THE IMPACT OF FOSSIL STRATIGRAPHIC RANGES ON TIP-CALIBRATION, AND THE ACCURACY AND PRECISION OF DIVERGENCE TIME ESTIMATES

by HANS P. PÜSCHEL1,2, JOSEPH E. O’REILLY1,3, DAVIDE PISANI1,4 and PHILIP C. J. DONOGHUE1

1School of Earth Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol, BS8 1TQ, UK; phil.donoghue@bristol.ac.uk
2School of Geosciences Grant Institute, University of Edinburgh, The King’s Buildings, James Hutton Road, Edinburgh, EH9 3FE, Scotland, UK; s1874123@sms.ed.ac.uk
3MRC Human Genetics Unit, MRC Institute of Genetics & Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, Scotland, UK; joseph.oreilly@igmm.ed.ac.uk
4School of Biological Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol, BS8 1TQ, UK; davide.pisani@bristol.ac.uk

Typescript received 12 December 2018; accepted in revised form 9 May 2019

Abstract: The molecular clock provides the only viable means of establishing realistic evolutionary timescales but it remains unclear how best to calibrate divergence time analyses. Calibrations can be applied to the tips and/or to the nodes of a phylogeny. Tip-calibration is an attractive approach since it allows fossil species to be included alongside extant relatives in molecular clock analyses. However, most fossil species are known from multiple stratigraphical horizons and it remains unclear how such age ranges should be interpreted to codify tip-calibrations. We use simulations and empirical data to explore the impact on precision and accuracy of different approaches to informing tip-calibrations. In particular, we focus on the effect of using tip-calibrations defined using the oldest vs youngest stratigraphic occurrences, the full stratigraphical range, as well as confidence intervals on these data points. The results of our simulations show that using different calibration approaches leads to different divergence-time estimates and demonstrate that concentrating tip-calibrations near the root of the dated phylogeny improves both precision and accuracy of estimated divergence times. Finally, our results indicate that the highest levels of accuracy and precision are achieved when fossil tips are calibrated based on the fossil occurrence from which the morphological data were derived. These trends were corroborated by analysis of an empirical dataset for Ursidae. Overall, we conclude that tip-dating analyses should, in particular, employ tip calibrations close to the root of the tree and they should be calibrated based on the age of the fossil used to inform the morphological data used in Total Evidence Dating.

Key words: molecular clock, simulation, fossil, divergence time, tip-calibration, Ursidae.

Calibrating the tree of life to geological time is a core aim of biology, facilitating temporal tests of evolutionary hypotheses, inference of evolutionary rates and patterns, as well as an understanding of the coevolution of life and the environment. The molecular clock (Zuckerkandl & Pauling 1965) affords the only viable means of establishing an evolutionary timescale, a prospect that has become increasingly tangible with the development of Bayesian relaxed clock methods (Donoghue & Yang 2016; dos Reis et al. 2016; Kumar & Hedges 2016; Bromham et al. 2017). A key aspect of the Bayesian divergence time estimation framework is the use of priors on divergence times which include information on the age of calibrations that is usually provided by fossil evidence. Traditionally, molecular clocks have been node-calibrated, whereby clade ages are constrained minimally by their oldest fossil member; clade age maxima are established based on a diversity of approaches ranging from statistical analysis of occurrence data, through heuristic analysis of a variety of evidence, to arbitrary probabilities that express some visceral perception of the true time of divergence (Donoghue & Benton 2007; Parham et al. 2012). Node calibrations are transformed into a joint time prior for the tree which may not always reflect the original node-calibrations (Inoue et al. 2010; Warnock et al. 2012, 2015). Many view node-calibrations as unsatisfactory because of this transformation of the original fossil evidence which is affected by prior assumptions on the phylogenetic affinity of the calibrating fossils, makes no use of fossil age information in topology estimation, strongly limits the number and nature of fossil species that can be used to calibrate the clock, and because of the diverse and ad hoc approaches that have been proposed to formulating probabilistic calibrations (Gavryushkina et al. 2017).
Tip-calibration has been introduced specifically to overcome the limitations of node-calibration (Pyron 2011; Ronquist et al. 2012a), affording a more direct approach to calibration, where fossils are included analytically en par with their living relatives, avoiding the need for the definition of maximum and minimum constraints on clade ages in the divergence-time analysis (Ronquist et al. 2012a). Tip-calibration is achieved by supplementing the molecular sequence data from living species with morphological data from both living and fossil species. These molecular–morphological datasets are then analysed using a partitioning scheme allowing the concurrent application of molecular and morphological models to the data. The calibration information is provided by the age information associated with the fossil taxa which disambiguates the morphological distances in terms of their otherwise confounded evolutionary rates and absolute times, that spread for the rest of the internal nodes of the tree characterized by molecular data (Barba-Montoya et al. 2017). Tip-calibration requires no prior assumptions of phylogenetic affinity and so it allows fossils to be more readily used in the co-estimation of divergence-times and tree topology and, at least theoretically, there are no restrictions on the number of fossil species that can be included in an analysis. Therefore, this approach allows the inclusion of older and sometimes fragmentary fossils typically excluded in node-calibration analyses because of their uncertain phylogenetic placement (Ronquist et al. 2012a; O’Reilly et al. 2015).

