topology and divergence times. However, the framework can also accommodate uncertainty about source tree topology by including the entire posterior distribution estimated from a given dataset in the profile, as proposed by

Ronquist et al. (2004) and similar to the "metatree" pipeline of Lloyd and Slater (2021). Unlike the backbone-and-patch approach, it is capable of efficiently propagating this uncertainty into the posterior supertree distribution by allowing for an arbitrary degree of overlap between the source trees. Re-use of the same data by multiple source trees is allowed but not required, and can be addressed in a nonparametric fashion by means of optional downweighting of non-independent source trees, without treating each one as providing an independent contribution to the likelihood term. Moreover, the use of a birth-death tree prior allows BLeSS to incorporate other sources of information in addition to the profile, such as node calibrations or topological constraints. We validate our implementation of BLeSS using simulation-based calibration checking, employ a comprehensive simulation scheme to assess its accuracy and precision, benchmark its performance across a wide range of supertree sizes, and finally demonstrate its utility by inferring a time-calibrated supertree for extant members of the mammalian order Carnivora.

1.3 Materials and Methods

1.3.1 Theory

The BLeSS method combines elements from the average consensus (Lapointe and Levasseur, 2004) and maximum likelihood (Steel and Rodrigo, 2008) approaches to supertree inference, and consists of two main components: an average distance matrix summarizing the topological and divergence time information contained in a profile of source trees, and an exponential error model used to evaluate the likelihood of a proposed supertree conditional on this matrix. Like any probabilistic model, the approach outlined here can be used in both maximum-likelihood and Bayesian settings. Here, we choose to focus on the Bayesian implementation, which provides a natural way of quantifying the uncertainty associated with parameter estimates in the form of the posterior distribution. In addition, previous at-

tempts to use the exponential error model for supertree inference revealed that the sampling techniques used for Bayesian inference were more efficient than the optimization techniques employed by maximum likelihood, making the former approach better suited for analyses of large datasets (Akanni et al., 2014, 2015).

Average distance matrix

Consider a profile of source trees $\mathcal{P} = \{\mathcal{T}_1, \ldots, \mathcal{T}_K\}$ defined over a set of leaves $\mathcal{L}$, where $\mathcal{L}(\mathcal{T}_k)$ is the leaf set of tree $\mathcal{T}_k$ such that $\mathcal{L}(\mathcal{T}_k) \subseteq \mathcal{L}$. Each tree is assumed to be ultrametric, with branch lengths in units of calendar time. We exploit the fact that each tree $\mathcal{T}_k$ can be converted to a unique path-length distance matrix $\mathbf{D}_k$ (Lapointe and Levasseur, 2004), such that the path-length distance $d_k(i, j)$ of any two leaves i and j is computed by taking the sum of the lengths of the intervening branches, and corresponds to twice the time elapsed since their divergence. If we denote by $\pi_{i,j}(\mathcal{P})$ the (possibly empty) set of indices of all trees in $\mathcal{P}$ that contain both i and j, that is, $\pi_{i,j}(\mathcal{P}) = \{k \in \{1, \ldots, K\} \ : \ i \in \mathcal{L}(\mathcal{T}_k) \wedge j \in \mathcal{L}(\mathcal{T}_k)\}$, the *average distance matrix* $\bar{\mathbf{D}}$ can be calculated element-wise as follows:

$$\bar{d}(i,j) = \begin{cases} \dfrac{1}{|\pi_{i,j}(\mathcal{P})|} \displaystyle\sum_{k \in \pi_{i,j}(\mathcal{P})} d_k(i,j) & \text{if } \pi_{i,j}(\mathcal{P}) \neq \varnothing \\ \text{undefined} & \text{if } \pi_{i,j}(\mathcal{P}) = \varnothing \end{cases} \tag{1}$$

The undefined entries in the resulting matrix, corresponding to missing distances, can either be treated as such and disregarded in subsequent steps (the "direct" approach of Levasseur et al., 2003; see Makarenkov and Leclerc, 1999; Kettleborough et al., 2015), or they can be filled in with estimated distances using a variety of imputation schemes (the "indirect" approach; see Landry and Lapointe, 1997 and the references in Criscuolo and Gascuel, 2008).

Compared to related methods (Lapointe and Levasseur, 2004; Criscuolo et al., 2006), several simplifying assumptions are made here. The distances $d_k(i,j)$ are obtained by transforming trees rather than estimated from sequence data, and as such are guaranteed to be ultrametric. Branch length heterogeneity was found to negatively impact the accuracy of supertree inference by Lapointe and Levasseur (2004), leading to attempts to "standardize" (Lapointe and Levasseur, 2004) or "deform" (Criscuolo et al., 2006) individual distance matrices by using a linear combination of the original distances to minimize the among-matrix least-squares difference. This step is unnecessary here, since all source trees are assumed to have branch lengths measured in time units (e.g., millions of years), and as long as these units are consistent, calendar time itself provides a natural common scale for all path-length matrices.

Exponential error model

Once $\bar{\mathbf{D}}$ has been calculated, the goal is to find a supertree $\hat{\mathcal{T}}$ with the full leaf set $\mathcal{L}$ that minimizes the least-squares difference between its own path-length matrix $\hat{\mathbf{D}}$ and $\bar{\mathbf{D}}$:

$$\delta(\hat{\mathbf{D}}, \bar{\mathbf{D}}) = \sum_{i=1}^{|\mathcal{L}|} \sum_{j=1}^{|\mathcal{L}|} \left[\hat{d}(i,j) - \bar{d}(i,j) \right]^2 \tag{2}$$

However, the problem of minimizing $\delta(\hat{\mathbf{D}}, \bar{\mathbf{D}})$ is NP-hard (Day, 1987; Makarenkov and Lapointe, 2004). Consequently, attempts to estimate supertrees from the average distance matrix, or more generally, to infer phylogenies from distance matrices, have generally relied on heuristic optimization to obtain a point estimate of $\hat{\mathcal{T}}$ (Makarenkov and Leclerc, 1999; Lapointe and Levasseur, 2004). In contrast to these earlier approaches, BLeSS recasts the problem as one of sampling the posterior distribution of $\hat{\mathcal{T}}$, and accordingly requires an explicit error model to estimate the likelihood of a given supertree given the average distance matrix. To this end, Steel and Rodrigo (2008) proposed the following exponential error model:

$$P(\mathcal{T}_k \,|\, \hat{\mathcal{T}}, \beta_k) = \alpha_k \exp\Big[- \beta_k \, \delta\big(\mathcal{T}_k, \, \hat{\mathcal{T}}|\mathcal{L}(\mathcal{T}_k)\big)\Big], \tag{3}$$

where $\hat{\mathcal{T}}|\mathcal{L}(\mathcal{T}_k)$ is the restriction of $\hat{\mathcal{T}}$ to the leaf set of source tree $\mathcal{T}_k$, and β_k is a tree-specific positive constant that can be assigned to each source tree to reflect data quality, or inferred using maximum likelihood. This model is not immediately applicable to BLeSS, since it evaluates the likelihood of $\hat{\mathcal{T}}$ conditional on the untransformed tree profile $\mathcal{P}$ rather than the average distance matrix $\bar{\mathbf{D}}$. Moreover, since the model was originally developed to estimate supertree topologies rather than supertrees with branch lengths, δ was assumed to be a discretely varying metric (such as the Robinson-Foulds distance; Robinson and Foulds, 1981), and $P(\,\cdot\,)$ was assumed to be a discrete probability distribution. This requires the inclusion of a normalizing term α_k, which proved to be dependent on the shape of $\hat{\mathcal{T}}$ (Bryant and Steel, 2009). Evaluating α_k is extremely computationally expensive, and while it is possible to choose the β_k coefficients in such a way that the ranking of possible supertrees in terms of their likelihood remains unaffected by α_k, this property is only sufficient for maximum likelihood inference (Bryant and Steel, 2009). However, previous attempts to perform maximum likelihood estimation under the model of Steel and Rodrigo (2008) proved incapable of efficient tree searches for large values of $|\mathcal{L}|$ (Akanni et al., 2014), whereas the Bayesian MCMC implementation of Akanni et al. (2015) performed better but could no longer be guaranteed to be valid given its reliance on inaccurately normalized likelihoods (Bryant and Steel, 2009).

Here, we overcome the normalization problem by replacing the discrete distribution describing the probability of observing tree $\mathcal{T}_k$ with a density function describing the probability of observing an average distance matrix $\bar{\mathbf{D}}$, using a continuously varying metric δ corresponding to Eq. 2:

$$
\begin{aligned}
f(\bar{\mathbf{D}} \,|\, \hat{\mathcal{T}}, \lambda_e) &= \lambda_e e^{-\lambda_e \delta(\hat{\mathbf{D}}, \bar{\mathbf{D}})} \\
&= \lambda_e \exp\left\{ -\lambda_e \sum_{i=1}^{|\mathcal{L}|} \sum_{j=1}^{|\mathcal{L}|} \left[\hat{d}(i,j) - \bar{d}(i,j) \right]^2 \right\}
\end{aligned}
\tag{4}
$$

The rate parameter λ_e of the exponential determines how strongly trees with larger values of δ will be penalized. Conveniently, actual supertree evaluation is likely to take place in terms of log-likelihoods, which are simply proportional (up to a multiplicative and additive constant) to the raw least-squares difference between the average distance matrix and the path-length distance matrix of the proposed supertree:

$$
\log f(\bar{\mathbf{D}} \,|\, \hat{\mathcal{T}}, \lambda_e) = -\lambda_e \sum_{i=1}^{|\mathcal{L}|} \sum_{j=1}^{|\mathcal{L}|} \left[\hat{d}(i,j) - \bar{d}(i,j) \right]^2 + \log \lambda_e \tag{5}
$$

Using the likelihood function above, Bayesian inference becomes possible if an appropriate prior is specified. Since the estimated supertree $\hat{\mathcal{T}}$ is itself a time tree, any joint prior distribution over topologies and node heights can be used for this purpose, including the phenomenological "uniform" prior of Ronquist et al. (2012a) or priors induced by the birth-death family of models (Gernhard, 2008). A realistic case of BLeSS supertree estimation under a birth-death prior with incomplete sampling and a uniform hyperprior on the time of origin is shown in Fig. 1.1. If the parameters of the birth-death process are jointly denoted as $\mathbf{\Phi}$ and assumed to be estimated from the data along with the supertree, the joint posterior distribution of $(\mathcal{T}, \mathbf{\Phi})$ is given by

$$
P(\mathcal{T}, \mathbf{\Phi} \,|\, \bar{\mathbf{D}}, \lambda_e) = \frac{f(\bar{\mathbf{D}} \,|\, \lambda_e, \mathcal{T})\, P(\mathcal{T} \,|\, \mathbf{\Phi})\, P(\mathbf{\Phi})}{\displaystyle\int_{\mathcal{T}} f(\bar{\mathbf{D}} \,|\, \lambda_e, \mathcal{T}) \left[\int_{\mathbf{\Phi}} P(\mathcal{T} \,|\, \mathbf{\Phi})\, P(\mathbf{\Phi})\, \mathrm{d}\mathbf{\Phi} \right] \mathrm{d}\mathcal{T}}. \tag{6}
$$

Alternatively, instead of setting λ_e to a fixed value as in Fig. 1.1, one can also assign it a prior distribution and use MCMC sampling to marginalize it out of the model:

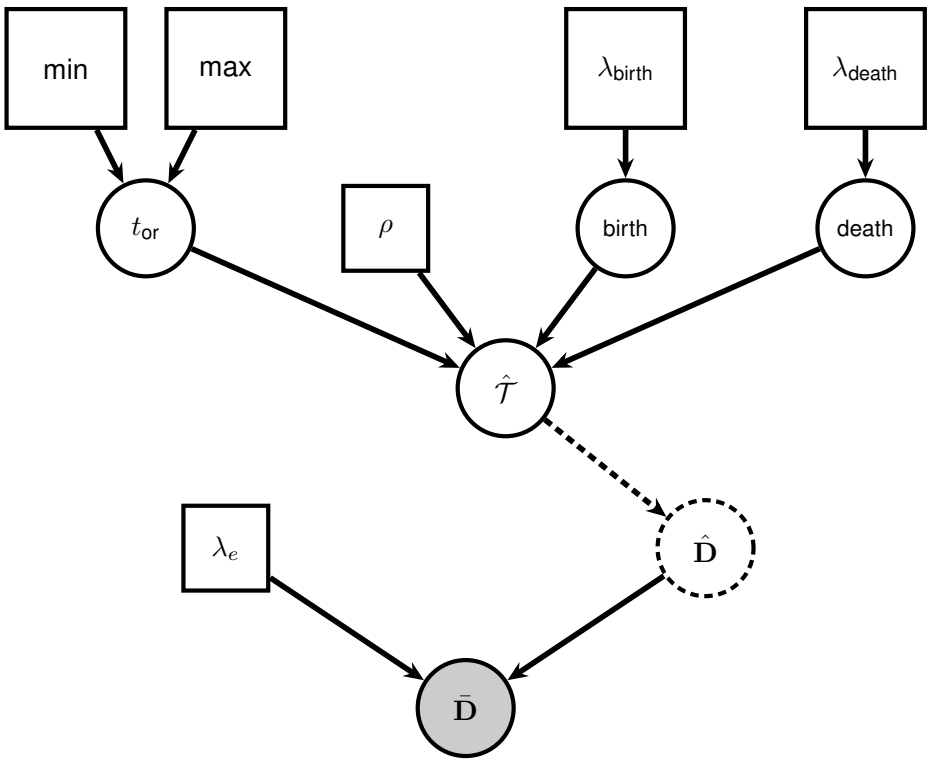

Figure 1.1: Probabilistic graphical model representation of a hypothetical BLeSS analysis. Notation after Höhna et al. (2014). t_{or} = time of origin of the birth-death process; ρ = extant sampling fraction; λ_{birth} = rate of the exponential hyperprior on the birth rate; λ_{death} = rate of the exponential hyperprior on the death rate.

$$
\begin{aligned}
P(\mathcal{T}, \mathbf{\Phi} \,|\, \bar{\mathbf{D}}) &= \int_{\lambda_e} P(\mathcal{T}, \mathbf{\Phi}, \lambda_e \,|\, \bar{\mathbf{D}}) \, \mathrm{d}\lambda_e \\
&= \frac{\displaystyle\int_{\lambda_e} f(\bar{\mathbf{D}} \,|\, \lambda_e, \mathcal{T}) \, P(\lambda_e) \, P(\mathcal{T} \,|\, \mathbf{\Phi}) \, P(\mathbf{\Phi}) \, \mathrm{d}\lambda_e}{\displaystyle\int_{\lambda_e} \int_{\mathcal{T}} f(\bar{\mathbf{D}} \,|\, \lambda_e, \mathcal{T}) \, P(\lambda_e) \left[\int_{\mathbf{\Phi}} P(\mathcal{T} \,|\, \mathbf{\Phi}) \, P(\mathbf{\Phi}) \, \mathrm{d}\mathbf{\Phi} \right] \mathrm{d}\mathcal{T} \, \mathrm{d}\lambda_e}.
\end{aligned} \tag{7}
$$

1.3.2 Implementation

We implemented BLeSS in RevBayes (Höhna et al., 2016), open-source Bayesian phylogenetics software with C++ backend and an R-like scripting language called Rev. To enable the "direct" approach, in which a supertree is inferred from a sparse average distance matrix

without any intermediate distance imputation step, we re-implemented the solution previously adopted in the fitch program (Felsenstein, 1997) from the PHYLIP suite (Felsenstein, 1989) and the MW algorithm (Makarenkov and Lapointe, 2004) from the T-REX software (Makarenkov, 2001), which deals with missing distances by generalizing Eq. 2 to a weighted least-squares case. Specifically, we introduced a new data type, `AverageDistanceMatrix`, which combines an underlying distance matrix with a Boolean mask matrix of the same dimensions. The mask assigns a unit weight to the known entries of the distance matrix and a null weight to its missing entries, ensuring that the latter do not contribute to the error function. An `AverageDistanceMatrix` object can be created from a vector of distance matrices using an `fnAverageDistanceMatrix()` utility function, which also accepts an optional vector of weights equal in length to the vector of matrices, allowing the user to account for possible differences in quality among the various source datasets. The BLeSS error function is implemented as an `ExponentialErrorDistribution` over distance matrices, which is parameterized by λ_e and can be "clamped" (*sensu* Höhna et al., 2014) to an `AverageDistanceMatrix` object to evaluate the likelihood of a proposed supertree. All of the above-described constructs can be freely combined with other functions, distributions, and MCMC moves already available in RevBayes, allowing BLeSS analyses to make use of a wide range of features including topological constraints, node age calibrations, and precomputed matrices with imputed distances.

1.3.3 Simulation Tests

Simulation-based calibration checking

To validate our implementation of BLeSS in an easily interpretable setting, we conducted a series of simulations involving 4-tip and 5-tip trees and used simulation-based calibration checking (SBC; Talts et al., 2020; Modrák et al., 2023) to test whether the method produces unbiased posteriors. SBC builds upon the earlier posterior quantile test of Cook et al. (2006),

which has seen previous use in phylogenetic settings (Slater et al., 2012b; Uyeda and Harmon, 2014; McHugh et al., 2022). Like the posterior quantile test, SBC exploits the self-recovering properties of Bayesian inference, and can be understood as extending the usual assessment of credible interval coverage (based on the expectation that, e.g., a 95% credible interval should contain the true value in 95% of simulations) by performing it for all credible interval widths simultaneously. However, compared to the posterior quantile test, SBC has the advantage of explicitly accounting for MCMC autocorrelation and the fact that the finite size of posterior samples results in discrete rather than continuous cumulative distribution functions (Gelman, 2017; Talts et al., 2020).

Simulation-based calibration checking employs a generator that repeatedly draws i.i.d. "true" values $\tilde{\theta}^{(1)}, \ldots, \tilde{\theta}^{(N)}$ of a parameter from its prior distribution $p(\theta)$, and then simulates data $y^{(1)}, \ldots, y^{(N)}$ conditional on each true value according to some sampling distribution $f(y \,|\, \tilde{\theta})$. The algorithm to be validated is then used to re-infer the parameter from the simulated data, producing N posteriors $p(\theta \,|\, y)$ of L samples each. Because the vectors of simulated data and parameter values represent draws from the joint distribution of the model, i.e., $(\tilde{\theta}, y) \sim f(\theta, y)$, and because $f(\theta, y) = p(\theta \,|\, y)\, f(y)$ by the axiomatic definition of conditional probability, it follows that the true values $\tilde{\theta}^{(1)}, \ldots, \tilde{\theta}^{(N)}$ represent draws from $p(\theta \,|\, y)$; that is, the same posterior distribution that the algorithm is attempting to sample from. Accordingly, if the model, generator, and posterior-sampling algorithm are all correctly matched, the rank statistic of $\tilde{\theta}$ with respect to the posterior sample $\theta_1, \ldots, \theta_L$, computed for the i-th case as $r^{(i)} = \sum_{l=1}^{L} \mathbb{I}\left[\theta_l^{(i)} < \tilde{\theta}^{(i)}\right]$, should be uniformly distributed on $[0, L]$.

We implemented the SBC procedure for five different scenarios corresponding to two tree sizes (4 or 5 tips) and all possible rooted tree shapes for a given size (two shapes for the 4-tip trees, three shapes for the 5-tip trees). To facilitate interpretation, we further required that the trees conform to specific labeled tree shapes (= topologies) as well. For each scenario, we drew 100 trees from a birth-death process with a high extinction fraction,

complete sampling, and an Exp(1.0) prior placed on the age of origin; following Slater et al. (2012a), the birth and death rates were set to 0.1 and 0.09, respectively. We conditioned on the number of taxa, and employed rejection sampling to ensure the draws were of the desired topology. Each simulated tree was then converted to a distance matrix $\bar{\mathbf{D}}$ (in this case obtained directly, without averaging), which was analyzed using BLeSS with all priors set to their true values and $\lambda_e = 1$. Given the small tree sizes employed in the simulation, this value was chosen to ensure that the BLeSS log likelihoods would primarily reflect $\delta(\hat{\mathbf{D}}, \bar{\mathbf{D}})$ instead of being dominated by the $\log \lambda_e$ term (Eq. 5). In addition to analyzing the complete distance matrix, we also aimed to evaluate the effect of missing distances. To this end, we removed one above-diagonal entry $[i, j]$ from $\bar{\mathbf{D}}$ at a time, producing 6 additional incomplete matrices per replicate in the 4-tip scenarios and 10 additional matrices per replicate in the 5-tip scenarios. Each of these was then analyzed under the same settings as the complete matrix. Accordingly, we carried out a total of $2 \times 7 \times 100$ (4-tip trees) $+ 3 \times 11 \times 100$ (5-tip trees) $= 4700$ analyses. Initially, each analysis was run with a pre-burnin period of 5000 iterations and a sampling period of 10,000 iterations, recording 1 sample every 25 iterations.

For SBC to be feasible, the parameter of interest θ has to be a univariate scalar, or transformable to such by some test quantity $f(\,\cdot\,)$ (Modrák et al., 2023). To validate BLeSS, we chose to focus on branch durations, or branch lengths in units of calendar time, such that for an n-tip tree, $\theta \in \{\ell_1, \ldots, \ell_{2n-2}\}$. Although time trees are commonly described in terms of node heights (divergence times), we found branch durations to be more convenient for the purposes of localizing the impact of missing distances. For a parent node A at height t_A and its child node B at height t_B, the duration of the connecting branch $\ell_{A \to B}$ is simply computed as $t_A - t_B$ (Cheon and Liang, 2014). To identify matching branches, we adopted the criterion proposed by Carruthers et al. (2022), whereby a branch in one tree can be considered equivalent to a branch in another tree only if the sets of tips descended from their parent nodes are identical and the sets of tips descended from their child nodes are identical

as well. This definition does not require the two branches to be present in topologically identical trees, although some branches may only be compatible with a single topology in practice.

To mitigate the impact of MCMC autocorrelation, we followed Algorithm 2 of Talts et al. (2020) and standardized the effective sample size (N_{eff}) of each branch duration to L states, with L chosen to be one less than a large power of 2 to facilitate re-binning when plotting the rank statistic distribution as a histogram. Specifically, we set $L = 255$, or one less than the first power of 2 to exceed the arbitrary but widely used N_{eff} threshold of 200 (Lanfear et al., 2016b; Rambaut et al., 2018). Although recent results indicate that substantially higher N_{eff} values may be necessary to reach a desired level of precision (Fabreti and Höhna, 2021), we consider our choice of L to represent an acceptable trade-off between precision and chain length (directly proportional to the computational cost of the analyses). To extract effective samples from a given analysis, we first calculated N_{eff} using the `coda` R package (Plummer et al., 2006). If the N_{eff} value exceeded L, the chain was truncated; if it failed to reach L, the analysis was re-run with the original number of iterations multiplied by L/N_{eff}. To keep the calculation of N_{eff} computationally feasible, we set the maximum allowed sampling period to 400,000 iterations. Analyses that failed to reach L even with this much greater chain length were repeatedly re-run for the maximum allowed number of iterations until the $N_{\text{eff}} \geq L$ condition was satisfied. Next, we uniformly thinned the full, correlated MCMC sample of size N_s by the autocorrelation time N_s/N_{eff}. More sophisticated thinning strategies are possible (Talts et al., 2020; Säilynoja et al., 2022), but their performance does not appreciably differ from the traditional approach (Säilynoja et al., 2022).

While Talts et al. (2020) recommended thinning the correlated sample only once by the longest observed autocorrelation time when multiple parameters are of interest, this guidance is only applicable when every parameter is sampled at every iteration. In our case, thinning the chain just once to a single set of L states would result in effective sample sizes

of less than L for at least some of the branch durations, as not every sampled tree contained all the branches of interest. Accordingly, we first extracted all samples containing a given branch, and only then applied the truncation, thinning, and rank calculation steps to the branch-specific subsample. While performing SBC separately for each branch provides a way of dealing with the interdependence between branch durations and topologies, it limits our evaluation of BLeSS to marginal rather than joint branch duration distributions, and fails to address parameter interactions (Yao and Domke, 2023).

To assess the uniformity of the rank statistics, we implemented the graphical methods proposed by Talts et al. (2020) using R code adapted from Säilynoja et al. (2022). We generated rank histograms with 20 bins and 95% confidence bands, whose lower and upper bounds were computed as the 2.5th and 97.5th percentiles of the Binom(100, $\frac{1}{20}$) distribution, respectively. This corresponds to an expectation that, on average, no more than 1 of the 20 bins should fall outside the band. While intuitive and easy to interpret, rank histograms are known to be sensitive to the number and placement of individual bins (Säilynoja et al., 2022). We therefore also assessed the results by transforming the $r^{(i)}$ values to fractional ranks (distributed on the unit interval), computing the empirical cumulative distribution function (ECDF) of the latter, subtracting the uniform expectation (i.e., the identity function on the unit interval), and plotting the resulting difference. To visualize expected deviations from exact uniformity, we followed Säilynoja et al. (2022) and computed 95% simultaneous confidence bands at 100 fixed evaluation quantiles (i.e., one for every ECDF value), based on inverse binomial distributions parameterized by a coverage adjustment value found via numerical optimization.

Visualizations are the preferred way of summarizing SBC results, as they make it possible to distinguish between different types and degrees of miscalibration (Modrák et al., 2023); however, more formal tests of deviation from uniformity may be desired. In their pioneering application of the posterior quantile test to phylogenetics, Slater et al. (2012b)

used the Kolmogorov-Smirnov (KS) test for this purpose, but the traditional KS test is not valid when the reference distribution is discrete. To address this shortcoming, we used the recently developed exact Kolmogorov-Smirnov fast Fourier transform (Exact KS-FFT) method (implemented by the authors in the R package KSgeneral; Dimitrova et al., 2020), which generalizes the KS test to mixed and purely discrete reference distributions and allows calculating exact p-values for arbitrarily large sample sizes. We note that this generalization does not address other weaknesses of the KS test, such as its low power to detect deviations in the tails (Säilynoja et al., 2022). Moreover, since we applied the Exact KS-FFT method separately to each component of the branch duration vector and to different matrix completeness scenarios, the resulting p-values are subject to the multiple comparisons problem (Yao and Domke, 2023). To avoid an excessive decrease in test power and discourage a binary interpretation of the results, we opted to use uncorrected p-values with multiple thresholds instead of applying a formal multiple testing correction.

200-tip simulations

To evaluate BLeSS in a setting more closely approaching a typical supertree analysis in terms of tree size, we used a hierarchical simulation scheme involving trees of 200 tips (Fig. 1.2). Specifically, we addressed four aspects of the method's performance: (1) its sensitivity to source tree estimation error, which may cause the distances in $\bar{\mathbf{D}}$ to deviate from ultrametricity; (2) its robustness to different ways of sampling source trees, which induce different proportions and distributions of missing entries in $\bar{\mathbf{D}}$; (3) the relative success of the direct and indirect (imputation-based) approaches to dealing with missing distances; and (4) the impact of the choice of the error-penalizing rate parameter λ_e. We used RevBayes to simulate 100 birth-death trees conditioned on the number of tips under a speciation rate (λ) of 0.1, an extinction rate (μ) of 0.09, complete sampling, and a broad $\mathrm{U}(0, 200)$ prior on the time of origin (t_{or}).

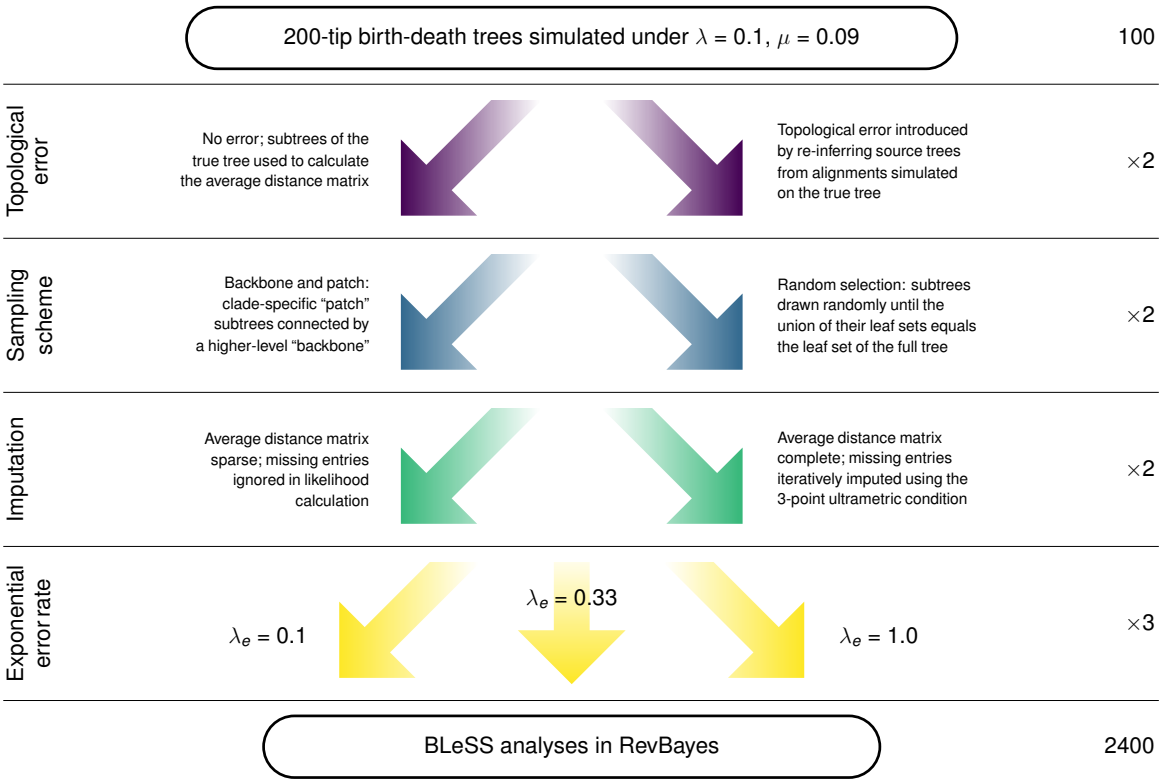

Figure 1.2: Flowchart illustrating the simulation scheme used to evaluate BLeSS.

To address the impact of source tree error, we carried out two sets of simulations. In the first, source trees were obtained simply by dropping tips from each simulated phylogeny, amounting to the assumption that their topologies and divergence times were known without error. The purpose of these analyses was to identify possible biases intrinsic to the method and establish a baseline for more complex tests. In the second scenario, topological and node age error was introduced by simulating DNA sequence data on each complete tree. For each of the 100 replicates, multiple subalignments were extracted from the simulated dataset so that the subsets of included taxa corresponded to the leaf sets of the source trees created for the simulations without error. Finally, source trees were re-inferred from these subalignments using RevBayes. The strict correspondence between the leaf sets of the source trees used in the simulations with and without error helped ensure that any differences in BLeSS performance

between the two scenarios were due to source tree quality rather than differential sampling. To make it computationally feasible to analyze the subsampled alignments in large numbers, and to generate realistic amounts of error across the full range of source tree sizes (3–60 tips), we employed a RevBayes re-implementation of the protocol outlined by Zhang et al. (2016). A strict clock with a rate of 0.003 substitutions per unit time was applied to each of the 100 synthetic trees to convert its branch lengths from units of calendar time to substitutions per site. The JC69 model (Jukes and Cantor, 1969) was then used to simulate 500-bp alignments on the resulting phylograms by using MCMC to sample from the unclamped phylogenetic continuous-time Markov chain distribution. A custom R script was used to break up the resulting data into subalignments, which were subsequently re-analyzed in RevBayes. Priors with a mean equal to the true value were used for the rate of speciation and the clock rate (r), the latter specified so as to ensure that the 95% highest prior density interval spanned two orders of magnitude: $\lambda \sim \text{Exp}(10)$, $r \sim \text{Lognorm}(-6.499, 1.175)$. The priors on the rate of extinction and origin time were slightly misspecified: $\mu \sim \text{Exp}(10)$, $t_{\text{or}} \sim \text{U}(10, 200)$. The substitution process was modeled using JC69, the true model.

At the second level of the simulation scheme (Fig. 1.2), we further varied the manner in which each true (simulated) tree was broken up into individual source trees. First, we used a "backbone-and-patch" protocol designed to mimic the common phylogenomic approach of the same name (Jetz et al., 2012; Upham et al., 2019), in which the complete tree was subdivided into "patch" trees by making a cut at a particular height above the root. The height of the cut was randomly drawn from a uniform distribution bounded by 0 and the root height of the full tree, with rejection sampling used to ensure that the cut resulted in at least 3 patch clades of 3 or more tips each. Any tip that did not belong to a sufficiently large patch clade was instead added to the backbone, which also included one randomly selected tip from each patch clade (Fig. 1.3a). Second, we employed a "random selection" protocol, in which subsets of 20–60 leaves were sampled with replacement from the full leaf

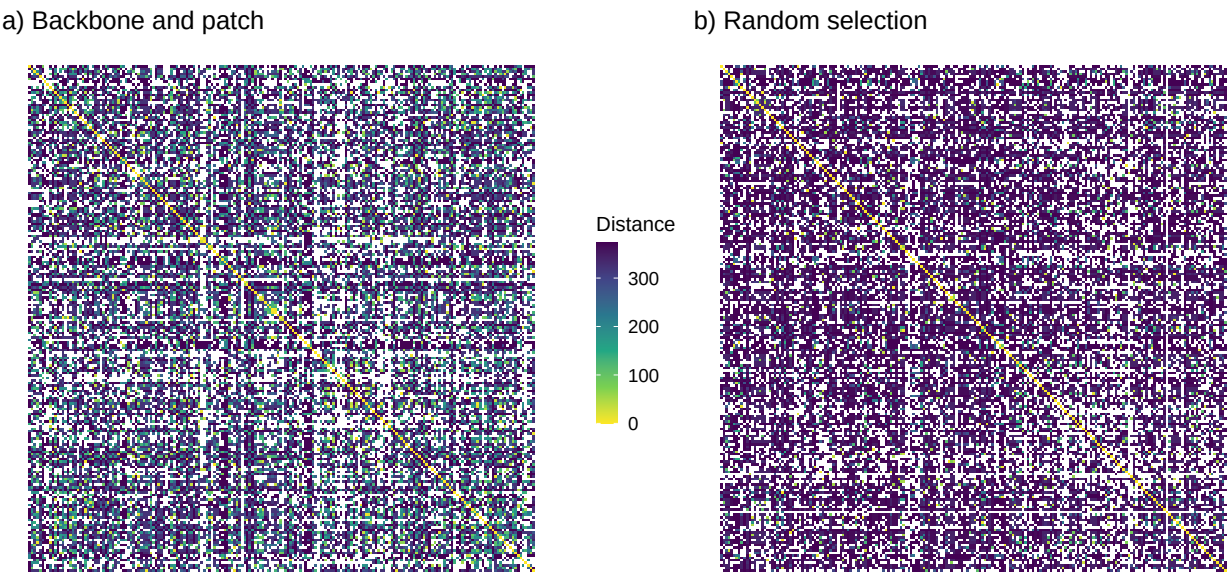

Figure 1.3: Distribution of known (filled cells) and missing (blank cells) distances in arbitrary time units under the (a) backbone-and-patch and (b) random selection sampling schemes. The replicate with median completeness was chosen to represent each scheme.

set of each synthetic tree until their union reached complete coverage of the original full set. The source trees were then obtained by dropping all leaves not included in a given subset from the full tree (Fig. 1.3b). The two sampling schemes gave rise to different distributions of missing distances in the resulting average distance matrices (Fig. 1.3), and to different levels of completeness (backbone-and-patch: 0.069–0.778, mean = 0.311; random selection: 0.517–0.890, mean = 0.685; see Supplementary Information, Table A.2).

In the next stage, we repeated all previously conducted simulations on average distance matrices in which the missing entries were estimated from the available data. We used a variant of the ultrametric imputation approach (De Soete, 1984) following Landry and Lapointe (1997) and Levasseur et al. (2003), who found this algorithm to outperform the more complex additive method when dealing with higher fractions of missing data, missing distances between sister tips, or near-ultrametric matrices. The algorithm relies on the three-point condition satisfied by ultrametric distances (Hartigan, 1967):

$$d(i,j) \leq \max\left[d(i,k), d(j,k)\right] \; \forall \{i,j,k\} \in \mathcal{L}^3 \tag{8}$$

Accordingly, a missing distance between i and j can be estimated by minimizing over all triplets $\{i,j,k\}$ for which $d(i,k)$ and $d(j,k)$ are known, such that

$$\hat{d}(i,j) = \min_k \left\{ \max\left[d(i,k), d(j,k)\right] \right\} \forall k \in \mathcal{L} \; : \; \exists d(i,k) \wedge \exists d(j,k). \tag{9}$$

Instead of applying Eq. 9 in a fully recursive fashion, we performed at most two passes across the matrix. In the first, Eq. 9 was independently applied to each missing distance for which at least one k could be found, so that the imputation of one distance did not immediately affect the number of $\{i,j,k\}$ triplets available for the imputation of the remaining distances. If no k could be found, a second pass was performed in which the results of the first pass were added to the known distances. The imputation algorithm was implemented in a custom R script, with precomputed average distance matrices passed to RevBayes as text files formatted using phangorn (Schliep, 2010). The imputed entries were assigned the same unit weight as the known distances.

Lastly, at the fourth and final level of the simulation scheme, we varied the value of the rate parameter of the exponential error model, λ_e (Fig. 1.2). In effect, λ_e determines the ruggedness of the likelihood surface, or the extent to which suboptimal supertrees are penalized. Smaller values of the parameter (corresponding to a larger mean) flatten the likelihood function and make it more difficult to reject estimates that poorly fit the average distance matrix, whereas larger values increase the ability to discriminate between alternative supertrees and prevent their path lengths from deviating too far from $\bar{\mathbf{D}}$. Our preliminary simulations suggested that large values of λ_e (10–100) led to low MCMC acceptance ratios, resulting in likelihood traces marked by long plateaus and extremely slow convergence. To keep the analyses computationally tractable, we restricted our exploration of the penalty

parameter to three smaller values spanning one order of magnitude: $\lambda_e \in \{0.1, 0.33, 1.0\}$. Accordingly, we carried out a total of $2 \times 2 \times 2 \times 3 \times 100 = 2400$ analyses. Initially, each analysis was run with a pre-burnin period of 40,000 iterations and a sampling period of 225,000 iterations, recording 1 sample every 500 iterations.

To assess MCMC convergence, we first determined the burnin proportion individually for each replicate using the previously suggested criterion (Beiko et al., 2006; Harrington et al., 2021) of discarding all samples taken before the point at which the chain first reached a likelihood higher than the mean of the last 10% of the run. After removing the burnin, we calculated the effective sample sizes (ESS) of all monitored scalar parameters using the coda R package. We further divided the post-burnin trace into 20 intervals of equal length, and calculated the average standard deviation of split frequencies (Lakner et al., 2008) between the last two intervals using R code adapted from the rwty package (Warren et al., 2017). Finally, we used rwty to calculate approximate topological ESS (Lanfear et al., 2016b), employing the squared path difference to quantify topological distances. A given replicate was regarded as having reached convergence if ESS > 200 for all sampled parameters (including topology) and ASDSF < 0.05.

The post-burnin tree traces of all replicates that satisfied the convergence criteria described above were summarized as maximum clade credibility (MCC) and maximum *a posteriori* (MAP) supertrees. These summary supertrees were then compared to the corresponding true tree using a variety of statistics: the Robinson-Foulds (RF; Robinson and Foulds, 1981) and Kuhner-Felsenstein (KF; Kuhner and Felsenstein, 1994) distances, as implemented in phangorn (Schliep, 2010); the path length difference (Steel and Penny, 1993), as implemented in the R package castor (Louca and Doebeli, 2017); the matching split (Bogdanowicz and Giaro, 2012; Lin et al., 2012) and clustering information (Smith, 2020a) distances, as implemented in the package TreeDist (Smith, 2020b); the Billera- Holmes-Vogtmann (BHV) metric (Billera et al., 2001), as implemented in the package distory (Chakerian and Holmes,

2012); the symmetric quartet divergence (Smith, 2019a), as implemented in the package Quartet (Sand et al., 2014; Smith, 2019b); the Kendall- Colijn (KC) metric (Kendall and Colijn, 2016), as implemented in the package treespace (Jombart et al., 2017) and with the contribution of branch lengths to the distance ranging from 0 to 1 in increments of 0.1; and the difference between the true and inferred root height, calculated using the package phytools (Revell, 2012).

To effectively summarize the results, which involved assessing the relative importance and marginal effects of potentially covarying discrete and continuous predictors, we employed Bayesian additive regression trees (BART; Chipman et al., 2010; Bleich et al., 2014; Hill et al., 2020) as implemented in the R package bartMachine (Kapelner and Bleich, 2016). BART is a nonparametric sum-of-trees machine learning method, in which every individual regression tree represents a "weak learner" whose contribution to explaining the variation in the response variable is constrained by a regularization prior to prevent overfitting. We used the RF as well as and KF distances between the estimated and true supertrees as response variables, as both metrics are only imperfectly correlated (Supplementary Information, Fig. A.16) and present distinct advantages. In contrast to the RF distance, the KF distance quantifies both topological and branch length estimation error, but unlike the former, it cannot be normalized, and is therefore less easy to interpret. In addition to the four categorical variables described above (topological error, source tree sampling scheme, missing data treatment, and λ_e value), we also used two continuous predictors: true tree imbalance, as quantified by the Colless index (Colless, 1982), and $\bar{\mathbf{D}}$ completeness, defined as the fraction of non-missing entries prior to imputation. Only the replicates that reached convergence according to the criteria listed above were used for the BART analyses, resulting in highly unequal sample sizes for the different scenarios. For both analyses (of RF and KF distances), we randomly allocated 80% of the response and predictor values to the training set and the remaining 20% to the testing set. We used out-of-sample root-mean-square error to choose

the optimum number of trees (testing numbers from 20 to 120 in increments of 2), and conducted an MCMC analysis with a burnin period of 120,000 iterations and a sampling period of 120,000 iterations under the best tree number and with all other hyperparameters set to their default values ($\alpha = 0.95$, $\beta = 2$, $k = 2$, $q = 0.9$, $\nu = 3$; Kapelner and Bleich, 2016). Following standard practice (Chipman et al., 2010), we used the proportion of times each predictor was chosen for a splitting rule (out of all splitting rules and across the entire posterior sample) as a measure of that predictor's importance, and partial dependence plots to visualize the marginal effects of continuous predictors.

In addition to evaluating point estimates (summary supertrees), we also aimed to assess each posterior sample as a whole in terms of node age coverage. Since our simulations involved the simultaneous inference of tree topology and divergence times, we specifically focused on the age of the root, which represented the only internal node that was guaranteed to be present in all sampled supertrees regardless of their topology. Coverage probability was calculated as the proportion of replicates in which the true root age fell into the corresponding 95% highest posterior density (HPD) interval, out of all those that reached convergence. Additionally, we computed the 95% HPD intervals for all nodes in a given MCC tree that were also present in the true tree, and determined what percentage of these contained the corresponding true ages. Naïvely, one may expect this to be the case for 95% of nodes on average if the method is well-behaved, with higher proportions indicating overly diffuse posteriors (too much uncertainty) and lower proportions indicating overly narrow posteriors (too much certainty). However, since ancestors necessarily predate their descendants, individual nodes of the same tree do not represent independent replicates in terms of age, and topological mismatches between the true tree and the summary tree may cause the same proportion to correspond to very different rates of overall inferential success. We mitigated the latter problem by calculating the relevant proportions both out of correctly recovered nodes only and out of all internal nodes present in the tree ($n = 199$), but while the resulting

percentages remain informative of the overall performance of the method, they cannot be interpreted strictly as coverage probabilities.

Scalability tests

In addition to testing for implementation correctness and accuracy, we performed one final set of simulations focusing on computational efficiency to evaluate how well BLeSS scales with respect to tree size. Specifically, we extended one of the scenarios explored in the 200-tip simulations (no topological error, backbone-and-patch sampling scheme) to trees of 500, 1000, 5000, and 10,000 tips. As before, true trees were simulated under $\lambda = 0.1$, $\mu = 0.09$, and complete sampling, and conditioned on the number of tips. Since the simulations were aimed at assessing computational speed rather than accuracy, only a single tree was generated for each scenario (Supplementary Information, Table A.9). The process of creating the backbone and patch trees was similar to that used in the 200-tip simulations, except that a minimum of 10 patch clades of at least 10 tips each was required.

First, we performed MCMC-free benchmarks to determine the amount of time required for those operations that are carried out upfront and only once per analysis (e.g., computing $\bar{\mathbf{D}}$, drawing a tree from a prior distribution), as well as those operations that collectively make up a single MCMC iteration (converting a proposed tree to a path-length matrix, and computing the natural logarithm of its probability density under the exponential error distribution). All measurements employed the internal RevBayes `time()` function, which provides a Rev wrapper around the microsecond clock from the Boost C++ library (`http://www.boost.org`). For each combination of tree size (500, 1000, 5000, and 10,000 tips) and missing data treatment (imputation vs. no imputation), 10 replicates with different random seeds were performed. The benchmarks were conducted with millisecond precision on a single compute node consisting of Intel Xeon Gold 6248R processors, with one thread per replicate. To confirm the current implementation of BLeSS gains no performance boost

from parallelization, we repeated the benchmarks using the MPI (Message Passing Interface) version of RevBayes with 32 threads per replicate (Supplementary Information, Table A.11).

Second, we ran 5 chains with different random seeds on each of the 8 datasets (4 tree sizes $\times$ 2 missing data treatments) for 5 million iterations or 180 hours, whichever was reached first. In contrast to the settings used for the SBC and 200-tip analyses, the `moveschedule` argument of the MCMC function was set to `"single"` so that each proposed move would be counted as a separate iteration (Höhna et al., 2017). Sampling frequencies were scaled so as to obtain a comparable number of samples per unit runtime for all tree sizes (Supplementary Information, Table A.10). The `mnScreen` monitor was used to write elapsed time to the standard output, which was redirected to a plain-text file. These files were subsequently parsed and used to compute the number of moves per unit runtime. In addition, we computed the previously proposed statistic of ESS increment per unit runtime (Aberer et al., 2016; Fourment et al., 2018) for all monitored scalar parameters by applying the `effectiveSize` function from the coda package to cumulative subsamples of each MCMC run. Due to limitations imposed by the cluster on which the analyses were performed, the 180-hour period consisted of 36-hour windows, with the RevBayes checkpointing functionality used to restart each chain from the last state saved before a timeout. Extra iterations that were run after the last saved state but before the timeout were included when calculating the total number of moves per unit runtime but omitted from the calculation of effective sample sizes. All MCMC analyses were carried out on the same hardware as the MCMC-free benchmarks.

Third, we evaluated sampling efficiency by recording the acceptance rates of the various MCMC moves used to update both the supertree itself and the hyperparameters of the birth-death prior from which the supertree is drawn. Deviations from the default RevBayes target rate of 0.44 can be due to tree size-related problems with the exploration of the parameter space, which may in turn reflect genuine features of the posterior: low acceptance rates of topology moves may be indicative of the chain having reached a tree with a high posterior

probability that is difficult to improve upon, while high acceptance rates may occur when the data are insufficient to discriminate among different topologies (Harrington et al., 2021). The acceptance rates were calculated from chains of different lengths for different tree sizes, without a dedicated pre-burnin period or autotuning.

1.3.4 Empirical Application to the Order Carnivora

As an empirical demonstration, we used BLeSS to sample the posterior distribution of time-calibrated supertrees for extant representatives of the mammalian order Carnivora based on a set of published time-scaled phylogenies. Carnivora is a compelling system in which to explore our approach for two reasons: several relatively complete time-scaled molecular phylogenies are available for comparison (Agnarsson et al., 2010; Slater and Friscia, 2019; Upham et al., 2019), and the order has been the focus of several previous supertree efforts (Bininda-Emonds et al., 1999; Nyakatura and Bininda-Emonds, 2012; Akanni et al., 2015).

We first performed a literature review to identify time-calibrated source trees that could be included in the profile. We ignored studies that obtained complete or near-complete sampling of the order to ensure that not all entries in the average distance matrix were filled, though studies that achieved relatively complete sampling of families or superfamilies were included. In order to be considered for inclusion, we required that one of two possible sources of data be associated with the paper: a tree file containing the topology annotated with branch lengths, or else a figure of the topology with mean divergence times provided in the text or in an associated table that would enable manual transcription of the time tree. If divergence times were provided only for a subset of figured nodes, we entered the topology corresponding to that subset of bipartitions into the profile. In cases where multiple sets of divergence times were provided, typically corresponding to different calibration schemes, we recorded all. We did not require the source trees to be based on any particular data type or time calibration approach, although we did note both. Our decision to reject source studies

that did not provide a time tree in a usable format, rather than downloading, aligning, and reanalyzing the sequence data used in them, was motivated by a desire to replicate the situations in which BLeSS, or any other supertree method, might most profitably be used in place of analysis of primary data.

Our literature review identified 31 studies for which at least one tree file was available or from which at least one time tree could be manually reconstructed based on figures and tables provided. Our final profile included 33 trees from 30 source studies, with one outlier study excluded because it yielded unusually young divergence times. After removing duplicates and domesticated taxa, the union of the leaf sets of these trees comprised a total of 260 species, including a few latest Pleistocene taxa that we treated as extant for the purposes of this study. Sampling was variable among families. Based on the American Society of Mammalogists' Mammalian Diversity Database (Burgin et al., 2018, v1.91; downloaded July 8, 2022), we achieved complete sampling of extant Hyaenidae ($n = 4$), Mephitidae ($n = 14$), Phocidae ($n = 19$), Prionodontidae ($n = 2$), and Ursidae ($n = 8$) among families with two or more species. The sampling fraction for other families containing more than two species was variable, ranging from as high as 0.96 in Felidae to as low as 0.29 in Herpestidae. Low sampling in some families is not due to a lack of available studies, but instead reflects a lack of associated tree files or divergence time estimates for figured trees in those studies that are available. A full description of source studies evaluated for inclusion is provided in the Supplementary Information.

As in our simulation study, we performed BLeSS analyses with and without imputation. Several of the source studies in our profile of carnivoran time trees are based on mitochondrial DNA data alone, which have been suggested to yield divergence time estimates that are too old for younger nodes and too young for older nodes (Fulton and Strobeck, 2010). We therefore performed a second set of BLeSS analyses in which the contribution of mitogenome-based source trees to the average distance matrix was downweighted by a factor of 10.

These downweighted distance matrices were again analyzed both with and without imputed distances. Based on our simulation studies, we set $\lambda_e = 0.1$ for the main analyses. However, as mentioned above, an alternative strategy is to marginalize over λ_e. We therefore performed a final analysis using the average distance matrix derived from downweighted mitochondrial data and with imputed missing distances (see Results for rationale) while integrating out λ_e, which was drawn from an Exp(1.0) prior.

For each analysis, we placed exponential priors on the speciation and extinction rates, $\lambda \sim \text{Exp}(10)$, $\mu \sim \text{Exp}(10)$. We used an offset exponential prior for the root age, $t_{\text{or}} \sim \text{Exp}(0.1645, \text{offset} = 38)$, where the offset approximately corresponds to the first appearance dates of the canid *Hesperocyon gregarius* and the amphicyonid *Daphoenus lambei* (Tomiya, 2011), and the rate parameter is scaled so as to place the 99th percentile at the Cretaceous–Paleogene boundary (66 Ma). We performed 4 simultaneous runs to ensure convergence, with a pre-burnin period of 40,000 iterations and a sampling period of 1,000,000 iterations. Parameter posteriors were visually inspected as trace plots, and convergence was further assessed based on the ESS and ASDSF values calculated for scalar parameters and sampled topologies, respectively. Combining all 4 runs almost always resulted in extremely low ESS values for the log likelihood (LnL), log prior (LnPr), and their sum (the log posterior), but moderate or high effective sample sizes for the remaining scalar parameters and low ASDSF values, in agreement with previous observations that there is little correlation between LnL ESS and topological convergence (Harrington et al., 2021). In such cases, we combined the posterior samples from all 4 runs as long as their ASDSF did not exceed 0.01, and computed MAP and MCC trees both from the pooled sample and from the single run with the highest post-burnin LnL values. For those analyses where ASDSF was > 0.01, only the run exhibiting the highest LnL values was summarized.

To determine how the constrained least-squares estimator underlying BLeSS compares to a simple application of existing distance-based methods to an average distance matrix,

we also inferred carnivoran supertrees using neighbor-joining (NJ; Saitou and Nei, 1987) as implemented in phangorn (Schliep, 2010). NJ represents a natural point of comparison, as it was historically viewed as a fast approximation to ordinary least squares (OLS) (Felsenstein, 2004), although more recent studies demonstrated that it actually seeks to optimize a different criterion (Gascuel and Steel, 2006) which allows it to outperform OLS in practice (Willson, 2005). Since existing implementations of NJ can deal neither with the ultrametricity constraint nor with sparse matrices (though see Criscuolo and Gascuel, 2008), we only estimated NJ supertrees from imputed average distance matrices (both with and without downweighting mtDNA-based source trees) and focused only on the topology of the inferred supertree when comparing the results to those of BLeSS. The latter was necessary because the lack of ultrametricity prevents the branch lengths recovered by NJ from being interpreted as amounts of time elapsed between successive divergences, thus rendering them essentially meaningless.

A common problem with evaluating method performance in empirical settings is the fact that in contrast to simulations, the true topology remains unknown. We adopted the approach of Agnarsson and May-Collado (2008) and evaluated the analyses by computing the fraction of "benchmark" clades recovered in the summary supertrees. Such clades have to be of undisputed monophyly, supported by multiple lines of evidence, and selected prior to the analysis. We recognized a total of 53 benchmark clades with ranks ranging from suborder to subtribe; the full list is given in the Supplementary Information (Table A.14) along with the expected composition of each clade in terms of our species sample.

1.4 Results

1.4.1 Simulation Tests

Simulation-based calibration checking

The number of iterations needed to reach the target effective sample size for all branch lengths of interest varied as a function of tree size, tree symmetry, and matrix completeness. Across all five scenarios, reaching the target ESS took a greater number of iterations on average for analyses performed on the complete $\bar{\mathbf{D}}$ than for analyses with missing distances. On average, the shortest runs were required for the 4-tip balanced-topology scenario (i.e., ((1, 2), (3, 4)); ∼49,930 iterations), while the longest runs characterized the 5-tip scenario with a single tip sister to a balanced subtree (i.e., (1, ((2, 3), (4, 5))); ∼233,830 iterations). For the latter scenario, achieving high ESS values proved to be particularly challenging, and a number of replicates had to be re-run multiple times with the maximum allowed chain length (400,000 iterations, corresponding to 16,000 total samples). This difficulty in inducing BLeSS to sample the true imbalanced topology in a trivial case is surprising given the previously reported bias of average distance matrix methods in favor of pectinate source trees (Wilkinson et al., 2005). However, it is not unexpected if the posterior ends up being dominated by the birth-death prior, which is uniform over labeled histories, and accordingly favors balanced topologies, which are compatible with more labeled histories than imbalanced ones (Velasco, 2008).

The graphical summaries generally indicate that the posteriors inferred by BLeSS are well-calibrated, with the empirical CDF mostly falling within the envelope of theoretical expectation (Fig. 1.4) and only a small number of rank histogram bins extending outside the approximate confidence interval. In the 4-tip balanced-topology scenario, a near-exact match between the empirical and theoretical CDFs was observed when analyzing the complete $\bar{\mathbf{D}}$. Even in the cases of mismatch, the observed failure modes were predictable from

the properties of the underlying data, such as when the removal of a distance between two terminal sister taxa resulted in excessive certainty in the duration of their branches, producing a strongly significant deviation from uniformity according to the Exact KS-FFT test (the "no $d(1,2)$" case in Fig. 1.4). This illustrates the well-known problem that missing distances between terminal sister taxa ("cherries" *sensu* Erdős et al., 1997) pose to all average consensus methods, as there is no information in the rest of the matrix that can be exploited to estimate the length of the corresponding branches (Lapointe and Kirsch, 1995; Lapointe and Levasseur, 2004). As expected, the posteriors and their summaries (i.e., rank histograms and ECDF plots) were identical for each pair of branches forming a cherry, reflecting the fact that their durations were not estimated independently but simply obtained by halving the corresponding distance.

When the true tree is imbalanced, BLeSS posteriors largely remain free of bias but may exhibit excessive uncertainty, as indicated by concave rank histograms and S-shaped ECDF curves around the diagonal. In fully pectinate 4-tip and 5-tip trees, this problem was particularly severe for the branch subtending the earliest-diverging tip (Supplementary Information: ℓ_5 in Figs. A.8, A.9 and ℓ_7 in Figs. A.12, A.13), whose duration is equivalent to the root height of the tree. The next most strongly affected branch subtended the second earliest-diverging tip (Supplementary Information: ℓ_3 in Figs. A.8, A.9 and ℓ_5 in Figs. A.12, A.13), suggesting a trend of excessive uncertainty progressively accumulating with increasing node depth. Compared to the 4-tip balanced-topology scenario, the matrices computed from fully pectinate trees were much less sensitive to the removal of individual entries, including even the distance between the two tips of the only cherry present in such trees (Supplementary Information: the "no $d(3,4)$" case in Figs. A.8, A.9 and the "no $d(4,5)$" case in Figs. A.12, A.13). The remaining two scenarios, which involved 5-tip trees intermediate between fully balanced and fully pectinate topologies, also showed minimal sensitivity to distance removal but less clear overall trends (Supplementary Information, Figs. A.10, A.11, A.14, A.15). In

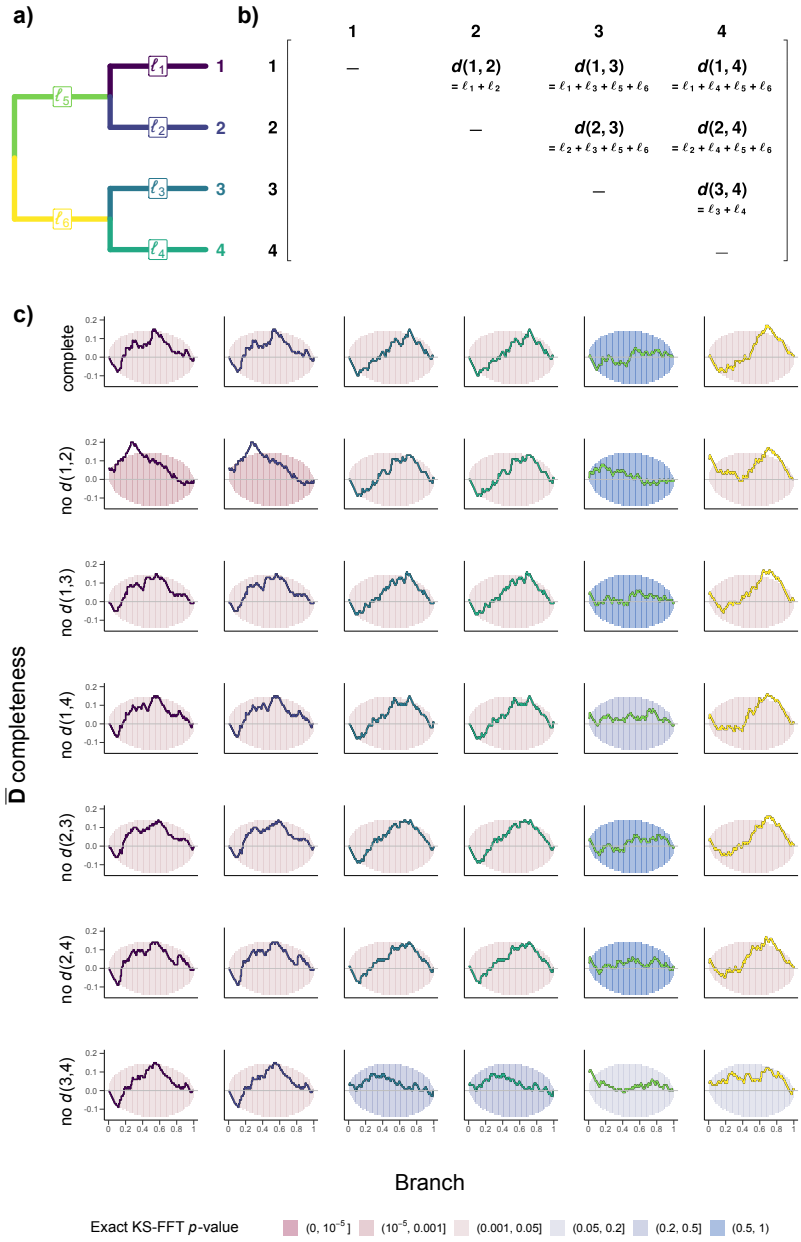

Figure 1.4: Results of simulation-based calibration checking for the 4-tip balanced tree scenario. (a) The topology with labeled branches. (b) The corresponding distance matrix, with observed distances expressed as sums of the underlying branch durations. (c) The difference between the empirical cumulative distribution function (ECDF) and the discrete uniform expectation (stepwise identity function) with 95% simultaneous confidence bands for individual branch durations and different average distance matrix ($\bar{\mathbf{D}}$) completeness scenarios. The confidence bands are colored according to the p-values from uniformity tests conducted using the exact Kolmogorov-Smirnov fast Fourier transform (Exact KS-FFT) method.

particular, the scenario with a single tip sister to a balanced subtree gave rise to the only instance of bias observed in our SBC analyses, which affected the branches subtending the two cherries (Supplementary Information: ℓ_5 and ℓ_6 in Figs. A.10, A.11). The durations of these two branches exhibited posteriors with an overabundance of particularly high estimates, and thus low ranks (Supplementary Information, Fig. A.11). Plausibly, this bias may be related to the convergence issues reported above for the same scenario, or to the solution we adopted in response to these issues (i.e., re-running the inference until a sufficient number of effectively independent samples was obtained within a specified chain length).

200-tip simulations

The 2,400 BLeSS analyses of simulated large-scale trees were run for 225,000–1,350,000 generations, with 1,432 analyses reaching convergence according to the predefined criteria (see Materials and Methods). The proportion of replicates satisfying our convergence metrics was subject to considerable variation, ranging from 7 out of 100 to 98 out of 100. The rate of achieving convergence was generally higher for analyses with rather than without imputation, and for analyses conducted under lower values of λ_e (Fig. 1.5). A preliminary examination of the results showed that the MCC trees slightly but consistently outperformed the MAP trees (Supplementary Information, Tables A.3–A.5), and we therefore restricted subsequent evaluation to the former. For ease of interpretation, we primarily focused on those distance metrics that can be normalized to the unit interval (with 0 indicating perfect agreement and 1 indicating no agreement) when comparing the inferred and true supertrees (Fig. 1.5); the values for the remaining metrics are given in the Supplementary Information (Tables A.6–A.8; Fig. A.27). The RF and KF distances were additionally investigated using BART analyses. These reached excellent convergence (Supplementary Information, Figs. A.19, A.20) when run with the optimum number of regression trees (RF distances: $m = 56$; KF distances: $m = 62$), but their residual distributions exhibited significant devi-

ations from normality (Supplementary Information, Fig. A.21), suggesting that prediction intervals and other output that assumes normally distributed noise may be of limited reliability. Indeed, while the 95% prediction intervals contain the actual response values from the testing set in 89.2% of cases for the RF model, indicating good predictive performance, this is only true in 74.1% of cases for the KF model (Supplementary Information, A.22).

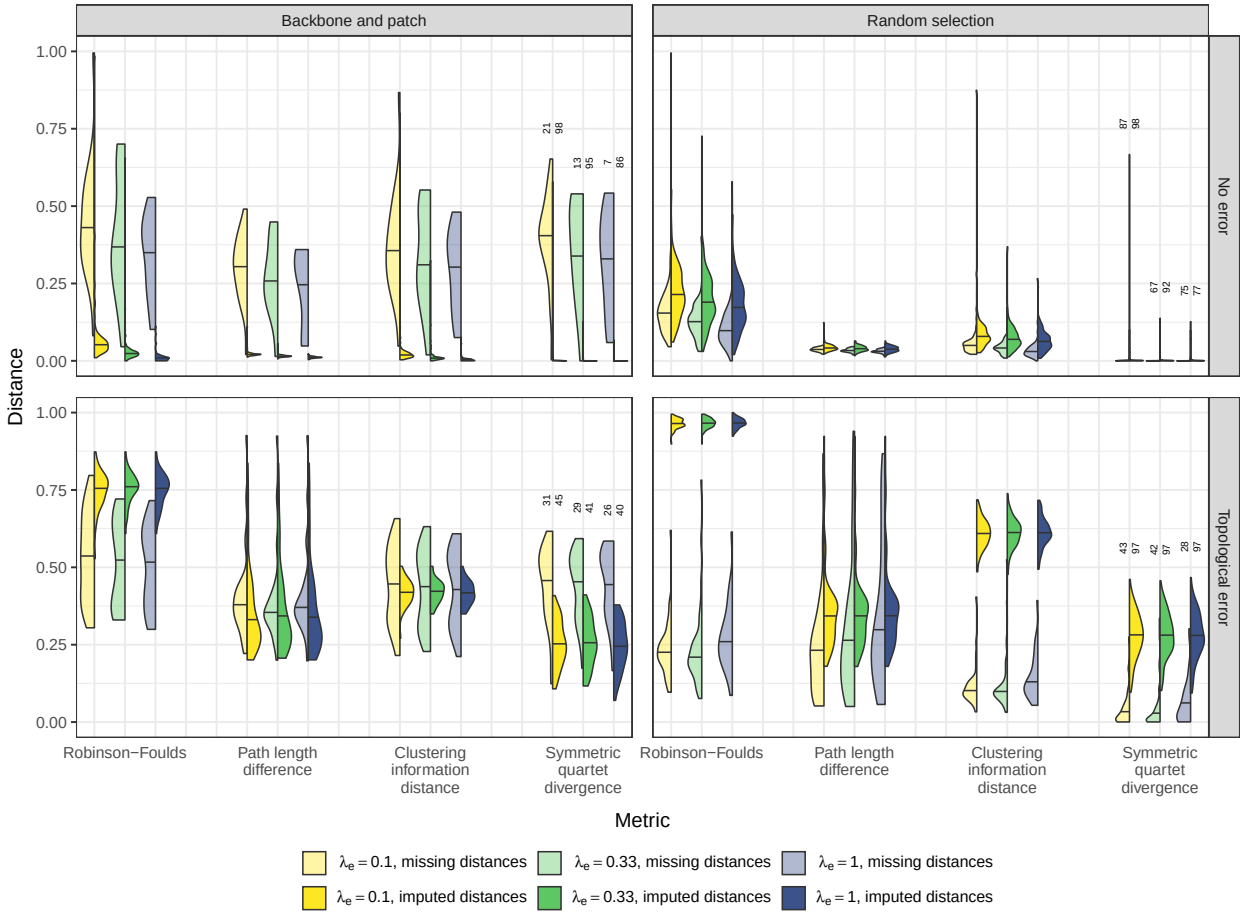

Figure 1.5: Split violin plots showing the distributions of accuracy metrics across the 24 scenarios illustrated in Fig. 1.2. All metrics are normalized and expressed as distances (rather than similarities) between the true (simulated) supertree and the BLeSS maximum clade credibility estimate, so that 0 indicates perfect agreement and 1 indicates no agreement. The numbers above the rightmost set of violin plots indicate the number of replicates that reached convergence for a given scenario.

The mean values of normalizable distance metrics vary widely across the 24 simulation scenarios, ranging from 0.012 to 0.965 for the RF distance, from 0.012 to 0.408 for path

length difference, from 0.006 to 0.613 for the clustering information distance, and from 0.001 to 0.446 for the symmetric quartet divergence (Fig. 1.5). When accounting for the well-known tendency of the RF metric to rapidly saturate (Steel and Penny, 1993; Smith, 2020a), which helps explain the very large distances observed in several scenarios, these values generally indicate moderate to excellent accuracy. At the same time, the corresponding scenario-specific distributions exhibit a considerable overlap, suggesting that the variation in the properties of individual simulated datasets (replicates) was at least as impactful as the differences in analytical choices (Fig. 1.5).

This is confirmed by the BART analyses, which show the imbalance of the true tree and the completeness of the corresponding average distance matrix to be the two most important predictors of both RF and KF distances, as determined by the frequency with which they were chosen for regression tree splitting rules (Supplementary Information, Fig. A.23). Similarly, all three variable selection procedures suggested by Bleich et al. (2014) consistently identify tree imbalance and $\bar{\mathbf{D}}$ completeness as the only two predictors to be included in both the RF model and the KF model (Supplementary Information, Fig. A.26). Nevertheless, comparing the pseudo-R^2 value of the original model to a null distribution derived from 50 permuted datasets shows that all predictors (including the four categorical variables corresponding to the different treatments outlined in Fig. 1.2) have a significant effect on the RF distance between the true and inferred supertrees, and all but two (λ_e: $p = 0.98$; and source tree sampling scheme: $p = 0.96$) have a significant effect on the corresponding KF distance. As expected, the partial dependence function shows that the RF distance between the true and inferred supertrees decreases with increasing $\bar{\mathbf{D}}$ completeness in an almost linear fashion, but the relationship is less straightforward for the corresponding KF distance (Supplementary Information, Fig. A.25). After marginalizing out the contributions of the remaining predictors, the contribution of true tree imbalance to the two accuracy metrics is also highly non-monotonic (Supplementary Information, A.25). While the average consen-

sus method was previously shown to resolve within-profile conflicts in favor of imbalanced source trees (Wilkinson et al., 2005), our estimates of the partial dependence function average the effect of true tree imbalance over supertrees derived from both perfectly congruent and conflicting source trees, possibly accounting for the complex relationship observed.

Qualitative comparisons show that in the absence of source tree conflict, λ_e has the expected effect on topological accuracy, with larger values corresponding to smaller distances between the true tree and the BLeSS summary supertree (Fig. 1.5). However, this effect is attenuated when topological error is introduced into the profile, likely because forcing the estimated supertrees to adhere more closely to $\bar{\mathbf{D}}$ is of lesser utility when the $\bar{\mathbf{D}}$ entries themselves deviate from the branch durations of the true tree because of source tree conflict. The BART analyses of the RF distances between the true and estimated supertrees confirm that of all categorical predictors (i.e., variables other than $\bar{\mathbf{D}}$ completeness or true tree imbalance), λ_e interacts most strongly with the presence or absence of source tree error, although the same is not true of the corresponding KF distances (Supplementary Information, Fig. A.24). In the latter case, interactions of λ_e with categorical predictors are infrequent, reflecting the fact that the penalty parameter has no significant effect on the response and is rarely used as a splitting variable in the first place (Supplementary Information, Figs. A.23, A.24).

Complex interactions between the presence or absence of error in the profile, source tree sampling scheme, and missing data treatment are apparent from the scenario-specific distributions of accuracy metrics (Fig. 1.5) as well as the BART analyses (Supplementary Information, Fig. A.24). The best-performing combinations of treatments include backbone-and-patch analyses with imputation and no source tree error, followed by random-selection analyses with no source tree error or imputation (see Fig. 1.6 for an example). Under the backbone-and-patch sampling scheme, the imputation of missing $\bar{\mathbf{D}}$ entries strongly improves BLeSS accuracy in the absence of source tree error, but its effect is more ambiguous when error is introduced into the profile, with most metrics (especially the symmetric quarter

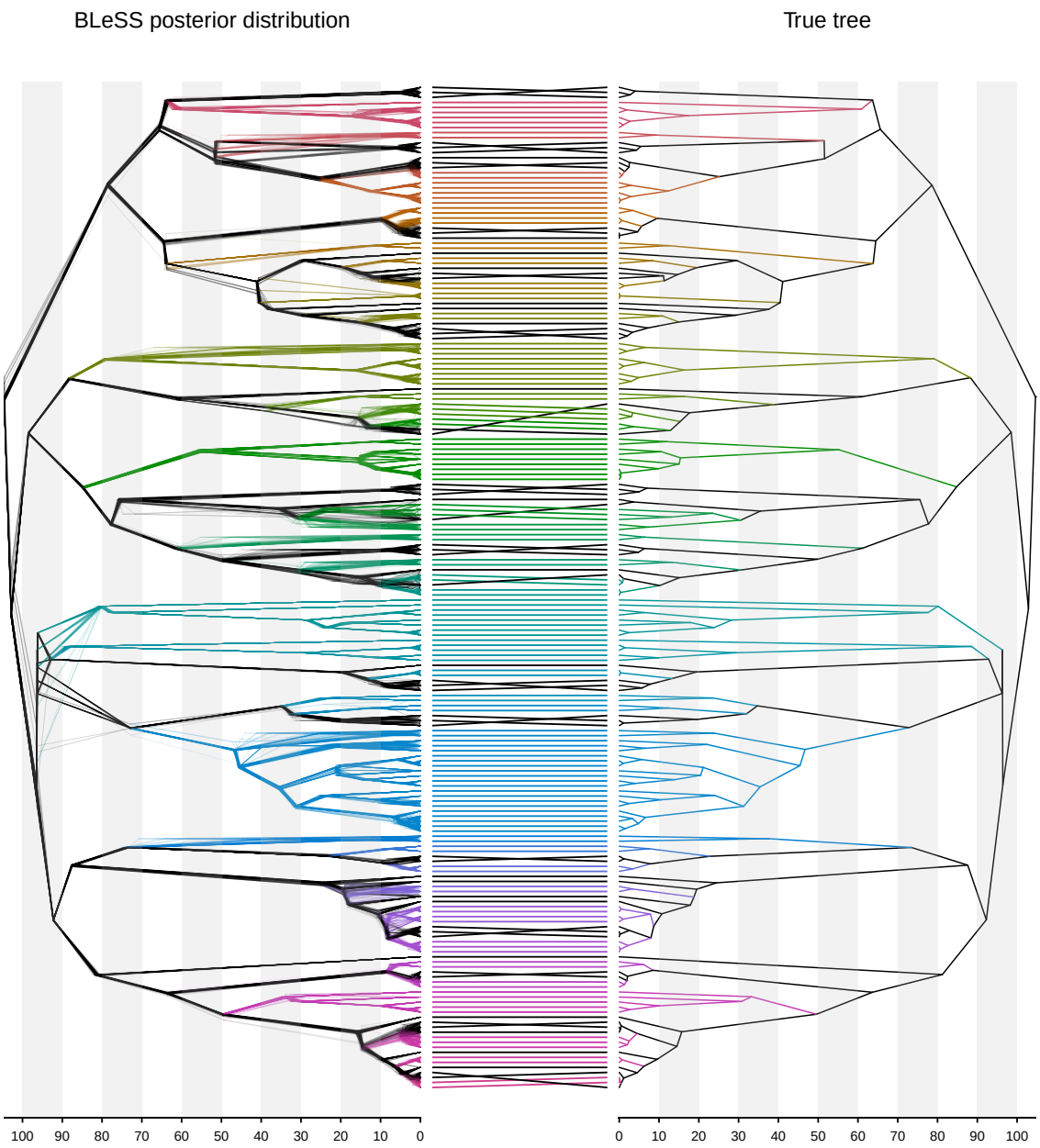

Figure 1.6: Tanglegram illustrating the topological and node height accuracy of a random BLeSS replicate conducted without imputation on randomly selected error-free source trees under $\lambda_e = 0.1$. A DensiTree (Bouckaert, 2010) representation of 50 samples from the BLeSS posterior is superimposed onto the maximum clade credibility (MCC) supertree computed from all samples on the left, and compared to the true tree on the right. Edges of the same color indicate the most inclusive non-nested subtrees that are congruent between the true tree and the MCC supertree; the inter-tree edges connect corresponding tips, and their crossings indicate topological conflict. The numbers along the x-axis represent arbitrary units of node height.

Sampling scheme	Imputation	λ_e	Robinson-Foulds distance		
			True vs. est. source	Est. source vs. est. supertree	True vs. est. supertree
Backbone and patch	No	0.1	0.123 [0.029, 0.367]	0.122 [0.041, 0.185]	0.536 [0.331, 0.778]
Backbone and patch	No	0.33	0.114 [0.028, 0.377]	0.118 [0.027, 0.171]	0.514 [0.334, 0.71]
Backbone and patch	No	1	0.121 [0.029, 0.382]	0.133 [0.083, 0.197]	0.503 [0.309, 0.697]
Backbone and patch	Yes	0.1	0.125 [0.024, 0.349]	0.64 [0.471, 0.813]	0.749 [0.622, 0.847]
Backbone and patch	Yes	0.33	0.126 [0.024, 0.352]	0.645 [0.468, 0.814]	0.755, [0.65, 0.838]
Backbone and patch	Yes	1	0.129 [0.024, 0.345]	0.643 [0.475, 0.814]	0.749 [0.649, 0.838]
Random selection	No	0.1	0.077 [0.016, 0.272]	0.064 [0.016, 0.244]	0.235 [0.102, 0.39]
Random selection	No	0.33	0.092 [0.02, 0.438]	0.083 [0.017, 0.456]	0.235 [0.097, 0.585]
Random selection	No	1	0.107 [0.029, 0.367]	0.095 [0.017, 0.369]	0.274 [0.114, 0.501]
Random selection	Yes	0.1	0.081 [0.02, 0.31]	0.78 [0.613, 0.874]	0.965 [0.934, 0.988]
Random selection	Yes	0.33	0.081 [0.02, 0.31]	0.785 [0.62, 0.876]	0.965 [0.934, 0.99]
Random selection	Yes	1	0.081 [0.02, 0.31]	0.785 [0.616, 0.874]	0.965 [0.934, 0.99]

Table 1.1: The relative contributions of source tree error and BLeSS error to the topological error in the estimated supertree ($\hat{\mathcal{T}}$) in those simulations where source tree error is present. The contribution of source tree error is quantified as the normalized Robinson-Foulds (RF) distance between a source tree $\mathcal{T}_k$ and $\tilde{\mathcal{T}}|\mathcal{L}(\mathcal{T}_k)$ (the restriction of the true tree to the leaf set of $\mathcal{T}_k$), averaged across all source trees ("True vs. est. source"). The contribution of BLeSS error is quantified as the normalized RF distance between $\mathcal{T}_k$ and $\hat{\mathcal{T}}|\mathcal{L}(\mathcal{T}_k)$, averaged across all source trees ("Est. source vs. est. supertree"). The overall error is quantified as the normalized RF distance between $\hat{\mathcal{T}}$ and $\tilde{\mathcal{T}}$ ("True vs. est. supertree"). Every value is reported as the mean and 95% interpercentile range over all replicates that reached convergence. While the distances between $\mathcal{T}_k$ and $\tilde{\mathcal{T}}|\mathcal{L}(\mathcal{T}_k)$ do not vary with imputation or λ_e, the values for different scenarios are averaged over different sets of replicates, resulting in different means and 95% ranges.

divergence) still showing improvement but the RF distance indicating strongly degraded performance. In contrast, under the random-selection sampling scheme, imputation always results in lower accuracy, with the difference being relatively minor in the absence of source tree error but increasing to such an extent in its presence that the distributions of several metrics are completely disjunct between the two missing data treatments (Fig. 1.5). The BART analyses of both RF and KF distances show the interactions of source tree error and missing data treatment to be the most important pairwise interactions between any two categorical predictors, followed by interactions between missing data treatment and sampling scheme, and finally by interactions between sampling scheme and source tree error (Supplementary Information, Fig. A.24).

The observed interplay between imputation and source tree error may appear intuitive: when the available $\bar{\mathbf{D}}$ entries are free of error (perfectly ultrametric), the missing entries can

be imputed without error as well, thus increasing the completeness of $\bar{\mathbf{D}}$ while still allowing it to perfectly fit the true supertree. A more complete average distance matrix can better discriminate between near-optimal estimates, leading to more accurate inference. In contrast, using ultrametric estimation to impute missing entries based on distances that deviate from ultrametricity is likely to compound pre-existing error. This expectation is partly borne out by decomposing the topological error in the inferred supertrees into two components: the mean topological distance between the source trees re-estimated from simulated sequences and the restrictions of the true supertree to their leaf sets, and the mean topological distance between the source trees and the restrictions of the inferred supertree to their leaf sets. The second component quantifies additional error introduced by the BLeSS inference itself, and is dramatically increased by imputation (Table 1.1). However, this expected relationship is modulated by the influence of the source tree sampling scheme. When source trees are sampled at random, overall error as well as its second component are low when missing entries are treated as such, and increase especially drastically with imputation. The backbone-and-patch sampling scheme is less sensitive to imputation, as both types of error are higher in its absence, but increase to a lesser extent when it is employed (Fig. 1.5, Table 1.1). Plausibly, this differential effect of imputation might not be directly due to the sampling scheme itself, but rather to its covariates such as $\bar{\mathbf{D}}$ completeness, which strongly interacts with it (Supplementary Information, A.24) and whose distribution differs markedly between the two schemes (Supplementary Information, Table A.2).

In terms of root age accuracy and coverage, the single largest difference is observed between the simulations conducted with and without source tree error, with further influence exerted by the sampling scheme (Table 1.2). The correlation between true and estimated root ages is strong under all scenarios (R^2 range: 0.68–1), but introducing error into the profile leads to the overestimation of small root heights (by a factor of up to 21.2) and underestimation of large root heights (by a factor of up to 1.6; Supplementary Information,

			Root age posterior coverage (n)		
Topological error	Sampling scheme	Imputation	$\lambda_e = 0.1$	$\lambda_e = 0.33$	$\lambda_e = 1$
No	Backbone and patch	No	0.095 (21)	0.077 (13)	0.286 (7)
No	Backbone and patch	Yes	0.99 (98)	1 (95)	1 (86)
No	Random selection	No	0.989 (87)	1 (67)	1 (75)
No	Random selection	Yes	1 (98)	1 (92)	1 (77)
Yes	Backbone and patch	No	0 (31)	0 (29)	0 (26)
Yes	Backbone and patch	Yes	0 (45)	0 (41)	0 (40)
Yes	Random selection	No	0 (43)	0 (42)	0 (28)
Yes	Random selection	Yes	0 (97)	0 (97)	0 (97)

Table 1.2: Root age coverage probability (proportion of simulation replicates in which the true root age falls within the 95% highest posterior density interval) across the 24 scenarios illustrated in Fig. 1.2. The number of replicates which reached convergence for a given scenario and from which the coverage probability was calculated is given in parentheses next to the respective value.

Fig. A.28). In the absence of source tree error, the ratio of estimated to true root ages tends to approximate 1 extremely closely (Supplementary Information, Tables A.3–A.5), and the relative precision of the estimates (width of the 95% HPD interval divided by the true age) is extremely high (range of scenario-specific means: 0.0003–0.0060). While these values mostly translate to high coverage probabilities, this is not the case for analyses conducted under the backbone-and-patch sampling scheme and without imputation, for which the normalized 95% HPD interval widths are still narrow (mean values of 0.0060, 0.0040, and 0.0034 for $\lambda_e = 0.1$, 0.33, and 1, respectively) but the ratio of estimated to true root ages appreciably deviates from 1 (mean values of 0.803, 0.849, and 0.939 for $\lambda_e = 0.1$, 0.33, and 1, respectively). The introduction of source tree error does little to reduce relative precision (range of scenario-specific means: 0.0006–0.0103) but results in drastically degraded accuracy (Supplementary Information, Tables A.3–A.5). Since younger root ages tend to be overestimated to a much greater extent than old root ages tend to be underestimated, the scenario-specific mean ratios of estimated to true values always exceed 1 (range: 1.419–2.109). This combination of inaccurate but extremely precise estimates yields coverage probabilities that are

uniformly zero, meaning that the true values are never included in the corresponding 95% HPD intervals (Table 1.2). For the ages of the remaining nodes, the trend is similar but somewhat attenuated; in those replicates where at least some of the nodes present in the true tree are recovered in the MCC supertree, a nonzero number of such correctly recovered nodes is often also associated with 95% HPD age intervals that contain the true value (Supplementary Information, Fig. A.29). This may be due to the tendency of age estimates to become less precise for younger nodes (see below and Fig. 1.7).

Scalability tests

Encouragingly, our benchmarks show that the most time-intensive tasks needed for conducting BLeSS inference are those that only need to be performed once at the beginning of each analysis, including the calculation of the average distance matrix $\bar{\mathbf{D}}$, the initial draw from the birth-death prior, and fixing (or "clamping"; Höhna et al., 2014) a realization of the exponential error model to an observed value in the form of $\bar{\mathbf{D}}$ (Table 1.3). In actual analyses of very large average distance matrices, the real computational bottleneck is likely to consist of the conversion of a proposed supertree to a distance matrix ($\hat{\mathcal{T}} \rightarrow \hat{\mathbf{D}}$); at 10,000 tips, this operation is 13.5 times (imputation) to 35 times (no imputation) more time-consuming than the calculation of the least-squares difference that underlies the BLeSS likelihood function (Table 1.3). None of the steps involved in BLeSS inference currently benefits from the native RevBayes parallelization via MPI (Supplementary Information, Table A.11). As expected, the wall-clock runtime of BLeSS scales approximately with the square of tree size, with little difference between missing data treatments (i.e., sparse or dense average distance matrices; Supplementary Information, Table A.12).

At 34.5–38.8 and 8.8–9.2 moves per second, respectively, BLeSS analyses of 500-tip and 1000-tip supertrees take less than one week to complete 5 million single-move MCMC iterations (Supplementary Information, Fig. A.30), a number that is generally sufficient to attain

Operation	Time (s); no imputation			
	500 tips	1000 tips	5000 tips	10,000 tips
Convert source trees to distance matrices	0.118	0.213	0.816	2.385
Compute the average distance matrix $\bar{\mathbf{D}}$	0.022	0.086	1.908	7.805
Draw a tree $\hat{\mathcal{T}}$ from a birth-death process	0.032	0.123	3.122	12.646
Convert the proposed tree $\hat{\mathcal{T}}$ to a distance matrix $\hat{\mathbf{D}}$	0.033	0.138	4.161	18.634
Draw from the exp. error distribution conditional on $\hat{\mathbf{D}}$	0.071	0.289	7.996	33.968
Clamp the distribution to $\bar{\mathbf{D}}$	0.016	0.066	1.558	6.460
Compute the log probability density	0.002	0.006	0.133	0.532
	Time (s); imputation			
Read in an imputed distance matrix	0.098	0.442	38.145	289.265
Convert the imputed matrix to type `AverageDistanceMatrix`	0.031	0.143	5.459	18.266
Draw a tree $\hat{\mathcal{T}}$ from a birth-death process	0.032	0.124	3.097	12.580
Convert the proposed tree $\hat{\mathcal{T}}$ to a distance matrix $\hat{\mathbf{D}}$	0.033	0.140	4.181	18.534
Draw from the exp. error distribution conditional on $\hat{\mathbf{D}}$	0.071	0.294	7.985	33.776
Clamp the distribution to $\bar{\mathbf{D}}$	0.028	0.130	4.327	17.408
Compute the log probability density	0.003	0.012	0.262	1.374

Table 1.3: Wall-clock time spent on individual operations underlying BLeSS as a function of supertree size and missing data treatment. All values are averaged over 10 replicates and measured with millisecond precision for the single-core version of RevBayes v1.2.2 (commit b083532).

moderate (>200) to good (>625; Fabreti and Höhna, 2021) effective sample sizes for LnL and the root age, though not for the hyperparameters of the birth-death process (Supplementary Information, Fig. A.31). Note, however, that because of computational constraints, we did not calculate approximate topological ESS (Lanfear et al., 2016b) per unit runtime, so sampling efficiency remains unclear for a key parameter of interest. Moreover, even for the remaining parameters, our simplistic calculation of ESS increments per unit runtime likely underestimated sampling efficiency by not treating any of the initial iterations as pre-burnin (a period used for move tuning rather than sampling) or burnin (early samples to be discarded). Except for the speciation rate scaler, most moves exhibit acceptance rates that increase as a function of supertree size (Supplementary Information, Table A.13). However, based on an examination of the corresponding LnL traces, this relationship largely reflects the fact that the sampler had not yet burned it for very large supertrees (of 5000 or 10,000

tips), and proposed moves were highly likely to increase the likelihood of the initial low-probability parameter samples. For supertrees of 500 and 1000 tips, the acceptance rates of almost all moves fall short of the default target of 0.44 (by a factor of up to ∼5400 for the root age), resulting in relatively inefficient sampling (Supplementary Information, Table A.13). In practice, this problem can likely be mitigated by autotuning.

1.4.2 Empirical Application to the Order Carnivora

Some general trends were observable in the convergence diagnostics for the carnivoran analyses with and without imputed distances, regardless of whether source studies based on mitochondrial data were downweighted or not. Topological convergence across runs, as measured by ASDSF, was very good (< 0.01) for analyses without imputation, but less so when missing distances were imputed ($0.01 <$ ASDSF < 0.015). However, analyses without imputation took longer to reach stationarity, yielding lower per-run effective sample sizes for most parameters. In general, within-chain post-burnin ESS values were moderate (>200) for speciation and extinction rates as well as the root age, regardless of missing data treatment. Due to the large burnin proportions required for analyses without imputed distances, within-chain ESS values were low (~ 100) for LnL, LnPr, and their sum (log posterior), while they exceeded 200 in analyses based on the imputation of missing distances.

A common pattern where ESS values dropped for all but birth-death parameters and the root age when combining runs showed that each chain sampled a slightly different part of the joint posterior distribution. This effect was weakest when missing distances were treated as such (ESS > 100 in most cases), and strongest when they were imputed (ESS < 10 for some parameters). We therefore summarized only the single run with the highest post-burnin LnL values for analyses based on imputed distances, while both combining chains and summarizing the highest-likelihood chain alone for those without imputation (Supplementary Information, Table A.15). For the analysis in which λ_e was estimated, a visual inspection

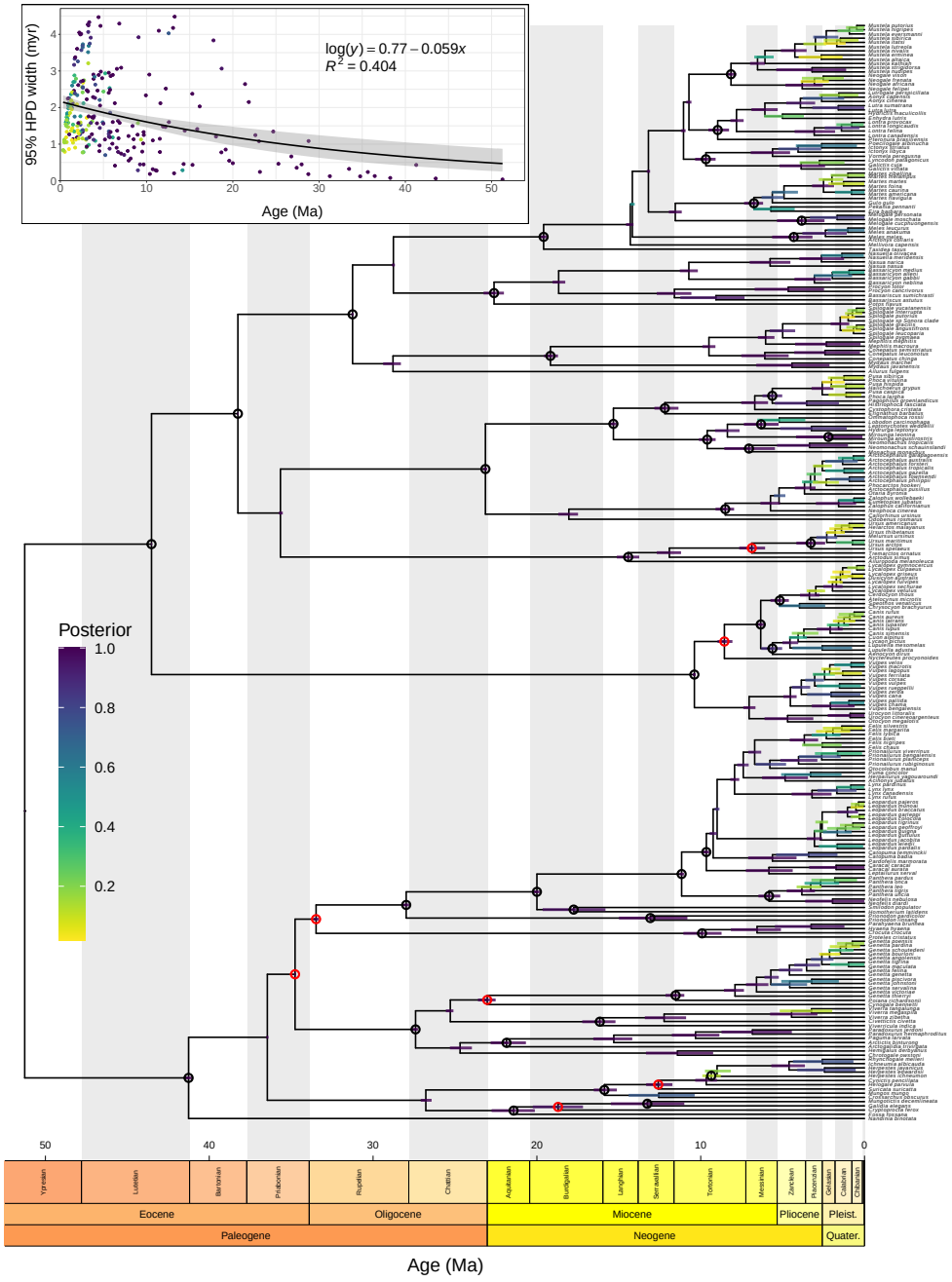

Figure 1.7: Results of a BLeSS analysis of 260 species of extant carnivorans. Main panel: the maximum clade credibility tree with mean node heights. The horizontal bars denote the 95% highest posterior density (HPD) intervals for the ages of internal nodes, and are colored by their posterior probability. The circles indicate nodes congruent with (black) or contradicting (red) the 53 predefined benchmark clades. Inset: the relationship between the precision of node age estimates (quantified as the width of the corresponding 95% HPD intervals) and their depth in the tree.

of the LnL traces indicated there was no need to discard initial samples as burnin, and combining all 4 chains resulted in ESS > 450 for all scalar parameters as well as an ASDSF of < 0.001. We therefore combined the chains for this analysis prior to summarizing. As in the 200-tip simulations, we found that when the MAP and MCC supertrees differed from each other, it was more often the latter that represented the more accurate point estimate (in terms of the number of benchmark clades recovered; Supplementary Information, Table A.15). Accordingly, we focus on the MCC supertrees when describing the relevant results.

Evaluation of benchmark clade recovery in MCC supertrees revealed that inference based on average distance matrices without imputed distances performed poorly. Regardless of whether we evaluated supertrees from combined runs or the single highest-likelihood run, the number of benchmark clades recovered was low (30–32%; $n = 16$–17). Missing benchmark clades spanned a range of taxonomic levels; at higher levels, we failed to recover a monophyletic Feliformia, Caniformia or Arctoidea, while at lower levels we missed a number of tribes and subtribes. Downweighting source trees inferred from mtDNA sequences had a limited effect on the MCC supertree derived from combined chains (36%; $n = 19$), but substantially increased the number of benchmark clades recovered in the summary tree from the highest-likelihood run (58%; $n = 31$), albeit with a similar hierarchical spread of missing clades. In contrast, supertrees inferred from average distance matrices with imputed missing distances consistently yielded a large number of benchmark clades (85–87%; $n =$ 45–46), with the highest number found in the MCC tree based on downweighted mitochondrial source trees (Fig. 1.7). This point estimate narrowly outperformed the NJ supertrees inferred with and without downweighting mtDNA-based source trees, both of which were consistent with the monophyly 44 benchmark clades (83%) when appropriately rerooted. The clades that BLeSS failed to recover even with imputation included not only subfamilies (Euplerinae, Hemigalinae, Mungotinae, as well as Herpestinae in the analysis without downweighting) and tribes (Arctotheriini, Vulpini) but also the infraorder Viverroidea (Viverridae

+ Herpestoidea) and superfamily Herpestoidea (Hyaenidae, Eupleridae, Herpestidae), which were contradicted by an unusual set of interfamilial relationships within feliforms. In contrast, both NJ supertrees additionally lack a monophyletic Galidinae, Canini, Phocini, and Canina, but do recover Arctotheriini, Viverroidea, and Herpestoidea (Supplementary Information, Figs. A.32, A.32). No benchmark clades were recovered in the analysis where λ_e was estimated.

The effect of missing data treatment also extended to precision of divergence time estimation. For analyses that treated missing distances as such, the mean precision of node age estimates is lower in both absolute (mean width of the 95% HPD intervals: 2.01 myr) and relative (mean ratio of 95% HPD interval widths to the corresponding posterior means: 0.881) terms, and there is a weak negative relationship (mean slope = -0.026; mean R^2 = 0.014) between posterior mean estimates and their absolute precision, which yields a slow decrease in expected HPD interval width from $\sim$ 2.2 myr (mean intercept) at shallow nodes to $\sim$ 0.8 myr at the root node. However, a stronger negative relationship of greater magnitude is found for inference based on imputed missing distances (mean slope = -0.047; mean $R^2 = 0.145$), which also yielded more precise estimates (mean absolute precision = 1.83 myr, mean relative precision = 0.783). Our preferred estimate illustrates that the actual difference between the widest (Galidinae: 4.48 myr; relative precision = 0.338) and narrowest (root: 0.0347 myr; relative precision = 6.77×10^{-4}) HPD intervals is larger than predicted by linear regression, and exponential regression produces a better fit ($R^2 = 0.404$; Fig. 1.7). A similar relationship is observed between node depth and node support, with the youngest and oldest nodes receiving the lowest and highest posterior probabilities, respectively (Fig. 1.7). Both correlations are expected, as deep nodes are informed by a greater number of $\bar{\mathbf{D}}$ entries than shallow ones, making it more difficult to fit a supertree containing a contradictory branching order or divergence times. The imputation of missing distances further increases this number, rendering the correlation even stronger.

The posterior mean root ages estimated by BLeSS (43.4–56.1 Ma) are plausible in light of the carnivoran fossil record and previous molecular dating studies. In particular, our preferred estimate of 51.3 Ma (associated with the supertree estimated from downweighted mitochondrial source trees and with imputed missing distances; Fig. 1.7) falls within the 95% HPD interval yielded by a recent phylogenomic study that was not included in the profile (Álvarez-Carretero et al., 2022). The extremely narrow 95% HPD interval about the root age never includes the prior mean (44.1 Ma) in the analyses where λ_e was fixed; however, in the analysis where λ_e was estimated as a free parameter, the posterior (43.4 Ma) and prior mean almost perfectly coincide, further suggesting that the resulting estimates are driven by the birth-death prior.

1.5 Discussion

BLeSS marries two previously introduced approaches to supertree inference (Lapointe and Levasseur, 2004; Steel and Rodrigo, 2008) into a single flexible framework that allows previously inferred time trees to be combined in a principled and statistically rigorous manner. Because it recasts the problem in a Bayesian setting, BLeSS addresses the two main shortcomings of existing supertree methods – the lack of branch lengths and the reliance on point estimates without accounting for the associated uncertainty – by providing the joint posterior distribution of supertree topologies and divergence times. This property allows it to move beyond the traditional dichotomy between "veto" and "voting" methods that has long featured in the supertree literature (Ranwez et al., 2007; Brinkmeyer et al., 2011), since a posterior sample of BLeSS supertrees needs neither to represent topological conflict as a lack of resolution nor to eliminate it by deciding in favor of one topology or the other. We provide a performant implementation of BLeSS in RevBayes (Höhna et al., 2016), validate it using a rigorous simulation-based calibration checking pipeline, and demonstrate its practicality for problems involving up to > 1000 tips. Using an extensive set of 200-tip

simulations, we further investigated the influence of a variety of analytical settings. Over the range of values we tested, the penalty parameter λ_e has only a weak effect on accuracy that is offset by its impact on the rate of MCMC convergence (Fig. 1.5; see below). The ability of BLeSS to incorporate source trees that overlap more extensively than those employed by the backbone-and-patch approach (Jetz et al., 2012; Tonini et al., 2016; Upham et al., 2019) represents a clear advantage, as shown by the fact that simulations with source trees whose leaf sets comprised randomly selected subsets of the full leaf set yielded more complete average distance matrices (Supplementary Information, Table A.2) that led to more accurate inference of both topologies and divergence times (Fig. 1.5; Table 1.1). The effect of imputing missing distances via ultrametric estimation (De Soete, 1984; Lapointe and Kirsch, 1995) is less straightforward and mediated by a number of other variables; however, at least in the empirical application of BLeSS presented here, imputation appears to outperform inference from a sparse average distance matrix even when using conflicting source trees.