Though tip-calibration overcomes some of the drawbacks of node-calibration, it presents a series of new concerns, including systematic biases in the preservation of morphological characters in the fossil record (Sansom & Wills 2013), the efficacy of the Mk model of morphological character evolution, as well as its extreme sensitivity to the prior on divergence times or the branching model used (O’Reilly et al. 2015; Donoghue & Yang 2016; dos Reis et al. 2016). Regarding this last aspect, it is striking that tip-calibration was initially promoted on the promise that it would deliver more accurate and precise divergence time estimates less sensitive to prior assumptions than node-calibration (Ronquist et al. 2012a) since the opposite has been shown to be the case in most tip-calibrated divergence time studies (Wood et al. 2013; Arcila et al. 2015; O’Reilly et al. 2015; Matzke & Wright 2016; Puttick et al. 2016; Ronquist et al. 2016). Moreover, divergence time studies employing tip-calibration have displayed a consistent tendency towards estimating older divergence-times than node-calibrated analyses (Ronquist et al. 2012a, 2016; Schrago et al. 2013; Slater 2013, 2015; Wood et al. 2013; Beck & Lee 2014; Sharma & Giribet 2014; Tseng et al. 2014; Arcila et al. 2015; Dornburg et al. 2015; O’Reilly et al. 2015; Winterton & Ware 2015; Bapst et al. 2016; Matzke & Wright 2016; Puttick et al. 2016; Saladin et al. 2017). This latter phenomenon has been termed deep root attraction (Ronquist et al. 2016).

One possible explanation for the unrealistically ancient estimates from tip-calibrated divergence time analyses lies with the unresolved and little explored manner in which tip-calibrations are formulated (O’Reilly et al. 2015; Donoghue & Yang 2016). Most studies that have employed tip-calibration have assumed errorless point ages for the calibrating fossils based on a single age sample from the uncertainty associated with their geological age (Ronquist et al. 2012a; Schrago et al. 2013; Lee et al. 2014; Sharma & Giribet 2014; Arcila et al. 2015), if at all. This is surprising since it does not consider the stratigraphic uncertainty of a fossil age which is usually defined in terms of a minimum–maximum age interval (O’Reilly et al. 2015). Other studies have modelled tip-calibration age uncertainty as a uniform distribution between the oldest and youngest fossil stratigraphic occurrence (Wood et al. 2013; Dornburg et al. 2015; Marx & Fordyce 2015; Heritage et al. 2016; Lee 2016; Puttick et al. 2016; Sallam & Seiffert 2016; Vea & Grimaldi 2016; Zhang et al. 2016; Borths & Stevens 2017; Gavryushkina et al. 2017; Harrington & Reeder 2017; Seiffert et al. 2017; Turner et al. 2017; Vinther et al. 2017; Wood 2017; Wright & Toom 2017). Nevertheless, fossil species that are distributed across stratigraphical intervals imply morphological stasis which in itself informs the rate of evolution. Marx & Fordyce (2015) asserted that in cases of multiple fossil occurrences of markedly different ages, the oldest fossil occurrences should be considered exclusively and all other data should be discarded. Thus, it is not at all clear which of these approaches is best to implement tip-calibrations, let alone whether choice among these models has an impact on divergence time estimation.