Our analyses of both synthetic and empirical datasets show BLeSS to generally perform well under a wide range of conditions, albeit with several caveats. These are illustrated especially well by the empirical case, in which both the width of the 95% HPD intervals on node ages and topological uncertainty declines from the tips to the root (Fig. 1.7). The high precision of the age estimates for deep nodes may result in overconfidence and low coverage (i.e., lower than nominal probability of the true value falling within a credible interval). The tendency to overestimate the ages of shallow nodes and underestimate the ages of deep nodes in the presence of source tree conflict, which was revealed by our simulations (Supplementary Information, Fig. A.28), is also concerning and merits further investigation. At the same time, the flexibility of our implementation leaves open the possibility that this bias might be partially or completely mitigated by measures such as node calibration or downweighting of outlier source trees. Such analytical practices are particularly well-suited to the empirical application illustrated in this study, which involves estimating a time

tree that is more densely sampled and/or spans a more inclusive clade from pre-existing time trees that are less densely sampled and/or span less inclusive clades. However, it is worth noting that BLeSS may also be applicable to other problems for which Bayesian supertrees have been proposed as a solution, such as devising divide-and-conquer strategies for *de novo* analysis of large datasets (Karcher et al., 2021) or the inference of species trees from incongruent gene trees (De Oliveira Martins et al., 2016). Investigating the statistical properties of BLeSS in the latter setting may be particularly worthwhile given recent results showing that relatively simple species tree estimators perform well even in circumstances that render more sophisticated supermatrix analyses statistically inconsistent (Dasarathy et al., 2015; Allman et al., 2019). Below, we outline several other analytical considerations, practical recommendations, and possible extensions that are broadly applicable to a variety of BLeSS use cases.

1.5.1 Influence of the exponential error rate parameter

Despite exploring a relatively narrow range of possible λ_e values, our analyses of simulated and empirical data show this penalty term to exert influence on supertree estimation in terms of both accuracy and MCMC performance. While treated here as a parameter that is internal to the BLeSS model and determines the shape of its posterior, λ_e is closely related to auxiliary variables employed in MCMC tempering techniques to sample from distributions other than the posterior, which helps explain its behavior and provide guidance regarding its treatment.

To the extent that the $\log \lambda_e$ term in Eq. 5 is negligible relative to the $\lambda_e \delta(\hat{\mathbf{D}}, \bar{\mathbf{D}})$ term, λ_e approximates the temperature parameter t used to construct "power posterior" distributions of the form $f_t(\theta|y) = p(\theta) f(y|\theta)^t$ between the unnormalized posterior ($t = 1$) and the prior ($t = 0$) (Gelman and Meng, 1998; Friel and Pettitt, 2008). Such power posteriors are commonly used in Bayesian model comparison to estimate marginal likelihoods using path-sampling (Gelman and Meng, 1998; Lartillot and Philippe, 2006) and stepping-stone

(Fan et al., 2011; Xie et al., 2010) techniques. Like the temperature parameter, λ_e scales the likelihood relative to the prior. This feature is potentially problematic, since manipulating λ_e could in theory be used to induce a posterior that is dominated to an arbitrary extent either by the user's prior beliefs about the branching process, or by the information contained in the average distance matrix. In practice, however, MCMC mixing imposes constraints on the range of values that are viable for any given analysis (see below).

While the arbitrariness in the choice of λ_e can conceivably be circumvented by estimating it from the data as a free parameter, our attempts to do so in preliminary analyses of the carnivoran dataset yielded unsatisfactory results. On the one hand, the posterior, which was approximately lognormal with a narrow peak close to 0 (mean = 5.4×10^{-8}, 95% HPD = $[6.60 \times 10^{-10}, 1.37 \times 10^{-7}]$ based on the pooled posterior sample from 4 runs), diverged sharply from the prior (a diffuse exponential with a mean of 1), suggesting that the average distance matrix is informative with respect to the value of λ_e. This is not unexpected, as distances averaged across source trees with a high degree of conflict will substantially deviate from ultrametricity, and any ultrametric supertree will thus only fit them imperfectly, favoring low values of λ_e. Conversely, low amounts of conflict allow the supertree to strictly adhere to the near-ultrametric average distance matrix, making it possible to impose a heavier penalty (corresponding to a high value of λ_e) for deviations therefrom. On the other hand, because the estimated penalty was extremely low, the posterior was entirely dominated by the birth-death prior, whose contribution exceeded that of the BLeSS likelihood term by more than 1800 log units. As a result, the analysis in which the supertree and λ_e were co-estimated performed extremely poorly in practice, failing to recover even the best-supported benchmark clades that received a posterior probability of 1 in the analyses where λ_e was fixed.

In addition to determining the prior-to-likelihood ratio, λ_e also modulates the ruggedness of the likelihood surface. To the extent that the prior is negligible relative to the likelihood, its behavior approximates the β parameter (also referred to as "temperature") employed in Metropolis-coupled MCMC (also known as MC3 or parallel tempering; Geyer, 1991). Unlike power posterior sampling, in which only the likelihood term is exponentiated, MC3 raises the entire posterior to a given temperature, producing a "heated" or "tempered" distribution of the form $f_\beta(\theta|y) = p(\theta)^\beta f(y|\theta)^\beta$. Choosing $\beta \in (0, 1)$ has the effect of flattening the target distribution, making it easier to explore multimodal posteriors by traversing the low-probability valleys between local optima (Altekar et al., 2004). If swaps are attempted between the chain sampling from the posterior and parallel chains sampling from the tempered distributions, the former can benefit from the rapid mixing of the latter (Gilks and Roberts, 1996). These improvements in mixing closely resemble the behavior we observed with low values of λ_e. In particular, our 200-tip simulations show that decreasing λ_e improves mixing (quantified by the number of replicates satisfying our convergence criteria) at the cost of accuracy (quantified by the distance between the true and estimated supertrees) (Fig. 1.5). The same trade-off was also apparent in the empirical analyses. When λ_e was freely estimated, its extremely low values made it easy for the 4 chains to converge to the prior-dominated posterior (ESS > 450, potential scale reduction factor < 1.015 for all scalar parameters; ASDSF < 0.001). In contrast, analyses performed with fixed values of λ_e often yielded parameter traces in which sudden shifts punctuated long periods characterized by a very low amount of variation. Similarly, different chains almost always exhibited non-overlapping traces, with among-chain variation exceeding within-chain variation by several orders of magnitude. As a result, pooling the resulting samples resulted in extremely low ESS values. In our final analyses, this problem mostly affected LnL traces, whereas the chains mixed relatively well for the main parameters of interest such as the root height and (based on the ASDSF) topology. However, in several preliminary analyses ($\lambda_0 = 0.1$–0.33,

∼325,000 iterations), the same behavior was also observed for the root age, with different chains sampling from different distributions even though the difference between the means of these distributions (< 0.2 myr) was well within the desired margin of error.

By analogy with t and β, we believe that λ_e is better treated as an auxiliary variable exterior to the model, and that $\lambda_e = 1$ represents the most natural choice when the unmodified posterior is of interest. If this relatively large value leads to convergence issues, these should be addressed using techniques such as MC3 rather than by manipulating λ_e itself. One consequence of always setting the penalty term to a unit value regardless of tree size is that the posterior will be dominated by the prior for small supertrees (as observed in the SBC analyses with 4–5 tips) and by the likelihood for large supertrees (as observed in the benchmark analyses with > 1000 tips). This fact is probably best interpreted as a genuine feature of the relationship between data and evidence under the BLeSS model: larger average distance matrices can discriminate among trees much more effectively than smaller ones. A fruitful avenue for further research would be to investigate the sensitivity of BLeSS to the choice of the time unit (Groussin et al., 2011). The functional forms of both the BLeSS likelihood (Eq. 5) and the birth-death prior contain exponentiated branch durations, and both will therefore respond to a change of the time unit, possibly in ways that affect their relative weights.

1.5.2 Potential efficiencies

Our simulations suggest that the current implementation of BLeSS is practical for supertrees of up to ∼1000 tips. This is approximately equivalent to the largest tree sizes at which full Bayesian co-estimation of tree topology and divergence times from sequence data remains feasible (e.g., du Plessis et al., 2021; Wisniewski et al., 2022), and substantially smaller than the recent megaphylogenies inferred by the time-scaling of maximum-likelihood

trees or backbone-and-patch techniques (Jetz et al., 2012; Tonini et al., 2016; Rabosky et al., 2018; Upham et al., 2019). At first, BLeSS would therefore seem to offer little advantage over existing methods. Our benchmarks show it is the conversion of a proposed tree to a distance matrix that represents the main computational bottleneck in the current implementation of the method (Table 1.3). Accordingly, future efforts to improve the performance of the method should focus on exploiting efficiencies that can be achieved at this step.

Importantly, standard tree-perturbing MCMC moves implemented in RevBayes, such as nearest neighbor interchange (NNI), subtree prune and regraft (SPR), a whole-tree scaler, or a node time slider, either only affect several branches at a time, leaving large subtrees of the original tree intact, or scale all branches by the same factor. In both cases, recomputing $\hat{\mathbf{D}}$ every time a tree has been updated is clearly wasteful. In the analyses of character data, this problem is circumvented by reusing partial likelihoods stored at the corresponding nodes or branches (Smith et al., 2024). However, unlike the likelihoods that result from modeling character substitution along the branches of a tree, which are computed using Felsenstein's pruning algorithm (Felsenstein, 1981), the BLeSS likelihoods (Eq. 5) are not evaluated by means of a tree traversal. This solution is therefore not applicable to them, since calculating subtree-specific likelihoods would increase (rather than reduce) the total number of computations performed. Nevertheless, considerable economies could be achieved by storing submatrices of $\hat{\mathbf{D}}$, rather than partial likelihoods, at specific nodes or branches. This would make it possible to update only those distances that changed after a given move, and reuse the submatrices corresponding to identical subtrees. However, storing submatrices at each of the $n - 2$ non-root nodes or internal branches would drastically increase the memory footprint of BLeSS analyses, which is already substantial at 10,000 tips. A more efficient approach may be to only extract the submatrix after a move picks the node or branch on which to perform a given operation. This solution would require less far-reaching changes to the RevBayes code base (only involving a re-implementation of common tree moves as

overloaded functions that operate differently when applied to tree objects or distance matrix representations thereof), and may offer an ideal trade-off between CPU time and memory usage.

Although the conversion of trees to distance matrices is by far the most computationally demanding step in BLeSS inference, our results show that the burden associated with likelihood evaluation proper (i.e., computing the squared differences of the non-missing elements of the two distance matrices) also becomes considerable at large tree sizes. Even if the time spent on all other operations could be reduced to zero, it would still take ∼379 hours (over two weeks) to run an MCMC analysis of 10^6 iterations on a dense (i.e., imputed) 10,000-by-10,000 average distance matrix (Table 1.3). However, as an element-wise operation, matrix subtraction is both low in computational complexity (being solvable in $\mathcal{O}(n^2)$ time) and embarrassingly parallel (Hunold et al., 2011; Chang et al., 2023). The task can be divided into as many submatrices as there are available processors, and assuming that both the average distance matrix $\bar{\mathbf{D}}$ and the path-length matrix $\hat{\mathbf{D}}$ of the proposed tree can be distributed among the processors in alignment (i.e., that for all i, j, $\bar{\mathbf{D}}_{i,j}$ and $\hat{\mathbf{D}}_{i,j}$ are assigned to the same processor), no interprocessor communication is required between the initial distribution step and the final re-assembly of the difference matrix (Nayak, 1999; Hunold et al., 2011; Gu et al., 2015; see Östermark, 2000 for the total cost incurred by communication overhead on a 2D mesh of processors). As a result, the use of p processors reduces the computational cost of the problem to $\mathcal{O}(n^2/p)$, potentially leading to near-linear speedup (Nayak, 1999; Hunold et al., 2011; Chang et al., 2023). Because of this feature, large matrix subtraction is particularly well-suited to graphics processing units (GPUs), which are designed to support a large number of lightweight concurrent threads (Le, 2012). Thanks to recent efforts (Smith et al., 2024), RevBayes can already exploit GPUs for calculations of standard phylogenetic likelihoods by offloading them to the high-performance BEAGLE library (Suchard and Rambaut, 2009; Ayres et al., 2012), which supports both CUDA and OpenCL platforms (Ayres et al.,

2019). Extending its GPU support to the evaluation of BLeSS likelihoods would represent a fruitful avenue for increasing the efficiency of the method.

1.5.3 Generalizing BLeSS to non-ultrametric trees

Throughout this study, we have assumed that both the source trees and the estimated supertree represented chronograms of extant taxa, and as such were ultrametric. However, this restriction is not inherent to the method, and a way to relax it was devised early on in the development of average consensus approaches (Lapointe and Cucumel, 1997). As a result, it became possible to estimate average consensus supertrees from source trees that were merely additive rather than ultrametric. These were originally supposed to represent phylograms, but in the context of BLeSS, they can be interpreted as serially sampled time trees, i.e., time trees with non-contemporaneous tips.

There is a growing interest in extending comparative methods that were originally developed for extant taxa to paleontological data (Slater, 2013, 2014; Bapst, 2014; Soul and Wright, 2021), but the ability to investigate clade-wide evolutionary trends is severely hampered by the lack of fossil trees comparable in size to extant megaphylogenies inferred from molecular data. Phylogenetic analyses of extinct taxa are limited in size by the costly, time-consuming, and labor-intensive nature of collecting morphological data, which have to be manually scored from the relevant specimens (often spread across many different collections worldwide) based on first-hand observations or published illustrations (Burleigh et al., 2013; Eliason et al., 2019). This problem has not yet been alleviated by new techniques such as crowdsourcing (Chang and Alfaro, 2016; O'Leary et al., 2018) or deep learning (Hunt and Pedersen, 2022; Hodel et al., 2023; Tsutsumi et al., 2023; Weaver and Smith, 2023), which largely remain at the proof-of-concept stage. As a result, even the largest morphological phylogenetic datasets usually do not include more than 200 operational taxonomic units (OTUs) (Supplementary Information, Fig. A.34); the most taxon-rich character matrices of which we

are aware (Hartman et al., 2019; Cau, 2024) represent clear outliers with ∼500 OTUs. Moreover, since different datasets may encode the same morphological features using different subjectively delimited characters and character states, combining them will typically require an extensive process of character redefinition and rescoring whose difficulty does not substantially differ from collecting data *de novo* (Lloyd and Slater, 2021). Accordingly, despite attempts to promote the reuse of previously assembled morphological datasets (O'Leary and Kaufman, 2011), these cannot be easily concatenated into a single taxon-rich supermatrix in a manner reminiscent of molecular alignments, and supertrees remain the only viable solution to the problem of generating comprehensive large-scale fossil phylogenies (Lloyd and Slater, 2021).

A number of comparative and macroevolutionary paleontological studies have adapted to this constraint by constructing informal (i.e., hand-drawn) supertrees and subjecting them to *post hoc* time-scaling using simple techniques based on first appearance dates (Benson et al., 2014; Famoso et al., 2016; Sakamoto et al., 2016; MacLaren, 2021; Kiat and O'Connor, 2024). This procedure has become common enough that dedicated software packages are now available to facilitate it (Castiglione et al., 2022). However, regardless of their plausibility, the resulting time trees amount to little more than expert opinions, and the process of generating them lacks the basic features of scientific workflow, including testability (since there are no underlying data to which competing expert opinions could be fitted) and reproducibility. Matrix representation with parsimony (MRP) supertrees avoid this weakness but suffer from other shortcomings, particularly the lack of a satisfying treatment of branch length information and topological uncertainty (see Introduction and Lloyd and Slater, 2021 for a recent review).

BLeSS represents an attractive alternative to both informal and MRP supertrees, and even its current implementation can considerably increase the size of phylogenies available to paleontologists. Its generalization to the non-ultrametric case would simply require replac-

ing the simple birth-death prior used throughout this study with an alternative branching-process prior accounting for serial sampling, such as the uniform prior of Ronquist et al. (2012a) or the fossilized birth-death (FBD) process (Stadler, 2010; Gavryushkina et al., 2014; Heath et al., 2014). When imputing missing entries in $\bar{\mathbf{D}}$, the three-point ultrametric condition (Hartigan, 1967) would further need to be replaced by the four-point additive condition (Buneman, 1974) to account for the fact that root-to-tip path lengths can no longer be assumed equal (Lapointe and Levasseur, 2004). A potential weakness of BLeSS in paleophylogenetic settings is its need for time-calibrated source trees, which may limit its applicability given that fossil phylogenies have historically been predominantly inferred using time-free methods. However, this problem is rapidly disappearing thanks to the growing popularity of Bayesian tip-dating under the FBD model (Wright et al., 2022) and availability of rigorous *post hoc* time-scaling methods such as *cal3* (Bapst, 2013). We expect BLeSS to provide a viable and powerful solution to the problem of inferring large fossil and total-evidence phylogenies, facilitating the use of information from extinct taxa in downstream analyses that currently rely mainly or exclusively on data from extant lineages.

1.6 Acknowledgements

We are indebted to Tracy Heath for her help in clarifying the conceptual basis of the method, and to Walker Pett and Joëlle Barido-Sottani for their invaluable assistance with the implementation of BLeSS in RevBayes. We thank Michael Landis for help with simulating sequence data, Luke Kelly for advice on diagnosing MCMC convergence, and Sebastian Höhna for guidance on designing RevBayes validation tests. We further wish to thank Michael Foote, Andrew Magee, Fernando Meléndez-Vazquez, Erin Molloy, Jonathan Nations, C. Tomomi Parins-Fukuchi, Benjamin Redelings, Richard Ree, and Ulises Rosas-Puchuri for enlightening discussions and encouragement. We are grateful to Aaron Schein for his suggestion to use Bayesian additive regression trees to evaluate our simulation results, and to Claudio

Kozický for his help with making the RevBayes implementation of BLeSS more efficient, as well as for insightful comments about possible future directions. All simulations described in this study were performed on the Midway2 and Midway3 Research Computing Clusters at the University of Chicago.

1.7 Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.[NNNN].

1.8 Funding

This work was supported by a Neubauer Family Distinguished Doctoral Fellowship from the University of Chicago and a Graduate Student Research Award from the Society of Systematic Biologists to D.Č.

CHAPTER 2

STATISTICAL EVALUATION OF CHARACTER SUPPORT REVEALS THE INSTABILITY OF HIGHER-LEVEL DINOSAUR PHYLOGENY

2.1 Abstract

The interrelationships of the three major dinosaur clades (Theropoda, Sauropodomorpha, and Ornithischia) have come under increased scrutiny following the recovery of conflicting phylogenies by a large new character matrix and its extensively modified revision. Here, we use tools derived from recent phylogenomic studies to investigate the strength and causes of this conflict. Using maximum likelihood as an overarching framework, we examine the global support for alternative hypotheses as well as the distribution of phylogenetic signal among individual characters in both the original and rescored dataset. We find the three possible ways of resolving the relationships among the main dinosaur lineages (Saurischia, Ornithischiformes, and Ornithoscelida) to be statistically indistinguishable and supported by nearly equal numbers of characters in both matrices. While the changes made to the revised matrix increased the mean phylogenetic signal of individual characters, this amplified rather than reduced their conflict, resulting in greater sensitivity to character removal or coding changes and little overall improvement in the ability to discriminate between alternative topologies. We conclude that early dinosaur relationships are unlikely to be resolved without fundamental changes to both the quality of available datasets and the techniques used to analyze them.

2.2 Introduction

The relationships between the three major clades that comprise Dinosauria (Theropoda, Sauropodomorpha, and Ornithischia) have historically been uncertain. In his seminal work that coined the name Ornithischia for herbivorous dinosaurs characterized by an opisthopubic pelvis, Seeley (1887) regarded this taxon as only distantly related to the theropods and sauropodomorphs, which he grouped together as Saurischia (Fig. 2.1). This view of dinosaur polyphyly later extended to Saurischia itself (Romer, 1956; Charig et al., 1965; Reig, 1970; Thulborn, 1975; Charig, 1982), leaving the ornithischians, theropods, and sauropodomorphs as three lineages independently descended from non-dinosaurian archosaurs of uncertain interrelationships.

The recognition of dinosaur monophyly failed to immediately clarify the relationships among the three clades, as the earliest arguments in favor of a monophyletic Dinosauria were coupled not only with the continued use of the Saurischia–Ornithischia dichotomy, but also with the seemingly contradictory suggestion that the ornithischians may have evolved from "prosauropods" (= early sauropodomorphs) (Bakker and Galton, 1974; Bonaparte, 1976; Cooper, 1981). In the mid-1980s, the latter scenario was formalized as a phylogenetic hypothesis linking Ornithischia and Sauropodomorpha to the exclusion of Theropoda (Paul, 1984, 1988; Sereno, 1984) in a group variously termed Ornithischiformes (Cooper, 1985) or Phytodinosauria (Bakker, 1986) (Fig. 2.1). However, this hypothesis was abandoned after the first rigorous application of algorithmic phylogenetics to nonavian dinosaurs by Gauthier (Gauthier, 1986), who provided detailed character evidence uniting the theropods and sauropodomorphs into a monophyletic Saurischia, cementing a view of dinosaur phylogeny that would remain uncontested for the following three decades (Novas, 1996; Benton, 2004; Nesbitt, 2011).

Using a large new dataset, Baron et al. (2017a; henceforth BEA) recently destabilized this view by lending support to the third possible (and previously unforeseen) way of resolving the

branch in question, namely, a clade formed by Theropoda and Ornithischia to the exclusion of Sauropodomorpha, for which the authors adopted Huxley's (1870) name Ornithoscelida (Fig. 2.1). Seven months later, Langer et al. (2017; henceforth LEA) published a response using a rescored version of BEA's character matrix with 9 taxa added, which recovered the traditional Saurischia–Ornithischia dichotomy, albeit with virtually no statistical support. The resulting controversy surrounding early dinosaur phylogeny has not been resolved by subsequent attempts to further rescore or add supposed key taxa such as *Pisanosaurus* (Baron et al., 2017b; Baron, 2019) and *Chilesaurus* (Baron and Barrett, 2017; Müller et al., 2018; Baron and Barrett, 2018; Müller and Dias-da Silva, 2019), nor by the use of more sophisticated phylogenetic methods such as time-free Bayesian inference (Parry et al., 2017) and Bayesian tip-dating (Griffin et al., 2022).

Despite the considerable interest generated by BEA's contribution, there have been few attempts to quantify the relative support for each of the candidate topologies, distinguish between low information content and internal conflict, or identify the characters that may drive such conflict – lines of investigation that are more typical of phylogenomics (Reddy et al., 2017; Shen et al., 2017; Pease et al., 2018) than morphological phylogenetics. LEA took early steps in this direction by demonstrating that when their data was analyzed in a parsimony framework, the reinstated Saurischia hypothesis was statistically indistinguishable from either of the alternatives (Langer et al., 2017). However, subsequent studies have mostly reverted to reporting a single point estimate of the phylogeny, without testing whether its support significantly exceeded that of the next best hypothesis. The occasional attempts to use the number of extra steps relative to the most parsimonious tree for this purpose (Baron et al., 2017a; Baron and Barrett, 2017) suffer from the fact that this quantity has no statistical interpretation (Felsenstein, 2004). Moreover, despite the acknowledged centrality of character scoring differences to the conflict among the resulting topologies (Langer et al., 2017; Baron et al., 2017b; Müller et al., 2018; Baron and Barrett, 2018), only one study to date has

attempted to determine which rescorings drove the difference between the hypotheses of early dinosaur phylogeny favored by BEA's and LEA's datasets (Goloboff and Sereno, 2021), and its methodological scope was limited to parsimony.

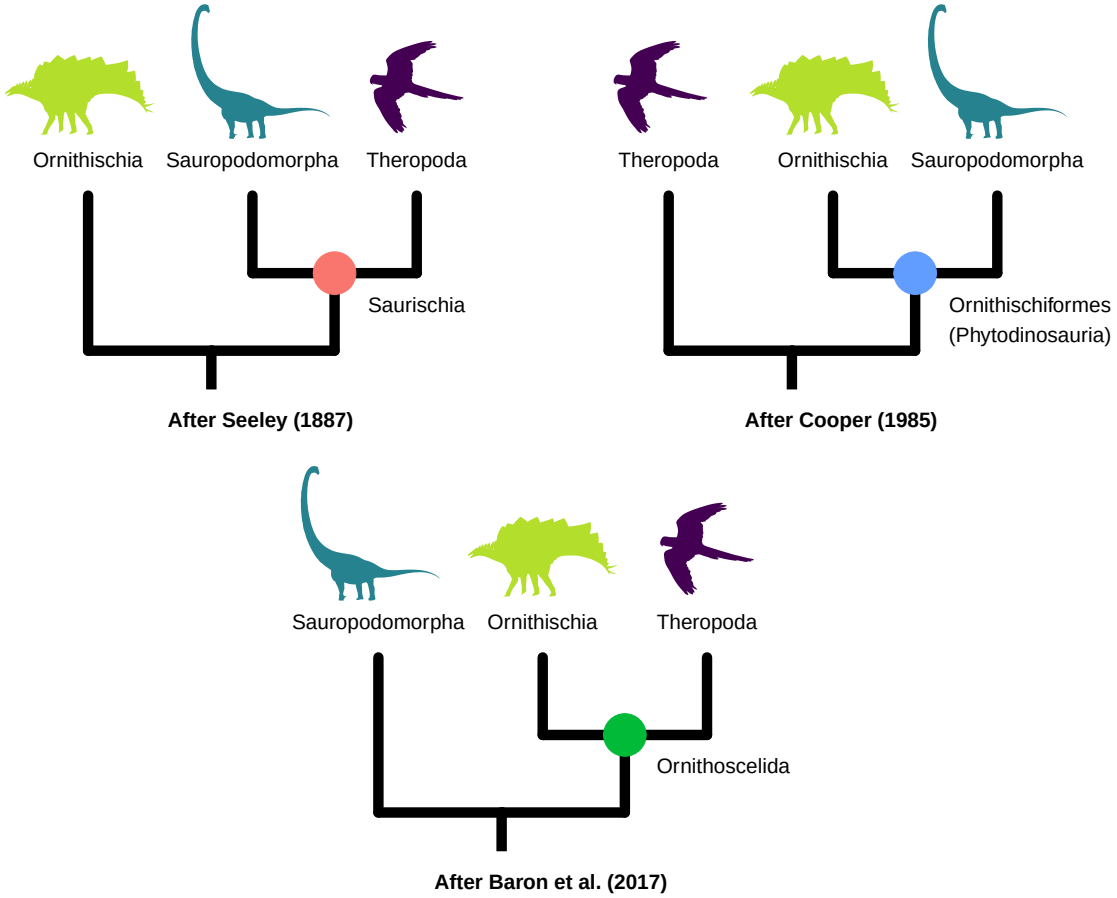

Figure 2.1: Three alternative hypotheses of early dinosaur phylogeny evaluated in this study. Dinosaur silhouettes from PhyloPic (`http://phylopic.org`) courtesy of Matthew Dempsey (CC BY 4.0), Jagged Fang Designs (CC0 1.0), and Sharon Wegner-Larsen (CC0 1.0). The figure was generated in R v4.2.2 (R Core Team, 2021) using the packages *ape* v5.6-2 (Paradis et al., 2004; Paradis and Schliep, 2018), *ggtree* v3.6.2 (Yu et al., 2017), *ggplot2* v3.4.0 (Wickham et al., 2022a), *ggforce* v0.4.1 (Pedersen, 2022a), *rphylopic* v1.0.0 (Gearty et al., 2023), and *viridisLite* v0.4.1 (Garnier et al., 2022).

Here, we use statistical tools drawn from phylogenomics to evaluate the relative support for the three hypotheses of large-scale dinosaur phylogeny in the BEA and LEA datasets, and to conduct detailed assessments of character support. We demonstrate the presence of per-

vasive conflict both across each dataset as a whole and among the subsets of characters with the strongest phylogenetic signal. We further show that although LEA's extensive changes to BEA's character coding dramatically altered the distribution of phylogenetic signal across the matrix, they did little to help discriminate among the three alternative topologies, which remain indistinguishable both before and after LEA's recoding. Our results suggest that there are many more plausible hypotheses of early dinosaur phylogeny than usually acknowledged, and that selecting between them may be beyond the reach of current character matrices and the techniques used to analyze them. We provide recommendations pertaining to both data-related and methodological aspects of the problem, and conclude that care should be taken to properly account for the uncertainty surrounding higher-level dinosaur phylogeny in downstream analyses.

2.3 Methods

2.3.1 Data

Our analyses were performed on the original character matrices used by BEA and LEA, whose properties are summarized in Table 2.1. Both were obtained from Graeme T. Lloyd's database of previously published character matrices with standardized formatting (`http://graemetlloyd.com/matrdino.html`; last accessed May 1, 2022). Both BEA and LEA used *Euparkeria capensis* and *Postosuchus kirkpatricki* as outgroup taxa. The two datasets differed in 3366 scorings (10.0% of the total number of overlapping cells) when polymorphic codings were treated as distinct from missing data, and in 3350 scorings (9.9% of the total number of overlapping cells) when treating the two as equivalent (as in the maximum likelihood analyses). Only 4 of the 74 overlapping taxa (*Dromomeron gigas*, *Dromomeron gregorii*, *Dromomeron romerii*, *Postosuchus kirkpatricki*) and 10 of the 457 characters (50–52, 118, 121–123, 152, 250, 344) were unaffected by the changes. The much higher number of differ-

ences previously reported in the literature (8050 scorings, or 21.2% of the total number of LEA's cells; Goloboff and Sereno, 2021) also reflects non-overlapping cells (representing the extra taxa added by LEA) as well as the notational distinction between missing data ("?") and inapplicables ("-"), which has no analytical significance in currently used phylogenetic algorithms (but see Goloboff et al., 2021; Hopkins and St. John, 2021; Tarasov, 2023). Both datasets are organized by anatomical region, with cranial characters (1–146) followed by dental (147–185), axial (186–236), and pectoral (237–250) characters, and finally by characters pertaining to the forelimb (251–291), pelvis (292–351), and the hindlimb (352–457).

2.3.2 Maximum likelihood analyses

We used maximum likelihood (ML) in our analyses of BEA's and LEA's datasets for several reasons. First, we found it useful to explore how the resulting topologies might change when using a parametric rather than nonparametric approach to phylogenetic inference, since method-dependent results may indicate the presence of within-dataset conflict (Betancur-R. et al., 2019). Second, despite its continued widespread use in the paleontological literature, maximum parsimony is well-known for its undesirable statistical properties compared to model-based methods (Felsenstein, 1978), including in the context of morphological phylogenetics (Wright and Hillis, 2014; O'Reilly et al., 2016, 2018; Puttick et al., 2019). Third, topologically constrained ML analyses allowed us to directly compare support between the three hypotheses using a number of well-established and easily interpretable frequentist tests.

All maximum likelihood analyses were performed using IQ-TREE v2.1.3 (Minh et al., 2020) with the default mix of starting trees (1 BioNJ + 99 parsimony trees). Both datasets were partitioned first into ordered and unordered characters and further by the number of character states, for a total of 6 partitions. In contrast to the number reported by the original authors (Table 2.1), only 36 and 37 characters from the BEA and LEA datasets were treated as ordered, respectively, since for characters 24, 334, and (for BEA) 180, only

states 0 and 1 were observed. Branch lengths were treated as proportional among partitions (the `-p` command-line option in IQ-TREE), and unordered characters were assigned the Mk model (Lewis, 2001) with k ranging from 2 to 5. Constant characters (BEA: 29, 59, 150, 245, 248, 268; LEA: 29, 59, 75, 112, 139, 150, 248, 268, 288) were excluded, and an ascertainment bias correction (Lewis, 2001) was applied to all partitions. Unlinked discrete gamma models of among-character rate heterogeneity were added to every substitution model except those applied to the character-poor unordered 5-state and ordered 4-state partitions.

2.3.3 Difficulty assessment and tree searches

A number of methods have been developed to quantify the expected difficulty of phylogenetic analysis or the amount of data needed to resolve a particular node (Braun and Kimball, 2001; Poe and Chubb, 2004; Verbruggen et al., 2010); however, these often rely on models that are inapplicable to morphological evolution, such as the multispecies coalescent (Sayyari and Mirarab, 2018). To evaluate how challenging it would be to estimate early dinosaur interrelationships from the BEA and LEA datasets in a maximum-likelihood framework, we performed 100 topologically unconstrained tree searches on each dataset, and used the resulting ML trees to calculate a nonparametric difficulty measure recently proposed by Haag et al. (2022):

$$\text{difficulty} = \frac{1}{5}\left[\bar{d}_{\text{all}} + \bar{d}_{\text{pl}} + \frac{n'_{\text{all}}}{n_{\text{all}}} + \frac{n'_{\text{pl}}}{n_{\text{pl}}} + \left(1 - \frac{n_{\text{pl}}}{n_{\text{all}}}\right)\right], \tag{1}$$

where $\bar{d}_{\text{all}}$ is the average pairwise normalized Robinson-Foulds (RF; Robinson and Foulds, 1981) distance between the $n_{\text{all}} = 100$ inferred trees, n'_{all} is the number of unique topologies among the 100 ML trees, $\bar{d}_{\text{pl}}$ is the average pairwise normalized RF distance within a subset of plausible trees, n_{pl} is the number of trees included in this set, and n'_{pl} is the number of

unique topologies present in the plausible set. Each term of Eq. (1) ranges from 0 to 1, as does the overall difficulty score equal to their unweighted mean. Depending on the resulting value, ML phylogenetic inference can range from trivial (0) to effectively impossible (1). Following Morel et al. (2020), the plausible tree set was constructed from all trees that were not found to be significantly worse than the best-scoring tree by any of the likelihood-based tests implemented in IQ-TREE. These were conducted using 10,000 approximate-bootstrap replicates (`-zb 10000 -zw -au`) generated by the resampling estimated log-likelihood method (Kishino et al., 1990) and included the Kishino-Hasegawa test (Kishino and Hasegawa, 1989), the unweighted and weighted SH test (Shimodaira and Hasegawa, 1999), the approximately unbiased test (Shimodaira, 2002), and expected likelihood weights (Strimmer and Rambaut, 2002). Postprocessing was carried out in the R statistical computing environment (R Core Team, 2021) using the packages *phytools* v1.2-0 (Revell, 2012), *TreeTools* v1.9.0 (Smith, 2019c), and their respective dependencies.

After the difficulty assessment, we performed one more round of unconstrained ML searches consisting of 10 runs (each with 100 starting trees). The overall ML estimate was obtained by selecting the best-scoring tree from the pooled sample of the 100 exploratory and 10 final runs. To assess clade support, we additionally performed ultrafast bootstrap approximation (UFBoot) with 1000 replicates (`-B 1000`), either after the fact (if the best-scoring tree was found during the exploratory runs) or simultaneously with the main tree search (for the last 10 runs). In addition to being relatively robust to model misspecification, the ultrafast bootstrap is less biased than the standard nonparametric bootstrap (Hoang et al., 2017), and we also found it to be more numerically stable. Finally, we carried out topologically constrained analyses enforcing those hypotheses that were not supported by the unconstrained tree. Similar to the main analyses, each consisted of 10 runs performed simultaneously with 1000 UFBoot replicates.

2.3.4 Character-wise support

We followed the protocol of Shen et al. (2017) to assess how support for the three competing hypotheses (S = Saurischia, Of = Ornithischiformes, Os = Ornithoscelida) was distributed across characters in both matrices, and to identify potential outliers. Using IQ-TREE, we estimated character-wise log-likelihood values for the ML trees yielded by both unconstrained and topologically constrained searches (`-wslr`). We first calculated the number of characters in either matrix that preferred a given hypothesis (i.e., yielded the least negative log-likelihood value under it), and repeated this calculation for individual anatomical partitions. Multinomial tests were carried out using the R package *EMT* v1.2 (Menzel, 2021) to determine whether the resulting distributions were significantly different from uniform. We further determined the number of characters that ranked the three hypotheses in the same way in terms of their log likelihoods between the two matrices. Next, we calculated the phylogenetic signal (PS) of the i-th character (C_i) in a given matrix following Eq. (5) of Shen et al. (2017):

$$\mathrm{PS}_i = \frac{|\ln L(C_i \,|\, \mathrm{S}) - \ln L(C_i \,|\, \mathrm{Of})| \;+\; |\ln L(C_i \,|\, \mathrm{S}) - \ln L(C_i \,|\, \mathrm{Os})| \;+\; |\ln L(C_i \,|\, \mathrm{Of}) - \ln L(C_i \,|\, \mathrm{Os})|}{3} \qquad (2)$$

Since the observed distributions of PS values were heavier-tailed than those shown in Shen et al. (2017), we used a different criterion to identify outliers, defining them as those characters whose PS was more than three standard deviations above the mean (Francis and Canfield, 2020). To assess the influence exerted by characters with strong phylogenetic signal on the resulting early dinosaur topologies, we generated 8 subsampled datasets by removing 1, 5, and 10 characters with the highest PS values from either matrix, as well as those characters whose PS values represented outliers according to the above criterion (BEA: 14 characters, LEA: 8 characters). These subsampled matrices were then subject to

unconstrained ML searches under the same settings as the original datasets (10 runs of 100 starting trees each + 1000 UFBoot replicates).

The definition of phylogenetic signal outlined above has recently been criticized for conflating cases in which one topology is strongly favored relative to either of the alternatives, and cases in which one topology is strongly disfavored relative to both alternatives that may nevertheless remain nearly indistinguishable from each other (Francis and Canfield, 2020). To tease apart these two scenarios, we further evaluated the difference in log-likelihood scores (ΔCLS) separately for each pair of hypotheses. For example, following Eq. (2) of Shen et al. (2017), the support afforded to Saurischia over Ornithoscelida by the i-th character was calculated as:

$$\Delta\text{CLS}(\text{S, Os})_i = \ln L(C_i|\text{S}) - \ln L(C_i|\text{Os}) \tag{3}$$

We then applied the same criterion for outliers to absolute ΔCLS values derived from each pairwise comparison, and identified those characters that represented outliers in at least two of the three comparisons. These corresponded to characters that either strongly favored or strongly disfavored a given topology relative to both of the alternatives. We also applied the criterion suggested by Francis and Canfield (2020) and identified those characters for which the log-likelihood difference between the best and second best hypotheses exceeded 0.5.

2.3.5 Rescoring of individual characters

To identify characters whose rescoring may have had disproportionate impact on the resulting topology, we ran further ML analyses on modified versions of both datasets in which we successively recoded one character at a time to its scoring in the opposite dataset. We excluded from consideration those characters whose coding did not change between the two

matrices (see "Data") as well as those that would be rendered constant by reverting their scoring to that of the opposite dataset (see "Maximum likelihood analyses"), resulting in a total of 879 analyses. All of these were conducted under the same settings as the analyses of the original datasets (10 runs of 100 starting trees each + 1000 UFBoot replicates). Using a custom R script employing the package *phangorn* v2.10.0 (Schliep, 2010), we scored each resulting ML topology for the recovery of Saurischia, Ornithischiformes, or Ornithoscelida, and extracted the UFBoot value of whichever of these three clades was present in the tree. To facilitate this process, we treated the names Saurischia, Ornithischiformes, and Ornithoscelida as referring to node-based clades, operationally defining them as (*Dilophosaurus* + *Plateosaurus*), (*Scelidosaurus* + *Plateosaurus*), and (*Dilophosaurus* + *Scelidosaurus*), respectively.

2.4 Results

2.4.1 Maximum likelihood analyses

Our initial round of 100 maximum likelihood (ML) searches suggested that estimating early dinosaur phylogeny from the two datasets (Table 2.1) would present considerable difficulty due to the ruggedness of the resulting tree space. The high difficulty scores (BEA = 0.596, LEA = 0.649) were mostly driven by a lack of topological congruence, since each of the 100 ML estimates had a unique topology that slightly (average pairwise normalized Robinson-Foulds distance: BEA = 0.330, LEA = 0.395) but appreciably differed from those of the remaining trees (Supplementary Table 1). For the BEA dataset, the best-scoring tree was generally similar to the original parsimony estimate (Supplementary Fig. 1) and strongly supported a monophyletic Ornithoscelida (ultrafast bootstrap [UFBoot] = 99%). In contrast, the LEA dataset supported a version of the Ornithischiformes hypothesis which nested Ornithischia within Sauropodomorpha (Supplementary Fig. 2). The sister-group relationship

between ornithischians and a clade of derived sauropodomorphs corresponding in content to Bagualosauria of Langer et al. (2019) again received substantial support (UFBoot = 97%), while the broader clade uniting Ornithischiformes proper with early sauropodomorphs was less robust (UFBoot = 94%).

	BEA	**LEA**
Taxa	74	83
Characters	457	457
Ordered	39	39
Constant	6	9
Binary	375	374
Of those autapomorphic:	45	43
Of those parsimony-informative:	330	331
Multistate	76	74
Of those autapomorphic:	2	1
Of those parsimony-informative:	74	73
Proportion missing	58.01%	55.58%

Table 2.1: Properties of the phylogenetic datasets analyzed in this study. The number of ordered characters is given as reported by BEA and LEA. For LEA, the proportion of missing data also includes the 109 polymorphic codings present in the original matrix, which are functionally equivalent to an unknown state under maximum likelihood.

Trees that were constrained to alternative early dinosaur topologies (Saurischia and Ornithischiformes for BEA, Saurischia and Ornithoscelida for LEA) exhibited log-likelihoods that fell well within the range yielded by the initial 100 unconstrained ML searches, suggesting that the three hypotheses were statistically indistinguishable for both datasets. We formally corroborated this result using multiple likelihood-based topology tests, all of which indicated that none of the three topologies was significantly better or worse than the others (Tables 2.2, 2.3). The topological constraints were accommodated by the two datasets in markedly different ways. When their monophyly was enforced, Saurischia and Ornithischiformes were subtended by near-zero-length branches in the trees produced by the BEA dataset (Supplementary Figs. 3, 4), indicating a lack of characters that could be

convincingly interpreted as saurischian or ornithischiform synapomorphies. For the LEA dataset, enforcing the monophyly of Saurischia yielded a topology that closely resembled the original parsimony estimate (cf. Supplementary Fig. 5 and Fig. 1 of Langer et al., 2017), while the tree optimized under an Ornithoscelida constraint showed an idiosyncratic topology that nested Ornithischia deep within theropods (Supplementary Fig. 6), reminiscent of hypotheses recently proposed by Baron (2019) to account for stratigraphic incongruence in ornithischian origins.

Hypothesis	**ln *L***	**bp (RELL)**	***p* (KH)**	***p* (SH)**	***p* (wSH)**	***p* (AU)**	***c* (ELW)**
Ornithoscelida	−6368.110126	0.754 +	0.826 +	1 +	0.946 +	0.824 +	0.748 +
Saurischia	−6377.249639	0.162 +	0.174 +	0.26 +	0.254 +	0.218 +	0.163 +
Ornithischiformes	−6377.484497	0.0844 +	0.114 +	0.247 +	0.194 +	0.222 +	0.0888 +

Table 2.2: Relative fit of the three alternative early dinosaur topologies to the BEA dataset. Best (unconstrained) and constrained ML trees are listed in order of decreasing likelihood (L). The bootstrap proportion (bp) from the resampling estimated log-likelihood (RELL) method is shown along with the p-values from the Kishino-Hasegawa (KH), Shimodaira-Hasegawa (SH), weighted Shimodaira-Hasegawa (wSH), and approximately unbiased (AU) tests, as well as the confidence value (c) based on expected likelihood weight (ELW). Significant rejection of (−) or the failure to reject (+) a given topology are indicated next to each test statistic.

2.4.2 Character-wise support

In both datasets, the distribution of character support for each of the three hypotheses was indistinguishable from uniform, both across the matrix as a whole (multinomial test; BEA: $p = 0.063$, LEA: $p = 0.668$) and within most of the individual anatomical partitions (Fig. 2.2b,d). Despite the overall similarity of character support distributions between the two matrices, we found substantial differences at the level of individual characters. Only 70 out of the 447 overlapping non-constant characters ranked the three hypotheses in the same way in terms of their log-likelihoods, less than the one-sixth expected by chance. Similarly, only 145 characters preferred the same hypothesis, a number that is also indistinguishable

from the one-third expected by chance (multinomial test: $p = 0.726$). While the number of characters supporting Ornithoscelida decreased as a result of LEA's rescoring (from 174 to 147) and the number of characters supporting Saurischia marginally increased (from 141 to 143), even in the LEA dataset, more characters supported Ornithoscelida than Saurischia, although the difference was negligible (Fig. 2.2b,d).

Examining support for competing hypotheses in terms of simple preference (i.e., by recording which tree yields the highest log-likelihood score for a given character) fails to account for the fact that most characters do not strongly prefer any of the three alternatives, rendering the resulting log-likelihood differences negligible and overly sensitive to minor branch length differences. Indeed, the character-wise log-likelihood difference (ΔCLS) between the best and second best hypotheses exceeds a threshold of 0.5 for less than 10% of characters in the BEA dataset and 30% of characters in the LEA dataset, although we note that these proportions are still substantially higher than is typical of the phylogenomic datasets for which the threshold was originally defined (Shen et al., 2017; Francis and Canfield, 2020). In the BEA dataset, the majority of these "strong" characters (29 out of 45) support Ornithoscelida, which is also the case for a plurality (63 out of 137) of such characters in the LEA dataset.

Hypothesis	**ln *L***	**bp (RELL)**	***p* (KH)**	***p* (SH)**	***p* (wSH)**	***p* (AU)**	***c* (ELW)**
Ornithischiformes	−7023.11918	0.723 +	0.882 +	1 +	0.962 +	0.832 +	0.718 +
Saurischia	−7034.61052	0.102 +	0.118 +	0.29 +	0.195 +	0.234 +	0.106 +
Ornithoscelida	−7035.848825	0.175 +	0.181 +	0.273 +	0.246 +	0.208 +	0.177 +

Table 2.3: Relative fit of the three alternative early dinosaur topologies to the LEA dataset. Best (unconstrained) and constrained ML trees are listed in order of decreasing likelihood. Abbreviations as in Table 2.2.

To obtain a more fine-grained view of key characters driving the support for particular topologies, we employed the method suggested by Shen et al. (2017) and calculated the phylogenetic signal (PS) of each character as the mean of the absolute values of the three pairwise

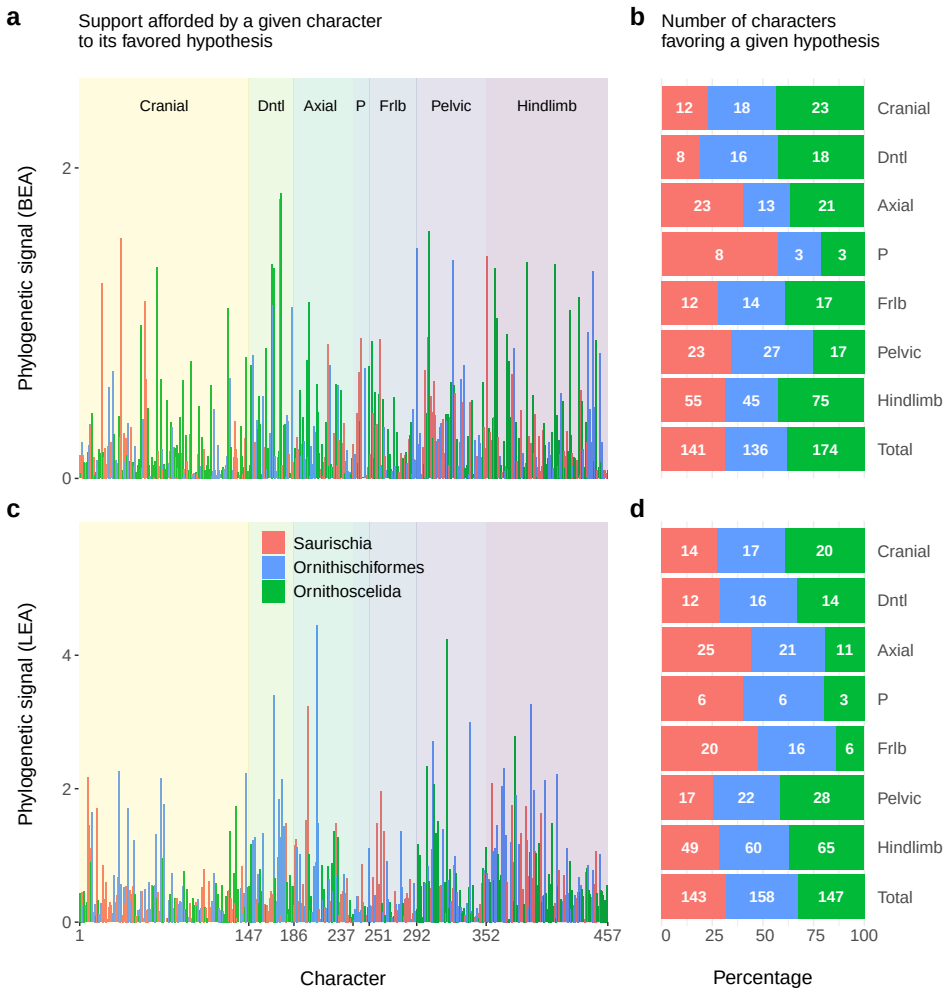

Figure 2.2: Distribution of phylogenetic signal across the two datasets. Phylogenetic signal of every non-constant character from the BEA (a) and LEA (c) dataset is plotted against its placement in the matrix and colored by the favored (highest-likelihood) hypothesis. The numbers and proportions of characters favoring a given hypothesis are further shown for the BEA (b) and LEA (d) datasets as well as the individual anatomical regions by which they are organized. Abbreviations: Dntl = dental, P = pectoral, Frlb = forelimb. The plots were generated in R v4.2.2 (R Core Team, 2021) using the packages *scales* v1.2.1 (Wickham and Seidel, 2022), *ggplot2* v3.4.0 (Wickham et al., 2022a), *cowplot* v1.1.1 (Wilke, 2020), *patchwork* v1.1.2 (Pedersen, 2022b), *ggnewscale* v0.4.8 (Campitelli, 2022), and *viridisLite* v0.4.1 (Garnier et al., 2022).

log-likelihood differences (Saurischia vs. Ornithischiformes, Saurischia vs. Ornithoscelida, Ornithischiformes vs. Ornithoscelida). Relative to the BEA dataset, the LEA dataset displays substantially higher mean (0.612 vs. 0.295) as well as maximum (4.454 vs. 1.840) phyloge-

netic signal values, consistent with its higher number of strong characters. The PS values of individual characters also differ considerably between the two datasets (Fig. 2.2a,c). In particular, among the characters with outlier PS values (more than three standard deviations above the mean), only one (character 169, serrations of maxillary and dentary teeth) is shared between the BEA matrix (in descending order of PS: 175, 174, 303, 37, 292, 353, 323, 387, 167, 411, 68, 360, 169, 444) and the LEA matrix (206, 318, 169, 391, 198, 338, 377, 306). Gradual removal of 1, 5, or 10 characters with the highest PS values or of all characters whose PS values represented outliers never caused either matrix to switch from the preferred topology to an alternative one (Supplementary Figs. 7–14). The exclusion of high-PS characters made little difference to the high statistical support for Ornithoscelida in the BEA dataset (UFBoot = 96–99%; Supplementary Figs. 7–10) but caused a drastic erosion of support for Ornithischiformes in the LEA dataset (Ornithischia + Bagualosauria: UFBoot = 69–93%; Ornithischiformes proper + early sauropodomorphs: UFBoot = 52–79%; Supplementary Figs. 11–14).

To further distinguish among different ways in which a character can attain a high PS value, we separately compared each pair of hypotheses (Fig. 2.3), focusing on those characters that also emerged as outliers in at least two of the three pairwise comparisons. These corresponded to characters that strongly favored a particular hypothesis, strongly disfavored one, or both. The results reveal substantial conflict among the high-PS characters, as each matrix contains characters that both strongly favor and strongly disfavor its globally preferred topology (Ornithoscelida for BEA, Ornithischiformes for LEA) as well as one or both of its alternatives (Fig. 2.3, Table 2.4). Neither of the characters that strongly favor Ornithoscelida in the BEA dataset was among the 21 synapomorphies of this clade originally identified by BEA, although one such synapomorphy (character 360, state 1: medial bowing of the femur forming a gentle curve) was found to strongly disfavor Saurischia in the present analysis (Table 2.4). Similarly, there is virtually no overlap between the characters strongly

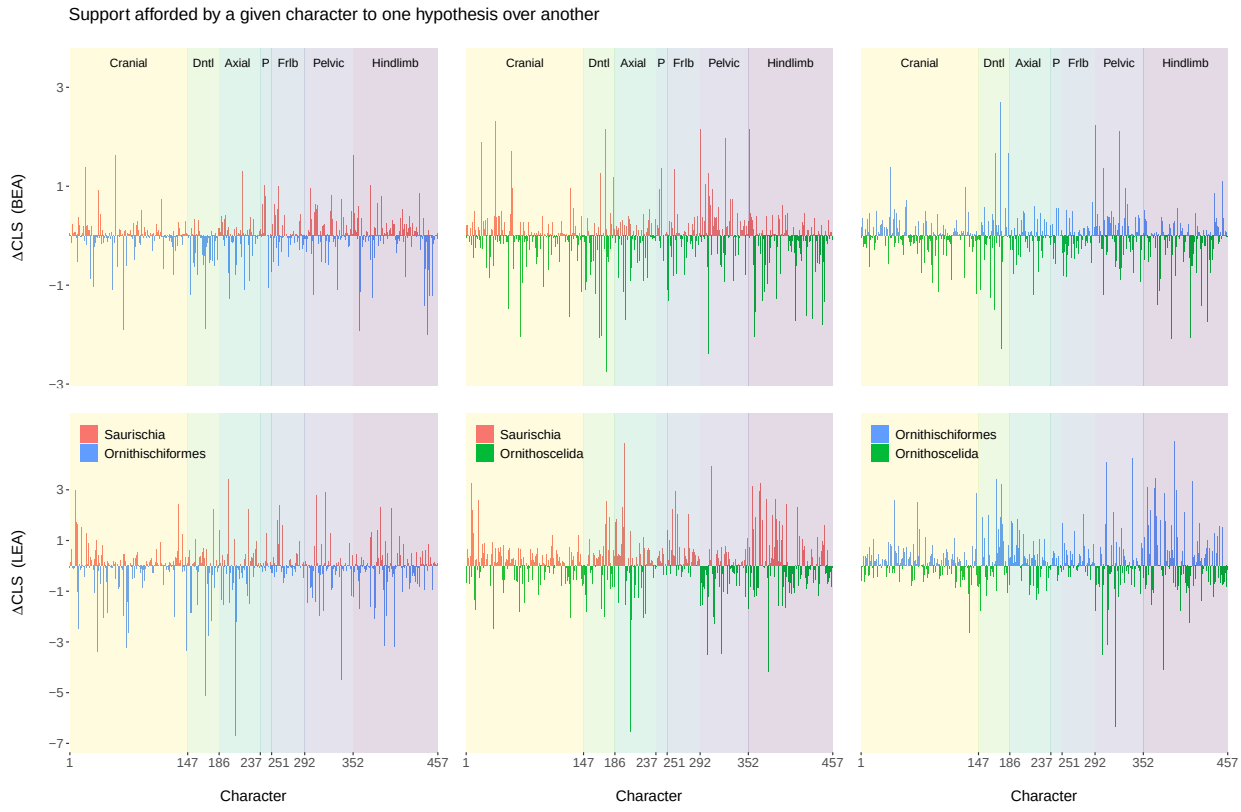

Figure 2.3: Pairwise comparisons of character support for the three alternative early dinosaur topologies. For each pair of hypotheses, the difference in character-wise log-likelihood values (ΔCLS) of every non-constant character from the BEA (top) and LEA (bottom) datasets is plotted against its placement in the matrix. Anatomical region abbreviations as in Fig. 2.2. The plots were generated in R v4.2.2 (R Core Team, 2021) using the packages *scales* v1.2.1 (Wickham and Seidel, 2022), *ggplot2* v3.4.0 (Wickham et al., 2022a), *cowplot* v1.1.1 (Wilke, 2020), *patchwork* v1.1.2 (Pedersen, 2022b), *ggnewscale* v0.4.8 (Campitelli, 2022), and *viridisLite* v0.4.1 (Garnier et al., 2022).

favoring or disfavoring one of the three hypotheses in the present study, and the "keystone" characters recently identified using a parsimony-based approach (Goloboff and Sereno, 2021).

2.4.3 Rescoring of individual characters

To determine which of LEA's coding changes most contributed to the difference between the topologies yielded by the original and rescored datasets, we successively replaced the scoring of each character in either matrix by its scoring from the opposite matrix, and

checked whether this change was sufficient for the re-estimated ML tree to switch to a different topology (Fig. 2.4). The procedure had little impact on the BEA dataset's support for Ornithoscelida, which proved robust to the rescoring of any one of the 438 applicable characters and remained high on average (mean UFBoot = 98.8%; Fig. 2.4a). Only three characters (110, 114, 387) caused the support for Ornithoscelida to drop below the 95% threshold when changed to their scorings in the LEA dataset. Of these, none was optimized as an ornithoscelidan synapomorphy in BEA's original analysis (Baron et al., 2017a), and while all favored Ornithoscelida over the alternatives in the original BEA matrix, only character 387 did so strongly (ΔCLS > 0.5 relative to the next best hypothesis). In the LEA matrix, all three characters ranked Saurischia and Ornithoscelida as the best- and worst-supported hypothesis, respectively, but the log-likelihood difference between the two exceeded the 0.5 threshold only for character 387.

In contrast, reverting four characters to their original scoring in the BEA matrix proved sufficient to flip the ML result from the LEA dataset (Ornithischiformes) either to Ornithoscelida (characters 148, 363, 370) or to Saurischia (character 77) (Fig. 2.4b), producing highly idiosyncratic topologies (Supplementary Figs. 15–18). In the ornithoscelidan trees, Ornithischia was deeply nested within theropods (with *Panguraptor* and *Zupaysaurus* consistently recovered closer to Ornithischia than to other theropods), similar to the results obtained from a constrained analysis of the unmodified LEA matrix (Supplementary Fig. 6). The support for the nodes uniting Ornithischia with its successive theropod outgroups was generally low, although one such node received a UFBoot value of 97% in the analysis based on rescoring character 370 (Supplementary Fig. 18), consistent with the fact that state 2 of this character (prominent, wing-like anterior trochanter) represented an ornithoscelidan synapomorphy under BEA's original scoring (Baron et al., 2017a; see also Baron and Barrett, 2017). The only analysis to recover a monophyletic Saurischia did so with poor support, and found unconventional relationships elsewhere in the tree (e.g., ornithischian

herrerasaurids; Supplementary Fig. 15). None of the characters that caused a switch from Ornithischiformes to Ornithoscelida favored Ornithischiformes in the LEA dataset, and only two of them (characters 148 and 370) favored Ornithoscelida in the BEA dataset. Notably, character 77 consistently ranked Saurischia as the worst of the three hypotheses under both BEA and LEA scorings; the fact that its rescoring was nevertheless sufficient for Saurischia to emerge as the preferred hypothesis indicates an extreme degree of instability across the LEA matrix. This is also borne out by the fact that even among those trees which continued to support Ornithischiformes, altering the scoring of a single character caused the UFBoot support for this clade to drop from 97% to (on average) 89.5% (Fig. 2.4b). Only 105 of the 441 applicable characters (23.8%) upheld Ornithischiformes with UFBoot support greater than 95% upon rescoring.

2.5 Discussion

By repurposing a protocol originally developed for phylogenomic data (Shen et al., 2017), we found that both the BEA and LEA datasets are unable to conclusively resolve the interrelationships of major dinosaur clades. All three hypotheses of overall dinosaur phylogeny – Saurischia, Ornithischiformes, and Ornithoscelida – remain plausible, and neither dataset shows any of these to be significantly better or worse than the alternatives (Tables 2.2, 2.3). Our results suggest that this is not due to low information content of the two matrices; in fact, the proportion of characters that strongly discriminate between the best and second best hypothesis (ΔCLS > 0.5) is far higher (10–30%) than typical for phylogenomic data (0.1–7%; Shen et al., 2017; Francis and Canfield, 2020). Similarly, although both matrices exhibit a highly uneven distribution of phylogenetic signal and contain several outlier character strongly favoring or disfavoring particular topologies, the results were not exclusively driven by a handful of such outliers, since their removal had limited impact on the support for one topology or another (Supplementary Figs. 7–14). Instead, we hypothesize that the

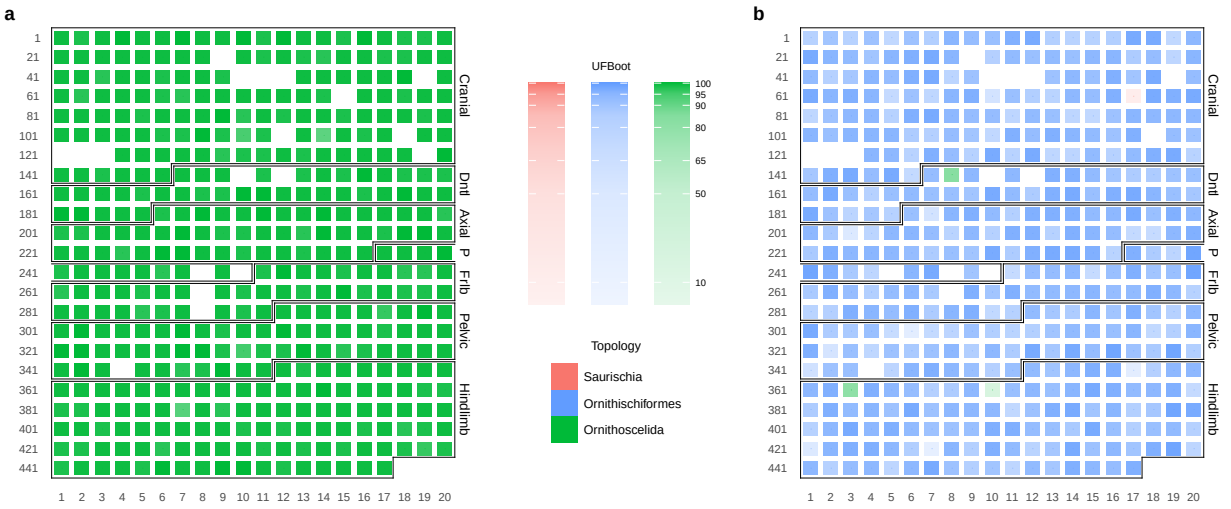

Figure 2.4: Support for alternative topologies after rescoring each matrix one character at a time. Characters changed from BEA's to LEA's coding (a) or from LEA's to BEA's coding (b) are arranged by row following their placement in either dataset; black boxes denote different anatomical regions. Color and opacity of individual heatmap cells denote the topology of the ML tree inferred after rescoring the character in question and the ultrafast bootstrap (UFBoot) support for that topology's focal clade, respectively. Gaps correspond to characters whose coding did not differ between the two matrices or characters that would be rendered constant by having their coding changed to that of the opposite matrix. Anatomical region abbreviations as in Fig. 2.2. The plots were generated in R v4.2.2 (R Core Team, 2021) using the packages *ape* v5.6-2 (Paradis et al., 2004; Paradis and Schliep, 2018), *dplyr* v1.0.10 (Wickham et al., 2022b), *phangorn* v2.10.0 (Schliep, 2010), *scales* v1.2.1 (Wickham and Seidel, 2022), *ggplot2* v3.4.0 (Wickham et al., 2022a), *cowplot* v1.1.1 (Wilke, 2020), *patchwork* v1.1.2 (Pedersen, 2022b), and *ggnewscale* v0.4.8 (Campitelli, 2022).

lack of meaningful statistical support for any of the three hypotheses (partially obscured by high UFBoot values) is due to pervasive conflict among individual characters. Limiting the focus to the characters with the highest PS values revealed patterns of conflict (Table 2.4) that were similar to those observed across each matrix as a whole (Fig. 2.2), explaining why their removal did little to change the underlying distribution of support.

According to almost every metric employed in this study, the LEA dataset produces less stable results than the BEA dataset. It yields a higher phylogenetic difficulty score (Supplementary Table 1), exhibits a more uniform distribution of character support for the three hypotheses (Fig. 2.2c, d), and is more sensitive to the exclusion of outlier characters (Supple-

mentary Figs. 11–14) despite containing fewer of them (Table 2.4). The LEA matrix was also less robust to changes in character coding, as demonstrated by its tendency to flip the ML tree from one topology to another after reverting a single character to its original scoring by BEA (Fig. 2.4). Consistent with the findings above, this greater degree of instability was not caused by weaker phylogenetic signal in the LEA matrix. In fact, although LEA's rescorings and taxon additions only marginally improved the completeness of the matrix (Table 2.1), they resulted in a markedly higher mean character-wise phylogenetic signal (Supplementary Fig. 19), and amplified the log-likelihood differences between competing hypotheses both for the dataset as a whole (Table 2.3) and for individual characters (Supplementary Fig. 20). In effect, the stronger phylogenetic signal present in the recoded and expanded matrix only served to amplify, rather than eliminate, underlying conflict within the dataset. In light of the failure to resolve this pervasive conflict by extensive coding changes, we outline several alternative recommendations for identifying its sources and assessing the relative support for competing topologies in its presence.

First, we recommend re-examining the original dataset at a deeper level, as the recovery of divergent yet statistically indistinguishable topologies may serve to highlight fundamental issues in the underlying character data. Poorly formulated characters should ideally be redefined rather than simply rescored. For example, LEA's coding changes that caused character 174 (recurvature of maxillary and dentary teeth) to lose the outlier status it originally had in the BEA dataset (Table 2.4) still took place in the framework of the vague definition inherited from BEA, who in turn modified it from an even earlier study (Butler et al., 2008). The character description provides no quantitative criterion for differentiating between teeth possessing strong, weak, or no recurvature, allowing for more or less arbitrary coding changes. Indeed, the scoring of *Efraasia* was changed by LEA from no recurvature to weak recurvature without explicit justification or photographic evidence, and on the basis of the same published sources which BEA cited in support of their own original coding. The problem of

Character	PS in BEA	PS in LEA	Outlier in BEA	Outlier in LEA
68	1.360	0.352	S (−)	—
167	1.383	0.409	Os (+)	—
169	1.352	3.407	S (−)	Of (+)
174	1.801	1.275	Os (−)	—
175	1.840	0.103	Os (+)	—
198	0.139	3.228	—	S (+)
206	0.605	4.454	—	S (−)
292	1.483	0.193	Os (−)	—
306	0.262	2.716	—	Os (−)
318	0.370	4.231	—	Os (+), Of (−)
323	1.403	0.500	Os (−)	—
338	0.493	2.998	—	Of (+)
353	1.429	0.738	S (+)	—
360	1.356	1.112	S (−)	—
377	0.057	2.787	—	Os (+)
391	0.226	3.267	—	Of (+)
444	1.337	0.780	S (−)	—

Table 2.4: Outlier characters in the BEA and LEA matrices. Of = Ornithischiformes, Os = Ornithoscelida, S = Saurischia. Characters were only included if both their phylogenetic signal (PS) and at least two of the three pairwise log-likelihood differences [ΔCLS(S, Of), ΔCLS(S, Os), ΔCLS(Of, Os)] were more than three standard deviations above the mean. Hypotheses that are strongly favored (+) or disfavored (−) relative to both alternatives are indicated next to each character.

vague character descriptions and subjective scoring decisions is widespread in the two matrices (e.g., characters 114, 216, 266, 337) and compounded by a number of additional issues, some of which were noted by BEA and LEA themselves. These include multiple instances of scoring taxa for characters that cannot be ascertained from their known material (Baron et al., 2017b). While this practice may occasionally be justifiable (e.g., using the length of the mandible to estimate the length of the skull), coding taxa based on assumed rather than observed morphologies represents a potentially serious biasing factor (Giribet, 2010; Gee, 2021).

Second, some of the issues previously suggested to represent fundamental problems with the data may in fact be mitigated at the methodological level. For example, character non-

independence (Langer et al., 2017) violates the assumptions of the methods currently used to analyze paleontological datasets (including those employed in this study), but frameworks capable of dealing with hierarchical and correlated characters are under development (Tarasov, 2023), as are ontology-based methods for their semi-automated detection and characterization (Eliason et al., 2019; Porto et al., 2022). More advanced methods may ultimately also alleviate other problems with existing matrices. The BEA and LEA datasets contain instances of problematic state delimitation that make it impossible to assign certain morphologies to any existing state (e.g., neither of the two states defined for character 28 can account for an antorbital fenestra equal in size to 10–15% of skull length), or merge distinct morphologies into a single state (e.g., character 77, state 0: paraquadratic foramen small or absent). Both exemplify a more general problem, namely the arbitrary discretization of characters that would be more naturally treated as continuous (Poe and Wiens, 2000). This ubiquitous feature of morphological phylogenetic datasets reflects long-standing methodological limitations, which have only recently been overcome by phylogenetic software packages that simultaneously implement models for discrete and continuous characters, making it possible to combine both types of data in a single analysis (Höhna et al., 2016; Bouckaert et al., 2019).

Third, instead of attempting to rescore or redefine an entire dataset in an indiscriminate "shotgun" approach, it is prudent to determine which characters are responsible for the signal in that dataset's results. While LEA re-examined the putative ornithoscelidan synapomorphies identified by BEA at an admirable level of detail (Langer et al., 2017), both previous findings (Goloboff and Sereno, 2021) and our own results demonstrate that the characters supporting a particular topology do not always coincide with the characters that map as the synapomorphies of that topology's focal clade. Methods for identifying such critical characters are now available in parsimony (Goloboff and Sereno, 2021), maximum-likelihood (Shen et al., 2017; Francis and Canfield, 2020), and Bayesian (Porto et al., 2022) frameworks, and

can be profitably used to narrow the focus of potential rescoring efforts, which often involve the time- and labor-intensive process of gathering data from first-hand observations of multiple museum specimens. The benefits of comprehensively revising a pre-existing character matrix have to be weighed against the costs inherent to such an effort, as well as the risk of introducing new errors into the data. In morphological phylogenetics, extensive reuse and iterative expansion of pre-existing matrices gives rise to complicated dataset genealogies (Gee, 2021; Regalado Fernández and Werneburg, 2022) in which coding error can propagate and compound over time. On the other hand, our results suggest that the effort invested into the comprehensive rescoring of a large pre-existing dataset can be difficult to justify: the three hypotheses of early dinosaur phylogeny were statistically indistinguishable based on the original BEA dataset, and remain such after LEA's extensive revision of it.

Fourth, we urge paleontologists to quantify the uncertainty associated with their phylogenetic hypotheses using well-characterized tools with a clear statistical interpretation. When alternative ways of resolving a given branch are of interest, as in the controversy surrounding early dinosaur phylogeny, we encourage the community to move beyond the mere reporting of a phylogenetic point estimate toward explicitly testing it against the next best alternative. Following recent practice (Wu et al., 2023), we used a variety of likelihood ratio tests (LRTs) to this end (Shimodaira and Hasegawa, 1999; Shimodaira, 2002; Strimmer and Rambaut, 2002), but other approaches are possible. Bayesian inference differs from maximum likelihood in its treatment of nuisance parameters such as branch lengths or the parameters of the substitution and rate heterogeneity models, which are jointly optimized with the parameter of interest (topology) in maximum likelihood but marginalized over in Bayesian methods (Huelsenbeck et al., 2002). Both approaches have their advantages (Holder and Lewis, 2003), and as a result, Bayes factors – a Bayesian equivalent of LRTs, relying on marginal rather than joint likelihoods (Kass and Raftery, 1995) – can represent a useful alternative way of evaluating competing topologies (Suchard et al., 2005; Bergsten et al., 2013). Although

most of the best-performing marginal likelihood estimators are much more computationally demanding than joint likelihood inference (Fourment et al., 2019), the minute size of phylogenetic datasets employed by paleontologists makes their application relatively easy, and Bayes factor topological comparisons are accordingly starting to see use in dinosaur phylogenetics (O'Connor et al., 2020).

Taken together, our results suggest that large-scale dinosaur phylogeny is much more poorly understood than commonly acknowledged. In particular, despite the widespread perception that saurischian monophyly is challenged primarily by the recently proposed Ornithoscelida hypothesis (Felice et al., 2020; Müller and Garcia, 2020; Castiglione et al., 2022), the earlier Ornithischiformes hypothesis receives comparable support, and is in fact weakly preferred when the LEA dataset is analyzed using maximum likelihood (Table 2.3). Additional support for Ornithischiformes was also recently detected in an independent dataset (Baron, 2022). Moreover, not only the three major hypotheses – Saurischia, Ornithischiformes, and Ornithoscelida – but also a number of their variations nesting the ornithischians deep within Sauropodomorpha or Theropoda (Supplementary Figs. 2, 6, 16–18) cannot be ruled out at present. Indeed, the specific variation on the Ornithischiformes topology recovered in this study shows the ornithischians to be not just sister to, but rather nested within Sauropodomorpha, with a clade of Carnian to early Norian sauropodomorphs (approximately corresponding to Guaibasauridae of Ezcurra, 2010 or Saturnaliidae of Langer et al., 2019) branching off before the ornithischians. This result is consistent with a phylogeny inferred from the LEA dataset by Parry et al. (2017) using time-free Bayesian inference, showing that the two model-based methods yield topologies that are much more similar to each other than to those favored by parsimony. By positing an early divergence of the early Late Triassic sauropodomorphs, this scenario helps reduce the temporal gap between the first appearance of Ornithischia and its sister group (Baron, 2019), and shows remarkable congruence with

the early suggestions that the ornithischians may have arisen from within "prosauropods" (Bakker and Galton, 1974; Bonaparte, 1976; Cooper, 1981; Bakker, 1986).

Our findings suggest that higher-level dinosaur interrelationships represent a phylogenetic problem of considerable difficulty that is unlikely to be conclusively resolved by minor additions and superficial modifications to the datasets currently in use. While our investigation was limited to two such datasets (Baron et al., 2017a; Langer et al., 2017), there are few reasons to believe that other matrices currently employed by dinosaur paleontologists are free of the problems identified here. Indeed, the main alternative to the datasets examined here has repeatedly lent support to yet another nonstandard hypothesis nesting the putatively non-dinosaurian Silesauridae within Ornithischia (Cabreira et al., 2016; Müller and Garcia, 2020; Norman et al., 2022), indicating that the number of plausible early dinosaur phylogenies proliferates even further when not only the three major clades, but also species-poor lineages such as Herrerasauridae and Silesauridae are taken into consideration. As a result, the increasingly common practice of repeating comparative analyses under the Saurischia and Ornithoscelida topologies (Felice et al., 2020; Castiglione et al., 2022; Hendrickx et al., 2022) most likely severely understates the uncertainty associated with early dinosaur phylogeny, potentially leading to biased or overconfident conclusions.

2.6 Data availability

All data matrices, tree files, log files, R code, and shell scripts are available from Zenodo: `https://doi.org/10.5281/zenodo.7945723`.

2.7 Acknowledgements

We thank Graham J. Slater for valuable comments and advice, and Joseph F. Walker for the suggestion to apply phylogenomic methods of investigating site-wise support for

alternative topologies to morphological datasets. We are indebted to Robin M. D. Beck and two anonymous reviewers for constructive criticism that helped improve the quality of the manuscript.

2.8 Author contributions

D.Č. and A.L.S. conceived the study, performed the phylogenetic analyses, and interpreted the results. A.L.S. compiled the data and carried out detailed examination of character coding. D.Č. designed the methodology, scripted tree post-processing and statistical analyses, and generated the figures. D.Č. wrote the paper with input from A.L.S.

2.9 Competing interests

The authors declare no competing interests.

CHAPTER 3

EMPIRICAL AND METHODOLOGICAL CHALLENGES TO THE MODEL-BASED INFERENCE OF DIVERSIFICATION RATES IN EXTINCT CLADES

3.1 Abstract

Changes in speciation and extinction rates are key to the dynamics of clade diversification, but attempts to infer them from phylogenies of extant species face challenges. Methods capable of synthesizing information from extant and fossil species have yielded novel insights into diversification rate variation through time, but little is known about their behavior when analyzing entirely extinct clades. Here, we use empirical and simulated data to assess how two popular methods, PyRate and Fossil BAMM, perform in this setting. We inferred the first tip-dated trees for ornithischian dinosaurs and combined them with fossil occurrence data to test whether the clade underwent an end-Cretaceous decline. We then simulated phylogenies and fossil records under empirical constraints to determine whether macroevolutionary and preservation rates can be teased apart under paleobiologically realistic conditions. We obtained discordant inferences about ornithischian macroevolution including a long-term speciation rate decline (BAMM), mostly flat rates with a steep diversification drop (PyRate) or without one (BAMM), and episodes of implausibly accelerated speciation and extinction (PyRate). Simulations revealed little to no conflation between speciation and preservation, but yielded spuriously correlated speciation and extinction estimates while time-smearing tree-wide shifts (BAMM) or overestimating their number (PyRate). Our results indicate that the small phylogenetic data sets available to vertebrate paleontologists and the assumptions made by current model-based methods combine to yield potentially unreliable inferences about the diversification of extinct clades. We provide guidelines for

interpreting the results of the existing approaches in light of their limitations and suggest how the latter may be mitigated.

3.2 Introduction

Variation in speciation and extinction rates through time and across lineages can establish a key causal link between patterns of biological diversity and the variety of factors that ultimately shape them (Sepkoski, 1998; Foote, 2000; Rabosky et al., 2012; Schluter and Pennell, 2017). Consequently, accurate estimation of macroevolutionary rates is essential for understanding the uneven distribution of species richness across the tree of life and its temporal dynamics, including post-extinction recovery (Stanley, 2007; Brayard et al., 2009), the identification of clades in decline (Quental and Marshall, 2011; Sakamoto et al., 2016; Burin et al., 2019; Billaud et al., 2020), or time and diversity dependence (Foote et al., 2018; Henao Diaz et al., 2019; Pannetier et al., 2021). Historically, the problem of macroevolutionary rate estimation has been addressed using two broad classes of approaches, relying either on the quantitative analysis of the fossil record (Raup, 1985; Foote, 2000, 2003; Alroy, 2008, 2014; Liow and Finarelli, 2014) or time-calibrated molecular phylogenies of extant taxa (Harvey et al., 1994; Nee et al., 1994; Pybus and Harvey, 2000; Alfaro et al., 2009; Morlon, 2014; Rabosky, 2014; Maliet et al., 2019). Both types of methods have seen extensive development and deployment, but often led to drastically different inferences about the underlying macroevolutionary dynamics (Quental and Marshall, 2009, 2010; Liow et al., 2010; Morlon et al., 2011; Etienne et al., 2012; Mitchell et al., 2018).

In recent years, concerns about estimating extinction rates in the absence of fossil data (Rabosky, 2010, 2016) and the widely noted discrepancy between diversification rates estimated from paleobiological and neontological data (Hunt and Slater, 2016; Marshall, 2017; Silvestro et al., 2018) have fueled the development of approaches that can incorporate both extant and extinct taxa. Examples of such frameworks include PyRate (Silvestro et al.,

2014a,b, 2019), developed as an extension of quantitative paleobiological methods, and Fossil BAMM (Bayesian Analysis of Macroevolutionary Mixtures) (Mitchell et al., 2018), which extends extant-only phylogenetic approaches to extinct lineages. Despite their origins in different research traditions, the two methods share fundamental similarities in being Bayesian, utilizing birth-death models, and averaging the parameters of interest across models of different dimensionality. These properties make them well suited for the study of diversification dynamics through time, and both frameworks are consequently beginning to see widespread adoption (Silvestro et al., 2015; Condamine et al., 2016; Raia et al., 2016; Pires et al., 2018; Crouch et al., 2019; Crouch, 2020; Lloyd and Slater, 2021).

Recently, several lines of evidence have cast doubt on the ability of birth-death approaches to tease apart variation in rates of speciation, extinction, and fossil sampling. Diversification scenarios involving time-varying speciation and extinction rates are not identifiable from time trees of extant taxa alone (Louca and Pennell, 2020), and while the ability of PyRate and Fossil BAMM to incorporate fossil taxa could potentially alleviate this problem, the fossil preservation rate may itself be subject to similar identifiability issues. Covariation between extinction and fossil preservation has long been recognized in paleobiology (Sepkoski, 1975; Foote, 2003, 2007; Foote et al., 2019) and motivated the development of approaches that jointly estimate the rates of speciation, extinction, and preservation (Connolly and Miller, 2001; Foote, 2003; Liow and Finarelli, 2014), but even these approaches can be misled by heterogeneity in the rate of sampling and introduce spurious shifts into macroevolutionary rates as a result (Smiley, 2018). Similarly, episodes of exceptional sampling were suggested to inflate PyRate estimates of net diversification rates (Condamine et al., 2016). Moreover, while both frameworks were tested and validated on empirical and simulated data upon their introduction (Silvestro et al., 2014b, 2019; Mitchell et al., 2018), these tests did not cover the full range of potential use cases. In particular, the availability of time trees for extinct clades (Bapst et al., 2016; Paterson et al., 2019) makes Fossil BAMM applicable not only

to groups with living representatives, as in the original simulations and current empirical applications (Mitchell et al., 2018; Crouch et al., 2019; Lloyd and Slater, 2021), but also to clades that are wholly extinct, where its performance remains unknown.

In this study, we apply Fossil BAMM and PyRate to the clade Ornithischia, one of the three major radiations of dinosaurs (Gauthier, 1986; Baron et al., 2017a; Langer et al., 2017), whose fossil record is ideally suited to testing the performance of the two frameworks. First, the ornithischians represent a wholly extinct, species-rich clade with an evolutionary history spanning at least a ∼134 myr-long period from the Early Jurassic at the latest (Agnolín and Rozadilla, 2018; Baron, 2019) to the Cretaceous–Paleogene (K–Pg) extinction event (Archibald and Fastovsky, 2004; Brusatte et al., 2015). Second, the ornithischians have been subject to extensive phylogenetic research (Boyd, 2015; Han et al., 2018; Dieudonné et al., 2020), allowing phylogenetic information to be used in diversification rate estimation. Third, the ornithischians have a fossil record of adequate but not exceptional quality, marked by global episodes of poor sampling, patchy local records, and a substantial proportion of singletons (taxa known from a single stratigraphically unique occurrence) (Pol et al., 2011; Tennant et al., 2018). In terms of size and completeness, ornithischian phylogenetic and occurrence data are thus typical of the data sets widely in use by vertebrate paleontologists, whose broad availability may lead them to be repurposed for diversification rate estimation as methods like Fossil BAMM and PyRate grow in popularity. Finally, there is an ongoing controversy concerning ornithischian diversity trends and diversification rates, especially those prior to the K–Pg extinction (Wang and Dodson, 2006; Barrett et al., 2009; Lloyd, 2012; Brusatte et al., 2015; Sakamoto et al., 2016; Chiarenza et al., 2019; Bonsor et al., 2020), further motivating the application of state-of-the-art diversification rate analysis to their fossil record.

Here, we infer the first Bayesian tip-dated phylogenies for Ornithischia based on multiple character matrices. We then use the resulting time trees and an expert-curated set of fossil

occurrences to assess the performance of the diversification rate estimation implemented in Fossil BAMM and PyRate in an entirely extinct clade and under a wide range of analytical settings. We evaluate the congruence and robustness of the resulting diversification scenarios, and examine how model, prior, and algorithm choice affect the magnitude of speciation and extinction rate estimates as well as the number and position of inferred rate shifts. To assess the overall plausibility of our results, we additionally simulate birth-death trees and fossil records under empirically informed constraints to determine the extent to which the methods in question can disentangle the rates of speciation, extinction, and fossil sampling. We conclude by discussing the diversification scenarios inferred by our analyses in light of previous hypotheses about the clade's macroevolutionary dynamics, and urge caution when applying current model-based methods to the low-power data sets common in vertebrate paleontology.

3.3 Materials and Methods

3.3.1 Occurrence Data

We downloaded all ornithischian species-level body fossil occurrences that were present in the Paleobiology Database (PBDB; `http://www.paleobiodb.org`) on 1 August 2019; the full search settings and URL are provided in the Supplementary Information available on Dryad at `http://dx.doi.org/10.5061/dryad.sbcc2fr4x`. The manually curated data set comprised 1240 occurrences from 398 species (occurrences per species: median = 1, standard deviation = 5.22); 60.8% of the species were singletons.

3.3.2 Character Data and Tip Ages

We updated and expanded three of the largest character matrices previously used for phylogenetic analyses of ornithischian dinosaurs: Han et al. (2018), Herne et al. (2019), and Madzia et al. (2018), hereafter referred to as "HaEA", "HeEA", and "MEA", respectively

(Supplementary Information, Phylogenetic Analyses available on Dryad). Following Barido-Sottani et al. (2019a), our time tree analyses sampled tip dates from their stratigraphic age ranges. Where possible, we extracted these from the PBDB data. The stratigraphic position of most occurrences could not be constrained beyond the stage level, leading to relatively wide ranges. We used the maximum first appearance date and the minimum last appearance date for taxa represented by multiple occurrences, and the maximum and minimum ages of the only occurrence for singletons. Several tips corresponded to unnamed specimens or recently described taxa absent from the PBDB; for these, age ranges were taken from the literature.

3.3.3 Phylogenetic Analyses and Tip-Dating

Following preliminary maximum likelihood analyses using RAxML (Stamatakis, 2014), we performed 16 tip-dating analyses using BEAST 2.6 (Bouckaert et al., 2019), varying the treatment of topology (unconstrained vs. fixed to the RAxML estimate), the extant sampling parameter ρ (zero or nonzero), and (for the HeEA data set) the partitioning scheme. All analyses used the sampled- ancestor fossilized birth–death model (Gavryushkina et al., 2014) as the tree prior and a lognormal relaxed clock model with empirically derived hyperpriors (Fig. 3.1). Full details of phylogenetic analyses are provided in the Supplementary Information available on Dryad.

3.3.4 Diversification Rate Analyses

Fossil BAMM

We inferred the diversification dynamics of the ornithischians using BAMM v2.6 (Fossil BAMM; Mitchell et al., 2018). Like prior versions of BAMM (Rabosky, 2014), Fossil BAMM uses a time-scaled phylogenetic tree to identify shifts separating distinct diversi-

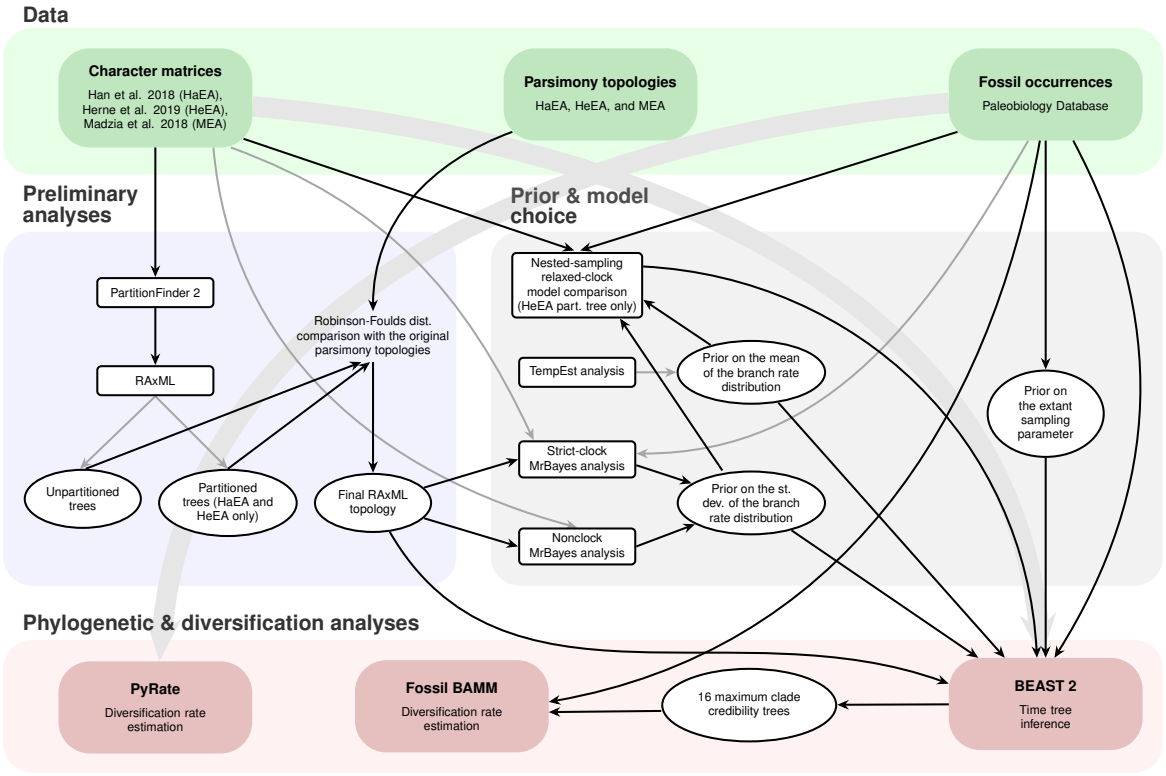

Figure 3.1: Flowchart showing the workflow of the phylogenetic and diversification rate analyses performed in this study. The protocol used for diversification rate estimation is further described below; for a detailed explanation and justification of the phylogenetic pipeline, see the Supplementary Information available on Dryad.

fication regimes, without requiring the user to *a priori* specify their number or locations. Unlike the original version of the method, Fossil BAMM does not require the tree to be ultrametric, enabling the inclusion of non-contemporaneous tips, and its likelihood function incorporates a fossil preservation rate parameter estimated from the total number of stratigraphically unique species-level occurrences associated with the lineages represented in the tree (Mitchell et al., 2018). The method automatically detects clade-specific rate shifts by using reversible-jump Markov chain Monte Carlo (rjMCMC) to move between models with different numbers of parameters, which in this case represent different diversification rates. Speciation rates may be constant or vary over time following an exponential change func-

tion (Rabosky, 2014), while the extinction rate is assumed to be time-constant within each regime, and the rate of fossil preservation is held constant throughout the tree (Mitchell et al., 2018).

Before each BAMM analysis, we set priors on the initial rates of speciation and extinction and on the exponential rate change parameter using the `setBAMMpriors()` function from the R (R Core Team, 2019) package BAMMtools v2.1.6 (Rabosky et al., 2014). Each analysis was run for 20 million generations, sampling every 10,000 generations and excluding the first 10% of samples as burnin following a visual inspection of the posterior traces. Markov chain Monte Carlo (MCMC) convergence was diagnosed using the package coda v0.19-2 (Plummer et al., 2006) by ensuring that the effective sample size (ESS) of all parameters exceeded 200, an arbitrary but widely used threshold (Lanfear et al., 2016b; Ali et al., 2017). Using BAMMtools, we compared the prior and posterior support of models involving different numbers of rate shifts, and calculated their Bayes factor (BF) relative to the no-shift model. We also extracted the 95% credible set of configurations differing in the number and location of "core" shifts, or shifts that were sampled more often than expected under the prior alone. We used a marginal odds ratio of 5 (default BAMMtools threshold) to distinguish between core and non-core shifts. To determine overall trends in marginal rates, we used a modified BAMMtools function to plot the medians of the rate distributions through time across multiple analyses.

To accommodate topological and divergence-time uncertainty, we ran Fossil BAMM on each of the 16 BEAST maximum clade credibility trees, using the PBDB data to calculate the number of occurrences associated with their tips. Taxa present in the trees but not in the PBDB data were assumed to have a single occurrence each. The global sampling fraction (Chang et al., 2020) was calculated as the number of ingroup species in each tree divided by the total number of valid species in our occurrence data (HaEA, MEA: 0.186; HeEA: 0.136). For each tree, we conducted both a time-constant and time-varying analysis.

To determine the strength of the signal for rate variation through time, we modified the R scripts of Friedman et al. (2019) to calculate 95% credible intervals (CIs) about the means of the speciation and net diversification rates at the root and at the tips and assessed their overlap. To evaluate prior sensitivity (Mitchell and Rabosky, 2017), we additionally ran each analysis under three different values of the expected number of rate shifts (0.1, 1, 10), for a total of 96 analyses.

PyRate

We used PyRate (Silvestro et al., 2014a,b, 2019) to estimate ornithischian macroevolutionary rates from fossil occurrence data. PyRate implements a hierarchical Bayesian approach that estimates the speciation and extinction times of every lineage based on the temporal distribution of fossil occurrences, a Poisson fossil sampling process, and a birth-death prior (Silvestro et al., 2014b). The speciation and extinction hyperparameters of the birth-death prior are assumed to be only piecewise-constant, allowing for an arbitrary number of rates separated by tree-wide shifts, and are themselves estimated from the data along with the speciation and extinction times and the fossil preservation rate using a transdimensional MCMC sampler. All lineages are assumed to be connected by an unknown underlying phylogeny in which every species has been sampled at least once (Silvestro et al., 2014b). The method makes use of all fossil occurrences, including those that fall between the first and last appearance, and those that constitute the only known record of a given lineage (singletons) (Silvestro et al., 2019).

We created two types of data sets: one in which the age of every occurrence was independently drawn from the corresponding range, and one "site-linked" data set. Site-linking accommodates the uncertainty regarding the age of a given fossil site while simultaneously accounting for the fact that the ages of all fossils from that site are approximately equal (King and Rücklin, 2020). In a site-linked analysis, a single age is drawn from the temporal

range of every site and assigned to all the occurrences from the same site. We treated collection number metadata as a proxy for unique sites; manually added entries without metadata were assumed to each come from a unique site, resulting in 1087 sites per 1240 occurrences. We accounted for occurrence age uncertainty by replicating this procedure 10 times. Following Silvestro et al. (2019), we used the PyRate model selection tool based on the corrected Akaike information criterion (AICc) to compare the fit of the homogeneous Poisson preservation (HPP) model, corresponding to the frequentist null hypothesis; the nonhomogeneous HPP (NHPP) model, which allows the rate of preservation to change over the lifespan of each lineage; and two time-varying (TPP) models that were piecewise-homogeneous within each geochronological period (5 time bins: Early Jurassic through Late Cretaceous) or age (23 time bins: Hettangian through Maastrichtian). Based on the results, we limited all subsequent analyses to the HPP, NHPP, and TPP-age models, as the TPP-period model was intermediate between the above in terms of both fit and complexity (Supplementary Information, Table C.5 available on Dryad). All three models were coupled with a four-category discrete Γ model of among-lineage preservation rate heterogeneity.

A total of 120 PyRate analyses were performed, varying the data set (site-linked vs. site-unlinked), replicate (10 per data set), preservation model (HPP, NHPP, TPP- age), and model averaging algorithm: in addition to the rjMCMC algorithm described by Silvestro et al. (2019), we also used birth-death Markov chain Monte Carlo (bdMCMC; Silvestro et al., 2014b), an alternative transdimensional MCMC sampler in which new rate parameters are introduced and removed following a birth-death process. Our preliminary runs showed that rjMCMC took longer to reach stationarity than bdMCMC; therefore, rjMCMC analyses were run for 125 million generations (sampling every 50,000), as opposed to 75 million generations at the same sampling frequency for bdMCMC. We considered the analyses to have converged if, after removing the first 10% of the chain as burnin, the ESS values were <200 for no more than 10 parameters. We combined the post-burnin posteriors of all replicates that

reached convergence, and summarized the posterior probabilities of models involving different numbers of speciation and extinction rates. We further extracted the marginal posterior distributions of speciation, extinction, and net diversification rates within 1 myr time bins, and calculated their means and 95% CIs to build rates-through-time (RTT) plots.

3.3.5 Simulations

To assess how well PyRate and Fossil BAMM estimate the macroevolutionary and preservation rates, we simulated phylogenies and fossil records under two different scenarios. In the Decreasing Speciation scenario, the extinction and sampling rates were held constant at $\mu = 0.07\ \text{sp} \cdot \text{myr}^{-1}$, $\psi = 0.567\ \text{occ} \cdot \text{sp}^{-1} \cdot \text{myr}^{-1}$, while the rate of speciation underwent a 4-fold decrease from $\lambda_1 = 0.2\ \text{sp} \cdot \text{myr}^{-1}$ to $\lambda_2 = 0.05\ \text{sp} \cdot \text{myr}^{-1}$ 70 myr before the end of the process (Fig. 3.2). In the Decreasing Preservation scenario, the speciation and extinction rates were held constant at $\lambda = 0.1\ \text{sp} \cdot \text{myr}^{-1}$, $\mu = 0.07\ \text{sp} \cdot \text{myr}^{-1}$, while the rate of fossil preservation decreased 4-fold from $\psi_1 = 1.135\ \text{occ} \cdot \text{sp}^{-1} \cdot \text{myr}^{-1}$ to $\psi_2 = 0.284\ \text{occ} \cdot \text{sp}^{-1} \cdot \text{myr}^{-1}$ at the same time horizon (Fig. 3.2; see the Supplementary Information available on Dryad for a justification of the values used). This choice of simulation settings was motivated by several considerations. First, previous analyses employing PyRate (Condamine et al., 2016) as well as quantitative paleobiological methods (Smiley, 2018) showed speciation to track fossil preservation. This expected positive correlation facilitates interpretation: if the two rates are confounded, both scenarios should yield similar estimated speciation rates despite the different generating rates. Second, both scenarios tie in well with the empirical case, since a decline in the rate of speciation and its drop below the rate of extinction prior to the K–Pg boundary has been proposed for the ornithischians (Sakamoto et al., 2016), and the clade displays sharp variations in the preservation rate over its lifespan (Starrfelt and Liow, 2016).

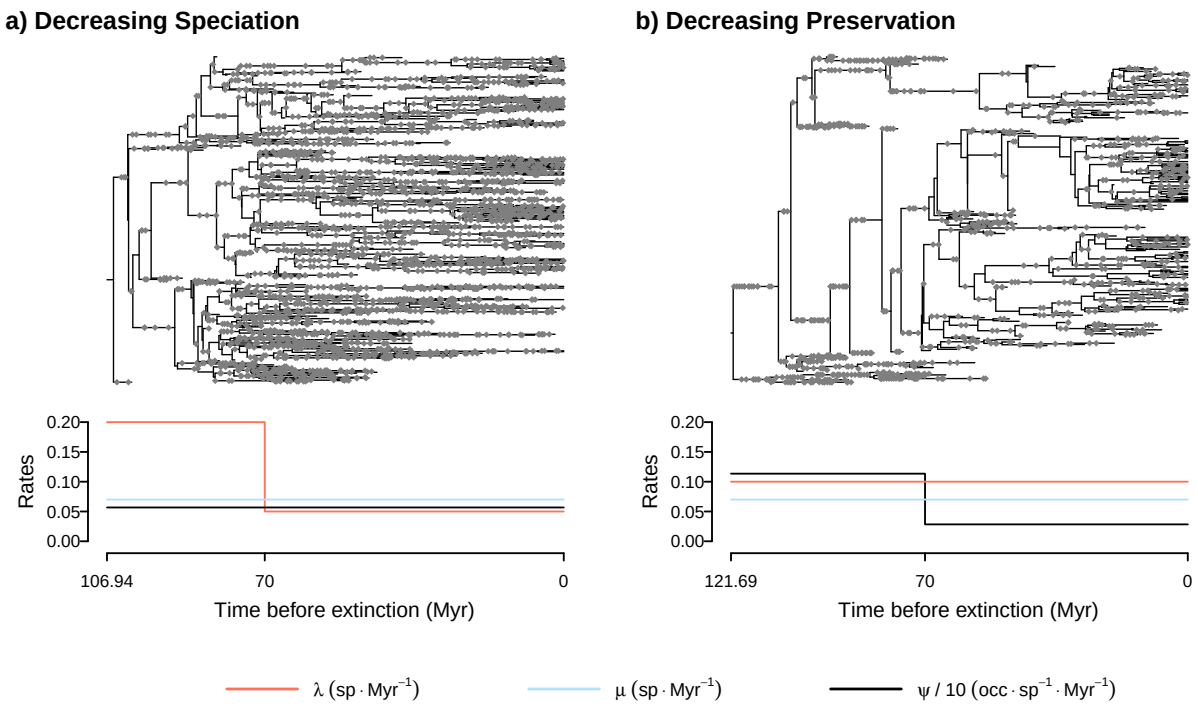

Figure 3.2: Examples of birth-death trees and the corresponding fossil records generated under the two simulation scenarios employed in this study. Each gray rectangle represents one fossil occurrence. a) Under Decreasing Speciation, the rates of extinction and preservation were held constant, while the rate of speciation underwent a 4-fold decrease. b) Under Decreasing Preservation, the rates of speciation and extinction were held constant, and the decrease affected the rate of preservation instead.

Using the R package TreeSim (Stadler, 2011b), we simulated 100 birth–death trees under each scenario and employed rejection sampling to ensure maximum congruence with the empirical data, selecting 10 trees per scenario (Supplementary Information, Table C.7 available on Dryad). We used the package FossilSim (Barido-Sottani et al., 2019b) to simulate fossil records on the synthetic phylogenies according to the preservation rates above. Exact sampling times were recorded for all occurrences to eliminate stratigraphic age uncertainty. To ensure that the rate estimates are not confounded by topological or divergence time error, Fossil BAMM was applied to the true (generating) phylogenies after pruning unsampled tips and truncating terminal edges to the age of their last occurrence. Additionally, we analyzed proportionally subsampled and 70-tip trees designed to mimic the phylogenetic data sets

available for Ornithischia, which only include a fraction of the known diversity of the clade (Supplementary Information, Fossil BAMM available on Dryad). Next, we used the AICc-based model selection implemented in PyRate to choose between the homogeneous Poisson preservation model ("HPP"), a time-varying model with a single shift at 136 Ma ("TPP-136"), and an arbitrarily overparameterized model allowing the preservation rate to change every 10 myr ("TPP-by10"), similar to the TPP-age model favored by the empirical analyses. We then analyzed each data set under both the best-performing model and the true model (HPP for Decreasing Speciation, TPP-136 for Decreasing Preservation) whenever the two differed (Supplementary Information, Table C.8 available on Dryad). All PyRate analyses were further repeated on subsampled data sets excluding sampled ancestors, which were abundant in the simulations but absent from the empirical phylogenies (Supplementary Information, Phylogenetic Analyses available on Dryad). For both Fossil BAMM and PyRate, we calculated the summary statistics described for the empirical analyses as well as the relative error and relative precision (width of the 95% CI divided by the mean rate) following Silvestro et al. (2019), averaging them first across time bins within a replicate and then across replicates.

3.4 Results

3.4.1 Diversification Rate Analyses

Fossil BAMM

We found no evidence of shifts between distinct diversification regimes throughout ornithischian evolutionary history. The 95% credible set of rate shift configurations contained a single configuration for all but two analyses (both based on unconstrained HaEA topologies, with constant within-regime speciation and one expected shift). The maximum *a posteriori* (MAP) configuration contained no shifts in any of the 96 analyses performed; in the two analyses where the 95% credible set included an additional configuration, this second-best

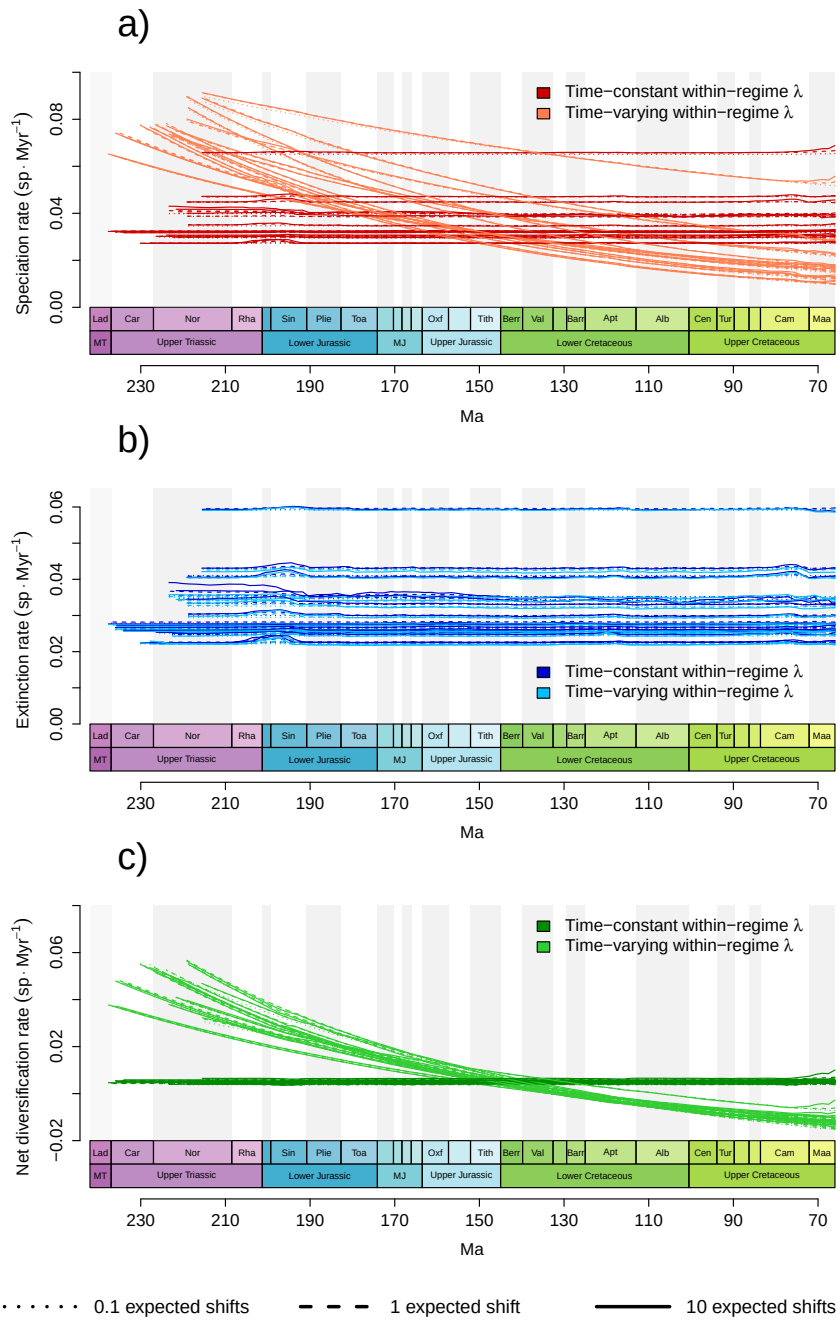

Figure 3.3: Diversification of Ornithischia estimated using Fossil BAMM, with rate-through-time plots (RTT) for a) speciation, b) extinction, and c) net diversification. Each curve represents the posterior mean from one analysis. The absence of support for multiple rate regimes results in net diversification rates that are either close to zero and constant (time-constant analyses), or gradually decay from the root toward the youngest tips (time-varying analyses).

configuration contained a single shift associated with the clade Ornithopoda. The posterior probability of the zero-shift configuration always exceeded its prior probability, even in cases where the zero-shift configuration was already strongly favored by the prior (expected number of shifts = 0.1). Model comparisons using BFs found no evidence in favor of models containing rate shifts relative to the zero-shift model, with BFs for alternative models ranging from 0.002 to 0.781.

Analyses assuming constant within-regime speciation yielded model-averaged rate estimates that were almost perfectly flat through time (Fig. 3.3); we thus report the average of mean rates at the root and at the youngest tips. Across the 48 time-constant analyses, the mean of root-tip mean net diversification rates was 0.00501 $\text{sp} \cdot \text{myr}^{-1}$ (range of root-tip mean rates = [0.0038, 0.0083] $\text{sp} \cdot \text{myr}^{-1}$), a low value reflecting low and nearly equal rates of speciation (0.0367 [0.0273, 0.0673] $\text{sp} \cdot \text{myr}^{-1}$), and extinction (0.0317 [0.0224, 0.0597] $\text{sp} \cdot \text{myr}^{-1}$). Analyses with time-varying within-regime speciation rates found strong support for a long-term decline, with 95% CIs showing no overlap between the root and the youngest tips in 41 and 39 out of 48 analyses for speciation and net diversification, respectively. In both cases, overlapping 95% CIs resulted from some but not all analyses based on zero-ρ HeEA and MEA trees, indicating that the estimated slowdown was due to neither taxon sampling nor topology. The mean of mean net diversification rates was 0.0463 $\text{sp} \cdot \text{myr}^{-1}$ (range of means = [0.0307, 0.0567] $\text{sp} \cdot \text{myr}^{-1}$) at the root and -0.0117 $\text{sp} \cdot \text{myr}^{-1}$ ([-0.0153,-0.0028] $\text{sp} \cdot \text{myr}^{-1}$) at the youngest tips. The 95% CIs were notably wider at the root (mean width = 0.0758 $\text{sp} \cdot \text{myr}^{-1}$) than at the youngest tips (0.0233 $\text{sp} \cdot \text{myr}^{-1}$) for speciation but not for extinction (root: 0.0196 $\text{sp} \cdot \text{myr}^{-1}$, youngest tips: 0.0199 $\text{sp} \cdot \text{myr}^{-1}$).

PyRate

The parameter-rich TPP-age preservation model exhibited the best fit to both site-linked and site-unlinked data (Supplementary Information, Table C.5 available on Dryad).

The fossil sampling rates estimated under the TPP model were broadly consistent between the bdMCMC and rjMCMC analyses, showing episodes of increased preservation in the Kimmeridgian–Tithonian, the Albian, and the Maastrichtian, with substantial uncertainty through much of the Early and Middle Jurassic (Supplementary Information, Figs. C.20–C.23 available on Dryad). In analyses subsequent to the model selection step, we experienced substantial convergence issues with the rjMCMC algorithm. In 2 out of the 12 analytical scenarios (site-linked and site-unlinked rjMCMC+HPP), all replicates failed to reach the ESS threshold for >10 parameters, and we thus excluded these scenarios from further comparisons.

The bdMCMC analyses inferred a diversification pattern similar to the results from time-constant BAMM, with the exception of a sharp drop in the speciation rate and an uptick in the extinction rate 10–15 myr before the K–Pg boundary. These shifts may represent edge effects, a known artifact in which the clustering of first appearance dates at the beginning of the studied interval and of last appearance dates toward its end biases the estimated rates of speciation and extinction (Foote, 2000). However, the absolute magnitudes of pre-80 Ma rates of speciation (range of time-averaged mean rates: [0.1737, 0.2684] $\text{sp} \cdot \text{myr}^{-1}$) and extinction ([0.1276, 0.1926] $\text{sp} \cdot \text{myr}^{-1}$) as well as their difference were an order of magnitude higher for PyRate than for BAMM, with the poorly fitting NHPP model favoring slightly lower rates than the other models (Fig. 3.4). Consistent with the rate drops observed in the RTT plots, all our bdMCMC analyses inferred one shift in speciation and one shift in extinction (Supplementary Information, Table C.6 available on Dryad). The second most frequently sampled speciation model involved one additional shift, yielding a slight decrease in the marginal rate around 160 Ma (Fig. 3.4).

Our rjMCMC analyses found two major periods of elevated speciation and extinction rates in the late Early through Middle Jurassic and in the early Late Cretaceous, characterized by biologically implausible rates (up to >5 $\text{sp} \cdot \text{myr}^{-1}$ for the Jurassic peak) (Fig. 3.4).

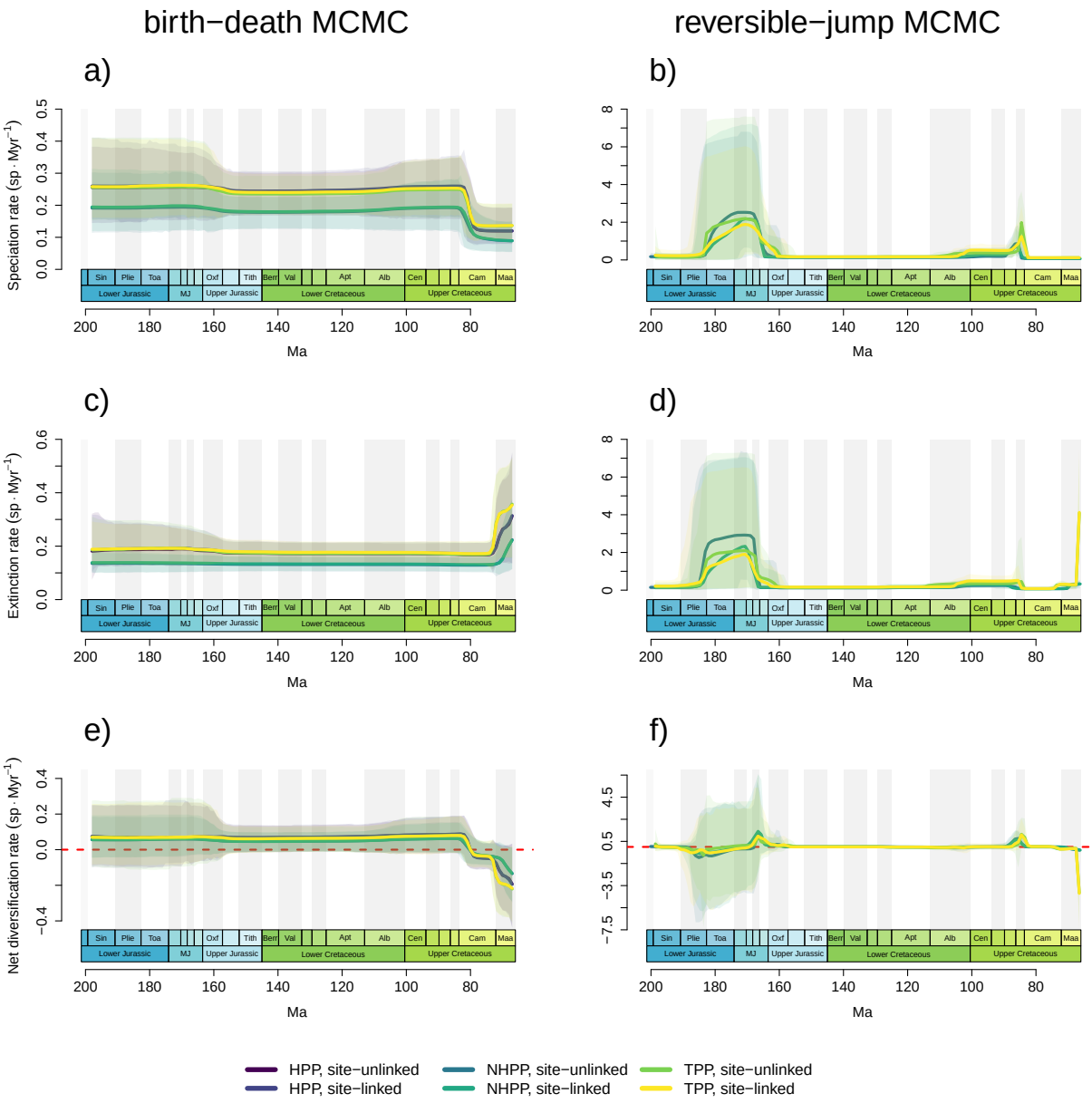

Figure 3.4: Diversification of Ornithischia estimated using PyRate under two different model-averaging algorithms, with RTT plots for (a, b) speciation, (c, d) extinction, and (e, f) net diversification. Each curve represents the posterior mean averaged across all replicates that reached convergence ($\leq$10 parameters with effective sample sizes of $<$200). The shaded areas show the associated 95% credibility intervals. The dashed line highlights the boundary between positive and negative net diversification. Aside from edge effects near the K–Pg boundary, the bdMCMC results resemble those obtained with BAMM, whereas the rjMCMC algorithm yields highly correlated speciation and extinction rates that accelerate to implausible values in the late Early Jurassic and the early Late Cretaceous.

The support for all four corresponding shifts (at the beginning and end of either period) was always positive (BF $>$ 2) and often strong (BF $>$ 6) for both speciation (13 out of the 19 rjMCMC analyses that reached convergence) and extinction (14 analyses). The peaks were close to, but did not reach, the edges of the analyzed time interval, and we therefore consider them distinct from the usual edge effects, which we also observed in some analyses (Fig. 3.4). Between the peaks, the inferred rates were nearly constant and similar to the bdMCMC estimates for speciation (range of time-averaged mean rates between 160 and 90 Ma: [0.1329, 0.2684] sp$\cdot$myr^{-1}) but slightly higher for extinction ([0.1458, 0.2608] sp$\cdot$myr^{-1}), occasionally producing negative net diversification rates. As in the bdMCMC analyses, the marginal rates were strongly correlated, with the pulses of speciation almost entirely offset in magnitude by pulses of extinction; however, due to slight differences in timing, the net diversification rate obtained by their subtraction displayed a series of short-lived peaks and dips (Fig. 3.4f). The rjMCMC algorithm consistently supported more rate shifts compared to bdMCMC in both speciation and extinction. The latter algorithm never sampled more than 7 extinction rates, while the rjMCMC analyses favored up to 8 rates for extinction, and sampled as many as 13 (Supplementary Information, Table C.6 available on Dryad).

3.4.2 Simulations

Fossil BAMM

BAMM analyses consistently inferred a single rate regime under Decreasing Speciation, with only 6 out of 40 analyses including at least one shift in their MAP configuration. The analytical scenarios in which the MAP configurations did contain one or more shifts also tended to produce larger 95% credible shift sets (Supplementary Information, Table C.9 available on Dryad); however, these never included configurations with more than 6 shifts, even though all Decreasing Speciation trees had 7–13 well-sampled branches crossing the time

	Fossil BAMM		PyRate	
Variable monitored	Decreasing speciation	Decreasing preservation	Decreasing speciation	Decreasing preservation
Number and timing of λ shifts	True tree-wide shift approximated by a single; regime with declining λ; occasional spurious shifts	Constant λ almost always correctly inferred	Number overestimated (bdMCMC) or correctly inferred (rjMCMC); true shift accurately placed	Constant λ correctly inferred
Number and timing of μ shifts	Constant μ correctly inferred	Constant μ correctly inferred	Spurious μ shifts not always attributable to edge effects	Spurious μ shifts attributable to edge effects
Number and timing of ψ shifts	Not applicable	Not inferred due to inherent limitations of the method	Time-variable ψ model incorrectly chosen over a time-homogeneous one	ψ model with the true shift correctly chosen
Marginal λ rates	Moderate (time-varying BAMM) to low (time-constant BAMM) accuracy, moderate precision, no bias	High accuracy, moderate precision, slightly biased toward overestimates	Low accuracy, moderate to low precision, consistently overestimated	Low accuracy, moderate to low precision, consistently overestimated
Marginal μ rates	High accuracy, moderate precision, consistently overestimated	Moderate accuracy, moderate precision, consistently overestimated	Low accuracy, moderate to low precision, consistently overestimated	Low accuracy, moderate to low precision, consistently overestimated
Marginal ψ rates	High accuracy, high precision, poor coverage, consistently overestimated	Intermediate between the two true rates	Correctly inferred as near-constant; occasional spurious shifts	High accuracy, moderate precision, high coverage
Spurious correlation between λ and ψ	Not applicable	Not applicable	In a minority of analyses	Absent
Spurious correlation between λ and μ	Present in marginal RTT plots and spurious shifts	Present in marginal RTT plots	Present and inducing a spurious shift in μ	Present; $(\lambda - \mu)$ estimated accurately despite both λ and μ being overestimated
Effect of subsampling	λ and μ estimated with lower accuracy and lower precision; both consistently underestimated	λ but not μ estimated with lower accuracy; both less precise and consistently underestimated	λ and μ estimated with lower precision; more among-replicate variation close to the origin under bdMCMC	λ and μ estimated with lower precision; more among-replicate variation close to the origin under bdMCMC
Effect of analytical settings	Weak for the expected number of shifts, moderate for within-regime speciation (time-constant vs. time-variable)		Weak for the preservation model, substantial for model-averaging algorithm (bdMCMC vs. rjMCMC)	

Table 3.1: Overview of the various aspects of Fossil BAMM and PyRate performance based on the analyses of simulated data. Accuracy and precision are measured by relative error and the relative width of the 95% credible interval (CI), respectively. Coverage is measured by the proportion of cases in which the true value is included in the 95% CI. "Subsampling" refers to the 70-tip and proportionally subsampled schemes for BAMM (Supplementary Information, Fossil BAMM available on Dryad) and to the exclusion of sampled ancestors for PyRate.

horizon of the shift (Supplementary Information, Table C.7 available on Dryad). When the MAP configuration did include one (five analyses) or two (one analysis) shifts, these were generally associated with accelerated rather than decelerated speciation and postdated the true tree-wide speciation drop (Supplementary Information, Figs. S29–S31 available on Dryad), suggesting that the method picked up stochastic variation rather than the generating birth-death process. Despite the absence of discrete shifts, the general tendency for the speciation rate to decrease over time was correctly inferred (Table 3.1). Seventeen out of 20 analyses with time-varying within-regime speciation exhibited non-overlapping 95% CIs about the net diversification rate at the root and at the youngest tips. The rates at the root and at the youngest tips were close to their true values in time-varying analyses, and the mean speciation rates (averaged over the entire history of each tree) inferred by time-constant analyses were intermediate between the two generating rates (Fig. 3.5). In contrast, subsampled analyses underestimated the speciation rate in both cases (Supplementary Information, Fig. C.25 available on Dryad).

The extinction rate was consistently overestimated (Fig. 3.5), and its high relative precision corresponded to low coverage, with the true value included in the 95% CI in just 4 out of 40 cases both at the root and at the youngest tips. In the time-constant analyses, the 95% CI did not contain the true value at any point throughout the history of the clade. The preservation rate estimates were fairly accurate but biased and overconfident; the rate was slightly but consistently overestimated (range of means: [0.595, 0.637] $\text{occ} \cdot \text{sp}^{-1} \cdot \text{myr}^{-1}$), and the narrow 95% CIs (relative precision: 0.069) never included the true value.

Under Decreasing Preservation, BAMM correctly inferred the absence of speciation rate shifts, with only 1 out of 40 analyses containing a shift in its MAP configuration. In time-varying analyses, the 95% CIs about the net diversification rate at the root and at the youngest tips overlapped in 19 out of 20 cases. The marginal speciation rates were estimated with a high degree of accuracy (Fig. 3.5; Supplementary Information, Table C.10 available

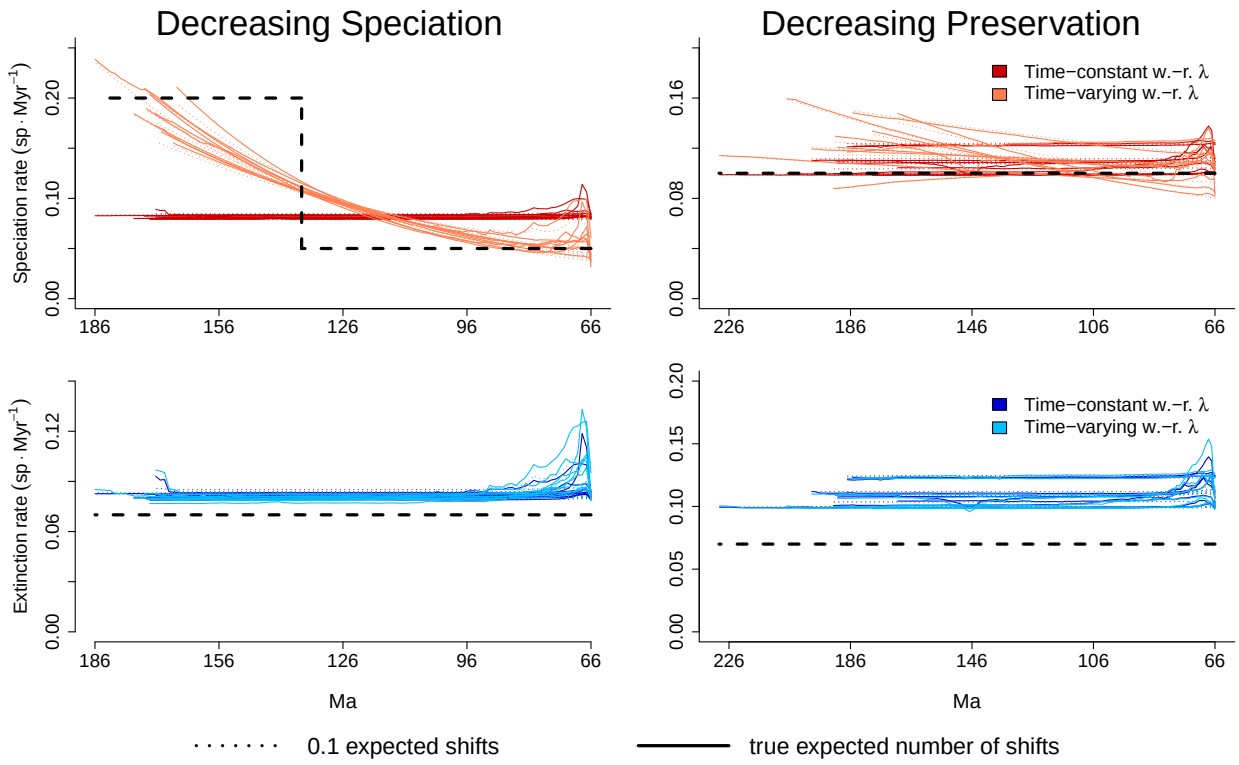

Figure 3.5: Speciation and extinction RTT plots from the Fossil BAMM analyses of simulated data. Each curve represents the posterior mean from one analysis; thick dashed lines represent the true values. The shift in the Decreasing Speciation scenario is not identified as a discrete event but approximated by the declining rates in the time-varying analyses.

on Dryad). On average, the 95% CIs were wider at the root (0.888) than at the youngest tips (0.531), but both values corresponded to comparably high coverage (true value included in 17 and 19 out of 20 cases at the root and at the youngest tips, respectively). Similarly, the true value fell within the CI at least at one point in the history of the clade in 17 out of the 20 time-constant analyses. In the subsampled analyses, accuracy, precision, and coverage were markedly poorer (Supplementary Information, Fossil BAMM available on Dryad), suggesting a positive relationship between performance and tree size.

As in the Decreasing Speciation scenario, the extinction rate was systematically overestimated (Fig. 3.5). In contrast to speciation, the width of the 95% CIs did not differ appreciably between the root (0.310) and the youngest tips (0.341) in the time-varying analyses, resulting

in similarly poor coverage in both cases (true value included in 1 and 0 out of 20 cases at the root and at the youngest tips, respectively). Low coverage also characterized the extinction rate estimates from time-constant analyses, where the 95% CIs included the true value at least once in just 2 out of 20 cases. The preservation rate was intermediate between the two true values (range of means: [0.416, 0.602] $\text{occ} \cdot \text{sp}^{-1} \cdot \text{myr}^{-1}$).

PyRate

Our PyRate simulations suffered from fewer convergence issues than the empirical analyses. As a result, only 5 replicates out of 122 (3 for bdMCMC+HPP and 2 for bdMCMC+TPP, both in the Decreasing Speciation scenario) had to be excluded from the following comparisons. Under Decreasing Speciation, model selection failed to identify the true preservation model (HPP), favoring the more parameter-rich TPP-136 and TPP-by10 models in 13 and 7 out of 20 cases, respectively (Supplementary Information, Table C.8 available on Dryad). Nevertheless, the marginal preservation rates were mostly flat through time even in the TPP analyses, with most variation restricted to a period of uncertainty early in the history of the clade (Fig. 3.6). An exception consisted of a small but occasionally significant decrease at the 136 Ma time horizon; in 7 out of the 38 TPP analyses that reached convergence, there was no overlap of the 95% CIs for the time intervals immediately preceding and following the 136 Ma rate shift, indicating possible conflation of speciation and preservation rate changes. Analyses run under the true (HPP) preservation model slightly but systematically overestimated the true rate (range of means: [0.569, 0.642] $\text{occ} \cdot \text{sp}^{-1} \cdot \text{myr}^{-1}$), with the 95% CI including the true value in just 2 out of the 37 HPP analyses that reached convergence.

In contrast to the empirical analyses, the rjMCMC algorithm favored models with fewer rate shifts for both speciation and extinction. The rjMCMC analyses correctly inferred two distinct speciation rates under Decreasing Speciation, while the bdMCMC runs suppor- ted a third rate (Supplementary Information, Table C.11 available on Dryad). Both algorithms

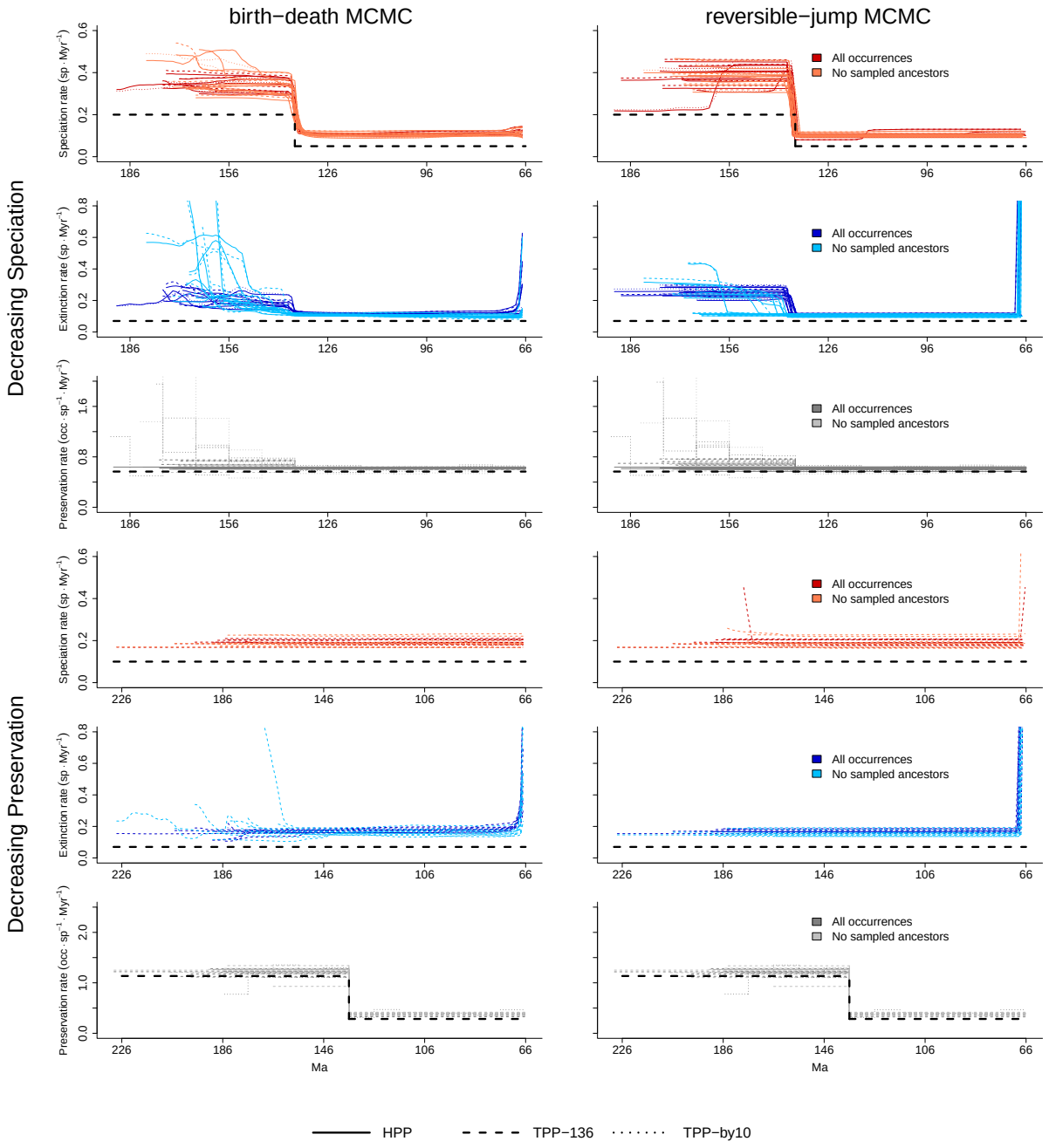

Figure 3.6: Speciation, extinction, and preservation RTT plots from the PyRate analyses of simulated data (cf. Fig. 3.5). The rate drop in the Decreasing Speciation scenario is correctly inferred but counterbalanced by a spurious drop in the rate of extinction. The rates under Decreasing Preservation are overestimated in absolute terms, but correctly inferred as constant through time despite the presence of a potentially confounding drop in the rate of preservation.

overestimated the number of extinction rates. The RTT plots accurately reflect the overall speciation dynamics, with a sharp drop at 136 Ma successfully identified by all 80 analyses. In the rjMCMC analyses, which allowed estimating marginal posterior rate shift probabilities (Silvestro et al., 2019), the probability of this drop usually approached 1. Spurious speciation rate shifts were sometimes inferred in the early history of the clade, especially in the analyses excluding sampled ancestors (Fig. 3.6). This is congruent with the expectation that sampled ancestors should form a higher proportion of lineages close to the origin of a clade, and their removal should thus disproportionately affect rate inference therein.

While the number of spurious speciation shifts was relatively low, a spurious drop in the extinction rate temporally coinciding with that in the rate of speciation appeared in most of the 80 analyses. In the rjMCMC analyses, its posterior probability was similar to that of the speciation shift (i.e., close to 1) when sampled ancestors were included but lower (<0.6) when they were excluded. The analyses without this spurious drop yielded other erroneous inferences about the extinction dynamics, including drops predating the 136 Ma time horizon, gradual decline between the origin of the clade and the 136 Ma boundary, or multiple shifts (Fig. 3.6). The magnitudes of the speciation rates both before and after the shift were systematically overestimated, as was the extinction rate. Interestingly, the highly correlated and overestimated rates combined to produce remarkably accurate estimates of the net diversification rate (Table 3.1), which were generally very close to the true values both before ($0.13\ \text{sp} \cdot \text{myr}^{-1}$) and after ($-0.02\ \text{sp} \cdot \text{myr}^{-1}$) the shift (Supplementary Information, Fig. C.27 available on Dryad). However, because the temporal correlation between the speciation and extinction rate shifts was imperfect, and because of edge effects as well as the spurious extinction rate fluctuations, even the net diversification RTT plots usually failed to capture the true clade dynamics. Typical spurious results included a brief period of negative rates following the origin of the clade (common in bdMCMC analyses without

sampled ancestors), and an uptick immediately preceding the drop at 136 Ma (common in rjMCMC analyses).

Under Decreasing Preservation, PyRate favored the true (TPP-136) model in 19 out of 20 cases; in one case, the more parameter-rich TPP-by10 model was preferred. The simulated drop in the preservation rate at 136 Ma was accurately detected in all cases (Fig. 3.6), with none of the 42 analyses showing an overlap between the 95% CIs before and after the shift. When the true model was used, the true value was usually included in the 95% CI about the pre-shift rate (30 out of 40 cases) but never in the 95% CI about the post-shift rate, which was slightly but consistently overestimated (range of means: [0.325, 0.418] $\text{occ} \cdot \text{sp}^{-1} \cdot \text{myr}^{-1}$). In the two TPP-by10 analyses, the 95% CIs contained the true values in all five pre-shift time periods, but only in the first of the seven post-shift periods.

Both the bdMCMC and rjMCMC algorithms correctly favored the model with a single speciation rate, and both again overestimated the number of distinct extinction rates (Supplementary Information, Table C.11 available on Dryad). The extinction model misspecification was caused by an edge effect consisting of a sharp rise in the extinction rate shortly before the end of the observed time interval. Spurious shifts in the speciation rate were rare and again attributable to edge effects at the beginning or the end of the clade's lifespan (Fig. 3.6). Aside from these, no shift had a posterior probability of >0.25 in the rjMCMC analyses (for which this metric was available), and no analysis placed a speciation rate drop at 136 Ma, indicating that there was no conflation of speciation with preservation in this scenario. Similarly, the extinction rates were correctly inferred as flat throughout much of the clade's history, with rare exceptions limited to a short period following the origin of the clade and bdMCMC analyses without sampled ancestors (Fig. 3.6).

In absolute terms, the rates of speciation and extinction again displayed a bias toward higher values. While errors in the inference of the overall extinction dynamics were less serious under Decreasing Preservation than under Decreasing Speciation, they ultimately resulted

in even lower levels of accuracy when quantified using the relative error (Supplementary Information, Table C.12 available on Dryad). The accuracy of the speciation rate estimates was slightly higher and comparable between the two scenarios; however, their higher precision under Decreasing Preservation resulted in especially poor coverage, with only 2 out of 42 replicates containing the true value in their 95% CIs at any point in the history of the clade. As with Decreasing Speciation, the biases in the estimates of both rates largely canceled out when the two were subtracted to obtain the net diversification rate, whose true value (0.03 $\text{sp} \cdot \text{myr}^{-1}$) was inferred with high accuracy except when subject to edge effects.

3.5 Discussion

We inferred the first time trees for Ornithischia based on three large morphological data sets, and combined them with curated occurrence data to infer the dynamics of ornithischian diversification using Fossil BAMM (Mitchell et al., 2018) and PyRate (Silvestro et al., 2014a,b, 2019). We found results that were markedly inconsistent between both methods as well as between different implementations of each approach (time-constant vs. time-varying within-regime speciation in BAMM, bdMCMC vs. rjMCMC algorithms in PyRate). To investigate the causes of these discrepancies, we simulated phylogenies and fossil records approximating the ornithischian data under two simple scenarios. These were designed to test the ability of both methods to disentangle time-varying speciation, extinction, and preservation rates under conditions typical of data sets used in vertebrate paleontology. Analyses of the synthetic data led to the overestimation of the number of rate shifts (PyRate) or their smearing across the history of the clade (Fossil BAMM), and revealed cases of inaccurate or spuriously correlated marginal rate estimates (Table 3.1). Our results suggest that the data sets available for many extinct clades may not be sufficient to provide reliable inferences about macroevolutionary dynamics when analyzed using existing methods, and offer insights into how the applicability of such methods to fossil data may be improved.

3.5.1 Implications for Ornithischian Diversification

Previous analyses of ornithischian macroevolutionary dynamics have mostly focused on absolute species richness rather than speciation or net diversification rates (Barrett et al., 2009; Lloyd, 2012; Brusatte et al., 2015; Chiarenza et al., 2019), hindering comparisons of the present results with earlier literature. Multiple studies have hypothesized that Ornithischia and other Mesozoic dinosaurs entered a period of decline prior to the K–Pg mass extinction (Fastovsky et al., 2004; Wang and Dodson, 2006; Chiarenza et al., 2019). This period is often regarded as restricted to the Maastrichtian (Archibald, 1996, 2011; Barrett et al., 2009), though a longer-term decline spanning much of the Cretaceous has been suggested (Lloyd, 2012) and supported by one of the few attempts to directly estimate ornithischian diversification rates (Sakamoto et al., 2016). The scenario of a diversity diminution restricted to the last 10 myr of the Cretaceous is broadly compatible with our bdMCMC PyRate results, which show negative net diversification rates in this time interval. In contrast to the gradual waning proposed by earlier studies (Archibald, 1996, 2011), our bdMCMC PyRate analyses support a two- stage decline comprising a discrete drop in speciation in the early Campanian, followed by a rising extinction rate in the Maastrichtian (Fig. 3.4). However, given their proximity to the K–Pg boundary, these rate changes possibly represent edge effects and should not be overinterpreted. The analogous drop in net diversification observed in the rjMCMC analyses was driven solely by extinction, while speciation held steady through the Campanian–Maastrichtian (Fig. 3.4).

The results of our time-varying Fossil BAMM analyses do not support a discrete pre-K–Pg speciation drop or extinction spike, but are congruent with the longer-term decline inferred by Sakamoto et al. (2016) using a regression-based framework. Both methods show that the ornithischians crossed the zero net diversification line in the Early Cretaceous and continued to decline until the K–Pg boundary (Fig. 3.3). The absolute net diversification rate estimates are also similar between the two approaches, with maximum rates of $\sim$0.06

$\text{sp}\cdot\text{myr}^{-1}$ attained early in the history of the clade (Fig. 3.3; Sakamoto et al., 2016). However, the results of Sakamoto et al. (2016) were recently interpreted as a possible artifact of failing to sample divergences that narrowly predate the K–Pg boundary (Wagner, 2019). Fossil BAMM, which also relies on the temporal distribution of divergences throughout a phylogeny, may be subject to the same issue, suggesting that an explicit parameterization of mass extinctions may be needed to capture the diversification dynamics of clades like Ornithischia.

Compared to Fossil BAMM and PyRate with the bdMCMC algorithm, the rjMCMC PyRate analyses revealed a much more complex diversification history for Ornithischia, marked by multiple correlated shifts in speciation and extinction (Fig. 3.4). However, the magnitudes of the shifted rates estimated by rjMCMC cast doubt on the reliability of the results. The mean speciation rates of 2–3 $\text{sp} \cdot \text{myr}^{-1}$ inferred for the Early–Middle Jurassic spike (Fig. 3.6) rival the fastest rates ever recorded for vertebrates, including those of the white-eyes (1.95–2.63 $\text{sp} \cdot \text{myr}^{-1}$; Moyle et al., 2009), *Todiramphus* kingfishers (2.01–4.49 $\text{sp} \cdot \text{myr}^{-1}$; Andersen et al., 2015), and haplochromine cichlids (tip rates of up to 3.13–3.42 $\text{sp} \cdot \text{myr}^{-1}$; Chang et al., 2019); and the values included in the 95% CI (up to 7.5 $\text{sp} \cdot \text{myr}^{-1}$) substantially exceed them. The speciation rates during the Coniacian–Santonian spike are only slightly lower, and the extinction rates are similarly high for both peaks (Fig. 3.4).

The strong intra-method and inter-method incongruence calls into question the reliability of the mac- roevolutionary inferences obtained from our data. In particular, the question of a pre-K–Pg decline in speciation remains unresolved, with flat rates, a two-stage drop, or prolonged slowdown all receiving support depending on the method used. The small data sets employed here (∼50–70 tips, 1200 occurrences) may not contain a signal of the clade's diversification history that is strong enough to be independently picked up under the different modeling assumptions adopted by the two methods. Given the unexceptional quality of the ornithischian fossil record, and the fact that the ornithischian time trees generated for this

study are fairly typical of paleontological phylogenies in terms of the number of tips, this conclusion is likely generalizable to numerous other extinct clades. We thus recommend that whenever possible, phylogeny-based inferences about macroevolutionary rates in extinct clades should be corroborated using multiple analytical frameworks with distinct modeling assumptions.

3.5.2 Rate Shift Models vs. Model-Averaged Rates

One of the attractive features of Fossil BAMM and PyRate consists in their ability to fit models of varying complexity and average the estimates of shared parameters across the different models, each weighted by its posterior probability (Rabosky, 2014; Silvestro et al., 2019). Therefore, both the relative probabilities of different models and the model-averaged parameter values are of interest. Using simulations, we found that PyRate frequently failed to assign the highest posterior probability to the model containing the true number of speciation and extinction rate shifts, while Fossil BAMM accounted for the tree-wide shift in the generating model by fitting a single regime and smearing the rate decline across its duration (Fig. 3.5), a tendency also noted by Mitchell et al. (2018). The failure of the tree-wide shift to induce a large number of simultaneous clade-specific shifts in BAMM analyses is unsurprising given the previously noted difficulty of detecting small rate regimes (Mitchell and Rabosky, 2017; Mitchell et al., 2018), with the little-discussed bias of BAMM toward preferentially inferring shifts to regimes with increased rather than decreased rates (cf. Mitchell et al., 2018, Fig. 7b) as another potential contributing factor. Arguably, the tree-wide decline in speciation inferred by BAMM is preferable to selectively mapping discrete rate drops onto sufficiently speciose clades only, and provides an imprecise but basically correct assessment of the overall diversification dynamics.

The overestimation of the number of rate shifts by PyRate appears to be due in part to edge artifacts, the causes of which are well-understood and can potentially be corrected

for (Pires et al., 2018). Like Silvestro et al. (2019), we find rjMCMC to outperform bdMCMC when estimating the number of distinct speciation rates. However, in contrast to the findings of Silvestro et al. (2019), this is because bdMCMC overestimates rather than underestimates rate heterogeneity through time (Supplementary Information, Table C.11 available on Dryad). In PyRate, the number of shifts is assigned a Poisson prior with rate parameter r, where r is fixed under bdMCMC and estimated under rjMCMC (Silvestro et al., 2014b, 2019). Since the rjMCMC prior adapts to the data, while the bdMCMC one does not, the difference in findings likely reflects the smaller number of shifts in our generating model compared to those of Silvestro et al. (2019). Interestingly, our simulation results also contrast with our empirical bdMCMC analyses, which tended to include fewer shifts than the corresponding rjMCMC runs (Supplementary Information, Table C.6 available on Dryad). This may indicate genuine support for multiple shifts in our data that were better picked up by the flexible rjMCMC prior. However, it is also conceivable that when the ability of the data to accurately infer r is low, bdMCMC analyses that fix r to a plausible value can outperform the rjMCMC analyses that estimate it. The previously hypothesized conflation of macroevolutionary and preservation rates (Condamine et al., 2016) may represent another factor contributing to the appearance of spurious shifts. This effect was negligible in our simulations (Table 3.1; Fig. 3.6), but the highest-probability speciation and extinction rate shifts inferred by the rjMCMC analyses of the empirical data corresponded to an uptick in the number of geological sampling proxies (Supplementary Information, Fig. C.2 available on Dryad) and a concomitant cluster of first appearance dates (Supplementary Information, Fig. C.1 available on Dryad).

We corroborate the previously reported "paradox" of Bayesian diversification rate inference estimating marginal rates more successfully than the exact number of rate shifts (Mitchell et al., 2018). Our model-averaged rate estimates are robust to most sources of variation introduced into the simulations. The marginal rates inferred by PyRate are largely

unaffected by the choice of the preservation model and the exclusion of sampled ancestors (Fig. 3.6). The latter is particularly remarkable, since the removal of sampled ancestors not only violates PyRate's assumption that every species has been sampled at least once (Warnock et al., 2020), but does so in a nonrandom manner. For BAMM we find that, similar to model probabilities (Mitchell and Rabosky, 2017; Rabosky et al., 2017), the model-averaged rates are insensitive to the prior on the expected number of shifts (Figs. 3.3 and 3.5). Similar to rate variance (Rabosky, 2019) and the power to detect rate shifts (Rabosky et al., 2017; Kodandaramaiah and Murali, 2018), the accuracy of marginal rates depends more on tree size, as indicated by analyses performed on artificially subsampled trees of 23–70 tips (Supplementary Information, Fossil BAMM available on Dryad). However, this relationship is not straightforward, as shown by the extinction rate inference under Decreasing Preservation (cf. Fig. 3.5 and Supplementary Information, Fig. C.25 available on Dryad), and should be explored over a wider range of tree sizes. More worryingly, we find evidence for spurious correlations between the speciation and extinction rates inferred by both PyRate and BAMM (Table 3.1), hinting at a possibility that only quantities that confound both rates, such as net diversification or the extinction fraction, may be estimable from low-information data sets (Supplementary Information, Results and Discussion available on Dryad).

3.5.3 Challenges to Estimating the Diversification Dynamics of Extinct Clades

The diversification dynamics of extinct groups, marked by symmetric phases of diversity expansion and contraction (Foote, 2007), may differ from that of extant clades, which appear to steadily accumulate diversity from their origin to the present (Rabosky et al., 2012; Burin et al., 2019; but see Quental and Marshall, 2010). The fact that BAMM and PyRate have been developed or predominantly used for the analysis of extant groups raises the question of whether their performance in the present study is not simply an effect of stretching their

application beyond the use cases they were intended for. However, this is arguably not the case. The ornithischian dinosaurs, used here as an empirical system, were extirpated in the near-instantaneous K–Pg extinction event (Brusatte et al., 2015) at high standing diversity (a phenomenon commonly observed in mass extinctions; Jablonski, 1986, and it is unclear from this study whether they entered a period of decline prior to their demise (see also Wang and Dodson, 2006; Barrett et al., 2009; Sakamoto et al., 2016). Similarly, our synthetic data sets were generated under the birth-death process, which is invariant to translation in time. The trees were conditioned on a given number of "extant" tips and shifted into the past by an arbitrary number of time units. In the case of Fossil BAMM analyses, this shift was purely formal: the method does not independently account for the ages of the root and the youngest known occurrence, but only for their difference ("observationTime"; Mitchell et al., 2018), which remained unaffected by the shift.

We suggest that rather than exhibiting fundamentally different macroevolutionary dynamics, extinct clades suffer from problems related to both data quantity and quality that render them ill-suited to analyses using currently available diversification rate estimation methods. The inferior performance of Fossil BAMM on subsampled phylogenies (Supplementary Information, Table C.10 available on Dryad) suggests the need for large trees that are rare in paleontology; indeed, we are aware of only one matrix (Hartman et al., 2019) in which the number of taxa exceeds the average number of sampled tips in our simulated trees ($n = 291$). Moreover, morphological character matrices are often designed to resolve relationships at specific taxonomic levels, with comprehensive sampling often limited to narrow taxonomic scales. In contrast, data sets covering broader groups (such as Ornithischia) tend to adopt a diversified sampling strategy (Höhna et al., 2011), in which each subclade is represented by a small sample of well-preserved species that may not reflect its overall diversity. Accordingly, the absence of shifts from our Fossil BAMM analyses is possibly explained by the fact that the character matrices used for this study included a wealth of early-diverging

taxa relevant to the higher-order ornithischian phylogeny but few or no representatives of Hadrosauriformes and Ceratopsidae, the two deeply nested clades previously associated with diversification rate shifts (Lloyd et al., 2008; Sakamoto et al., 2016). While the number of tips does not account for the ambiguous inferences obtained with PyRate, which does not use phylogenies as its input, data quality is also of concern. Compared to the rhinocerotid and marine mammal data sets previously used to explore the performance of PyRate (Silvestro et al., 2014b, 2019), our data included fewer occurrences (1240 vs. 2463 and 4740) spanning a longer period of time (135 myr vs. 45 and 66 myr), a lower average number of occurrences per taxonomic unit (3.12 vs. 15.02 and 8.86), and a much higher proportion of singletons (0.61 vs. 0.21 and 0.38). If the rich record of Cenozoic fossil mammals is the exception rather than the rule (Marshall, 2017), our results suggest that PyRate may not perform well in settings representative of a typical vertebrate fossil record.

3.5.4 Best Practices

We recommend that for extinct taxa, the use of Fossil BAMM be limited to the basic assessment of overall clade dynamics. Our simulations suggest that the method can reliably detect declining speciation rates and accurately infer the magnitude of the decline, but its focus on clade-specific events makes Fossil BAMM unable, at present, to distinguish between gradual and sharp transitions. This precludes inference of the onset and duration of any tree-wide decline, rendering the method unsuited for addressing many questions of paleobiological interest, including that of a possible pre-extinction drop in ornithischian diversification investigated here. While the ability of BAMM to infer clade-specific shifts allows valuable insights into the macroevolutionary dynamics of extant taxa (Friedman et al., 2019; Henao Diaz et al., 2019), the species-poor phylogenetic data sets available for extinct clades are unlikely to afford BAMM sufficient power to detect lineage-specific events. This difficulty may be further compounded by diversified sampling (Höhna et al., 2011), which obscures

the differences between species-rich and species-poor clades and remains a common feature of paleontological data sets (Simões et al., 2020). If Fossil BAMM analyses are conducted on such data sets, the lack of rate shifts should not be interpreted as positive evidence for uniform diversification.

The epoch model of diversification implemented in PyRate is more appropriate when coordinated shifts in speciation or extinction rates are expected to have occurred across multiple lineages, as may be the case during mass extinctions and rapid radiations typical of post-extinction recovery. Our results suggest that real shifts can be accurately inferred (Fig. 3.6), but the converse is not true: inferred shifts may not always be real. We thus suggest that PyRate analyses be coupled with (1) a careful re-examination of the original occurrence data to determine whether the inferred shifts coincide with conspicuous clusters of first or last appearance dates (cf. Supplementary Information, Fig. C.1 available on Dryad), and (2) independent assessments of the rate of preservation to gain insight into the extent to which such clusters reflect preservational rather than macroevolutionary dynamics. Such assessments may include simple geological proxies (e.g., number of rock units or localities per time bin; Butler et al., 2012; Holland, 2016), or estimates from more traditional paleobiological methods such as the capture-mark-recapture (Connolly and Miller, 2001) or three-timer (Alroy, 2008) techniques. We further suggest exploring (3) the influence of the model-averaging algorithm (bdMCMC vs. rjMCMC) and (4) the effect of the Poisson prior (or the corresponding hyperprior in the case of rjMCMC) on the number of shifts. The issue of the potential sensitivity of the inferred number of shifts to the corresponding prior has received attention in the context of BAMM analyses of extant data (Moore et al., 2016; Rabosky et al., 2017), and its importance may be even higher for the low-power data sets typically available for fossil vertebrates.

The ability of Fossil BAMM and PyRate to provide checks on each other's results is limited by their focus on fundamentally different types of rate shifts (clade-specific vs. tree-

wide, respectively). However, we still suggest that both methods may be profitably employed together. This is particularly true when marginal rates are of interest, as BAMM estimates speciation rates with moderate to high accuracy (Fig. 3.5; Supplementary Information, Table C.10 available on Dryad), while PyRate performs better at estimating the rate of net diversification (Supplementary Information, Fig. C.27 available on Dryad), suggesting possible complementarity of the two approaches. The preservation rates estimated by Fossil BAMM and PyRate also allow useful comparisons, and evidence for preservational heterogeneity from the latter method, such as the preference for a TPP model and non-overlapping credibility intervals for interval-specific rates, can reveal cases where the time-constant preservation model employed by BAMM may be misspecified, potentially causing erroneous inferences about diversification. Additionally, while Fossil BAMM should not be expected to replicate the tree-wide shifts inferred by PyRate, it can capture the dynamics of a shift (or a monotonic series of shifts) in the form of non- overlapping credibility intervals for the rates at the root and at the tips, providing for a limited but important range of diversification scenarios that can be tested by assessing congruence between the two methods.

3.5.5 Future Directions

The single empirical test case and simple simulation scenarios explored in this study do not exhaust the range of conditions likely to be encountered in diversification rate analyses of extinct clades. Our synthetic data were generated under a bifurcating model of speciation, following the complete absence of ancestor-descendant pairs from our ornithischian time trees. However, budding and anagenetic speciation are common features of the fossil record (Raup, 1985; Foote, 1996; Hopkins, 2011; Parins-Fukuchi, 2021), and can be simulated using the same tools we employed here (Barido-Sottani et al., 2019b). Exploring the sensitivity of diversification rate estimates to the mode of speciation can thus represent a fruitful subject for future research. Another problem left unaddressed by the current study is the amount of

data needed to distinguish discrete tree-wide shifts from prolonged rate changes. While our simulations show that Fossil BAMM tends to approximate tree-wide speciation rate shifts by fitting a single exponentially varying rate rather than multiple clade-specific shifts, it may be worthwhile to fit both models and record their likelihood difference as a function of tree size and rate change magnitude, so as to estimate the amount of evidence required for Fossil BAMM to detect simultaneous shifts along multiple branches. Valuable insights may also be obtained from more complex simulations in which clade-specific shifts are clustered in a short period of time rather than perfectly simultaneous, or in which the shift magnitude varies across branches in a probabilistic fashion. Besides representing a more realistic description of adaptive radiation-like phenomena, such simulations would have the added advantage of being able to tease apart the relative contributions of low power and model violation to the Fossil BAMM performance issues reported here.

The model-based framework adopted by Fossil BAMM and PyRate lends itself easily to empirically motivated extensions, holding out promise that continued model development may mitigate the issues reported here and improve the applicability of both methods to entirely extinct taxa. We suggest that high priority should be given to the development of approaches that accommodate both tree-wide and clade-specific shifts in a single framework (cf. Laurent et al., 2015), possibly including mass extinctions as a third and separate class of events (Höhna, 2015; May et al., 2016). It may be argued that true tree-wide shifts should virtually never be expected, since events acting across the tree, like mass extinctions and subsequent recoveries, should not change the diversification rates of all lineages to exactly the same extent; stochastic variation should be expected even in the absence of life history differences or other factors that render some lineages more prone to speciation or extinction than others. However, we argue that while they may not perfectly capture the underlying diversification dynamics, discrete tree-wide shifts represent a useful compromise between precision (which compares favorably to smearing a shift across the entire duration of a rate

regime) and the generally low power to detect lineage-specific shifts subtending tip-poor clades (Rabosky et al., 2017)—a point borne out by the multiple methods already dedicated to inferring such events (Rabosky, 2006; Morlon et al., 2011; Stadler, 2011a; May et al., 2016). A further argument could be made for treating mass extinctions as fundamentally different from rate shifts of any kind; we refer readers to May et al. (2016) for a defense of this approach.

When analyzing extinct taxa, fossil preservation has to be modeled with the same care as lineage diversification (Foote, 2003, 2007; Liow and Finarelli, 2014). Currently, PyRate holds an advantage over Fossil BAMM in allowing greater flexibility in its preservation models, the most general of which can account for variation both through time and among lineages (Silvestro et al., 2019). Fossil BAMM may benefit from adopting similarly flexible models, although it is notable that our simulations did not find its performance to be seriously compromised by the assumption of rate constancy (Fig. 3.5), and that the preservation rate inferred by BAMM for the ornithischians was closer to an independently derived estimate than the PyRate results based on a more complex model (Supplementary Information, Preservation Rates available on Dryad). The latter finding points to the importance of balancing the realism of a model against the ability to estimate all of its parameters with sufficient accuracy, and indicates that model selection tools may be as important as the models themselves. PyRate currently relies on the relatively liberal corrected Akaike information criterion to avoid the high computational costs associated with full BF comparisons (Silvestro et al., 2019), underscoring the importance of developing more efficient ways of estimating marginal likelihoods (Fourment et al., 2019). While the risk of overparameterization remains salient, occurrence-rich data sets may benefit from preservation models even more complex than those currently implemented in PyRate. Such models could account for geographic and environmental rate variation (Wagner and Marcot, 2013; Holland, 2016) or use empirically informed descriptions of how preservation potential changes over the lifespan of a lineage

(Marshall, 2019). Alternatively, fossil preservation could be treated in the same manner as diversification, allowing for both temporal (tree-wide) rate shifts and lineage-specific events, such as the colonization of environments conferring increased preservation potential.

We encourage researchers to take advantage of the parametric framework common to the methods evaluated here and continue the development of models that better account for the idiosyncrasies of fossil data, as well as the exploration of their performance under paleobiologically realistic conditions. In light of the issues highlighted here, it is imperative that these steps be taken before the widespread application of existing methods to extinct clades. It is equally important to recognize the limitations inherent to most paleontological phylogenetic data sets, which were typically not assembled with diversification rate estimation in mind, and to interrogate their suitability to address the ques- tion under examination. Combined with the flexibility of Bayesian model-averaging approaches, such restricted but informed use of fossil data can help shed light on macroevolutionary dynamics throughout the tree of life (Silvestro et al., 2018; Louca and Pennell, 2020; Lloyd and Slater, 2021).

3.6 Supplementary Material

Data available from the Dryad Digital Repository: `http://dx.doi.org/10.5061/dryad.sbcc2fr4x`.

3.7 Acknowledgments

We thank Matt Friedman, Peter J. Wagner, and three anonymous reviewers for helpful comments and constructive criticism. We benefited from PBDB data compiled by numerous enterers and authorizers, especially M.T. Carrano, P.D. Mannion, A.K. Behrensmeyer, G.T. Lloyd, and J. Alroy. D.Č. further thanks the instructors and participants of the Taming the BEAST, eh! Workshop (11–16 August 2019, Squamish, British Columbia), Jonathan

Chang, and Michael Foote for valuable advice, and April Wright for stimulating discussions. Michael Foote and Rossy Natale provided helpful comments on an early draft of the article. All simulations were run on the Midway2 Research Computing Cluster at the University of Chicago. This is Paleobiology Database publication number 401.

3.8 Funding

This work was supported by the National Science Centre, Poland (2020/37/B/NZ8/01321 to D.M.).

3.9 Competing Interests

There are no conflicts of interest to disclose.

CONCLUSIONS

Model-based phylogenetics and diversification rate estimation represent an important framework for integrating fossil data into statistically rigorous inferences about macroevolution. The research described in this dissertation highlights the limitations of existing approaches (Chapter 3), and shows that the common tendency to exclusively report point estimates can often obscure the true degree of uncertainty associated with them (Chapter 2). One of the key issues facing current attempts to estimate diversification rates in extinct clades using explicit statistical models of the branching process is the lack of sufficiently large and densely sampled time trees, a problem that stems from difficulties inherent to the collection of morphological data from which paleontological phylogenies are inferred. However, supertree techniques represent a promising a way of circumventing this problem, as long as they can be made statistically rigorous and satisfy the requirements of modern comparative methods. To this purpose, I have developed and implemented a new supertree approach, and used both synthetic and empirical data to demonstrate its good performance under a wide range of conditions (Chapter 1). Although further computational efficiencies will be necessary for this approach to become a viable alternative to the currently used backbone-and-patch techniques, its emphasis on accommodating uncertainty as well as simultaneous inference of topology and divergence times allows it to address important weaknesses of current fossil-based phylogenetics.

APPENDIX A: SUPPLEMENTARY INFORMATION FOR CHAPTER ONE

A.10 Simulation Tests

A.10.1 Simulation-based calibration checking

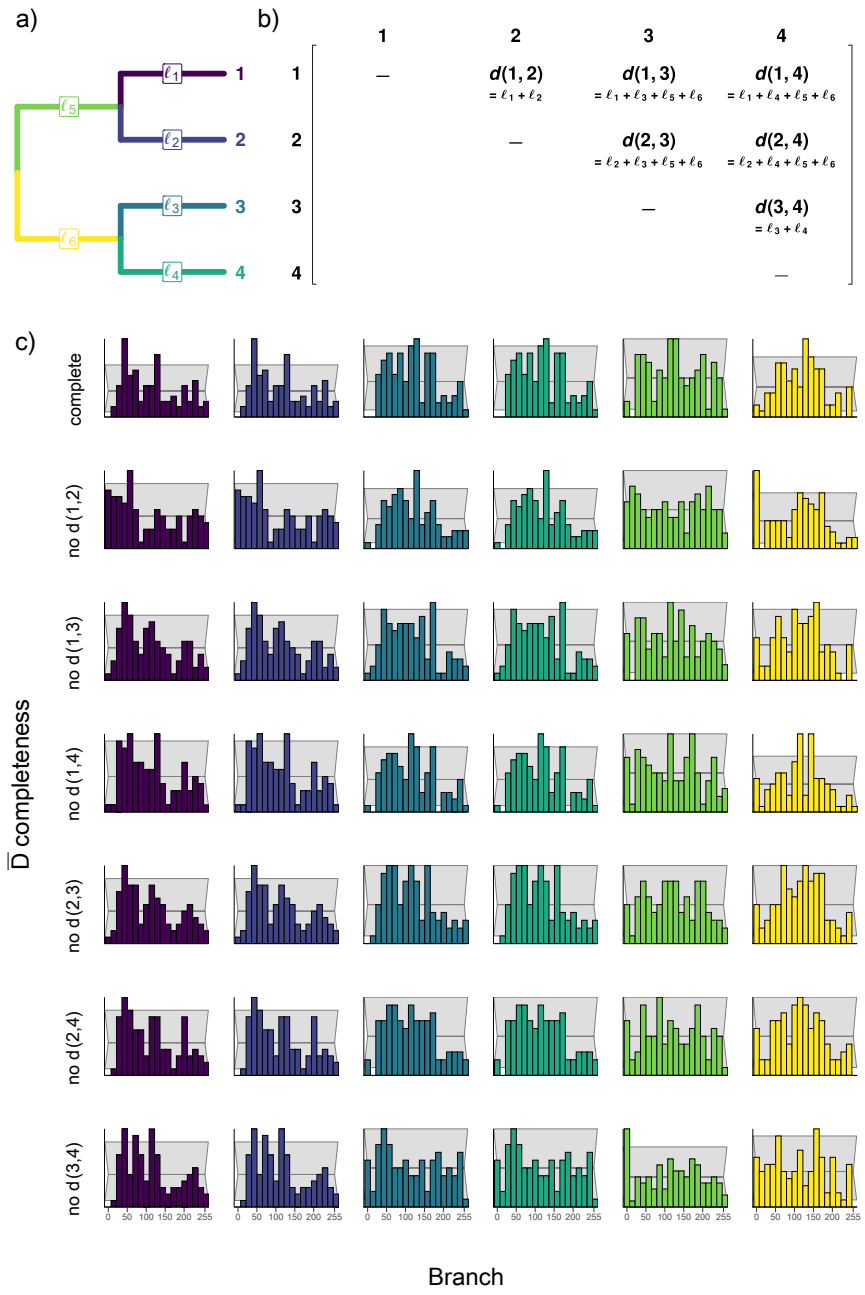

Figure A.7: The 4-tip balanced tree with labeled branches (a) and its distance matrix, with observed distances expressed as sums of the underlying branch durations (b). Panel (c) shows the rank histograms with 95% confidence intervals (indicated by the light gray wedged rectangles in the background) for individual branch durations and different average distance matrix ($\bar{\mathbf{D}}$) completeness scenarios. See Figure 4 in the main text for the corresponding ECDF difference plots.

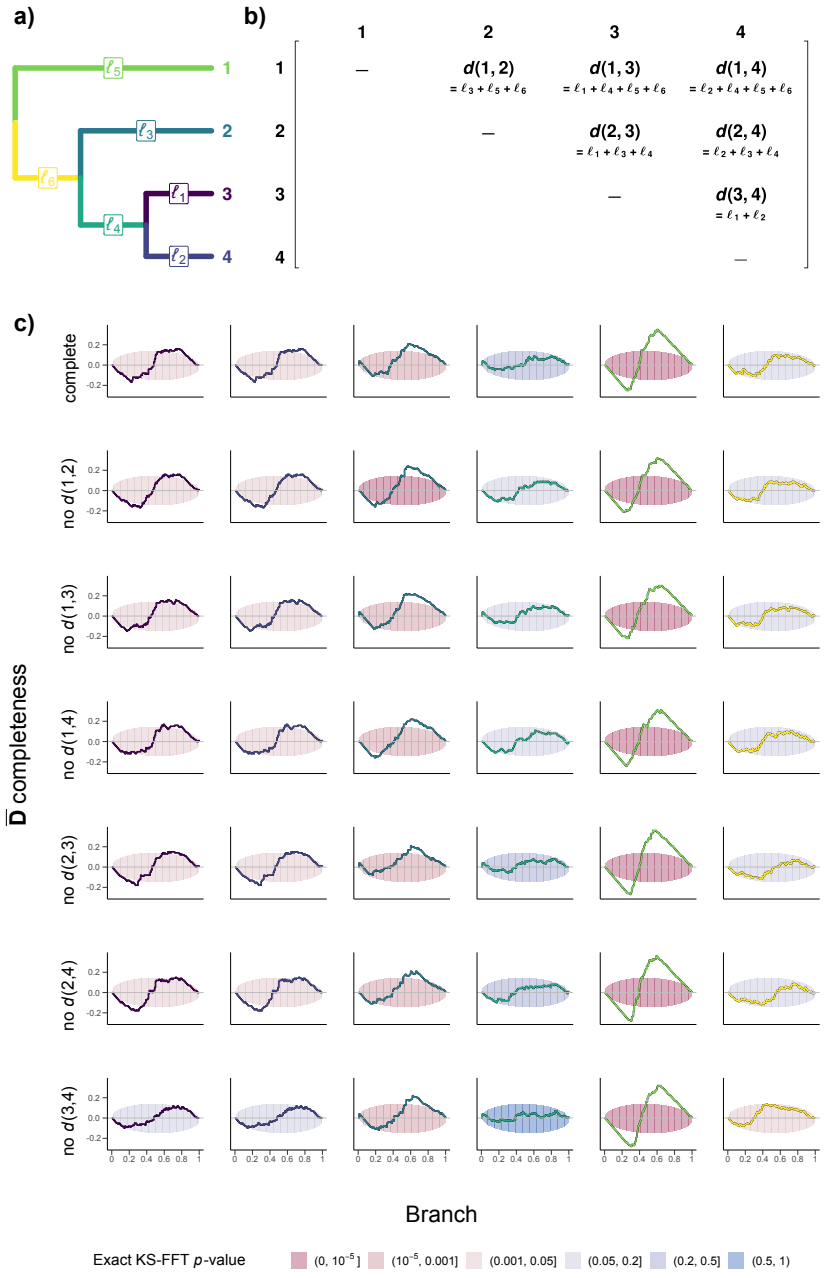

Figure A.8: The 4-tip pectinate tree with labeled branches (a) and its distance matrix, with observed distances expressed as sums of the underlying branch durations (b). Panel (c) shows the difference between the empirical cumulative distribution function (ECDF) and the discrete uniform expectation (stepwise identity function) with 95% simultaneous confidence bands for individual branch durations and different average distance matrix ($\bar{\mathbf{D}}$) completeness scenarios. The confidence bands are colored according to the p-values from uniformity tests conducted using the exact Kolmogorov-Smirnov fast Fourier transform (Exact KS-FFT) method.

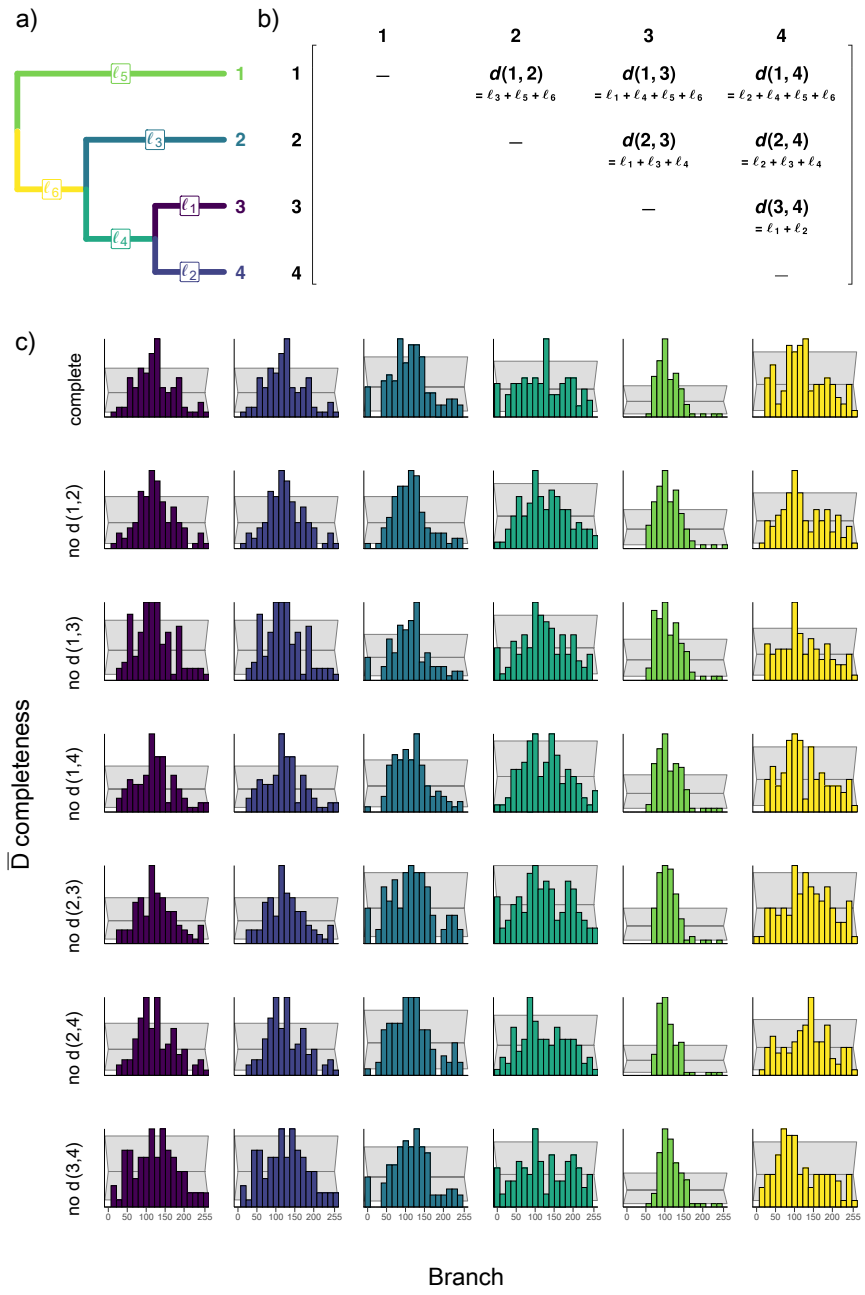

Figure A.9: Results of simulation-based calibration checking for the 4-tip pectinate tree, visualized as rank histograms. Panel descriptions as in Figure A.7.

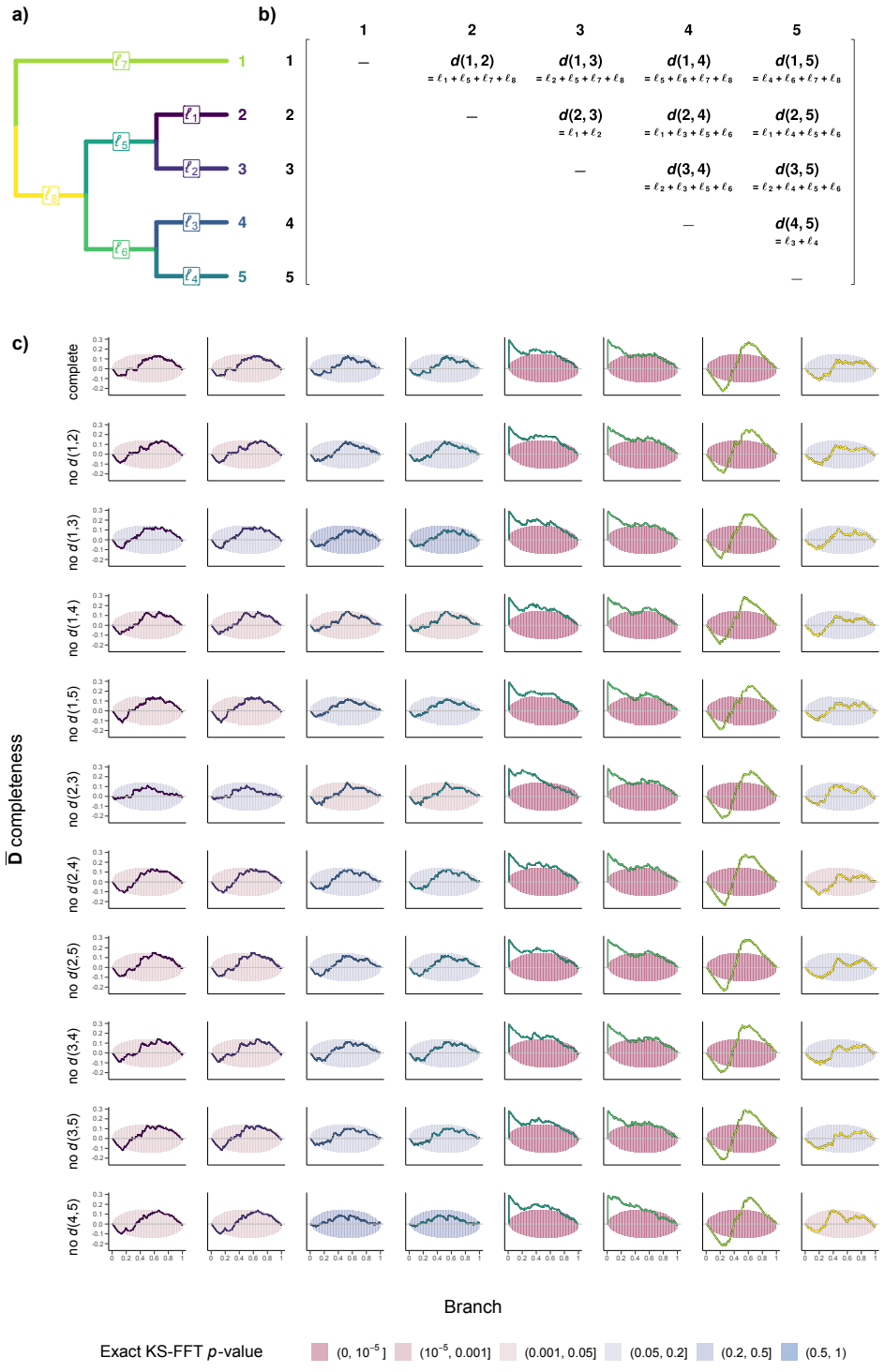

Figure A.10: Results of simulation-based calibration checking for the 5-tip tree with a single tip sister to a balanced subtree, visualized as ECDF difference plots. Panel descriptions as in Figure A.8.

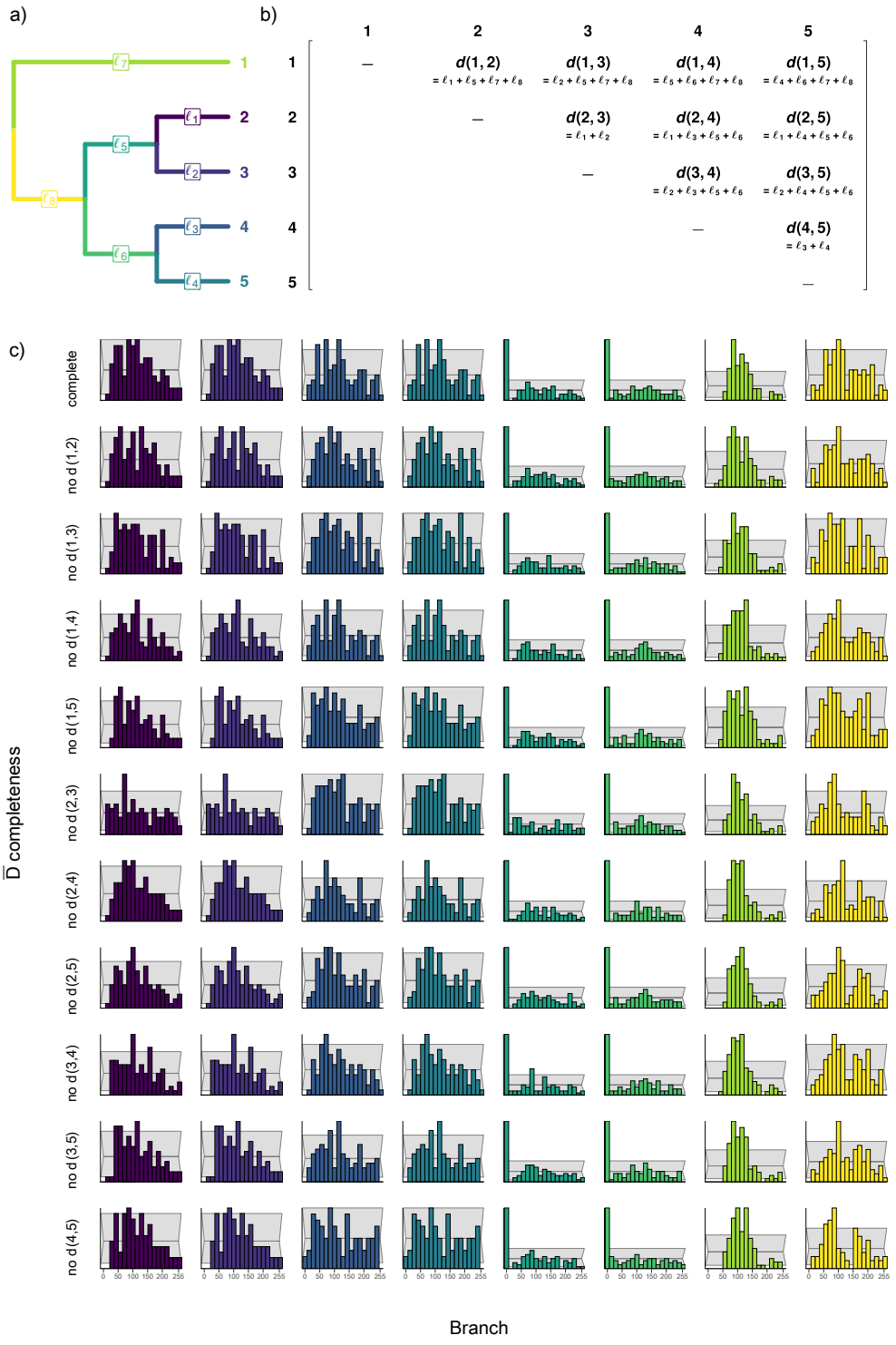

Figure A.11: Results of simulation-based calibration checking for the 5-tip tree with a single tip sister to a balanced subtree, visualized as rank histograms. Panel descriptions as in Figure A.7.

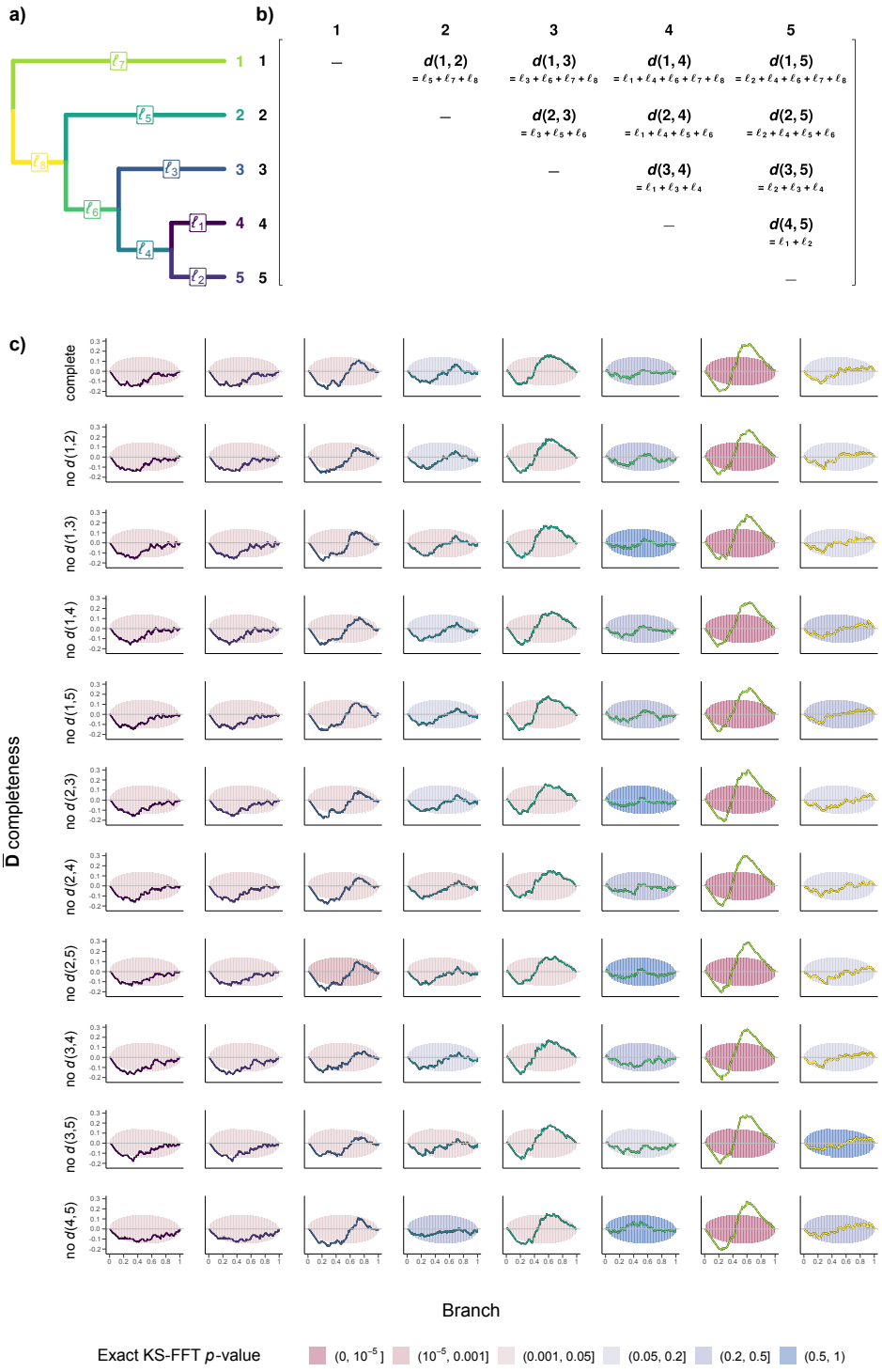

Figure A.12: Results of simulation-based calibration checking for the fully pectinate 5-tip tree, visualized as ECDF difference plots. Panel descriptions as in Figure A.8.

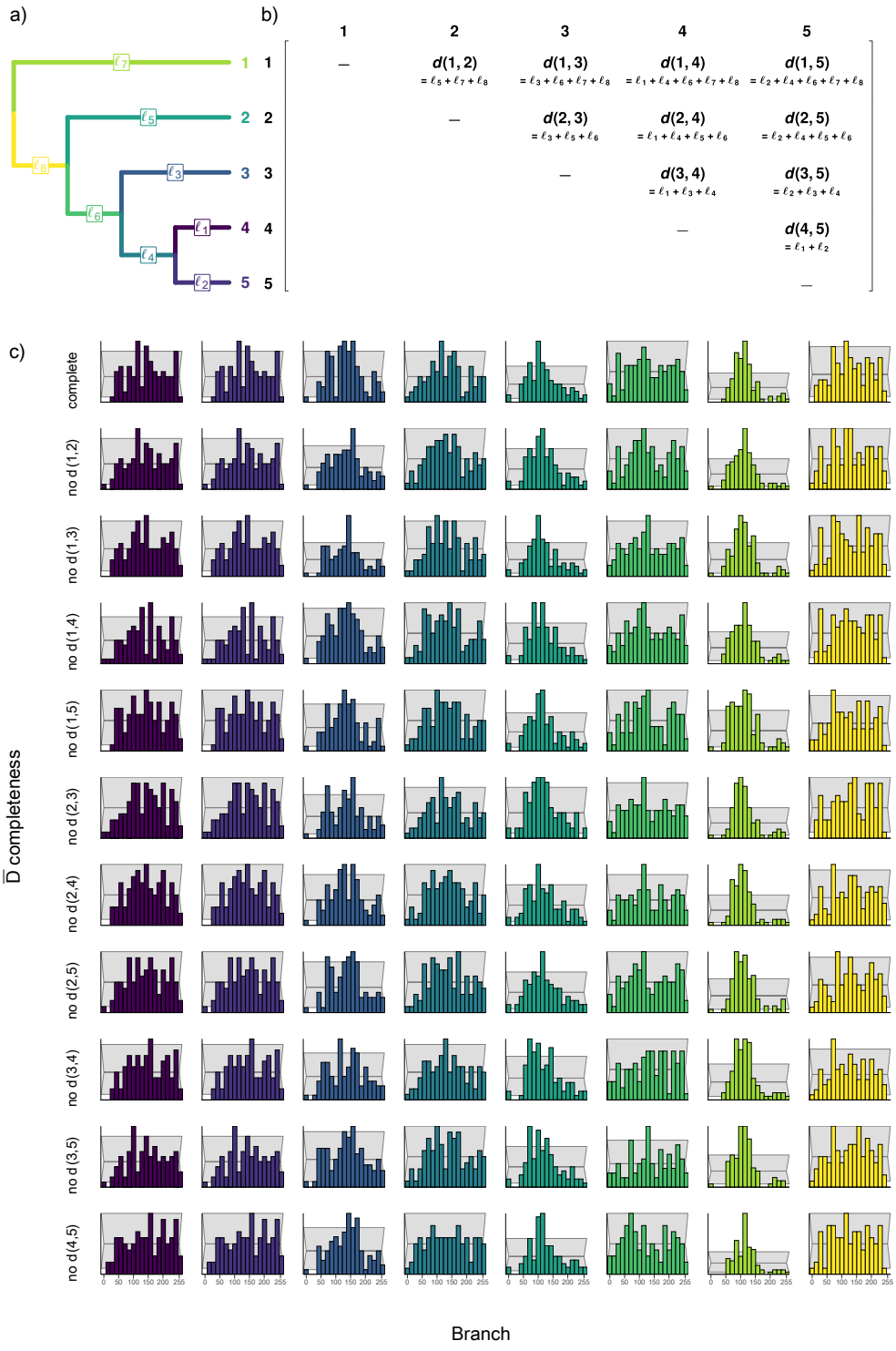

Figure A.13: Results of simulation-based calibration checking for the fully pectinate 5-tip tree, visualized as rank histograms. Panel descriptions as in Figure A.7.

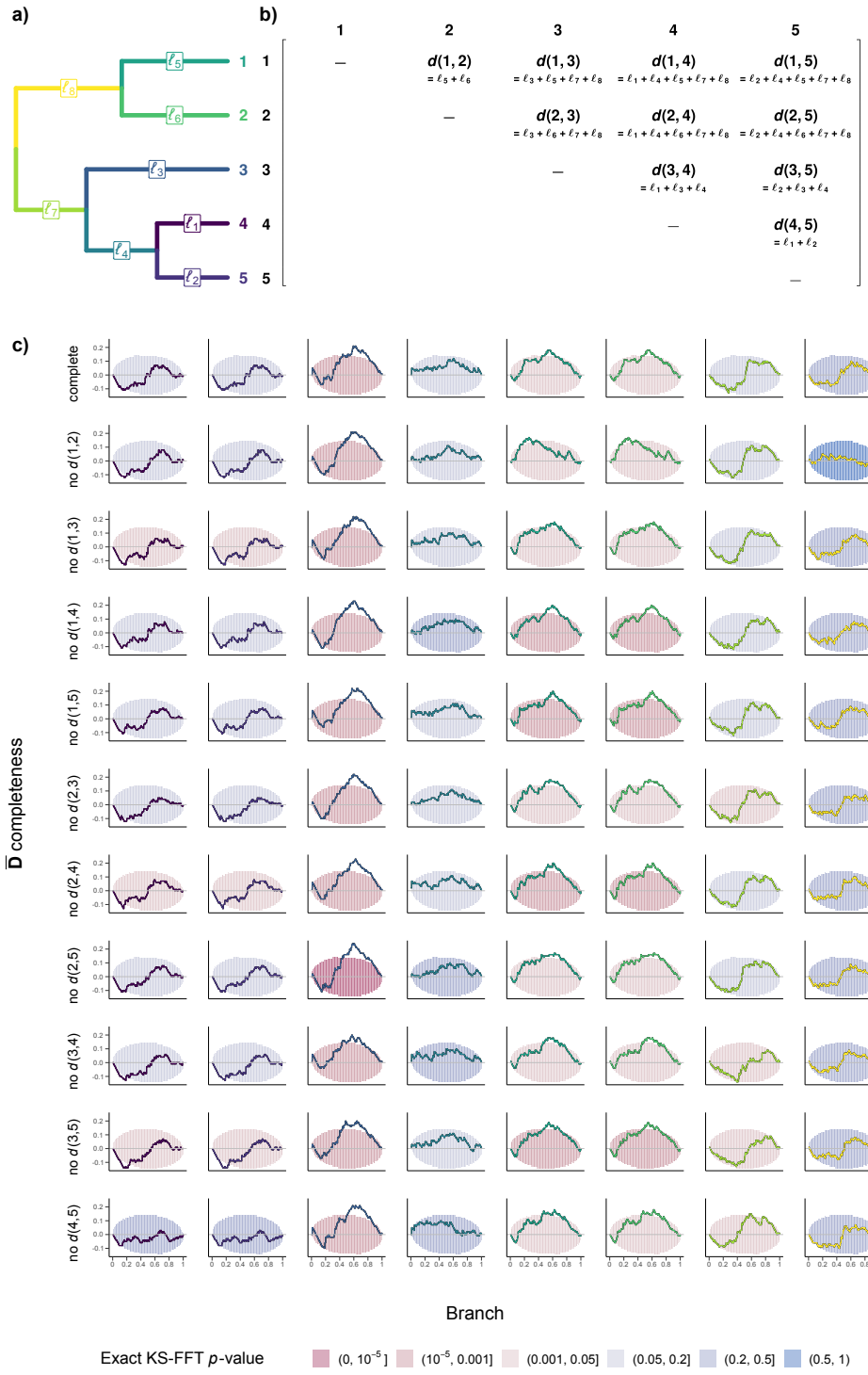

Figure A.14: Results of simulation-based calibration checking for the 5-tip tree with a 2-tip clade sister to an imbalanced subtree, visualized as ECDF difference plots. Panel descriptions as in Figure A.8.

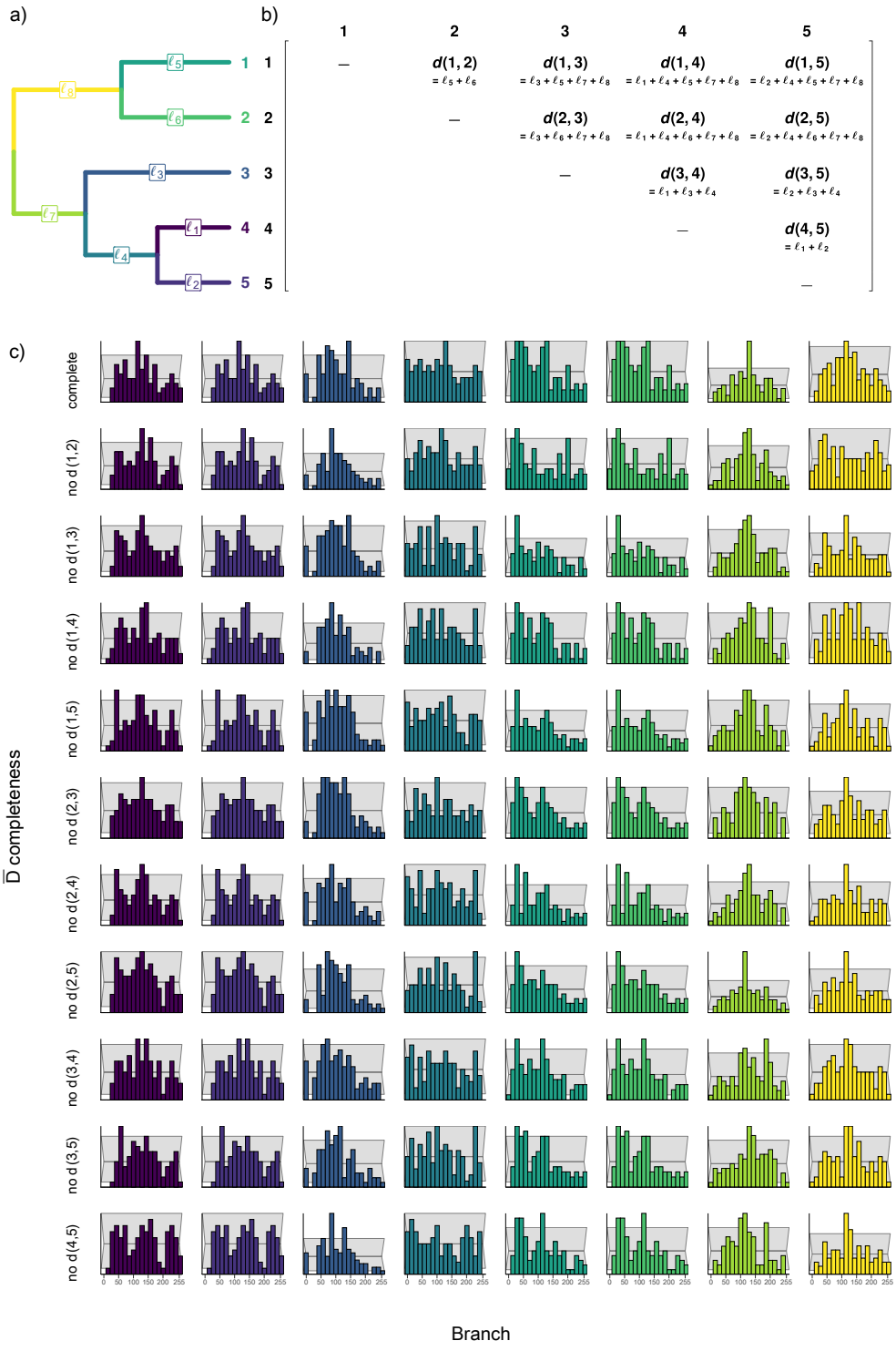

Figure A.15: Results of simulation-based calibration checking for the 5-tip tree with a 2-tip clade sister to an imbalanced subtree, visualized as rank histograms. Panel descriptions as in Figure A.7.

A.10.2 200-tip simulations

Table A.2: Properties of the 200-tip simulated trees and of the average distance matrices ($\bar{\mathbf{D}}$) obtained by breaking them up into subtrees under the backbone-and-patch (BP) and random selection (RS) tip sampling schemes. Root height is given in arbitrary units.

Tree	Root height	Colless index	Subtree number (BP)	$\bar{\mathbf{D}}$ completeness (BP)	Subtree number (RS)	$\bar{\mathbf{D}}$ completeness (RS)
1	191.66	0.0324	4	0.3577	18	0.5172
2	146.87	0.0521	9	0.3972	28	0.6863
3	11.68	0.0306	9	0.1958	24	0.6992
4	67.68	0.0586	4	0.5046	28	0.6763
5	193.82	0.0453	8	0.375	36	0.7446
6	128.73	0.0333	8	0.1614	32	0.7561
7	47.7	0.0404	27	0.297	23	0.6455
8	10.06	0.0283	32	0.1119	32	0.7732
9	95.48	0.0548	7	0.38	31	0.6664
10	157.06	0.0423	12	0.1402	25	0.6581
11	104.62	0.0337	4	0.3703	29	0.6789
12	64.79	0.045	8	0.2582	27	0.7264
13	131.73	0.041	10	0.1516	33	0.8163
14	180.26	0.0389	4	0.3666	26	0.6462
15	5.94	0.0385	9	0.1666	21	0.6196
16	59.27	0.0475	15	0.0969	30	0.703
17	43.63	0.0457	18	0.1009	28	0.6227
18	105.94	0.0325	7	0.2483	30	0.6952
19	113.23	0.0343	4	0.4293	29	0.7608
20	56	0.0373	12	0.1674	28	0.7493
21	157.67	0.0358	8	0.2257	17	0.5768
22	133.89	0.0338	4	0.3354	26	0.6963
23	172.42	0.0393	4	0.4478	35	0.7478
24	187.71	0.0358	4	0.4626	23	0.6354
25	126.88	0.0545	4	0.6037	25	0.651
26	195.83	0.0663	4	0.59	19	0.6432
27	197.24	0.0397	8	0.261	23	0.5939
28	64.9	0.0456	10	0.2152	49	0.8902
29	2.1	0.0338	4	0.3748	24	0.6284
30	65	0.0588	11	0.1437	24	0.6086
31	166.48	0.0399	4	0.4546	16	0.6012

Table A.2 – *Continued from previous page*

Tree	Root height	Colless index	Subtree number (BP)	$\bar{\mathbf{D}}$ completeness (BP)	Subtree number (RS)	$\bar{\mathbf{D}}$ completeness (RS)
32	98.35	0.0378	4	0.7575	27	0.6368
33	86	0.0436	7	0.2862	25	0.6538
34	197.84	0.0352	4	0.3558	17	0.6776
35	189.31	0.0504	7	0.2372	27	0.6922
36	84.75	0.0585	8	0.3275	32	0.7789
37	142.13	0.0357	4	0.5095	31	0.7152
38	196.95	0.0442	5	0.3114	24	0.6911
39	58.75	0.0638	7	0.3404	23	0.5986
40	50.14	0.0442	10	0.1794	35	0.8174
41	190.81	0.0325	4	0.3439	25	0.6463
42	27.57	0.0385	7	0.2193	20	0.5936
43	44.03	0.046	5	0.476	29	0.7102
44	181.96	0.0508	4	0.4982	16	0.5988
45	45.42	0.0378	8	0.195	22	0.6847
46	47.41	0.0593	8	0.2322	22	0.5376
47	178.36	0.0375	6	0.2498	26	0.7404
48	154.12	0.0371	4	0.4703	23	0.5978
49	134.95	0.0497	20	0.0877	28	0.7398
50	22.96	0.0269	4	0.3546	21	0.6536
51	29.12	0.0406	11	0.1568	25	0.6989
52	54.08	0.0456	8	0.2628	29	0.7327
53	161.31	0.0387	9	0.1953	25	0.7458
54	153.85	0.0523	4	0.7781	40	0.8354
55	50.26	0.0441	4	0.4357	38	0.847
56	129.34	0.0411	31	0.0908	38	0.8336
57	69.06	0.0387	6	0.391	28	0.6944
58	86.19	0.0333	13	0.1057	20	0.6369
59	161.55	0.0467	4	0.4137	29	0.7272
60	105.57	0.0421	7	0.2388	20	0.642
61	128.38	0.0479	5	0.4062	25	0.7028
62	88.76	0.0518	10	0.1624	22	0.607
63	178.6	0.0367	6	0.266	20	0.6359
64	127.09	0.0397	15	0.0989	34	0.8026
65	62.6	0.0354	10	0.1388	25	0.6769
66	71.45	0.0505	12	0.1719	19	0.5826

Table A.2 – *Continued from previous page*

Tree	Root height	Colless index	Subtree number (BP)	$\bar{\mathbf{D}}$ completeness (BP)	Subtree number (RS)	$\bar{\mathbf{D}}$ completeness (RS)
67	124.11	0.0331	5	0.2652	39	0.8078
68	58.03	0.0385	10	0.1706	23	0.5929
69	83.93	0.0395	6	0.2462	29	0.7422
70	131.84	0.0568	7	0.3083	25	0.7079
71	188.87	0.0475	9	0.1832	38	0.7819
72	187.32	0.0374	7	0.2492	27	0.6246
73	159.59	0.0384	7	0.2052	25	0.7172
74	25.25	0.0407	7	0.2021	27	0.6826
75	69.58	0.0447	4	0.6073	24	0.6563
76	83.82	0.0497	5	0.4503	29	0.6496
77	67.01	0.0472	4	0.4024	27	0.7412
78	153.93	0.0381	7	0.241	18	0.5868
79	43.29	0.0349	9	0.1941	34	0.7364
80	29.3	0.0622	4	0.6328	22	0.6323
81	119.91	0.0548	10	0.1846	22	0.64
82	45.55	0.038	6	0.3384	19	0.592
83	111.18	0.0388	12	0.1696	27	0.741
84	185.11	0.0354	8	0.1871	22	0.6239
85	109.34	0.0522	4	0.3822	25	0.6541
86	105.81	0.0421	8	0.1864	25	0.74
87	162.96	0.0608	5	0.4956	36	0.7817
88	162.29	0.032	8	0.2058	26	0.6335
89	185.93	0.035	6	0.2757	28	0.6888
90	14.24	0.0363	4	0.4427	29	0.6824
91	135.74	0.0356	5	0.339	34	0.7338
92	172.67	0.0517	4	0.5664	18	0.5885
93	176.52	0.0368	5	0.3308	22	0.5901
94	174.7	0.0346	7	0.2299	26	0.6276
95	105.21	0.0481	4	0.412	26	0.6704
96	125.04	0.0388	4	0.4562	42	0.8293
97	53.68	0.0388	5	0.4122	24	0.655
98	30.97	0.0486	7	0.3959	27	0.6348
99	140.2	0.0356	4	0.4786	27	0.7048
100	32.01	0.0465	25	0.0692	29	0.7518

Table A.3: Topological and divergence time accuracy of the maximum *a posteriori* (MAP) and maximum clade credibility (MCC) summary supertrees inferred under $\lambda_e = 0.1$. For each of the n replicates that reached convergence, the posterior summary was compared to the corresponding true tree in terms of Robinson-Foulds (RF) distance and normalized root age difference (Δ root age $= |a_e - a_t| \,/\, a_t$, where a_t is the true root age and a_e is the posterior mean). Every statistic is reported as the mean $\pm$ standard deviation over all n replicates.

		MAP tree		**MCC tree**	
Scenario	n	**RF dist.**	**Δ root age**	**RF dist.**	**Δ root age**
no error, backbone and patch	21	0.4643 ± 0.1798	0.1970 ± 0.1314	0.4462 ± 0.1856	0.1970 ± 0.1314
no error, random selection	87	0.2166 ± 0.1181	0.0004 ± 0.0032	0.1718 ± 0.1216	0.0004 ± 0.0032
no error, backbone and patch, imputed	98	0.1141 ± 0.1066	0.0003 ± 0.0025	0.0709 ± 0.1054	0.0003 ± 0.0025
no error, random selection, imputed	98	0.2454 ± 0.0844	0 ± 0.0001	0.2188 ± 0.0819	0 ± 0.0001
top error, backbone and patch	31	0.5618 ± 0.1410	1.1117 ± 3.0300	0.5359 ± 0.1440	1.1117 ± 3.0300
top error, random selection	43	0.2624 ± 0.0795	0.5964 ± 1.4833	0.2347 ± 0.0911	0.5964 ± 1.4833
top error, backbone and patch, imputed	45	0.7533 ± 0.0641	0.6819 ± 1.0366	0.7489 ± 0.0623	0.6819 ± 1.0366
top error, random selection, imputed	97	0.9651 ± 0.0159	0.6516 ± 1.6782	0.9645 ± 0.0159	0.6516 ± 1.6782

Table A.4: Topological and divergence time accuracy of the MAP and MCC summary supertrees inferred under $\lambda_e = 0.33$. Abbreviations as in Table A.3.

		MAP tree		**MCC tree**	
Scenario	n	**RF dist.**	**Δ root age**	**RF dist.**	**Δ root age**
no error, backbone and patch	13	0.4205 ± 0.2046	0.1508 ± 0.1434	0.4034 ± 0.2046	0.1508 ± 0.1434
no error, random selection	67	0.1590 ± 0.0664	0 ± 0	0.1278 ± 0.0573	0 ± 0
no error, backbone and patch, imputed	95	0.0690 ± 0.0741	0 ± 0.0001	0.0349 ± 0.0685	0 ± 0.0001
no error, random selection, imputed	92	0.2240 ± 0.0914	0 ± 0.0001	0.1993 ± 0.0939	0 ± 0.0001
top error, backbone and patch	29	0.5351 ± 0.1425	1.0427 ± 2.9626	0.5137 ± 0.1417	1.0427 ± 2.9626
top error, random selection	42	0.2499 ± 0.1248	1.2046 ± 3.3431	0.2350 ± 0.1297	1.2046 ± 3.3431
top error, backbone and patch, imputed	41	0.7558 ± 0.0526	0.7407 ± 1.0791	0.7555 ± 0.0525	0.7407 ± 1.0791
top error, random selection, imputed	97	0.9655 ± 0.0162	0.6484 ± 1.6782	0.9654 ± 0.0159	0.6484 ± 1.6782

Table A.5: Topological and divergence time accuracy of the MAP and MCC summary supertrees inferred under $\lambda_e = 1$. Abbreviations as in Table A.3.

		MAP tree		**MCC tree**	
Scenario	n	**RF dist.**	**Δ root age**	**RF dist.**	**Δ root age**
no error, backbone and patch	7	0.3561 ± 0.1366	0.0611 ± 0.0760	0.3387 ± 0.1450	0.0611 ± 0.0760
no error, random selection	75	0.1340 ± 0.0779	0 ± 0	0.1076 ± 0.0752	0 ± 0
no error, backbone and patch, imputed	86	0.0301 ± 0.0278	0 ± 0	0.0116 ± 0.0169	0 ± 0
no error, random selection, imputed	77	0.1993 ± 0.0844	0 ± 0	0.1775 ± 0.0803	0 ± 0
top error, backbone and patch	26	0.5203 ± 0.1348	1.2330 ± 3.1857	0.5027 ± 0.1312	1.2330 ± 3.1857
top error, random selection	28	0.2808 ± 0.1096	1.1226 ± 1.9651	0.2745 ± 0.1098	1.1226 ± 1.9651
top error, backbone and patch, imputed	40	0.7490 ± 0.0547	0.7205 ± 1.0890	0.7490 ± 0.0547	0.7205 ± 1.0890
top error, random selection, imputed	97	0.9651 ± 0.0161	0.6511 ± 1.6788	0.9653 ± 0.0160	0.6511 ± 1.6788

Table A.6: Accuracy of the MCC summary supertrees inferred under $\lambda_e = 0.1$. For each of the n replicates that reached convergence, the posterior summary was compared to the corresponding true tree in terms of Kuhner-Felsenstein (KF), Billera-Holmes-Vogtmann (BHV), and matching split distances. Every statistic is reported as the mean and 95% interpercentile range over all n replicates.

Scenario	n	**KF distance**	**BHV distance**	**Matching split dist.**
no error, backbone and patch	21	175.7 [10.25, 276.29]	222.14 [12.22, 356.5]	978.38 [263.5, 1839.5]
no error, random selection	87	74.6 [21.01, 163.07]	83.91 [23.2, 182.42]	113.06 [35.45, 320.5]
no error, backbone and patch, imputed	98	7.06 [5.57, 11.65]	7.61 [5.92, 12.99]	62 [10, 213.35]
no error, random selection, imputed	98	165.34 [35.28, 387.57]	178.4 [37.79, 440.19]	154.92 [53.27, 312.2]
top error, backbone and patch	31	240.16 [136.06, 344.01]	296 [147.11, 448.14]	1180.87 [477.5, 1806]
top error, random selection	43	156.69 [65.56, 251.32]	168.22 [71.35, 280.02]	247.53 [90.6, 640.5]
top error, backbone and patch, imputed	45	381.99 [193.69, 508.57]	460.11 [210.62, 654.69]	869.31 [669.4, 1158.1]
top error, random selection, imputed	97	448.5 [274.35, 592.01]	531.88 [286.96, 727.7]	1394.81 [1028.2, 1961.8]

Table A.7: Accuracy of the MCC summary supertrees inferred under $\lambda_e = 0.33$. Abbreviations as in Table A.6.

Scenario	n	**KF distance**	**BHV distance**	**Matching split dist.**
no error, backbone and patch	13	153.14, [25.92, 257.11]	191.57, [32.28, 329.21]	880.23, [108, 1549.5]
no error, random selection	67	72.91, [19.9, 157.58]	80.62, [21.96, 176.86]	73.57, [19, 173.1]
no error, backbone and patch, imputed	95	3.01, [2.24, 4.35]	3.18, [2.36, 4.76]	27.92, [0.7, 104.3]
no error, random selection, imputed	92	165.09, [25.06, 391.58]	178.8, [27, 432.65]	141.77, [40.83, 299.25]
top error, backbone and patch	29	236.24, [138.59, 346.89]	291.95, [148.16, 441.64]	1158.76, [579, 1821.8]
top error, random selection	42	156.53, [65.42, 246.92]	165.38, [70.9, 267.03]	277.12, [67.88, 820.73]
top error, backbone and patch, imputed	41	377.32, [191.47, 519.57]	454.93, [208.67, 661.57]	870.39, [685, 1078]
top error, random selection, imputed	97	449.84, [265.72, 612.32]	533.64, [278.85, 738.85]	1414.72, [1030.2, 2068]

Table A.8: Accuracy of the MCC summary supertrees inferred under $\lambda_e = 1$. Abbreviations as in Table A.6.

Scenario	n	**KF distance**	**BHV distance**	**Matching split dist.**
no error, backbone and patch	7	121.72, [19.87, 235.45]	156.84, [24.98, 301.41]	790.29, [217.85, 1353.2]
no error, random selection	75	63.94, [16.04, 135.94]	71.1, [17.43, 150.62]	58.97, [11.95, 210.25]
no error, backbone and patch, imputed	86	1.3, [0.96, 1.8]	1.34, [0.96, 2.01]	11.17, [0, 96.25]
no error, random selection, imputed	77	145.22, [21.82, 295.24]	155.9, [23.01, 324.07]	125.47, [28.6, 277]
top error, backbone and patch	26	229.44, [138.34, 337.68]	276.1, [145.44, 420.67]	1178.08, [531.25, 1819]
top error, random selection	28	189.18, [116.11, 263.13]	199.53, [121.3, 273.98]	325.04, [135.4, 704.52]
top error, backbone and patch, imputed	40	378.64, [191.23, 509.49]	454.9, [208.13, 655.52]	864.72, [676.78, 1082.88]
top error, random selection, imputed	97	449.35, [271.36, 590.53]	532.95, [279.87, 726.98]	1418.04, [1028.4, 2087.4]

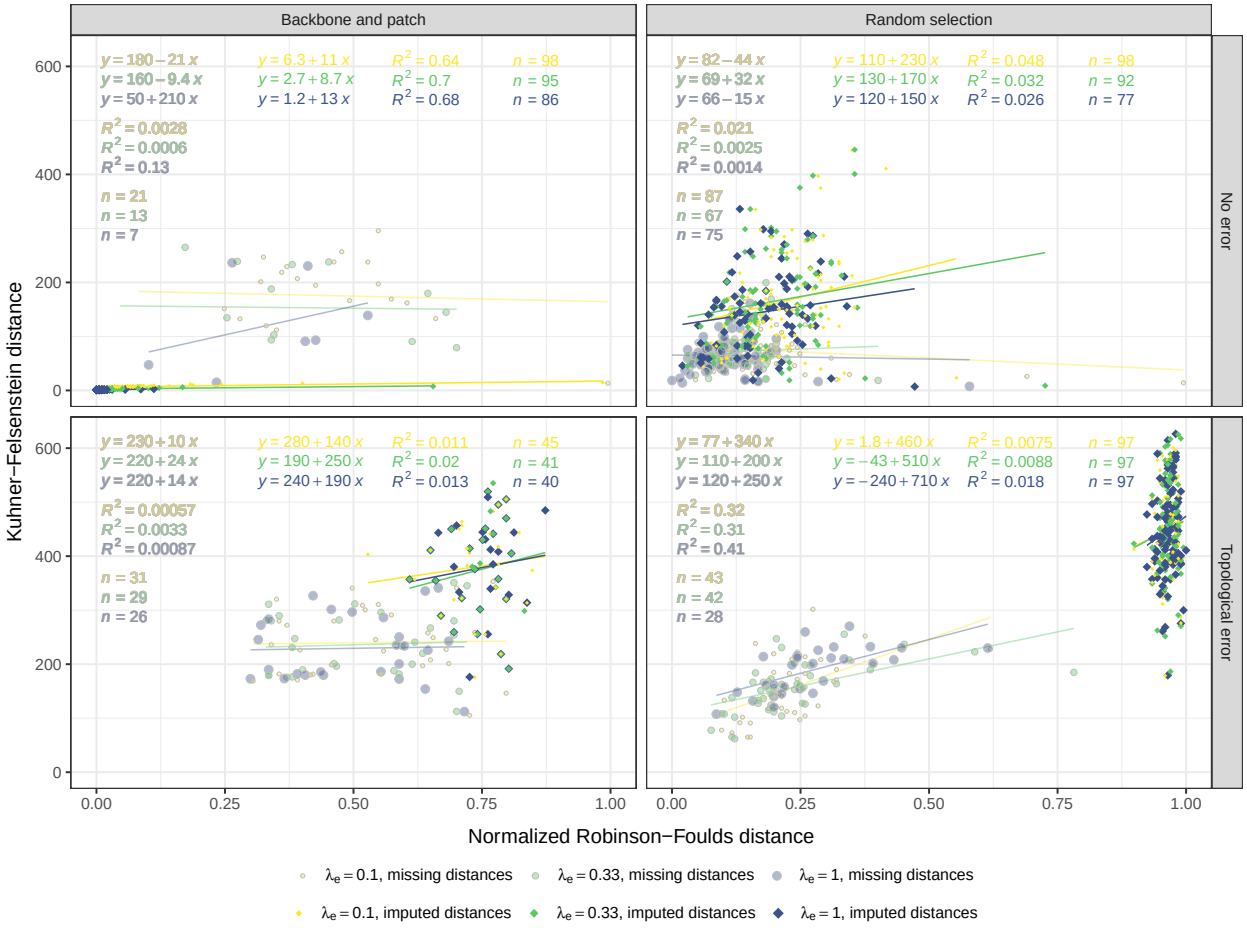

Figure A.16: Scenario-specific relationships of the Robinson-Foulds (x) and Kuhner-Felsenstein (y) distances between the true (simulated) and estimated (BLeSS MCC) supertrees, computed from all n replicates that reached convergence. These two accuracy metrics were used as response variables in the Bayesian additive regression tree (BART) analyses of the 200-tip simulation results. While exhibiting a strong and positive correlation globally (i.e., across scenarios: $y = 28 + 430x$; $R^2 = 0.8$), they show no or even an inverse relationship in several simulation scenarios (e.g., no error, random selection, missing distances), motivating the use of both metrics.

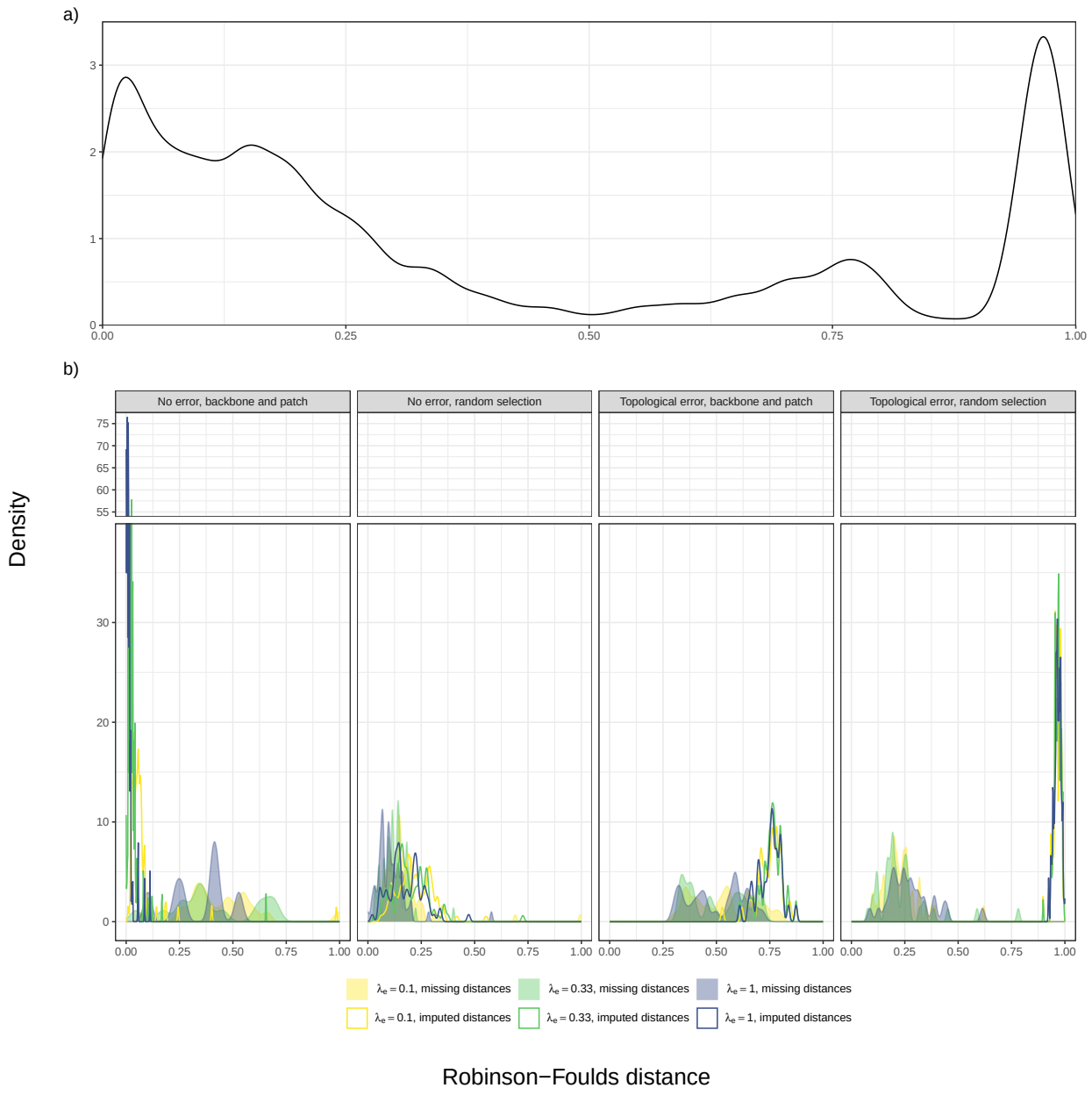

Figure A.17: Global (a) and scenario-specific (b) distributions of the Robinson-Foulds distance between the true (simulated) and estimated (BLeSS MCC) supertree, one of the accuracy metrics used as response variables in the BART analyses of the 200-tip simulation results. Note the y-axis break from 38 to 55 in panel (b).

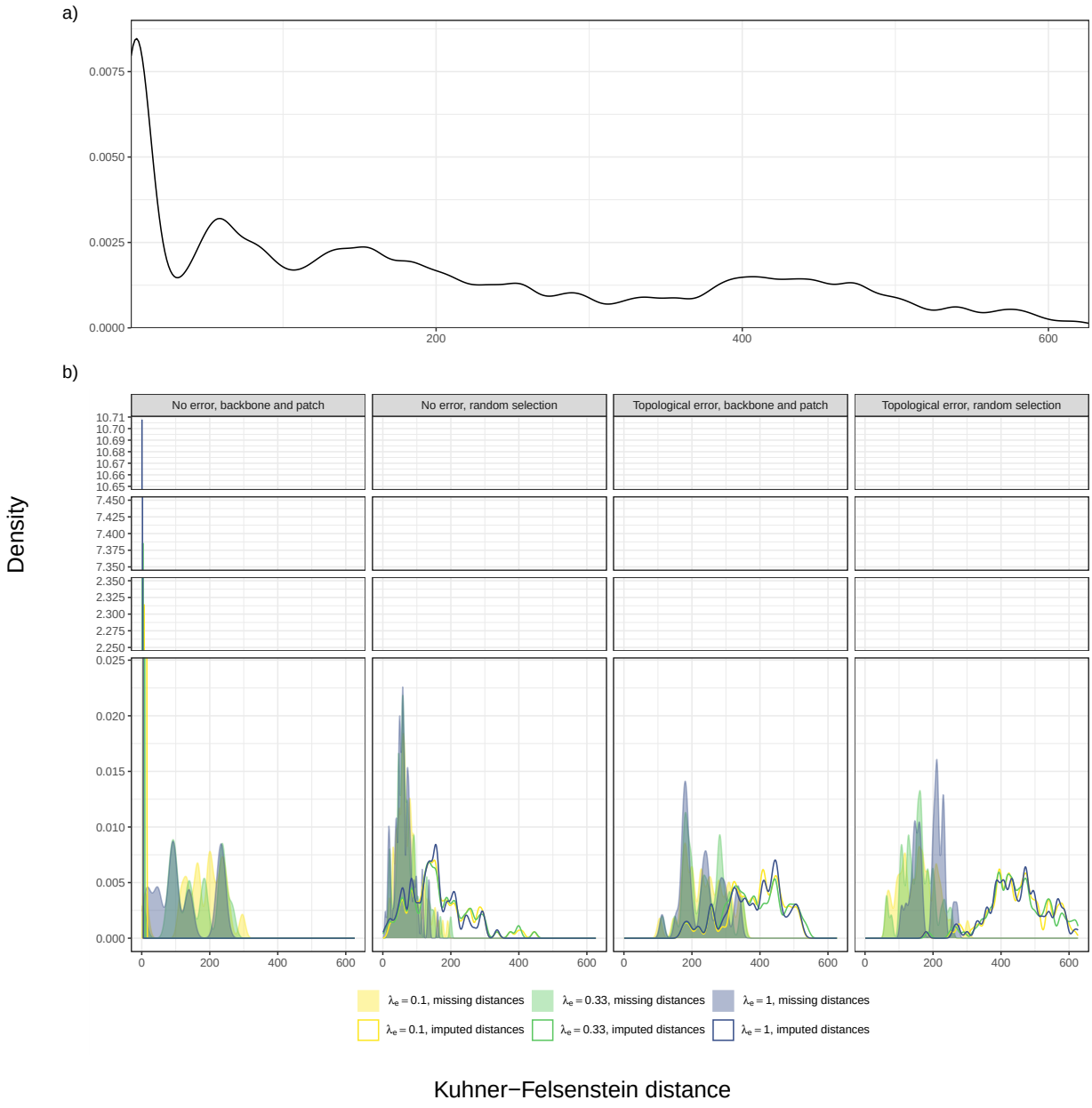

Figure A.18: Global (a) and scenario-specific (b) distributions of the Kuhner-Felsenstein distance between the true (simulated) and estimated (BLeSS MCC) supertree, one of the accuracy metrics used as response variables in the BART analyses of the 200-tip simulation results. Note the y-axis breaks from 0.024 to 2.25, from 2.35 to 7.35, and from 7.45 to 10.65 in panel (b).

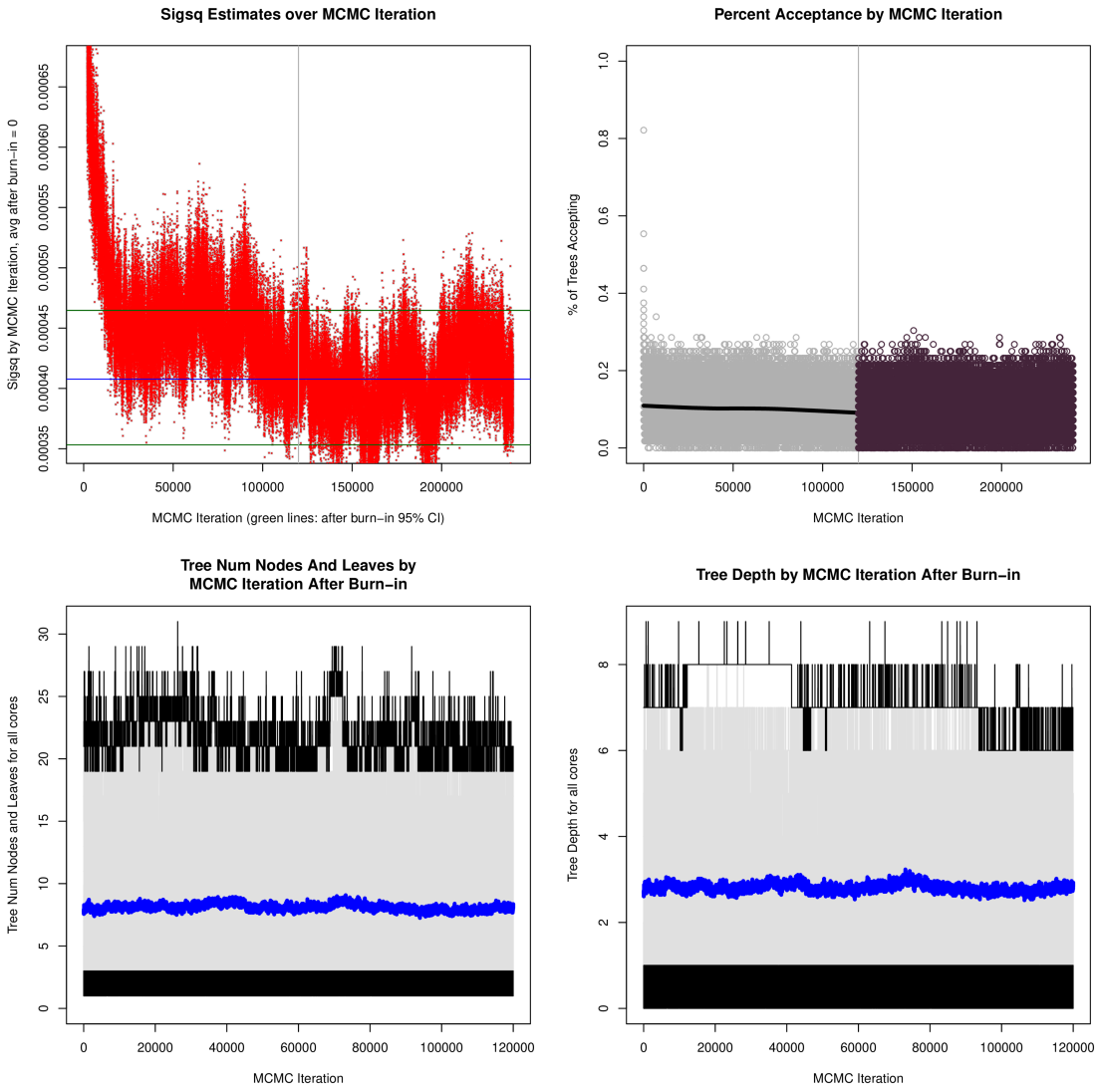

Figure A.19: Convergence diagnostics for the BART analysis of the 200-tip simulation results with the RF distance between the true (simulated) and estimated (BLeSS MCC) supertree as the response variable. Top left: posterior estimate of σ^2 (error variance; "sigsq") by MCMC iteration. Note that due to the Metropolis-within-Gibbs sampling employed by BART, each MCMC iteration corresponds to one Gibbs sample but $2m + 1$ Metropolis-Hastings (MH) steps, where m is the number of trees (56 for the analysis of RF distances). The blue and green lines indicate the post-burnin mean and 95% credible interval (CI), respectively. Top right: proportion of MH steps accepted for each Gibbs sample, calculated as the number of accepted steps divided by m. The gray vertical line separates the pre-burnin and post-burnin iterations; the thick black line shows the LOESS smoothing results. Bottom left: number of nodes (including terminal nodes, i.e., leaves) per tree in each post-burnin Gibbs sample. Gray lines correspond to individual trees, black lines denote the trees with the minimum and maximum number of nodes, and the thick blue line shows the mean number of nodes

over all $m = 56$ trees. Bottom right: tree depth (i.e., number of edges between the root and the most distant leaf) per tree in each post-burnin Gibbs sample. Gray lines correspond to individual trees, black lines denote the trees with the minimum and maximum depth, and the thick blue line shows the mean depth over all $m = 56$ trees. Note that the analysis was conducted on a single core to ensure deterministic output.

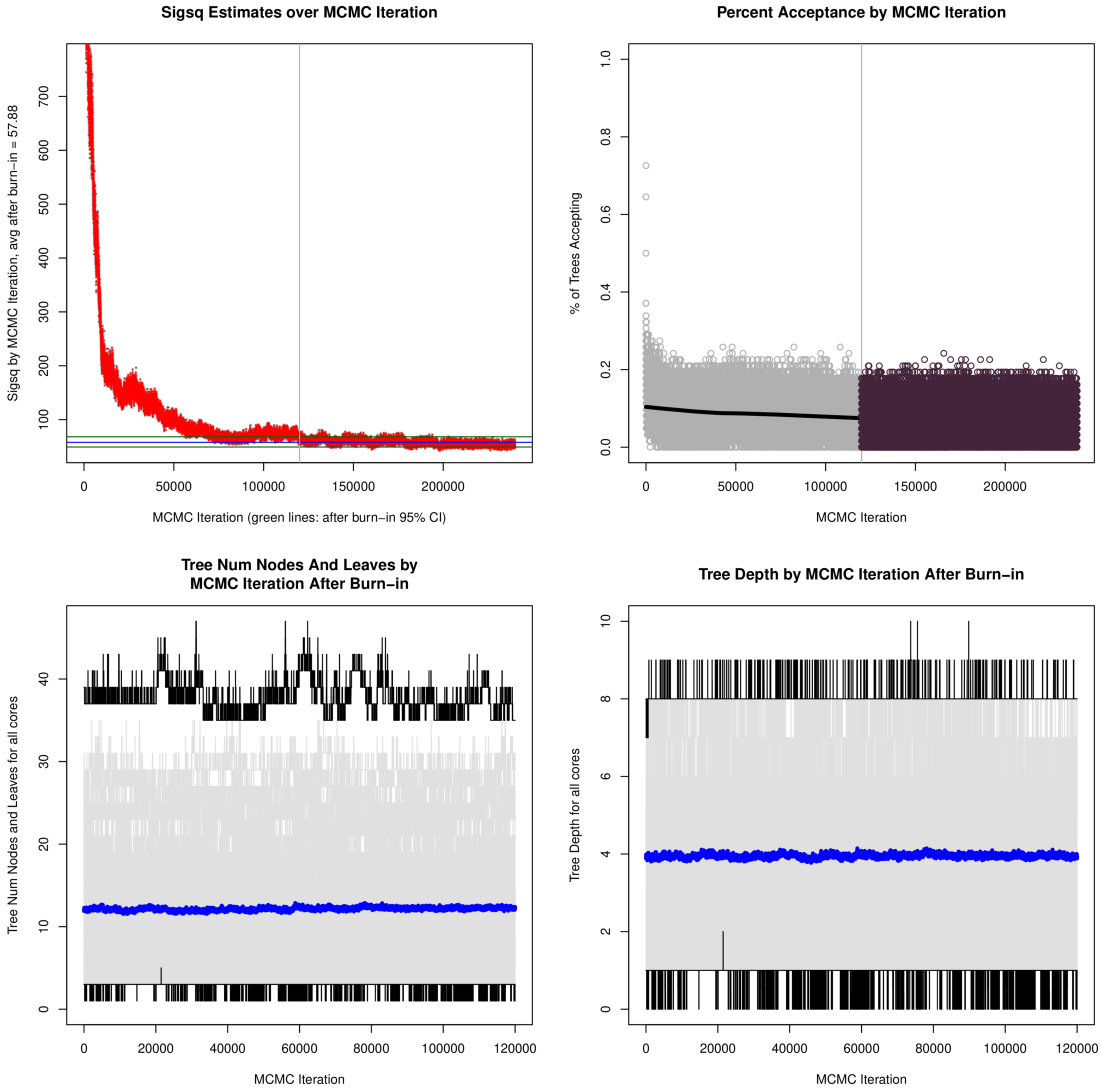

Figure A.20: Convergence diagnostics for the BART analysis of the 200-tip simulation results with the KF distance between the true (simulated) and estimated (BLeSS MCC) supertree as the response variable. Panel descriptions as in Figure A.19. The analysis employed $m = 62$ trees and was conducted on a single core to ensure deterministic output.

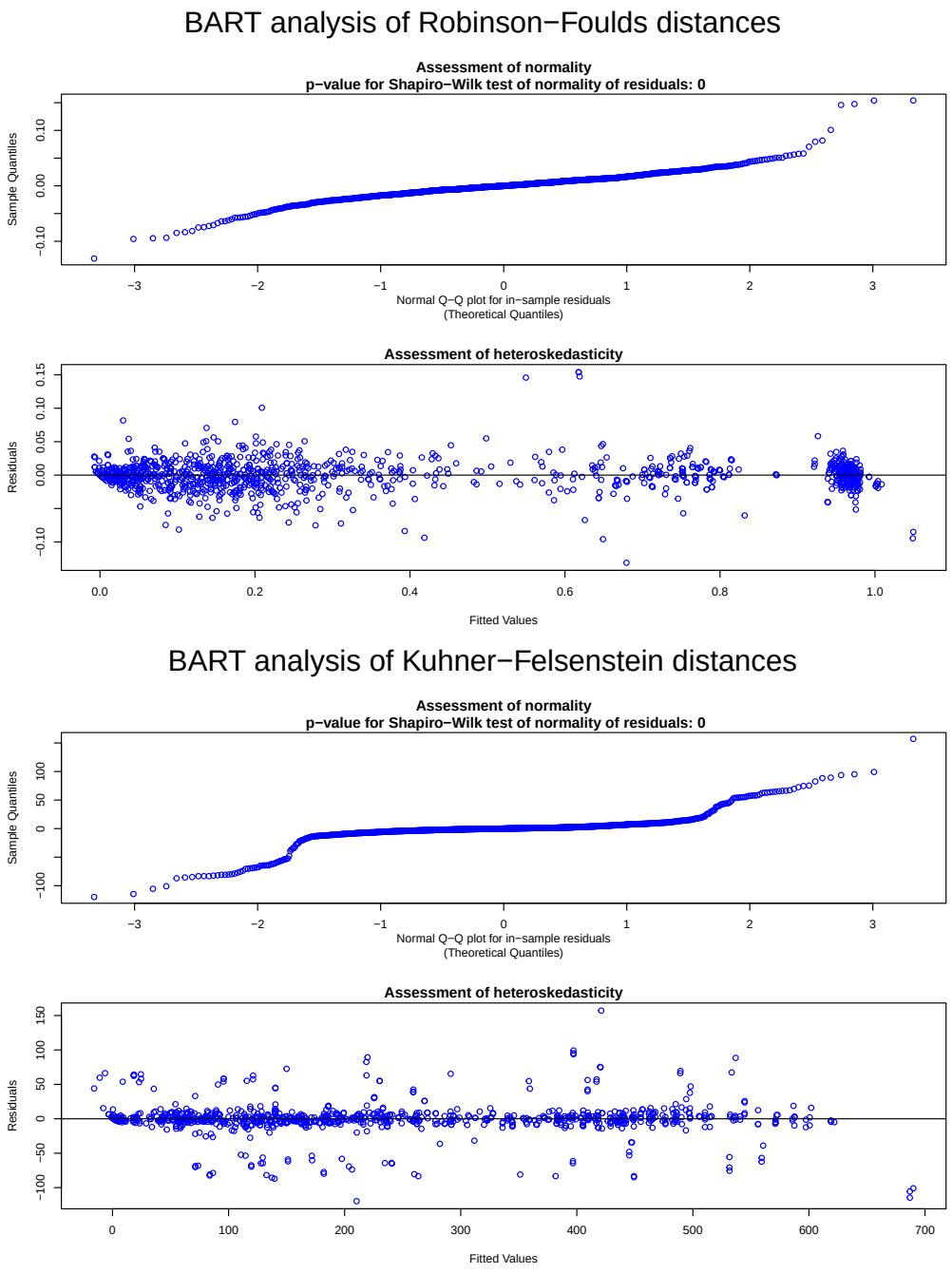

Figure A.21: Error assumption checks for the BART analyses of the RF (top) and KF (bottom) distances between the true (simulated) and estimated (BLeSS MCC) supertrees. The model assumes that the errors are normally distributed and homoskedastic; the former is assessed using the Shapiro-Wilk test and a quantile-quantile plot of the actual residual distribution against the normal expectation, while the latter is tested by plotting the residuals against the fitted values of the response. In both analyses, the normality assumption is violated.

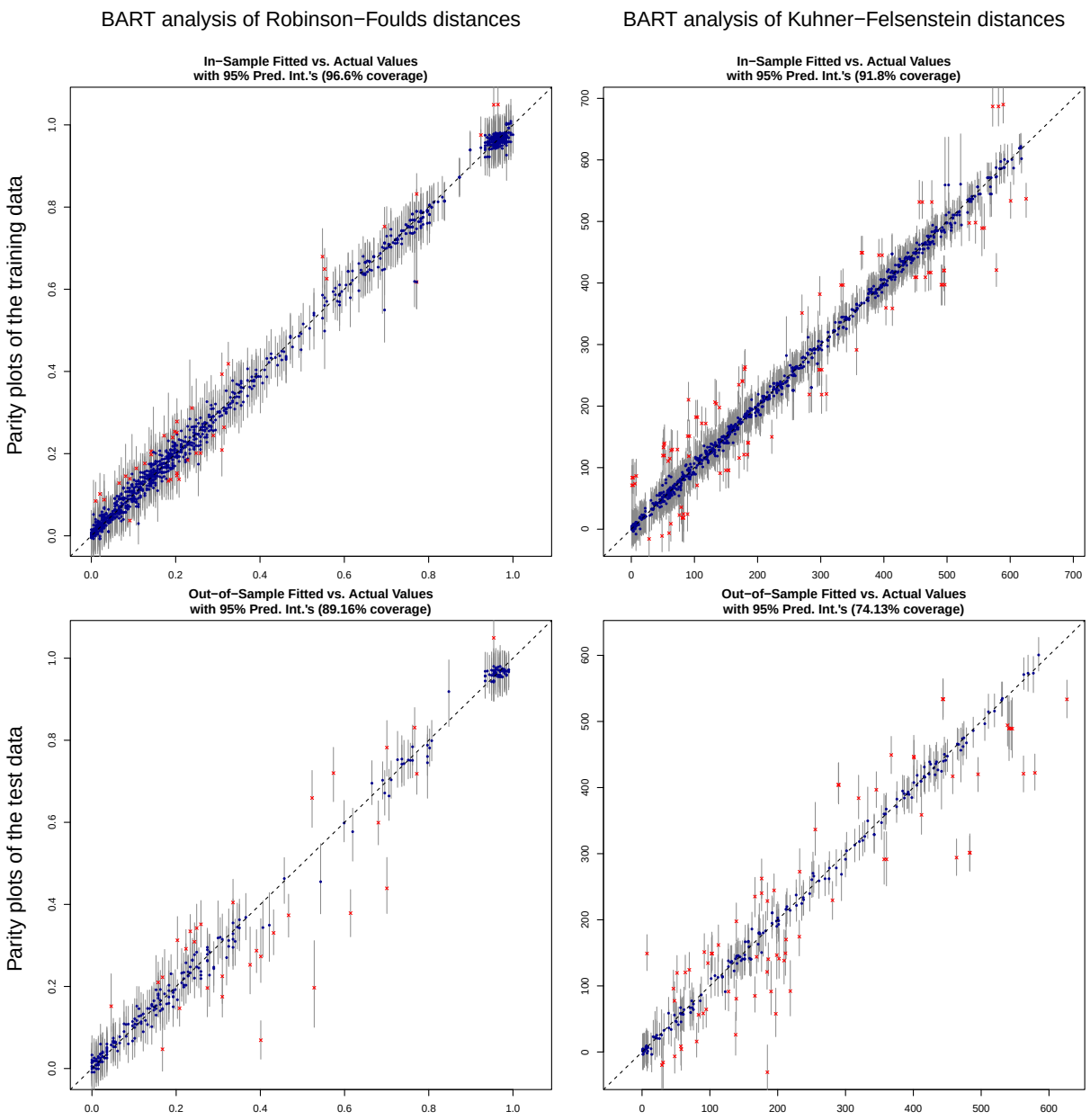

Figure A.22: Parity plots of the fitted vs. actual response values for the training (top) and testing (bottom) sets from the BART analyses of the RF (left) and KF (right) distances between the true (simulated) and estimated (BLeSS MCC) supertrees. The gray vertical segments represent prediction intervals; blue dots and red crosses indicate actual response values that do and do not fall within the corresponding prediction interval, respectively.

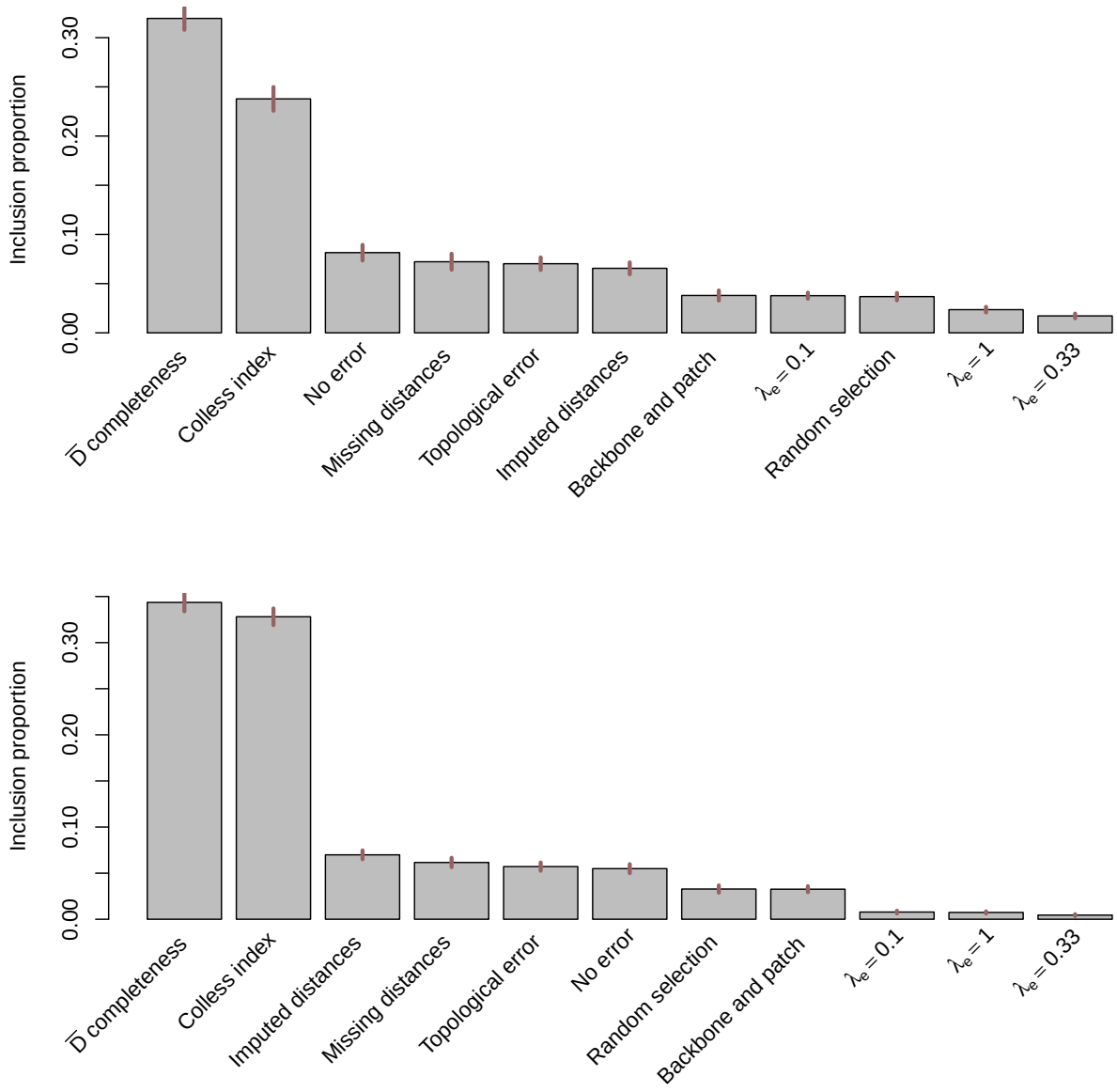

Figure A.23: Predictor importance in the BART analyses of the RF (top) and KF (bottom) distances between the true (simulated) and estimated (BLeSS MCC) supertrees, as quantified by the number of splits that use a given predictor as a splitting variable divided by the total number of splits in the ensemble ("inclusion proportion"). This proportion is averaged across all m trees (here, $m = 20$ to force the predictors to compete for inclusion), all post-burnin MCMC iterations (Gibbs samples), and 20 runs to ensure the sampler does not get stuck in a local maximum. The red segments represent 95% confidence intervals.

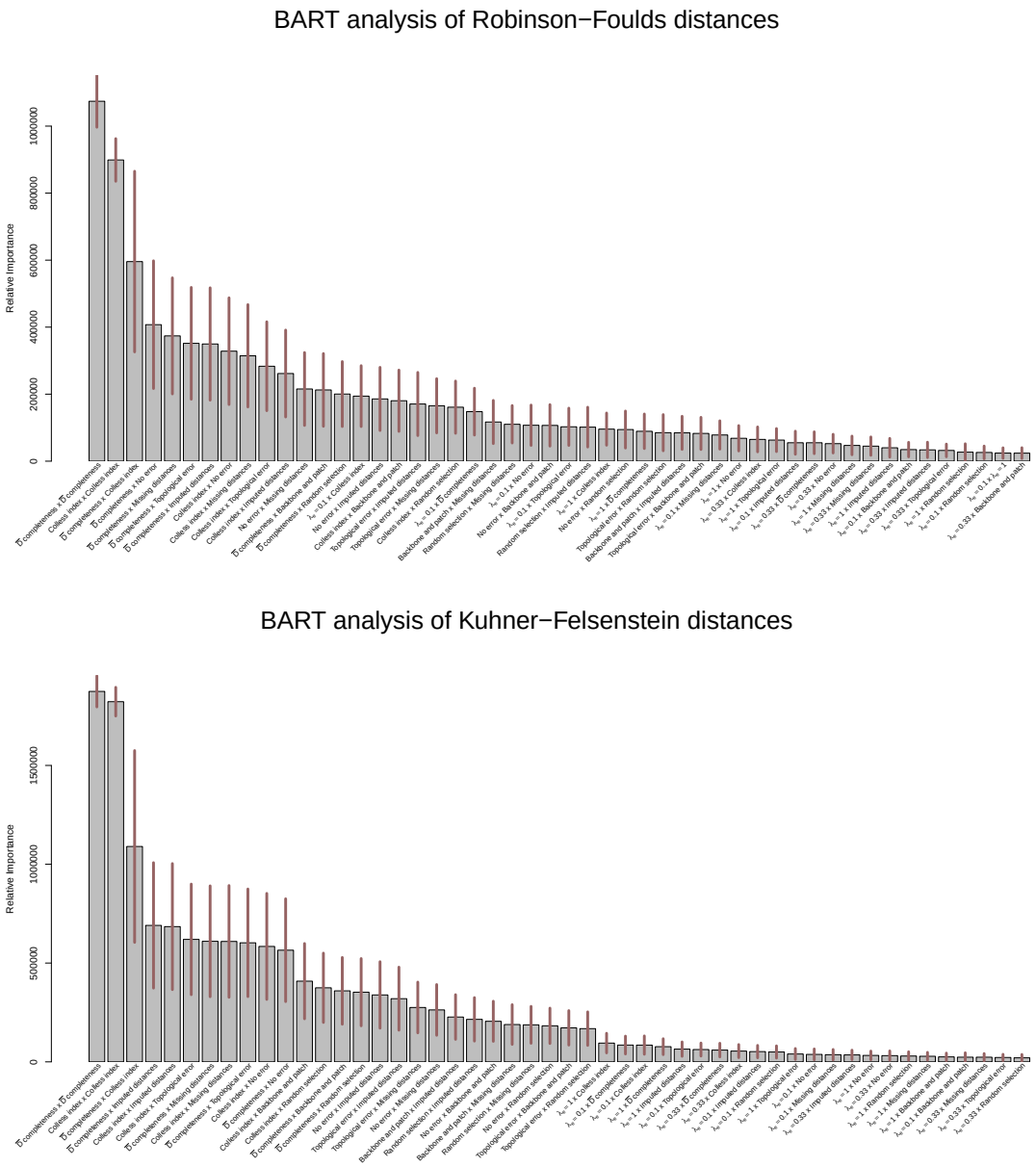

Figure A.24: The top 50 predictor interactions in the BART analyses of the RF (top) and KF (bottom) distances between the true (simulated) and estimated (BLeSS MCC) supertrees. Two predictors are considered to interact if splits that use them as splitting variables appear in the same root-to-leaf path. The strength of interaction ("relative importance") is quantified by the number of times this occurs, summed over all m trees (here, $m = 20$ to force the predictors to compete for inclusion), all post-burnin MCMC iterations (Gibbs samples), and 20 runs to ensure the sampler does not get stuck in a local maximum. The red segments represent 95% confidence intervals. Note that a predictor can interact with itself, such as when adjacent splits involve different thresholds for the same continuous variable.

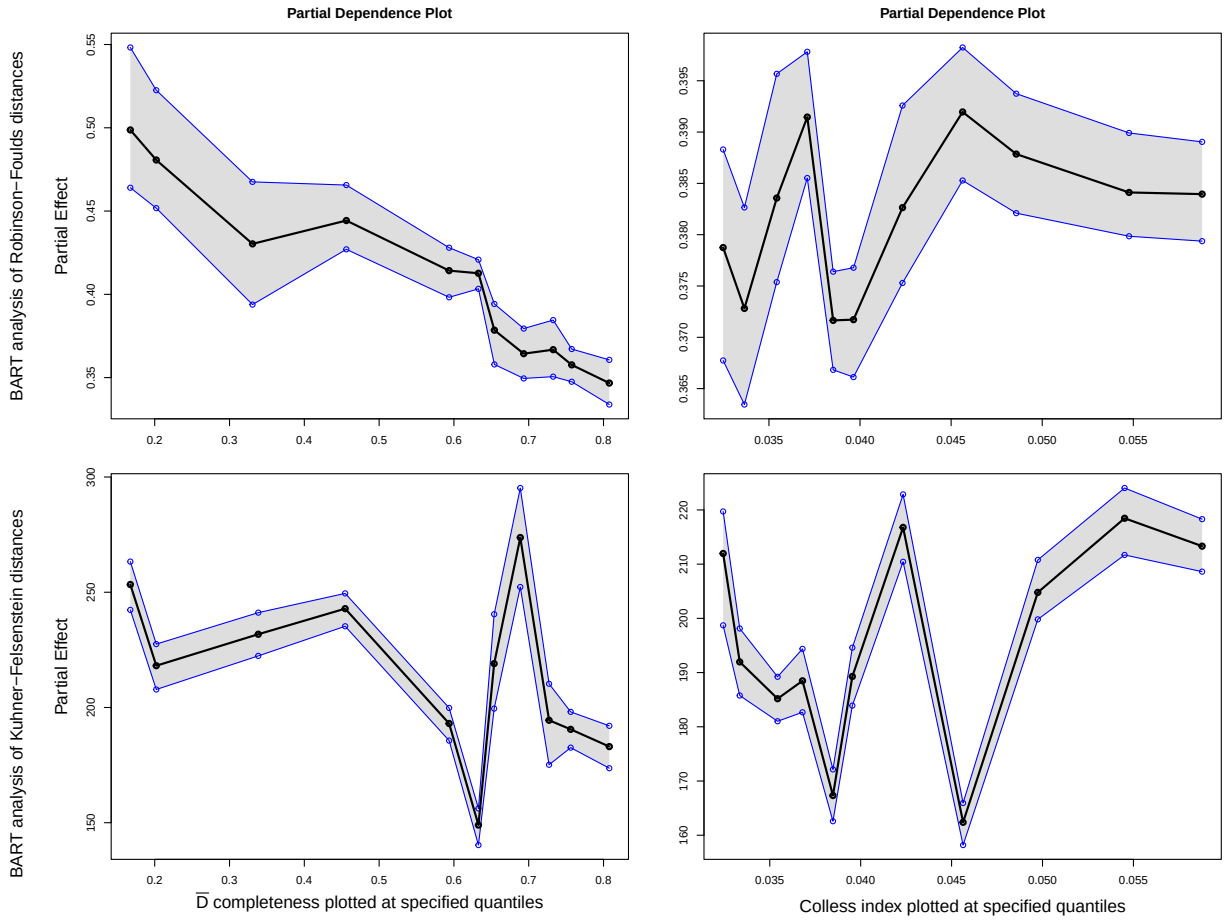

Figure A.25: Partial dependence plots for the two continuous predictors in the BART analyses of the RF (top) and KF (bottom) distances between the true (simulated) and estimated (BLeSS MCC) supertrees. The partial effect of a given predictor represents the average contribution to the predicted response after fixing the predictor to a given value and letting the remaining predictors vary over their marginal distribution. The mean estimate of the partial dependence function is plotted in black at the 5th, 10th, ..., 90th, and 95th percentiles, with 95% credible bands in gray and blue.

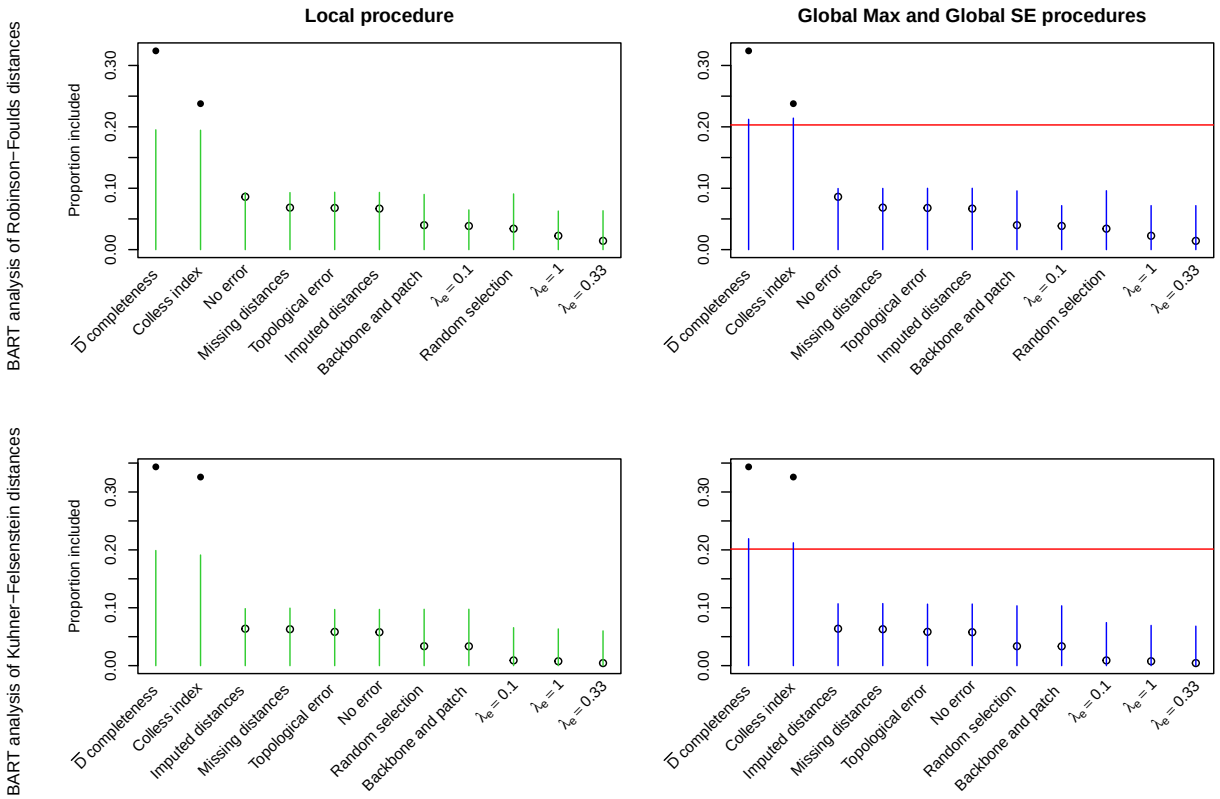

Figure A.26: Results of three variable selection procedures performed on the BART analyses of the RF (top) and KF (bottom) distances between the true (simulated) and estimated (BLeSS MCC) supertrees. Left: results of the Local procedure, with green segments representing the permutation-derived inclusion proportion (see Figure A.23) thresholds that a predictor has to exceed to be selected. Solid and open dots denote predictors that were and were not selected for inclusion in the model, respectively. Right: results of both the Global Max and Global SE procedures. The red line represents the Global Max cutoff, while the blue segments indicate the Global SE inclusion proportion thresholds. Solid dots denote predictors that exceeded both thresholds; open dots denote predictors that were not selected by either procedure. All thresholds were computed at $\alpha = 0.05$ using 100 permutations of the response and 25 BART runs with $m = 20$ trees each (to force the predictors to compete for inclusion). Note that all three procedures agree on the variables to be selected.

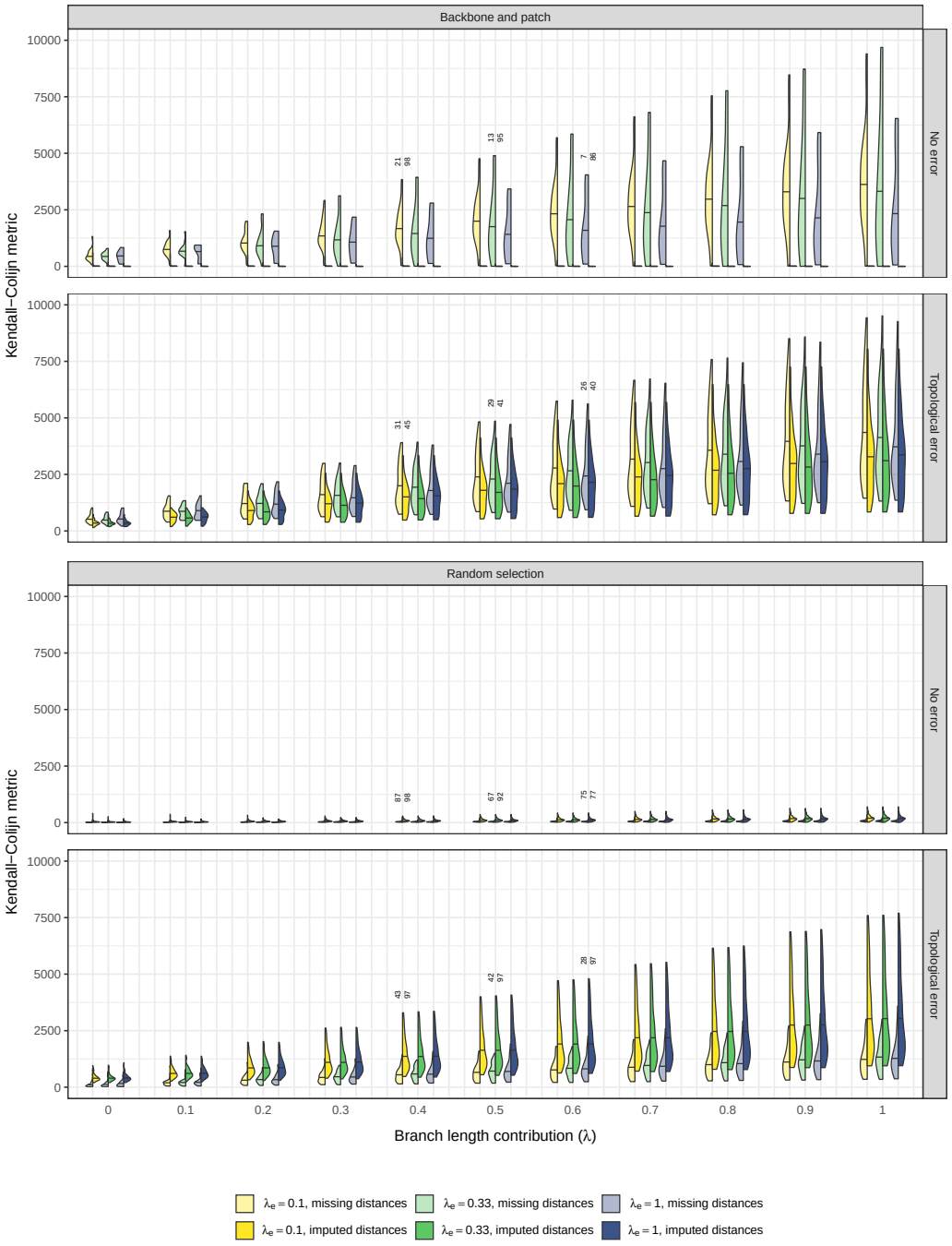

Figure A.27: Distributions of Kendall-Colijn (KC) distances between the true (simulated) and estimated (BLeSS MCC) supertrees, plotted for different simulation scenarios and different values of the parameter λ, which determines the relative contribution of branch lengths to the KC metric ($\lambda = 0$: topology only, $\lambda = 1$: branch lengths only). The values above the violin plots denote the number of replicates that reached convergence for a given simulation scenario.

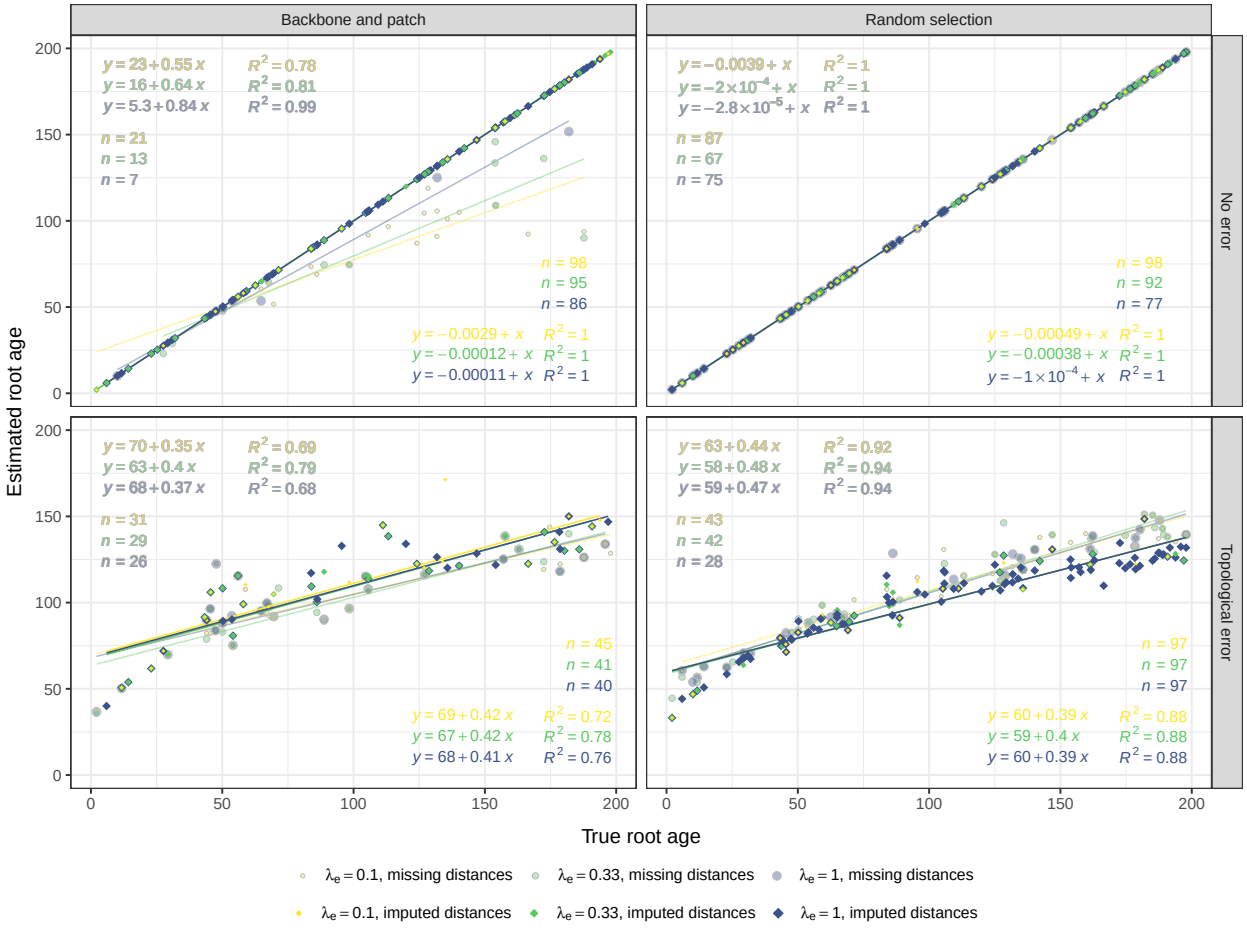

Figure A.28: Scenario-specific relationships of the true (simulated; x) and estimated (BLeSS posterior mean; y) root ages, computed from all n replicates that reached convergence. As in individual simulation scenarios, the two sets of ages also exhibit a strong and positive correlation globally (i.e., across scenarios: $y = 28 + 0.73x$; $R^2 = 0.83$).

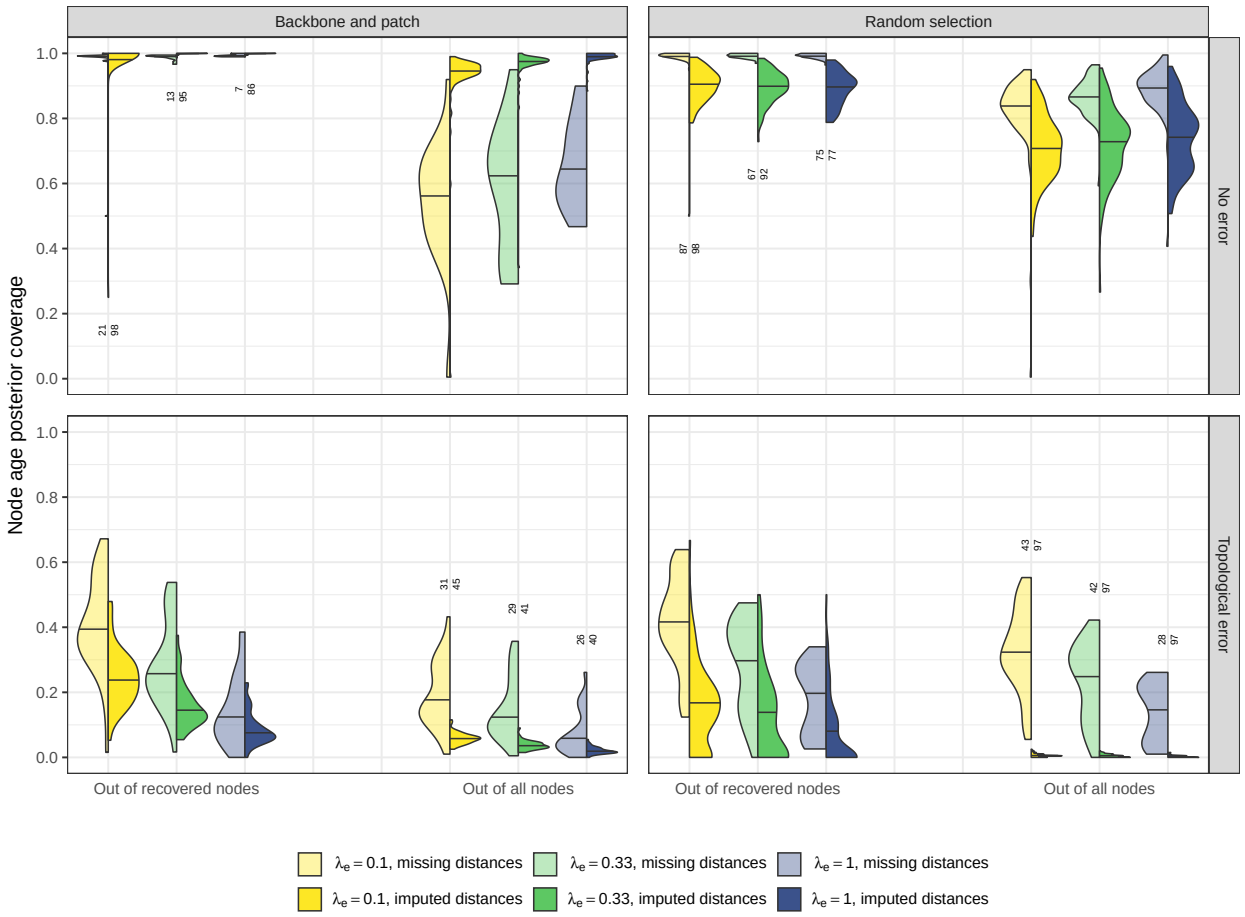

Figure A.29: The proportion of nodes in a given tree whose true (simulated) age falls within the corresponding 95% highest posterior density interval yielded by BLeSS, computed for different simulation scenarios and either out of only those nodes that were present both in the true tree and the BLeSS supertree (left group of violin plots), or all 199 nodes (right group of violin plots). The values above the violin plots denote the number of replicates that reached convergence for a given simulation scenario. Note that the resulting proportions should not be strictly interpreted as posterior coverages, because the ages of individual nodes are not mutually independent.

A.10.3 Scalability tests

Table A.9: Properties of the large-scale simulated trees and their corresponding average distance matrices ($\bar{\mathbf{D}}$). Root height is given in arbitrary units.

Tip number	Root height	Colless index	Subtree number	$\bar{\mathbf{D}}$ completeness
500	552.15	0.0208	11	0.8893
1000	621.46	0.012	17	0.8578
5000	782.40	0.0031	18	0.9125
10,000	851.72	0.0016	22	0.9081

Table A.10: Model and MCMC settings used for the large-scale tree analyses. Sampling freq. = frequency at which samples of scalar parameters and trees are extracted from the chain and logged; Printing freq. = frequency at which elapsed time and select parameter values are written to the standard output. Chain lengths and frequencies are given in MCMC moves.

Tip number	λ_e	Target chain length	Sampling freq.	Printing freq.	Checkpointing freq.
500	1	5×10^6	2.5×10^3	200	5×10^4
1000	1	5×10^6	500	50	2.5×10^4
5000	1	5×10^6	100	4	1.6×10^4
10,000	1	5×10^6	50	2	8×10^3

Table A.11: Wall-clock time spent on individual operations underlying BLeSS as a function of tree size and missing data treatment. All values are averaged over 10 replicates and measured with millisecond precision for the MPI version of RevBayes v1.2.2 (commit b083532) run with 32 threads per replicate on a compute cluster consisting of Intel Xeon Gold 6248R processors.

	Time (s); no imputation			
Operation	500 tips	1000 tips	5000 tips	10,000 tips
Convert source trees to distance matrices	0.118	0.225	0.868	2.341
Compute the average distance matrix $\bar{\mathbf{D}}$	0.021	0.087	1.919	7.765
Draw a tree $\hat{\mathcal{T}}$ from a birth-death process	0.032	0.125	3.100	12.566
Convert the proposed tree $\hat{\mathcal{T}}$ to a distance matrix $\hat{\mathbf{D}}$	0.033	0.139	4.156	18.872
Draw from the exp. error distribution conditional on $\hat{\mathbf{D}}$	0.071	0.293	8.022	33.942
Clamp the distribution to $\bar{\mathbf{D}}$	0.016	0.072	1.630	6.867
Compute the log probability density	0.003	0.006	0.143	0.612
	Time (s); imputation			
Read in an imputed distance matrix	0.108	0.486	38.935	291.262
Convert the imputed matrix to type `AverageDistanceMatrix`	0.032	0.141	4.486	19.944
Draw a tree $\hat{\mathcal{T}}$ from a birth-death process	0.032	0.124	3.131	12.630
Convert the proposed tree $\hat{\mathcal{T}}$ to a distance matrix $\hat{\mathbf{D}}$	0.033	0.140	4.156	18.874
Draw from the exp. error distribution conditional on $\hat{\mathbf{D}}$	0.071	0.291	7.972	34.345
Clamp the distribution to $\bar{\mathbf{D}}$	0.028	0.126	4.220	20.428
Compute the log probability density	0.003	0.011	0.249	1.452

Table A.12: Wall-clock runtime per 1 million MCMC iterations (under `moveschedule="single"`, i.e., with a single Metropolis-Hastings move proposed per iteration) as a function of tree size and missing data treatment. All values are averaged over 5 replicates and were calculated for the single-core version of RevBayes v1.2.2 (commit ed422ba) on a compute cluster consisting of Intel Xeon Gold 6248R processors. See also Figure A.30.

	Time (s)	
Tree size	No imputation	Imputation
500 tips	26,482 s $\approx$ 7.4 hours	26,097 s $\approx$ 7.2 hours
1,000 tips	110,954 s $\approx$ 30.8 hours	108,426 s $\approx$ 30.1 hours
5,000 tips	3.251×10^6 s $\approx$ 5.4 weeks (projected)	3.231×10^6 s $\approx$ 5.3 weeks (projected)
10,000 tips	1.384×10^7 s $\approx$ 5.3 months (projected)	1.464×10^7 s $\approx$ 5.6 months (projected)

Table A.13: Acceptance rates of different MCMC moves applied during BLeSS inference either directly to the supertree (`tree`) or to the hyperparameters of the birth-death prior from which the supertrees are drawn (`speciation`, `extinction`, `tree_height`), tabulated as a function of tree size and missing data treatment. The rates were calculated from the following total numbers of proposed moves: 5×10^6 (500 tips), 1.15×10^6 (1000 tips), 40,000 (5000 tips), and 8000 (10,000 tips). The weight of a particular move represents the relative frequency at which it is proposed (with the actual frequency obtained by dividing its weight by the sum of all move weights).

			Acceptance rate (no imputation)			
Move	Parameter	Weight	500 tips	1000 tips	5000 tips	10,000 tips
`Scaling`	`speciation`	3	0.2131	0.1646	0.0955	0.0533
`Scaling`	`extinction`	3	0.2185	0.2241	0.9428	0.9203
`Scaling`	`tree_height`	2	0.0007	0.0025	0.0113	0.0392
`NarrowExchange`	`tree`	10	0.1509	0.2285	0.4870	0.4949
`NNI`	`tree`	10	0.1178	0.1836	0.4522	0.4581
`FNPR`	`tree`	10	0.0034	0.0235	0.4238	0.4974
`SubtreeScale`	`tree`	10	0.2404	0.3157	0.5624	0.5458
`NodeTimeSlideUniform`	`tree`	15	0.3145	0.4152	0.7511	0.7533
			Acceptance rate (imputation)			
`Scaling`	`speciation`	3	0.4271	0.4003	0.1039	0.0545
`Scaling`	`extinction`	3	0.4279	0.4115	0.9580	0.9212
`Scaling`	`tree_height`	2	0.0001	0.0004	0.0054	0.0183
`NarrowExchange`	`tree`	10	0.0331	0.0866	0.2904	0.3604
`NNI`	`tree`	10	0.0185	0.0496	0.2591	0.3263
`FNPR`	`tree`	10	0.0046	0.0264	0.4715	0.4912
`SubtreeScale`	`tree`	10	0.0756	0.1016	0.3120	0.3579
`NodeTimeSlideUniform`	`tree`	15	0.0938	0.1714	0.4999	0.5598

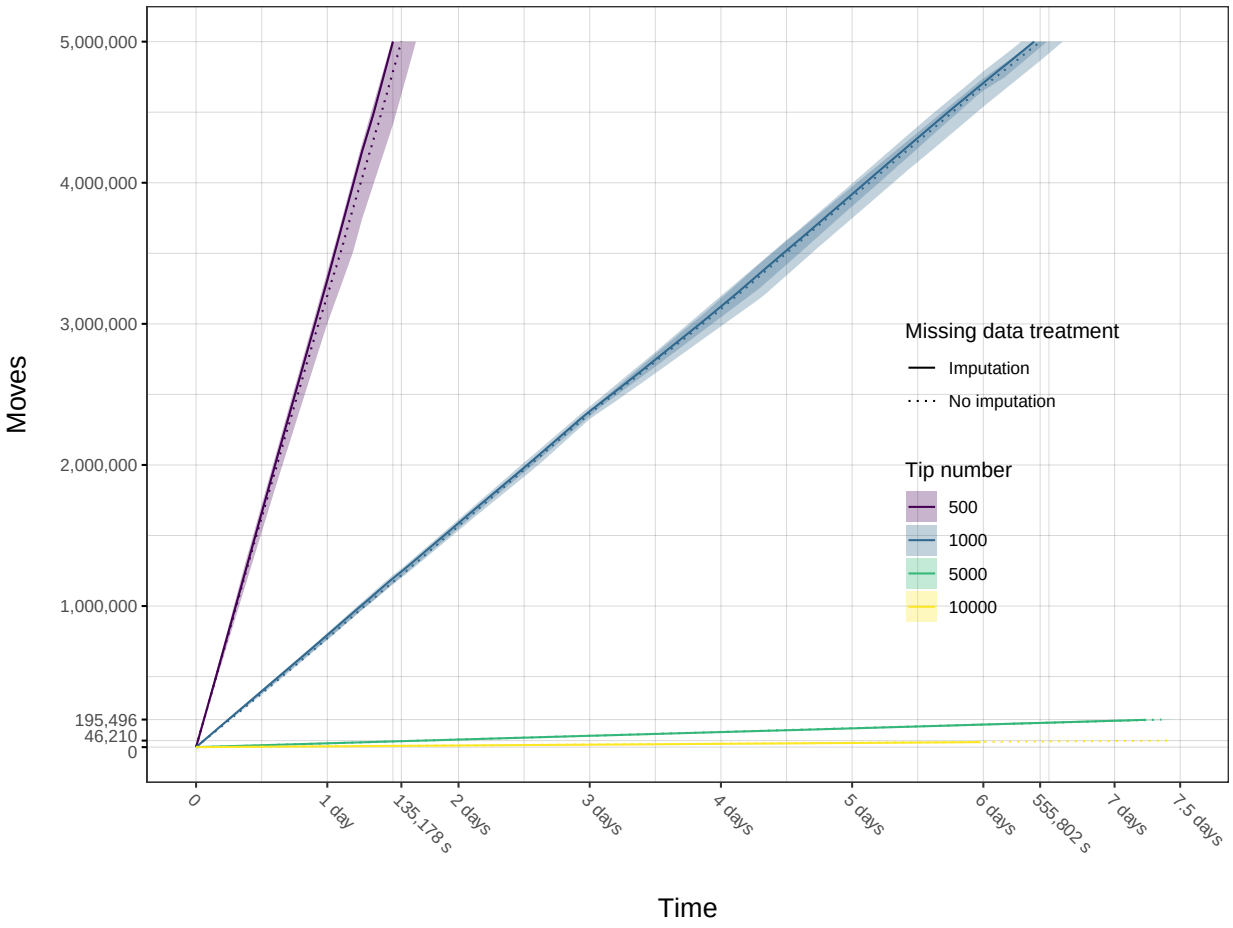

Figure A.30: Number of MCMC iterations (under `moveschedule="single"`, i.e., with a single Metropolis-Hastings move proposed per iteration) as a function of wall-clock time, plotted for different tree sizes and missing data treatments. Lines denote mean values across 5 replicates (performed on the same simulated dataset but under different random seeds); shaded bands indicate the corresponding range. The maximum runtime of the analyses that reached the target chain length of 5 million iterations within the time limit of 7.5 days (500 tips, 1000 tips) is recorded along the x-axis. For the analyses that failed to do so (5000 tips, 10,000 tips), the maximum number of iterations completed before the time limit was reached is recorded along the y-axis. When restarting analyses from a checkpoint file, the iterations performed after the last saved state but before a timeout were counted toward the total.

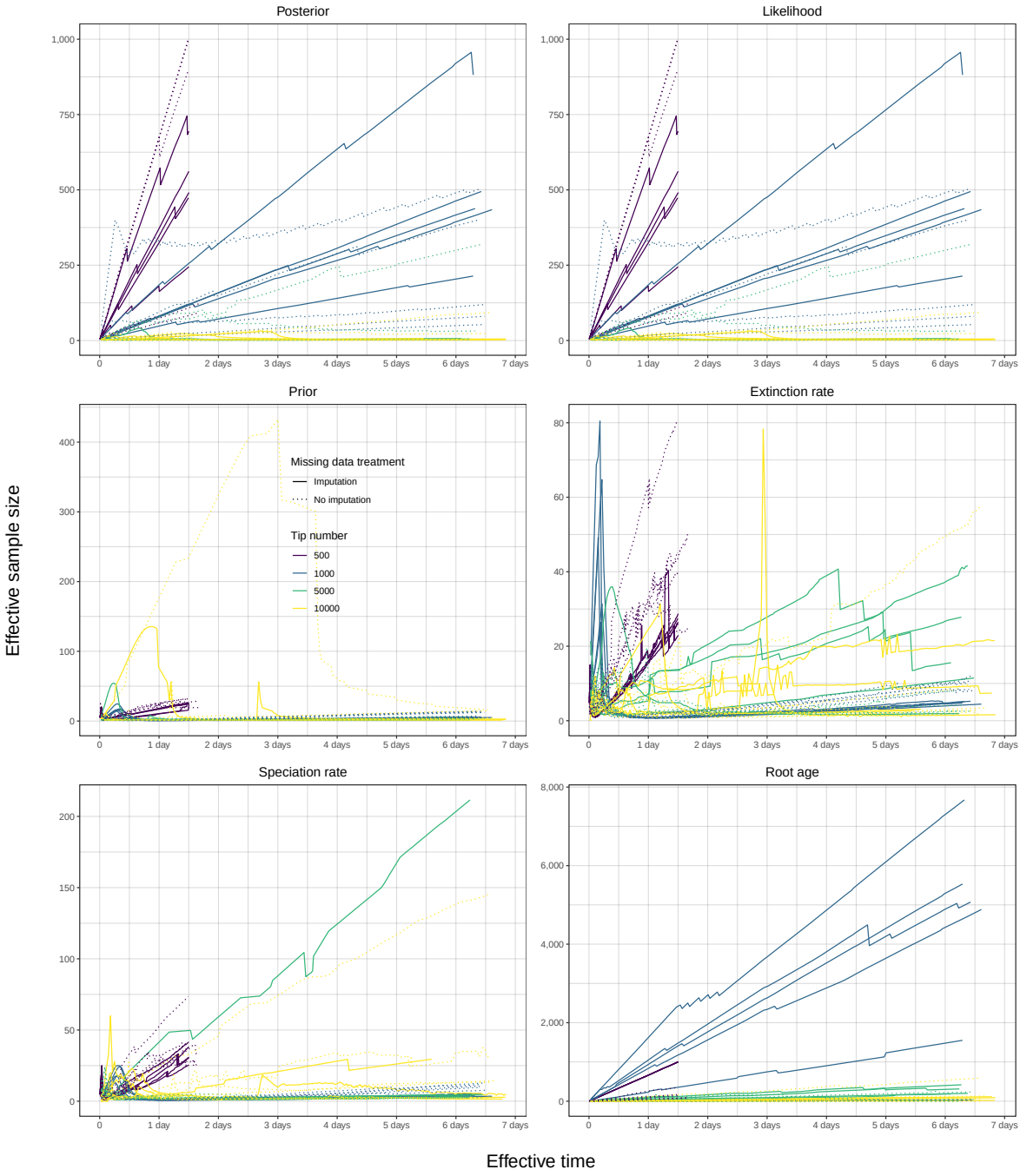

Figure A.31: Effective sample sizes of scalar parameters as a function of wall-clock time, plotted for different tree sizes and missing data treatments. Each of the 5 replicates (performed on the same simulated dataset but under different random seeds) is plotted as a separate line. When restarting analyses from a checkpoint file, the iterations performed after the last saved state but before a timeout were not counted toward the total; as a result, the effective time was less than the total runtime.

A.11 Empirical Application to the Order Carnivora

Our literature search yielded 41 candidate studies that carried out divergence time estimation for the order as a whole or various subclades thereof. We evaluated these studies for their suitability as BLeSS input, which was largely determined either by the availability of the resulting time tree(s) in a machine-parsable format (Newick or Nexus) or by the possibility of assembling such a time tree from other information provided (e.g., a figure illustrating the topology of the tree plus a table of divergence times). Below, we provide a brief description of each study, including its focal clade and (if applicable) the number of source trees extracted from it. Included studies are indicated by an asterisk (*).

Bagatharia et al. (2013)* *Felidae*. Estimated divergence times for a small sample of felids using whole mitochondrial genomes and a relaxed uncorrelated molecular clock in BEAST v1.7.1. They used a single calibration point at 10.78 Ma for Felidae. Although no tree file is available, we could reconstruct the topology and divergence times from the paper.

Barrett et al. (2021) *Carnivora*. Used BEAST v2.6.3 to infer a total-evidence tip-dated tree of carnivorans using 223 morphological characters, three nuclear, and 13 mitochondrial protein-coding loci (14,624 bp) for 13 extant and 32 extinct ingroup species as well as 3 outgroups. They did not provide a summary tree but did provide a posterior sample of trees in the supplement that were used for macroevolutionary analyses. Within this sample, extant tips are found to be non-contemporaneous because of the use of stratigraphic ranges for taxa with fossil records. Rather than extending tip branches or re-inferring the tree with these tips treated as extant, we opted not to use to this dataset.

Bon et al. (2008)* *Ursidae*. Used BEAST to infer phylogeny and divergence times for extant ursids and the extinct cave bear from mitochondrial genomes and two fossil

calibrations under an uncorrelated lognormal relaxed clock. No tree file is available and the tree is reconstructed from Figure 5, with divergence times taken from the associated figure legend.

Davis et al. (2010)* *Felidae*. Generated novel sequence data and employed Bayesian species tree analysis, followed by time scaling in PAML/multidivtime, for the big cats *Panthera* plus one outgroup. They used three minimum age calibrations and investigated the effects of moving each. As results are similar, we use the divergence times from analysis with the complete set of priors here. A tree file is not provided, so we used their Figure 8 to reconstruct the chronogram.

dos Reis et al. (2012) *Carnivora*. Used mitochondrial and nuclear loci to infer a dated phylogeny of 274 placental mammal species, including a number of carnivorans. However, no tree file or table of divergence times is available, with a figure showing topology being all that is available in the supplement.

Eizirik et al. (2010)* *Carnivora*. Sampled 14 genes across Carnivora and inferred the time scale of evolution using the Thorne-Kishino autocorrelated-rates clock model in divtime5b and a relaxed uncorrelated clock in BEAST v1.5.2, with 25 fossil calibrations. No tree files are available but we were able to manually digitize the tree and annotated it with mean ages across the divtime and BEAST analyses, where both were provided (their Table 6). For nodes not represented in Table 6, we used the divtime dates reported in Table S1 that were obtained from analyses with a 55 Ma root prior.

Fulton and Strobeck (2010)* *Phocidae*. Used fifteen nuclear genes and mitochondrial genomes and 8 fossil calibrations to infer topology and branch lengths in BEAST v1.4.8 for all extant phocid species as well as the walrus, three otariids and seven carnivore outgroups (two felids, two canids, three mustelids). We transcribed their time tree from Figure 2 and Table 4.

Gaubert and Begg (2006)* *Viverridae*. Employed two mitochondrial loci (cytochrome *b* and 5' domain of the control region) and one Y-linked (terminal exon of zinc-finger) gene to infer the phylogeny of the genus *Genetta* using likelihood and Bayesian methods. They then time-scaled the CytB tree under a relaxed clock using PAML/multidivtime with a single calibration point assigned to the root. We transcribed their time tree from Figure 2, omitting duplicated individuals within a single species and treating the divergence of *G. maculata* and *G. tigrina* as the older of the two ages.

Gaubert and Cordeiro-Estrela (2006)* *Viverridae*. Used penalized likelihood, as implemented in r8s v1.7, to time-scale a tree of feliform carnivorans, with sampling focused on Viverrinae, inferred from 2 nuclear and one mitochondrial gene. No tree file is available, so we digitized the tree plotted in Figure 4, which is described as a linearized tree showing minimum divergence time estimates. Several taxa (*Cryptoprocta ferox*, some *Genetta* species) were pruned from the tree as divergence times were not provided.

Hassanin et al. (2021) *Carnivora*. Added 51 new mitogenomes from 13 families to previously sequenced mitogenomes for an alignment of 220 taxa. They time-calibrated their tree using BEAST v2.4.7 with 21 fossil calibration points. Due to the completeness of this data set it is not used here for assessing the performance of our supertree method.

Helgen et al. (2013) *Procyonidae*. Used sequence data analyzed in BEAST to infer a time-calibrated phylogeny of Procyonidae while also describing a new species of olingo *Bassaricyon neblina*. However, no tree file is available and the descriptions of divergence times in the text are insufficient to reconstruct the chronogram manually. This dataset is therefore excluded.

Higdon et al. (2007)* *Pinnipedia*. Assembled a supertree of pinnipeds, plus a canid and ursid outgroup, from a weighted matrix representation with parsimony (MRP) analysis

of 50 gene trees inferred under maximum likelihood, and time-scaled the resulting topology using a perl script to calibrate relative branch lengths from the gene trees to a set of fossils. We took the topology from Figure 1 and the divergence time estimates from Table 2.

Johnson et al. (2006)* *Felidae*. Used 22,789 base pairs of nuclear and mitochondrial data from 37 extant felid species along with several outgroups to infer a maximum likelihood phylogeny using PAUP*. They then employed 16 fossil calibrations to timescale their tree using the nuclear loci alone in the program estbranches. A tree file is not available but we were able to reconstruct the tree topology and branch lengths from their Figure 1 and Table 1.

Koepfli et al. (2006)* *Hyaenidae*. Used 7 nuclear loci and 1 mitochondrial gene to infer a maximum likelihood phylogeny for the four extant species of Hyaenidae and a set of feliform outgroups, and then time-scaled this topology using a relaxed molecular clock as implemented in divtime5b based on 9 fossil constraints. We manually digitized the tree topology from Figure 2 and added mean node ages from Table 6.

Koepfli et al. (2007)* *Procyonidae*. Used data from 9 nuclear and 2 mitochondrial genes to infer the phylogeny of Procyonidae and estimated divergence times under a relaxed molecular clock using divtime5b with a variable set of fossil calibrations. We manually reconstructed the maximum likelihood topology from Figure 3 and sampled all three sets of divergence times from Table 4.

Koepfli et al. (2008)* *Mustelidae*. Used 22 gene segments for 44 mustelids plus 2 procyonid outgroups to infer a time-scaled phylogeny using BEAST v1.4.2. They employed several calibration strategies, reporting the mean and 95% highest posterior density (HPD) intervals for each. We used the mean of the mean ages for each node. We note that divergence times were not provided for the splits between *Bassariscus astutus* and

Procyon lotor, *Enhydra lutris* and *Hydrictis maculicollis*, and *Martes americana* and *Martes melampus*. In each case, we deleted the first named taxon from the tree.

Krause et al. (2008)* *Ursidae*. Also used mitochondrial sequences for extant bears, plus two extinct taxa (short faced and cave bears) and inferred phylogeny and branch lengths using BEAST v1.4.4 under a strict molecular clock. The tree is manually reconstructed from Figure 1 of their paper.

Kutschera et al. (2014)* *Ursidae*. Used *BEAST to jointly estimate gene trees for 14 autosomal introns and 9 Y-chromosome genes, and the species tree for all 8 extant bear species under a multilocus coalescence approach. The topology with highest posterior probability (their Figure 2A) is used here, with posterior mean divergence times taken from Table 1.

Law et al. (2018)* *Musteloidea*. Used a 46-gene dataset in combination with 74 fossil ages to infer a time-scaled phylogeny for 75 species of musteloid carnivores under a relaxed uncorrelated lognormal molecular clock and fossilized birth-death tree prior in BEAST v2.3.2. We use the maximum clade credibility tree from this study.

Li et al. (2016)* *Felidae*. Used phylogenomic data to investigate phylogeny, divergence times and introgression in felids. No tree files are provided so we used the topology and mean branch lengths derived from multidivtime analyses performed on the biparental nuclear DNA genome partition, with different partitions and constraints, as reconstructed from figure 5A and mean ages from Table S3.

Lindblad-Toh et al. (2005) *Canidae*. Presented a time-scaled phylogeny of canids based on ∼15 kbp of exon and intron sequences. However, the figure does not include all node ages and a tree file is not available. This dataset is not included here.

Lopes et al. (2021)* *Otariidae*. Used whole-genome sequencing for 14 of 15 extant species

of Otariidae and inferred time-calibrated species trees using ASTRAL and StarBeast 2.

Lynch (2020)* *Mustelidae*. Used 12S, 16S, cytochrome b, and D-loop sequences for *Martes americana*, *M. caurina*, and *M. foina*, combined with three fossil calibrations, to infer the timescale of divergence using BEAST v1.10.4. The tree was manually transcribed from Figure 3 with mean divergence times recorded from the text.

McDonough et al. (2022)* *Mephitidae*. Used a multilocus nuclear and mitochondrial dataset to recognize seven distinct species of *Spilogale*. They estimated divergence times in BEAST v2.6.1 based on 13 protein-coding genes from the mitochondrial genome and calibrated the molecular clock using one fossil calibration and one molecular date for the origin of American mephitids.

Meredith et al. (2011)* *Carnivora*. Used a molecular dataset comprising 26 genes with a variety of treatments and fossil calibration schemes to infer the timescale of mammalian diversification. We extracted the carnivoran portion of their trees and scaled branches to mean divergence times reported over calibration schemes in the supplementary materials. The topology of the tree differed in resolution at the base of Arctoidea between analysis of amino acids vs nucleotides (ursids as sister to pinnipeds in the former vs. ursids as sister to pinnipeds plus musteloids in the latter), so we included two versions of the tree, with divergence times taken from the appropriate dataset.

Mitra et al. (2019)* *Feliformia*. Used protein coding loci from complete mitochondrial genomes and BEAST v1.7 to infer the timescale of feliform evolution from 24 species, plus a canid outgroup (*Cuon*). We transcribed their maximum clade credibility tree from Figure 2 of their paper.

Nascimento et al. (2021)* *Felidae*. Compiled previously published mitochondrial sequences covering ATP8, control region, CytB, and ND5 for for the pampas cat complex and inferred topology and divergence times under a relaxed clock in BEAST v1.8.2 using

2 calibrations. Based on the resulting clades, combined with morphological features, they recognized five species of pampas cat. A tree file was not available for this study but the topology and branch lengths could be manually reconstructed based on the mean ages labeled in Figure 8a of their paper.

Nigenda-Morales et al. (2019) *Procyonidae*. Used three mitochondrial loci genotypes from 11 microsatellite loci and BEAST v2.3.1 to examine the phylogeographic history of the white-nosed coati *Nasua narica* but also included samples from other procyonid species. However, no supplementary tree files are available and divergence times are not listed or described in the text. This dataset therefore cannot be used.

Ogilvie et al. (2022)* *Canidae*. Introduced a new model, the FBD-MSC, that combines the multispecies coalescent (MSC) and the fossilized birth–death (FBD) processes. They used the approach to infer the phylogeny and divergence times of all extant Canidae using a combined dataset of 58 nuclear genes and 230 morphological characters. Following their methods, we combined the sampled trees from four posterior chains (sb2-extant-only) provided in their online materials after removing the first 2% of each as burnin and produced a maximum clade credibility tree from the resulting output, annotated with mean divergence times.

Paijmans et al. (2017)* *Felidae*. Used mitogenomes to investigate the phylogeny and divergence times of sabertoothed cats (Machairodontinae) with respect to a sample of extant felids and feliform carnivores. They do not provide a tree file and, although the time-scaled topology is figured in their Figure 1, they do not provide a table of divergence times. We were able to extract the topology for Felidae that encompassed the two genera of sabertoothed cats and two felines spanning the root node.

Patou et al. (2008)* *Viverridae*. Used two mitochondrial and two nuclear loci, combined with a relaxed molecular clock to infer the phylogeny and divergence times of a set

of hemigaline and paradoxurine viverrids. No tree file is available but we were able to manually reconstruct the tree topology and branch lengths using the divergence times estimated from the complete dataset. Note that the authors omitted the otter civet *Cynogale bennetti* from their combined analyses.

Patou et al. (2009) *Herpestidae*. Used 2 mitochondrial (CytB and ND2) and 2 nuclear (β-fibrinogen intron 7 and transthyretin intron 1) loci from almost all of the recognized mongoose species to produce a phylogeny of the Herpestidae. Divergence ages were estimated using a relaxed molecular clock approach using PAML/multidivtime. TTRi1 was excluded from divergence time estimation as it was missing for too many taxa. The tree is not available online and the figure is insufficient to reconstruct branch lengths. Therefore, this study is not included.

Patou et al. (2010) *Viverridae*. Used mitochondrial sequence data to estimate the phylogeny and divergence times of the genus *Paradoxurus*, including the identification of putative distinct, species-level OTUs within *P. hermaphroditus*. The did not provide a tree file, and while a time tree figure is shown in the paper, insufficient detail is provided to transcribe it. We therefore omitted this dataset.

Perri et al. (2021)* *Canidae*. Time-scaled a maximum likelihood topology inferred for 10 extant canids plus the dire wolf *Aenoycon dirus* using a nuclear supermatrix in PAML/MCMCTree. A time tree file is not available, so we manually transcribed the tree and branch lengths from Figure 2a. We retain the dire wolf here as its terminal age is within the rounding error of the branch durations.

Püschel et al. (2020) *Ursidae*. Used a combination of molecular and morphological data for extant and extinct ursids, as well as a canid outgroup to infer branch lengths under an independent gamma rates (IGR) relaxed clock in MrBayes. Although tree files are available, the resulting branching time estimates are unusually young, and are therefore

excluded here.

Rao et al. (2020) *Hyaenidae*. Used proteomic data from 3 extant species of hyena as well as the extinct cave hyena (Eurasian *Crocuta crocuta*) to infer the timescale of hyaenid evolution using BEAST v2.6.2. Although XML files are available and a time-scaled phylogeny is plotted (their Figure 4), no tree file or description of divergence times are provided, and we therefore exclude this dataset.

Scheel et al. (2014)* *Phocidae*. Added CytB sequence data for the Neotropical monk seal *Neomonachus tropicalis* to the alignment of Fulton and Strobeck (2010) and jointly inferred the tree topology and divergence times using BEAST v1.7.4. They do not provide a tree file or a set of divergence time estimates for most nodes, but they do provide the latter for the node uniting the monk seals and the node uniting the two New World species. Accordingly, we assembled a 3-taxon tree from these two nodes.

Trindade et al. (2021)* *Felidae*. Generated a maximum likelihood tree from 60,931 SNPs to investigate the relationships of South American felids. They time-scaled this tree using PAML/MCMCTree under an autocorrelated-rates clock with a single calibration applied to the root node. There was no tree file available with the paper but we were able to manually digitize the topology and branch lengths present in Figure 1B of their paper.

Tseng et al. (2014) *Felidae*. Combined 43 kbp of molecular data with morphological data for 10 extant and recently extinct felid species with morphological data coded for those taxa plus an additional two fossil species and inferred a total-evidence tip-dated tree using MrBayes under a uniform tree prior and IGR clock. Although a MrBayes input file is available on MorphoBank (project number P898) there is no tree file and the divergence times are not provided in the paper. We therefore opted not to use this dataset.

Westbury et al. (2021)* *Hyaenidae*. Used ASTRAL to infer a species tree for the four extant species of Hyaenidae from maximum likelihood gene trees estimated using IQ-TREE. They then dated this tree using 3 fossil calibrations with soft bounds in the software PAML/MCMCTree. We took the tree topology from Figure 1 of their paper and posterior mean divergence times from the text. Outgroups were not included as dates are not available in the paper.

Wolsan and Sato (2020)* *Pinnipedia*. Used three taste receptor genes and six fossil constraints to sample time-calibrated phylogenies for 16 pinnipeds and 8 caniform outgroups. We transcribed the maximum likelihood tree shown in Supplementary Figure 5 and the posterior mean divergence time estimates from Supplementary Table 8.

Table A.14: Composition of the 53 benchmark clades used to evaluate the performance of the empirical BLeSS analyses, ordered by taxonomic rank from most inclusive to least inclusive.

Clade	Rank	Included species
Caniformia	Suborder	*Aenocyon dirus*, *Ailuropoda melanoleuca*, *Ailurus fulgens*, *Aonyx capensis*, *Aonyx cinerea*, *Arctocephalus australis*, *Arctocephalus forsteri*, *Arctocephalus galapagoensis*, *Arctocephalus gazella*, *Arctocephalus philippii*, *Arctocephalus pusillus*, *Arctocephalus townsendi*, *Arctocephalus tropicalis*, *Arctodus simus*, *Arctonyx collaris*, *Atelocynus microtis*, *Bassaricyon alleni*, *Bassaricyon gabbii*, *Bassaricyon medius*, *Bassaricyon neblina*, *Bassariscus astutus*, *Bassariscus sumichrasti*, *Callorhinus ursinus*, *Canis aureus*, *Canis latrans*, *Canis lupaster*, *Canis lupus*, *Canis rufus*, *Canis simensis*, *Cerdocyon thous*, *Chrysocyon brachyurus*, *Conepatus chinga*, *Conepatus leuconotus*, *Conepatus semistriatus*, *Cuon alpinus*, *Cystophora cristata*, *Dusicyon australis*, *Eira barbara*, *Enhydra lutris*, *Erignathus barbatus*, *Eumetopias jubatus*, *Galictis cuja*, *Galictis vittata*, *Gulo gulo*, *Halichoerus grypus*, *Helarctos malayanus*, *Histriophoca fasciata*, *Hydrictis maculicollis*, *Hydrurga leptonyx*, *Ictonyx libyca*, *Ictonyx striatus*, *Leptonychotes weddellii*, *Lobodon carcinophaga*, *Lontra canadensis*, *Lontra felina*, *Lontra longicaudis*, *Lontra provocax*, *Lupulella adusta*, *Lupulella mesomelas*, *Lutra lutra*, *Lutra sumatrana*, *Lutrogale perspicillata*, *Lycalopex culpaeus*, *Lycalopex fulvipes*, *Lycalopex griseus*, *Lycalopex gymnocercus*, *Lycalopex sechurae*, *Lycalopex vetulus*, *Lycaon pictus*, *Lyncodon patagonicus*, *Martes americana*, *Martes caurina*, *Martes flavigula*, *Martes foina*, *Martes martes*, *Martes melampus*, *Martes zibellina*, *Meles anakuma*, *Meles leucurus*, *Meles meles*, *Mellivora capensis*, *Melogale cucphuongensis*, *Melogale moschata*, *Melogale personata*, *Melursus ursinus*, *Mephitis macroura*, *Mephitis mephitis*, *Mirounga angustirostris*, *Mirounga leonina*, *Monachus monachus*, *Mustela altaica*, *Mustela erminea*, *Mustela eversmanni*, *Mustela itatsi*, *Mustela kathiah*, *Mustela lutreola*, *Mustela nigripes*, *Mustela nivalis*, *Mustela nudipes*, *Mustela putorius*, *Mustela sibirica*, *Mustela strigidorsa*, *Mydaus javanensis*, *Mydaus marchei*, *Nasua narica*, *Nasua nasua*, *Nasuella meridensis*, *Nasuella olivacea*, *Neogale africana*, *Neogale felipei*, *Neogale frenata*, *Neogale vison*, *Neomonachus schauinslandi*, *Neomonachus tropicalis*, *Neophoca cinerea*, *Nyctereutes procyonoides*, *Odobenus rosmarus*, *Ommatophoca rossii*, *Otaria byronia*, *Otaria flavescens*, *Otocyon megalotis*, *Pagophilus groenlandicus*, *Pekania pennanti*, *Phoca largha*, *Phoca vitulina*, *Phocarctos hookeri*, *Poecilogale albinucha*, *Potos flavus*, *Procyon cancrivorus*, *Procyon lotor*, *Pteronura brasiliensis*, *Pusa caspica*, *Pusa hispida*, *Pusa sibirica*, *Speothos venaticus*, *Spilogale angustifrons*, *Spilogale gracilis*, *Spilogale interrupta*, *Spilogale leucoparia*, *Spilogale putorius*, *Spilogale pygmaea*, *Spilogale* sp., *Spilogale yucatanensis*, *Taxidea taxus*, *Tremarctos ornatus*, *Urocyon cinereoargenteus*, *Urocyon littoralis*, *Ursus americanus*, *Ursus arctos*, *Ursus maritimus*, *Ursus spelaeus*, *Ursus thibetanus*, *Vormela peregusna*, *Vulpes bengalensis*, *Vulpes cana*,

Table A.14 – *Continued from previous page*

Clade	Rank	Included species
		Vulpes chama, *Vulpes corsac*, *Vulpes ferrilata*, *Vulpes lagopus*, *Vulpes macrotis*, *Vulpes pallida*, *Vulpes rueppellii*, *Vulpes velox*, *Vulpes vulpes*, *Vulpes zerda*, *Zalophus californianus*, *Zalophus wollebaeki*
Feliformia	Suborder	*Acinonyx jubatus*, *Arctictis binturong*, *Arctogalidia trivirgata*, *Caracal aurata*, *Caracal caracal*, *Catopuma badia*, *Catopuma temminckii*, *Chrotogale owstoni*, *Civettictis civetta*, *Crocuta crocuta*, *Crossarchus obscurus*, *Cryptoprocta ferox*, *Cynictis pencillata*, *Cynogale bennetti*, *Felis bieti*, *Felis chaus*, *Felis lybica*, *Felis margarita*, *Felis nigripes*, *Felis silvestris*, *Fossa fossana*, *Galidia elegans*, *Genetta angolensis*, *Genetta bourloni*, *Genetta felina*, *Genetta genetta*, *Genetta johnstoni*, *Genetta maculata*, *Genetta pardina*, *Genetta piscivora*, *Genetta poensis*, *Genetta schoutedeni*, *Genetta servalina*, *Genetta thierryi*, *Genetta tigrina*, *Genetta victoriae*, *Helogale parvula*, *Hemigalus derbyanus*, *Herpailurus yagouaroundi*, *Herpestes edwardsii*, *Herpestes ichneumon*, *Herpestes javanicus*, *Homotherium latidens*, *Hyaena hyaena*, *Ichneumia albicauda*, *Leopardus braccatus*, *Leopardus colocola*, *Leopardus garleppi*, *Leopardus geoffroyi*, *Leopardus guigna*, *Leopardus guttulus*, *Leopardus jacobita*, *Leopardus munoai*, *Leopardus pajeros*, *Leopardus pardalis*, *Leopardus tigrinus*, *Leopardus wiedii*, *Leptailurus serval*, *Lynx canadensis*, *Lynx lynx*, *Lynx pardinus*, *Lynx rufus*, *Mungos mungo*, *Mungotictis decemlineata*, *Nandinia binotata*, *Neofelis diardi*, *Neofelis nebulosa*, *Otocolobus manul*, *Paguma larvata*, *Panthera leo*, *Panthera onca*, *Panthera pardus*, *Panthera tigris*, *Panthera uncia*, *Paradoxurus hermaphroditus*, *Paradoxurus jerdoni*, *Parahyaena brunnea*, *Pardofelis marmorata*, *Poiana richardsonii*, *Prionailurus bengalensis*, *Prionailurus planiceps*, *Prionailurus rubiginosus*, *Prionailurus viverrinus*, *Prionodon linsang*, *Prionodon pardicolor*, *Proteles cristatus*, *Puma concolor*, *Rhynchogale melleri*, *Smilodon populator*, *Suricata suricatta*, *Viverra megaspila*, *Viverra tangalunga*, *Viverra zibetha*, *Viverricula indica*
Arctoidea	Infraorder	*Ailuropoda melanoleuca*, *Ailurus fulgens*, *Aonyx capensis*, *Aonyx cinerea*, *Arctocephalus australis*, *Arctocephalus forsteri*, *Arctocephalus galapagoensis*, *Arctocephalus gazella*, *Arctocephalus philippii*, *Arctocephalus pusillus*, *Arctocephalus townsendi*, *Arctocephalus tropicalis*, *Arctodus simus*, *Arctonyx collaris*, *Bassaricyon alleni*, *Bassaricyon gabbii*, *Bassaricyon medius*, *Bassaricyon neblina*, *Bassariscus astutus*, *Bassariscus sumichrasti*, *Callorhinus ursinus*, *Conepatus chinga*, *Conepatus leuconotus*, *Conepatus semistriatus*, *Cystophora cristata*, *Eira barbara*, *Enhydra lutris*, *Erignathus barbatus*, *Eumetopias jubatus*, *Galictis cuja*, *Galictis vittata*, *Gulo gulo*, *Halichoerus grypus*, *Helarctos malayanus*, *Histriophoca fasciata*, *Hydrictis maculicollis*, *Hydrurga leptonyx*, *Ictonyx libyca*, *Ictonyx striatus*, *Leptonychotes weddellii*, *Lobodon carcinophaga*, *Lontra canadensis*, *Lontra felina*, *Lontra longicaudis*, *Lontra provocax*, *Lutra lutra*, *Lutra sumatrana*, *Lutrogale perspicillata*, *Lyncodon patagonicus*,

Table A.14 – *Continued from previous page*

Clade	Rank	Included species
		Martes americana, *Martes caurina*, *Martes flavigula*, *Martes foina*, *Martes martes*, *Martes melampus*, *Martes zibellina*, *Meles anakuma*, *Meles leucurus*, *Meles meles*, *Mellivora capensis*, *Melogale cucphuongensis*, *Melogale moschata*, *Melogale personata*, *Melursus ursinus*, *Mephitis macroura*, *Mephitis mephitis*, *Mirounga angustirostris*, *Mirounga leonina*, *Monachus monachus*, *Mustela altaica*, *Mustela erminea*, *Mustela eversmanni*, *Mustela itatsi*, *Mustela kathiah*, *Mustela lutreola*, *Mustela nigripes*, *Mustela nivalis*, *Mustela nudipes*, *Mustela putorius*, *Mustela sibirica*, *Mustela strigidorsa*, *Mydaus javanensis*, *Mydaus marchei*, *Nasua narica*, *Nasua nasua*, *Nasuella meridensis*, *Nasuella olivacea*, *Neogale africana*, *Neogale felipei*, *Neogale frenata*, *Neogale vison*, *Neomonachus schauinslandi*, *Neomonachus tropicalis*, *Neophoca cinerea*, *Odobenus rosmarus*, *Ommatophoca rossii*, *Otaria byronia*, *Otaria flavescens*, *Pagophilus groenlandicus*, *Pekania pennanti*, *Phoca largha*, *Phoca vitulina*, *Phocarctos hookeri*, *Poecilogale albinucha*, *Potos flavus*, *Procyon cancrivorus*, *Procyon lotor*, *Pteronura brasiliensis*, *Pusa caspica*, *Pusa hispida*, *Pusa sibirica*, *Spilogale angustifrons*, *Spilogale gracilis*, *Spilogale interrupta*, *Spilogale leucoparia*, *Spilogale putorius*, *Spilogale pygmaea*, *Spilogale* sp., *Spilogale yucatanensis*, *Taxidea taxus*, *Tremarctos ornatus*, *Ursus americanus*, *Ursus arctos*, *Ursus maritimus*, *Ursus spelaeus*, *Ursus thibetanus*, *Vormela peregusna*, *Zalophus californianus*, *Zalophus wollebaeki*
Viverroidea	Infraorder	*Arctictis binturong*, *Arctogalidia trivirgata*, *Chrotogale owstoni*, *Civettictis civetta*, *Crocuta crocuta*, *Crossarchus obscurus*, *Cryptoprocta ferox*, *Cynictis pencillata*, *Cynogale bennetti*, *Fossa fossana*, *Galidia elegans*, *Genetta angolensis*, *Genetta bourloni*, *Genetta felina*, *Genetta genetta*, *Genetta johnstoni*, *Genetta maculata*, *Genetta pardina*, *Genetta piscivora*, *Genetta poensis*, *Genetta schoutedeni*, *Genetta servalina*, *Genetta thierryi*, *Genetta tigrina*, *Genetta victoriae*, *Helogale parvula*, *Hemigalus derbyanus*, *Herpestes edwardsii*, *Herpestes ichneumon*, *Herpestes javanicus*, *Hyaena hyaena*, *Ichneumia albicauda*, *Mungos mungo*, *Mungotictis decemlineata*, *Paguma larvata*, *Paradoxurus hermaphroditus*, *Paradoxurus jerdoni*, *Parahyaena brunnea*, *Poiana richardsonii*, *Proteles cristatus*, *Rhynchogale melleri*, *Suricata suricatta*, *Viverra megaspila*, *Viverra tangalunga*, *Viverra zibetha*, *Viverricula indica*
Feloidea	Superfamily	*Acinonyx jubatus*, *Caracal aurata*, *Caracal caracal*, *Catopuma badia*, *Catopuma temminckii*, *Felis bieti*, *Felis chaus*, *Felis lybica*, *Felis margarita*, *Felis nigripes*, *Felis silvestris*, *Herpailurus yagouaroundi*, *Homotherium latidens*, *Leopardus braccatus*, *Leopardus colocola*, *Leopardus garleppi*, *Leopardus geoffroyi*, *Leopardus guigna*, *Leopardus guttulus*, *Leopardus jacobita*, *Leopardus munoai*, *Leopardus pajeros*, *Leopardus pardalis*, *Leopardus tigrinus*, *Leopardus wiedii*, *Leptailurus serval*, *Lynx canadensis*, *Lynx lynx*, *Lynx pardinus*, *Lynx rufus*, *Neofelis diardi*, *Neofelis nebulosa*, *Otocolobus manul*, *Panthera leo*,

Table A.14 – *Continued from previous page*

Clade	Rank	Included species
		Panthera onca, *Panthera pardus*, *Panthera tigris*, *Panthera uncia*, *Pardofelis marmorata*, *Prionailurus bengalensis*, *Prionailurus planiceps*, *Prionailurus rubiginosus*, *Prionailurus viverrinus*, *Prionodon linsang*, *Prionodon pardicolor*, *Puma concolor*, *Smilodon populator*
Herpestoidea	Superfamily	*Crocuta crocuta*, *Crossarchus obscurus*, *Cryptoprocta ferox*, *Cynictis pencillata*, *Fossa fossana*, *Galidia elegans*, *Helogale parvula*, *Herpestes edwardsii*, *Herpestes ichneumon*, *Herpestes javanicus*, *Hyaena hyaena*, *Ichneumia albicauda*, *Mungos mungo*, *Mungotictis decemlineata*, *Parahyaena brunnea*, *Proteles cristatus*, *Rhynchogale melleri*, *Suricata suricatta*
Musteloidea	Superfamily	*Ailurus fulgens*, *Aonyx capensis*, *Aonyx cinerea*, *Arctonyx collaris*, *Bassaricyon alleni*, *Bassaricyon gabbii*, *Bassaricyon medius*, *Bassaricyon neblina*, *Bassariscus astutus*, *Bassariscus sumichrasti*, *Conepatus chinga*, *Conepatus leuconotus*, *Conepatus semistriatus*, *Eira barbara*, *Enhydra lutris*, *Galictis cuja*, *Galictis vittata*, *Gulo gulo*, *Hydrictis maculicollis*, *Ictonyx libyca*, *Ictonyx striatus*, *Lontra canadensis*, *Lontra felina*, *Lontra longicaudis*, *Lontra provocax*, *Lutra lutra*, *Lutra sumatrana*, *Lutrogale perspicillata*, *Lyncodon patagonicus*, *Martes americana*, *Martes caurina*, *Martes flavigula*, *Martes foina*, *Martes martes*, *Martes melampus*, *Martes zibellina*, *Meles anakuma*, *Meles leucurus*, *Meles meles*, *Mellivora capensis*, *Melogale cucphuongensis*, *Melogale moschata*, *Melogale personata*, *Mephitis macroura*, *Mephitis mephitis*, *Mustela altaica*, *Mustela erminea*, *Mustela eversmanni*, *Mustela itatsi*, *Mustela kathiah*, *Mustela lutreola*, *Mustela nigripes*, *Mustela nivalis*, *Mustela nudipes*, *Mustela putorius*, *Mustela sibirica*, *Mustela strigidorsa*, *Mydaus javanensis*, *Mydaus marchei*, *Nasua narica*, *Nasua nasua*, *Nasuella meridensis*, *Nasuella olivacea*, *Neogale africana*, *Neogale felipei*, *Neogale frenata*, *Neogale vison*, *Pekania pennanti*, *Poecilogale albinucha*, *Potos flavus*, *Procyon cancrivorus*, *Procyon lotor*, *Pteronura brasiliensis*, *Spilogale angustifrons*, *Spilogale gracilis*, *Spilogale interrupta*, *Spilogale leucoparia*, *Spilogale putorius*, *Spilogale pygmaea*, *Spilogale* sp., *Spilogale yucatanensis*, *Taxidea taxus*, *Vormela peregusna*
Pinnipedia	Superfamily	*Arctocephalus australis*, *Arctocephalus forsteri*, *Arctocephalus galapagoensis*, *Arctocephalus gazella*, *Arctocephalus philippii*, *Arctocephalus pusillus*, *Arctocephalus townsendi*, *Arctocephalus tropicalis*, *Callorhinus ursinus*, *Cystophora cristata*, *Erignathus barbatus*, *Eumetopias jubatus*, *Halichoerus grypus*, *Histriophoca fasciata*, *Hydrurga leptonyx*, *Leptonychotes weddellii*, *Lobodon carcinophaga*, *Mirounga angustirostris*, *Mirounga leonina*, *Monachus monachus*, *Neomonachus schauinslandi*, *Neomonachus tropicalis*, *Neophoca cinerea*, *Odobenus rosmarus*, *Ommatophoca rossii*, *Otaria byronia*, *Otaria flavescens*, *Pagophilus groenlandicus*, *Phoca largha*, *Phoca vitulina*, *Phocarctos hookeri*, *Pusa caspica*, *Pusa hispida*, *Pusa sibirica*, *Zalophus californianus*, *Zalophus wollebaeki*

Table A.14 – *Continued from previous page*

Clade	Rank	Included species
Canidae	Family	*Aenocyon dirus*, *Atelocynus microtis*, *Canis aureus*, *Canis latrans*, *Canis lupaster*, *Canis lupus*, *Canis rufus*, *Canis simensis*, *Cerdocyon thous*, *Chrysocyon brachyurus*, *Cuon alpinus*, *Dusicyon australis*, *Lupulella adusta*, *Lupulella mesomelas*, *Lycalopex culpaeus*, *Lycalopex fulvipes*, *Lycalopex griseus*, *Lycalopex gymnocercus*, *Lycalopex sechurae*, *Lycalopex vetulus*, *Lycaon pictus*, *Nyctereutes procyonoides*, *Otocyon megalotis*, *Speothos venaticus*, *Urocyon cinereoargenteus*, *Urocyon littoralis*, *Vulpes bengalensis*, *Vulpes cana*, *Vulpes chama*, *Vulpes corsac*, *Vulpes ferrilata*, *Vulpes lagopus*, *Vulpes macrotis*, *Vulpes pallida*, *Vulpes rueppellii*, *Vulpes velox*, *Vulpes vulpes*, *Vulpes zerda*
Eupleridae	Family	*Cryptoprocta ferox*, *Fossa fossana*, *Galidia elegans*, *Mungotictis decemlineata*
Felidae	Family	*Acinonyx jubatus*, *Caracal aurata*, *Caracal caracal*, *Catopuma badia*, *Catopuma temminckii*, *Felis bieti*, *Felis chaus*, *Felis lybica*, *Felis margarita*, *Felis nigripes*, *Felis silvestris*, *Herpailurus yagouaroundi*, *Homotherium latidens*, *Leopardus braccatus*, *Leopardus colocola*, *Leopardus garleppi*, *Leopardus geoffroyi*, *Leopardus guigna*, *Leopardus guttulus*, *Leopardus jacobita*, *Leopardus munoai*, *Leopardus pajeros*, *Leopardus pardalis*, *Leopardus tigrinus*, *Leopardus wiedii*, *Leptailurus serval*, *Lynx canadensis*, *Lynx lynx*, *Lynx pardinus*, *Lynx rufus*, *Neofelis diardi*, *Neofelis nebulosa*, *Otocolobus manul*, *Panthera leo*, *Panthera onca*, *Panthera pardus*, *Panthera tigris*, *Panthera uncia*, *Pardofelis marmorata*, *Prionailurus bengalensis*, *Prionailurus planiceps*, *Prionailurus rubiginosus*, *Prionailurus viverrinus*, *Puma concolor*, *Smilodon populator*
Herpestidae	Family	*Crossarchus obscurus*, *Cynictis pencillata*, *Helogale parvula*, *Herpestes edwardsii*, *Herpestes ichneumon*, *Herpestes javanicus*, *Ichneumia albicauda*, *Mungos mungo*, *Rhynchogale melleri*, *Suricata suricatta*
Hyaenidae	Family	*Crocuta crocuta*, *Hyaena hyaena*, *Parahyaena brunnea*, *Proteles cristatus*
Mephitidae	Family	*Conepatus chinga*, *Conepatus leuconotus*, *Conepatus semistriatus*, *Mephitis macroura*, *Mephitis mephitis*, *Mydaus javanensis*, *Mydaus marchei*, *Spilogale angustifrons*, *Spilogale gracilis*, *Spilogale interrupta*, *Spilogale leucoparia*, *Spilogale putorius*, *Spilogale pygmaea*, *Spilogale* sp., *Spilogale yucatanensis*
Mustelidae	Family	*Aonyx capensis*, *Aonyx cinerea*, *Arctonyx collaris*, *Eira barbara*, *Enhydra lutris*, *Galictis cuja*, *Galictis vittata*, *Gulo gulo*, *Hydrictis maculicollis*, *Ictonyx libyca*, *Ictonyx striatus*, *Lontra canadensis*, *Lontra felina*, *Lontra longicaudis*, *Lontra provocax*, *Lutra lutra*, *Lutra sumatrana*, *Lutrogale perspicillata*, *Lyncodon patagonicus*, *Martes americana*, *Martes caurina*, *Martes flavigula*, *Martes foina*, *Martes martes*, *Martes melampus*, *Martes zibellina*, *Meles anakuma*, *Meles leucurus*, *Meles meles*, *Mellivora capensis*, *Melogale cucphuongensis*, *Melogale moschata*, *Melogale personata*, *Mustela altaica*, *Mustela erminea*, *Mustela eversmanni*, *Mustela itatsi*, *Mustela kathiah*, *Mustela lutreola*,

Table A.14 – *Continued from previous page*

Clade	Rank	Included species
		Mustela nigripes, *Mustela nivalis*, *Mustela nudipes*, *Mustela putorius*, *Mustela sibirica*, *Mustela strigidorsa*, *Neogale africana*, *Neogale felipei*, *Neogale frenata*, *Neogale vison*, *Pekania pennanti*, *Poecilogale albinucha*, *Pteronura brasiliensis*, *Taxidea taxus*, *Vormela peregusna*
Otariidae	Family	*Arctocephalus australis*, *Arctocephalus forsteri*, *Arctocephalus galapagoensis*, *Arctocephalus gazella*, *Arctocephalus philippii*, *Arctocephalus pusillus*, *Arctocephalus townsendi*, *Arctocephalus tropicalis*, *Callorhinus ursinus*, *Eumetopias jubatus*, *Neophoca cinerea*, *Otaria byronia*, *Otaria flavescens*, *Phocarctos hookeri*, *Zalophus californianus*, *Zalophus wollebaeki*
Phocidae	Family	*Cystophora cristata*, *Erignathus barbatus*, *Halichoerus grypus*, *Histriophoca fasciata*, *Hydrurga leptonyx*, *Leptonychotes weddellii*, *Lobodon carcinophaga*, *Mirounga angustirostris*, *Mirounga leonina*, *Monachus monachus*, *Neomonachus schauinslandi*, *Neomonachus tropicalis*, *Ommatophoca rossii*, *Pagophilus groenlandicus*, *Phoca largha*, *Phoca vitulina*, *Pusa caspica*, *Pusa hispida*, *Pusa sibirica*
Prionodontidae	Family	*Prionodon linsang*, *Prionodon pardicolor*
Procyonidae	Family	*Bassaricyon alleni*, *Bassaricyon gabbii*, *Bassaricyon medius*, *Bassaricyon neblina*, *Bassariscus astutus*, *Bassariscus sumichrasti*, *Nasua narica*, *Nasua nasua*, *Nasuella meridensis*, *Nasuella olivacea*, *Potos flavus*, *Procyon cancrivorus*, *Procyon lotor*
Ursidae	Family	*Ailuropoda melanoleuca*, *Arctodus simus*, *Helarctos malayanus*, *Melursus ursinus*, *Tremarctos ornatus*, *Ursus americanus*, *Ursus arctos*, *Ursus maritimus*, *Ursus spelaeus*, *Ursus thibetanus*
Viverridae	Family	*Arctictis binturong*, *Arctogalidia trivirgata*, *Chrotogale owstoni*, *Civettictis civetta*, *Cynogale bennetti*, *Genetta angolensis*, *Genetta bourloni*, *Genetta felina*, *Genetta genetta*, *Genetta johnstoni*, *Genetta maculata*, *Genetta pardina*, *Genetta piscivora*, *Genetta poensis*, *Genetta schoutedeni*, *Genetta servalina*, *Genetta thierryi*, *Genetta tigrina*, *Genetta victoriae*, *Hemigalus derbyanus*, *Paguma larvata*, *Paradoxurus hermaphroditus*, *Paradoxurus jerdoni*, *Poiana richardsonii*, *Viverra megaspila*, *Viverra tangalunga*, *Viverra zibetha*, *Viverricula indica*
Caninae	Subfamily	*Aenocyon dirus*, *Atelocynus microtis*, *Canis aureus*, *Canis latrans*, *Canis lupaster*, *Canis lupus*, *Canis rufus*, *Canis simensis*, *Cerdocyon thous*, *Chrysocyon brachyurus*, *Cuon alpinus*, *Dusicyon australis*, *Lupulella adusta*, *Lupulella mesomelas*, *Lycalopex culpaeus*, *Lycalopex fulvipes*, *Lycalopex griseus*, *Lycalopex gymnocercus*, *Lycalopex sechurae*, *Lycalopex vetulus*, *Lycaon pictus*, *Nyctereutes procyonoides*, *Otocyon megalotis*, *Speothos venaticus*, *Urocyon cinereoargenteus*, *Urocyon littoralis*, *Vulpes bengalensis*, *Vulpes cana*, *Vulpes chama*, *Vulpes corsac*, *Vulpes ferrilata*, *Vulpes lagopus*, *Vulpes macrotis*, *Vulpes pallida*, *Vulpes rueppellii*, *Vulpes velox*, *Vulpes vulpes*, *Vulpes zerda*
Euplerinae	Subfamily	*Cryptoprocta ferox*, *Fossa fossana*

Table A.14 – *Continued from previous page*

Clade	Rank	Included species
Felinae	Subfamily	*Acinonyx jubatus*, *Caracal aurata*, *Caracal caracal*, *Catopuma badia*, *Catopuma temminckii*, *Felis bieti*, *Felis chaus*, *Felis lybica*, *Felis margarita*, *Felis nigripes*, *Felis silvestris*, *Herpailurus yagouaroundi*, *Leopardus braccatus*, *Leopardus colocola*, *Leopardus garleppi*, *Leopardus geoffroyi*, *Leopardus guigna*, *Leopardus guttulus*, *Leopardus jacobita*, *Leopardus munoai*, *Leopardus pajeros*, *Leopardus pardalis*, *Leopardus tigrinus*, *Leopardus wiedii*, *Leptailurus serval*, *Lynx canadensis*, *Lynx lynx*, *Lynx pardinus*, *Lynx rufus*, *Neofelis diardi*, *Neofelis nebulosa*, *Otocolobus manul*, *Panthera leo*, *Panthera onca*, *Panthera pardus*, *Panthera tigris*, *Panthera uncia*, *Pardofelis marmorata*, *Prionailurus bengalensis*, *Prionailurus planiceps*, *Prionailurus rubiginosus*, *Prionailurus viverrinus*, *Puma concolor*
Galidinae	Subfamily	*Galidia elegans*, *Mungotictis decemlineata*
Genettinae	Subfamily	*Genetta angolensis*, *Genetta bourloni*, *Genetta felina*, *Genetta genetta*, *Genetta johnstoni*, *Genetta maculata*, *Genetta pardina*, *Genetta piscivora*, *Genetta poensis*, *Genetta schoutedeni*, *Genetta servalina*, *Genetta thierryi*, *Genetta tigrina*, *Genetta victoriae*, *Poiana richardsonii*
Guloninae	Subfamily	*Eira barbara*, *Gulo gulo*, *Martes americana*, *Martes caurina*, *Martes flavigula*, *Martes foina*, *Martes martes*, *Martes melampus*, *Martes zibellina*, *Pekania pennanti*
Helictidinae	Subfamily	*Melogale cucphuongensis*, *Melogale moschata*, *Melogale personata*
Hemigalinae	Subfamily	*Chrotogale owstoni*, *Cynogale bennetti*, *Hemigalus derbyanus*
Herpestinae	Subfamily	*Cynictis pencillata*, *Herpestes edwardsii*, *Herpestes ichneumon*, *Herpestes javanicus*, *Ichneumia albicauda*, *Rhynchogale melleri*
Ictonychinae	Subfamily	*Galictis cuja*, *Galictis vittata*, *Ictonyx libyca*, *Ictonyx striatus*, *Lyncodon patagonicus*, *Poecilogale albinucha*, *Vormela peregusna*
Lutrinae	Subfamily	*Aonyx capensis*, *Aonyx cinerea*, *Enhydra lutris*, *Hydrictis maculicollis*, *Lontra canadensis*, *Lontra felina*, *Lontra longicaudis*, *Lontra provocax*, *Lutra lutra*, *Lutra sumatrana*, *Lutrogale perspicillata*, *Pteronura brasiliensis*
Machairodontinae	Subfamily	*Homotherium latidens*, *Smilodon populator*
Melinae	Subfamily	*Arctonyx collaris*, *Meles anakuma*, *Meles leucurus*, *Meles meles*
Monachinae	Subfamily	*Hydrurga leptonyx*, *Leptonychotes weddellii*, *Lobodon carcinophaga*, *Mirounga angustirostris*, *Mirounga leonina*, *Monachus monachus*, *Neomonachus schauinslandi*, *Neomonachus tropicalis*, *Ommatophoca rossii*
Mungotinae	Subfamily	*Crossarchus obscurus*, *Helogale parvula*, *Mungos mungo*, *Suricata suricatta*
Mustelinae	Subfamily	*Mustela altaica*, *Mustela erminea*, *Mustela eversmanni*, *Mustela itatsi*, *Mustela kathiah*, *Mustela lutreola*, *Mustela nigripes*, *Mustela nivalis*, *Mustela nudipes*, *Mustela putorius*, *Mustela sibirica*, *Mustela strigidorsa*, *Neogale africana*, *Neogale felipei*, *Neogale frenata*, *Neogale vison*
Paradoxurinae	Subfamily	*Arctictis binturong*, *Arctogalidia trivirgata*, *Paguma larvata*, *Paradoxurus hermaphroditus*, *Paradoxurus jerdoni*

Table A.14 – *Continued from previous page*

Clade	Rank	Included species
Phocinae	Subfamily	*Cystophora cristata*, *Erignathus barbatus*, *Halichoerus grypus*, *Histriophoca fasciata*, *Pagophilus groenlandicus*, *Phoca largha*, *Phoca vitulina*, *Pusa caspica*, *Pusa hispida*, *Pusa sibirica*
Ursinae	Subfamily	*Ailuropoda melanoleuca*, *Arctodus simus*, *Helarctos malayanus*, *Melursus ursinus*, *Tremarctos ornatus*, *Ursus americanus*, *Ursus arctos*, *Ursus maritimus*, *Ursus spelaeus*, *Ursus thibetanus*
Viverrinae	Subfamily	*Civettictis civetta*, *Viverra megaspila*, *Viverra tangalunga*, *Viverra zibetha*, *Viverricula indica*
Arctotheriini	Tribe	*Arctodus simus*, *Tremarctos ornatus*
Canini	Tribe	*Aenocyon dirus*, *Atelocynus microtis*, *Canis aureus*, *Canis latrans*, *Canis lupaster*, *Canis lupus*, *Canis rufus*, *Canis simensis*, *Cerdocyon thous*, *Chrysocyon brachyurus*, *Cuon alpinus*, *Dusicyon australis*, *Lupulella adusta*, *Lupulella mesomelas*, *Lycalopex culpaeus*, *Lycalopex fulvipes*, *Lycalopex griseus*, *Lycalopex gymnocercus*, *Lycalopex sechurae*, *Lycalopex vetulus*, *Lycaon pictus*, *Speothos venaticus*
Felini	Tribe	*Acinonyx jubatus*, *Caracal aurata*, *Caracal caracal*, *Catopuma badia*, *Catopuma temminckii*, *Felis bieti*, *Felis chaus*, *Felis lybica*, *Felis margarita*, *Felis nigripes*, *Felis silvestris*, *Herpailurus yagouaroundi*, *Leopardus braccatus*, *Leopardus colocola*, *Leopardus garleppi*, *Leopardus geoffroyi*, *Leopardus guigna*, *Leopardus guttulus*, *Leopardus jacobita*, *Leopardus munoai*, *Leopardus pajeros*, *Leopardus pardalis*, *Leopardus tigrinus*, *Leopardus wiedii*, *Leptailurus serval*, *Lynx canadensis*, *Lynx lynx*, *Lynx pardinus*, *Lynx rufus*, *Otocolobus manul*, *Pardofelis marmorata*, *Prionailurus bengalensis*, *Prionailurus planiceps*, *Prionailurus rubiginosus*, *Prionailurus viverrinus*, *Puma concolor*
Lobodontini	Tribe	*Hydrurga leptonyx*, *Leptonychotes weddellii*, *Lobodon carcinophaga*, *Ommatophoca rossii*
Miroungini	Tribe	*Mirounga angustirostris*, *Mirounga leonina*
Monachini	Tribe	*Monachus monachus*, *Neomonachus schauinslandi*, *Neomonachus tropicalis*
Pantherini	Tribe	*Neofelis diardi*, *Neofelis nebulosa*, *Panthera leo*, *Panthera onca*, *Panthera pardus*, *Panthera tigris*, *Panthera uncia*
Phocini	Tribe	*Halichoerus grypus*, *Histriophoca fasciata*, *Pagophilus groenlandicus*, *Phoca largha*, *Phoca vitulina*, *Pusa caspica*, *Pusa hispida*, *Pusa sibirica*
Ursini	Tribe	*Helarctos malayanus*, *Melursus ursinus*, *Ursus americanus*, *Ursus arctos*, *Ursus maritimus*, *Ursus spelaeus*, *Ursus thibetanus*
Vulpini	Tribe	*Nyctereutes procyonoides*, *Otocyon megalotis*, *Urocyon cinereoargenteus*, *Urocyon littoralis*, *Vulpes bengalensis*, *Vulpes cana*, *Vulpes chama*, *Vulpes corsac*, *Vulpes ferrilata*, *Vulpes lagopus*, *Vulpes macrotis*, *Vulpes pallida*, *Vulpes rueppellii*, *Vulpes velox*, *Vulpes vulpes*, *Vulpes zerda*
Canina	Subtribe	*Aenocyon dirus*, *Canis aureus*, *Canis latrans*, *Canis lupaster*, *Canis lupus*, *Canis rufus*, *Canis simensis*, *Cuon alpinus*, *Lupulella adusta*, *Lupulella mesomelas*, *Lycaon pictus*

Table A.14 – *Continued from previous page*

Clade	Rank	Included species
Cerdocyonina	Subtribe	*Atelocynus microtis*, *Cerdocyon thous*, *Chrysocyon brachyurus*, *Dusicyon australis*, *Lycalopex culpaeus*, *Lycalopex fulvipes*, *Lycalopex griseus*, *Lycalopex gymnocercus*, *Lycalopex sechurae*, *Lycalopex vetulus*, *Speothos venaticus*

Table A.15: Mean posterior root ages and topological accuracy (as quantified by the number of recovered benchmark clades from Table A.14) of the summary supertrees inferred with BLeSS under different analytical settings. MAP = maximum *a posteriori* tree; MCC = maximum clade credibility tree; HL run = MCMC run with the highest post-burnin likelihoods. Note that the MAP and MCC trees were only computed from the pooled posterior sample of all 4 MCMC runs if the average standard deviation of split frequencies did not exceed 0.01; otherwise, only the highest-likelihood run was summarized.

Imputation	Downweighting of mtDNA trees	λ_e	Summary tree	Root age (Ma)	# of benchmark clades recovered
No	No	Fixed to 0.1	MAP, all runs	49.301	16
No	No	Fixed to 0.1	MCC, all runs	49.301	16
No	No	Fixed to 0.1	MAP, HL run	49.301	16
No	No	Fixed to 0.1	MCC, HL run	49.301	17
No	Yes	Fixed to 0.1	MAP, all runs	49.658	31
No	Yes	Fixed to 0.1	MCC, all runs	49.658	19
No	Yes	Fixed to 0.1	MAP, HL run	56.118	31
No	Yes	Fixed to 0.1	MCC, HL run	56.118	31
Yes	No	Fixed to 0.1	MAP, HL run	51.265	45
Yes	No	Fixed to 0.1	MCC, HL run	51.265	45
Yes	Yes	Fixed to 0.1	MAP, HL run	51.265	45
Yes	Yes	Fixed to 0.1	MCC, HL run	51.265	46
Yes	Yes	Estimated	MAP, all runs	43.374	0
Yes	Yes	Estimated	MCC, all runs	43.374	0

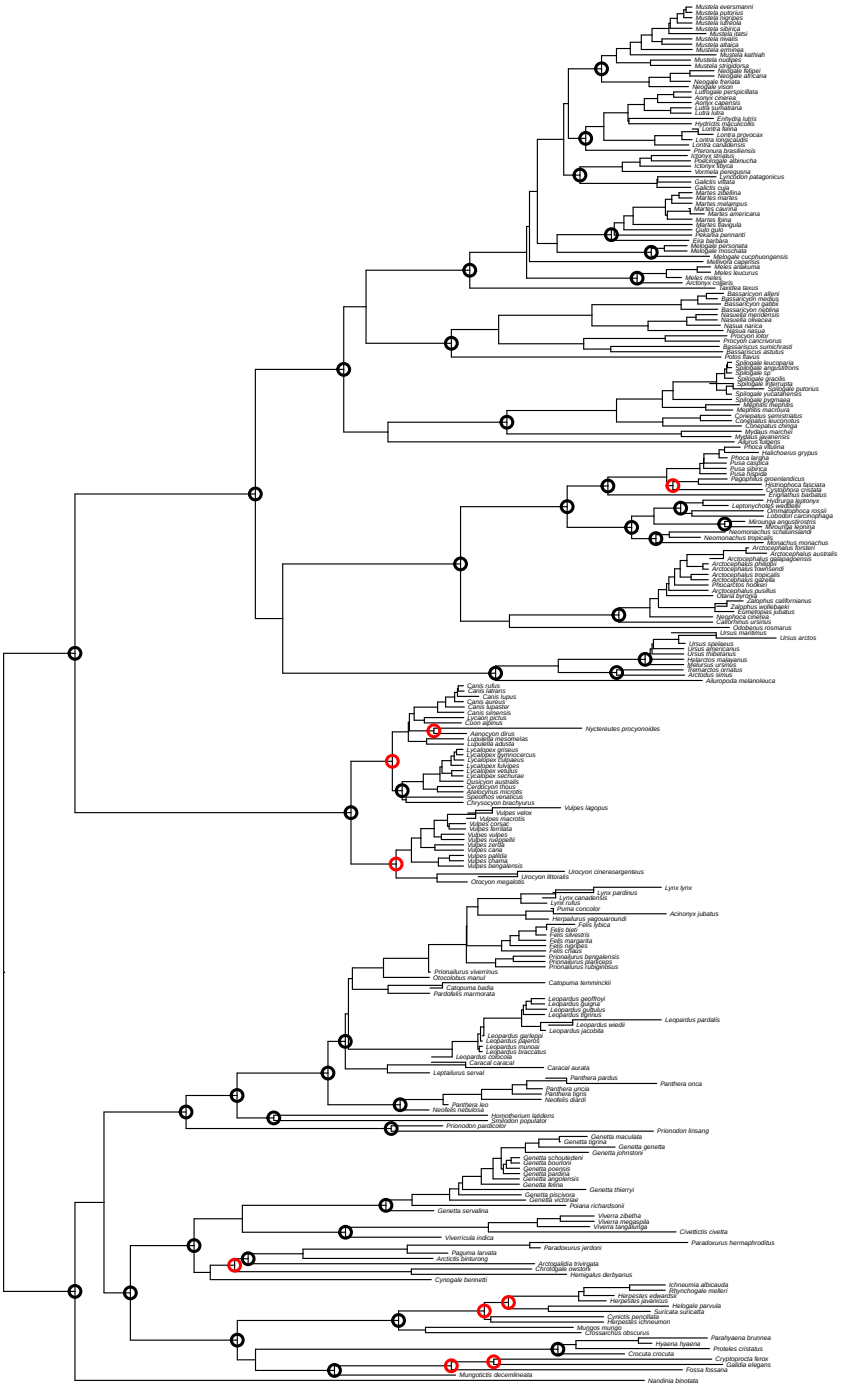

Figure A.32: Neighbor-joining supertree of Carnivora inferred from an average distance matrix with imputed missing entries and all source trees weighted equally. The circles indicate nodes congruent with (black) or contradicting (red) the 53 predefined benchmark clades. Note that the tree is not ultrametric, rendering the branch lengths uninterpretable.

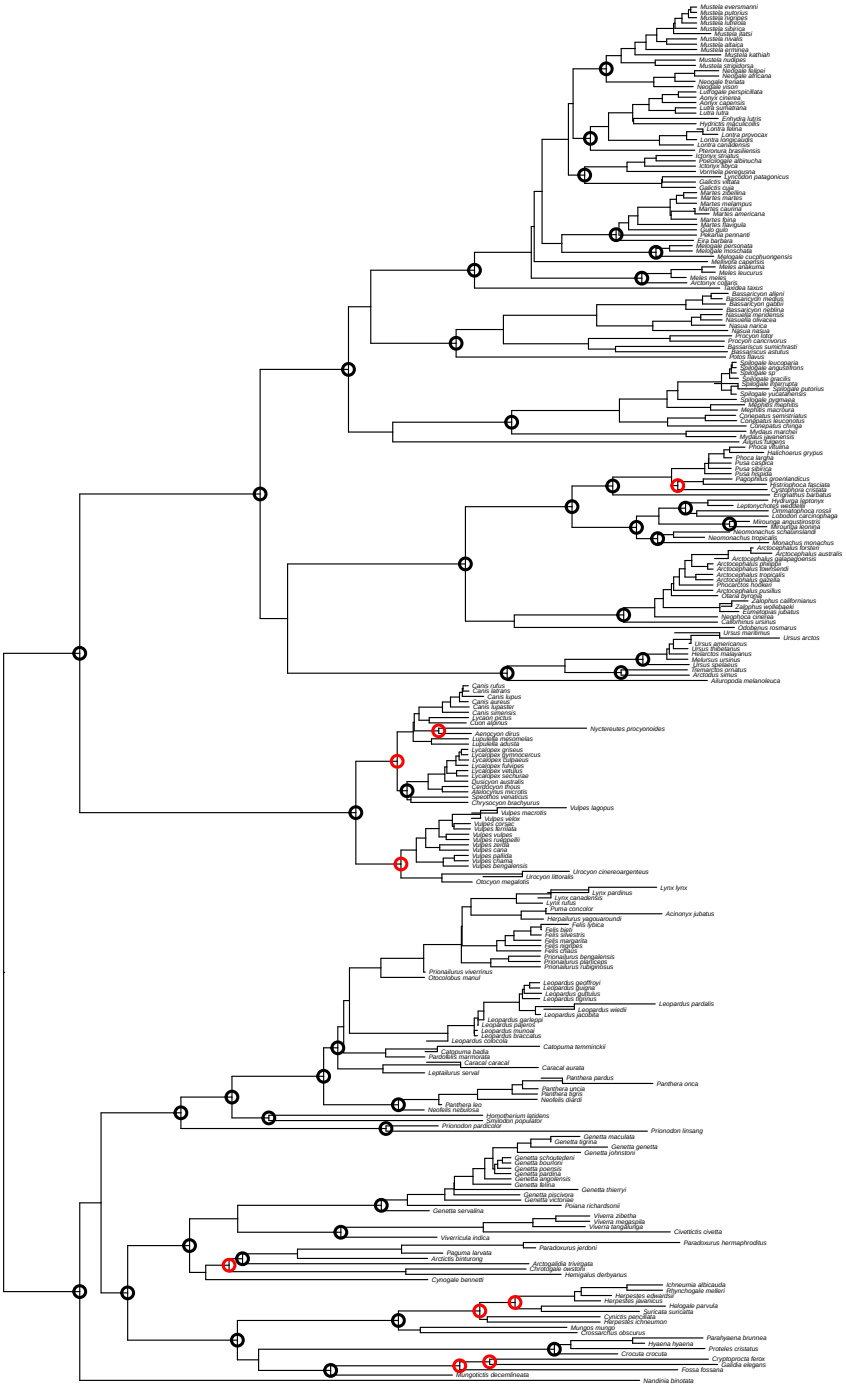

Figure A.33: Neighbor-joining supertree of Carnivora inferred from an average distance matrix with imputed missing entries and with mtDNA-based source trees downweighted by a factor of 10. The circles indicate nodes congruent with (black) or contradicting (red) the 53 predefined benchmark clades. Note that the tree is not ultrametric, rendering the branch lengths uninterpretable.

A.12 Additional Figures

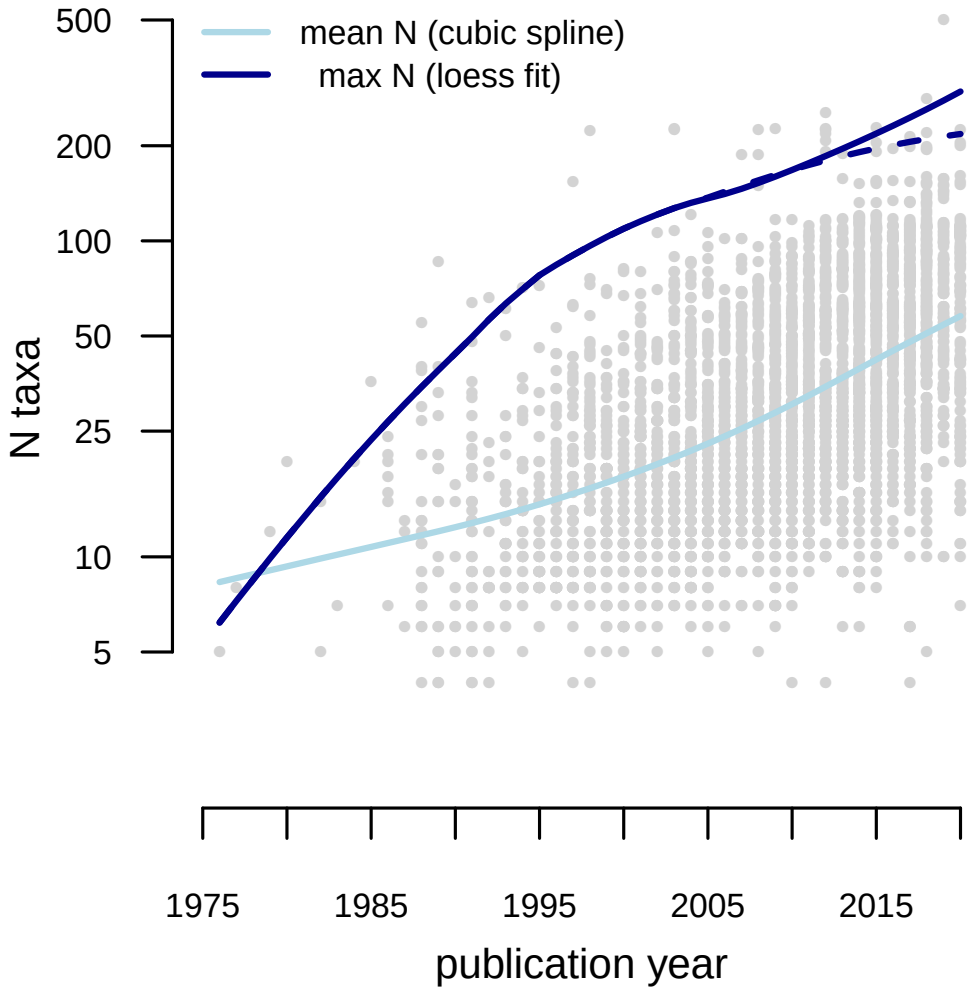

Figure A.34: Number of taxa (N) included in morphological phylogenetic datasets over time, based on 3671 studies published between 1975 and 2020 (inclusive) and sourced from `https://graemetlloyd.com/matr.html`. A cubic spline (light blue line) fitted to $\log N$ shows that the average dataset only increased from 8.3 to 58 taxa during this period (note the log scale on the y–axis). A loess fit (dark blue solid line) shows that the maximum number of taxa increased as well, but a marked slow-down in the rate of increase becomes apparent after the removal of a single outlier (Hartman et al., 2019; dark blue dashed line), which contains twice as many taxa as any other study published during the analyzed period. Figure reprinted with permission from Lloyd and Slater (2020, Fig. 1), the preprint version of a study later published as Lloyd and Slater (2021).

APPENDIX B: SUPPLEMENTARY INFORMATION FOR CHAPTER TWO

Table B.1: Metrics computed to evaluate the difficulty of estimating higher-level dinosaur phylogeny from the Baron et al. (BEA) and Langer et al. (LEA) datasets. RF distance = Robinson-Foulds distance.

Metric	BEA	LEA
# of trees (n_{all})	100	100
# of plausible trees (n_{pl})	67	51
Avg. pairwise normalized RF distance, all trees ($\bar{d}_{\text{all}}$)	0.3296	0.3952
Avg. pairwise normalize RF distance, plausible trees ($\bar{d}_{\text{pl}}$)	0.3203	0.3618
# of unique topologies, all trees (n'_{all})	100	100
# of unique topologies, plausible trees (n'_{pl})	67	51

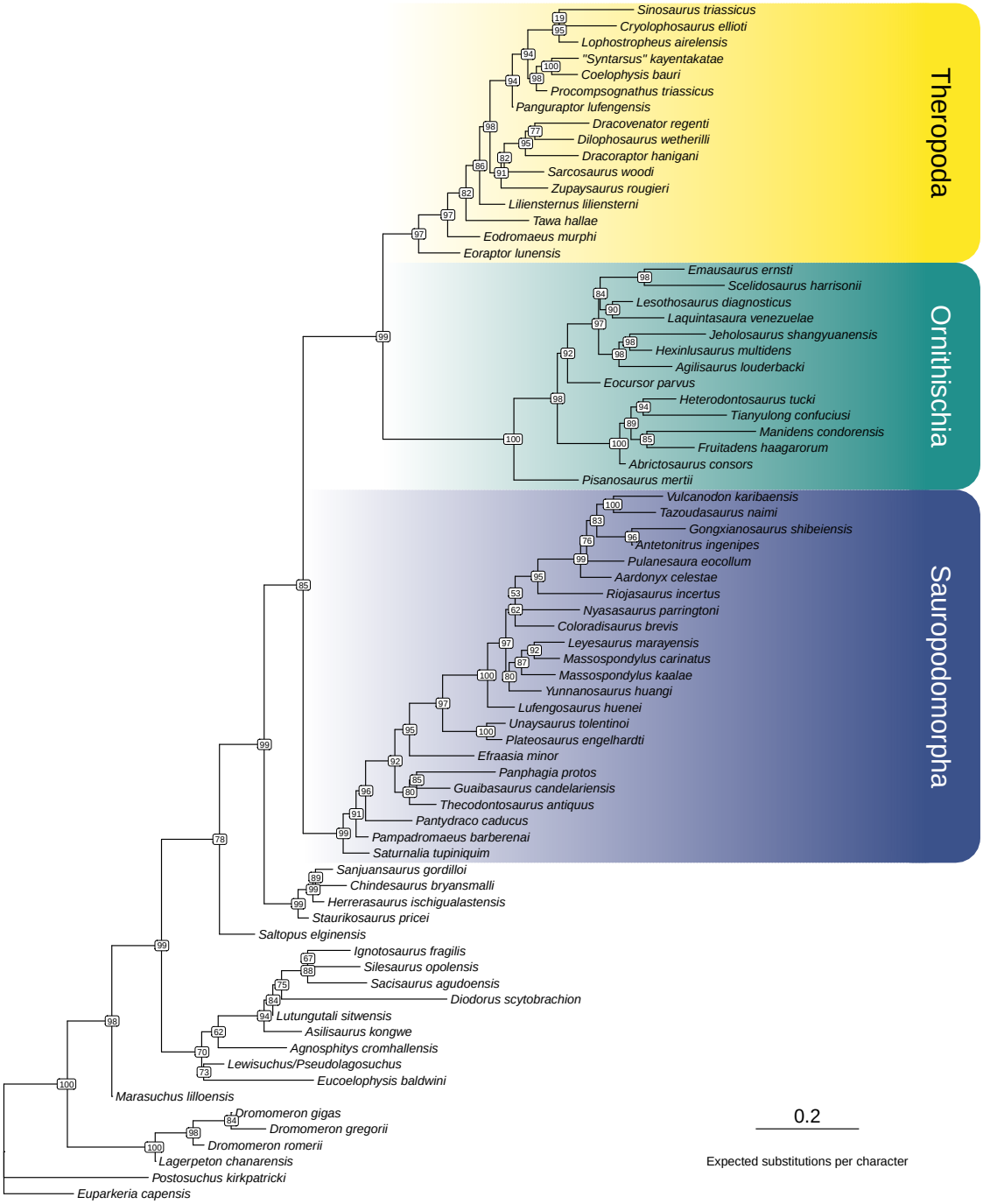

Figure B.1: Unconstrained maximum likelihood tree inferred for the original BEA matrix ($\ln L = -6367.975$). Node labels denote ultrafast bootstrap values computed from 1000 replicates.

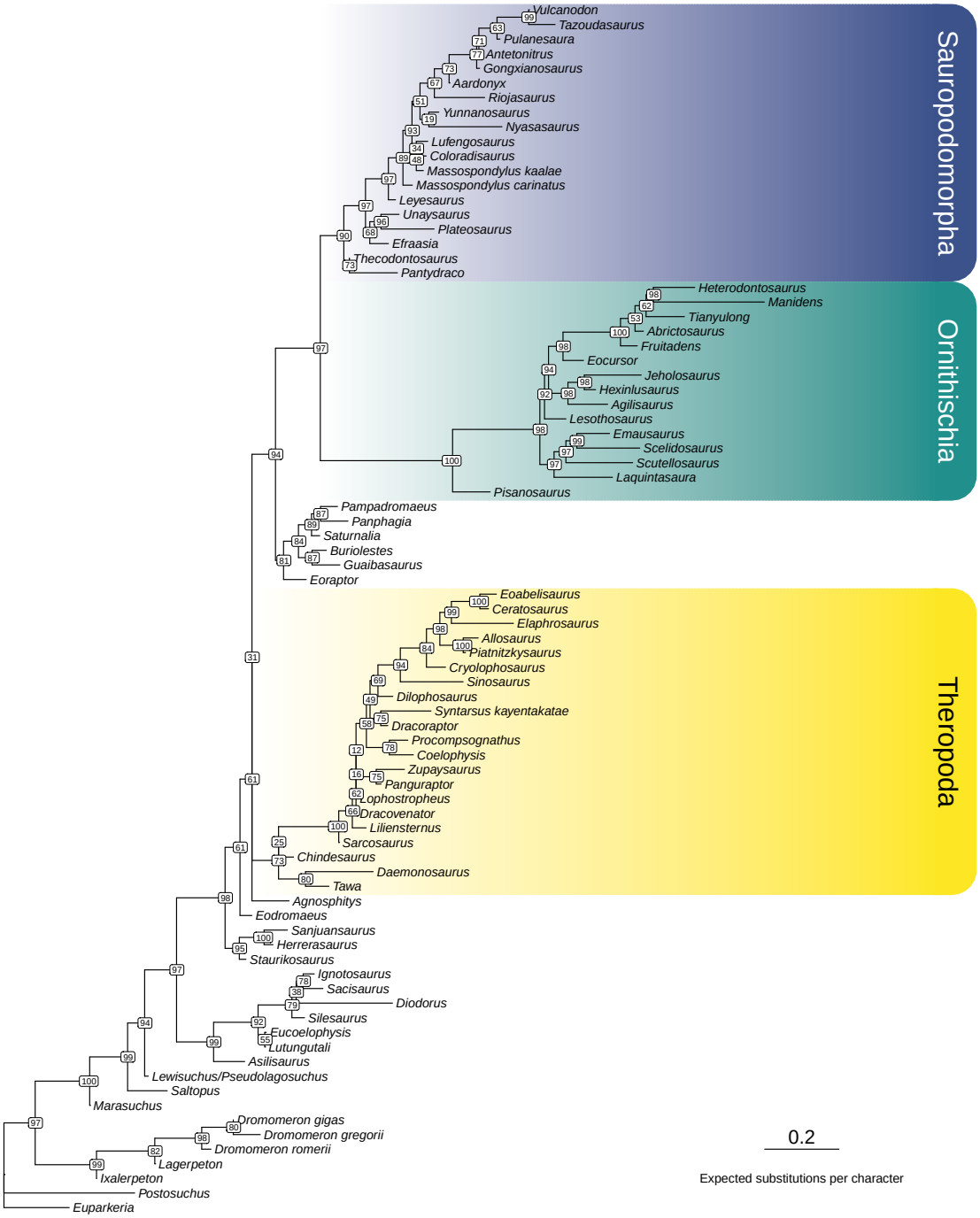

Figure B.2: Unconstrained maximum likelihood tree inferred for the original LEA matrix ($\ln L = -7007.117$). Node labels denote ultrafast bootstrap values computed from 1000 replicates.

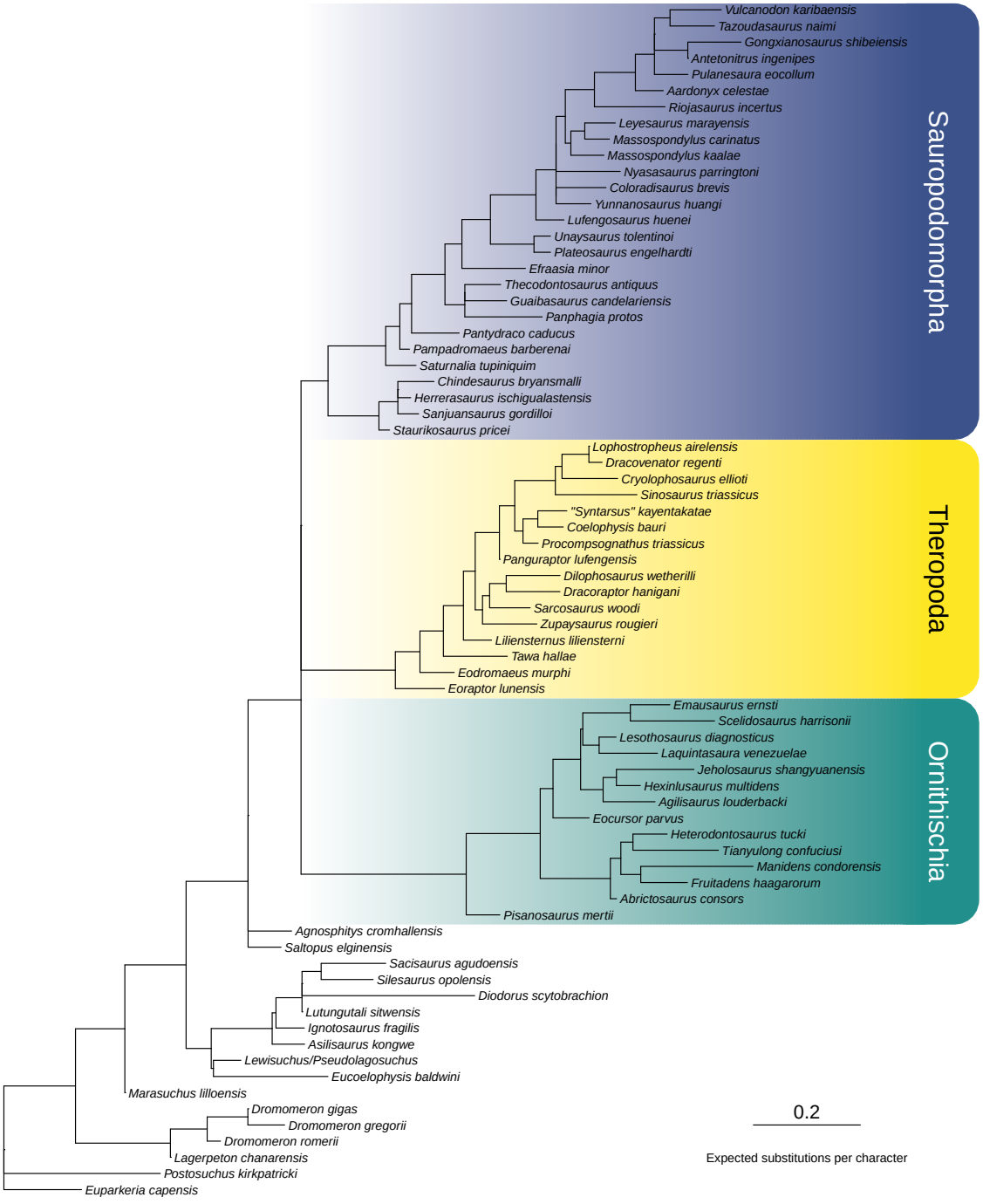

Figure B.3: Maximum likelihood tree inferred for the original BEA matrix and constrained to recover Saurischia ($\ln L = -6376.854$).

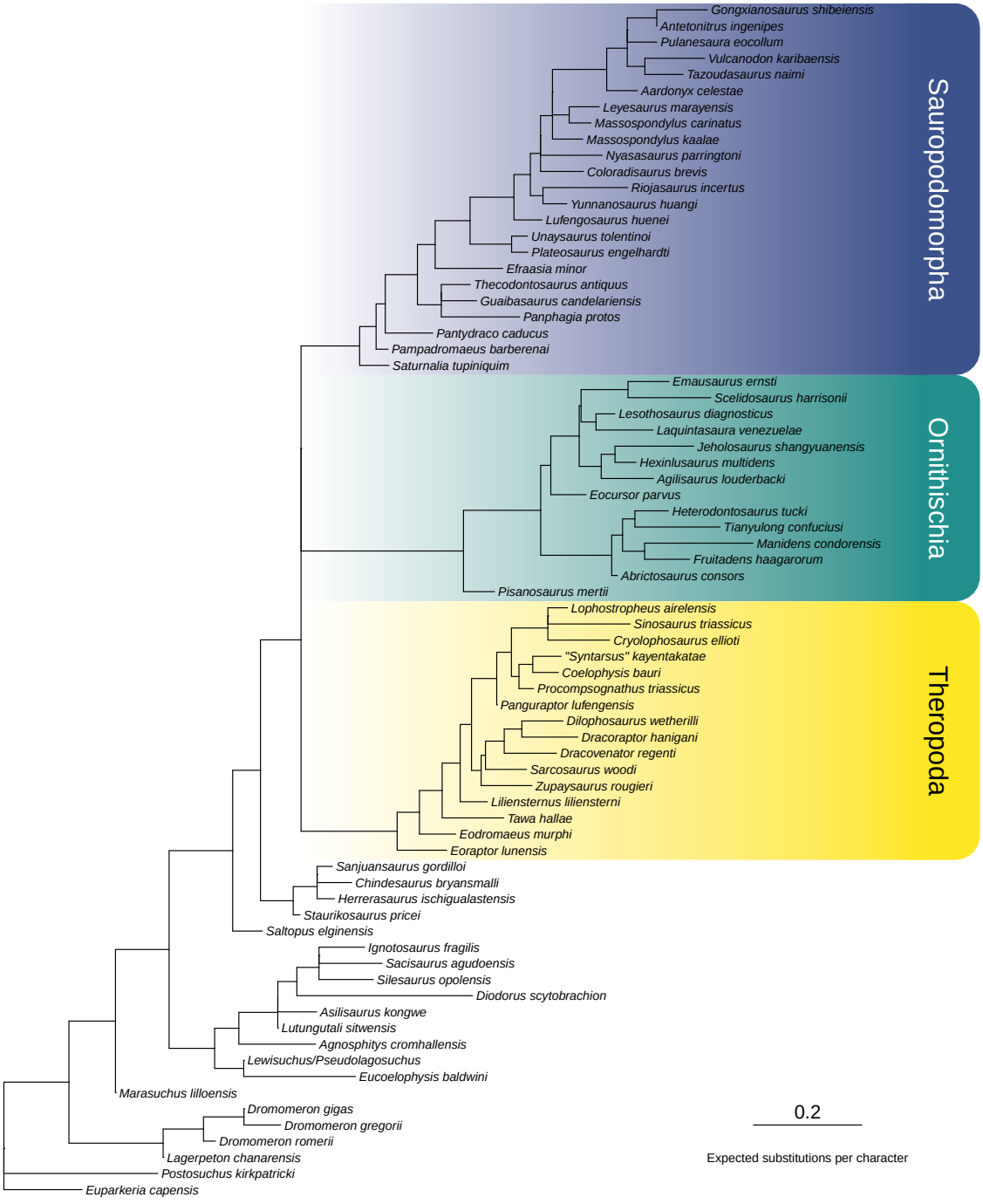

Figure B.4: Maximum likelihood tree inferred for the original BEA matrix and constrained to recover Ornithischiformes ($\ln L = -6377.425$).

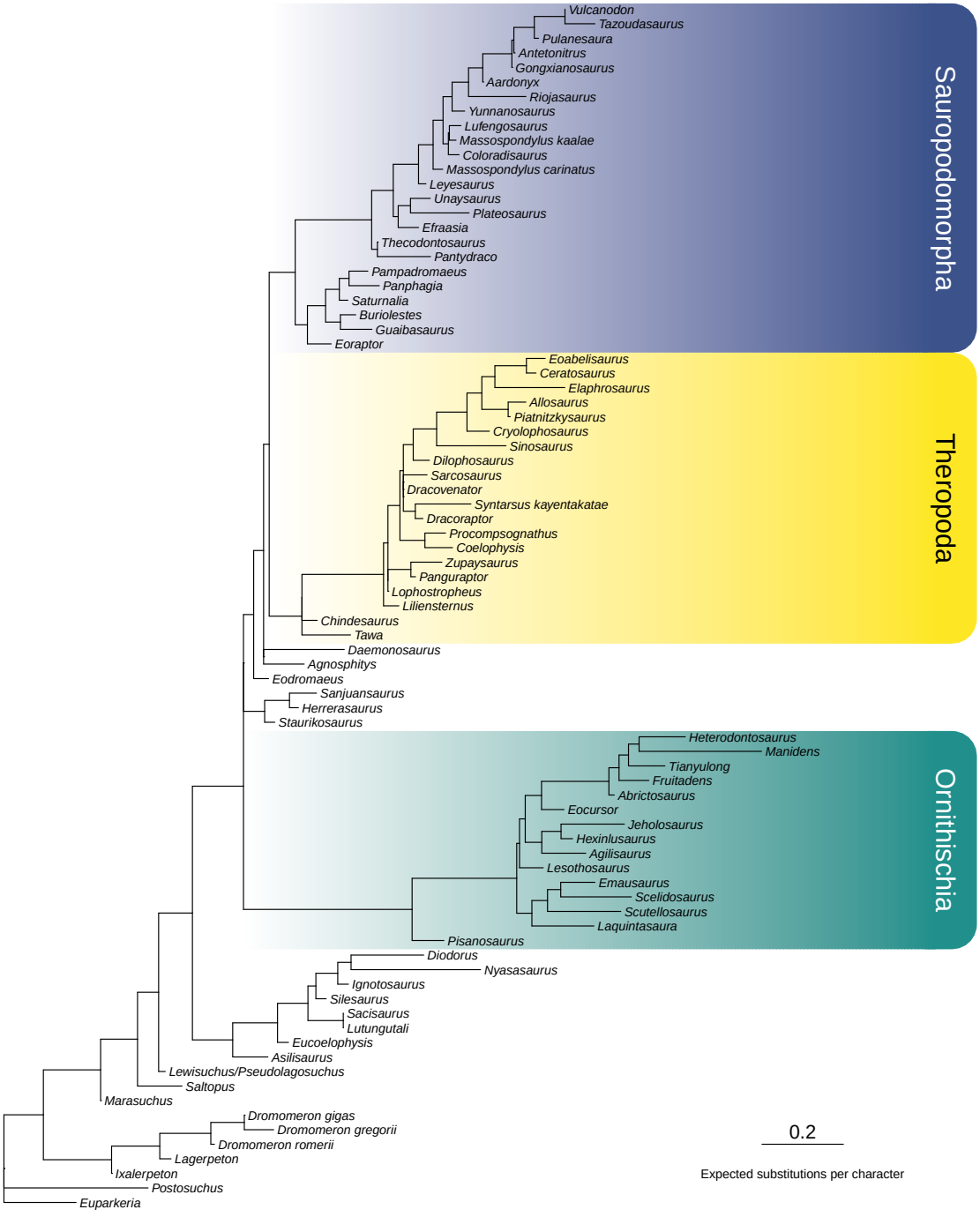

Figure B.5: Maximum likelihood tree inferred for the original LEA matrix and constrained to recover Saurischia ($\ln L = -7017.246$).

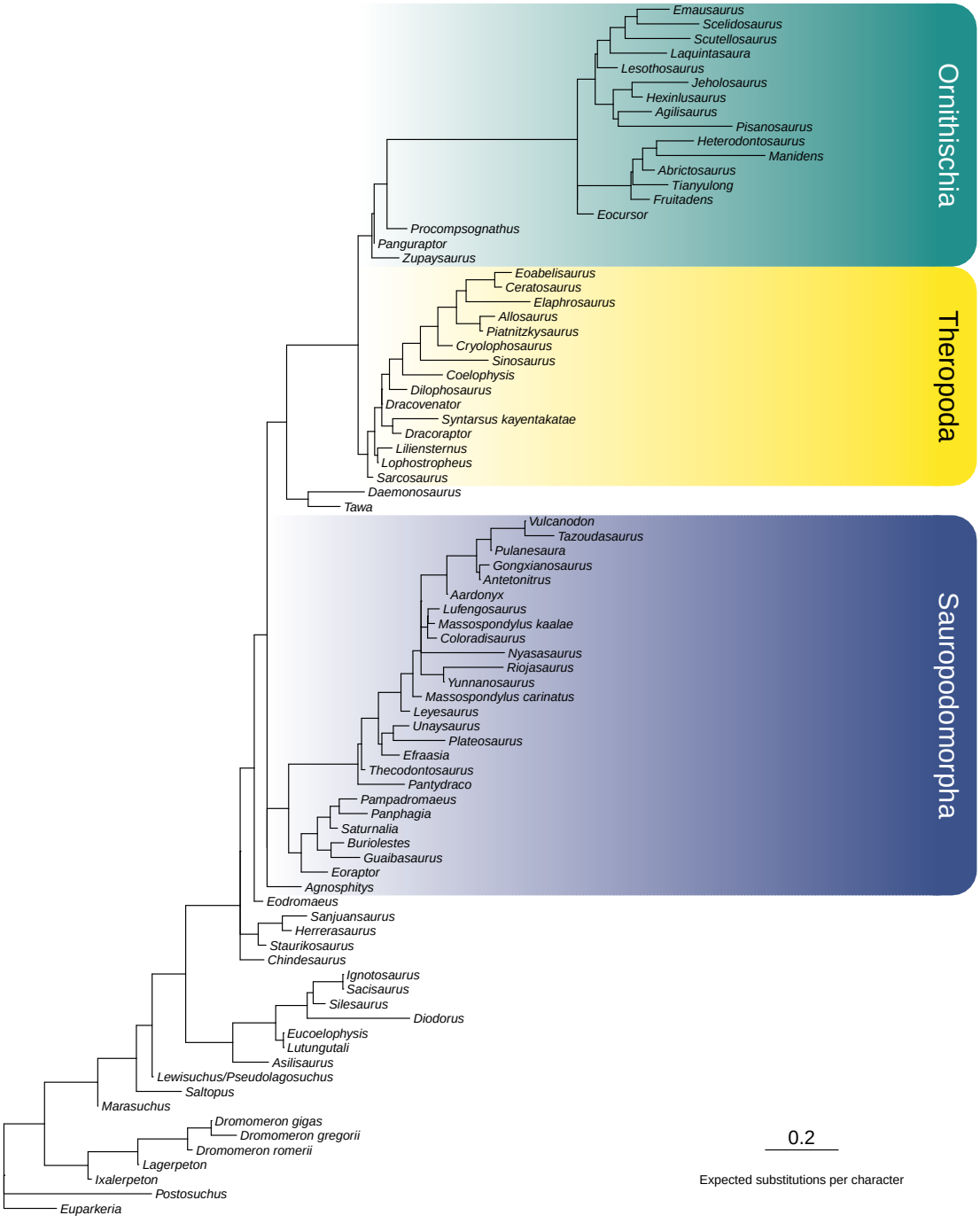

Figure B.6: Maximum likelihood tree inferred for the original LEA matrix and constrained to recover Ornithoscelida ($\ln L = -7017.424$).

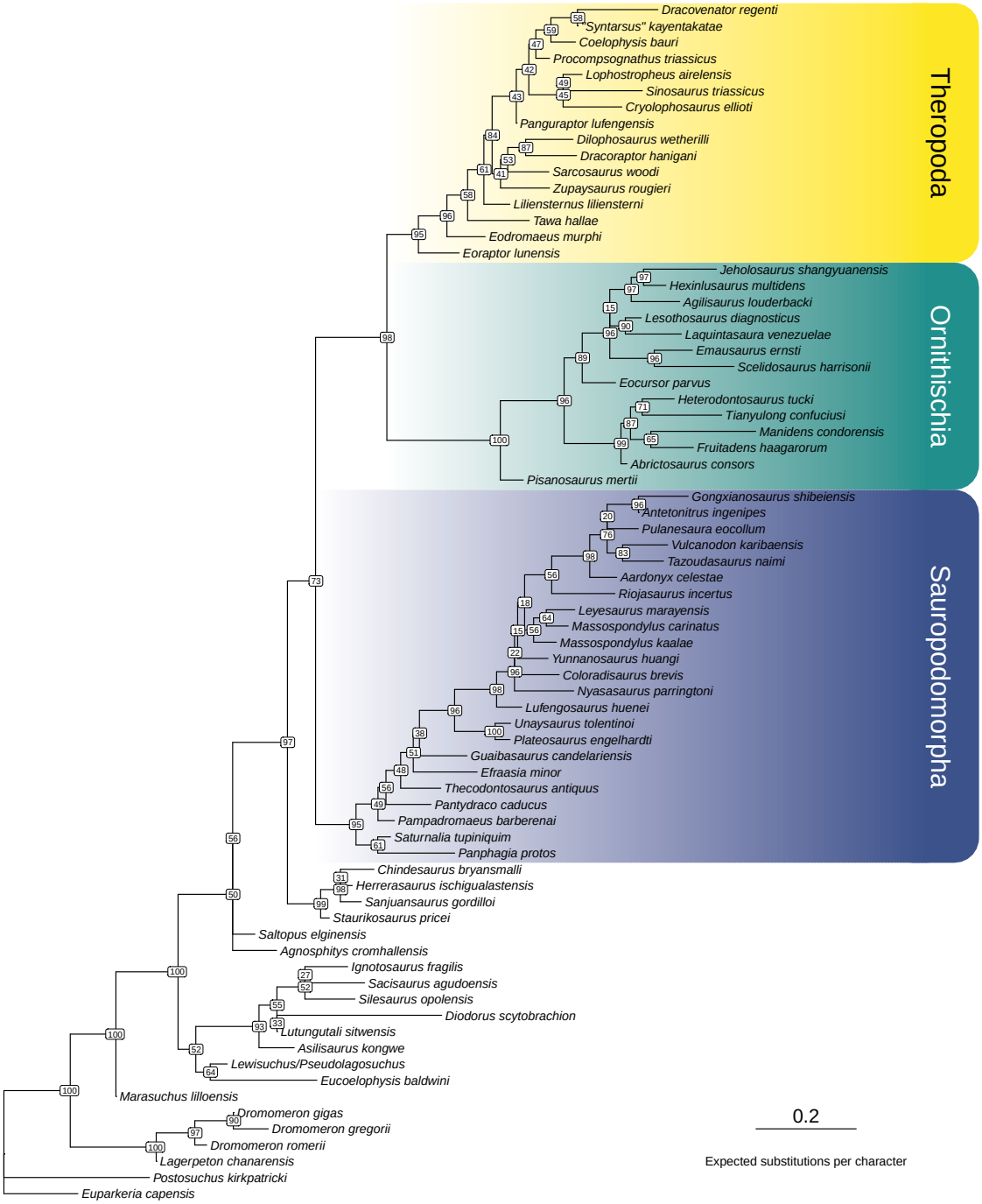

Figure B.7: Maximum likelihood tree inferred for the BEA matrix without the single most decisive character (character 175; PS = 1.840) ($\ln L = -6353.365$). Node labels denote ultrafast bootstrap values computed from 1000 replicates.

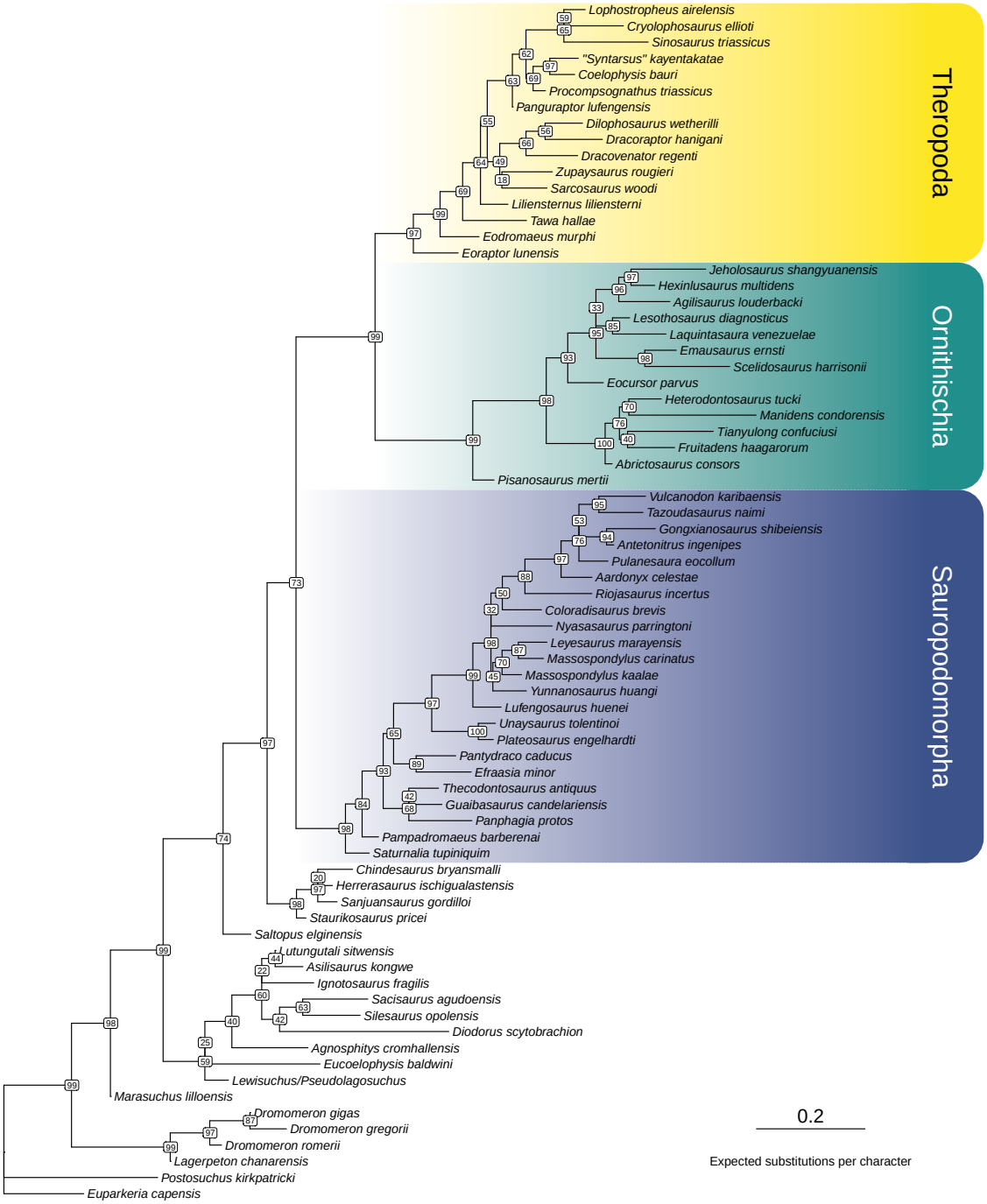

Figure B.8: Maximum likelihood tree inferred for the BEA matrix without the 5 most decisive characters (characters 175, 174, 303, 37, 292; PS = 1.483–1.840) ($\ln L = -6232.176$). Node labels denote ultrafast bootstrap values computed from 1000 replicates.

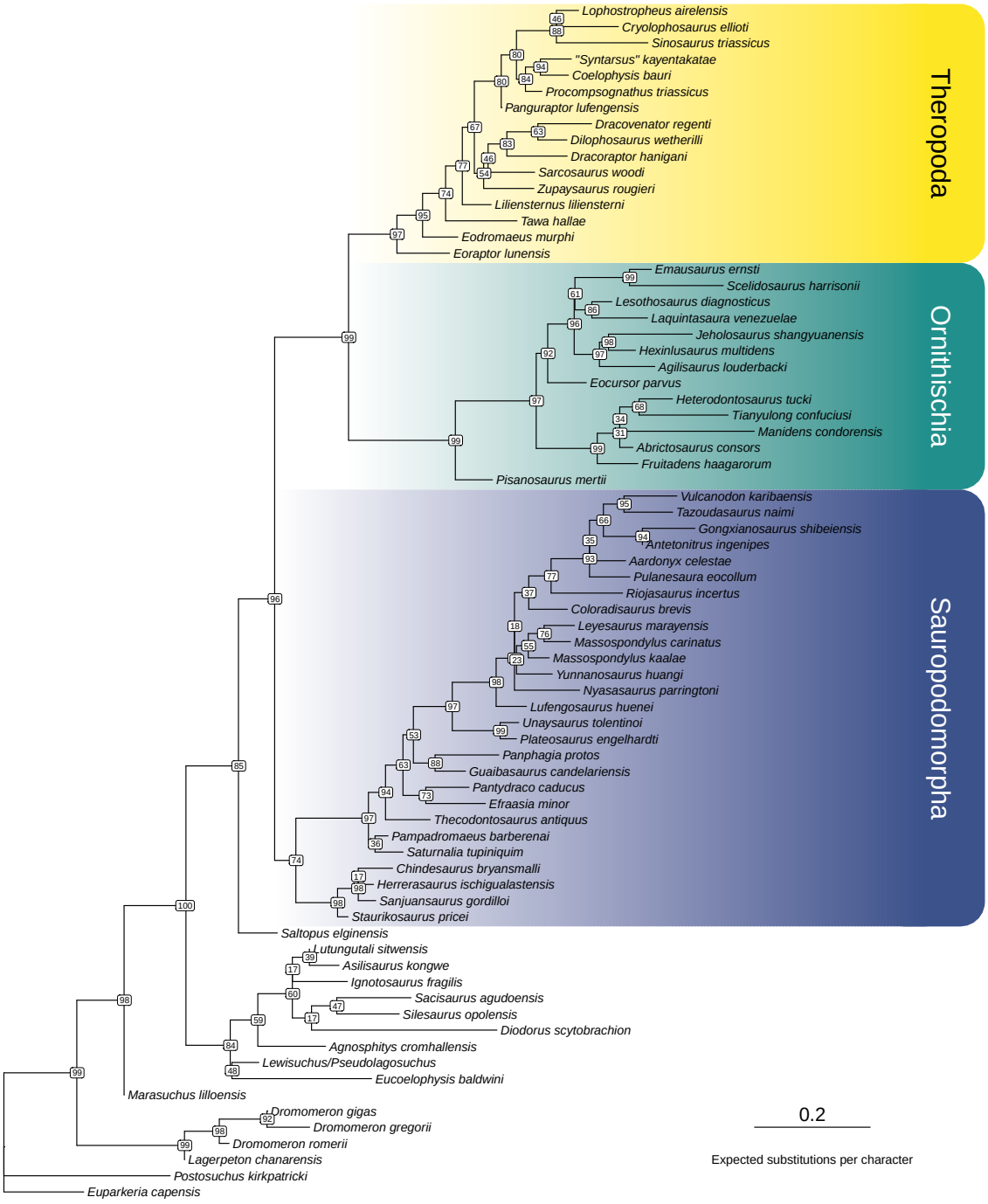

Figure B.9: Maximum likelihood tree inferred for the BEA matrix without the 10 most decisive characters (characters 175, 174, 303, 37, 292, 353, 323, 387, 167, 411; PS = 1.382–1.840) ($\ln L = -6103.143$). Node labels denote ultrafast bootstrap values computed from 1000 replicates.

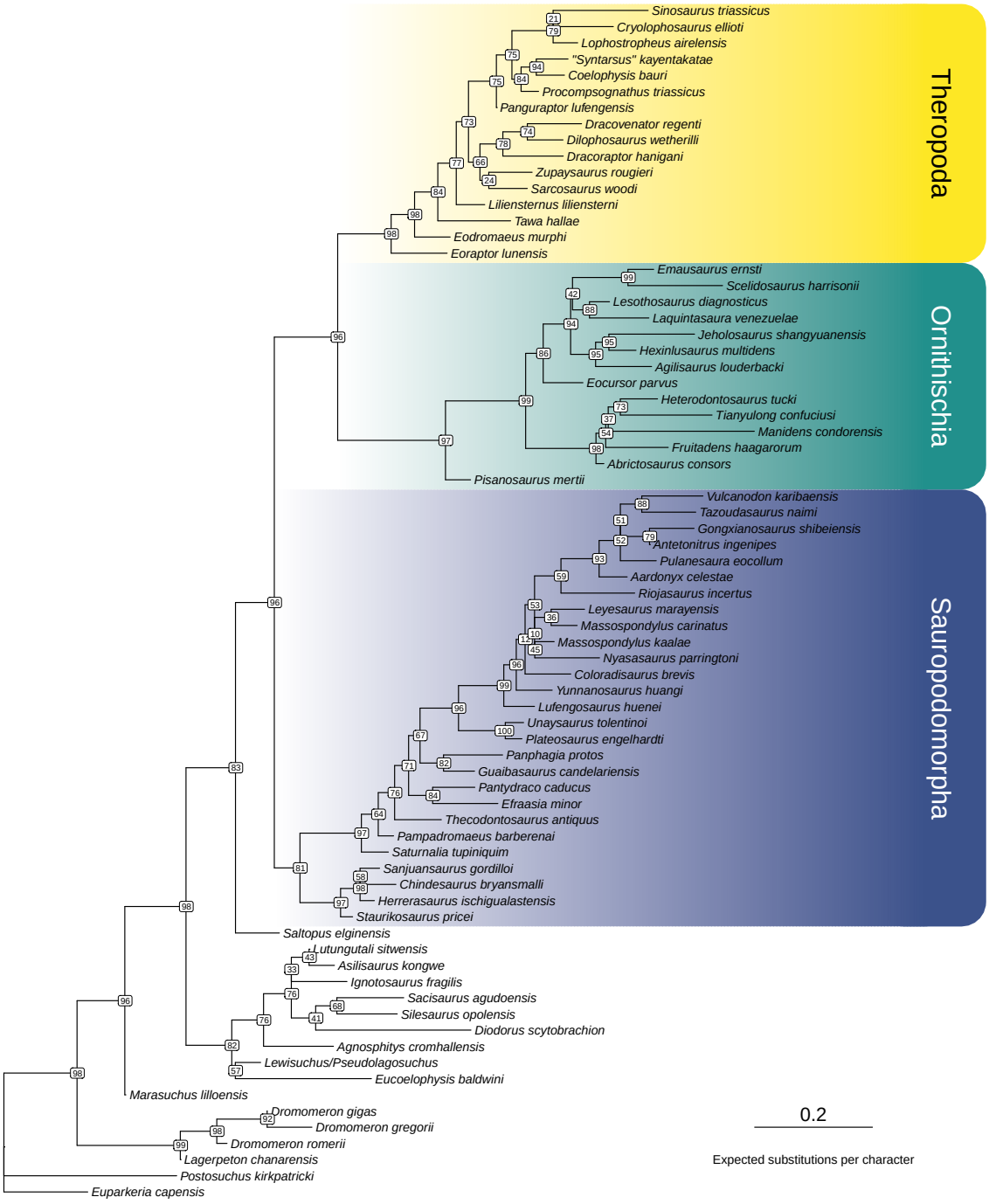

Figure B.10: Maximum likelihood tree inferred for the BEA matrix without all characters with outlier PS values (characters 175, 174, 303, 37, 292, 353, 323, 387, 167, 411, 68, 360, 169, 444; PS = 1.337–1.840) ($\ln L = -5992.320$). Node labels denote ultrafast bootstrap values computed from 1000 replicates.

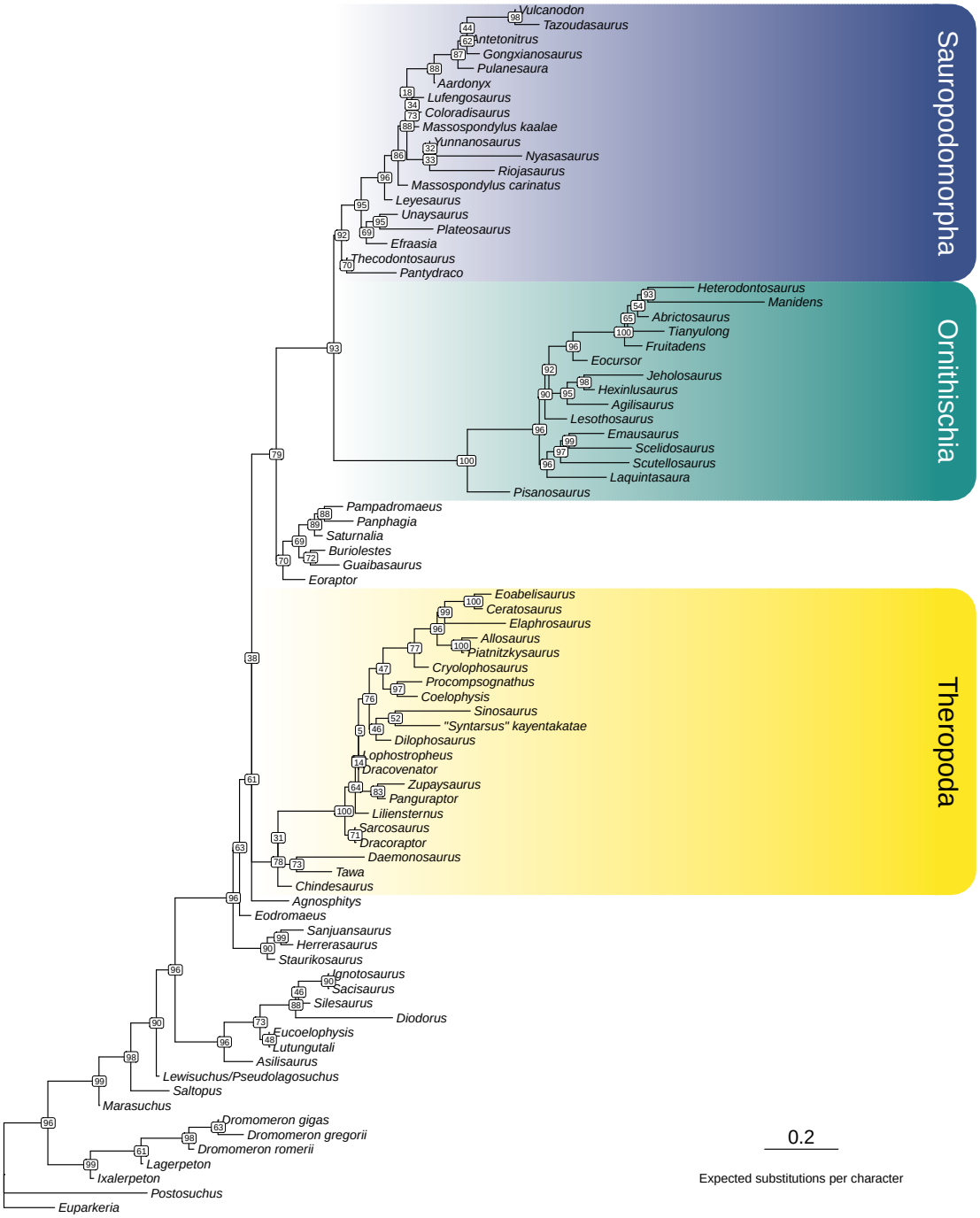

Figure B.11: Maximum likelihood tree inferred for the LEA matrix without the single most decisive character (character 206; PS = 4.454) ($\ln L = -6972.146$). Node labels denote ultrafast bootstrap values computed from 1000 replicates.

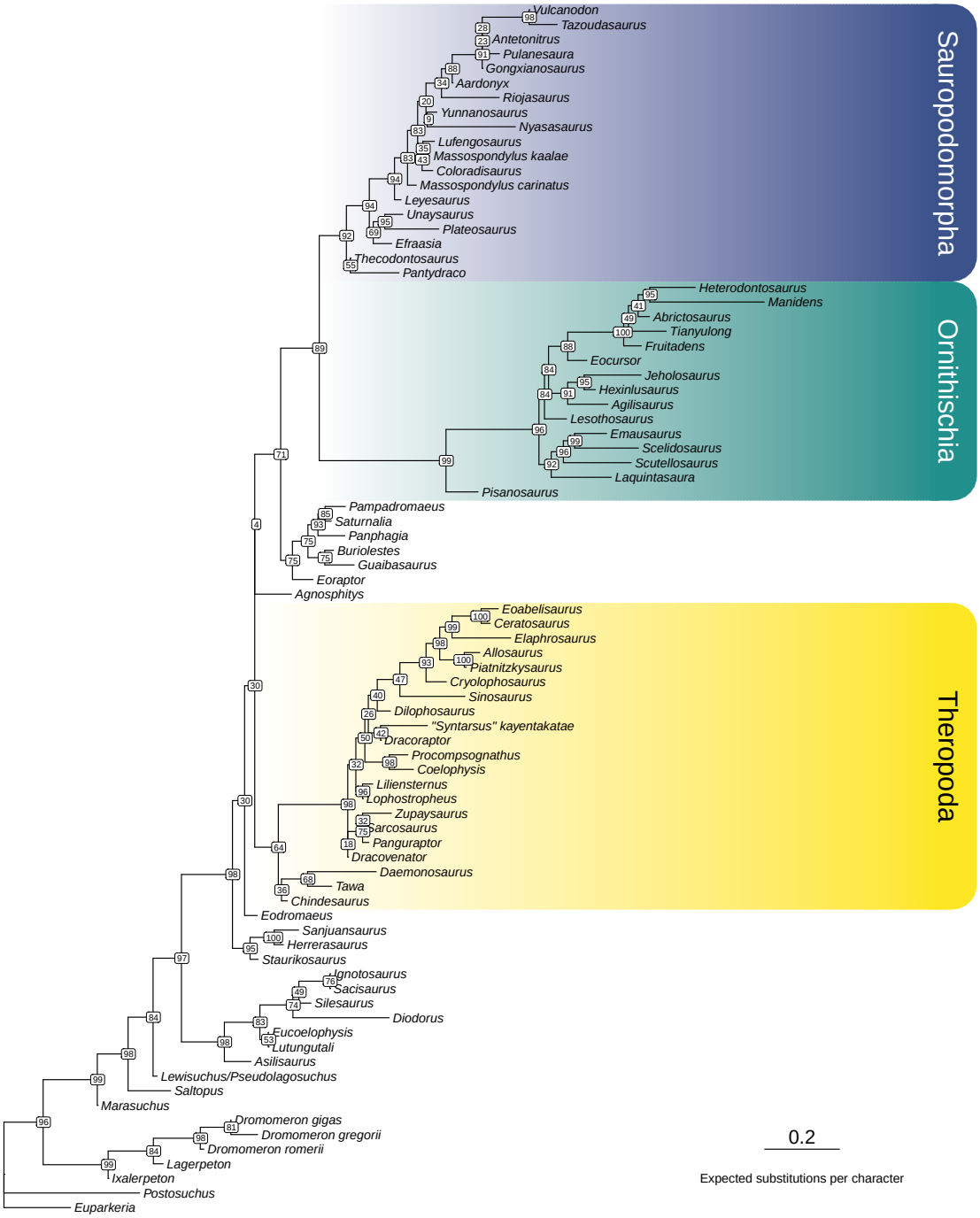

Figure B.12: Maximum likelihood tree inferred for the LEA matrix without the 5 most decisive characters (characters 206, 318, 169, 391, 198; PS = 3.228–4.454) ($\ln L = -6869.597$). Node labels denote ultrafast bootstrap values computed from 1000 replicates.

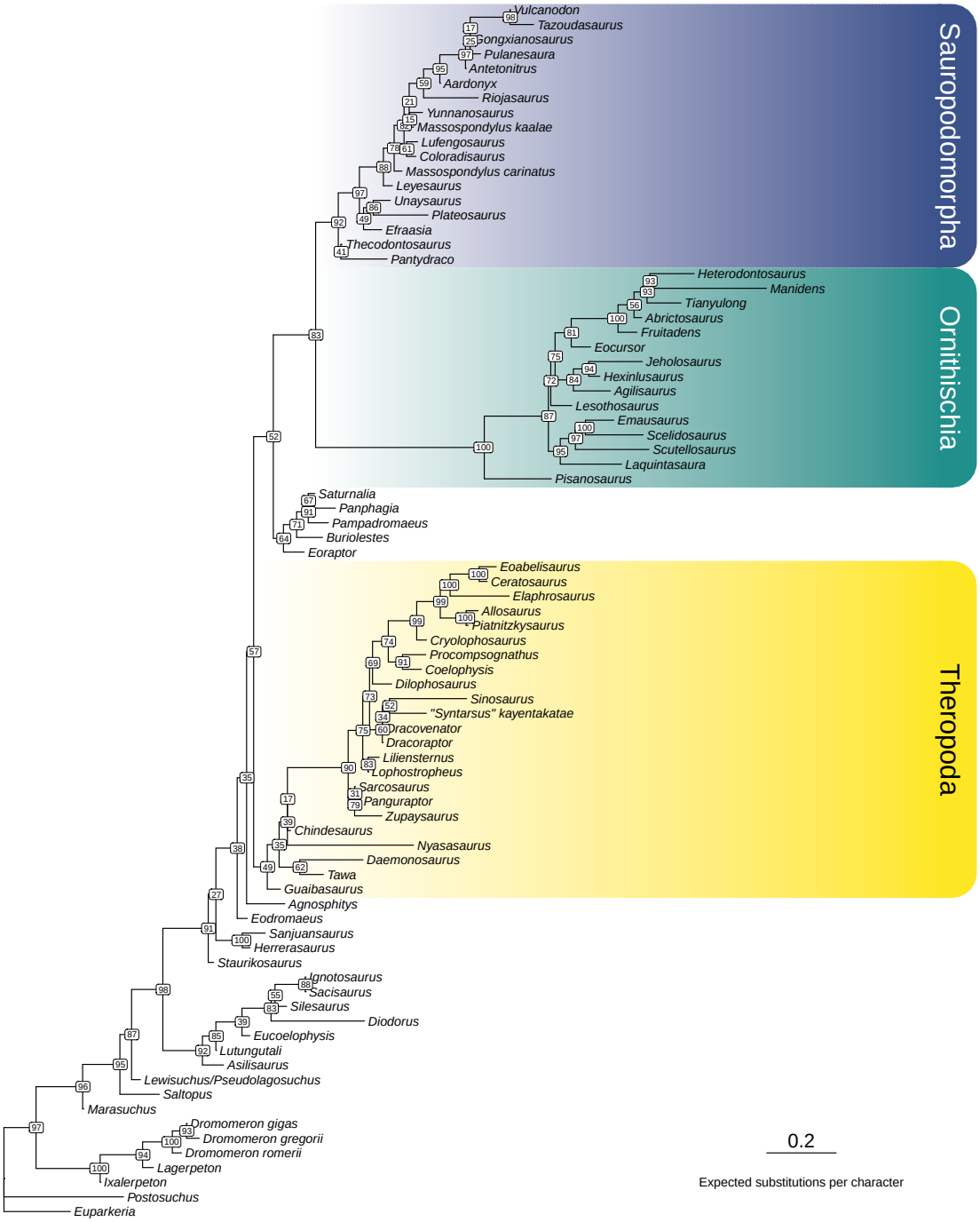

Figure B.13: Maximum likelihood tree inferred for the LEA matrix without the 10 most decisive characters (characters 206, 318, 169, 391, 198, 338, 377, 306, 301, 367; PS = 2.311–4.454) ($\ln L = -6795.273$). Node labels denote ultrafast bootstrap values computed from 1000 replicates.

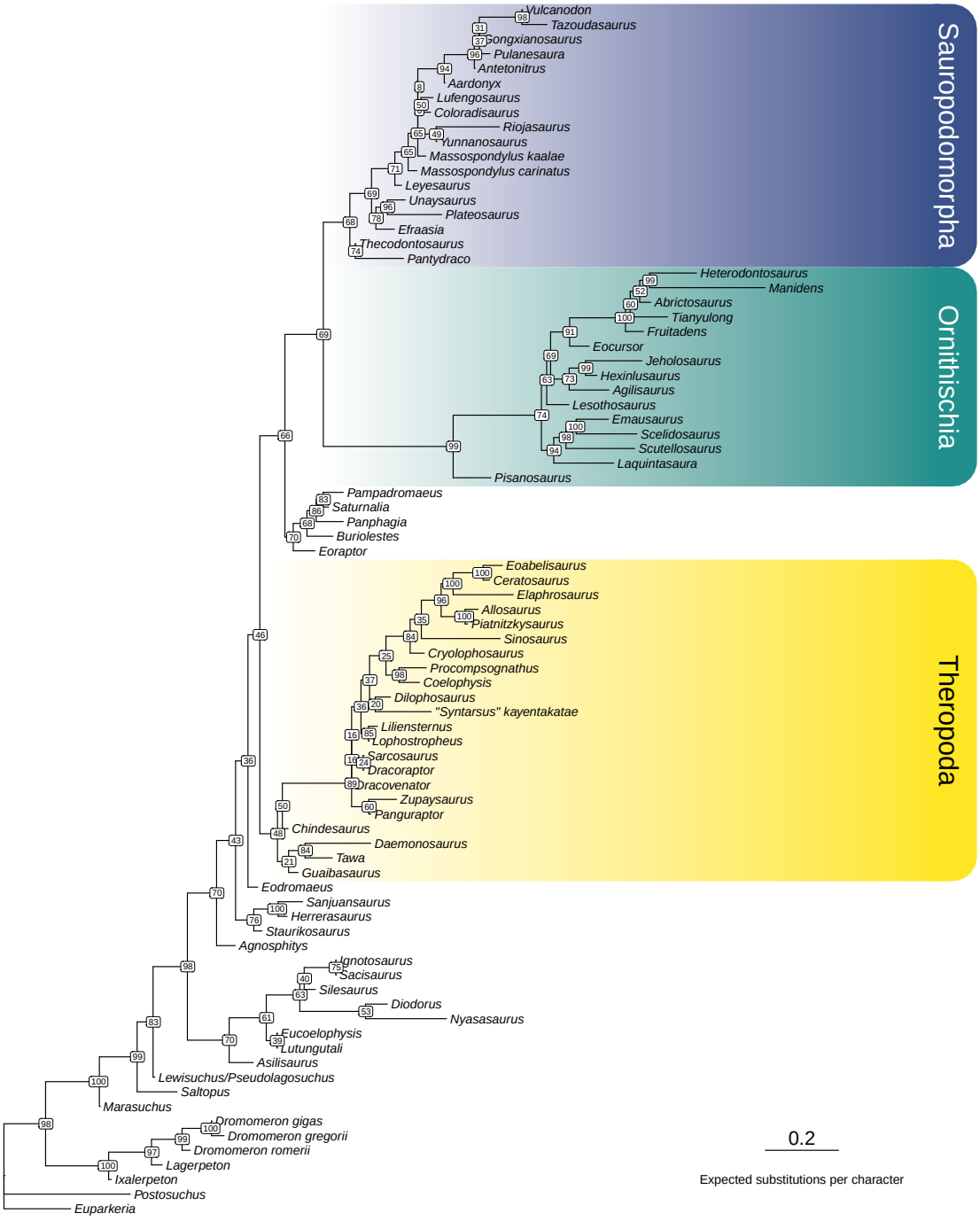

Figure B.14: Maximum likelihood tree inferred for the LEA matrix without all characters with outlier PS values (characters 206, 318, 169, 391, 198, 338, 377, 306; PS = 2.716–4.454) ($\ln L = -6825.859$). Node labels denote ultrafast bootstrap values computed from 1000 replicates.

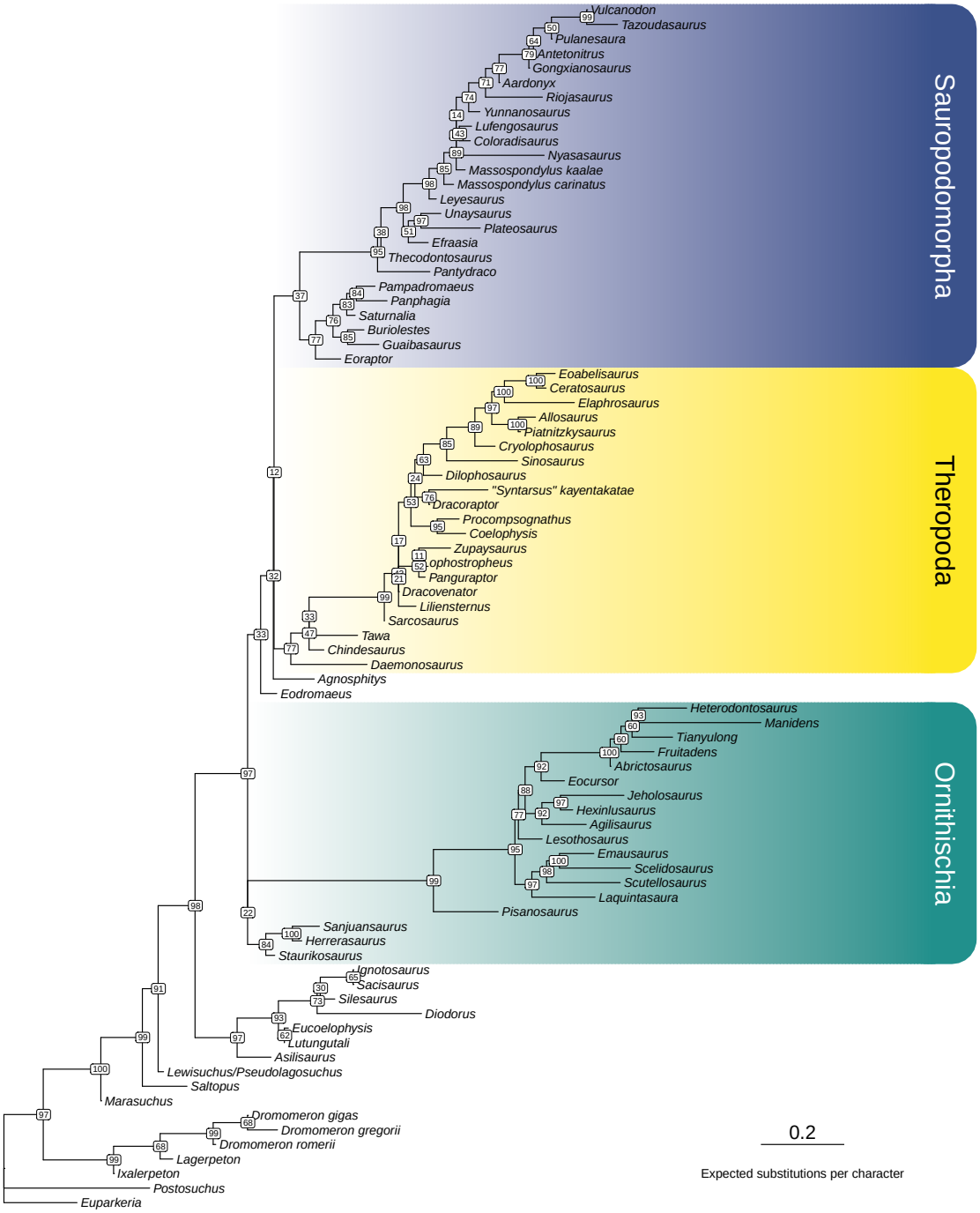

Figure B.15: Maximum likelihood tree inferred for the LEA matrix with character 77 reverted to its original coding in the BEA matrix ($\ln L = -7013.107$). Node labels denote ultrafast bootstrap values computed from 1000 replicates.

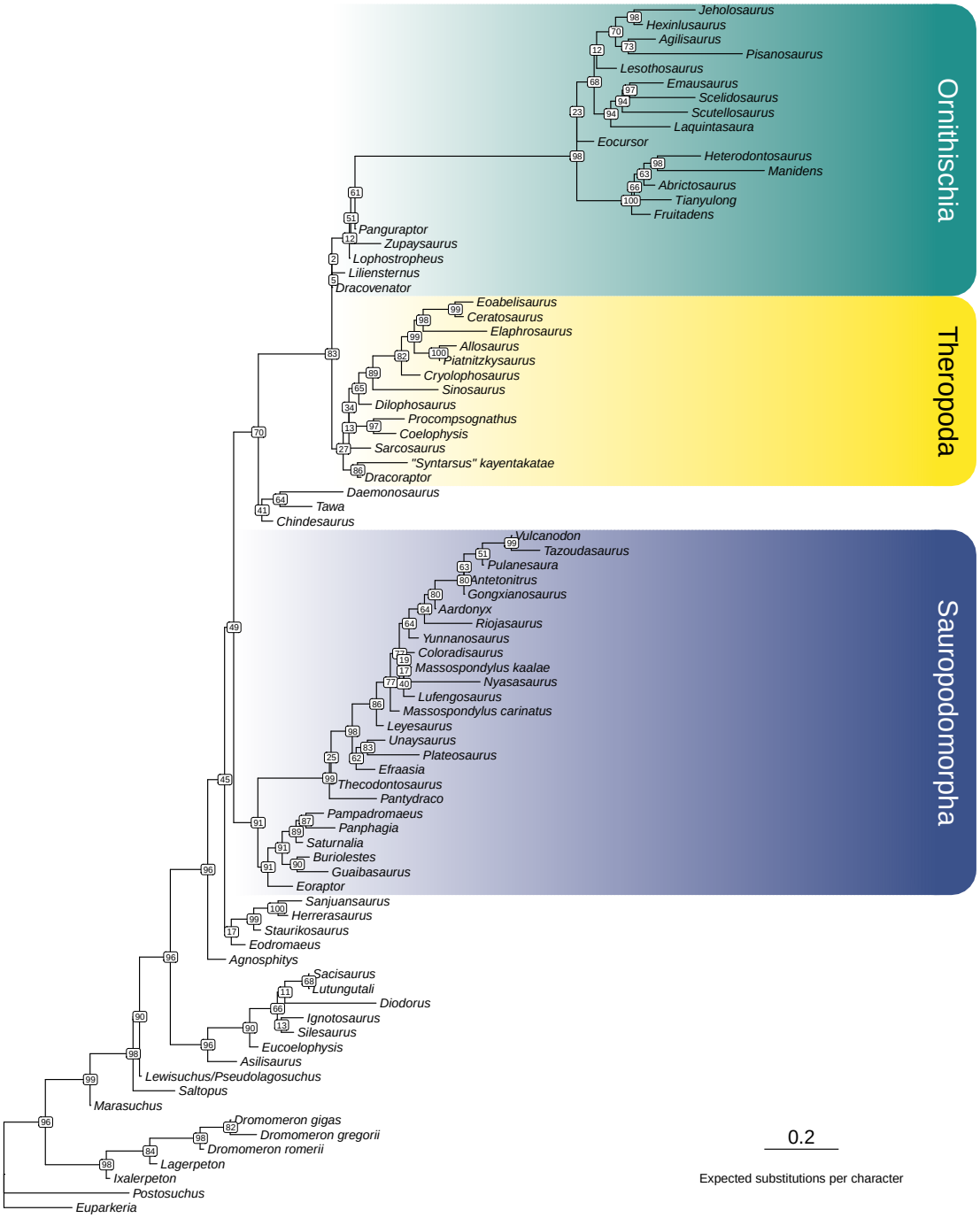

Figure B.16: Maximum likelihood tree inferred for the LEA matrix with character 148 reverted to its original coding in the BEA matrix ($\ln L = -7010.895$). Node labels denote ultrafast bootstrap values computed from 1000 replicates.

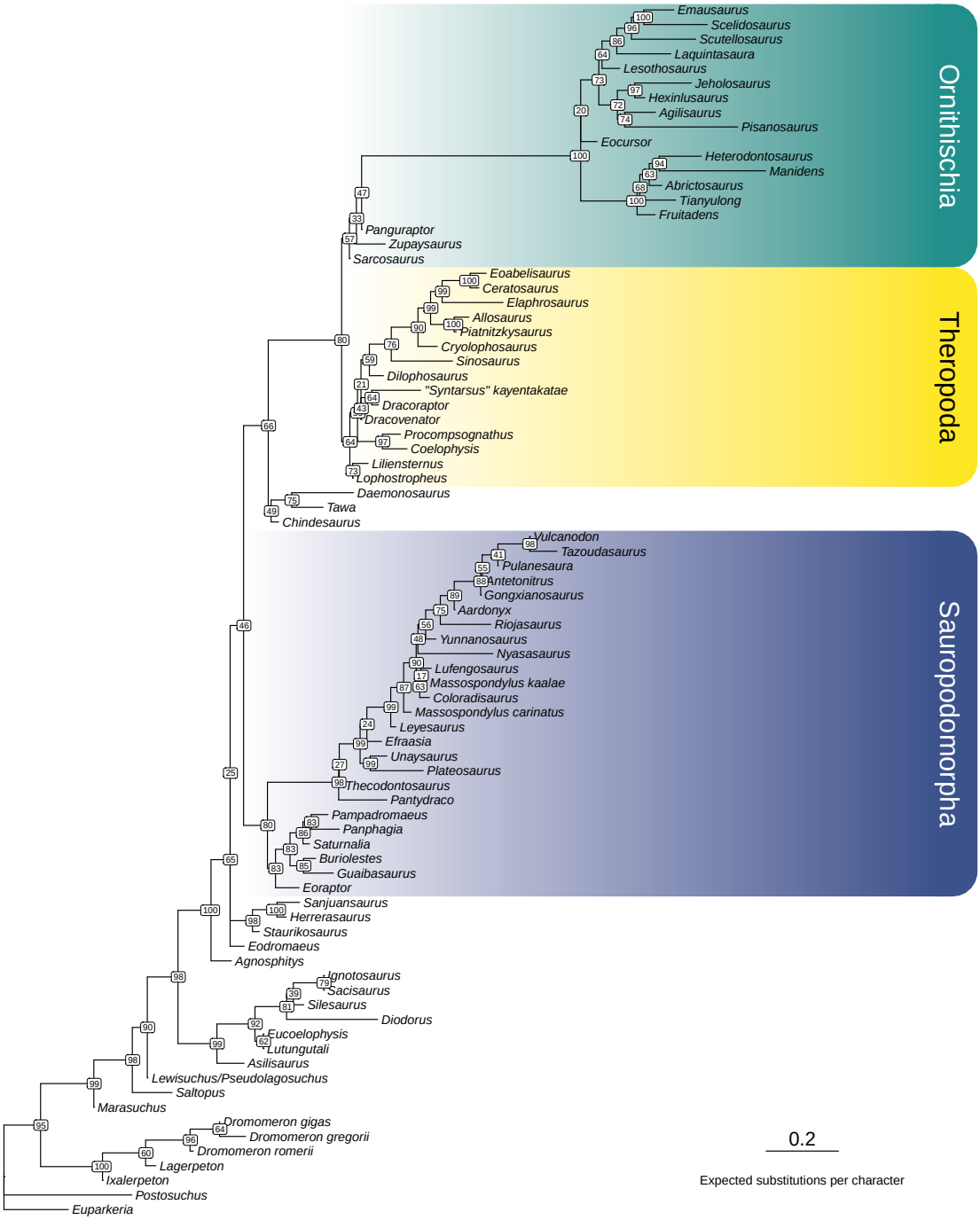

Figure B.17: Maximum likelihood tree inferred for the LEA matrix with character 363 reverted to its original coding in the BEA matrix ($\ln L = -7012.255$). Node labels denote ultrafast bootstrap values computed from 1000 replicates.

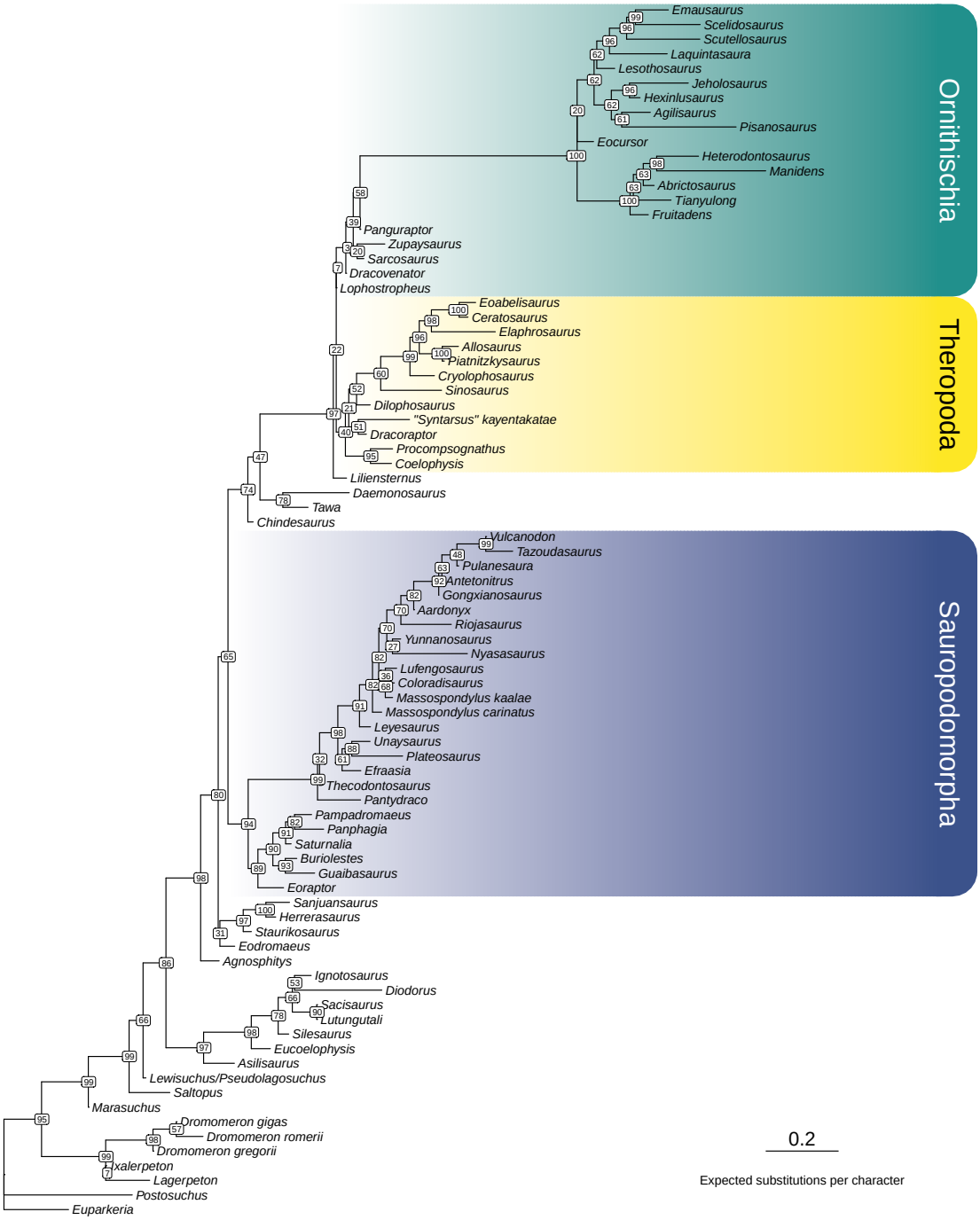

Figure B.18: Maximum likelihood tree inferred for the LEA matrix with character 370 reverted to its original coding in the BEA matrix ($\ln L = -7023.812$). Node labels denote ultrafast bootstrap values computed from 1000 replicates.

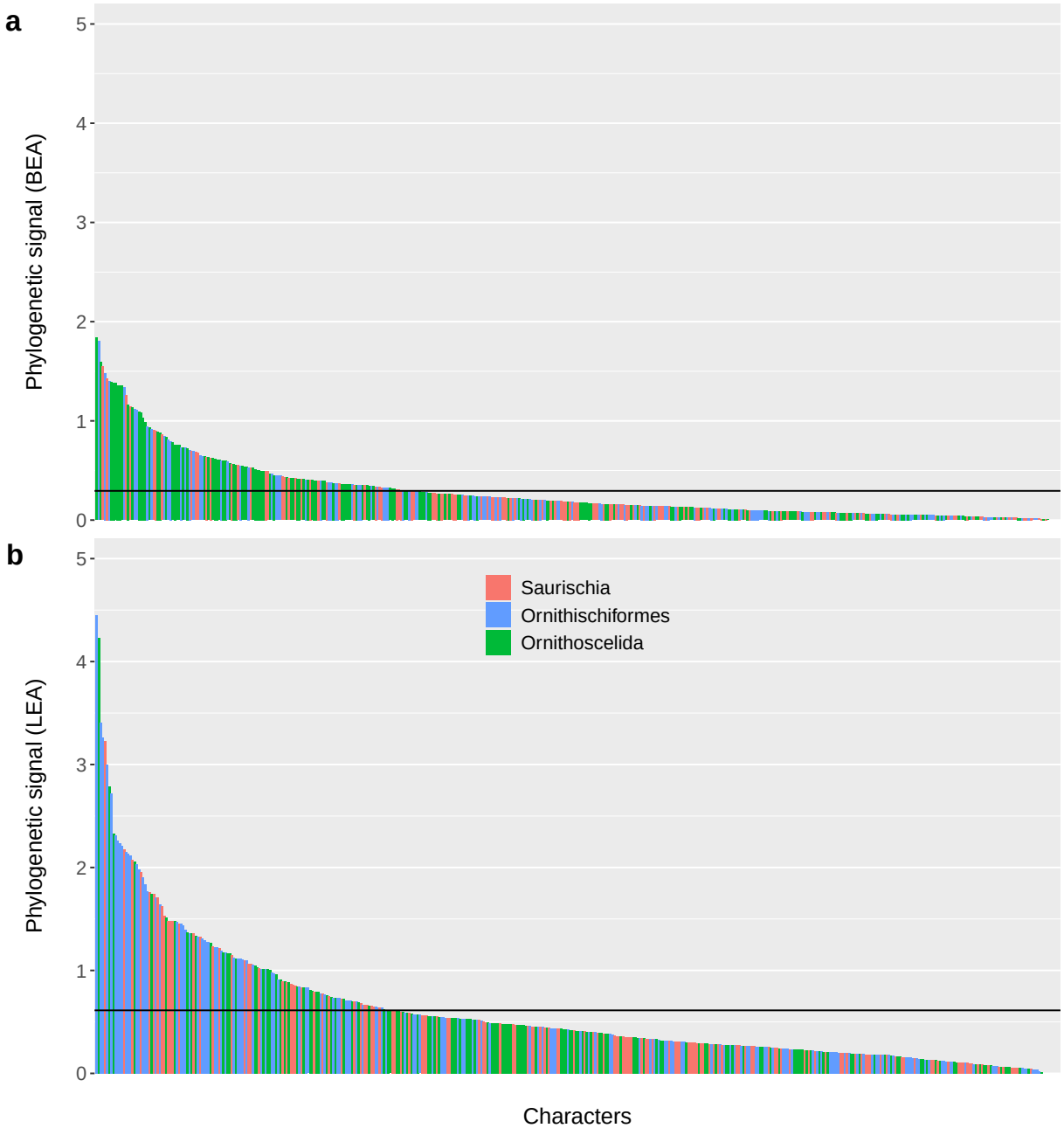

Figure B.19: Distribution of phylogenetic signal (PS) across the (a) BEA and (b) LEA matrices. The figure corresponds to panels (a) and (c) of Fig. 2.2 in the main text, except that individual characters are arranged in descending order of their PS values and a common scale is imposed for both datasets. Additionally, the mean PS value in each dataset is denoted by a horizontal line.

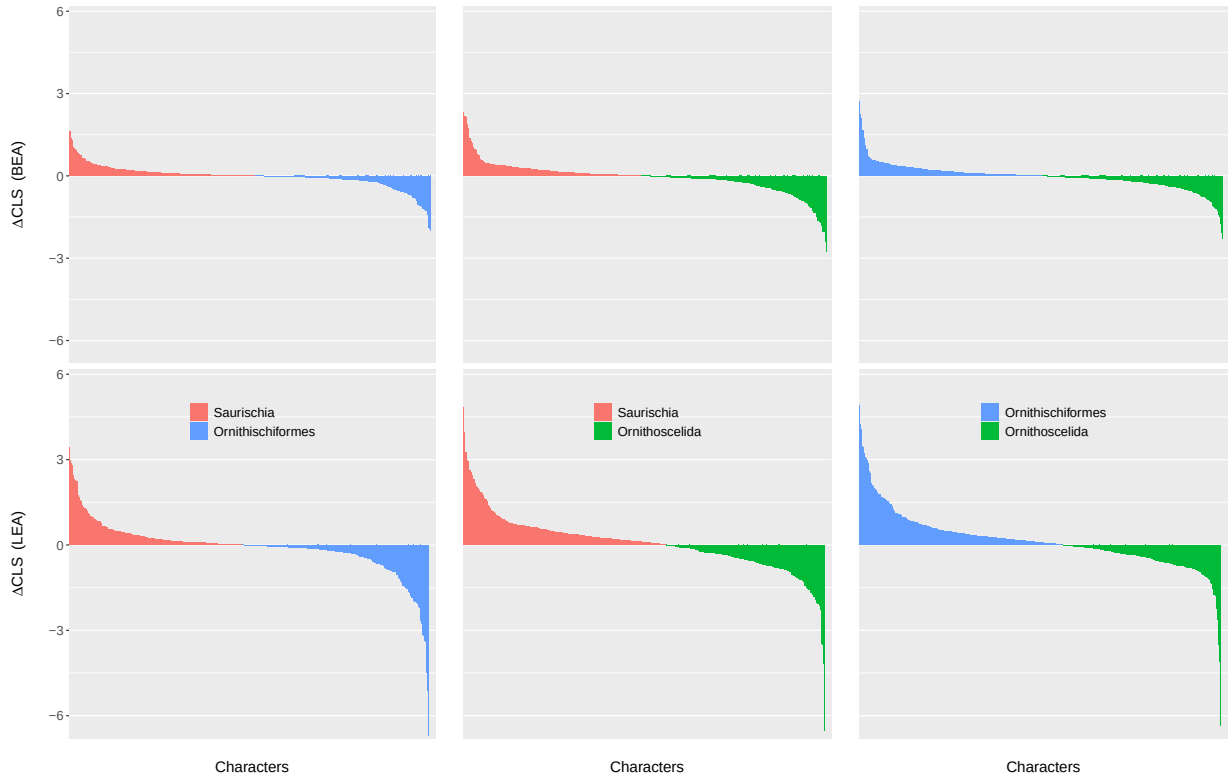

Figure B.20: Pairwise comparisons of character support for the three alternative early dinosaur topologies. The figure corresponds to Fig. 2.3 in the main text, except that individual characters are arranged in descending order of their ΔCLS values and a common scale is imposed for both datasets.

APPENDIX C: SUPPLEMENTARY INFORMATION FOR CHAPTER THREE

C.1 Occurrence Data

The occurrence data were downloaded from the Paleobiology Database (PBDB) under the following settings: taxon = Ornithischia, taxonomic resolution = species, preservation = regular taxa only, taxon status = all, identification = latest, time rule = major (URL: `http://paleobiodb.org/data1.2/occs/list.csv?datainfo&rowcount&base_name=Ornithischia&taxon_reso=species&pres=regular`). No additional filters were imposed. We manually curated the resulting dataset to remove taxa no longer considered to be ornithischians, occurrences no longer considered to be identifiable to the species level, and several ichnofossil occurrences misclassified as body fossils. We also updated the taxonomic identities of multiple records in line with recent revisions, and added 13 recently described species not yet entered into the database at the time: *Adynomosaurus arcanus*, *Aquilarhinus palimentus*, *Choyrodon barsboldi*, *Dryosaurus elderae*, *Fostoria dhimbangunmal*, *Galleonosaurus dorisae*, *Magnamanus soriaensis*, *Mahuidacursor lipanglef*, *Matheronodon provincialis*, *Orthomerus dolloi*, *Psittacosaurus amitabha*, *Sektensaurus sanjuanboscoi*, *Weewarrasaurus pobeni*. These species were treated as having a single occurrence each. The pre-existing occurrence data were sourced by 16 authorizers from a total of 701 studies, with eleven references contributing more than 10 occurrences each (Lull, 1915; Lambe, 1918; Hennig, 1925; Lull, 1933; Lull and Wright, 1942; Brown and Schlaikjer, 1943; Sternberg, 1950; Ostrom, 1970; Danis, 1986; Horner and Currie, 1994; Eberth and Getty, 2005).

C.1.1 Geological Sampling Proxies

The distribution of ornithischian species through geological time is highly uneven and characterized by a conspicuous clustering of first appearance dates at the Barremian–Aptian

and Santonian–Campanian boundaries (Figure C.1). However, it is unclear to what extent this heterogeneity is due to changes in the rate of net diversification or changes in the rate of fossil preservation (ψ). To assess the relative contributions of both factors, we used the counts of fossil collections and fossiliferous rock units in each chronostratigraphic stage as simple geological proxies of sampling. Since within-collection or within-formation species diversity is disregarded, a comparison of changes in the number of collections and rock units to changes in taxonomic diversity over time represents a simple means of disentangling the rates of preservation and diversification that is independent of the birth-death models implemented in Fossil BAMM and PyRate. While these proxies are not commensurable with the ψ estimates yielded by the latter two methods (measured in occurrences per species per 1 Myr), and thus cannot be employed for quantitative comparisons of rate magnitudes, they still allow for qualitative comparisons of overall trends, and can be used to evaluate the results of those analyses that allow ψ to vary over time. Moreover, the geological sampling proxies can provide a check on methods that assume a constant ψ (Fossil BAMM, PyRate under the HPP model) by revealing potential instances where the diversification shifts inferred by the analyses of ornithischian data are more likely to be due to changes in ψ.

First, we focused on the number of PBDB collections, a widely used sampling unit (e.g., Kiessling et al., 2007; Alroy, 2010; Butler et al., 2013) that should ideally represent an assemblage of fossils found in close geographical proximity to one another and recorded at the finest stratigraphic scale possible (e.g., a bed or a group of beds), although some collections may result from formation-level and basin-level surveys (Butler et al., 2012, 2013; Peters and McClennen, 2016). Each collection can be associated with further information about lithology, depositional environment, preservation type, and collection methods. To account for stratigraphic and temporal uncertainty, a single collection spanning n stages was treated as contributing $\frac{1}{n}$ to each of the stages spanned. All stage boundaries were based on the most recent version of the International Chronostratigraphic Chart (v2020/03;

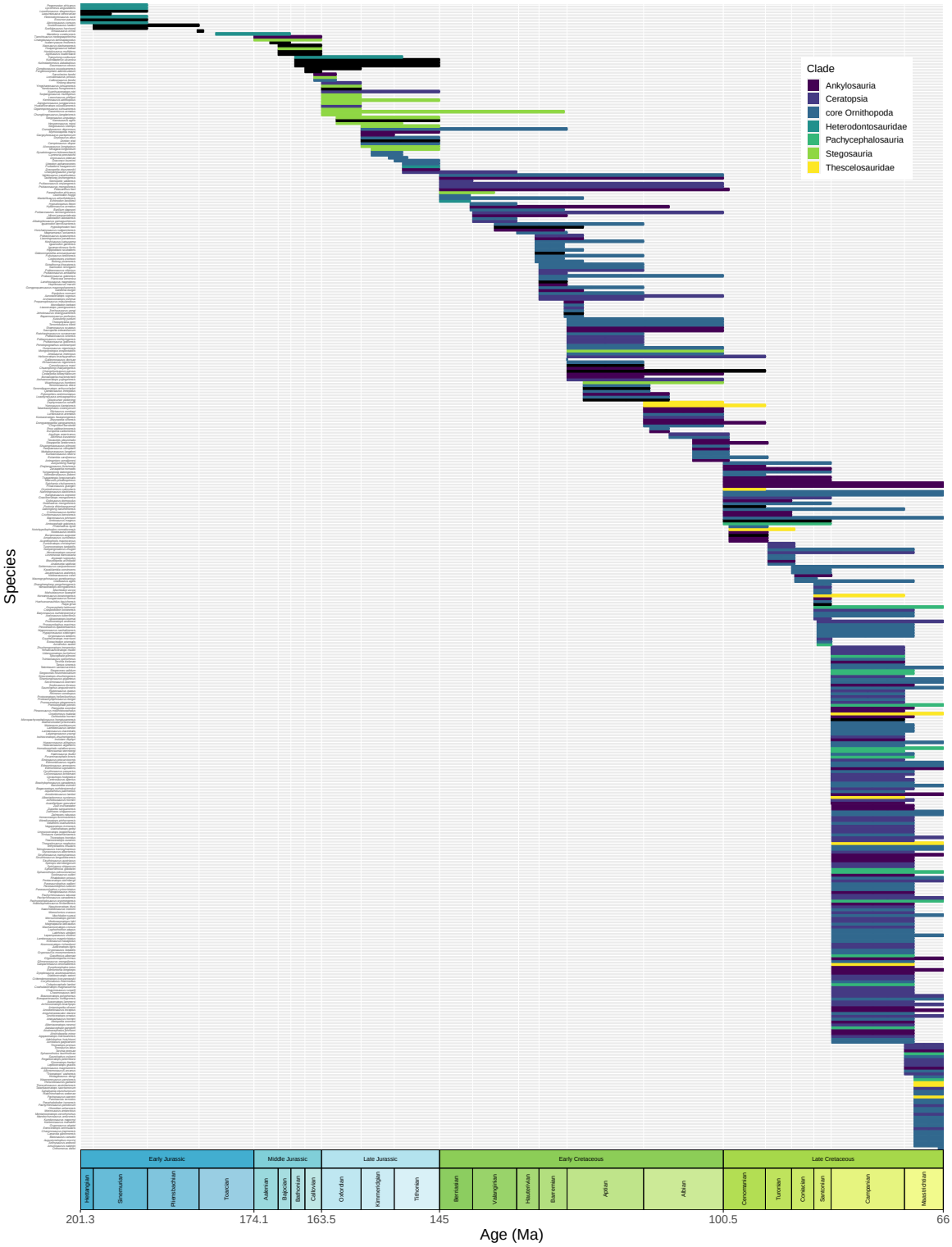

Figure C.1: Temporal distribution of the 398 valid ornithischian species analyzed in this study.

http://stratigraphy.org/ICSchart/ChronostratChart2020-03.pdf). We controlled for the fact that collections entered at different times used different numerical values for stage boundaries (e.g., those based on the time scales of Gradstein et al., 2005, 2012) to avoid erroneous assignment to multiple stages. The resulting per-stage collection counts ranged from 0.8 (Toarcian) to 395.1 (Maastrichtian), with a mean of 51.5 and a standard deviation of 95.6. Consistent with the observed patterns of taxonomic diversity, the number of collections underwent a sharp increase between the Santonian and the Campanian (Figure C.2), suggesting that a pulse of elevated preservation was the main driver behind the aggregation of first appearances at this boundary.

While the number of collections correlates with exposed sedimentary rock volume, it is also influenced by anthropogenic variables such as sampling effort (Butler et al., 2012). This is desirable in principle, since the ψ parameter inferred by BAMM and PyRate also confounds the geologically controlled preservation potential and the anthropogenically controlled rates of fossil recovery, identification, and description (Silvestro et al., 2014b; Mitchell et al., 2018). However, the process of converting identified and described fossil finds into PBDB collections may be subject to biases. These include "monographic" effects, in which a single detailed study listing all occurrences from a given time interval or region artificially inflates its diversity by introducing a number of species not recorded from any other place or time (Raup, 1976; Sepkoski, 1996; Butler et al., 2012), and binge sampling, in which taxa known from a single formation are included in a number of different collections by virtue of occurring in multiple layers of a single well-sampled section or multiple easily accessible outcrops (Plotnick and Wagner, 2018; Wagner et al., 2018).

To control for these biases, we used the number of ornithischian-bearing rock units per chronostratigraphic stage as our second sampling proxy. In keeping with other paleobiological studies (Peters and Foote, 2001; Butler et al., 2012; Darroch and Wagner, 2015; Wagner et al., 2018), we specifically focused on formations rather than beds, members, or

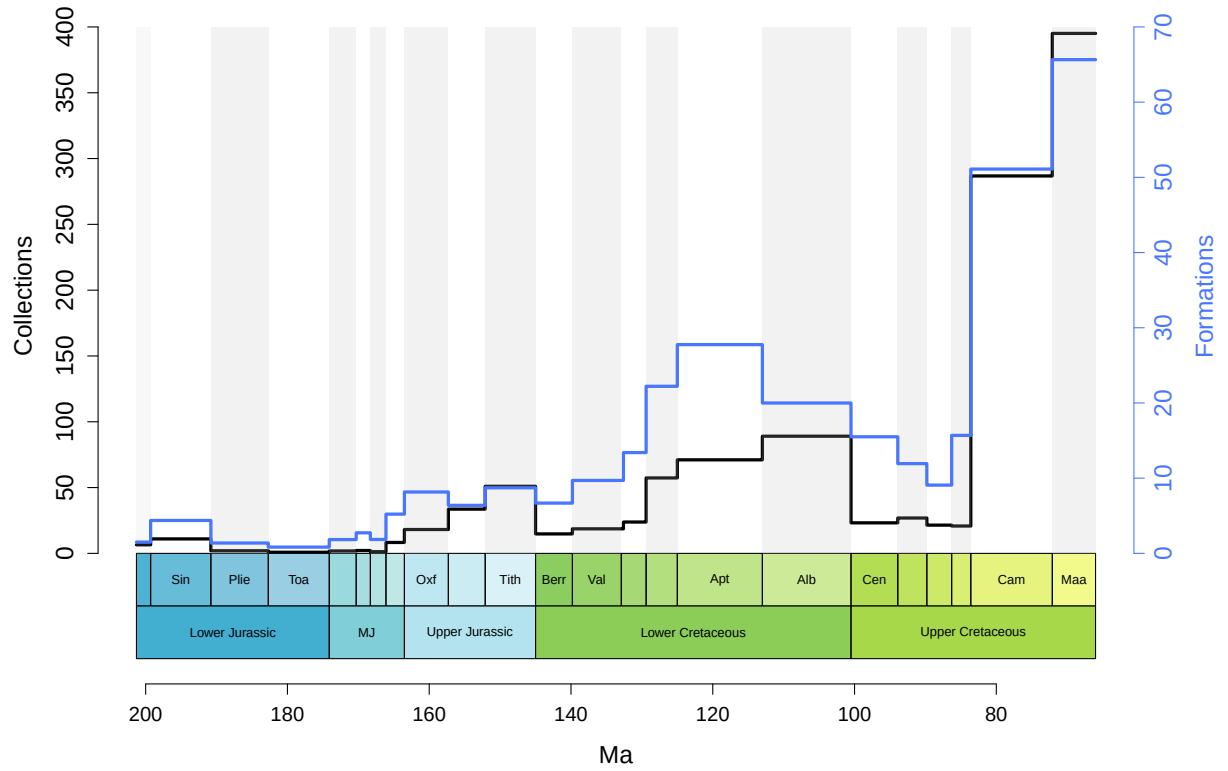

Figure C.2: Number of ornithischian-bearing collections (black) and formations (blue) in the Paleobiology Database plotted through time at the stage level of resolution.

groups. Formations represent the most fundamental stratigraphic unit and can be formally defined based on characteristic lithology and mappability, so that every formation should approximately correspond to a distinct paleoenvironment or biome (Peters and Foote, 2001; Darroch and Wagner, 2015). Moreover, of the three stratigraphic units recorded in the PBDB (group, formation, member), formation was the one for which information was available for the largest number of collections. Through a combination of manual queries and automated lookups performed using the paleobioDB package (Varela et al., 2015), which provides an R wrapper to the Paleobiology Database API, we were able to extract formation data from 1112 occurrences. In addition, we added the corresponding information for 12 out of the 13 manually entered species without collection numbers as well as 15 pre-existing collections not assigned to a named formation-level unit. Each of these collections either came from a

locality whose name could be associated with a known formation based on references other than that from which the collection was entered, or belonged to an as-yet-unnamed unit that was nevertheless clearly distinct from all named formations already present in the data based on its age and geographic location. In the latter case, we coined an informal designation for such a unit based on its lithology and geography (Wagner et al., 2018). The final dataset, obtained after discarding three ambiguous collections, included formation data for 1032 out of 1084 collections (95.2%) representing 1181 out of 1237 occurrences (95.5%), a success rate comparable to other PBDB-based formation-level analyses (e.g., Peters and Heim, 2010).

We employed the protocol outlined by Darroch and Wagner (2015) and vetted the stratigraphic data to reconcile variant spellings and standardize the rank of rock units variously considered to be members, formations, or groups according to the most recent opinion available. Following the advice of one of the reviewers (Peter J. Wagner), we treated the general case of a formation spanning n stages and comprising k sites (represented by collections) as contributing $1 - \left(\frac{n-1}{n}\right)^k$ to each stage, i.e., the complement of the probability that none of the collections belonged to the stage in question. Note that this formula reduces to the previously used fraction $\frac{1}{n}$ in the single-site case. As before, we controlled for the variation in the numerical ages of stage boundaries. The final per-stage formation counts ranged from 0.8 (Toarcian) to 65.7 (Maastrichtian), with a mean of 13.6 and a standard deviation of 16.0. The resulting curve closely resembled that of collection counts, including a period of low values in the early Late Cretaceous, followed by a steep increase at the Santonian–Campanian boundary (Figure C.2).

C.1.2 Sampling Intensity Analyses

To obtain an independent estimate of ψ that would allow for quantitative as well qualitative comparisons with the BAMM and PyRate results, we used a simple piecewise-homogeneous Poisson model to estimate the preservation rate from stage-specific sampling

intensities. We define sampling intensity as the ratio of the (known) observed number of species to the total species richness, which is unknown and has to be estimated. To do so, we used the "Chao2" nonparametric approach (Chao, 1984), representing a broader class of asymptotic extrapolators that have been historically more often used in ecology than in paleobiology (Close et al., 2018), but are now increasingly applied to fossil data as well (Rivadeneira, 2010; Vavrek and Larsson, 2010; Darroch and Wagner, 2015). We used the always-obtainable bias-corrected form of the Chao2 extrapolator (Gotelli and Colwell, 2011) to calculate an estimate of true richness $\hat{S}$ from incidence data as:

$$\hat{S} = S_{\text{obs}} + \left(\frac{c-1}{c}\right)\frac{q_1(q_1-1)}{2(q_2+1)}, \tag{1}$$

where S_{obs} = observed richness, c = the number of samples (here taken to represent collections), q_1 = the number of species recorded from one collection (singletons), and q_2 = the number of species recorded from two collections (doubletons). In addition, we calculated asymmetrical confidence intervals (CIs) based on the assumption that $\hat{S}$ is normally distributed and the approximate variance estimator derived by Chao (1987). The corresponding bias-corrected forms are given by Colwell (2013, Eqs. 10, 13; see also Rivadeneira, 2010):

$$\widehat{\text{Var}}(\hat{S}) = \hat{S} + \left(\frac{c-1}{c}\right)^2 \frac{q_1(2q_1-1)^2}{4(q_2+1)^2} + \left(\frac{c-1}{c}\right)^2 \frac{q_1^2\, q_2(q_1-1)^2}{4(q_2+1)^4} \tag{2}$$

$$\hat{S}_{\text{CI95, L}} = S_{\text{obs}} + \frac{\hat{S} - S_{\text{obs}}}{\exp\left\{1.96\sqrt{\log\left[1 + \dfrac{\widehat{\text{Var}}(\hat{S})}{(\hat{S}-S_{\text{obs}})^2}\right]}\right\}} \tag{3}$$

$$\hat{S}_{\text{CI95, U}} = S_{\text{obs}} + (\hat{S} - S_{\text{obs}})\exp\left\{1.96\sqrt{\log\left[1 + \frac{\widehat{\text{Var}}(\hat{S})}{(\hat{S}-S_{\text{obs}})^2}\right]}\right\}, \tag{4}$$

where $\hat{S}_{\text{CI95, L}}$ and $\hat{S}_{\text{CI95, U}}$ are the lower and upper bounds of the asymmetrical 95% confidence interval, respectively. We evaluated $\hat{S}$, $\hat{S}_{\text{CI95, L}}$, and $\hat{S}_{\text{CI95, U}}$ separately for each stage while making sure to discard unsampled range-throughs (i.e., taxa present in stages $i-1$ and $i+1$, but not in stage i) when determining S_{obs}. To convert from abundance to incidence data, we rarefied the occurrences by collection (i.e., counted only one occurrence per species per collection).

From the Chao2 richness estimates, we calculated sampling intensity as $\frac{S_{\text{obs}}}{\hat{S}}$, or the probability that a species will be sampled (will leave at least 1 occurrence) over the duration of a given stage. If fossil preservation is modeled using a Poisson process following common paleobiological practice (e.g., Solow and Smith, 1997; Starrfelt and Liow, 2016), these stage-specific sampling intensities can be set equal to the complement of the Poisson probability of failing to sample a species over the duration of the i-th stage:

$$\frac{S_{\text{obs},i}}{\hat{S}_i} = 1 - \text{Poiss}(0, \psi_i t_i) = 1 - e^{-\psi_i t_i}, \tag{5}$$

where ψ_i is the stage-specific rate of the Poisson preservation process and t_i is the duration of the i-th stage in millions of years. By rearranging, we obtain:

$$\psi_i = \frac{-\log\left(1 - \frac{S_{\text{obs},i}}{\hat{S}_i}\right)}{t_i}, \tag{6}$$

or the stage-specific preservation rate in the units of stratigraphically unique occurrences per species per 1 Myr (occ·sp^{-1}·Myr^{-1}), directly comparable to the BAMM and PyRate estimates. The lower and upper bounds of the confidence intervals for ψ_i were calculated from $\hat{S}_{\text{CI95, U}}$ and $\hat{S}_{\text{CI95, L}}$, respectively.

We were able to infer ψ for all stages except the Pliensbachian, where the extrapolated richness equaled the observed richness, resulting in a nominally infinite preservation rate. For the remaining 22 stages, ψ ranged from 0.0161 (Cenomanian) to 0.6303 (Hettangian)

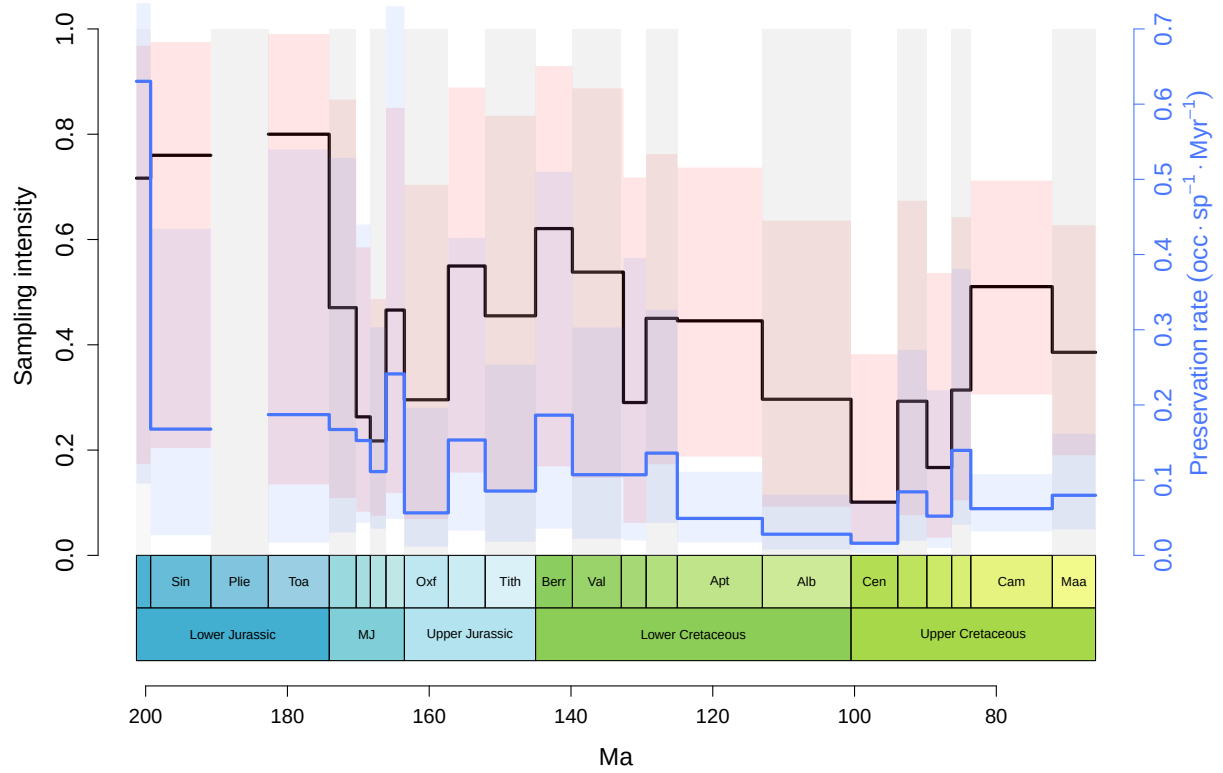

Figure C.3: Sampling intensity (observed over Chao2-extrapolated species richness) and Poisson preservation rate plotted through time at the stage level of resolution. The shaded areas show the associated 95% confidence intervals (red: sampling intensity, blue: preservation rate).

occ·sp^{-1}·Myr^{-1}, with a weighted mean (by stage duration) of 0.1077 occ·sp^{-1}·Myr^{-1}. Unlike the two sampling proxies, the estimated preservation rate did not increase at the Santonian–Campanian boundary, and even underwent a slight drop due to a corresponding rise in sampling intensity (Figure C.3). The precision of the estimates (quantified as the dimensionless ratio of the width of the 95% CI to the mean rate) was generally low and ranged from 1.232 (Campanian) to 4.329 (Cenomanian). Accordingly, despite the large variation in mean rates, our findings are largely consistent with ψ being constant given the overlap between the 95% CIs of all post-Hettangian stages (Figure C.3). The value for the Hettangian may itself be overestimated owing to the short duration of the stage and paucity of data, and its associated 95% CI still overlaps with those of most of the other stages. These results suggest

that the time-constant preservation model employed by BAMM is not seriously misspecified for the ornithischian data, despite the preference for more complex models by PyRate (Section C.4.1).

C.2 Character Matrices

For our analyses, we used three morphological character matrices obtained from Han et al. (2018), Madzia et al. (2018), and Herne et al. (2019), and modified them to include recently established taxa as well as the results of new descriptive studies.

The dataset of Han et al. (2018) was first expanded by the addition of a new character indicating the presence/absence of intercostal plates, taken from Cruzado-Caballero et al. (2019). This character replaced the dummy all-zero character that Han et al. (2018) used as the first column of their matrix. The matrix was further modified as follows: *Diluvicursor*, *Macrogryphosaurus*, *Mahuidacursor*, *Notohypsilophodon*, and *Talenkauen* were added using the scores of Cruzado-Caballero et al. (2019); *Sektensaurus* was scored based on Ibiricu et al. (2019); and *Burianosaurus* was added following personal observations by one of us (DM). Two additional taxa, *Marasuchus* and *Pisanosaurus*, were removed from the matrix. *Pisanosaurus* is no longer considered to be an ornithischian (Baron et al., 2017b; Agnolín and Rozadilla, 2018), and the basal dinosauriform *Marasuchus* was redundant with respect to three other more derived outgroups already included in the dataset (*Eoraptor*, *Herrerasaurus*, *Silesaurus*).

The changes to the dataset of Madzia et al. (2018) include the addition of *Weewarrasaurus* and *Fostoria* based on Bell et al. (2018); the addition of *Diluvicursor* and NMV P186047 (formerly the postcranial material of *Leaellynasaura*) as well as modifications to the scores of *Atlascopcosaurus*, *Leaellynasaura*, and *Quantasaurus* following Herne et al. (2018); the addition of *Kangnasaurus* and '*Fulgurotherium*' as well as modifications to the scores of *Trinisaura*, *Morrosaurus*, *Talenkauen*, *Macrogryphosaurus*, and *Notohypsilophodon* as in

Rozadilla et al. (2019); and modifications to the postcranial scores of *Yinlong* following the description by Han et al. (2018). Additionally, we renamed *Othnielosaurus* as *Nanosaurus* based on the reassessment of the taxon by Carpenter and Galton (2018), and removed *Marasuchus* and *Asilisaurus* as redundant outgroups.

The character matrix of Herne et al. (2019) has not been modified, as it was published shortly before we finished gathering data for our analyses.

	HaEA	HeEA	MEA
Taxa	77	57	78
Outgroups	3	1	4
Supraspecific	0	2	0
Characters	381	302	255
Multistate	55	43	40
Ordered	19	23	18
Constant (with outgroups)	0	0	3
Autapomorphies (with outgroups)	15	43	30
Constant (without outgroups)	19	16	17
Autapomorphies (without outgroups)	17	34	24

Table C.1: Properties of the phylogenetic datasets used for time tree inference. HaEA = Han et al. (2018); HeEA = Herne et al. (2019); MEA = Madzia et al. (2018). Outgroups and supraspecific taxa were included in the initial maximum likelihood analyses but excluded from time tree inference.

C.3 Phylogenetic Analyses

C.3.1 Preliminary Phylogenetic Analyses

We used PartitionFinder v2.1.1 (Lanfear et al., 2016a) to test for the presence of groups of characters marked by distinct patterns of evolution, using character-wise entropy values as the grouping criterion and the Bayesian Information Criterion (BIC) for model choice. Since the test was intended to inform our RAxML analyses (see below), which were used to determine the position of the root, PartitionFinder was run on full datasets including outgroups. We used the iterative k-means algorithm to search for groups of similarly evolving characters

without any prior partitioning information. The use of k-means clustering was recently criticized by Baca et al. (2017) for its tendency to artificially increase the fit of spurious partitioning schemes by grouping all invariant sites into a single partition. This problem is likely to be less relevant for morphological character matrices, such as ours (Table C.1), which tend to nearly or totally exclude constant (and by extension invariant) characters. However, since Baca et al. (2017) found that the k-means partitioning schemes tended to produce spurious downstream inferences of tree topologies and their associated nodal support values, we evaluated the plausibility of the k-means schemes in light of the resulting topologies (see below). Finally, we ran PartitionFinder with branch lengths both linked and unlinked across partitions to account for potential heterotachy, recently reported to be widespread in morphological data (Goloboff et al., 2018).

Next, we estimated phylogenetic trees with branch lengths in units of substitutions per character using RAxML v8.2.12 (Stamatakis, 2014). For datasets in which more than one partition was detected by PartitionFinder (HaEA and HeEA), we performed both partitioned and unpartitioned analyses to evaluate the differences between the resulting topologies. In both cases, characters were separated into binary and multi-state partitions. In the partitioned analyses, these were further split according to the assignment of their constituent characters to the PartitionFinder schemes. To assess nodal support, we ran bootstrapping analyses with the `-N autoMRE` autostopping criterion and a maximum of 1,000 pseudoreplicates. Each of the maximum likelihood trees inferred from the expanded and revised character matrices was then pruned down to the taxon set of the original matrix, and the ETE 3 Python library (Huerta-Cepas et al., 2016) was used to calculate its polytomy-corrected Robinson-Foulds (RF) distance from the published parsimony tree (as shown in Han et al., 2018, Figure 15; Herne et al., 2019, Figure 25.2; and Madzia et al., 2018, Figure 4A).

For the HaEA dataset, the partitioned and unpartitioned topologies exhibited few differences in the relationships among major clades, and their relatively large RF distance

(Table C.2) was mostly inflated by re-arrangements within major groups. Both topologies conflicted with the corresponding parsimony tree and with other recent ornithischian phylogenies in the position of Pachycephalosauria, which did not form the sister group of horned dinosaurs (Ceratopsia) but rather part of the sister group to all neornithischians other than *Lesothosaurus*, *Agilisaurus*, and *Hexinlusaurus*. The ceratopsians themselves were found to be closely related to heterodontosaurids as well as to the paraphyletic "jeholosaurids". In addition, the partitioned analysis found *Isaberrysaura* to be a basal thyreophoran, whereas both the unpartitioned analysis and the original parsimony tree placed it within Stegosauria. More pronounced differences were observed between the partitioned and unpartitioned analyses of the HeEA dataset. The partitioned tree was similar to the parsimony topology in placing psittacosaurids within Ceratopsia and in finding jeholosaurids and orodromines to be more closely related to each other than to other neornithischians, although the resulting clade did not include thescelosaurines or *Nanosaurus*, in contrast to the original results of Herne et al. (2019). The most prominent difference from the parsimony and unpartitioned topologies was the paraphyly of Heterodontosauridae. The unpartitioned HeEA tree retained monophyletic heterodontosaurids, but placed psittacosaurids in a sister-group relationship with Pachycephalosauria and reconstructed *Parksosaurus*, *Thescelosaurus*, orodromines, jeholosaurids, and *Nanosaurus* as the successive sister groups of (Cerapoda + *Hypsilophodon*). Both the partitioned and unpartitioned topologies differed from the parsimony results in finding the dryomorphs of Herne et al. (2019) to be paraphyletic with respect to Rhabdodontidae, and in supporting a non-cerapodan position for *Hypsilophodon*.

To minimize the risk of biasing downstream inferences of divergence times and diversification dynamics due to spurious k-means induced topologies, we only retained a partitioning scheme if its RAxML analysis yielded a tree whose distance from the parsimony topology was smaller than or equal to that of the corresponding unpartitioned tree. Based on this criterion, all subsequent analyses employed the unpartitioned HaEA and MEA matrices only and the

partitioned as well as unpartitioned versions of the HeEA matrix, for a total of 4 datasets (Table C.2).

Dataset	Partitioned–parsimony RF distance	Unpartitioned–parsimony RF distance	Partitioned–unpartitioned RF distance
HaEA	32	30	24
HeEA	46	46	22
MEA		48	

Table C.2: The polytomy-corrected Robinson-Foulds (RF) distances between the maximum likelihood trees based on the partitioned and unpartitioned RAxML analyses of the revised matrices, and the original trees based on the parsimony analyses of the unmodified matrices. Since the revised matrices were expanded by the addition of new taxa, each RAxML tree was pruned down to the taxon set of the unmodified matrix prior to the comparison. Note that the best scheme identified by PartitionFinder for the MEA dataset left the matrix unpartitioned.

C.3.2 Tip-Dating

Mean rate of character evolution

To place a hyperprior on the mean rate of the relaxed clock model in our tip-dating analyses, we conducted a series of analyses using TempEst v1.5.3 (Rambaut et al., 2016). First, we pruned the outgroups from our RAxML trees to make the resulting estimates applicable to tip-dating analyses including only the ingroup. Second, we made 10 random draws from the tip age ranges to account for age uncertainty. Since TempEst assumes that greater values correspond to younger ages (i.e., it measures root-to-tip distances, rather than tip ages in time units before the present), we subtracted all tip ages from the maximum age in each of the 10 replicates. We then used TempEst to regress the root-to-tip divergence measured in substitutions per character on tip sampling time. For each of the 10 replicates, we recorded the slope of the regression line, corresponding to the rate of character evolution measured in substitutions per character per million years.

Variance of the rate of character evolution

We followed the protocol established by Ronquist et al. (2012a) to estimate the variance of the rate of character evolution. Using MrBayes v3.2.6 (Ronquist et al., 2012b), we ran non-clock and strict-clock analyses on the same fixed topology, obtained by removing outgroups from the RAxML trees. We used the default Exp(10) branch length prior for the non-clock analyses. We ran 2 chains of 10 million generation each for both non-clock and strict-clock analyses. To maximize the number of branches that could be compared between the non-clock and strict-clock trees, we rooted the former in accordance with the RAxML analyses, but since the lengths of the two new branches created by rooting were not meaningful, we dropped them from further comparisons. Finally, we identified matching branches between the non-clock and strict-clock trees, and regressed the square of their length difference (corresponding to the squared deviation from the 1:1 expectation) on the strict-clock branch length, recording the regression line slope.

Treatment of the extant sampling parameter

The sampled-ancestor fossilized birth-death (SA-FBD) model is increasingly applied to analyses of wholly extinct clades (Bapst et al., 2016; Cau, 2017; Paterson et al., 2019). However, it was originally developed for clades comprising a mixture of extinct and extant taxa, and accordingly includes the extant sampling probability ρ as a parameter (Gavryushkina et al., 2014). Moreover, it is often this probability that is known with a reasonable degree of confidence, making it the parameter of choice to fix and condition the analysis on (Gavryushkina et al., 2014). Unfortunately, the best treatment of this parameter in the absence of any extant taxa remains unclear. One solution previously proposed in the literature (Cau, 2017) and informal settings (Gavryushkina, 2018) consists in fixing ρ to 0, but this step has been informally suggested to render the SA-FBD model non-identifiable (Bouckaert, 2014). Indeed, the transmission birth-death model, which can be derived from the SA-FBD

model by the removal of ρ, is not identifiable (Gavryushkina et al., 2014), but this model also relaxes the removal probability r that is fixed to 0 in the SA-FBD model, making the relative contributions of different parameters to non-identifiability unclear.

An alternative solution consists in creating a "pseudoextant" taxon (or taxa) by subtracting the age of the youngest tip(s) from all other tip dates. Notably, BEAST 2, one of the most popular software packages implementing the SA-FBD model, already performs this subtraction internally when computing likelihoods (Gavryushkina, 2018). Consequently, it is possible to interpret ρ as the proportion of taxa existing at the time horizon of the youngest tip that were sampled in the analysis. This option may be useful when analyzing clades such as the ornithischians, which suffered a near-instantaneous extinction with a considerable species richness still in place immediately beforehand. However, the total number of species extant at the time horizon of interest is unlikely to be known with accuracy due to the incompleteness of the fossil record.

To assess the effect of ρ on the resulting time trees, we ran each analysis with the parameter set to both 0 and a nonzero value. This value was calculated as the number of tips with a minimum age of 66.0 Ma, divided by the total number of such species in the PBDB data.

Fixed-topology analyses

To assess the influence of stratigraphic information on time tree topologies, we conducted two independent analyses on each combination of dataset, partitioning scheme (where applicable), and ρ treatment. In the unconstrained analysis, we allowed topology and divergence times to be jointly estimated, imposing no user-supplied starting tree or monophyly constraints. In the fixed-topology analysis, we used the RAxML tree stripped of branch lengths as the starting tree, and removed all operators perturbing tree topology or swapping tips for sampled ancestors.

Clock model comparison

In partitioned tip-dating analyses, individual partitions shared a common topology but were assigned separate substitution models. To determine whether relaxed clock models should be linked or unlinked across partitions, we performed Bayes factor (BF) model comparison, using nested sampling (Maturana Russel et al., 2018) as implemented in the BEAST 2 package NS v1.1.0 to estimate the marginal likelihoods. We chose the zero-ρ unconstrained-topology scenario to compare the two clock models. The nested-sampling analysis was performed using 4 particles (one particle per thread), a chain length of 20,000 points, a subchain length of 5,000 iterations per point, and a tolerance of 10^{-12}. To verify that the number of particles used was sufficient to distinguish between the two models, we checked that the difference of their log marginal likelihoods (93.9, corresponding to a BF of 187.8 in favor of the two-clock model) exceeded twice the square root of the sum of the variances of the two log marginal likelihood estimates (21.9). Following Kass and Raftery (1995), we interpreted the result as providing very strong support for assigning a separate relaxed clock model to each of the two partitions.

Prior choices

We used the SA-FBD model (Gavryushkina et al., 2014) as the tree prior in all analyses. Following Ronquist et al. (2016), we placed an Exp(10) hyperprior on the net diversification rate. The turnover rate received a $\beta(2, 1)$ hyperprior, consistent with Marshall's (2017) "paleobiological law" stating that the average rates of speciation and extinction are approximately equal, while the sampling proportion was assigned the default $U(0, 1)$ prior. The extant sampling parameter ρ was fixed in all analyses, either to 0 or to a nonzero value as described above (Table C.3).

The prior on the time of origin t_{or} was based on the oldest known records of the ornithischians and their successive sister groups. After the recent redescription of the putative Late

#	Dataset	Partitions	Topology	ρ	UCLD mean	UCLD st. dev.
1	HaEA	1	F	0	LN$(-6.113, 2.5)$	Exp(2.105)
2	HaEA	1	F	0.165	LN$(-6.113, 2.5)$	Exp(2.105)
3	HaEA	1	U	0	LN$(-6.113, 2.5)$	Exp(2.105)
4	HaEA	1	U	0.165	LN$(-6.113, 2.5)$	Exp(2.105)
5	HeEA	1	F	0	LN$(-4.767, 2.5)$	Exp(1.855)
6	HeEA	1	F	0.152	LN$(-4.767, 2.5)$	Exp(1.855)
7	HeEA	1	U	0	LN$(-4.767, 2.5)$	Exp(1.855)
8	HeEA	1	U	0.152	LN$(-4.767, 2.5)$	Exp(1.855)
9	HeEA	2	F	0	LN$(-4.386, 2.5)$	Exp(1.721)
10	HeEA	2	F	0.152	LN$(-4.386, 2.5)$	Exp(1.721)
11	HeEA	2	U	0	LN$(-4.386, 2.5)$	Exp(1.721)
12	HeEA	2	U	0.152	LN$(-4.386, 2.5)$	Exp(1.721)
13	MEA	1	F	0	LN$(-5.395, 2.5)$	Exp(1.862)
14	MEA	1	F	0.165	LN$(-5.395, 2.5)$	Exp(1.862)
15	MEA	1	U	0	LN$(-5.395, 2.5)$	Exp(1.862)
16	MEA	1	U	0.165	LN$(-5.395, 2.5)$	Exp(1.862)

Table C.3: Summary of the datasets and settings used for the BEAST 2 analyses of ornithischian phylogeny. Notes: ρ = extant sampling parameter; UCLD mean = hyperprior on the mean of the lognormal relaxed clock model; UCLD st. dev. = hyperprior on the standard deviation of the lognormal relaxed clock model; F = topology fixed to the RAxML point estimate; U = unconstrained topology; LN(m, s) = lognormal distibution with mean m and standard deviation s; Exp(λ) = exponential distribution with rate λ.

Triassic ornithischian *Pisanosaurus* as a non-dinosaurian (Baron et al., 2017b; Agnolín and Rozadilla, 2018) and a biostratigraphic reassessment of the South African Elliot Formation (McPhee et al., 2017), there remain no known ornithischians predating the Triassic–Jurassic boundary (Baron, 2019), the age of which (201.3 Ma) was thus used as the offset of a truncated exponential prior. The mean was scaled to place 99% probability on ages less than 244.6 Ma, the midpoint of the age range spanned by the Anisian, used to approximate the ages of *Nyasasaurus* and *Asilisaurus*. Both taxa were at least once reconstructed as ornithischians (Puttick et al., 2017; Marsola et al., 2018) and are reported to date from the "late Anisian" (Nesbitt et al., 2013). We consider this constraint to be highly permissive, as *Nyasasaurus* may represent a non-dinosaurian dinosauriform (Nesbitt et al., 2013; Puttick et al., 2017) or even a chimera (Lee et al., 2018), and *Asilisaurus* is not reconstructed as a dinosaur in most phylogenies (Nesbitt, 2011; Bittencourt et al., 2015; Langer et al., 2017; Nesbitt et al.,

2020). Finally, the t_{or} prior was truncated at 251.926 Ma, the age of the Permian–Triassic boundary, as none of the many (> 10) successive eucrocopodan sister groups to Ornithischia predate the end-Permian mass extinction (Ezcurra and Butler, 2018).

We used an uncorrelated lognormal relaxed clock model with the mean rate drawn from a diffuse lognormal prior (mean set to the TempEst estimate; Table C.3) and the standard deviation drawn from an exponential prior (median determined as described above; Table C.3). Default priors were applied to all substitution model parameters.

Analysis settings and convergence diagnostics.

Using BEAST 2.6 (Bouckaert et al., 2019), we ran 2 chains of 100 million generations (sampling every 50,000 generations) for each of a total of 16 analytical scenarios, differing by dataset, partitioning (for the HeEA dataset only), ρ treatment, and topology treatment (Table C.3). As all of our datasets contained constant characters after the exclusion of outgroups (Table C.1), we employed the Mk rather than Mkv model. Ordered characters were treated as such using custom rate matrices, while the Γ rate heterogeneity model was linked across the ordered and unordered partitions of a given number of states. To determine whether the analyses reached stationarity, we used Tracer v1.7 (Rambaut et al., 2018) to ensure that the effective sample sizes (ESS) of all scalar parameters exceeded 200 and to examine trends in parameter values over the course of each chain. For the unconstrained-topology analyses, we additionally used the RWTY package (Warren et al., 2017) to assess topological convergence using the average standard deviation of split frequencies (ASDSF < 0.01). To compare the posterior and prior distributions, we also ran each analysis without data for 100–500 million generations, depending on the time needed to attain sufficiently high ESS values for all scalar parameters.

Analysis	Robinson-Foulds distance		
	Parsimony–BEAST 2	**Parsimony–RAxML**	**RAxML–BEAST 2**
HaEA, nonzero ρ	37	29	80
HaEA, zero ρ	35	29	72
HeEA, partitioned, nonzero ρ	52	42	70
HeEA, partitioned, zero ρ	44	42	60
HeEA, unpartitioned, nonzero ρ	46	40	54
HeEA, unpartitioned, zero ρ	48	40	52
MEA, nonzero ρ	52	48	100
MEA, zero ρ	48	48	100

Table C.4: Polytomy-corrected RF distances between the BEAST 2, RAxML, and original parsimony trees. All trees were pruned down to a common subset of taxa, resulting in the removal of outgroups (excluded from the BEAST 2 analyses) and all taxa absent from the published matrices used to infer the original parsimony trees. Note that the parsimony–RAxML distances differ from those reported in Table C.2 due to the removal of the outgroups.

C.3.3 Results and Discussion

Topologically, our RAxML phylograms and BEAST 2 time trees are broadly congruent with other recent estimates of ornithischian phylogeny (Butler et al., 2008; Boyd, 2015; Dieudonné et al., 2016). However, the bootstrap support values and posterior probabilities of most nodes are extremely low, especially deep in the tree (Figs. C.4–C.19). In accordance with recent studies, heterodontosaurids form the sister group to Genasauria (Thyreophora + Neornithischia; Boyd, 2015; Dieudonné et al., 2016; Han et al., 2018; Yang et al., 2020) or alternatively fall in the vicinity of Ceratopsia (Figs. C.4–C.6; Xu et al., 2006; Dieudonné et al., 2020).

In a notable departure from recent phylogenies of the group, a subset of our analyses failed to infer a sister-group relationship between Ceratopsia and Pachycephalosauria (Marginocephalia). This is true of both unconstrained BEAST 2 chronograms (Figures C.6, C.7, C.19) and the RAxML analysis of the HaEA dataset (Figures C.4, C.5), indicating that this result is not due to the conflicting topological influences of the SA-FBD prior and morphological likelihood that may occur when character data imply deep divergences yielding no sampled ancestors or sister taxa (= ghost lineages/taxa *sensu* Norell, 1993) (Turner et al., 2017). However, we found indirect support for the latter effect as well, with RF-

distance comparisons consistently showing the original parsimony trees to be closer to the RAxML phylograms than to the topologically unconstrained tip-dated trees (Table C.4). While weak support for the monophyly of traditional Marginocephalia was also recently noted by Dieudonné et al. (2020), we note that one of the trees that did not include traditional Marginocephalia (Figure C.19) was based on the MEA dataset, which lacks the characters needed to test the group's monophyly. Moreover, the posterior probabilities of the nodes separating Ceratopsia from Pachycephalosauria were extremely low in the unconstrained analyses (Figures C.6, C.7, C.19), as was their bootstrap support in the RAxML analysis.

Our tip-dating results are broadly compatible with an early Late Triassic divergence of the ornithischians from their sister group, implied by a majority of current phylogenetic hypotheses (Baron, 2019; however, see also Pacheco et al., 2019 and Müller and Garcia, 2020 for alternative topologies that place silesaurids within Ornithischia). The root age of the ornithischians was Norian in all unconstrained analyses (range of posterior medians: 222.8–216.3 Ma), with the corresponding 95% highest posterior density (HPD) intervals covering most of the Late Triassic. The fixed-topology analyses yielded older dates (range of posterior medians: 237.6–224.6 Ma) and wider HPD intervals that extended as far back as the latest Early Triassic while still including much of the Norian. Analyses using different values of ρ produced closely matching point estimates of node ages and nearly perfectly overlapping HPD intervals when the topology was fixed; however, when no topological constraints were used, setting ρ to a nonzero value resulted in extremely diffuse posterior age distributions with HPD intervals up to >100 Myr wide and mean dates that were up to 40 Myr older than those of the equivalent nodes in the zero-ρ trees (Figs. C.4–C.19).

We found no ancestor-descendant pairs among the taxa represented in any of the three datasets (posterior mean count = 0 for all analyses) despite their abundance in the marginal prior tree distributions (range of prior mean counts: 41.7–63.7). Note that given the formulae

derived by Foote (1996), the expected number of sampled ancestors $\hat{n}_{\mathrm{SA}}$ can be computed from the number of taxa n in a given tree and the rates of speciation (λ), extinction (μ), and sampling (s) inferred using BEAST 2, with the first two obtained from the posterior means of net diversification (d) and turnover (r) rates by a change of variables:

$$\lambda = \frac{d}{1-r}, \quad \mu = \frac{dr}{1-r} \tag{7}$$

$$\hat{n}_{\mathrm{SA}} \approx \begin{cases} 2ns\left(\frac{\lambda}{\mu}\right) & \text{budding speciation} \\ 2ns\left(\frac{\lambda}{\lambda+\mu}\right) & \text{bifurcating speciation} \\ ns\left(\frac{\lambda}{\lambda+\mu}\right) & \text{anagenetic speciation} \end{cases} \tag{8}$$

Across the 16 BEAST 2 analyses, $\hat{n}_{\mathrm{SA}}$ ranges from 0.48 to 7.98 under budding, 0.24–3.85 under bifurcating speciation, and 0.12–1.92 under anagenetic speciation. We attribute the lower than expected number of sampled ancestors in our time trees to the abundance of autapomorphies in our data (Table C.1) and the use of diversified sampling (Höhna et al., 2011), which prioritizes the inclusion of phylogenetically disparate taxa over the dense sampling of closely related species most likely to represent one another's ancestors. As recently noted by Simões et al. (2020), it is extremely unlikely that the relatively few species sampled from across the entire temporal and geographic range of a long-lived and globally distributed clade could represent ancestor-descendant pairs, a consideration that holds true for our data ($\sim$70 species from 6 continents per 135 Myr). However, our results also indicate that the absence of sampled ancestors can be detected even if the model allows for them, alleviating concerns about the overestimation of branch lengths under the SA-FBD model (Simões et al., 2020).

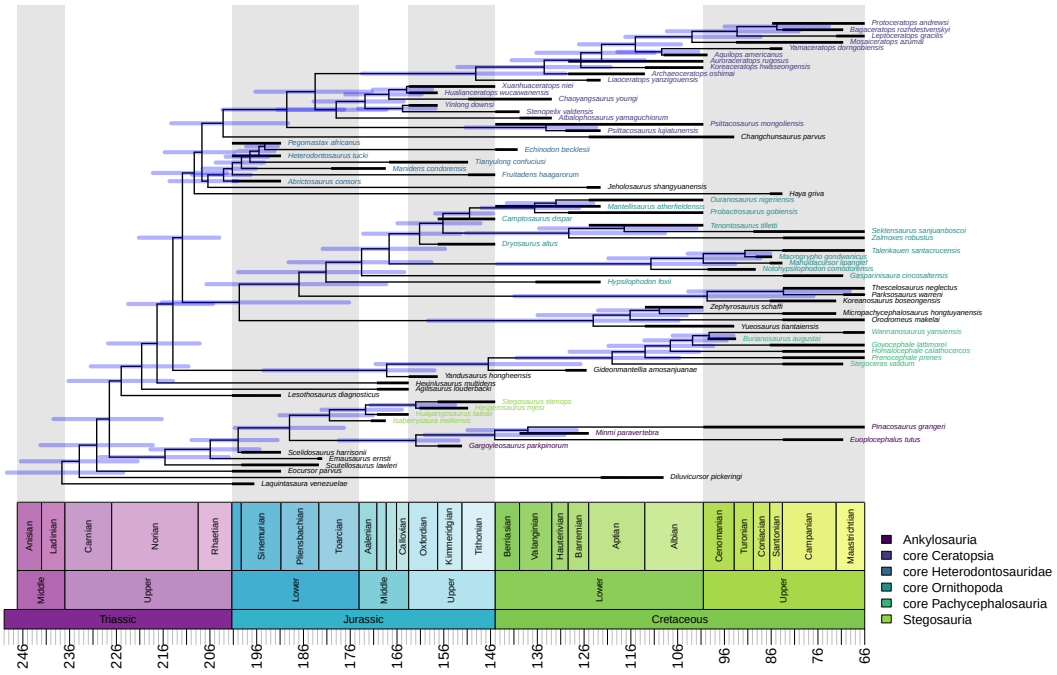

Figure C.4: Maximum clade credibility (MCC) tree from a BEAST 2 analysis of the HaEA dataset with a nonzero value of ρ and topology fixed to the RAxML estimate. The blue bars represent the 95% highest posterior density intervals for the divergence times; the black bars represent tip age uncertainty (for single-occurrence taxa) or tip age ranges inclusive of uncertainty (for taxa with multiple occurrences). All ages are in millions of years before the present (Ma). Unlabeled stages (from oldest to youngest): Hettangian, Bajocian, Bathonian.

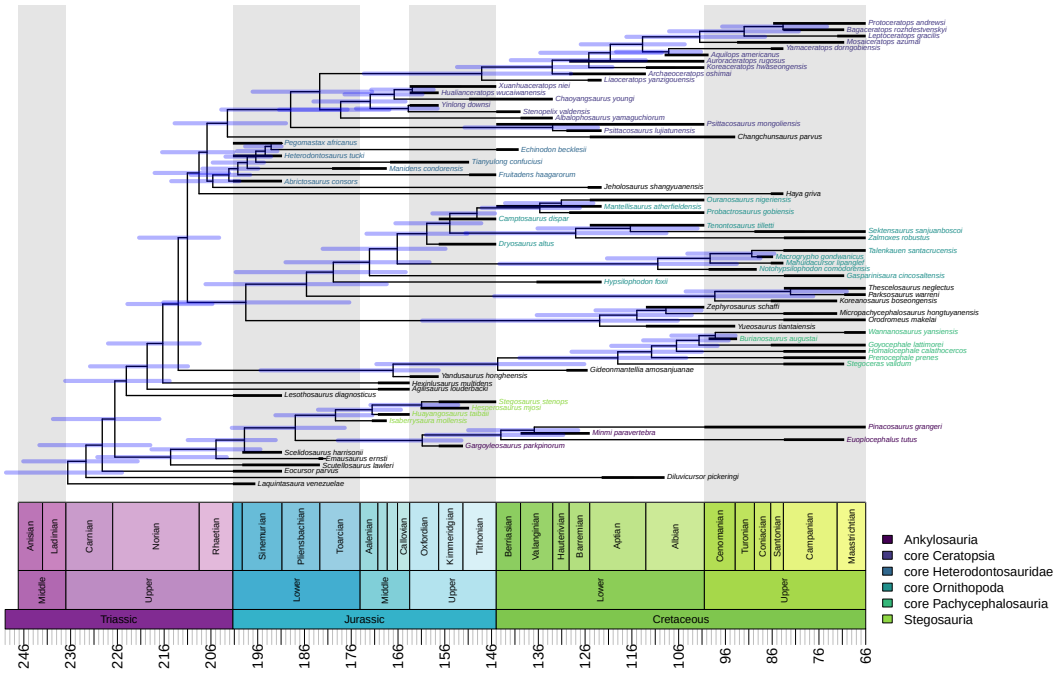

Figure C.5: MCC tree from a BEAST 2 analysis of the HaEA dataset with ρ set to 0 and topology fixed to the RAxML estimate.

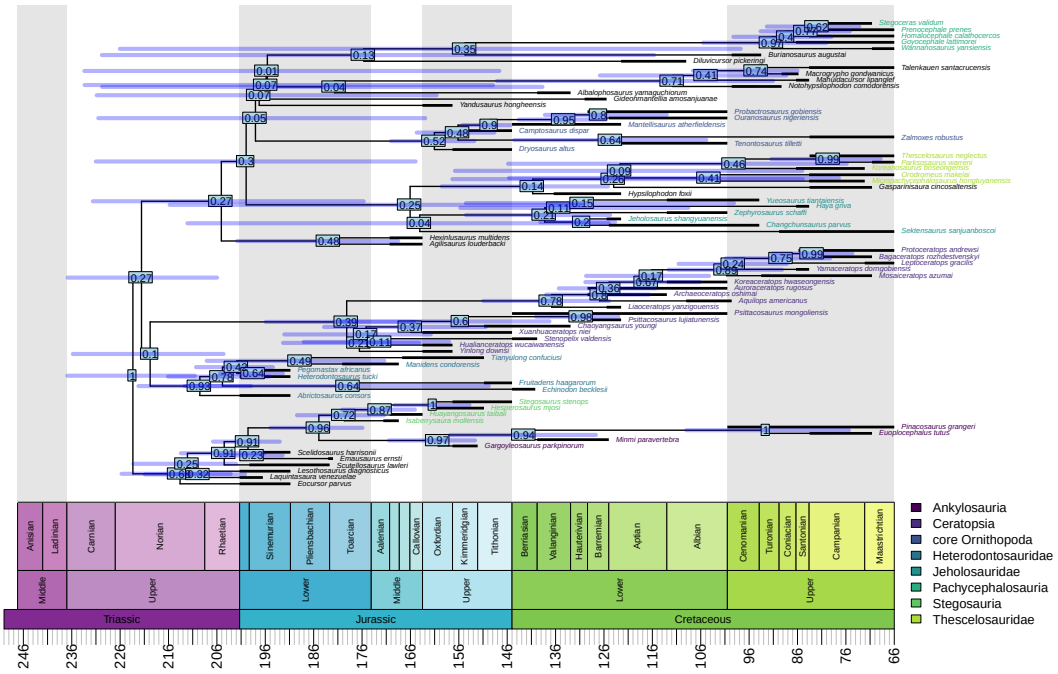

Figure C.6: MCC tree from a topologically unconstrained BEAST 2 analysis of the HaEA dataset with a nonzero value of ρ. The node labels denote posterior probabilities.

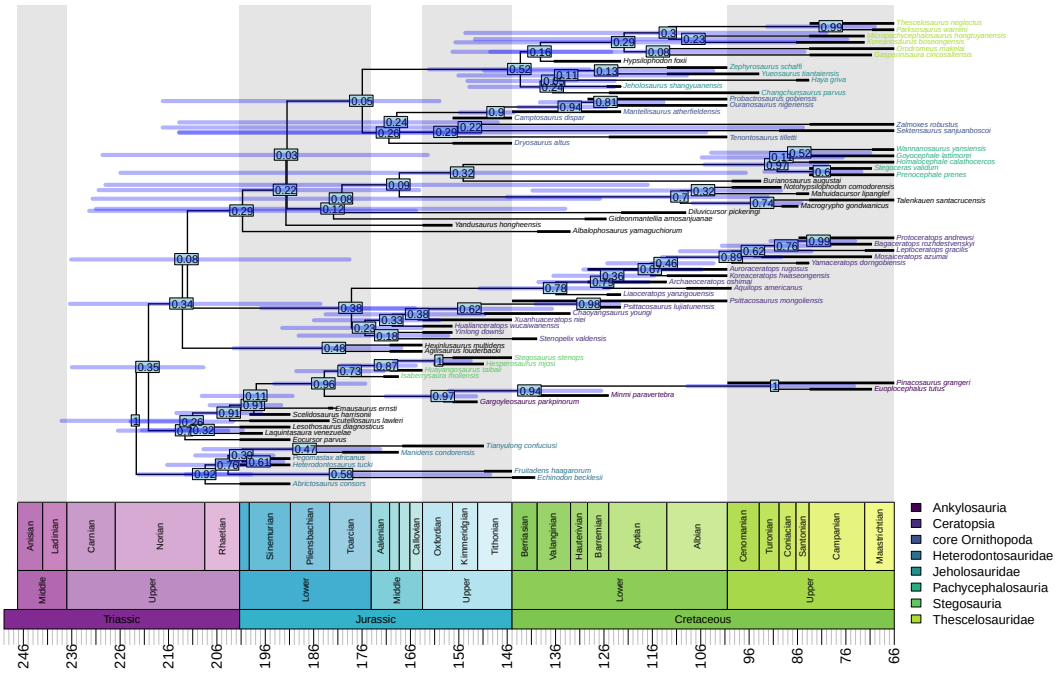

Figure C.7: MCC tree from a topologically unconstrained BEAST 2 analysis of the HaEA dataset with ρ set to 0.

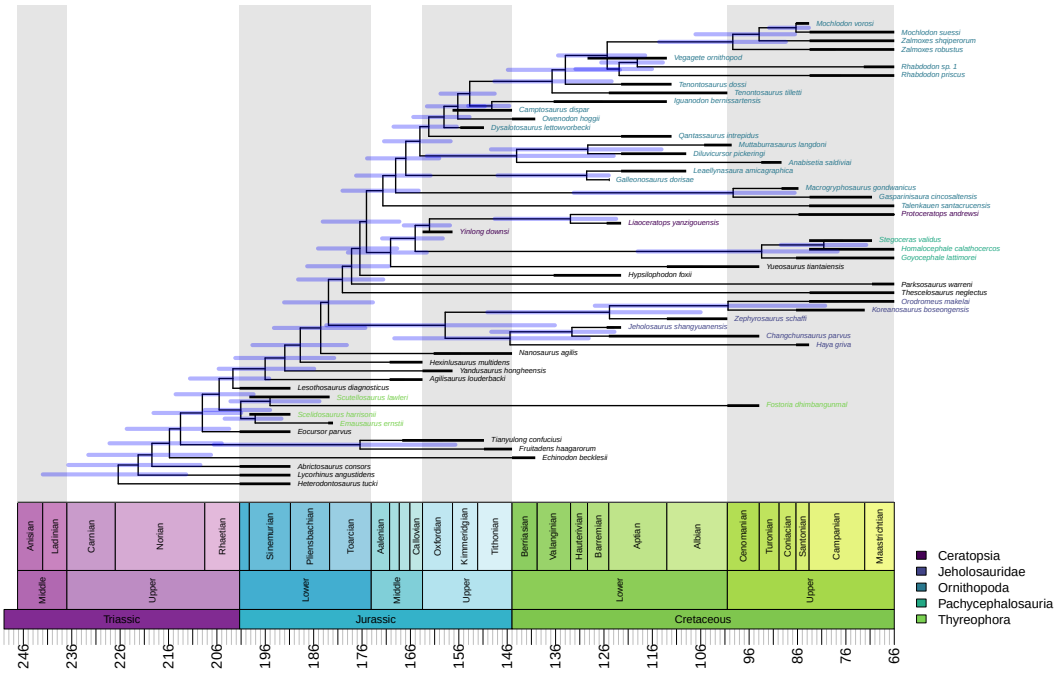

Figure C.8: MCC tree from a BEAST 2 analysis of the partitioned HeEA dataset with a nonzero value of ρ and topology fixed to the RAxML estimate.

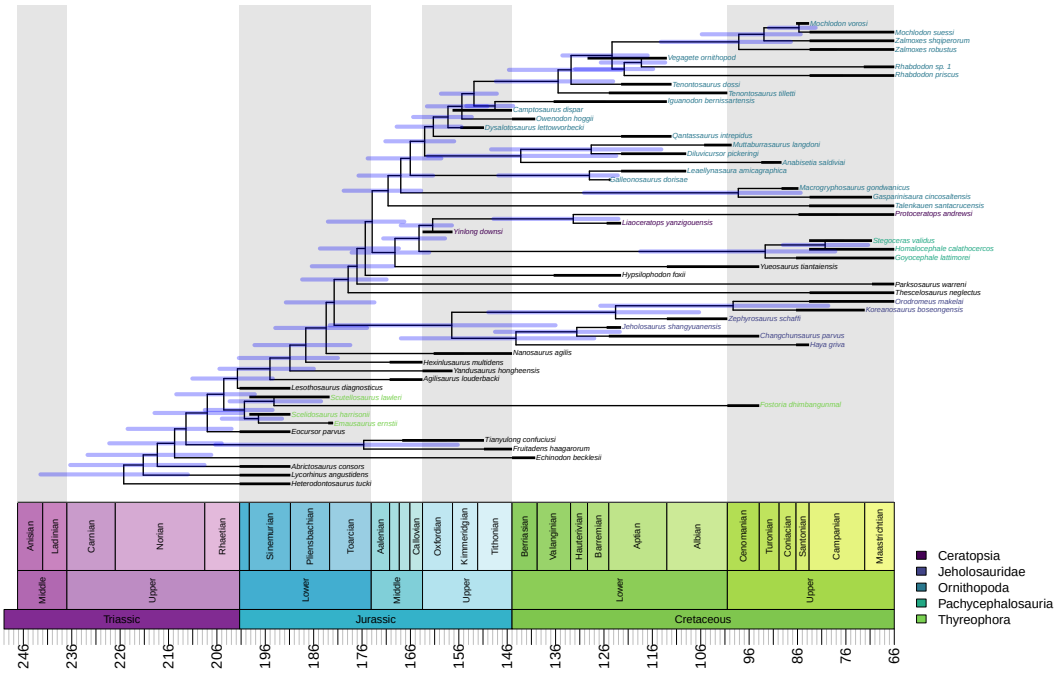

Figure C.9: MCC tree from a BEAST 2 analysis of the partitioned HeEA dataset with ρ set to 0 and topology fixed to the RAxML estimate.

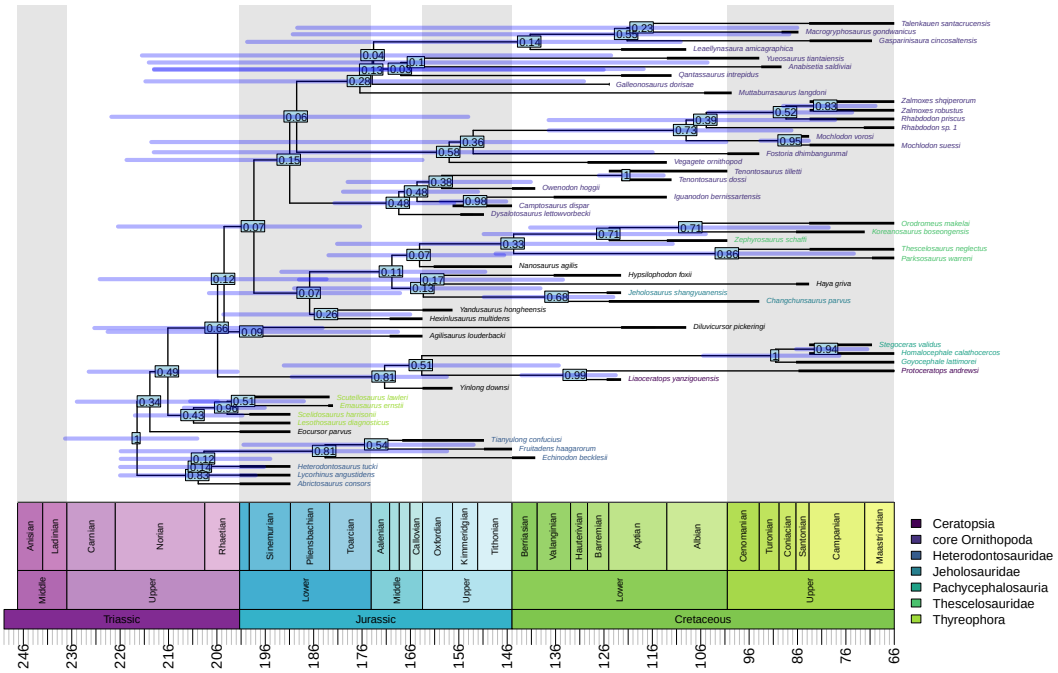

Figure C.10: MCC tree from a topologically unconstrained BEAST 2 analysis of the partitioned HeEA dataset with a nonzero value of ρ.

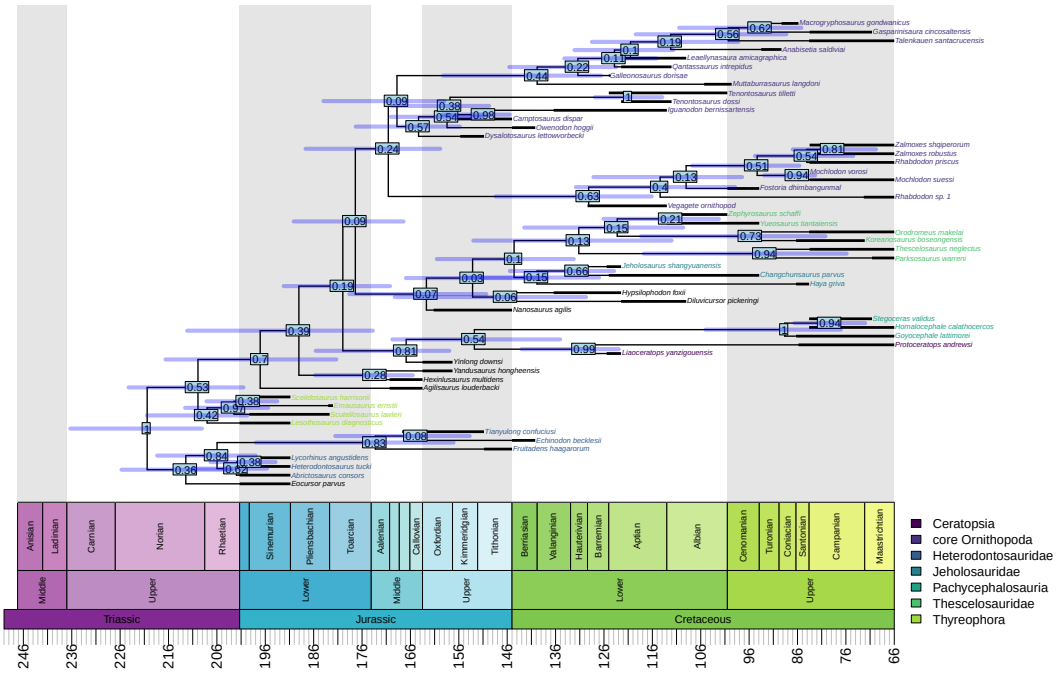

Figure C.11: MCC tree from a topologically unconstrained BEAST 2 analysis of the partitioned HeEA dataset with ρ set to 0.

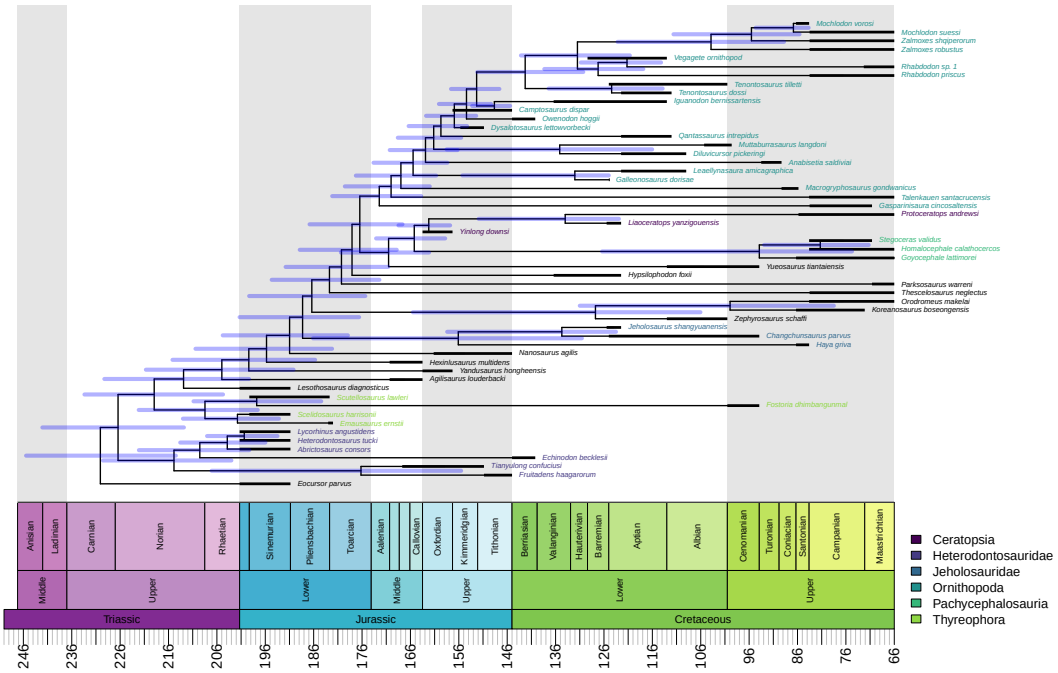

Figure C.12: MCC tree from a BEAST 2 analysis of the unpartitioned HeEA dataset with a nonzero value of ρ and topology fixed to the RAxML estimate.

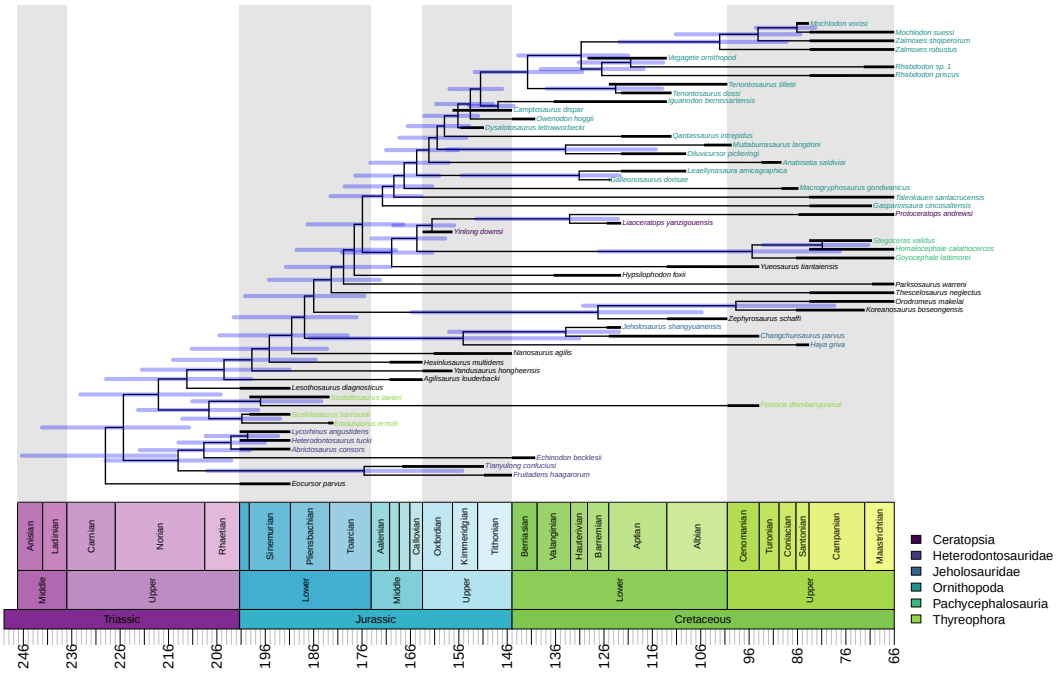

Figure C.13: MCC tree from a BEAST 2 analysis of the unpartitioned HeEA dataset with ρ set to 0 and topology fixed to the RAxML estimate.

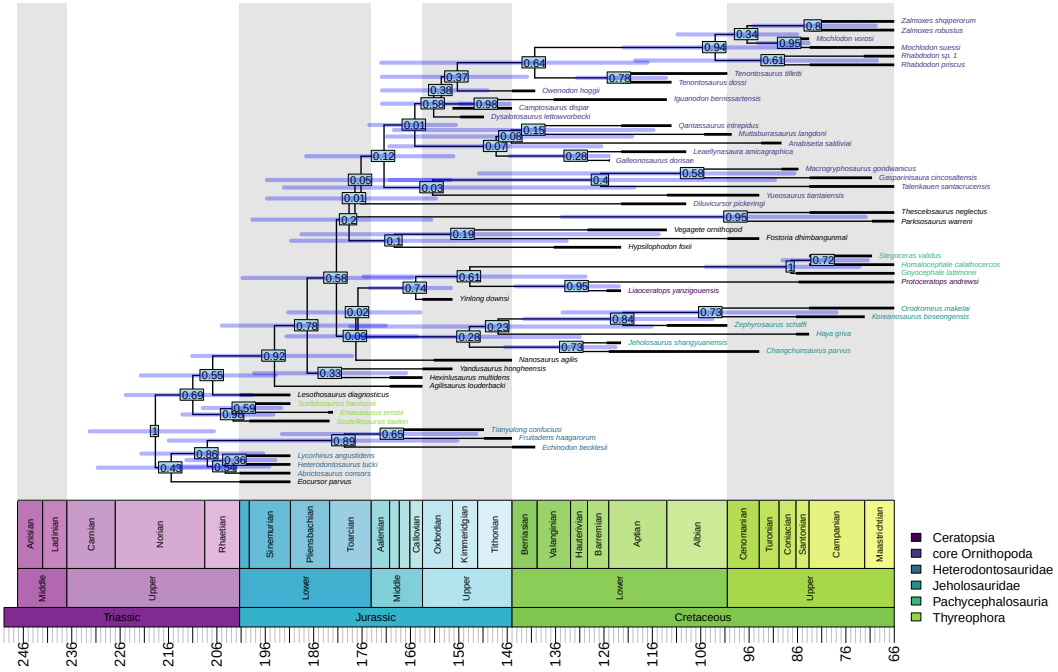

Figure C.14: MCC tree from a topologically unconstrained BEAST 2 analysis of the unpartitioned HeEA dataset with a nonzero value of ρ.

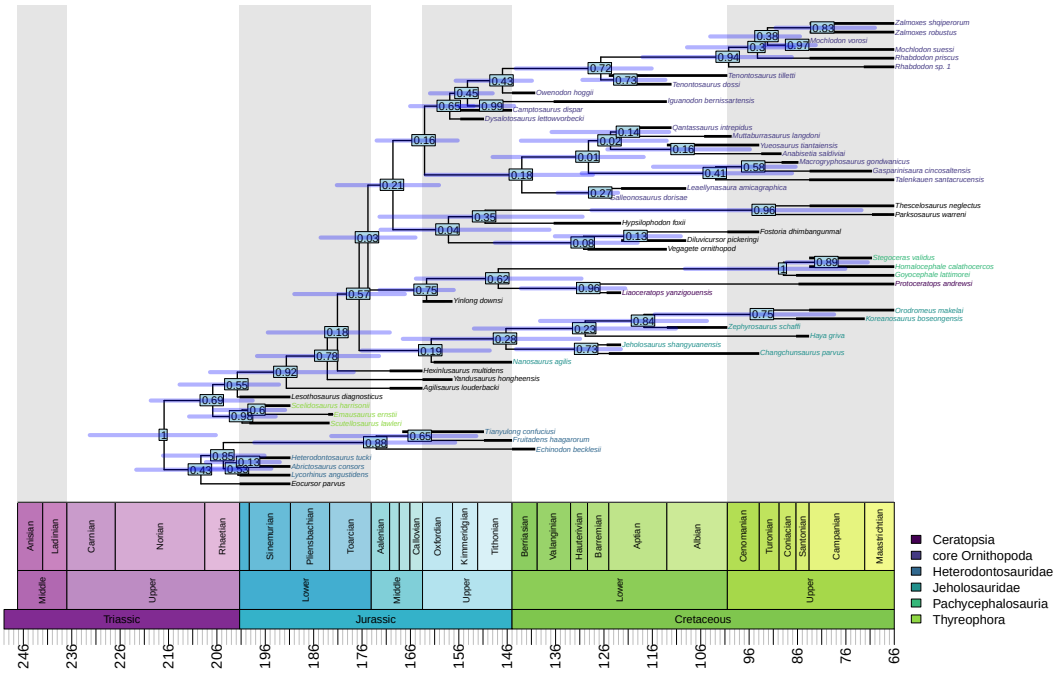

Figure C.15: MCC tree from a topologically unconstrained BEAST 2 analysis of the unpartitioned HeEA dataset with ρ set to 0.

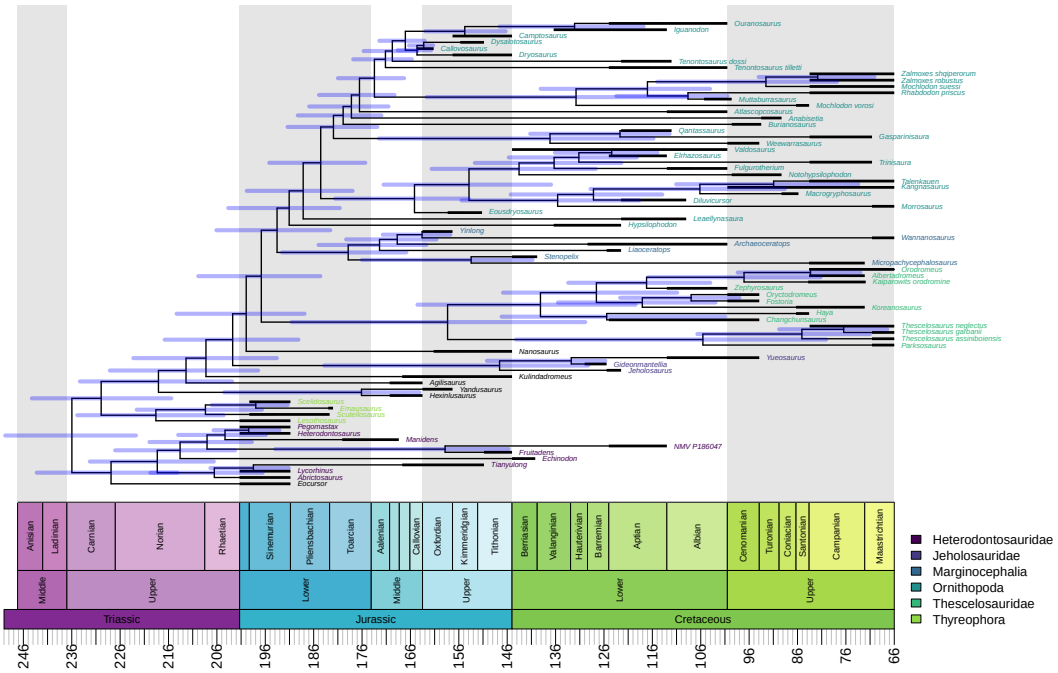

Figure C.16: MCC tree from a BEAST 2 analysis of the MEA dataset with a nonzero value of ρ and topology fixed to the RAxML estimate.

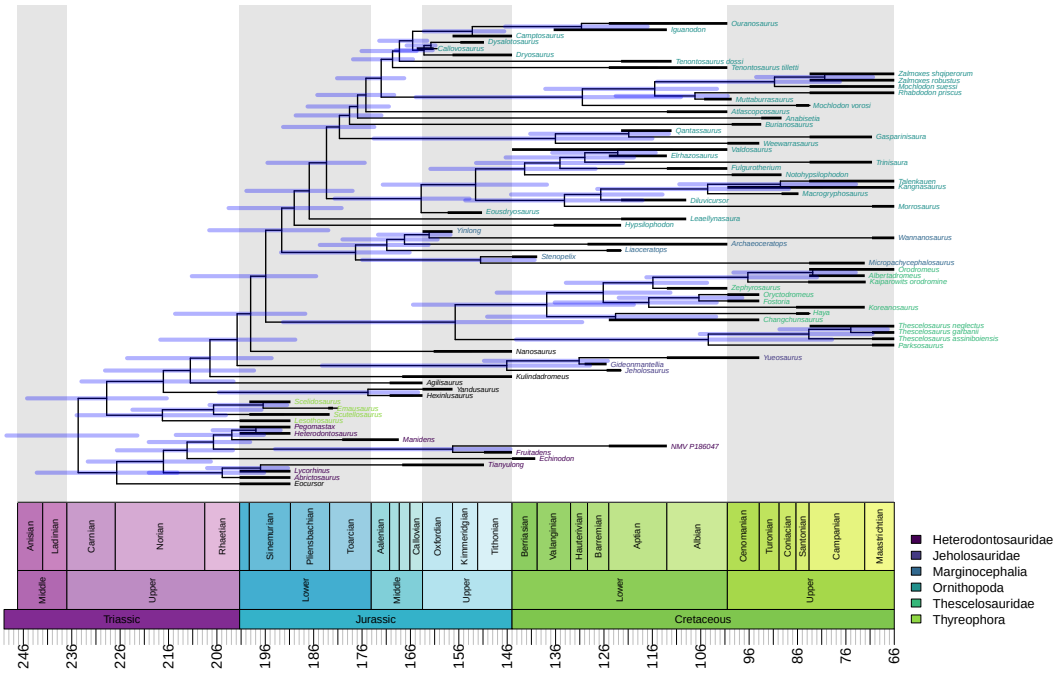

Figure C.17: MCC tree from a BEAST 2 analysis of the MEA dataset with ρ set to 0 and topology fixed to the RAxML estimate.

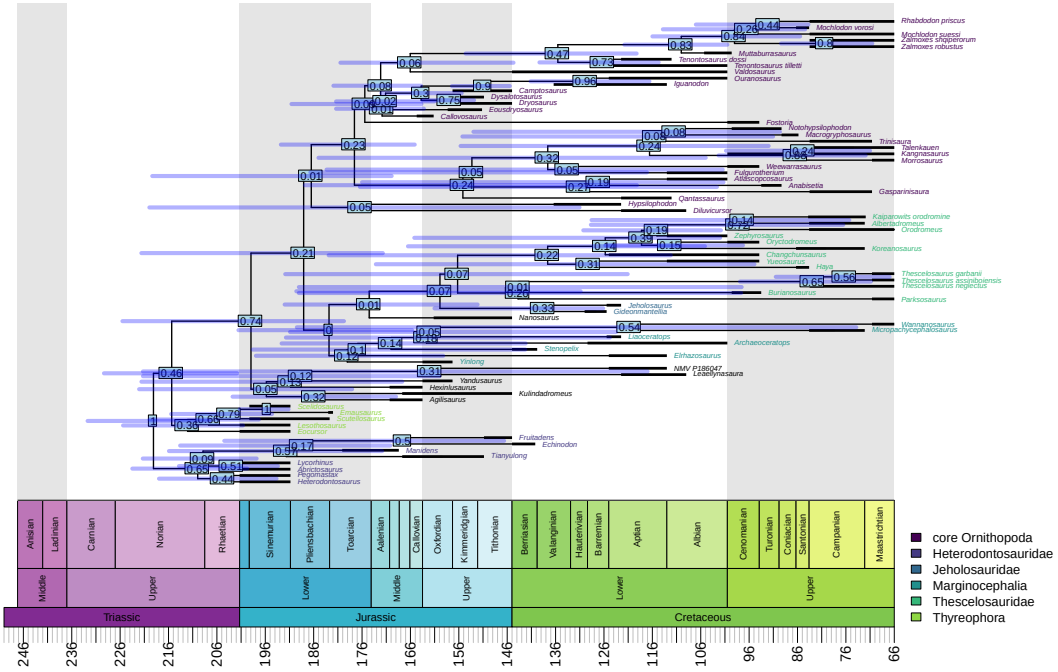

Figure C.18: MCC tree from a topologically unconstrained BEAST 2 analysis of the MEA dataset with a nonzero value of ρ.

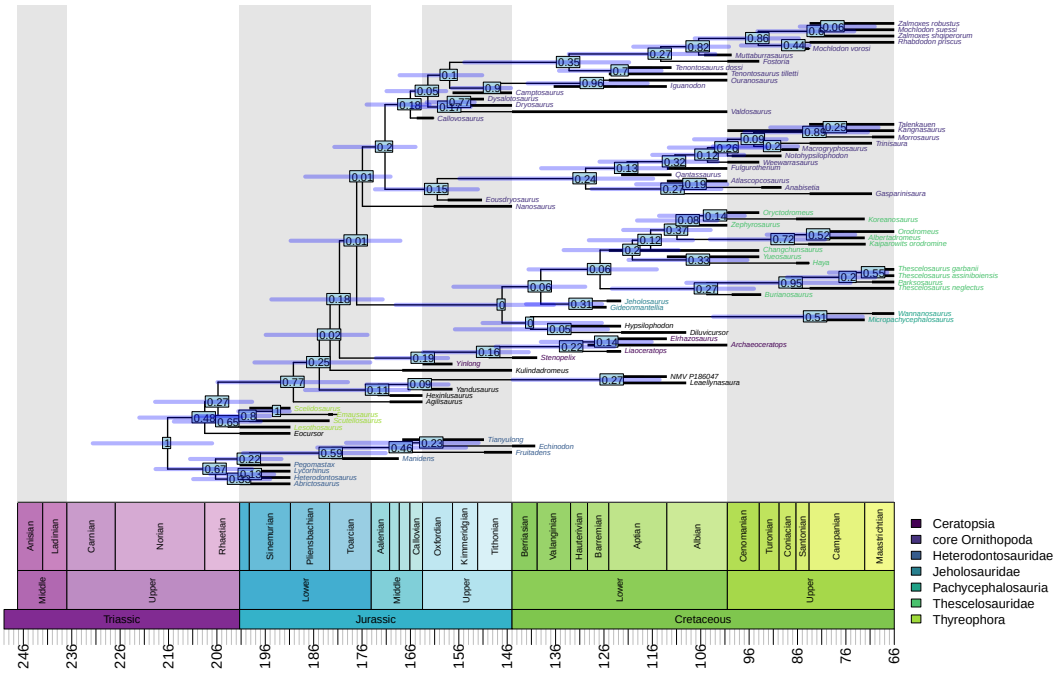

Figure C.19: MCC tree from a topologically unconstrained BEAST 2 analysis of the MEA dataset with ρ set to 0.

C.4 Diversification Rate Analyses

C.4.1 Preservation Rates

The Fossil BAMM estimates of the preservation rate ψ were nearly identical to the weighted mean (by stage duration) resulting from the sampling intensity analysis described in Section C.1.2 (mean of posterior means across all 96 BAMM analyses: 0.0993 occ·sp^{-1}·Myr^{-1}, range: [0.0653, 0.153] occ·sp^{-1}·Myr^{-1}). On average, the 95% credible intervals (CIs) inferred by BAMM fell outside of the sampling intensity-based CIs in just 1.9 out of the 22 stages for which the overlap could be assessed, further corroborating a high degree of agreement between both estimators. In contrast, the ψ estimates yielded by PyRate were substantially higher than the sampling intensity results regardless of how the rate was modeled. Similar to the corresponding rates of speciation and extinction (see Figure **??** in the main text), the preservation rates inferred under the nonhomogeneous Poisson (NHPP) model, which exhib-

ited the poorest fit to the ornithischian data (Table C.5), tended to be the lowest (range of time-averaged mean rates: [0.3371, 0.4071] occ·sp^{-1}·Myr^{-1}). The rates estimated under the homogeneous Poisson (HPP) model were consistently higher (range of means: [0.5356, 0.6109] occ·sp^{-1}·Myr^{-1}), while the best-fitting TPP model produced results that were intermediate between the other two models under the bdMCMC model-averaging algorithm (range of means weighted by stage duration: [0.4387, 0.5319] occ·sp^{-1}·Myr^{-1}), but exceeded both of them under the rjMCMC algorithm ([0.5633, 0.6716] occ·sp^{-1}·Myr^{-1}).

Preservation model	**Site-linked**	**Site-unlinked**
HPP	169.877	178.818
NHPP	204.761	218.099
TPP-age	0	0
TPP-period	137.463	144.499

Table C.5: AICc-based comparison of different preservation models implemented in PyRate, as applied to the ornithischian occurrence data. Values shown are the differences in the Akaike information criterion (ΔAICc) between different preservation models. HPP = homogeneous Poisson process; NHPP = nonhomogeneous Poisson process; TPP = time-variable Poisson process.

Unlike the sampling intensity analysis, PyRate estimated the preservation rate with high relative precision (determined as the width of the 95% CI divided by the mean), except when coupled with the TPP model. The average relative precisions under the HPP model (0.2446; note that only bdMCMC analyses reached convergence for this model) and the NHPP model (0.2247; bdMCMC: 0.2196, rjMCMC: 0.2339) were comparable to the corresponding BAMM value (mean across all 96 analyses: 0.2664), while the ψ values inferred under the TPP model were considerably less precise (average of mean relative precisions weighted by stage duration: 1.6087; bdMCMC: 1.6086, rjMCMC: 1.5885). This was due to the much wider 95% CIs yielded by the latter model, especially around the Early–Midle Jurassic boundary (Figs. C.20–C.23). Given the high complexity of the TPP model, it is plausible that the relatively small size of the ornithischian dataset limited the ability of PyRate to simultaneously estimate all of its parameters with high precision.

The combination of high mean rates coupled with relatively tight credibility intervals resulted in a poor overlap between the PyRate results and the ψ estimates resulting from the sampling intensity analysis (Section C.1.2). This incongruence was especially pronounced for the HPP model, where an overlap between the corresponding 95% CIs failed to occur in (on average) 17.6 out of the 22 or 21 stages for which it could be assessed; the values for the NHPP model (11.5) and the TPP model (5.5) were notably better, although none of the three models reached the same level of congruence as BAMM. Moreover, when coupled with the TPP model, PyRate also favored substantially higher levels of ψ heterogeneity than the sampling intensity analysis: the 95% CIs often showed little to no overlap between successive chronostratigraphic stages, with the gap between the Campanian and Maastrichtian CIs being especially prominent (Figs. C.20–C.23).

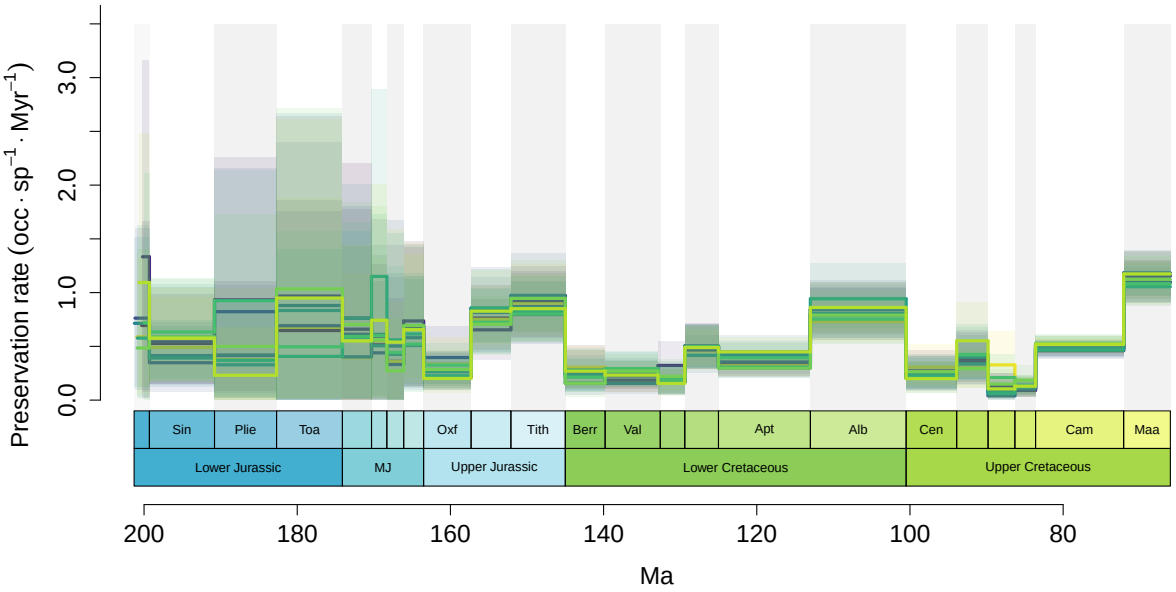

Figure C.20: Preservation rate-through-time (RTT) plot from the bdMCMC site-unlinked PyRate analyses of the ornithischian occurrence data under the time-variable Poisson preservation (TPP) model.

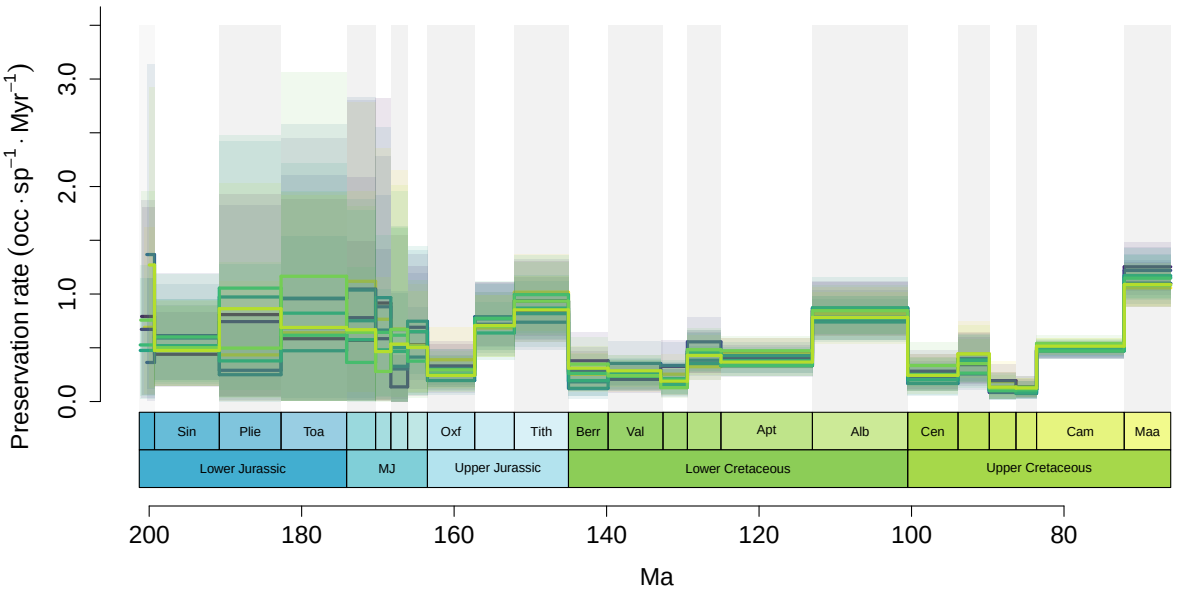

Figure C.21: Preservation RTT plot from the bdMCMC site-linked PyRate analyses of the ornithischian occurrence data under the TPP model.

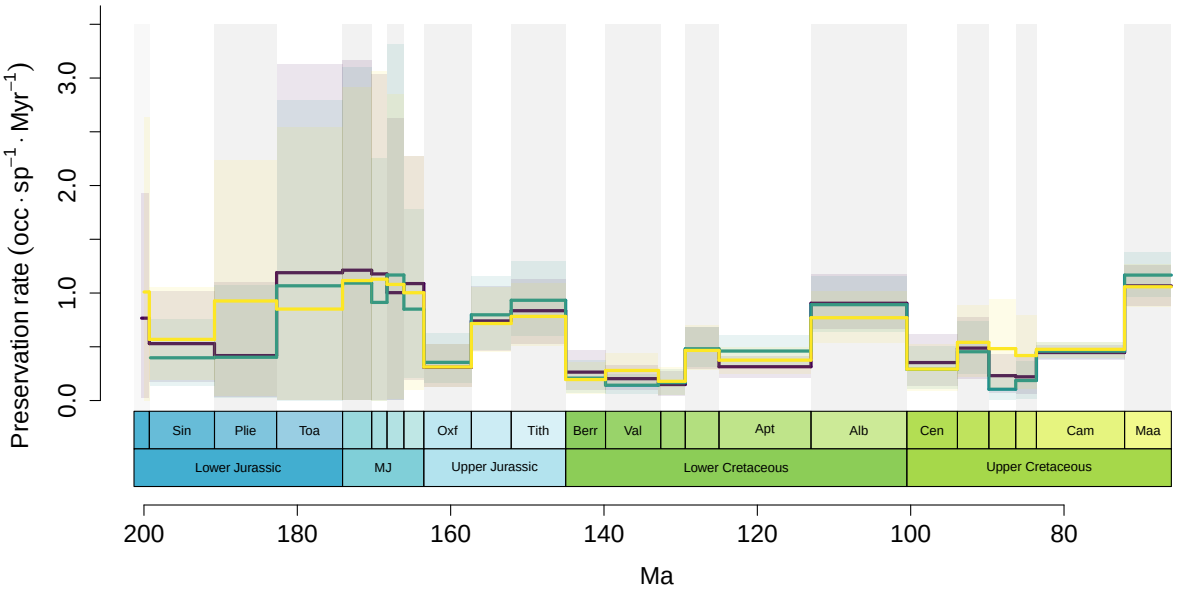

Figure C.22: Preservation RTT plot from the rjMCMC site-unlinked PyRate analyses of the ornithischian occurrence data under the TPP model.

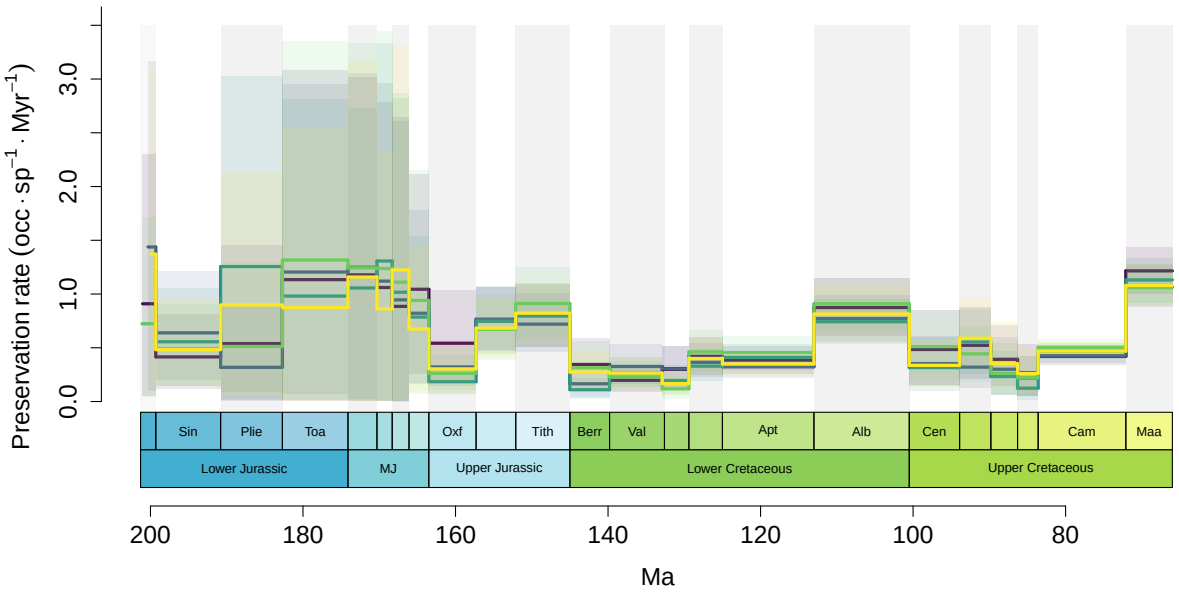

Figure C.23: Preservation RTT plot from the rjMCMC site-linked PyRate analyses of the ornithischian occurrence data under the TPP model.

C.4.2 PyRate Rate Shift Probabilities

The sampling frequencies of speciation and extinction rate shifts (approximating their posterior probabilities) can only be logged for rjMCMC analyses, which tended to infer substantially more shifts for the ornithischians than the corresponding bdMCMC runs (Table C.6). The rjMCMC rate shift probabilities were relatively low ($\sim$ 0.3–0.4) for the first period of elevated rates around the Early–Middle Jurassic boundary and for the edge effects associated with the clade's origin and extinction, but rose to values >0.8 for the end of the second period of elevated rates, coinciding with the Santonian–Campanian boundary (Figure C.24). The posterior probability of a shift was not consistently associated with its magnitude. Note that the Santonian–Campanian boundary is also associated with the largest cluster of first appearance dates (Figure C.1) and with the marked increase in the number of ornithischian-bearing formations and collections (Figure C.2), suggesting a possible inter-

action between heterogeneous preservation rates on the one hand and shifts in the rates of speciation and extinction on the other.

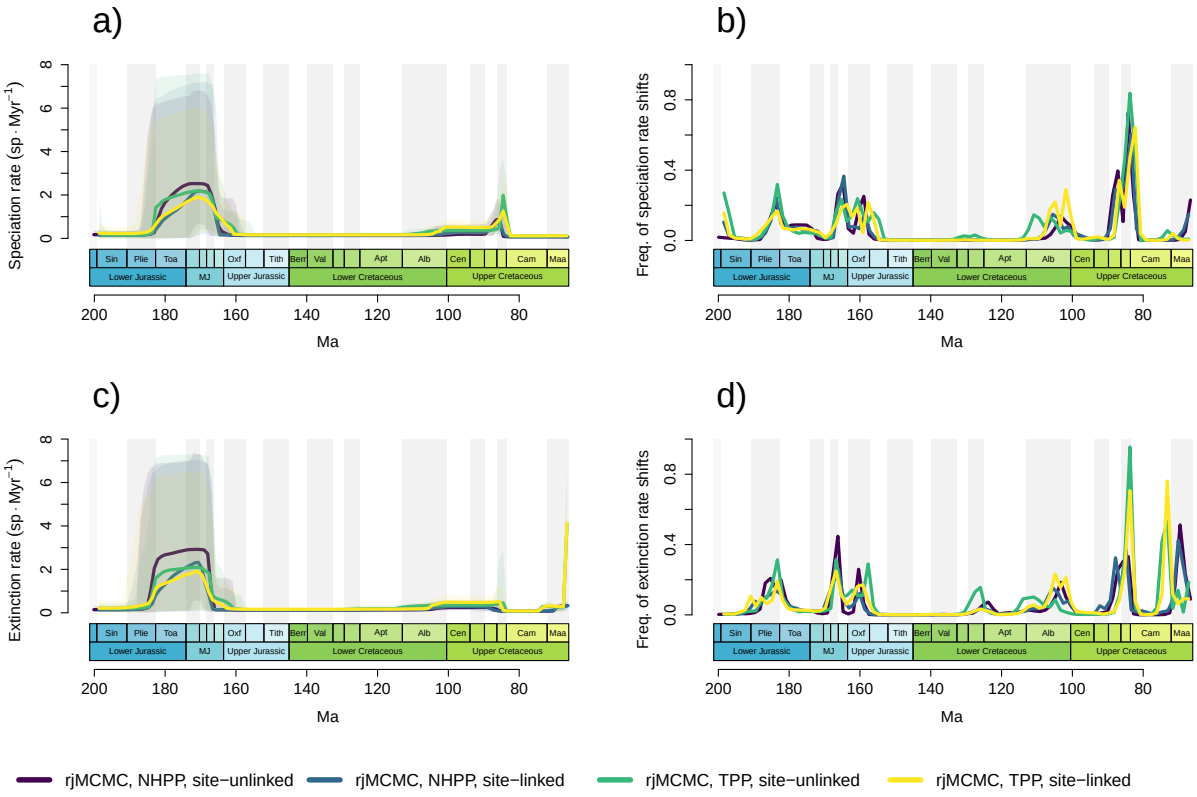

Figure C.24: Diversification of Ornithischia estimated using PyRate and the rjMCMC algorithm (cf. Figure 4 in the main text), with rate-through-time plots for a) speciation and c) extinction shown next to the posterior probabilities of b) speciation and d) extinction rate shifts.

	Site-linked					Site-unlinked				
λ rates	bdMCMC +HPP	bdMCMC +NHPP	bdMCMC +TPP	rjMCMC +NHPP	rjMCMC +TPP	bdMCMC +HPP	bdMCMC +NHPP	bdMCMC +TPP	rjMCMC +NHPP	rjMCMC +TPP
1	0.0553	0.0604	0.0819	0	0	0.0621	0.065	0.0872	0	0
2	**0.6403**	**0.5677**	**0.6825**	0	0	**0.6541**	**0.5894**	**0.6854**	0	0
3	0.257	0.3091	0.1975	0.0004	0	0.2426	0.2906	0.1933	0	0.0001
4	0.041	0.0522	0.0322	0.0124	0.0194	0.0354	0.0472	0.0289	0.0016	0.0009
5	0.0059	0.0092	0.0044	0.1956	0.1472	0.0057	0.0063	0.0047	0.2285	0.0142
6	0.0005	0.0012	0.0002	**0.3772**	**0.282**	0.0001	0.0015	0.0002	**0.3892**	0.1242
7	0	0.0002	0.0004	0.2757	0.2667	0.0001	0	0.0003	0.2735	**0.3826**
8	0	0	0.0008	0.1101	0.1854	0	0	0	0.0866	0.3119
9	0	0	0	0.0248	0.074	0	0	0	0.0176	0.1284
10	0	0	0	0.0034	0.0212	0	0	0	0.0028	0.0326
11	0	0	0	0.0005	0.0032	0	0	0	0.0001	0.0043
12	0	0	0	0	0.0006	0	0	0	0	0.0006
13	0	0	0	0	0.0001	0	0	0	0	0.0001
14	0	0	0	0	0.0001	0	0	0	0	0
μ rates										
1	0.1906	0.2363	0.1681	0	0	0.1754	0.2235	0.1553	0	0
2	**0.5172**	**0.5794**	**0.5885**	0	0	**0.5146**	**0.5831**	**0.5714**	0.0001	0
3	0.2258	0.1548	0.1922	0.0088	0	0.2362	0.1632	0.2175	0.0004	0
4	0.0533	0.0261	0.0433	0.0525	0	0.0585	0.0266	0.0462	0.0049	0
5	0.0107	0.003	0.0059	0.0992	0.0105	0.013	0.0031	0.009	0.0616	0.0043
6	0.0022	0.0002	0.0015	**0.4062**	0.0545	0.0024	0.0003	0.0005	**0.5227**	0.0093
7	0.0003	0.0001	0.0005	0.311	0.3382	0	0.0001	0.0001	0.3135	0.1673
8	0	0	0	0.0983	**0.3433**	0	0	0	0.0814	**0.4176**
9	0	0	0	0.0205	0.1751	0	0	0	0.0136	0.2802
10	0	0	0	0.0032	0.0601	0	0	0	0.0015	0.0946
11	0	0	0	0.0002	0.0149	0	0	0	0.0001	0.024
12	0	0	0	0.0001	0.0032	0	0	0	0	0.0024
13	0	0	0	0	0.0002	0	0	0	0	0.0003
14	0	0	0	0	0	0	0	0	0	0

Table C.6: Posterior probabilities of models containing different numbers of speciation and extinction rates in the PyRate analyses of the ornithischian occurrence data. For each combination of the model-averaging algorithm, preservation model, and site treatment used, the values are averaged across all replicates that reached convergence ($\leq$10 parameters with effective sample sizes of $<$200). The best-performing model is shown in bold type. λ = speciation rate; μ = extinction rate; bdMCMC = birth-death Markov chain Monte Carlo; rjMCMC = reversible-jump Markov chain Monte Carlo.

C.5 Simulations

C.5.1 Birth-Death Trees

We used the R package TreeSim v2.4 (Stadler, 2011b) to simulate 100 birth-death trees under the Decreasing Speciation (time-varying λ, constant ψ) and Decreasing Preservation (time-varying ψ, constant λ) scenarios, conditioning on the number of "extant" tips. This

number was set to 79, corresponding to the number of ornithischian species whose age ranges extend to the K–Pg boundary. We employed rejection sampling to ensure that the size and root age of the simulated trees differed by a factor of no more than 1.5 from the number of known ornithischian species (266–597 tips) and from the time span of the ornithischian fossil record ($\leq$203.0 Myr), respectively. We also imposed an additional requirement that all sampled trees be more than 80 Myr old to ensure that each tree underwent the 70 Ma rate shift and that the periods both before and after the shift were sufficiently sampled. The speciation and extinction rates were chosen to satisfy these criteria; we used the values of $\lambda_1 = 0.2$ sp·Myr^{-1}, $\lambda_2 = 0.05$ sp·Myr^{-1}, $\mu = 0.07$ sp·Myr^{-1} for the Decreasing Speciation scenario and $\lambda = 0.1$ sp·Myr^{-1}, $\mu = 0.07$ sp·Myr^{-1} the Decreasing Preservation scenario. In both scenarios, trees were simulated on the assumption of complete sampling of extant taxa ($\rho = 1$). We observed that the trees generated under the former scenario tended to be younger and more speciose than the ornithischian dataset, while the simulations performed under the latter scenario exhibited a wider range of root ages and tree sizes. For further analyses, we therefore chose the 10 smallest trees whose root age exceeded 100 Myr for Decreasing Speciation, and 10 random trees for Decreasing Preservation. This multi-stage selection procedure produced tree sets for both scenarios whose tip number and root age distributions had different central tendencies but overlapping ranges, with the empirical values included in the area of overlap (Table C.7).

C.5.2 Fossil Records

Using the R package FossilSim v2.1.1 (Barido-Sottani et al., 2019b), we simulated fossil data on the synthetic trees under a Poisson process with a constant (Decreasing Speciation) or piecewise-constant (Decreasing Preservation) preservation rate ψ. We imposed a one-to-one mapping between edges and species, equivalent to the assumption that speciation occurs only via bifurcation rather than anagenesis (multiple species per edge) or budding

	Empirical data	Decreasing Speciation											
Replicate		1	2	3	4	5	6	7	8	9	10	Mean	Median
Tips		484	466	473	487	500	395	479	489	515	492	478	486
Sampled tips		393	354	387	385	398	304	393	385	408	389	380	388
Sampled ancestors		352	346	350	378	376	302	360	361	376	360	356	360
Sampled species (total)	389	745	700	737	763	774	606	753	746	784	749	736	748
Sampled ancestors / sampled species		0.472	0.494	0.475	0.495	0.486	0.498	0.478	0.484	0.480	0.481	0.484	0.483
Tips extant at K–Pg		79	79	79	79	79	79	79	79	79	79	79	79
Sampled tips extant at K–Pg	79	64	63	74	66	63	60	64	67	66	66	65	65
Age of origin		126.3	105.6	103.3	112.3	106.3	106.9	110.6	107.9	124.6	112.2	111.6	109.2
Root age		120.0	105.4	101.0	110.6	100.2	105.4	108.0	107.6	106.7	111.2	107.6	107.2
Sampled age	135.3	125.8	105.2	100.6	111.9	106.1	104.5	110.4	106.0	123.7	111.1	110.5	108.2
Edges at 136 Ma with >10 sampled tips		9	11	7	11	11	9	8	9	13	9	10	9
Colless index		0.045	0.052	0.046	0.044	0.039	0.055	0.056	0.040	0.036	0.041	0.045	0.045
Occurrences	1240	3663	3302	3654	3677	3890	2970	3715	3604	3902	3651	3603	3658
Occurrences along external edges		2086	1670	2064	2071	2195	1557	1996	1909	2193	1951	1969	2030
Median # of occurrences per species	1	3	3	3	3	3	3	3	3	3	3	3	3
St. dev. of occurrences per species	5.223	4.68	4.367	4.460	4.592	4.746	4.303	4.480	4.815	4.625	4.66	4.573	4.608
Proportion of singletons	0.608	0.239	0.220	0.216	0.221	0.222	0.188	0.226	0.233	0.239	0.215	0.222	0.222

		Decreasing Preservation											
Replicate		1	2	3	4	5	6	7	8	9	10	Mean	Median
Tips		313	459	273	343	268	332	295	294	317	371	327	315
Sampled tips		195	304	146	217	172	195	178	190	191	236	202	193
Sampled ancestors		191	300	161	217	160	212	182	173	184	239	202	188
Sampled species (total)	389	386	604	307	434	332	407	360	363	375	475	404	381
Sampled ancestors / sampled species		0.495	0.497	0.524	0.5	0.482	0.521	0.506	0.477	0.491	0.503	0.5	0.499
Tips extant at K–Pg		79	79	79	79	79	79	79	79	79	79	79	79
Sampled tips extant at K–Pg	79	55	50	37	52	42	40	46	50	53	41	47	48
Age of origin		135.1	146.7	126	125.0	121.7	115.2	120.8	130.0	145.8	168.7	133.5	128
Root age		132.9	141.0	125.8	124.2	121.1	112.9	104.5	119.5	125.1	163.3	127.0	124.7
Sampled age	135.3	132.5	140.9	125.6	123.8	120.9	112.7	104.1	119.2	124.7	163.0	126.7	124.2
Edges at 136 Ma with >10 sampled tips		4	9	6	4	4	8	4	3	3	8	5	4
Colless index		0.101	0.078	0.085	0.094	0.101	0.083	0.093	0.079	0.101	0.103	0.092	0.094
Occurrences	1240	1362	1981	973	1461	1068	1345	1066	1092	1288	1977	1361	1316
Occurrences along external edges		685	974	439	783	505	582	522	539	634	944	661	608
Median # of occurrences per species	1	2	2	2	2	2	2	2	2	2	2	2	2
St. dev. of occurrences per species	5.223	3.701	3.374	3.885	4.084	3.41	3.047	2.696	3.207	4.265	4.873	3.654	3.556
Proportion of singletons	0.608	0.358	0.348	0.371	0.408	0.346	0.324	0.353	0.364	0.376	0.301	0.355	0.356

Table C.7: Properties of the birth-death trees and fossil records simulated under a time-variable speciation rate and a constant fossil sampling rate (Decreasing Speciation) or under a constant speciation rate and a time-variable fossil sampling rate (Decreasing Preservation). Sampled age = difference between the ages of the first and last occurrence; Edges at 136 Ma with >10 sampled tips = the number of edges crossing the 136 Ma time horizon that subtend clades of more than 10 sampled tips; Colless index = the tree imbalance index of Colless (1982) normalized for tree size; Proportion of singletons = the fraction of species represented by a single occurrence out of the total number of sampled species.

(multiple edges per species). We generated one fossil record per phylogeny. To obtain an empirically justified value of ψ, we extracted the posterior medians of the preservation rate parameter from all site-linked PyRate analyses of the ornithischian data which were run under the corresponding (i.e., homogeneous Poisson) preservation model and for which the effective sample size of the parameter exceeded 200. Since the preservation rate estimated by PyRate and the rate argument of the FossilSim simulation functions are both given in the units of occurrences per species per 1 Myr ($\text{occ}\cdot\text{sp}^{-1}\cdot\text{Myr}^{-1}$; Silvestro et al., 2014b; Barido-Sottani et al., 2019b), we used the median of the unscaled posterior medians as the

rate under Decreasing Speciation ($\psi = 0.567$ occ·sp^{-1}·Myr^{-1}) and as the geometric mean of the two rates used under Decreasing Preservation ($\psi_1 = 1.135$ occ·sp^{-1}·Myr^{-1}, $\psi_2 = 0.284$ occ·sp^{-1}·Myr^{-1}). Exact sampling times were recorded for all simulated occurrences to eliminate stratigraphic age uncertainty. To better approximate the empirical case, we added 66 Myr to the ages of all occurrences, effectively shifting the entire fossil record into the past and changing the age of the true shift to 136 Ma.

C.5.3 Fossil BAMM

In both scenarios, we applied Fossil BAMM to the true (generating) phylogenies under three sampling schemes to explore the impact of tree size on the accuracy of rate inference while eliminating phylogenetic error. In the basic ("B") analyses, only unsampled tips (without a fossil occurrence) were removed. In the proportionally subsampled ("PS") runs, only 15% of sampled tips were retained in the tree, approximately corresponding to the proportion of the known ornithischian species included in the HaEA (0.186), HeEA (0.136), and MEA (0.186) datasets. The tips to be retained were randomly selected under the constraint that the most recent common ancestor (MRCA) of the resulting taxon set be identical to the MRCA of the full sampled tree. Under the 70-tip ("T70") sampling scheme, said number of tips was kept in the tree to approximate the size of the ornithischian trees (HaEA: 74 tips, HeEA: 54 tips, MEA: 74 tips). Tip sampling choice was subject to the same constraint as in the PS scheme. In all three schemes, true values were used for the global sampling fraction (B: 1, PS: 0.15, T70: tree-specific). Terminal edges were truncated to the age of their last occurrence following Mitchell et al. (2018), and fossils associated with the internal edges (sampled ancestors) were included in the total number of occurrences specified for each analysis.

For each combination of tree and sampling scheme, we performed two analyses, with the expected number of rate shifts set either to 0.1 or to the number of those edges crossing

the 136 Ma time horizon that gave rise to more than 10 sampled tips. The former value placed high prior mass on the zero-shift configuration, corresponding to the true rate shift configuration in the Decreasing Preservation scenario. Under Decreasing Speciation, a single tree-wide shift in the speciation rate at 136 Ma should ideally induce a tree-specific number of near-simultaneous branch shifts in the Fossil BAMM analyses, with the caveat that the method can reliably detect only those shifts that give rise to regimes with a sufficient number of tips (e.g., $n > 10$; Rabosky et al., 2017). As a result, setting the expected number of shifts to the number of branches that cross the 136 Ma time horizon and subtend sufficiently large clades should favor the recovery of the true rate shift configuration under Decreasing Speciation. Finally, each combination of speciation rate scenario, tree (10 per scenario), sampling scheme (3 per tree), and expected number of shifts (2 per sampling scheme) was subject to both time-constant and time-varying analyses, for a total of 240 Fossil BAMM runs. The analyses were run for 50 million generations, the first 1 million of which were discarded as burnin following the quick convergence observed in the empirical analyses. To speed up post-processing, each posterior was thinned to 500 post-burnin samples. Convergence diagnostics were identical to those described for the empirical analyses.

C.5.4 PyRate

We applied two different treatments of sampled ancestors to each simulated fossil record. Sampled ancestors were common in the synthetic data (Table C.7), complicating our efforts to bring the simulations as close as possible to the ornithischian fossil record. First, sampled ancestors inflate the total number of species sampled beyond that known for the ornithischians, especially in the Decreasing Speciation scenario (Table C.7). Second, a fossil record rich in sampled ancestors may not be realistic for the ornithischians, where we found no evidence for them in our tip-dating analyses (Section C.3.3) and where species are often described based on autapomorphies, which sampled ancestors are expected to lack (though see Wagner,

1998). Therefore, we converted each simulated fossil record into two PyRate datasets: one that included all occurrences, and one that included only those associated with the tips.

Decreasing Speciation								
Replicate	Including sampled ancestors				Excluding sampled ancestors			
	HPP	NHPP	TPP-136	TPP-by10	HPP	NHPP	TPP-136	TPP-by10
1	19451.457	19763.332	19378.888	**19355.421**	11352.059	11535.677	11322.697	**11321.354**
2	17276.73	17565.398	**17255.763**	17264.721	8808.985	8970.566	**8802.647**	8812.629
3	19422.413	19727.451	19394.373	**19393.12**	11255.708	11430.632	**11246.866**	11255.66
4	19606.554	19845.254	19532.308	**19531.926**	11598.198	11759.951	**11580.492**	11591.794
5	20752.082	21097.369	**20721.148**	20724.058	12073.152	12307.303	**12063.327**	12070.289
6	15879.623	16077.559	**15811.648**	15817.035	8439.209	8555.25	**8419.826**	8429.594
7	19679.441	19959.479	**19626.123**	19627.926	10730.622	10901.157	**10724.782**	10734.433
8	19548.787	19857.259	19491.409	**19483.947**	10558.989	10739.483	10544.502	**10535.014**
9	21082.548	21416.274	**21032.414**	21039.181	12144.115	12381.002	12127.855	**12127.056**
10	19448.044	19769.055	**19389.565**	19398.912	10473.438	10648.189	**10452**	10457.836
Decreasing Preservation								
Replicate	Including sampled ancestors				Excluding sampled ancestors			
	HPP	NHPP	TPP-136	TPP-by10	HPP	NHPP	TPP-136	TPP-by10
1	6435.604	6502.513	**6230.615**	6242.321	3282.853	3322.468	**3204.704**	3222.404
2	9497.524	9581.382	**9221.043**	9243.706	4634.022	4654.287	**4507.178**	4530.64
3	4681.073	4722.852	**4553.131**	4556.665	2141.81	2147.376	**2086.248**	2102.408
4	6841.485	6896.308	**6659.753**	6672.205	3785.764	3812.319	**3685.47**	3700.143
5	4879.573	4919.365	**4749.382**	4758.882	2232.729	2240.458	**2178.873**	2184.876
6	6408.087	6471.124	**6242.48**	6243.237	2707.69	2728.251	**2653.722**	2668.306
7	5066.368	5103.07	**4900.752**	4902.606	2556.499	2579.169	**2510.79**	2517.552
8	5035.816	5081.689	4915.946	**4912.511**	2401.015	2416.739	**2346.701**	2359.922
9	6167.323	6218.775	**5979.116**	5985.898	3162.753	3194.467	**3041.369**	3050.292
10	9949.522	10052.242	**9595.599**	9609.477	4741.276	4790.92	**4557.72**	4575.408

Table C.8: AICc-based comparison of different preservation models implemented in PyRate, as applied to the synthetic data generated under Scenarios 1 and 2 before and after the removal of sampled ancestors. Values shown represent the Akaike information criterion (AICc) scores of different preservation models. For each combination of scenario, replicate, and sampled ancestor treatment, the true model is highlighted in blue, while the best-performing model is shown in bold type. HPP = homogeneous Poisson process; NHPP = nonhomogeneous Poisson process; TPP-136 = time-variable Poisson process with a single shift at 136 Ma, TPP-by10 = time-variable Poisson process with shifts occurring every 10 Myr.

For each of the 40 resulting datasets, we performed AICc-based model selection between the HPP, TPP-136, and TPP-by10 preservation models as described in the main text. The best-performing model differed from the true model (Decreasing Speciation: HPP, Decreasing Preservation: TPP-136) for all 10 replicates in the Decreasing Speciation scenario and one replicate in the Decreasing Preservation scenario (Table C.8). In these cases, analyses were performed under both preservation models. Among-lineage preservation rate heterogeneity

was modeled using a 4-category discrete Γ distribution, except in the Decreasing Speciation TPP analyses where its use caused overflow issues. Default priors were used for all analyses, including the $\Gamma(1.5, 1.5)$ prior on the vector of preservation rates under the TPP models and the $\Gamma(1.1, 1.1)$ prior on the vectors of speciation and extinction rates. Each analysis was repeated using the bdMCMC and rjMCMC algorithms, for a total of 122 PyRate runs. All MCMC settings, convergence diagnostics, and the calculation of summary statistics followed those described for the empirical analyses.

C.5.5 Supplementary Results and Discussion

Subsampled BAMM analyses

Subsampled BAMM analyses were even less successful in detecting the speciation rate shift under Decreasing Speciation than the analyses employing the B sampling scheme. Of the 80 PS and T70 analyses, only 1 included a shift in its MAP configuration (Table C.9). As in the B-scheme analyses, the shift comprised simultaneous accelerations in speciation and extinction, and occurred at a much later time than the true speciation drop (Figure S35). Similarly, configurations with larger numbers of shifts were included in the 95% credible set less often than under the B sampling scheme (Table C.9), suggesting that the power to detect shifts increases with tree size.

Similar to the results obtained under the B scheme, a majority of the subsampled analyses (17 out of 20 for both the PS and T70 schemes) correctly inferred the general tendency for the speciation rate to decay from the root toward the tips. However, in contrast to the generally accurate rate B-scheme estimates, the PS and T70 sampling schemes yielded underestimates of the speciation rates at the tips and at the root in time-varying analyses. Similarly, time-

Decreasing Speciation	Shift config. in the 95% credible set	Core shifts in the MAP configuration	Mean speciation rate ($sp \cdot Myr^{-1}$)	Mean extinction rate ($sp \cdot Myr^{-1}$)	Tip net div. rates significantly different from root net div. rates
Time-constant	4.1 [1–11]	0	0.0818	0.0819	—
Time-constant, true exp. # of shifts	1.9 [1–4]	0.1 [0–1]	0.0818	0.0818	—
Time-varying	11.1 [1–29]	0.1 [0–1]	0.1848 / 0.0426	0.0814 / 0.0814	− (10)
Time-varying, true exp. # of shifts	17.6 [1–79]	0.5 [0–2]	0.1915 / 0.0449	0.0825 / 0.0840	− (7), 0 (3)
PS, time-constant	1	0	0.0361	0.0316	—
PS, time-constant, true exp. # of shifts	1.1 [1–2]	0	0.0363	0.0317	—
PS, time-varying	1.1 [1–2]	0	0.1349 / 0.0087	0.0316 / 0.0315	− (10)
PS, time-varying, true exp. # of shifts	1.2 [1–2]	0.1 [0–1]	0.1369 / 0.0103	0.0350 / 0.0323	− (7), 0 (3)
T70, time-constant	1	0	0.0375	0.0324	—
T70, time-constant, true exp. # of shifts	1	0	0.0376	0.0323	—
T70, time-varying	1.1 [1–2]	0	0.1359 / 0.0109	0.0324 / 0.0326	− (9), 0 (1)
T70, time-varying, true exp. # of shifts	1.1 [1–2]	0	0.1380 / 0.0194	0.0329 / 0.0356	− (8), 0 (2)
Decreasing Preservation					
Time-constant	3.0 [1–13]	0	0.1085	0.1089	—
Time-constant, true exp. # of shifts	2.6 [1–8]	0	0.1080	0.1083	—
Time-varying	2.4 [1–9]	0	0.1260 / 0.1014	0.1088 / 0.1088	− (1), 0 (9)
Time-varying, true exp. # of shifts	4.6 [1–25]	0.1 [0–1]	0.1277 / 0.1001	0.1070 / 0.1082	0 (10)
PS, time-constant	1	0	0.0357	0.0294	—
PS, time-constant, true exp. # of shifts	1	0	0.0359	0.0294	—
PS, time-varying	1	0	0.0637 / 0.0256	0.0294 / 0.0294	− (1), 0 (9)
PS, time-varying, true exp. # of shifts	1	0	0.0644 / 0.0276	0.0303 / 0.0300	0 (10)
T70, time-constant	1.1 [1–2]	0	0.0563	0.0479	—
T70, time-constant, true exp. # of shifts	1.1 [1–2]	0	0.0566	0.0480	—
T70, time-varying	1	0	0.0912 / 0.0447	0.0478 / 0.0479	− (2), 0 (8)
T70, time-varying, true exp. # of shifts	1	0	0.0914 / 0.0454	0.0484 / 0.0475	− (2), 0 (8)

Table C.9: Summary of the Fossil BAMM simulations, indicating the number of shift configurations included in the 95% credible set, the number of core shifts (marginal odds ratio >5) present in the maximum *a posteriori* (MAP) configuration, the mean rates of speciation and extinction, and the presence or absence of an overlap between the 95% credibility intervals about the net diversification rate at the root and at the youngest tips. For the credible shift set size and the number of shifts in the MAP configuration, the values are averaged across 10 replicates and given along with the range (in brackets), unless the number was identical for all replicates. The mean rate values are given separately for the root / youngest tips for the analyses with time-varying within-regime speciation, and averaged across the entire clade history for the time-constant analyses. In both cases, the values reported are averaged across 10 replicates. In time-varying analyses, the net diversification rates can either be significantly lower (−) or higher (+) at the youngest tips than at the root if their 95% credibility intervals do not overlap, or there can be no significant difference between the two (0) if an overlap exists. The numbers in parentheses indicate the number of replicates with a given root–tip relationship.

analyses employing the PS and T70 schemes inferred mean speciation rates (i.e., rates averaged over the entire duration of the tree) that were lower than either of the two generating rates (Figure C.25). In time-varying analyses, relative precision (determined as the width of the 95% CI relative to the mean) was lower on average under the PS (root: 1.074, youngest tips: 1.742) and T70 (root: 1.016, youngest tips: 2.163) schemes than under the B scheme (root: 0.607, youngest tips: 0.950). However, even the relatively wide 95% CIs of the PS

and T70 analyses only contained the true value in 26 out of 40 cases for the root rate and 4 out of 40 cases for the rate at the youngest tips.

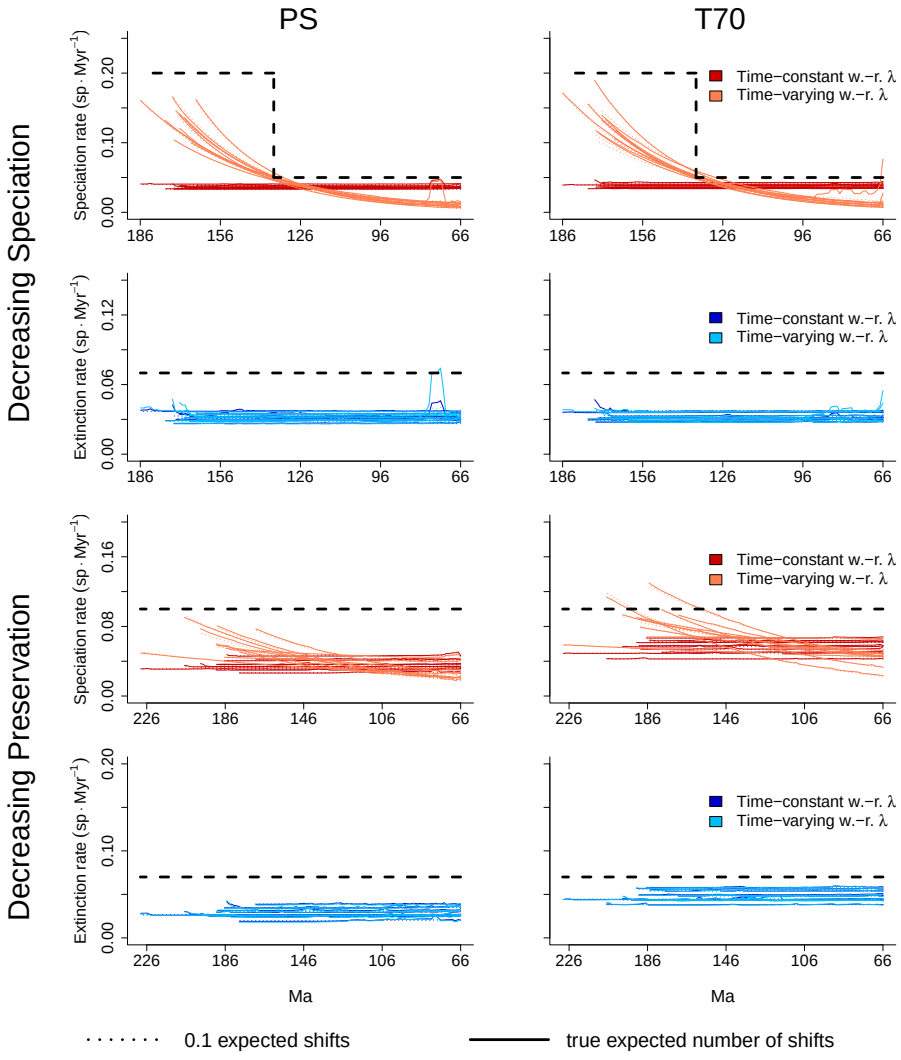

Figure C.25: Speciation and extinction rate-through-time plots from the Fossil BAMM analyses of simulated data. Each curve represents the posterior mean from one analysis; black dashed lines represent the true values. The proportionally subsampled (PS) analyses as well as those subsampled to 70 tips (T70) yield model-averaged rates that are both inaccurate and biased.

The extinction rate was estimated less accurately under the PS and T70 schemes than under the B scheme (Table C.10), and the estimates were biased in the opposite direction (Figure C.25). Wider credibility intervals and lower accuracy (Table C.10) translated to

low coverage values comparable to the B-scheme results: the true value fell within the 95% CI in 3 (root) or 4 (youngest tips) cases out of 40 in the time-varying analyses, and in 4 cases out of 40 at least once in the history of the clade in the time-constant analyses. In contrast, the preservation rates inferred by the artificially subsampled analyses were less biased ([0.554, 0.627] occ·sp^{-1}·Myr^{-1}), and their decreased relative precision (mean: 0.120) somewhat improved the coverage (95% CIs including the true value in 49 out of 80 cases).

The absence of speciation rate shifts in the Decreasing Preservation scenario was correctly inferred by subsampled BAMM analyses. None of the PS and T70 analyses contained a shift in their MAP configuration, and the 95% credible set only included shift-containing configurations in 2 out of 80 analyses (Table C.9), compared to 17 out of 40 for the B sampling scheme. Since the B scheme also most often led to the inclusion of shift-containing configurations in the 95% credible set under Decreasing Speciation, these results support the identification of the shifts estimated for the latter scenario as false positives. In analyses with time-varying within-regime speciation, there was an overlap between the 95% CIs about the net diversification rate at the root and at the youngest tips in 35 out of 40 cases. When the overlap failed to occur, the trend was always for the rate to decay from the root toward the tips (Table C.9).

The accuracy of marginal speciation rate estimates progressively decreased under the T70 and PS schemes (Table C.10; Figure C.25). Similar to the B-scheme results, the 95% CIs tended to be wider at the root (PS: 1.610, T70: 1.217) than at the youngest tips (PS: 1.304, T70: 0.841) in the time-varying analyses. The greater width corresponded to higher coverage in the artificially subsampled analyses (PS: true value included in 15 and 0 out of 20 cases at the root and at the youngest tips, respectively; T70: 20 and 0 out of 20 cases), while the B scheme exhibited no such relationship (17 and 19 out of 20 cases). In the time-constant analyses, the wider 95% CIs yielded by the PS and T70 schemes did not compensate for the lower accuracy (Table C.10) in terms of coverage. As a result, the true value fell within the

CI at least at one point in the history of the clade in just 1 out of 20 cases for the PS and T70 analyses.

	λ rates		μ rates	
Decreasing Speciation	Relative error	Precision	Relative error	Precision
Time-constant	0.609 [0.586, 0.637]	0.204 [0.184, 0.229]	0.157 [0.125, 0.195]	0.202 [0.183, 0.242]
Time-constant, true exp. # of shifts	0.608 [0.588, 0.640]	0.213 [0.182, 0.256]	0.155 [0.122, 0.191]	0.218 [0.188, 0.294]
Time-varying	0.340 [0.278, 0.431]	0.335 [0.293, 0.384]	0.156 [0.125, 0.193]	0.213 [0.185, 0.254]
Time-varying, true exp. # of shifts	0.329 [0.272, 0.423]	0.383 [0.326, 0.507]	0.154 [0.119, 0.198]	0.261 [0.212, 0.360]
PS, time-constant	0.692 [0.588, 0.779]	0.521 [0.498, 0.603]	0.757 [0.621, 0.904]	0.568 [0.520, 0.647]
PS, time-constant, true exp. # of shifts	0.688 [0.594, 0.770]	0.566 [0.501, 0.733]	0.755 [0.611, 0.893]	0.630 [0.550, 0.849]
PS, time-varying	0.814 [0.729, 0.870]	0.841 [0.759, 1.020]	0.759 [0.612, 0.913]	0.580 [0.522, 0.642]
PS, time-varying, true exp. # of shifts	0.798 [0.724, 0.845]	0.952 [0.800, 1.225]	0.755 [0.621, 0.905]	0.643 [0.560, 0.766]
T70, time-constant	0.663 [0.552, 0.762]	0.469 [0.428, 0.500]	0.739 [0.607, 0.856]	0.513 [0.485, 0.547]
T70, time-constant, true exp. # of shifts	0.661 [0.541, 0.763]	0.483 [0.449, 0.551]	0.741 [0.609, 0.861]	0.546 [0.508, 0.603]
T70, time-varying	0.761 [0.659, 0.881]	0.776 [0.735, 0.910]	0.741 [0.615, 0.861]	0.513 [0.486, 0.578]
T70, time-varying, true exp. # of shifts	0.744 [0.650, 0.877]	0.846 [0.755, 1.033]	0.739 [0.612, 0.864]	0.574 [0.507, 0.709]
Decreasing Preservation				
Time-constant	0.084 [0.005, 0.214]	0.274 [0.228, 0.314]	0.431 [0.342, 0.562]	0.276 [0.211, 0.313]
Time-constant, true exp. # of shifts	0.080 [0.006, 0.207]	0.308 [0.225, 0.373]	0.425 [0.345, 0.551]	0.306 [0.218, 0.389]
Time-varying	0.135 [0.045, 0.262]	0.455 [0.377, 0.503]	0.431 [0.343, 0.555]	0.280 [0.209, 0.333]
Time-varying, true exp. # of shifts	0.138 [0.053, 0.252]	0.463 [0.375, 0.520]	0.422 [0.340, 0.562]	0.316 [0.234, 0.393]
PS, time-constant	0.952 [0.731, 1.146]	0.723 [0.560, 0.825]	0.825 [0.572, 1.104]	0.813 [0.659, 0.927]
PS, time-constant, true exp. # of shifts	0.948 [0.732, 1.143]	0.739 [0.592, 0.844]	0.827 [0.567, 1.107]	0.915 [0.750, 1.051]
PS, time-varying	0.880 [0.709, 1.062]	0.992 [0.808, 1.133]	0.827 [0.560, 1.102]	0.809 [0.668, 0.893]
PS, time-varying, true exp. # of shifts	0.867 [0.705, 1.062]	1.023 [0.892, 1.154]	0.837 [0.588, 1.127]	0.913 [0.786, 1.079]
T70, time-constant	0.565 [0.403, 0.804]	0.476 [0.450, 0.500]	0.380 [0.185, 0.581]	0.527 [0.478, 0.567]
T70, time-constant, true exp. # of shifts	0.559 [0.396, 0.802]	0.497 [0.464, 0.550]	0.379 [0.184, 0.586]	0.579 [0.525, 0.685]
T70, time-varying	0.480 [0.343, 0.651]	0.699 [0.634, 0.754]	0.383 [0.184, 0.597]	0.541 [0.493, 0.579]
T70, time-varying, true exp. # of shifts	0.475 [0.344, 0.643]	0.715 [0.682, 0.745]	0.384 [0.189, 0.586]	0.598 [0.528, 0.697]

Table C.10: Comparison of accuracy and precision of the model-averaged rates estimated by Fossil BAMM from simulated data. Relative error = $|r_t - r_e|$ / $[(r_t + r_e)/2]$, where r_t is the true rate and r_e is the posterior mean; Precision = $(r_{0.975} - r_{0.025})/r_e$, where $r_{0.975}$ is the upper bound of the 95% credibility interval and $r_{0.025}$ is its lower bound. Both quantities are averaged across the entire history of the clade, with the resulting values further averaged across 10 replicates and given along with the across-replicate range (in brackets).

The positive relationship between accuracy and tree size observed for speciation did not hold for the extinction rate, where the estimates from 70-tip trees outperformed those obtained from the larger B-scheme trees while being less precise than the latter (Table C.10). The patterns of bias were identical to those observed for speciation, with the B scheme overestimating the true extinction rate and the artificially subsampled analyses consistently underestimating it (Figure C.25). Consistent with the results obtained under the B scheme, the relative precision of the extinction rates inferred by time-varying analyses was similar

between the root (PS: 0.999, T70: 0.630) and the youngest tips (PS: 0.882, T70: 0.564). The same is true of coverage, in which the T70 scheme (true value included in the 95% CI in 6 and 5 out of 20 cases at the root and at the youngest tips, respectively) again outperformed the PS (1 and 1 out of 20 cases) and B (1 and 0 out of 20 cases) schemes. Similarly, in the time-constant analyses, the 95% CIs included the true value at least once in 10 out of 20 cases under the T70 scheme, but in only 5 and 2 out of 20 cases under the PS and B

Decreasing Speciation												
	Including sampled ancestors						Excluding sampled ancestors					
	bdMCMC			rjMCMC			bdMCMC			rjMCMC		
λ rates	HPP (7)	TPP-136 (5)	TPP-by10 (3)	HPP (10)	TPP-136 (6)	TPP-by10 (4)	HPP (10)	TPP-136 (7)	TPP-by10 (3)	HPP (10)	TPP-136 (7)	TPP-by10 (3)
1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	**7**	**5**	**2**	1 (1)	1 (2)	0	**9** (1)	**7**	**2** (1)
3	**4** (1)	**3** (2)	**2** (1)	2 (8)	1 (5)	1 (3)	**6** (3)	**3** (4)	1 (2)	1 (9)	(7)	1 (2)
4	3 (4)	2 (3)	1 (2)	1 (2)	(1)	1 (1)	2 (6)	3 (1)	**2** (1)	0	0	0
5	(2)	0	0	0	0	0	1	0	0	0	0	0
μ rates												
1	0	0	0	0	0	0	0	0	0	0	0	0
2	(1)	0	(1)	0	0	0	0	0	0	2 (2)	1 (2)	1
3	**5**	2 (2)	**2**	**10**	**6**	**4**	3 (5)	2 (4)	(1)	**8** (2)	**6** (1)	**2** (1)
4	(6)	**2** (3)	(1)	(10)	(6)	(4)	**6** (1)	**3** (1)	**2** (1)	(6)	(4)	(2)
5	1	0	(1)	0	0	0	1 (3)	1 (1)	1	0	0	0
6	0	1	0	0	0	0	0	0	0	0	0	0
7	1	0	1	0	0	0	(1)	0	(1)	0	0	0
8	0	0	0	0	0	0	0	1	0	0	0	0

Decreasing Preservation						
	Including sampled ancestors				Excluding sampled ancestors	
	bdMCMC		rjMCMC		bdMCMC	rjMCMC
λ rates	TPP-136 (10)	TPP-by10 (1)	TPP-136 (10)	TPP-by10 (1)	TPP-136 (10)	TPP-136 (10)
1	**10**	**1**	**6** (2)	**1**	**10**	**6** (3)
2	(10)	(1)	4 (6)	(1)	(10)	4 (6)
3	0	0	(2)	0	0	(1)
μ rates						
1	0	0	0	0	0	0
2	3 (2)	0	**10**	**1**	(4)	**10**
3	**7** (3)	**1**	(10)	(1)	**7** (2)	(10)
4	(5)	(1)	0	0	2 (4)	0
5	0	0	0	0	1	0

Table C.11: Overview of the number of distinct rates included in the models most frequently sampled by the PyRate analyses of simulated data. The values indicate the number of replicates for which the most frequently sampled (highest-probability) model contained a given number of rates. The number of replicates for which the given number of rates occurred in the second most-probable model is given in parentheses; note that for some replicates, PyRate only sampled a single model. The number of rates that was most often included in the highest probability model is given in bold for each analysis; the true number of rates in the generating model is given in bold on the left. The total number of replicates per analysis is given in parentheses in the column headers. TPP-136 = time-variable Poisson process with a single preservation rate shift at 136 Ma; TPP-by10 = time-variable Poisson process in which the preservation rate shifts every 10 Myr.

schemes, respectively. The preservation rate estimates did not vary greatly depending on the sampling scheme, and were intermediate between the two true rates (range of means: [0.414, 0.589] occ·sp^{-1}·Myr^{-1}).

Supplementary PyRate results

In contrast to the empirical results (Table C.6), the PyRate analyses of synthetic data tend to infer more shifts in both speciation and extinction when using the bdMCMC model-averaging algorithm (Table C.11). There were no appreciable differences in accuracy between the bdMCMC and rjMCMC rate estimates, but the latter tended to have higher precision (Table C.12). The relative error of the PyRate estimates of speciation and extinction was broadly comparable to the subsampled BAMM analyses under both of the scenarios tested (Tables C.10, C.12).

	λ rates		μ rates	
Decreasing Speciation	Relative error	Precision	Relative error	Precision
bdMCMC, HPP	0.668 [0.640, 0.699]	0.686 [0.614, 0.842]	0.676 [0.601, 0.732]	0.750 [0.565, 0.954]
bdMCMC, TPP-136	0.686 [0.662, 0.702]	0.687 [0.620, 0.733]	0.688 [0.632, 0.741]	0.785 [0.678, 0.860]
bdMCMC, TPP-by10	0.661 [0.641, 0.681]	0.727 [0.613, 0.877]	0.678 [0.614, 0.719]	0.774 [0.569, 0.992]
rjMCMC, HPP	0.653 [0.531, 0.708]	0.347 [0.245, 0.726]	0.695 [0.615, 0.772]	0.376 [0.265, 0.725]
rjMCMC, TPP-136	0.681 [0.610, 0.715]	0.346 [0.271, 0.468]	0.700 [0.646, 0.737]	0.441 [0.284, 0.715]
rjMCMC, TPP-by10	0.644 [0.543, 0.704]	0.383 [0.230, 0.725]	0.711 [0.624, 0.788]	0.336 [0.290, 0.446]
bdMCMC, no samp. ancestors, HPP	0.664 [0.615, 0.724]	0.735 [0.611, 1.135]	0.568 [0.451, 0.776]	0.880 [0.735, 1.108]
bdMCMC, no samp. ancestors, TPP-136	0.669 [0.637, 0.724]	0.691 [0.620, 0.796]	0.565 [0.464, 0.708]	0.879 [0.747, 1.080]
bdMCMC, no samp. ancestors, TPP-by10	0.671 [0.631, 0.713]	0.825 [0.656, 1.100]	0.596 [0.478, 0.775]	0.891 [0.773, 1.048]
rjMCMC, no samp. ancestors, HPP	0.662 [0.620, 0.751]	0.450 [0.321, 0.729]	0.532 [0.398, 0.687]	0.759 [0.346, 1.194]
rjMCMC, no samp. ancestors, TPP-136	0.679 [0.643, 0.755]	0.427 [0.370, 0.626]	0.541 [0.406, 0.668]	0.756 [0.336, 1.214]
rjMCMC, no samp. ancestors, TPP-by10	0.648 [0.628, 0.666]	0.577 [0.327, 0.796]	0.556 [0.411, 0.716]	0.841 [0.456, 1.148]
Decreasing Preservation				
bdMCMC, TPP-136	0.622 [0.505, 0.711]	0.248 [0.190, 0.338]	0.845 [0.780, 0.900]	0.548 [0.373, 0.761]
bdMCMC, TPP-by10	0.688	0.235	0.884	0.601
rjMCMC, TPP-136	0.618 [0.504, 0.708]	0.323 [0.202, 0.439]	0.849 [0.753, 0.932]	0.355 [0.195, 0.649]
rjMCMC, TPP-by10	0.684	0.239	0.902	0.293
bdMCMC, no samp. ancestors, TPP-136	0.613 [0.509, 0.785]	0.339 [0.251, 0.417]	0.819 [0.662, 1.043]	0.843 [0.615, 1.013]
rjMCMC, no samp. ancestors, TPP-136	0.610 [0.491, 0.782]	0.671 [0.265, 2.482]	0.771 [0.646, 0.948]	0.656 [0.291, 2.022]

Table C.12: Comparison of accuracy and precision of the model-averaged rates estimated by PyRate from simulated data. Both quantities are averaged across the entire history of the clade, with the resulting values further averaged across all replicates that reached convergence ($\leq$10 parameters with effective sample sizes of <200) and given along with the across-replicate range (in brackets), unless only one replicate was run per analysis.

Rate identifiability.

The recent demonstration that a given molecular time tree can be generated with equal likelihood by an infinite number of time-varying birth-death processes (Louca and Pennell, 2020) provokes concern about inferring macroevolutionary dynamics from phylogenies of extant taxa alone. It has been suggested that fossil data, which carry information about past extinction rates, could alleviate this problem and shrink the space of plausible models (Louca and Pennell, 2020). This cautious optimism is tempered by the recently reported interactions between the rates of fossil preservation (ψ), speciation (λ), and extinction (μ) (Condamine et al., 2016; Smiley, 2018; Foote et al., 2019), raising the possibility that the parameters λ, μ, ψ are not identifiable. Our results lend some support to this possibility. We found no evidence of spurious shifts in speciation or extinction induced by shifts in preservation (*contra* Smiley, 2018), whereas the converse case of a spurious shift in preservation induced by a shift in speciation is impossible in Fossil BAMM (where shifts in preservation are currently disallowed altogether) and rare in PyRate (<20% of analyses with non-overlapping pre-shift and post-shift preservation rate CIs; see Figure 3.6 in the main text).

Despite their success in disentangling macroevolution from preservation, PyRate and Fossil BAMM yield highly correlated estimates of speciation and extinction rates. While this correlation may be real in at least some datasets (Crampton et al., 2020), our simulations infer it in spite of its absence from the generating data. This is best exemplified by the spurious extinction rate shift seen in the PyRate RTT plots under Decreasing Speciation (Figure 3.6 in the main text), which closely matches the speciation rate shift present in the generating model. The simultaneous pulses of speciation and extinction inferred by the rjMCMC PyRate analyses of the empirical data (Figure 3.4 in the main text) exhibit the same correspondence in time and magnitude. The phenomenon of speciation and extinction tracking each other is less conspicuous with Fossil BAMM, where the zero-shift model dominated marginal rates in both empirical and simulated analyses. However, it is still apparent in the spurious correlated

upticks in speciation and extinction present in the last 30 Myr (Figure 3.5 in the main text) and from the "phylorate" plots (time trees with branch segments colored by rate). In all the 8 simulations where the MAP configuration contained at least one shift, phylorate plots show the shifts to comprise simultaneous speed-ups in speciation and extinction (Figures S29–S31, S35, and S43), yielding net diversification rates that differ little from the background regime.

Correlations between parameter posteriors despite their independence under the prior are suggestive but not positively diagnostic of non-identifiability (Rannala, 2002; Ponciano et al., 2012). In particular, if parameters are i.i.d. under the prior (as was the case for the vectors of speciation and extinction rates in all our PyRate analyses), the correlation of their posterior estimates may be due to non-informative data rather than structural features of the model (Rannala, 2002). These two scenarios are referred to as non-estimability vs. non-identifiability (Ponciano et al., 2012) or practical vs. structural non-identifiability (Li and Vu, 2013). The varying strength of the correlation between λ and μ in our simulations, as well as the lack of such correlation in some previous analyses (Silvestro et al., 2014b; Condamine et al., 2016; Pires et al., 2018), suggests that the non-identifiability of the two rates is practical rather than structural, and should be attributed to the low information content of the data rather than the models with which the data are analyzed. The model used still plays an important role, as shown by the tendency of PyRate (but not Fossil BAMM) to estimate the difference $(\lambda - \mu)$ more accurately than either rate taken individually (Figures C.26, C.27). This may reflect the fact that the scenarios explored here (with tree-wide rather than clade-specific rate shifts) conformed to the model assumptions of PyRate better than to those of BAMM. However, the spurious shifts in μ inferred by PyRate still suggest that even with the "correct" model, only quantities that confound both rates may be estimable from low-power datasets, consistent with previous demonstrations that survivorship curves inferred from paleontological data are more sensitive to $(\lambda - \mu)$ than to the absolute magnitudes of λ and μ (Foote, 1988). Given the interest in teasing apart the relative contributions of

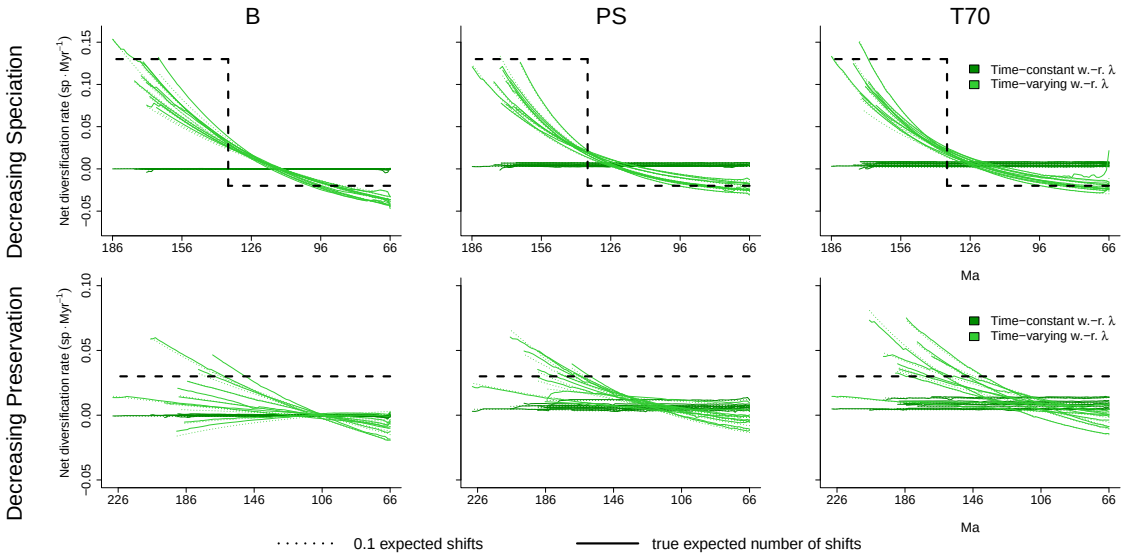

Figure C.26: Net diversification rate-through-time (RTT) plots from the Fossil BAMM analyses of simulated data. Each curve represents the posterior mean from one analysis; black dashed lines represent the true values. The accuracy of net diversification rate estimates is comparable to that of speciation rate estimates under Decreasing Speciation, but lower under Decreasing Preservation due to systematically overestimated rates of extinction (cf. Figure 3.5 in the main text).

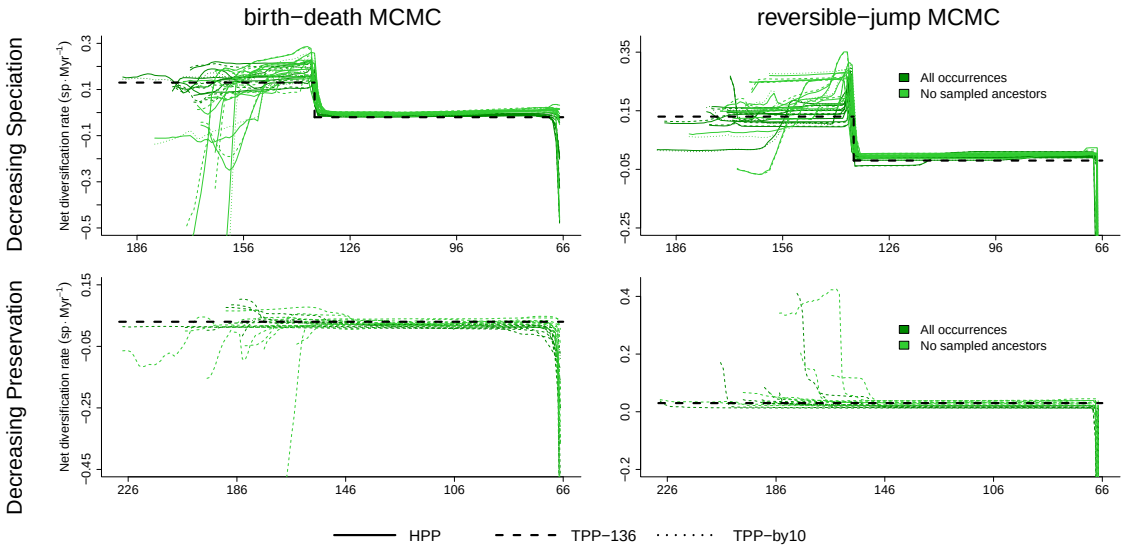

Figure C.27: Net diversification RTT plots from the PyRate analyses of simulated data. Each curve represents the posterior mean from one analysis; black dashed lines represent the true values. The inferred net diversification rates track the true values more closely than either speciation or extinction rate estimates (cf. Figure 3.6 in the main text).

speciation and extinction to macroevolutionary dynamics (Morlon et al., 2011; Quental and Marshall, 2013; Burin et al., 2019), the range of hypotheses that can be tested using current rate estimation methods may be considerably limited for many datasets.

C.5.6 List of BAMM Phylorate Plots

The supplementary figures referred to below are provided in separate PDF files.

Figure S28: Phylorate plots showing the instantaneous rates of speciation, extinction, and net diversification for the maximum *a posteriori* (MAP) rate shift configurations resulting from the Fossil BAMM analyses of 10 replicates under the Decreasing Speciation scenario, the basic (B) sampling scheme, time-constant within-regime speciation, and 0.1 expected shifts. The values shown in the legend are in the units of species per million years ($\text{sp}\cdot\text{Myr}^{-1}$). The green line represents the location of the tree-wide drop in the speciation rate present in the generating model.

Figure S29: Phylorate plots for the Fossil BAMM analyses performed under Decreasing Speciation, the B scheme, time-constant within-regime speciation, and the true expected number of shifts. The filled red circles denote the location of the rate shifts present in the MAP configuration.

Figure S30: Phylorate plots for the Fossil BAMM analyses performed under Decreasing Speciation, the B scheme, time-varying within-regime speciation, and 0.1 expected shifts. The filled red circles denote the location of the rate shifts present in the MAP configuration.

Figure S31: Phylorate plots for the Fossil BAMM analyses performed under Decreasing Speciation, the B scheme, time-varying within-regime speciation, and the true expected number

of shifts. The filled red circles denote the location of the rate shifts present in the MAP configuration.

Figure S32: Phylorate plots for the Fossil BAMM analyses performed under Decreasing Speciation, the proportionally subsampled (PS) sampling scheme, time-constant within-regime speciation, and 0.1 expected shifts.

Figure S33: Phylorate plots for the Fossil BAMM analyses performed under Decreasing Speciation, the PS scheme, time-constant within-regime speciation, and the true expected number of shifts.

Figure S34: Phylorate plots for the Fossil BAMM analyses performed under Decreasing Speciation, the PS scheme, time-varying within-regime speciation, and 0.1 expected shifts.

Figure S35: Phylorate plots for the Fossil BAMM analyses performed under Decreasing Speciation, the PS scheme, time-varying within-regime speciation, and the true expected number of shifts. The filled red circles denote the location of the rate shifts present in the MAP configuration.

Figure S36: Phylorate plots for the Fossil BAMM analyses performed under Decreasing Speciation, the 70-tip (T70) sampling scheme, time-constant within-regime speciation, and 0.1 expected shifts.

Figure S37: Phylorate plots for the Fossil BAMM analyses performed under Decreasing Speciation, the T70 scheme, time-constant within-regime speciation, and the true expected number of shifts.

Figure S38: Phylorate plots for the Fossil BAMM analyses performed under Decreasing Speciation, the T70 scheme, time-varying within-regime speciation, and 0.1 expected shifts.

Figure S39: Phylorate plots for the Fossil BAMM analyses performed under Decreasing Speciation, the T70 scheme, time-varying within-regime speciation, and the true expected number of shifts.

Figure S40: Phylorate plots for the Fossil BAMM analyses performed under the Decreasing Preservation scenario, the B scheme, time-constant within-regime speciation, and 0.1 expected shifts.

Figure S41: Phylorate plots for the Fossil BAMM analyses performed under Decreasing Preservation, the B scheme, time-constant within-regime speciation, and the true expected number of shifts.

Figure S42: Phylorate plots for the Fossil BAMM analyses performed under Decreasing Preservation, the B scheme, time-varying within-regime speciation, and 0.1 expected shifts.

Figure S43: Phylorate plots for the Fossil BAMM analyses performed under Decreasing Preservation, the B scheme, time-varying within-regime speciation, and the true expected number of shifts. The filled red circles denote the location of the rate shifts present in the MAP configuration.

Figure S44: Phylorate plots for the Fossil BAMM analyses performed under Decreasing Preservation, the PS scheme, time-constant within-regime speciation, and 0.1 expected shifts.

Figure S45: Phylorate plots for the Fossil BAMM analyses performed under Decreasing

Preservation, the PS scheme, time-constant within-regime speciation, and the true expected number of shifts.

Figure S46: Phylorate plots for the Fossil BAMM analyses performed under Decreasing Preservation, the PS scheme, time-varying within-regime speciation, and 0.1 expected shifts.

Figure S47: Phylorate plots for the Fossil BAMM analyses performed under Decreasing Preservation, the PS scheme, time-varying within-regime speciation, and the true expected number of shifts.

Figure S48: Phylorate plots for the Fossil BAMM analyses performed under Decreasing Preservation, the T70 scheme, time-constant within-regime speciation, and 0.1 expected shifts.

Figure S49: Phylorate plots for the Fossil BAMM analyses performed under Decreasing Preservation, the T70 scheme, time-constant within-regime speciation, and the true expected number of shifts.

Figure S50: Phylorate plots for the Fossil BAMM analyses performed under Decreasing Preservation, the T70 scheme, time-varying within-regime speciation, and 0.1 expected shifts.

Figure S51: Phylorate plots for the Fossil BAMM analyses performed under Decreasing Preservation, the T70 scheme, time-varying within-regime speciation, and the true expected number of shifts.

REFERENCES

Aberer, A. J., Stamatakis, A., and Ronquist, F. (2016). An efficient independence sampler for updating branches in Bayesian Markov chain Monte Carlo sampling of phylogenetic trees. *Systematic Biology*, 65(1):161–176.

Agnarsson, I., Kuntner, M., and May-Collado, L. J. (2010). Dogs, cats, and kin: a molecular species-level phylogeny of Carnivora. *Molecular Phylogenetics and Evolution*, 54(3):726–745.

Agnarsson, I. and May-Collado, L. J. (2008). The phylogeny of Cetartiodactyla: The importance of dense taxon sampling, missing data, and the remarkable promise of cytochrome *b* to provide reliable species-level phylogenies. *Molecular Phylogenetics and Evolution*, 48(3):964–985.

Agnolín, F. L. and Rozadilla, S. (2018). Phylogenetic reassessment of *Pisanosaurus mertii* Casamiquela, 1967, a basal dinosauriform from the Late Triassic of Argentina. *Journal of Systematic Palaeontology*, 16(10):853–879.

Akanni, W. A., Creevey, C. J., Wilkinson, M., and Pisani, D. (2014). L.U.St: a tool for approximated maximum likelihood supertree reconstruction. *BMC Bioinformatics*, 15(1):183.

Akanni, W. A., Wilkinson, M., Creevey, C. J., Foster, P. G., and Pisani, D. (2015). Implementing and testing Bayesian and maximum-likelihood supertree methods in phylogenetics. *Royal Society Open Science*, 2(8):140436.

Alfaro, M. E. and Holder, M. T. (2006). The posterior and the prior in Bayesian phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, 37(1):19–42.

Alfaro, M. E., Santini, F., Brock, C., Alamillo, H., Dornburg, A., Rabosky, D. L., Carnevale, G., and Harmon, L. J. (2009). Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proceedings of the National Academy of Sciences*, 106(32):13410–13414.

Ali, R. H., Bark, M., Miró, J., Muhammad, S. A., Sjöstrand, J., Zubair, S. M., Abbas, R. M., and Arvestad, L. (2017). VMCMC: a graphical and statistical analysis tool for Markov chain Monte Carlo traces. *BMC Bioinformatics*, 18(1):97.

Allman, E. S., Long, C., and Rhodes, J. A. (2019). Species tree inference from genomic sequences using the log-det distance. *SIAM Journal on Applied Algebra and Geometry*, 3(1):107–127.

Alroy, J. (2008). Dynamics of origination and extinction in the marine fossil record. *Proceedings of the National Academy of Sciences*, 105(supplement_1):11536–11542.

Alroy, J. (2010). Geographical, environmental and intrinsic biotic controls on Phanerozoic marine diversification. *Palaeontology*, 53(6):1211–1235.

Alroy, J. (2014). Accurate and precise estimates of origination and extinction rates. *Paleobiology*, 40(3):374–397.

Altekar, G., Dwarkadas, S., Huelsenbeck, J. P., and Ronquist, F. (2004). Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics*, 20(3):407–415.

Álvarez-Carretero, S., Tamuri, A. U., Battini, M., Nascimento, F. F., Carlisle, E., Asher, R. J., Yang, Z., Donoghue, P. C. J., and dos Reis, M. (2022). A species-level timeline of mammal evolution integrating phylogenomic data. *Nature*, 602(7896):263–267.

Andersen, M. J., Shult, H. T., Cibois, A., Thibault, J.-C., Filardi, C. E., and Moyle, R. G. (2015). Rapid diversification and secondary sympatry in Australo-Pacific kingfishers (Aves: Alcedinidae: *Todiramphus*). *Royal Society Open Science*, 2(2):140375.

Archibald, J. D. (1996). *Dinosaur Extinction and the End of an Era: What the Fossils Say*. Columbia University Press, New York, NY.

Archibald, J. D. (2011). *Extinction and Radiation: How the Fall of the Dinosaurs Led to the Rise of the Mammals*. The Johns Hopkins University Press, Baltimore, MD.

Archibald, J. D. and Fastovsky, D. (2004). Dinosaur extinction. In Weishampel, D. B., Dodson, P., and Osmólska, H., editors, *The Dinosauria, 2nd edition*, pages 672–684. University of California Press, Berkeley, CA.

Ayres, D. L., Cummings, M. P., Baele, G., Darling, A. E., Lewis, P. O., Swofford, D. L., Huelsenbeck, J. P., Lemey, P., Rambaut, A., and Suchard, M. A. (2019). BEAGLE 3: Improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics. *Systematic Biology*, 68(6):1052–1061.

Ayres, D. L., Darling, A., Zwickl, D. J., Beerli, P., Holder, M. T., Lewis, P. O., Huelsenbeck, J. P., Ronquist, F., Swofford, D. L., Cummings, M. P., Rambaut, A., and Suchard, M. A. (2012). BEAGLE: An application programming interface and high-performance computing library for statistical phylogenetics. *Systematic Biology*, 61(1):170–173.

Baca, S. M., Toussaint, E. F. A., Miller, K. B., and Short, A. E. Z. (2017). Molecular phylogeny of the aquatic beetle family Noteridae (Coleoptera: Adephaga) with an emphasis on data partitioning strategies. *Molecular Phylogenetics and Evolution*, 107:282–292.

Bagatharia, S. B., Joshi, M. N., Pandya, R. V., Pandit, A. S., Patel, R. P., Desai, S. M., Sharma, A., Panchal, O., Jasmani, F. P., and Saxena, A. K. (2013). Complete mitogenome of asiatic lion resolves phylogenetic status within *Panthera*. *BMC Genomics*, 14:572.

Bakker, R. T. (1986). *The Dinosaur Heresies: New Theories Unlocking the Mystery of the Dinosaurs and Their Extinction*. William Morrow & Co., New York.

Bakker, R. T. and Galton, P. M. (1974). Dinosaur monophyly and a new class of vertebrates. *Nature*, 248(5444):168–172.

Bapst, D. W. (2013). A stochastic rate-calibrated method for time-scaling phylogenies of fossil taxa. *Methods in Ecology and Evolution*, 4(8):724–733.

Bapst, D. W. (2014). Preparing paleontological datasets for phylogenetic comparative methods. In Garamszegi, L. Z., editor, *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice*, pages 515–544. Springer, Heidelberg.

Bapst, D. W., Wright, A. M., Matzke, N. J., and Lloyd, G. T. (2016). Topology, divergence dates, and macroevolutionary inferences vary between different tip-dating approaches applied to fossil theropods (Dinosauria). *Biology Letters*, 12(7):20160237.

Barido-Sottani, J., Aguirre-Fernández, G., Hopkins, M. J., Stadler, T., and Warnock, R. C. M. (2019a). Ignoring stratigraphic age uncertainty leads to erroneous estimates of species divergence times under the fossilized birth–death process. *Proceedings of the Royal Society B: Biological Sciences*, 286(1902):20190685.

Barido-Sottani, J., Pett, W., O'Reilly, J. E., and Warnock, R. C. M. (2019b). FossilSim: An R package for simulating fossil occurrence data under mechanistic models of preservation and recovery. *Methods in Ecology and Evolution*, 10(6):835–840.

Barido-Sottani, J., Vaughan, T. G., and Stadler, T. (2020). A multitype birth–death model for Bayesian inference of lineage-specific birth and death rates. *Systematic Biology*, 69(5):973–986.

Baron, M. G. (2019). *Pisanosaurus mertii* and the Triassic ornithischian crisis: could phylogeny offer a solution? *Historical Biology*, 31(8):967–981.

Baron, M. G. (2022). The effect of character and outgroup choice on the phylogenetic position of the Jurassic dinosaur *Chilesaurus diegosaurezi*. *Palaeoworld*, 33(1):142–151.

Baron, M. G. and Barrett, P. M. (2017). A dinosaur missing-link? *Chilesaurus* and the early evolution of ornithischian dinosaurs. *Biology Letters*, 13(8):20170220.

Baron, M. G. and Barrett, P. M. (2018). Support for the placement of *Chilesaurus* within Ornithischia: a reply to Müller *et al.* *Biology Letters*, 14(3):20180002.

Baron, M. G., Norman, D. B., and Barrett, P. M. (2017a). A new hypothesis of dinosaur relationships and early dinosaur evolution. *Nature*, 543(7646):501–506.

Baron, M. G., Norman, D. B., and Barrett, P. M. (2017b). Untangling the dinosaur family tree: Baron *et al.* reply. *Nature*, 551(7678):E4–E5.

Barrett, P. M., McGowan, A. J., and Page, V. (2009). Dinosaur diversity and the rock record. *Proceedings of the Royal Society B: Biological Sciences*, 276(1667):2667–2674.

Barrett, P. Z., Hopkins, S. S. B., and Price, S. A. (2021). How many sabertooths? Reevaluating the number of carnivoran sabretooth lineages with total-evidence Bayesian techniques and a novel origin of the Miocene Nimravidae. *Journal of Vertebrate Paleontology*, 41(1):e1923523.

Baum, B. R. and Ragan, M. A. (2004). The MRP method. In Bininda-Emonds, O. R. P., editor, *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pages 17–34. Kluwer Academic Publishers, Dordrecht, NL.

Beiko, R. G., Keith, J. M., Harlow, T. J., and Ragan, M. A. (2006). Searching for convergence in phylogenetic Markov chain Monte Carlo. *Systematic Biology*, 55(4):553–565.

Bell, P. R., Herne, M. C., Brougham, T., and Smith, E. T. (2018). Ornithopod diversity in the Griman Creek Formation (Cenomanian), New South Wales, Australia. *PeerJ*, 6:e6008.

Benson, R. B. J., Campione, N. E., Carrano, M. T., Mannion, P. D., Sullivan, C., Upchurch, P., and Evans, D. C. (2014). Rates of dinosaur body mass evolution indicate 170 million years of sustained ecological innovation on the avian stem lineage. *PLoS Biology*, 12(5):e1001853.

Benton, M. J. (2004). Origin and relationships of Dinosauria. In Weishampel, D. B., Dodson, P., and Osmólska, H., editors, *The Dinosauria, 2nd edition*, pages 7–19. University of California Press, Berkeley, CA.

Bergsten, J., Nilsson, A. N., and Ronquist, F. (2013). Bayesian tests of topology hypotheses with an example from diving beetles. *Systematic Biology*, 62(5):660–673.

Betancur-R., R., Arcila, D., Vari, R. P., Hughes, L. C., Oliveira, C., Sabaj, M. H., and Ortí, G. (2019). Phylogenomic incongruence, hypothesis testing, and taxonomic sampling: The monophyly of characiform fishes. *Evolution*, 73(2):329–345.

Billaud, O., Moen, D. S., Parsons, T. L., and Hélène (2020). Estimating diversity through time using molecular phylogenies: old and species-poor frog families are the remnants of a diverse past. *Systematic Biology*, 69(2):363–383.

Billera, L. J., Holmes, S. P., and Vogtmann, K. (2001). Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4):733–767.

Binet, M., Gascuel, O., Scornavacca, C., Douzery, E. J. P., and Pardi, F. (2016). Fast and accurate branch lengths estimation for phylogenomic trees. *BMC Bioinformatics*, 17(1):23.

Bininda-Emonds, O. R. P. (2014). An introduction to supertree construction (and partitioned phylogenetic analyses) with a view toward the distinction between gene trees and species trees. In Garamszegi, L. Z., editor, *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice*, chapter 3, pages 49–76. Springer-Verlag, Berlin, Germany.

Bininda-Emonds, O. R. P., Gittleman, J. L., and Purvis, A. (1999). Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). *Biological Reviews*, 74(2):143–175.

Bininda-Emonds, O. R. P., Gittleman, J. L., and Steel, M. A. (2002). The (super)tree of life: procedures, problems, and prospects. *Annual Review of Ecology and Systematics*, 33:265–289.

Bittencourt, J. S., Arcucci, A. B., Marsicano, C. A., and Langer, M. C. (2015). Osteology of the Middle Triassic archosaur *Lewisuchus admixtus* Romer (Chañares Formation, Argentina), its inclusivity, and relationships amongst early dinosauromorphs. *Journal of Systematic Palaeontology*, 13(3):189–219.

Bleich, J., Kapelner, A., George, E. I., and Jensen, S. T. (2014). Variable selection for BART: An application to gene regulation. *Annals of Applied Statistics*, 8(3):1750–1781.

Bogdanowicz, D. and Giaro, K. (2012). Matching split distance for unrooted binary phylogenetic trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(1):150–160.

Bon, C., Caudy, N., de Dieuleveult, M., Fosse, P., Philippe, M., Maksud, F., Beraud-Colomb, É., Bouzaid, E., Kefi, R., Laugier, C., Rousseau, B., Casane, D., van der Plicht, J., and Elalouf, J.-M. (2008). Deciphering the complete mitochondrial genome and phylogeny of the extinct cave bear in the Paleolithic painted cave of Chauvet. *Proceedings of the National Academy of Sciences*, 105(45):17447–17452.

Bonaparte, J. F. (1976). *Pisanosaurus mertii* Casamiquela and the origin of the Ornithischia. *J. Paleontol.*, 50(5):808–820.

Bonsor, J. A., Barrett, P. M., Raven, T. J., and Cooper, N. (2020). Dinosaur diversification rates were not in decline prior to the K-Pg boundary. *Royal Society Open Science*, 7(11):201195.

Bouckaert, R. R. (2010). DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics*, 26(10):1372–1373.

Bouckaert, R. R. (2014). Sampled ancestor trees in BEAST. `http://www.beast2.org/2014/09/02/sampled-ancestor-trees-in-beast.html`. Accessed December 28, 2019.

Bouckaert, R. R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., Matschiner, M., Mendes, F. K., Müller, N. F., Ogilvie, H. A., du Plessis, L., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., Suchard, M. A., Wu, C.-H., Xie, D., Zhang, C., Stadler, T., and Drummond, A. J. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Computational Biology*, 15(4):1–28.

Boyd, C. A. (2015). The systematic relationships and biogeographic history of ornithischian dinosaurs. *PeerJ*, 3:e1523.

Braun, E. L. and Kimball, R. T. (2001). Polytomies, the power of phylogenetic inference, and the stochastic nature of molecular evolution: A comment on Walsh et al. (1999). *Evolution*, 55(6):1261–1263.

Brayard, A., Escarguel, G., Bucher, H., Monnet, C., Brühwiler, T., Goudemand, N., Galfetti, T., and Guex, J. (2009). Good genes and good luck: ammonoid diversity and the end-Permian mass extinction. *Science*, 325(5944):1118–1121.

Brinkmeyer, M., Griebel, T., and Böcker, S. (2011). Polynomial supertree methods revisited. *Advances in Bioinformatics*, 2011(1):524182.

Brown, B. and Schlaikjer, E. M. (1943). A study of the troödont dinosaurs with the description of a new genus and four new species. *Bulletin of the American Museum of Natural History*, 82(5):115–150.

Brusatte, S. L., Butler, R. J., Barrett, P. M., Carrano, M. T., Evans, D. C., Lloyd, G. T., Mannion, P. D., Norell, M. A., Peppe, D. J., Upchurch, P., and Williamson, T. E. (2015). The extinction of the dinosaurs. *Biological Reviews*, 90(2):628–642.

Bryant, D. and Steel, M. (2009). Computing the distribution of a tree metric. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(3):420–426.

Buneman, P. (1974). A note on the metric properties of trees. *Journal of Combinatorial Theory, Series B*, 17(1):48–50.

Burgin, C. J., Colella, J. P., Kahn, P. L., and Upham, N. S. (2018). How many species of mammals are there? *Journal of Mammalogy*, 99(1):1–14.

Burin, G., Alencar, L. R. V., Chang, J., Alfaro, M. E., and Quental, T. B. (2019). How well can we estimate diversity dynamics for clades in diversity decline? *Systematic Biology*, 68(1):47–62.

Burleigh, J. G., Alphonse, K., Alverson, A. J., Bik, H. M., Blank, C., Cirranello, A. L., Cui, H., Daly, M., Dietterich, T. G., Gasparich, G., Irvine, J., Julius, M., Kaufman, S., Law, E., Liu, J., Moore, L., O'Leary, M. A., Passarotti, M., Ranade, S., Simmons, N. B., Stevenson, D. W., Thacker, R. W., Theriot, E. C., Todorovic, S., Velazco, P. M., Walls, R. L., Wolfe, J. M., and Yu, M. (2013). Next-generation phenomics for the Tree of Life. *PLoS Currents: Tree of Life*, 5. doi:10.1371/currents.tol.085c713acafc8711b2ff7010a4b03733.

Burleigh, J. G., Driskell, A. C., and Sanderson, M. J. (2006). Supertree bootstrapping methods for assessing phylogenetic variation among genes in genome-scale data sets. *Systematic Biology*, 55(3):426–440.

Butler, R. J., Benson, R. B. J., and Barrett, P. M. (2013). Pterosaur diversity: Untangling the influence of sampling biases, Lagerstätten, and genuine biodiversity signals. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 372:78–87.

Butler, R. J., Brusatte, S. L., Andres, B., and Benson, R. B. J. (2012). How do geological sampling biases affect studies of morphological evolution in deep time? A case study of pterosaur (Reptilia: Archosauria) disparity. *Evolution*, 66(1):147–162.

Butler, R. J., Upchurch, P., and Norman, D. B. (2008). The phylogeny of the ornithischian dinosaurs. *Journal of Systematic Palaeontology*, 6(1):1–40.

Cabreira, S. F., Kellner, A. W. A., Dias-da Silva, S., Roberto da Silva, L., Bronzati, M., Marsola, J. C. d. A., Müller, R. T., Bittencourt, J. d. S., Batista, B. J., Raugust, T., Carrilho, R., Brodt, A., and Langer, M. C. (2016). A unique Late Triassic dinosauromorph assemblage reveals dinosaur ancestral anatomy and diet. *Current Biology*, 26(22):3090–3095.

Caetano, D. S. and Harmon, L. J. (2018). Estimating correlated rates of trait evolution with uncertainty. *Systematic Biology*, 68(3):412–429.

Campitelli, E. (2022). *ggnewscale: Multiple fill and colour scales in 'ggplot2'*. R package version 0.4.8.

Carpenter, K. and Galton, P. M. (2018). A photo documentation of bipedal ornithischian dinosaurs from the Upper Jurassic Morrison Formation, USA. *Geology of the Intermountain West*, 5:167–207.

Carruthers, T., Sun, M., Baker, W. J., Smith, S. A., de Vos, J. M., and Eiserhardt, W. L. (2022). The implications of incongruence between gene tree and species tree topologies for divergence time estimation. *Systematic Biology*, 71(5):1124–1146.

Castiglione, S., Serio, C., Mondanaro, A., Melchionna, M., and Raia, P. (2022). Fast production of large, time-calibrated, informal supertrees with `tree.merger`. *Palaeontology*, 65(1):e12588.

Cau, A. (2017). Specimen-level phylogenetics in paleontology using the Fossilized Birth-Death model with sampled ancestors. *PeerJ*, 5:e3055.

Cau, A. (2024). A unified framework for predatory dinosaur macroevolution. *Bollettino della Società Paleontologica Italiana*, 63. 10.4435/BSPI.2024.08.

Černý, D., Madzia, D., and Slater, G. J. (2022). Empirical and methodological challenges to the model-based inference of diversification rates in extinct clades. *Systematic Biology*, 71(1):153–171.

Černý, D. and Simonoff, A. L. (2023). Statistical evaluation of character support reveals the instability of higher-level dinosaur phylogeny. *Scientific Reports*, 13(1):9273.

Chakerian, J. and Holmes, S. (2012). Computational tools for evaluating phylogenetic and hierarchical clustering trees. *Journal of Computational and Graphical Statistics*, 21(3):581–599.

Chang, J. and Alfaro, M. E. (2016). Crowdsourced geometric morphometrics enable rapid large-scale collection and analysis of phenotypic data. *Methods in Ecology and Evolution*, 7(4):472–482.

Chang, J., Rabosky, D. L., and Alfaro, M. E. (2020). Estimating diversification rates on incompletely sampled phylogenies: theoretical concerns and practical solutions. *Systematic Biology*, 69(3):602–611.

Chang, J., Rabosky, D. L., Smith, S. A., and Alfaro, M. E. (2019). An R package and online resource for macroevolutionary studies using the ray-finned fish tree of life. *Methods in Ecology and Evolution*, 10(7):1118–1124.

Chang, S. Y., Chen, Y., Kao, Y.-C., and Chen, H. H. (2023). Distributed IoT community detection via Gromov-Wasserstein metric. *IEEE Internet of Things Journal*. doi:10.1109/JIOT.2023.3332740.

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, 11(4):265–270.

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 43(4):783–791.

Charig, A. J. (1982). Problems in dinosaur phylogeny: a reasoned approach to their attempted resolution. *Geobios*, 15:113–126.

Charig, A. J., Attridge, J., and Crompton, A. W. (1965). On the origin of the sauropods and the classification of the Saurischia. *Proceedings of the Linnean Society of London*, 176(2):197–221.

Cheon, S. and Liang, F. (2014). Bayesian phylogeny analysis. In Chen, M.-H., Kuo, L., and Lewis, P. O., editors, *Bayesian Phylogenetics: Methods, Algorithms, and Applications*, chapter 7, pages 129–162. Chapman & Hall/CRC Press, Boca Raton, FL.

Chiarenza, A. A., Mannion, P. D., Lunt, D. J., Farnsworth, A., Jones, L. A., Kelland, S.-J., and Allison, P. A. (2019). Ecological niche modelling does not support climatically-driven dinosaur diversity decline before the Cretaceous/Paleogene mass extinction. *Nature Communications*, 10(1):1091.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4(1):266–298.

Close, R. A., Evers, S. W., Alroy, J., and Butler, R. J. (2018). How should we estimate diversity in the fossil record? Testing richness estimators using sampling-standardised discovery curves. *Methods in Ecology and Evolution*, 9(6):1386–1400.

Colless, D. H. (1982). [Review of] *Phylogenetics: the theory and practice of phylogenetic systematics*. *Systematic Zoology*, 31(1):100–104.

Colwell, R. K. (2013). EstimateS: Statistical estimation of species richness and shared species from samples. Version 9. User's Guide. `http://viceroy.colorado.edu/estimates/EstimateSPages/EstSUsersGuide/EstimateSUsersGuide.htm#Chao2`. Accessed March 9, 2021.

Condamine, F. L., Clapham, M. E., and Kergoat, G. J. (2016). Global patterns of insect diversification: towards a reconciliation of fossil and molecular evidence? *Scientific Reports*, 6(1):19208.

Connolly, S. R. and Miller, A. I. (2001). Joint estimation of sampling and turnover rates from fossil databases: capture-mark-recapture methods revisited. *Paleobiology*, 27(4):751–767.

Cook, S. R., Gelman, A., and Rubin, D. B. (2006). Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, 15(3):675–692.

Cooper, M. R. (1981). The prosauropod dinosaur *Massospondylus carinatus* Owen from Zimbabwe: its biology, mode of life and phylogenetic significance. *Occasional Papers of the National Museums and Monuments of Rhodesia, series B, Natural sciences*, 6(10):689–840.

Cooper, M. R. (1985). A revision of the ornithischian dinosaur *Kangnasaurus coetzeei* Haughton, with a classification of the Ornithischia. *Annals of the South African Museum*, 95(8):281–317.

Cotton, J. A., Slater, C. S. C., and Wilkinson, M. (2006). Discriminating supported and unsupported relationships in supertrees using triplets. *Systematic Biology*, 55(2):345–350.

Crampton, J. S., Cooper, R. A., Foote, M., and Sadler, P. M. (2020). Ephemeral species in the fossil record? Synchronous coupling of macroevolutionary dynamics in mid-Paleozoic zooplankton. *Paleobiology*, 46(1):123–135.

Criscuolo, A., Berry, V., Douzery, E. J. P., and Gascuel, O. (2006). SDM: A fast distance-based approach for (super)tree building in phylogenomics. *Systematic Biology*, 55(5):740–755.

Criscuolo, A. and Gascuel, O. (2008). Fast NJ-like algorithms to deal with incomplete distance matrices. *BMC Bioinformatics*, 9(1):166.

Crouch, N. M. A. (2020). Extinction rates of non-avian dinosaur species are uncorrelated with the rate of evolution of phylogenetically informative characters. *Biology Letters*, 16(6):20200231.

Crouch, N. M. A., Ramanauskas, K., and Igić, B. (2019). Tip-dating and the origin of Telluraves. *Molecular Phylogenetics and Evolution*, 131:55–63.

Cruzado-Caballero, P., Gasca, J. M., Filippi, L. S., Cerda, I. A., and Garrido, A. C. (2019). A new ornithopod dinosaur from the Santonian of Northern Patagonia (Rincón de los Sauces, Argentina). *Cretaceous Research*, 98:211–229.

Danis, J. (1986). Quarries of Dinosaur Provincial Park. In Naylor, B. G., editor, *Dinosaur Systematics Symposium: Field Trip Guidebook to Dinosaur Provincial Park, 2 June 1986*, pages 43–51. Tyrrell Museum of Palaeontology, Drumheller, Alberta.

Darroch, S. A. F. and Wagner, P. J. (2015). Response of beta diversity to pulses of Ordovician-Silurian mass extinction. *Ecology*, 96(2):532–549.

Dasarathy, G., Nowak, R., and Roch, S. (2015). Data requirement for phylogenetic inference from multiple loci: a new distance method. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(2):422–432.

Davis, B. W., Li, G., and Murphy, W. J. (2010). Supermatrix and species tree methods resolve phylogenetic relationships within the big cats, *Panthera* (Carnivora: Felidae). *Molecular Phylogenetics and Evolution*, 56:64–76.

Day, W. H. E. (1987). Computational complexity of inferring phylogenies from dissimilarity matrices. *Bulletin of Mathematical Biology*, 49(4):461–467.

De Oliveira Martins, L., Mallo, D., and Posada, D. (2016). A Bayesian supertree model for genome-wide species tree reconstruction. *Systematic Biology*, 65(3):397–416.

de Queiroz, A. and Gatesy, J. (2007). The supermatrix approach to systematics. *Trends in Ecology & Evolution*, 22(1):34–41.

De Soete, G. (1984). Ultrametric tree representations of incomplete dissimilarity data. *Journal of Classification*, 1(1):235–242.

Delsuc, F., Brinkmann, H., and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews Genetics*, 6(5):361–375.

Dieudonné, P.-E., Cruzado-Caballero, P., Godefroit, P., and Tortosa, T. (2020). A new phylogeny of cerapodan dinosaurs. *Historical Biology*, 33(10):2335–2355.

Dieudonné, P.-E., Tortosa, T., Torcida Fernández-Baldor, F., Canudo, J. I., and Díaz-Martínez, I. (2016). An unexpected early rhabdodontid from Europe (Lower Cretaceous of Salas de los Infantes, Burgos Province, Spain) and a re-examination of basal iguanodontian relationships. *PLOS ONE*, 11(6):1–40.

Dimitrova, D. S., Kaishev, V. K., and Tan, S. (2020). Computing the Kolmogorov-Smirnov distribution when the underlying CDF is purely discrete, mixed, or continuous. *Journal of Statistical Software*, 95(10):1–42.

dos Reis, M., Donoghue, P. C. J., and Yang, Z. (2016). Bayesian molecular clock dating of species divergences in the genomics era. *Nature Reviews Genetics*, 17(2):71–80.

dos Reis, M., Inoue, J., Hasegawa, M., Asher, R. J., Donoghue, P. C. J., and 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(1742):3491–3500.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLOS Biology*, 4(5):e88.

du Plessis, L., McCrone, J. T., Zarebski, A. E., Hill, V., Ruis, C., Gutierrez, B., Raghwani, J., Ashworth, J., Colquhoun, R., Connor, T. R., Faria, N. R., Jackson, B., Loman, N. J., O'Toole, A., Nicholls, S. M., Parag, K. V., Scher, E., Vasylyeva, T. I., Volz, E. M., Watts, A., Bogoch, I. I., Khan, K., Consortium, C.-. G. U. C.-U., Aanensen, D. M., Kraemer, M. U. G., Rambaut, A., and Pybus, O. G. (2021). Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science*, 371(6530):708–712.

Eberth, D. A. and Getty, M. A. (2005). Ceratopsian bonebeds: occurrence, origins, and significance. In Currie, P. J. and Koppelhus, E. B., editors, *Dinosaur Provincial Park: A Spectacular Ancient Ecosystem Revealed*, pages 501–536. Indiana University Press, Bloomington, IN.

Eizirik, E., Murphy, W. J., Koepfli, K.-P., Johnson, W. E., Dragoo, J. W., Wayne, R. K., and O'Brien, S. J. (2010). Pattern and timing of diversification of the mammalian order Carnivora inferred from multiple nuclear gene sequences. *Molecular Phylogenetics and Evolution*, 56:49–63.

Eliason, C. M., Edwards, S. V., and Clarke, J. A. (2019). `phenotools`: An R package for visualizing and analysing phenomic datasets. *Methods in Ecology and Evolution*, 10(9):1393–1400.

Erdős, P. L., Steel, M. A., Székely, L. A., and Warnow, T. J. (1997). Local quartet splits of a binary tree imply all quartet splits under one dyadic inference rule. *Computers and Artificial Intelligence*, 16(2):217–227.

Etienne, R. S., Haegeman, B., Stadler, T., Aze, T., Pearson, P. N., Purvis, A., and Phillimore, A. B. (2012). Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. *Proceedings of the Royal Society B: Biological Sciences*, 279(1732):1300–1309.

Ezcurra, M. D. (2010). A new early dinosaur (Saurischia: Sauropodomorpha) from the Late Triassic of Argentina: a reassessment of dinosaur origin and phylogeny. *Journal of Systematic Palaeontology*, 8(3):371–425.

Ezcurra, M. D. and Butler, R. J. (2018). The rise of the ruling reptiles and ecosystem recovery from the Permo-Triassic mass extinction. *Proceedings of the Royal Society B: Biological Sciences*, 285(1880):20180361.

Fabreti, L. G. a. and Höhna, S. (2021). Convergence assessment for Bayesian phylogenetic analysis using MCMC simulation. *Methods in Ecology and Evolution*, 13(1):77–90.

Famoso, N. A., Davis, E. B., Feranec, R. S., Hopkins, S. S. B., and Price, S. A. (2016). Are hypsodonty and occlusal enamel complexity evolutionarily correlated in ungulates? *Journal of Mammalian Evolution*, 23(1):43–47.

Fan, Y., Wu, R., Chen, M.-H., Kuo, L., and Lewis, P. O. (2011). Choosing among partition models in Bayesian phylogenetics. *Molecular Biology and Evolution*, 28(1):523–532.

Fastovsky, D. E., Huang, Y., Hsu, J., Martin-McNaughton, J., Sheehan, P. M., and Weishampel, D. B. (2004). Shape of Mesozoic dinosaur richness. *Geology*, 32(10):877–880.

Felice, R. N., Watanabe, A., Cuff, A. R., Hanson, M., Bhullar, B.-A. S., Rayfield, E. R., Witmer, L. M., Norell, M. A., and Goswami, A. (2020). Decelerated dinosaur skull evolution with the origin of birds. *PLoS Biology*, 18(8):e3000801.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, 27(4):401–410.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376.

Felsenstein, J. (1985). Phylogenies and the comparative method. *American Naturalist*, 125(1):1–15.

Felsenstein, J. (1989). PHYLIP—Phylogeny inference package (version 3.2). *Cladistics*, 5(2):163–166.

Felsenstein, J. (1997). An alternating least squares approach to inferring phylogenies from pairwise distances. *Systematic Biology*, 46(1):101–111.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.

Foote, M. (1988). Survivorship analysis of Cambrian and Ordovician trilobites. *Paleobiology*, 14(3):258–271.

Foote, M. (1996). On the probability of ancestors in the fossil record. *Paleobiology*, 22(2):141–151.

Foote, M. (2000). Origination and extinction components of taxonomic diversity: general problems. *Paleobiology*, 26(S4):74–102.

Foote, M. (2003). Origination and extinction through the Phanerozoic: a new approach. *Journal of Geology*, 111(2):125–148.

Foote, M. (2007). Symmetric waxing and waning of marine invertebrate genera. *Paleobiology*, 33(4):517–529.

Foote, M., Cooper, R. A., Crampton, J. S., and Sadler, P. M. (2018). Diversity-dependent evolutionary rates in early Palaeozoic zooplankton. *Proceedings of the Royal Society B: Biological Sciences*, 285(1873):20180122.

Foote, M., Sadler, P. M., Cooper, R. A., and Crampton, J. S. (2019). Completeness of the known graptoloid palaeontological record. *Journal of the Geological Society*, 176(6):1038–1055.

Fourment, M., Claywell, B. C., Dinh, V., McCoy, C., Matsen, F. A., and Darling, A. E. (2018). Effective online Bayesian phylogenetics via sequential Monte Carlo with guided proposals. *Systematic Biology*, 67(3):490–502.

Fourment, M., Magee, A. F., Whidden, C., Bilge, A., Matsen, Frederick A, I., and Minin, V. N. (2019). 19 dubious ways to compute the marginal likelihood of a phylogenetic tree topology. *Systematic Biology*, 69(2):209–220.

Francis, W. R. and Canfield, D. E. (2020). Very few sites can reshape the inferred phylogenetic tree. *PeerJ*, 8:e8865.

Friedman, M., Feilich, K. L., Beckett, H. T., Alfaro, M. E., Faircloth, B. C., Černý, D., Miya, M., Near, T. J., and Harrington, R. C. (2019). A phylogenomic framework for pelagiarian fishes (Acanthomorpha: Percomorpha) highlights mosaic radiation in the open ocean. *Proceedings of the Royal Society B: Biological Sciences*, 286(1910):20191502.

Friel, N. and Pettitt, A. N. (2008). Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(3):589–607.

Fulton, T. L. and Strobeck, C. (2010). Multiple fossil calibrations, nuclear loci and mitochondrial genomes provide new insight into biogeography and divergence timing for true seals (Phocidae, Pinnipedia). *Journal of Biogeography*, 37(5):814–829.

Garnier, S., Ross, N., Rudis, B., Sciaini, M., Camargo, A. P., and Scherer, C. (2022). *viridisLite: Colorblind-friendly color maps (lite version)*. R package version 0.4.1.

Gascuel, O. and Steel, M. (2006). Neighbor-joining revealed. *Molecular Biology and Evolution*, 23(11):1997–2000.

Gaubert, P. and Begg, C. M. (2006). Re-assessed molecular phylogeny and evolutionary scenario within genets (Carnivora, Viverridae, Genettinae). *Molecular Phylogenetics and Evolution*, 44(2):920–927.

Gaubert, P. and Cordeiro-Estrela, P. (2006). Phylogenetic systematics and tempo of evolution of the Viverrinae (Mammalia, Carnivora, Viverridae) within feliformians: implications for faunal exchanges between Asia and Africa. *Molecular Phylogenetics and Evolution*, 41(2):266–278.

Gauthier, J. A. (1986). Saurischian monophyly and the origin of birds. *Memoirs of the California Academy of Sciences*, 8:1–55.

Gavryushkina, A. (2018). Re: Conditioning on rho and/or the root for a wholly extinct phylogeny. http://groups.google.com/forum/#!topic/beast-users/I-FvZgACAOQ. Accessed December 28, 2019.

Gavryushkina, A., Welch, D., Stadler, T., and Drummond, A. J. (2014). Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLoS Computational Biology*, 10(12):e1003919.

Gearty, W., Jones, L. A., and Chamberlain, S. (2023). *rphylopic: an R package for accessing and plotting PhyloPic silhouettes*. R package version 1.0.0.

Gee, B. M. (2021). Returning to the roots: resolution, reproducibility, and robusticity in the phylogenetic inference of Dissorophidae (Amphibia: Temnospondyli). *PeerJ*, 9:e12423.

Gelman, A. (2017). Correction to Cook, Gelman, and Rubin (2006). *Journal of Computational and Graphical Statistics*, 26(4):940.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science*, 13(2):163–185.

Gernhard, T. (2008). The conditioned reconstructed process. *Journal of Theoretical Biology*, 253(4):769–778.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In Keramidas, E. M., editor, *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 156–163. Interface Foundation, Fairfax Station, VA.

Gilks, W. R. and Roberts, G. O. (1996). Strategies for improving MCMC. In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors, *Markov Chain Monte Carlo in Practice*, pages 89–114. Chapman & Hall, London, UK.

Giribet, G. (2010). A new dimension in combining data? The use of morphology and phylogenomic data in metazoan systematics. *Acta Zoologica*, 91(1):11–19.

Goloboff, P. A., De Laet, J., Ríos-Tamayo, D., and Szumik, C. A. (2021). A reconsideration of inapplicable characters, and an approximation with step-matrix recoding. *Cladistics*, 37(5):596–629.

Goloboff, P. A., Pittman, M., Pol, D., and Xu, X. (2018). Morphological data sets fit a common mechanism much more poorly than DNA sequences and call into question the Mkv model. *Systematic Biology*, 68(3):494–504.

Goloboff, P. A. and Sereno, P. C. (2021). Comparative cladistics: identifying the sources for differing phylogenetic results between competing morphology-based datasets. *Journal of Systematic Palaeontology*, 19(11):761–786.

Gotelli, N. J. and Colwell, R. K. (2011). Estimating species richness. In Magurran, A. E. and McGill, B. J., editors, *Biological Diversity: Frontiers in Measurement and Assessment*, pages 39–54. Oxford University Press, Oxford, UK.

Gradstein, F. M., Ogg, J. G., Schmitz, M. D., and Ogg, G. M., editors (2012). *The Geologic Time Scale 2012*. Elsevier, Boston.

Gradstein, F. M., Ogg, J. G., and Smith, A. G., editors (2005). *A Geologic Time Scale 2004*. Cambridge University Press, Cambridge, UK.

Griffin, C. T., Wynd, B. M., Munyikwa, D., Broderick, T. J., Zondo, M., Tolan, S., Langer, M. C., Nesbitt, S. J., and Taruvinga, H. R. (2022). Africa's oldest dinosaurs reveal early suppression of dinosaur distribution. *Nature*, 609(7926):313–319.

Groussin, M., Pawlowski, J., and Yang, Z. (2011). Bayesian relaxed clock estimation of divergence times in foraminifera. *Molecular Phylogenetics and Evolution*, 61(1):157–166.

Gu, R., Tang, Y., Wang, Z., Wang, S., Yin, X., Yuan, C., and Huang, Y. (2015). Efficient large scale distributed matrix computation with Spark. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 2327–2336, Washington DC. IEEE Computer Society.

Haag, J., Höhler, D., Bettisworth, B., and Stamatakis, A. (2022). From easy to hopeless—Predicting the difficulty of phylogenetic analyses. *Molecular Biology and Evolution*, 39(12):msac254.

Han, F., Forster, C. A., Xu, X., and Clark, J. M. (2018). Postcranial anatomy of *Yinlong downsi* (Dinosauria: Ceratopsia) from the Upper Jurassic Shishugou Formation of China and the phylogeny of basal ornithischians. *Journal of Systematic Palaeontology*, 16(14):1159–1187.

Harrington, S. M., Wishingrad, V., and Thomson, R. C. (2021). Properties of Markov chain Monte Carlo performance across many empirical alignments. *Molecular Biology and Evolution*, 38(4):1627–1640.

Hartigan, J. A. (1967). Representation of similarity matrices by trees. *Journal of the American Statistical Association*, 62(320):1140–1158.

Hartman, S., Mortimer, M., Wahl, W. R., Lomax, D. R., Lippincott, J., and Lovelace, D. M. (2019). A new paravian dinosaur from the Late Jurassic of North America supports a late acquisition of avian flight. *PeerJ*, 7:e7247.

Harvey, P. H., May, R. M., and Nee, S. (1994). Phylogenies without fossils. *Evolution*, 48(3):523–529.

Hassanin, A., Veron, G., Ropiquet, A., Jansen van Vuuren, B., Lécu, A., Goodman, S. M., Haider, J., and Nguyen, T. T. (2021). Evolutionary history of Carnivora (Mammalia, Laurasiatheria) inferred from mitochondrial genomes. *PLoS ONE*, 16(2):e0240770.

Heath, T. A., Huelsenbeck, J. P., and Stadler, T. (2014). The fossilized birth–death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences of the USA*, 111(29):E2957–E2966.

Helgen, K. M., Pinto, C. M., Kays, R. W., Helgen, L. E., Tsuchiya, M. T. N., Quinn, A., Wilson, D. E., and Maldonado, J. E. (2013). Taxonomic revision of the olingos (*Bassaricyon*), with description of a new species, the Olinguito. *ZooKeys*, 324:1–83.

Henao Diaz, L. F., Harmon, L. J., Sugawara, M. T. C., Miller, E. T., and Pennell, M. W. (2019). Macroevolutionary diversification rates show time dependency. *Proceedings of the National Academy of Sciences*, 116(15):7403–7408.

Hendrickx, C., Bell, P. R., Pittman, M., Milner, A. R. C., Cuesta, E., O'Connor, J., Loewen, M., Currie, P. J., Mateus, O., Kaye, T. G., and Delcourt, R. (2022). Morphology and distribution of scales, dermal ossifications, and other non-feather integumentary structures in non-avialan theropod dinosaurs. *Biological Reviews*, 97(3):960–1004.

Hennig, E. (1925). *Kentrurosaurus aethiopicus*. Die Stegosaurierfunde vom Tendaguru, Deutsch-Ostafrika. *Palaeontographica – Supplementbände*, SVII(2):101–254.

Herne, M. C., Nair, J. P., Evans, A. R., and Tait, A. M. (2019). New small-bodied ornithopods (Dinosauria, Neornithischia) from the Early Cretaceous Wonthaggi Formation (Strzelecki Group) of the Australian-Antarctic rift system, with revision of *Qantassaurus intrepidus* Rich and Vickers-Rich, 1999. *Journal of Paleontology*, 93(3):543–584.

Herne, M. C., Tait, A. M., Weisbecker, V., Hall, M., Nair, J. P., Cleeland, M., and Salisbury, S. W. (2018). A new small-bodied ornithopod (Dinosauria, Ornithischia) from a deep, high-energy Early Cretaceous river of the Australian–Antarctic rift system. *PeerJ*, 5:e4113.

Higdon, J. W., Bininda-Emonds, O. R. P., Beck, R. M. D., and Ferguson, S. H. (2007). Phylogeny and divergence of the pinnipeds (Carnivora: Mammalia) assessed using a multigene dataset. *BMC Evolutionary Biology*, 7:216.

Hill, J., Linero, A., and Murray, J. (2020). Bayesian additive regression trees: a review and look forward. *Annual Review of Statistics and Its Application*, 7:251–278.

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2017). UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, 35(2):518–522.

Hodel, R. G. J., Winslow, S. K., Liu, B.-B., Johnson, G., Trizna, M., White, A. E., Dikow, R. B., Potter, D., Zimmer, E. A., and Wen, J. (2023). A phylogenomic approach, combined with morphological characters gleaned via machine learning, uncovers the hybrid origin and biogeographic diversification of the plum genus. *bioRχiv*. doi:10.1101/2023.09.13.557598.

Höhna, S. (2015). The time-dependent reconstructed evolutionary process with a key-role for mass-extinction events. *Journal of Theoretical Biology*, 380:321–331.

Höhna, S., Heath, T. A., Boussau, B., Landis, M. J., Ronquist, F., and Huelsenbeck, J. P. (2014). Probabilistic graphical model representation in phylogenetics. *Systematic Biology*, 63(5):753–771.

Höhna, S., Landis, M. J., and Heath, T. A. (2017). Phylogenetic inference using `RevBayes`. *Current Protocols in Bioinformatics*, 57(1):6.16.1–6.16.34.

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P., and Ronquist, F. (2016). RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology*, 65(4):726–736.

Höhna, S., Stadler, T., Ronquist, F., and Britton, T. (2011). Inferring speciation and extinction rates under different sampling schemes. *Molecular Biology and Evolution*, 28(9):2577–2589.

Holder, M. and Lewis, P. O. (2003). Phylogeny estimation: traditional and Bayesian approaches. *Nature Reviews Genetics*, 4(4):275–284.

Holland, S. M. (2016). The non-uniformity of fossil preservation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1699):20150130.

Hopkins, M. J. (2011). Species-level phylogenetic analysis of pterocephaliids (Trilobita, Cambrian) from the Great Basin, western USA. *Journal of Paleontology*, 85(6):1128–1153.

Hopkins, M. J. and Smith, A. B. (2015). Dynamic evolutionary change in post-Paleozoic echinoids and the importance of scale when interpreting changes in rates of evolution. *Proceedings of the National Academy of Sciences USA*, 112(12):3758–3763.

Hopkins, M. J. and St. John, K. (2021). Incorporating hierarchical characters into phylogenetic analysis. *Systematic Biology*, 70(6):1163–1180.

Horner, J. R. and Currie, P. J. (1994). Embryonic and neonatal morphology and ontogeny of a new species of *Hypacrosaurus* (Ornithischia, Lambeosauridae) from Montana and Alberta. In Carpenter, K., Hirsch, K. F., and Horner, J. R., editors, *Dinosaurs Eggs and Babies*, pages 312–336. Cambridge University Press, Cambridge, UK.

Huelsenbeck, J. P., Larget, B., Miller, R. E., and Ronquist, F. (2002). Potential applications and pitfalls of Bayesian inference of phylogeny. *Systematic Biology*, 51(5):673–688.

Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Molecular Biology and Evolution*, 33(6):1635–1638.

Hunold, S., Casanova, H., and Suter, F. (2011). From simulation to experiment: a case study on multiprocessor task scheduling. In *2011 IEEE International Symposium on Parallel and Distributed Processing Workshops and PhD Forum (IPDPSW)*, pages 665–672, Washington DC. IEEE Computer Society.

Hunt, G. and Slater, G. J. (2016). Integrating paleontological and phylogenetic approaches to macroevolution. *Annual Review of Ecology, Evolution, and Systematics*, 47:189–213.

Hunt, R. and Pedersen, K. S. (2022). Rove-Tree-11: The not-so-Wild Rover. A hierarchically structured image dataset for deep metric learning research. In *Computer Vision – ACCV 2022. 16th Asian Conference on Computer Vision, Macao, China, December 4–8, 2022, Proceedings, Part V*, pages 425–441. Springer, Cham, Switzerland.

Huxley, T. H. (1870). On the classification of the Dinosauria, with observations on the Dinosauria of the Trias. *Quarterly Journal of the Geological Society*, 26(1-2):32–51.

Ibiricu, L. M., Casal, G. A., Martínez, R. D., Luna, M., Canale, J. I., Álvarez, B. N., and González-Riga, B. (2019). A new ornithopod dinosaur (Dinosauria: Ornithischia) from the Late Cretaceous of central Patagonia. *Cretaceous Research*, 98:276–291.

Jablonski, D. (1986). Background and mass extinctions: the alternation of macroevolutionary regimes. *Science*, 231(4734):129–133.

Janssens, S. B., Couvreur, T. L., Mertens, A., Dauby, G., Dagallier, L.-P. M. J., Vanden Abeele, S., Vandelook, F., Mascarello, M., Beeckman, H., Sosef, M., Droissart, V., van der Bank, M., Maurin, O., Hawthorne, W., Marshall, C., Réjou-Méchain, M., Beina, D., Baya, F., Merckx, V., Verstraete, B., and Hardy, O. (2020). A large-scale species level dated angiosperm phylogeny for evolutionary and ecological analyses. *Biodiversity Data Journal*, 8:e39677.

Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K., and Mooers, A. O. (2012). The global diversity of birds in space and time. *Nature*, 491(7424):444–448.

Johnson, W. E., Eizirik, E., Pecon-Slattery, J., Murphy, W. J., Antunes, A., Teeling, E., and O'Brien, S. J. (2006). The late Miocene radiation of modern Felidae: a genetic assessment. *Science*, 311(5757):73–77.

Jombart, T., Kendall, M., Almagro-Garcia, J., and Colijn, C. (2017). TREESPACE: Statistical exploration of landscapes of phylogenetic trees. *Molecular Ecology Resources*, 17(6):1385–1392.

Jukes, T. H. and Cantor, C. R. (1969). Evolution of protein molecules. In Munro, H. N., editor, *Mammalian Protein Metabolism, Vol. III*, chapter 24, pages 21–132. Academic Press, New York.

Kapelner, A. and Bleich, J. (2016). bartMachine: machine learning with Bayesian additive regression trees. *Journal of Statistical Software*, 70(4):1–40.

Karcher, M., Zhang, C., and Matsen, F. A. (2021). Variational Bayesian supertrees. *arχiv*, page 2104.11191v1.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kendall, M. and Colijn, C. (2016). Mapping phylogenetic trees to reveal distinct patterns of evolution. *Molecular Biology and Evolution*, 33(10):2735–2743.

Kettleborough, G., Dicks, J., Roberts, I. N., and Huber, K. T. (2015). Reconstructing (super)trees from data sets with missing distances: not all is lost. *Molecular Biology and Evolution*, 32(6):1628–1642.

Kiat, Y. and O'Connor, J. K. (2024). Functional constraints on the number and shape of flight feathers. *Proceedings of the National Academy of Sciences*, 121(8):e2306639121.

Kiessling, W., Aberhan, M., Brenneis, B., and Wagner, P. J. (2007). Extinction trajectories of benthic organisms across the Triassic–Jurassic boundary. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 244(1):201–222.

Kimball, R. T., Oliveros, C. H., Wang, N., White, N. D., Barker, F. K., Field, D. J., Ksepka, D. T., Chesser, R. T., Moyle, R. G., Braun, M. J., Brumfield, R. T., Faircloth, B. C., Smith, B. T., and Braun, E. L. (2019). A phylogenomic supertree of birds. *Diversity*, 11(7):109.

King, B. and Rücklin, M. (2020). Tip dating with fossil sites and stratigraphic sequences. *PeerJ*, 8:e9368.

Kishino, H. and Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *Journal of Molecular Evolution*, 29(2):170–179.

Kishino, H., Miyata, T., and Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution*, 31(2):151–160.

Kodandaramaiah, U. and Murali, G. (2018). What affects power to estimate speciation rate shifts? *PeerJ*, 6:e5495.

Koepfli, K.-P., Deere, K. A., Slater, G. J., Begg, C., Begg, K., Grassman, L., Lucherini, M., Veron, G., and Wayne, R. K. (2008). Multigene phylogeny of the Mustelidae: resolving relationships, tempo and biogeographic history of a mammalian adaptive radiation. *BMC Biology*, 6:10.

Koepfli, K.-P., Gompper, M. E., Eizirik, E., Ho, C.-C., Linden, L., Maldonado, J. E., and Wayne, R. K. (2007). Phylogeny of the Procyonidae (Mammalia: Carnivora): molecules, morphology and the great American interchange. *Molecular Phylogenetics and Evolution*, 43(3):1076–1095.

Koepfli, K.-P., Jenks, S. M., Eizirik, E., Zahirpour, T., Van Valkenburgh, B., and Wayne, R. K. (2006). Molecular systematics of the Hyaenidae: relationships of a relictual lineage resolved by a molecular supermatrix. *Molecular Phylogenetics and Evolution*, 38(3):603–620.

Krause, J., Unger, T., Noçon, A., Malaspinas, A.-S., Kolokotronis, S.-O., Stiller, M., Soibelzon, L., Spriggs, H., Dear, P. H., Briggs, A. W., Bray, S. C. E., O'Brien, S. J., Rabeder, G., Matheus, P., Cooper, A., Slatkin, M., Pääbo, S., and Hofreiter, M. (2008). Mitochondrial genomes reveal an explosive radiation of extinct and extant bears near the Miocene-Pliocene boundary. *BMC Evolutionary Biology*, 8:220.

Kuhner, M. K. and Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, 11(3):459–468.

Kutschera, V. E., Bidon, T., Hailer, F., Rodi, J. L., Fain, S. R., and Janke, A. (2014). Bears in a forest of gene trees: phylogenetic inference is complicated by incomplete lineage sorting and gene flow. *Molecular Biology and Evolution*, 31(8):2004–2017.

Lakner, C., van der Mark, P., Huelsenbeck, J. P., Larget, B., and Ronquist, F. (2008). Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics. *Systematic Biology*, 57(1):86–103.

Lambe, L. M. (1918). The Cretaceous genus *Stegoceras* typifying a new family referred provisionally to the Stegosauria. *Proceedings and Transactions of the Royal Society of Canada*, Series 3, 12:23–36.

Landis, M. J., Matzke, N. J., Moore, B. R., and Huelsenbeck, J. P. (2013). Bayesian analysis of biogeography when the number of areas is large. *Systematic Biology*, 62(6):789–804.

Landry, P.-A. and Lapointe, F.-J. (1997). Estimation of missing distances in path-length matrices: problems and solutions. In Mirkin, B., McMorris, F. R., Roberts, F. S., and Rzhetsky, A., editors, *Mathematical Hierarchies and Biology: DIMACS Series in Discrete Mathematics and Theoretical Computer Science Volume 37*, pages 209–218, Providence, RI. American Mathematical Society.

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., and Calcott, B. (2016a). PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution*, 34(3):772–773.

Lanfear, R., Hua, X., and Warren, D. L. (2016b). Estimating the effective sample size of tree topologies from Bayesian phylogenetic analyses. *Genome Biology and Evolution*, 8(8):2319–2332.

Langer, M. C., Ezcurra, M. D., Rauhut, O. W. M., Benton, M. J., Knoll, F., McPhee, B. W., Novas, F. E., Pol, D., and Brusatte, S. L. (2017). Untangling the dinosaur family tree. *Nature*, 551(7678):E1–E3.

Langer, M. C., McPhee, B. W., Marsola, J. C. d. A., Roberto-da Silva, L., and Cabreira, S. F. (2019). Anatomy of the dinosaur *Pampadromaeus barberenai* (Saurischia—Sauropodomorpha) from the Late Triassic Santa Maria Formation of southern Brazil. *PLoS ONE*, 14(2):e0212543.

Lapointe, F.-J. and Cucumel, G. (1997). The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa. *Systematic Biology*, 46(2):306–312.

Lapointe, F.-J. and Kirsch, J. A. W. (1995). Estimating phylogenies from lacunose distance matrices, with special reference to DNA hybridization data. *Molecular Biology and Evolution*, 12(2):266–284.

Lapointe, F.-J. and Levasseur, C. (2004). Everything you always wanted to know about the average consensus, and more. In Bininda-Emonds, O. R. P., editor, *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pages 87–105. Kluwer Academic Publishers, Dordrecht, NL.

Lartillot, N. and Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Systematic Biology*, 55(2):195–207.

Laurent, S., Robinson-Rechavi, M., and Salamin, N. (2015). Detecting patterns of species diversification in the presence of both rate shifts and mass extinctions. *BMC Evolutionary Biology*, 15(1):157.

Law, C. J., Slater, G. J., and Mehta, R. S. (2018). Lineage diversity and size disparity in Musteloidea: testing patterns of adaptive radiation using molecular and fossil-based methods. *Systematic Biology*, 67(1):127–144.

Le, N. V. (2012). High-resolution optogenetic functional magnetic resonance imaging powered by compressed sensing and parallel processing. Master's thesis, University of California, Los Angeles, Los Angeles, CA.

Lee, M. S. Y., Baron, M. G., Norman, D. B., and Barrett, P. M. (2018). Dynamic biogeographic models and dinosaur origins. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 109(1-2):325–332.

Levasseur, C., Landry, P.-A., Makarenkov, V., Kirsch, J. A. W., and Lapointe, F.-J. (2003). Incomplete distance matrices, supertrees and bat phylogeny. *Molecular Phylogenetics and Evolution*, 27(2):239–246.

Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, 50(6):913–925.

Li, G., Davis, B. W., Eizirik, E., and Murphy, W. J. (2016). Phylogenomic evidence for ancient hybridization in the genomes of living cats (Felidae). *Genome Research*, 26(1):1–11.

Li, P. and Vu, Q. D. (2013). Identification of parameter correlations for parameter estimation in dynamic biological models. *BMC Systems Biology*, 7(1):91.

Lin, Y., Rajan, V., and Moret, B. M. E. (2012). A metric for phylogenetic trees based on matching. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(4):1014–1022.

Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., Clamp, M., Chang, J. L., Kulbokas, E. J., Zody, M. C., Mauceli, E., Xie, X., Breen, M., Wayne, R. K., Ostrander, E. A., Ponting, C. P., Galibert, F., Smith, D. R., deJong, P. J., Kirkness, E., Alvarez, P., Biagi, T., Brockman, W., Butler, J., Chin, C.-W., Cook, A., Cuff, J., Daly, M. J., DeCaprio, D., Gnerre, S., Grabherr, M., Kellis, M., Kleber, M., Bardeleben, C., Goodstadt, L., Heger, A., Hitte, C., Kim, L., Koepfli, K.-P., Parker, H. G., Pollinger, J. P., Searle, S. M. J., Sutter, N. B., Thomas, R., Webber, C., Baldwin, J., Abebe, A., Abouelleil, A., Aftuck, L., Ait-zahra, M., Aldredge, T., Allen, N., An, P., Anderson, S., Antoine, C., Arachchi, H., Aslam, A., Ayotte, L., Bachantsang, P., Barry, A., Bayul, T., Benamara, M., Berlin, A., Bessette, D., Blitshteyn, B., Bloom, T., Blye, J., Boguslavskiy, L., Bonnet, C., Boukhgalter, B., Brown, A., Cahill, P., Calixte, N., Camarata, J., Cheshatsang, Y., Chu, J., Citroen, M., Collymore, A., Cooke, P., Dawoe, T., Daza, R., Decktor, K., DeGray, S., Dhargay, N., Dooley, K., Dooley, K., Dorje, P., Dorjee, K., Dorris, L., Duffey, N., Dupes, A., Egbiremolen, O., Elong, R., Falk, J., Farina, A., Faro, S., Ferguson, D., Ferreira, P., Fisher, S., FitzGerald, M., Foley, K., Foley, C., Franke, A., Friedrich, D., Gage, D., Garber, M., Gearin, G., Giannoukos, G., Goode, T., Goyette, A., Graham, J., Grandbois, E., Gyaltsen, K., Hafez, N., Hagopian, D., Hagos, B., Hall, J., Healy, C., Hegarty, R., Honan, T., Horn, A., Houde, N., Hughes, L., Hunnicutt, L., Husby, M., Jester, B., Jones, C., Kamat, A., Kanga, B., Kells, C., Khazanovich, D., Kieu, A. C., Kisner, P., Kumar, M., Lance, K., Landers, T., Lara, M., Lee, W., Leger, J.-P., Lennon, N., Leuper, L., LeVine, S., Liu, J., Liu, X., Lokyitsang, Y., Lokyitsang, T., Lui, A., Macdonald, J., Major, J., Marabella, R., Maru, K., Matthews, C., McDonough, S., Mehta, T., Meldrim, J., Melnikov, A., Meneus, L., Mihalev, A., Mihova, T., Miller, K., Mittelman, R., Mlenga, V., Mulrain, L., Munson, G., Navidi, A., Naylor, J., Nguyen, T., Nguyen, N., Nguyen, C., Nguyen, T., Nicol, R., Norbu, N., Norbu, C., Novod, N., Nyima, T., Olandt, P., O'Neill, B., O'Neill, K., Osman, S., Oyono, L., Patti, C., Perrin, D., Phunkhang, P., Pierre, F., Priest, M., Rachupka, A., Raghuraman, S., Rameau, R., Ray, V., Raymond, C., Rege, F., Rise, C., Rogers, J., Rogov, P., Sahalie, J., Settipalli, S., Sharpe, T., Shea, T., Sheehan, M., Sherpa, N., Shi, J., Shih, D., Sloan, J., Smith, C., Sparrow, T., Stalker, J., Stange-Thomann, N., Stavropoulos, S., Stone, C., Stone, S., Sykes, S., Tchuinga, P., Tenzing, P., Tesfaye, S., Thoulutsang, D., Thoulutsang, Y., Topham, K., Topping, I., Tsamla, T., Vassiliev, H., Venkataraman, V., Vo, A., Wangchuk, T., Wangdi, T., Weiand, M., Wilkinson, J., Wilson, A., Yadav, S., Yang, S., Yang, X., Young, G., Yu, Q., Zainoun, J., Zembek, L., Zimmer, A., Lander, E. S., and members, B. S. P. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438(7069):803–819.

Liow, L. H. and Finarelli, J. A. (2014). A dynamic global equilibrium in carnivoran diversification over 20 million years. *Proceedings of the Royal Society B: Biological Sciences*, 281(1778):20132312.

Liow, L. H., Quental, T. B., and Marshall, C. R. (2010). When can decreasing diversification

rates be detected with molecular phylogenies and the fossil record? *Systematic Biology*, 59(6):646–659.

Lloyd, G. T. (2012). A refined modelling approach to assess the influence of sampling on palaeobiodiversity curves: new support for declining Cretaceous dinosaur richness. *Biology Letters*, 8(1):123–126.

Lloyd, G. T., Davis, K. E., Pisani, D., Tarver, J. E., Ruta, M., Sakamoto, M., Hone, D. W. E., Jennings, R., and Benton, M. J. (2008). Dinosaurs and the Cretaceous Terrestrial Revolution. *Proceedings of the Royal Society B: Biological Sciences*, 275(1650):2483–2490.

Lloyd, G. T. and Slater, G. J. (2020). A total-group phylogenetic metatree for Cetacea and the importance of fossil data in diversification analyses. *bioRχiv*. doi:10.1101/2020.06.24.169078.

Lloyd, G. T. and Slater, G. J. (2021). A total-group phylogenetic metatree for Cetacea and the importance of fossil data in diversification analyses. *Systematic Biology*, 70(5):922–939.

Lopes, F., Oliveira, L. R., Kessler, A., Beux, Y., Crespo, E., Cárdenas-Alayza, S., Majluf, P., Sepúlveda, M., Brownell Jr, R. L., Franco-Trecu, V., Páez-Rosas, D., Chaves, J., Loch, C., Robertson, B. C., Acevedo-Whitehouse, K., Elorriaga-Verplancken, F. R., Kirkman, S. P., Peart, C. R., Wolf, J. B. W., and Bonatto, S. L. (2021). Phylogenomic discordance in the eared seals is best explained by incomplete lineage sorting following explosive radiation in the Southern Hemisphere. *Systematic Biology*, 70(4):786–802.

Louca, S. and Doebeli, M. (2017). Efficient comparative phylogenetics on large trees. *Bioinformatics*, 34(6):1053–1055.

Louca, S. and Pennell, M. W. (2020). Extant timetrees are consistent with a myriad of diversification histories. *Nature*, 580(7804):502–505.

Lull, R. S. (1915). The mammals and horned dinosaurs of the Lance Formation of Niobrara County, Wyoming. *American Journal of Science*, Series 4, 40(238):319–348.

Lull, R. S. (1933). A revision of the Ceratopsia or horned dinosaurs. *Memoirs of the Peabody Museum of Natural History*, 3(3):1–175.

Lull, R. S. and Wright, N. E. (1942). Hadrosaurian dinosaurs of North America. *Geological Society of America Special Paper*, 40:1–242.

Lynch, L. M. (2020). Fossil calibration of mitochondrial phylogenetic relationships of North American pine martens, *Martes*, suggests an older divergence of *M. americana* and *M. caurina* than previously hypothesized. *Journal of Mammalian Evolution*, 27(3):535–548.

MacLaren, J. A. (2021). Biogeography a key influence on distal forelimb variation in horses through the Cenozoic. *Proceedings of the Royal Society B: Biological Sciences*, 288(1942):20202465.

Madzia, D., Boyd, C. A., and Mazuch, M. (2018). A basal ornithopod dinosaur from the Cenomanian of the Czech Republic. *Journal of Systematic Palaeontology*, 16(11):967–979.

Makarenkov, V. (2001). T-REX: reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics*, 17(7):664–668.

Makarenkov, V. and Lapointe, F.-J. (2004). A weighted least-squares approach for inferring phylogenies from incomplete distance matrices. *Bioinformatics*, 20(13):2113–2121.

Makarenkov, V. and Leclerc, B. (1999). An algorithm for the fitting of a tree metric according to a weighted least-squares criterion. *Journal of Classification*, 16(1):3–26.

Maliet, O., Hartig, F., and Morlon, H. (2019). A model with many small shifts for estimating species-specific diversification rates. *Nature Ecology & Evolution*, 3(7):1086–1092.

Marshall, C. R. (2017). Five palaeobiological laws needed to understand the evolution of the living biota. *Nature Ecology & Evolution*, 1(6):0165.

Marshall, C. R. (2019). Using the fossil record to evaluate timetree timescales. *Frontiers in Genetics*, 10:1049.

Marsola, J. C. A., Bittencourt, J. S., Butler, R. J., Da Rosa, A. A. S., Sayão, J. M., and Langer, M. C. (2018). A new dinosaur with theropod affinities from the Late Triassic Santa Maria Formation, south Brazil. *Journal of Vertebrate Paleontology*, 38(5):e1531878.

Maturana Russel, P., Brewer, B. J., Klaere, S., and Bouckaert, R. R. (2018). Model selection and parameter inference in phylogenetics using nested sampling. *Systematic Biology*, 68(2):219–233.

May, M. R., Höhna, S., and Moore, B. R. (2016). A Bayesian approach for detecting the impact of mass-extinction events on molecular phylogenies when rates of lineage diversification may vary. *Methods in Ecology and Evolution*, 7(8):947–959.

McDonough, M. M., Ferguson, A. W., Dowler, R. C., Gompper, M. E., and Maldonado, J. E. (2022). Phylogenomic systematics of the spotted skunks (Carnivora, Mephitidae, *Spilogale*): Additional species diversity and Pleistocene climate change as a major driver of diversification. *Molecular Phylogenetics and Evolution*, 167:107266.

McHugh, S. W., Espíndola, A., White, E., and Uyeda, J. (2022). Jointly modeling species niche and phylogenetic model in a Bayesian hierarchical framework. *bioRχiv*. doi:10.1101/2022.07.06.499056.

McPhee, B. W., Bordy, E. M., Sciscio, L., and Choiniere, J. N. (2017). The sauropodomorph biostratigraphy of the Elliot Formation of southern Africa: Tracking the evolution of Sauropodomorpha across the Triassic–Jurassic boundary. *Acta Palaeontologica Polonica*, 62(3):441–465.

Menzel, U. (2021). *EMT: Exact multinomial test: goodness-of-fit test for discrete multivariate data*. R package version 1.2.

Meredith, R. W., Janečka, J. E., Gatesy, J., Ryder, O. A., Fisher, C. A., Teeling, E. C., Goodbla, A., Eizirik, E., Simão, T. L. L., Stadler, T., Rabosky, D. L., Honeycutt, R. L., Flynn, J. J., Ingram, C. M., Steiner, C., Williams, T. L., Robinson, T. J., Burk-Herrick, A., Westerman, M., Ayoub, N. A., Springer, M. S., and Murphy, W. J. (2011). Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. *Science*, 334(6055):521–524.

Meseguer, A. S., Lobo, J. M., Ree, R., Beerling, D. J., and Sanmartín, I. (2014). Integrating fossils, phylogenies, and niche models into biogeography to reveal ancient evolutionary history: the case of *Hypericum* (Hypericaceae). *Systematic Biology*, 64(2):215–232.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, 37(5):1530–1534.

Mitchell, J. S., Etienne, R. S., and Rabosky, D. L. (2018). Inferring diversification rate variation from phylogenies with fossils. *Systematic Biology*, 68(1):1–18.

Mitchell, J. S. and Rabosky, D. L. (2017). Bayesian model selection with BAMM: effects of the model prior on the inferred number of diversification shifts. *Methods in Ecology and Evolution*, 8(1):37–46.

Mitra, S., Kunteepuram, V., Koepfli, K.-P., Mehra, N., Tabasum, W., Sreenivas, A., and Gaur, A. (2019). Characteristics of the complete mitochondrial genome of the monotypic genus *Arctictis* (Family: Viverridae) and its phylogenetic implications. *PeerJ*, 7:e8033.

Modrák, M., Moon, A. H., Kim, S., Bürkner, P., Huurre, N., Faltejsková, K., Gelman, A., and Vehtari, A. (2023). Simulation-based calibration checking for Bayesian computation: the choice of test quantities shapes sensitivity. *Bayesian Analysis*. doi:10.1214/23-BA1404.

Moore, B. R., Höhna, S., May, M. R., Rannala, B., and Huelsenbeck, J. P. (2016). Critically evaluating the theory and performance of Bayesian analysis of macroevolutionary mixtures. *Proceedings of the National Academy of Sciences*, 113(34):9569–9574.

Morel, B., Barbera, P., Czech, L., Bettisworth, B., Hübner, L., Lutteropp, S., Serdari, D., Kostaki, E.-G., Mamais, I., Kozlov, A. M., Pavlidis, P., Paraskevis, D., and Stamatakis, A. (2020). Phylogenetic analysis of SARS-CoV-2 data is difficult. *Molecular Biology and Evolution*, 38(5):1777–1791.

Morlon, H. (2014). Phylogenetic approaches for studying diversification. *Ecology Letters*, 17(4):508–525.

Morlon, H., Parsons, T. L., and Plotkin, J. B. (2011). Reconciling molecular phylogenies with the fossil record. *Proceedings of the National Academy of Sciences*, 108(39):16327–16332.

Moyle, R. G., Filardi, C. E., Smith, C. E., and Diamond, J. (2009). Explosive Pleistocene diversification and hemispheric expansion of a "great speciator". *Proceedings of the National Academy of Sciences*, 106(6):1863–1868.

Müller, R. T. and Dias-da Silva, S. (2019). Taxon sample and character coding deeply impact unstable branches in phylogenetic trees of dinosaurs. *Historical Biology*, 31(8):1089–1092.

Müller, R. T. and Garcia, M. S. (2020). A paraphyletic 'Silesauridae' as an alternative hypothesis for the initial radiation of ornithischian dinosaurs. *Biology Letters*, 16(8):20200417.

Müller, R. T., Pretto, F. A., Kerber, L., Silva-Neves, E., and Dias-da Silva, S. (2018). Comment on 'A dinosaur missing-link? *Chilesaurus* and the early evolution of ornithischian dinosaurs'. *Biology Letters*, 14(3):20170581.

Nascimento, F. O. D., Cheng, J., and Feijó, A. (2021). Taxonomic revision of the pampas cat *Leopardus colocola* complex (Carnivora: Felidae): an integrative approach. *Zoological Journal of the Linnean Society*, 191(2):575–611.

Nayak, A. (1999). Design and characterization of MATLAB functions on a distributed and heterogeneous system. Master's thesis, Northwestern University, Evanston, IL.

Nee, S., May, R. M., and Harvey, P. H. (1994). The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 344(1309):305–311.

Nesbitt, S. J. (2011). The early evolution of archosaurs: Relationships and the origin of major clades. *Bulletin of the American Museum of Natural History*, 352:1–292.

Nesbitt, S. J., Barrett, P. M., Werning, S., Sidor, C. A., and Charig, A. J. (2013). The oldest dinosaur? A Middle Triassic dinosauriform from Tanzania. *Biology Letters*, 9(1):20120949.

Nesbitt, S. J., Langer, M. C., and Ezcurra, M. D. (2020). The anatomy of *Asilisaurus kongwe*, a dinosauriform from the Lifua Member of the Manda Beds (∼Middle Triassic) of Africa. *The Anatomical Record*, 303(4):813–873.

Nguyen, N., Mirarab, S., and Warnow, T. (2012). MRL and SuperFine+MRL: new supertree methods. *Algorithms for Molecular Biology*, 7(1):3.

Nigenda-Morales, S. F., Gompper, M. E., Valenzuela-Galván, D., Lay, A. R., Kapheim, K. M., Hass, C., Booth-Binczik, S. D., Binczik, G. A., Hirsch, B. T., McColgin, M., Koprowski, J. L., McFadden, K., Wayne, R. K., and Koepfli, K.-P. (2019). Phylogeographic and diversification patterns of the white-nosed coati (*Nasua narica*): Evidence for south-to-north colonization of North America. *Molecular Phylogenetics and Evolution*, 131:149–163.

Norell, M. A. (1993). Tree-based approaches to understanding history; comments on ranks, rules and the quality of the fossil record. *American Journal of Science*, 293(A):407–417.

Norman, D. B., Baron, M. G., Garcia, M. S., and Müller, R. T. (2022). Taxonomic, palaeobiological and evolutionary implications of a phylogenetic hypothesis for Ornithischia (Archosauria: Dinosauria). *Zoological Journal of the Linnean Society*, 196(4):1273–1309.

Novas, F. E. (1996). Dinosaur monophyly. *Journal of Vertebrate Paleontology*, 16(4):723–741.

Nyakatura, K. and Bininda-Emonds, O. R. P. (2012). Updating the evolutionary history of Carnivora (Mammalia): a new species-level supertree complete with divergence time estimates. *BMC biology*, 10(1):1–31.

O'Connor, P. M., Turner, A. H., Groenke, J. R., Felice, R. N., Rogers, R. R., Krause, D. W., and Rahantarisoa, L. J. (2020). Late cretaceous bird from Madagascar reveals unique development of beaks. *Nature*, 588(7837):272–276.

Ogilvie, H. A., Mendes, F. K., Vaughan, T. G., Matzke, N. J., Stadler, T., Welch, D., and Drummond, A. J. (2022). Novel integrative modeling of molecules and morphology across evolutionary timescales. *Systematic Biology*, 71(1):208–220.

O'Leary, M. A., Alphonse, K., Mariangeles, A. H., Cavaliere, D., Cirranello, A., Dietterich, T. G., Julius, M., Kaufman, S., Law, E., Passarotti, M., Reft, A., Robalino, J., Simmons, N. B., Smith, S. Y., Stevenson, D. W., Theriot, E., Velazco, P. M., Walls, R. L., Yu, M., and Daly, M. (2018). Crowds replicate performance of scientific experts scoring phylogenetic matrices of phenotypes. *Systematic Biology*, 67(1):49–60.

O'Leary, M. A. and Kaufman, S. (2011). MorphoBank: phylophenomics in the "cloud". *Cladistics*, 27(5):529–537.

O'Reilly, J. E., Puttick, M. N., Parry, L., Tanner, A. R., Tarver, J. E., Fleming, J., Pisani, D., and 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(4):20160081.

O'Reilly, J. E., Puttick, M. N., Pisani, D., and Donoghue, P. C. J. (2018). Probabilistic methods surpass parsimony when assessing clade support in phylogenetic analyses of discrete morphological data. *Palaeontology*, 61(1):105–118.

Östermark, R. (2000). A flexible multicomputer algorithm for elementary matrix operations. *Computers & Operations Research*, 27(3):245–268.

Ostrom, J. H. (1970). Stratigraphy and paleontology of the Cloverly Formation (Lower Cretaceous) of the Bighorn Basin area, Wyoming and Montana. *Bulletin of the Peabody Museum of Natural History*, 35:1–234.

Pacheco, C., Müller, R. T., Langer, M., Pretto, F. A., Kerber, L., and Dias da Silva, S. (2019). *Gnathovorax cabreirai*: a new early dinosaur and the origin and initial radiation of predatory dinosaurs. *PeerJ*, 7:e7963.

Paijmans, J. L. A., Barnett, R., Gilbert, M. T. P., Zepeda-Mendoza, M. L., Reumer, J. W. F., de Vos, J., Zazula, G., Nagel, D., Baryshnikov, G. F., Leonard, J. A., Rohland, N., Westbury, M. V., Barlow, A., and Hofreiter, M. (2017). Evolutionary history of saber-toothed cats based on ancient mitogenomics. *Current Biology*, 27(21):3330–3336.

Pannetier, T., Martinez, C., Bunnefeld, L., and Etienne, R. S. (2021). Branching patterns in phylogenies cannot distinguish diversity-dependent diversification from time-dependent diversification. *Evolution*, 75(1):25–38.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2):289–290.

Paradis, E. and Schliep, K. (2018). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35(3):526–528.

Parins-Fukuchi, C. (2021). Morphological and phylogeographic evidence for budding speciation: an example in hominins. *Biology Letters*, 17(1):20200754.

Parry, L. A., Baron, M. G., and Vinther, J. (2017). Multiple optimality criteria support Ornithoscelida. *Royal Society Open Science*, 4(10):170833.

Paterson, J. R., Edgecombe, G. D., and Lee, M. S. Y. (2019). Trilobite evolutionary rates constrain the duration of the Cambrian explosion. *Proceedings of the National Academy of Sciences*, 116(10):4394–4399.

Patou, M.-L., Debruyne, R., Jennings, A. P., Zubaid, A., Rovie-Ryan, J. J., and Veron, G. (2008). Phylogenetic relationships of the Asian palm civets (Hemigalinae & Paradoxurinae, Viverridae, Carnivora). *Molecular Phylogenetics and Evolution*, 47(3):883–892.

Patou, M.-L., Mclenachan, P. A., Morley, C. G., Couloux, A., Jennings, A. P., and Veron, G. (2009). Molecular phylogeny of the Herpestidae (Mammalia, Carnivora) with a special emphasis on the Asian *Herpestes*. *Molecular Phylogenetics and Evolution*, 53:69–80.

Patou, M.-L., Wilting, A., Gaubert, P., Esselstyn, J. A., Cruaud, C., Jennings, A. P., Fickel, J., and Veron, G. (2010). Evolutionary history of the *Paradoxurus* palm civets – a new model for Asian biogeography. *Journal of Biogeography*, 37(11):2077–2097.

Paul, G. S. (1984). The segnosaurian dinosaurs: relics of the prosauropod–ornithischian transition? *Journal of Vertebrate Paleontology*, 4(4):507–515.

Paul, G. S. (1988). *Predatory Dinosaurs of the World: A Complete Illustrated Guide*. Simon and Schuster, New York.

Pease, J. B., Brown, J. W., Walker, J. F., Hinchliff, C. E., and Smith, S. A. (2018). Quartet Sampling distinguishes lack of support from conflicting support in the green plant tree of life. *American Journal of Botany*, 105(3):385–403.

Pedersen, T. L. (2022a). *ggforce: Accelerating 'ggplot2'*. R package version 0.4.1.

Pedersen, T. L. (2022b). *patchwork: The composer of plots*. R package version 1.1.2.

Perri, A. R., Mitchell, K. J., Mouton, A., Álvarez-Carretero, S., Hulme-Beaman, A., Haile, J., Jamieson, A., Meachen, J., Lin, A. T., Schubert, B. W., Ameen, C., Antipina, E. E., Bover, P., Brace, S., Carmagnini, A., Carøe, C., Samaniego Castruita, J. A., Chatters, J. C., Dobney, K., dos Reis, M., Evin, A., Gaubert, P., Gopalakrishnan, S., Gower, G., Heiniger, H., Helgen, K. M., Kapp, J., Kosintsev, P. A., Linderholm, A., Ozga, A. T., Presslee, S., Salis, A. T., Saremi, N. F., Shew, C., Skerry, K., Taranenko, D. E., Thompson, M., Sablin, M. V., Kuzmin, Y. V., Collins, M. J., Sinding, M.-H. S., Gilbert, M. T. P., Stone, A. C., Shapiro, B., Van Valkenburgh, B., Wayne, R. K., Larson, G., Cooper, A., and Frantz, L. A. F. (2021). Dire wolves were the last of an ancient New World canid lineage. *Nature*, 591(7848):87–91.

Peters, S. E. and Foote, M. (2001). Biodiversity in the Phanerozoic: a reinterpretation. *Paleobiology*, 27(4):583–601.

Peters, S. E. and Heim, N. A. (2010). The geological completeness of paleontological sampling in North America. *Paleobiology*, 36(1):61–79.

Peters, S. E. and McClennen, M. (2016). The Paleobiology Database application programming interface. *Paleobiology*, 42(1):1–7.

Pires, M. M., Rankin, B. D., Silvestro, D., and Quental, T. B. (2018). Diversification dynamics of mammalian clades during the K–Pg mass extinction. *Biology Letters*, 14(9):20180458.

Plotnick, R. E. and Wagner, P. J. (2018). The greatest hits of all time: the histories of dominant genera in the fossil record. *Paleobiology*, 44(3):368–384.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, 6:7–11.

Poe, S. and Chubb, A. L. (2004). Birds in a bush: five genes indicate explosive evolution of avian orders. *Evolution*, 58(2):404–415.

Poe, S. and Wiens, J. J. (2000). Character selection and the methodology of morphological phylogenetics. In Wiens, J. J., editor, *Phylogenetic Analysis of Morphological Data*, page 20–36. Smithsonian Institution Press, Washington, DC.

Pol, D., Rauhut, O. W. M., and Becerra, M. (2011). A Middle Jurassic heterodontosaurid dinosaur from Patagonia and the evolution of heterodontosaurids. *Naturwissenschaften*, 98(5):369–379.

Ponciano, J. M., Burleigh, J. G., Braun, E. L., and Taper, M. L. (2012). Assessing parameter identifiability in phylogenetic models using data cloning. *Systematic Biology*, 61(6):955–972.

Porto, D. S., Dahdul, W. M., Lapp, H., Balhoff, J. P., Vision, T. J., Mabee, P. M., and Uyeda, J. (2022). Assessing Bayesian phylogenetic information content of morphological data using knowledge from anatomy ontologies. *Systematic Biology*, 71(6):1290–1306.

Püschel, H. P., O'Reilly, J. E., Pisani, D., and Donoghue, P. C. J. (2020). The impact of fossil stratigraphic ranges on tip-calibration, and the accuracy and precision of divergence time estimates. *Palaeontology*, 63(1):67–83.

Puttick, M. N., O'Reilly, J. E., Pisani, D., and Donoghue, P. C. J. (2019). Probabilistic methods outperform parsimony in the phylogenetic analysis of data simulated without a probabilistic model. *Palaeontology*, 62(1):1–17.

Puttick, M. N., O'Reilly, J. E., Tanner, A. R., Fleming, J. F., Clark, J., Holloway, L., Lozano-Fernandez, J., Parry, L. A., Tarver, J. E., Pisani, D., and Donoghue, P. C. J. (2017). Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data. *Proceedings of the Royal Society B: Biological Sciences*, 284(1846):20162290.

Pybus, O. G. and Harvey, P. H. (2000). Testing macro-evolutionary models using incomplete molecular phylogenies. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 267(1459):2267–2272.

Quental, T. B. and Marshall, C. R. (2009). Extinction during evolutionary radiations: reconciling the fossil record with molecular phylogenies. *Evolution*, 63(12):3158–3167.

Quental, T. B. and Marshall, C. R. (2010). Diversity dynamics: molecular phylogenies need the fossil record. *Trends in Ecology & Evolution*, 25(8):434–441.

Quental, T. B. and Marshall, C. R. (2011). The molecular phylogenetic signature of clades in decline. *PLoS ONE*, 6(10):e25780.

Quental, T. B. and Marshall, C. R. (2013). How the Red Queen drives terrestrial mammals to extinction. *Science*, 341(6143):290–292.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabosky, D. L. (2006). Likelihood methods for detecting temporal shifts in diversification rates. *Evolution*, 60(6):1152–1164.

Rabosky, D. L. (2010). Extinction rates should not be estimated from molecular phylogenies. *Evolution*, 64(6):1816–1824.

Rabosky, D. L. (2014). Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PLoS ONE*, 9(2):e89543.

Rabosky, D. L. (2016). Challenges in the estimation of extinction from molecular phylogenies: a response to Beaulieu and O'Meara. *Evolution*, 70(1):218–228.

Rabosky, D. L. (2019). Phylogenies and diversification rates: variance cannot be ignored. *Systematic Biology*, 68(3):538–550.

Rabosky, D. L., Chang, J., Title, P. O., Cowman, P. F., Sallan, L., Friedman, M., Kaschner, K., Garilao, C., Near, T. J., Coll, M., and Alfaro, M. E. (2018). An inverse latitudinal gradient in speciation rate for marine fishes. *Nature*, 559(7714):392–395.

Rabosky, D. L., Grundler, M., Anderson, C., Title, P., Shi, J. J., Brown, J. W., Huang, H., and Larson, J. G. (2014). BAMMtools: an R package for the analysis of evolutionary dynamics on phylogenetic trees. *Methods in Ecology and Evolution*, 5(7):701–707.

Rabosky, D. L., Mitchell, J. S., and Chang, J. (2017). Is BAMM flawed? Theoretical and practical concerns in the analysis of multi-rate diversification models. *Systematic Biology*, 66(4):477–498.

Rabosky, D. L., Slater, G. J., and Alfaro, M. E. (2012). Clade age and species richness are decoupled across the eukaryotic tree of life. *PLoS Biology*, 10(8):e1001381.

Raia, P., Carotenuto, F., Mondanaro, A., Castiglione, S., Passaro, F., Saggese, F., Melchionna, M., Serio, C., Alessio, L., Silvestro, D., and Fortelius, M. (2016). Progress to extinction: increased specialisation causes the demise of animal clades. *Scientific Reports*, 6(1):30965.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., and Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology*, 67(5):901–904.

Rambaut, A., Lam, T. T., Max Carvalho, L., and Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution*, 2(1):vew007.

Rannala, B. (2002). Identifiability of parameters in MCMC Bayesian inference of phylogeny. *Systematic Biology*, 51(5):754–760.

Ranwez, V., Berry, V., Criscuolo, A., Fabre, P.-H., Guillemot, S., Scornavacca, C., and Douzery, E. J. P. (2007). PhySIC: a veto supertree method with desirable properties. *Systematic Biology*, 56(5):798–817.

Rao, H., Yang, Y., Liu, J., Westbury, M. V., Zhang, C., and Shao, Q. (2020). Palaeoproteomic analysis of Pleistocene cave hyenas from east Asia. *Scientific Reports*, 10(1):16674.

Raup, D. M. (1976). Species diversity in the Phanerozoic: an interpretation. *Paleobiology*, 2(4):289–297.

Raup, D. M. (1985). Mathematical models of cladogenesis. *Paleobiology*, 11(1):42–52.

Raup, D. M., Gould, S. J., Schopf, T. J. M., and Simberloff, D. S. (1973). Stochastic models of phylogeny and the evolution of diversity. *Journal of Geology*, 81(5):525–542.

Reddy, S., Kimball, R. T., Pandey, A., Hosner, P. A., Braun, M. J., Hackett, S. J., Han, K.-L., Harshman, J., Huddleston, C. J., Kingston, S., Marks, B. D., Miglia, K. J., Moore, W. S., Sheldon, F. H., Witt, C. C., Yuri, T., and Braun, E. L. (2017). Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. *Systematic Biology*, 66(5):857–879.

Ree, R. H. (2005). Detecting the historical signature of key innovations using stochastic models of character evolution and cladogenesis. *Evolution*, 59(2):257–265.

Ree, R. H. and Smith, S. A. (2008). Maximum likelihood inference of geographic range evolution by dispersal, local extinction, and cladogenesis. *Systematic Biology*, 57(1):4–14.

Regalado Fernández, O. R. and Werneburg, I. (2022). A new massopodan sauropodomorph from Trossingen Formation (Germany) hidden as '*Plateosaurus*' for 100 years in the historical Tübingen collection. *Vertebrate Zoology*, 72:771–822.

Reig, O. A. (1970). The Proterosuchia and the early evolution of the archosaurs; an essay about the origin of a major taxon. *Bulletin of the Museum of Comparative Zoology*, 139(5):229–292.

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2):217–223.

Rivadeneira, M. M. (2010). On the completeness and fidelity of the Quaternary bivalve fossil record from the temperate Pacific coast of South America. *Palaios*, 25(1):40–45.

Robinson, D. F. and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1):131–147.

Romer, A. S. (1956). *Osteology of the Reptiles*. University of Chicago Press, Chicago, IL.

Ronquist, F., Huelsenbeck, J. P., and Britton, T. (2004). Bayesian supertrees. In Bininda-Emonds, O. R. P., editor, *Phylogenetic Supertrees: Combining information to reveal the Tree of Life*, pages 193–224. Kluwer Academic Publishers, Dordrecht, NL.

Ronquist, F., Klopfstein, S., Vilhelmsen, L., Schulmeister, S., Murray, D. L., and Rasnitsyn, A. P. (2012a). A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. *Systematic Biology*, 61(6):973–999.

Ronquist, F., Lartillot, N., and 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(1699):20150136.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. (2012b). MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61(3):539–542.

Ross, H. A. and Rodrigo, A. G. (2004). An assessment of matrix representation with compatibility in supertree construction. In Bininda-Emonds, O. R. P., editor, *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pages 35–63. Kluwer Academic Publishers, Dordrecht, NL.

Rozadilla, S., Agnolín, F. L., and Novas, F. E. (2019). Osteology of the patagonian ornithopod *Talenkauen santacrucensis* (Dinosauria, Ornithischia). *Journal of Systematic Palaeontology*, 17(24):2043–2089.

Säilynoja, T., Bürkner, P.-C., and Vehtari, A. (2022). Graphical test for discrete uniformity and its applications in goodness-of-fit evaluation and multiple sample comparison. *Statistics and Computing*, 32(2):32.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425.

Sakamoto, M., Benton, M. J., and Venditti, C. (2016). Dinosaurs in decline tens of millions of years before their final extinction. *Proceedings of the National Academy of Sciences*, 113(18):5036–5040.

Sand, A., Holt, M. K., Johansen, J., Brodal, G. S., Mailund, T., and Pedersen, C. N. S. (2014). tqDist: a library for computing the quartet and triplet distances between binary or general trees. *Bioinformatics*, 30(14):2079–2080.

Sanderson, M. J., Purvis, A., and Henze, C. (1998). Phylogenetic supertrees: assembling the trees of life. *Trends in Ecology & Evolution*, 13(3):105–109.

Sayyari, E. and Mirarab, S. (2018). Testing for polytomies in phylogenetic species trees using quartet frequencies. *Genes*, 9(3):132.

Scheel, D.-M., Slater, G. J., Kolokotronis, S.-O., Potter, C. W., Rotstein, D. S., Tsangaras, K., Greenwood, A. D., and Helgen, K. M. (2014). Biogeography and taxonomy of extinct and endangered monk seals illuminated by ancient DNA and skull morphology. *ZooKeys*, 409:1–33.

Schliep, K. P. (2010). phangorn: phylogenetic analysis in R. *Bioinformatics*, 27(4):592–593.

Schluter, D. and Pennell, M. W. (2017). Speciation gradients and the distribution of biodiversity. *Nature*, 546(7656):48–55.

Seeley, H. G. (1887). I. On the classification of the fossil animals commonly named Dinosauria. *Proceedings of the Royal Society of London*, 43(258-265):165–171.

Sepkoski, J. J. (1975). Stratigraphic biases in the analysis of taxonomic survivorship. *Paleobiology*, 1(4):343–355.

Sepkoski, J. J. (1996). Patterns of Phanerozoic extinction: a perspective from global data bases. In Walliser, O. H., editor, *Global Events and Event Stratigraphy in the Phanerozoic*, pages 35–51. Springer, Berlin.

Sepkoski, J. J. (1998). Rates of speciation in the fossil record. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 353(1366):315–326.

Sereno, P. C. (1984). The phylogeny of the Ornithischia: a reappraisal. In Reif, W.-E. and Westphal, F., editors, *Third Symposium on Mesozoic Terrestrial Ecosystems, Short Papers*, page 219–226. Attempto Verlag, Tübingen.

Shen, X.-X., Hittinger, C. T., and Rokas, A. (2017). Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology & Evolution*, 1(5):0126.

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, 51(3):492–508.

Shimodaira, H. and Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution*, 16(8):1114–1114.

Silvestro, D., Cascales-Miñana, B., Bacon, C. D., and Antonelli, A. (2015). Revisiting the origin and diversification of vascular plants through a comprehensive Bayesian analysis of the fossil record. *New Phytologist*, 207(2):425–436.

Silvestro, D., Salamin, N., Antonelli, A., and Meyer, X. (2019). Improved estimation of macroevolutionary rates from fossil data using a Bayesian framework. *Paleobiology*, 45(4):546–570.

Silvestro, D., Salamin, N., and Schnitzler, J. (2014a). PyRate: a new program to estimate speciation and extinction rates from incomplete fossil data. *Methods in Ecology and Evolution*, 5(10):1126–1131.

Silvestro, D., Schnitzler, J., Liow, L. H., Antonelli, A., and Salamin, N. (2014b). Bayesian estimation of speciation and extinction from incomplete fossil occurrence data. *Systematic Biology*, 63(3):349–367.

Silvestro, D., Warnock, R. C. M., Gavryushkina, A., and Stadler, T. (2018). Closing the gap between palaeontological and neontological speciation and extinction rate estimates. *Nature Communications*, 9(1):5237.

Simões, T. R., Caldwell, M. W., and Pierce, S. E. (2020). Sphenodontian phylogeny and the impact of model choice in Bayesian morphological clock estimates of divergence times and evolutionary rates. *BMC Biology*, 18(1):191.

Simpson, G. G. (1944). *Tempo and mode in evolution*. Columbia University Press, New York, NY.

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(8):734–744.

Slater, G. J. (2014). Correction to 'Phylogenetic evidence for a shift in the mode of mammalian body size evolution at the Cretaceous–Palaeogene boundary', and a note on fitting macroevolutionary models to comparative paleontological data sets. *Methods in Ecology and Evolution*, 5(7):714–718.

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*, 112(16):4897–4902.

Slater, G. J. and Friscia, A. R. (2019). Hierarchy in adaptive radiation: a case study using the Carnivora (Mammalia). *Evolution*, 73(3):524–539.

Slater, G. J. and Harmon, L. J. (2013). Unifying fossils and phylogenies for comparative analyses of diversification and trait evolution. *Methods in Ecology and Evolution*, 4(8):699–702.

Slater, G. J., Harmon, L. J., and Alfaro, M. E. (2012a). Integrating fossils with molecular phylogenies improves inference of trait evolution. *Evolution*, 66(12):3931–3944.

Slater, G. J., Harmon, L. J., Wegmann, D., Joyce, P., Revell, L. J., and Alfaro, M. E. (2012b). Fitting models of continuous trait evolution to incompletely sampled comparative data using approximate Bayesian computation. *Evolution*, 66(3):752–762.

Smiley, T. M. (2018). Detecting diversification rates in relation to preservation and tectonic history from simulated fossil records. *Paleobiology*, 44(1):1–24.

Smith, K., Ayres, D., Neumaier, R., Worheide, G., and Hohna, S. (2024). Bayesian phylogenetic analysis on multi-core compute architectures: implementation and evaluation of BEAGLE in RevBayes with MPI. *Systematic Biology*. doi:10.1093/sysbio/syae005.

Smith, M. R. (2019a). Bayesian and parsimony approaches reconstruct informative trees from simulated morphological datasets. *Biology Letters*, 15(2):20180632.

Smith, M. R. (2019b). Quartet: comparison of phylogenetic trees using quartet and split measures. *Comprehensive R Archive Network*. doi:10.5281/zenodo.2536318.

Smith, M. R. (2019c). *TreeTools: Create, modify and analyse phylogenetic trees*. R package version 1.9.0.

Smith, M. R. (2020a). Information theoretic generalized Robinson–Foulds metrics for comparing phylogenetic trees. *Bioinformatics*, 36(20):5007–5013.

Smith, M. R. (2020b). TreeDist: distances between phylogenetic trees. *Comprehensive R Archive Network*. doi:10.5281/zenodo.3528123.

Smith, S. A. and O'Meara, B. C. (2012). treePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics*, 28(20):2689–2690.

Solow, A. R. and Smith, W. (1997). On fossil preservation and the stratigraphic ranges of taxa. *Paleobiology*, 23(3):271–277.

Soul, L. C. and Wright, D. F. (2021). *Phylogenetic comparative methods: a user's guide for paleontologists*. Cambridge University Press, Cambridge, UK.

Stadler, T. (2010). Sampling-through-time in birth–death trees. *Journal of Theoretical Biology*, 267:396–404.

Stadler, T. (2011a). Mammalian phylogeny reveals recent diversification rate shifts. *Proceedings of the National Academy of Sciences*, 108(15):6187–6192.

Stadler, T. (2011b). Simulating trees with a fixed number of extant species. *Systematic Biology*, 60(5):676–684.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

Stanley, S. M. (2007). An analysis of the history of marine animal diversity. *Paleobiology*, 33(S4):1–55.

Starrfelt, J. and Liow, L. H. (2016). How many dinosaur species were there? Fossil bias and true richness estimated using a Poisson sampling model. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1691):20150219.

Steel, M. A. and Penny, D. (1993). Distributions of tree comparison metrics—some new results. *Systematic Biology*, 42(2):126–141.

Steel, M. A. and Rodrigo, A. G. (2008). Maximum likelihood supertrees. *Systematic Biology*, 57(2):243–250.

Sternberg, C. M. (1950). *Notes and annotated list of quarries. Map 969A. Steveville, west of fourth meridian*. Canada Department of Mines and Technical Surveys, Geological Survey of Canada, Ottawa.

Strimmer, K. and Rambaut, A. (2002). Inferring confidence sets of possibly misspecified gene trees. *Proceedings of the Royal Society B: Biological Sciences*, 269(1487):137–142.

Suchard, M. A. and Rambaut, A. (2009). Many-core algorithms for statistical phylogenetics. *Bioinformatics*, 25(11):1370–1376.

Suchard, M. A., Weiss, R. E., and Sinsheimer, J. S. (2005). Models for estimating Bayes factors with applications to phylogeny and tests of monophyly. *Biometrics*, 61(3):665–673.

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., and Gelman, A. (2020). Validating Bayesian inference algorithms with simulation-based calibration. *arχiv*. 1804.06788v2.

Tarasov, S. (2023). New phylogenetic Markov models for inapplicable morphological characters. *Systematic Biology*, 72(3):681–693.

Tennant, J. P., Chiarenza, A. A., and Baron, M. (2018). How has our knowledge of dinosaur diversity through geologic time changed through research history? *PeerJ*, 6:e4417.

Thulborn, R. A. (1975). Dinosaur polyphyly and the classification of archosaurs and birds. *Australian Journal of Zoology*, 23(2):249–270.

Tomiya, S. (2011). A new basal caniform (Mammalia: Carnivora) from the middle Eocene of North America and remarks on the phylogeny of early carnivorans. *PLoS ONE*, 6(9):e24146.

Tonini, J. F. R., Beard, K. H., Barbosa Ferreira, R., Jetz, W., and Pyron, R. A. (2016). Fully-sampled phylogenies of squamates reveal evolutionary patterns in threat status. *Biological Conservation*, 204:23 – 31.

Trindade, F. J., Rodrigues, M. R., Figueiró, H. V., Li, G., Murphy, W. J., and Eizirik, E. (2021). Genome-wide SNPs clarify a complex radiation and support recognition of an additional cat species. *Molecular Biology and Evolution*, 38(11):4987–4991.

Tseng, Z. J., Wang, X., Slater, G. J., Takeuchi, G. T., Li, Q., Liu, J., and Xie, G. (2014). Himalayan fossils of the oldest known pantherine establish ancient origin of big cats. *Proceedings of the Royal Society B: Biological Sciences*, 281(1774):20132686.

Tsutsumi, M., Saito, N., Koyabu, D., and Furusawa, C. (2023). A deep learning approach for morphological feature extraction based on variational auto-encoder: an application to mandible shape. *npj Systems Biology and Applications*, 9(1):30.

Turner, A. H., Pritchard, A. C., and Matzke, N. J. (2017). Empirical and Bayesian approaches to fossil-only divergence times: A study across three reptile clades. *PLOS ONE*, 12(2):e0169885.

Upham, N. S., Esselstyn, J. A., and Jetz, W. (2019). Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLOS Biology*, 17(12):e3000494.

Upham, N. S., Esselstyn, J. A., and Jetz, W. (2021). Molecules and fossils tell distinct yet complementary stories of mammal diversification. *Current Biology*, 31(19):4195–4206.e3.

Uyeda, J. C. and Harmon, L. J. (2014). A novel Bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data. *Systematic Biology*, 63(6):902–918.

Varela, S., González-Hernández, J., Sgarbi, L. F., Marshall, C., Uhen, M. D., Peters, S., and McClennen, M. (2015). paleobioDB: an R package for downloading, visualizing and processing data from the Paleobiology Database. *Ecography*, 38(4):419–425.

Vavrek, M. J. and Larsson, H. C. E. (2010). Low beta diversity of Maastrichtian dinosaurs of North America. *Proceedings of the National Academy of Sciences*, 107(18):8265–8268.

Velasco, J. D. (2008). The prior probabilities of phylogenetic trees. *Biology & Philosophy*, 23(4):455–473.

Verbruggen, H., Maggs, C. A., Saunders, G. W., Le Gall, L., Yoon, H. S., and De Clerck, O. (2010). Data mining approach identifies research priorities and data requirements for resolving the red algal tree of life. *BMC Evolutionary Biology*, 10(1):16.

Wagner, P. J. (1998). Phylogenetic analyses and the quality of the fossil record. In Donovan, S. K. and Paul, C. R. C., editors, *The Adequacy of the Fossil Record*, pages 165–187. John Wiley & Sons, Chichester, UK.

Wagner, P. J. (2019). On the probabilities of branch durations and stratigraphic gaps in phylogenies of fossil taxa when rates of diversification and sampling vary over time. *Paleobiology*, 45(1):30–55.

Wagner, P. J. and Marcot, J. D. (2013). Modelling distributions of fossil sampling rates over time, space and taxa: assessment and implications for macroevolutionary studies. *Methods in Ecology and Evolution*, 4(8):703–713.

Wagner, P. J., Plotnick, R. E., and Lyons, S. K. (2018). Evidence for trait-based dominance in occupancy among fossil taxa and the decoupling of macroecological and macroevolutionary success. *American Naturalist*, 192(3):E120–E138.

Wang, S. C. and Dodson, P. (2006). Estimating the diversity of dinosaurs. *Proceedings of the National Academy of Sciences*, 103(37):13601–13605.

Warnock, R. C. M., Heath, T. A., and Stadler, T. (2020). Assessing the impact of incomplete species sampling on estimates of speciation and extinction rates. *Paleobiology*, 46(2):137–157.

Warren, D. L., Geneva, A. J., and Lanfear, R. (2017). RWTY (R We There Yet): An R package for examining convergence of Bayesian phylogenetic analyses. *Molecular Biology and Evolution*, 34(4):1016–1020.

Weaver, W. N. and Smith, S. A. (2023). From leaves to labels: Building modular machine learning networks for rapid herbarium specimen analysis with LeafMachine2. *Applications in Plant Sciences*, 11(5):e11548.

Westbury, M. V., Le Duc, D., Duchêne, D. A., Krishnan, A., Prost, S., Rutschmann, S., Grau, J. H., Dalén, L., Weyrich, A., Norén, K., Werdelin, L., Dalerum, F., Schöneberg, T.,

and Hofreiter, M. (2021). Ecological specialization and evolutionary reticulation in extant Hyaenidae. *Molecular Biology and Evolution*, 38(9):3884–3897.

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C. O., Woo, K., Yutani, H., and Dunnington, D. (2022a). *ggplot2: Create elegant data visualisations using the grammar of graphics*. R package version 3.4.0.

Wickham, H., François, R., Henry, L., Müller, K., and Vaughan, D. (2022b). *dplyr: A grammar of data manipulation*. R package version 1.0.10.

Wickham, H. and Seidel, D. (2022). *scales: Scale functions for visualization*. R package version 1.2.1.

Wilke, C. O. (2020). *cowplot: Streamlined plot theme and plot annotations for 'ggplot2'*. R package version 1.1.1.

Wilkinson, M., Cotton, J. A., Creevey, C., Eulenstein, O., Harris, S. R., Lapointe, F.-J., Levasseur, C., Mcinerney, J. O., Pisani, D., and Thorley, J. L. (2005). The shape of supertrees to come: tree shape related properties of fourteen supertree methods. *Systematic Biology*, 54(3):419–431.

Wilkinson, M., Cotton, J. A., Lapointe, F.-J., and Pisani, D. (2007). Properties of supertree methods in the consensus setting. *Systematic Biology*, 56(2):330–337.

Willson, S. J. (2004). Constructing rooted supertrees using distances. *Bulletin of Mathematical Biology*, 66(6):1755–1783.

Willson, S. J. (2005). Minimum evolution using ordinary least-squares is less robust than neighbor-joining. *Bulletin of Mathematical Biology*, 67(2):261–279.

Wisniewski, A. L., Lloyd, G. T., and Slater, G. J. (2022). Extant species fail to estimate ancestral geographical ranges at older nodes in primate phylogeny. *Proceedings of the Royal Society B: Biological Sciences*, 289(1975):20212535.

Wolsan, M. and Sato, J. J. (2020). Parallel loss of sweet and umami taste receptor function from phocids and otarioids suggests multiple colonizations of the marine realm by pinnipeds. *Journal of Biogeography*, 47(1):235–249.

Wright, A. M., Bapst, D. W., Barido-Sottani, J., and Warnock, R. C. M. (2022). Integrating fossil observations into phylogenetics using the fossilized birth–death model. *Annual Review of Ecology, Evolution, and Systematics*, 53:251–273.

Wright, A. M. and Hillis, D. M. (2014). Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. *PLoS ONE*, 9(10):e109210.

Wu, R., Pisani, D., and Donoghue, P. C. J. (2023). The unbearable uncertainty of panarthropod relationships. *Biology Letters*, 19(1):20220497.

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., and Chen, M.-H. (2010). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic Biology*, 60(2):150–160.

Xu, X., Forster, C. A., Clark, J. M., and Mo, J. (2006). A basal ceratopsian with transitional features from the Late Jurassic of northwestern China. *Proceedings of the Royal Society B: Biological Sciences*, 273(1598):2135–2140.

Yang, Y., Wu, W., Dieudonné, P.-E., and Godefroit, P. (2020). A new basal ornithopod dinosaur from the Lower Cretaceous of China. *PeerJ*, 8:e9832.

Yao, Y. and Domke, J. (2023). Discriminative calibration: Check Bayesian computation from simulations and flexible classifier. *arχiv*. 2305.14593v2.

Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T.-Y. (2017). GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.*, 8(1):28–36.

Zhang, C., Stadler, T., Klopfstein, S., Heath, T. A., and Ronquist, F. (2016). Total-evidence dating under the fossilized birth–death process. *Systematic Biology*, 65(2):228–249.