In an attempt to explore the impact that different approaches to formulating tip-calibration has on divergence time estimation we took a simulation approach in which the performance of competing approaches can be assessed relative to a reference generating tree. This approach overcomes the impossible challenge to reconciling competing methodological approaches when analysing empirical data where the true timescale is unknowable (Bromham 2019). Using simulations we were able to test the credibility of the estimated diverge-times (Bromham et al. 2017) and to directly evaluate both accuracy and precision against known ages (Warnock et al. 2017). This was supplemented by analyses of an empirical dataset of Ursidae. We considered five different approaches to informing tip-calibrations and assessed their performance in terms of the absolute precision and accuracy of divergence-time estimates, as well as their coverage probability. Fossil occurrences were interpreted to inform tip-calibrations based on: (1) oldest or (2) youngest stratigraphical occurrences; (3) a uniform probability spanning
their full stratigraphical range, or a uniform probability spanning their full stratigraphical range plus a 95% confidence interval added to their (4) oldest or (5) youngest stratigraphical occurrences, based on their number of intervening fossiliferous horizons (Marshall 1990).

MATERIAL AND METHOD

Tree generation

An 18-tip tree with a Colless Index (Ic) of 0.5 was generated using the R packages ape v5.1 (Paradis et al. 2004), geiger v2.0.6 (Harmon et al. 2008) and apTreeshape v1.5.0 (Bortolussi et al. 2006). The Ic measures the asymmetry of the tree topology, ranging from zero to one with a value of zero representing a fully balanced tree and a value of one a fully imbalanced tree (Colless 1982). The branch lengths of this topology were manually modified to generate two different trees: a tree with fossil lineages nested along the stem and a tree with fossil lineages within the crown group. In both cases, six fossil taxa and 12 extant taxa were used (Fig. 1). The branch lengths were scaled so that the height of the tree was 250 myr, which was selected due to the 250 myr periodicity of a Phanerozoic Wilson cycle.

Simulation of fossil occurrences and character data generation

The R package phyclust v0.1.22 (Chen 2011) was used to simulate 1000 nucleotide sequences for each tree with the HKY nucleotide substitution model. Site-specific rate heterogeneity was modelled with a gamma distribution with $\alpha = \beta = 0.25$, the transition/transversion ratio $\kappa$ was set as 0.75 and the relative state frequencies were sampled from a Dirichlet distribution with $\alpha = (1,1,1,1)$. A strict clock with a rate of 0.0025 substitutions per site per million years was applied. For fossil lineages, simulated molecular information was entirely replaced with missing data. Fossil occurrence data was simulated with the R package Fossilsim (Barido-Sottani et al. 2019) using a non-uniform model across 125 stratigraphic intervals using $\text{PA} = 1$, $\text{PD} = 1$ and $\text{DT} = 1.5$ as values for the parameters of the model (Holland 1995). Through rejection sampling, each fossil lineage was enforced to possess at least two fossil occurrences and branches were trimmed such that they did not extend beyond the oldest fossil occurrence; since morphospecies exhibit morphological consistency it would have been inappropriate to create a second set of trees in which branches were trimmed to the youngest fossil occurrence. The trimmed trees were used to simulate morphological data using a method which matched the Mk model of morphological evolution (Lewis 2001). To achieve this, 1000 nucleotide sequences were simulated for each tree with a JC nucleotide substitution model. The site-specific rate heterogeneity and the clock rate kept the same values as in the molecular simulations. The nucleotides were then recoded as purines and pyrimidines, resulting in binary morphology-like character data (O’Reilly et al. 2016). Constant morphological characters were not assessed or removed.

Five separate tip-calibration strategies were then applied to the simulated data: (1) a point calibration representing the oldest occurrence of a lineage (Fig. 2A); (2) a point calibration representing the youngest occurrence of a lineage (Fig. 2B); (3) a uniform distribution spanning the oldest and youngest occurrences (Fig. 2C); (4) a uniform distribution spanning the oldest and youngest occurrences plus a 95% confidence interval added to the maximum age as described in Marshall (1990); and (5) a uniform distribution spanning the oldest and youngest occurrences plus the same confidence interval added to the minimum age (Fig. 2E). The span of this confidence interval depends on the number of fossil horizons and the width of the stratigraphical interval between the oldest and youngest occurrences with the only assumptions that these horizons are randomly distributed and uniformly collected over the whole stratigraphical range (Marshall 1990). For both trees this methodology was replicated 100 times for each calibration approach ($R = 100$).

Tip-calibration analyses of simulated data

The simulated datasets were analysed in MrBayes v3.2.2 (Ronquist et al. 2012b). The generating model (HKY + $\Gamma$) was applied to analyse the molecular data, the Mk + $\Gamma$ model was used to analyse morphological data (Lewis 2001), and ascertainment bias was not accounted for in the model. The topology was fixed to that of the generating tree and a strict uniform clock with an exponential prior with mean 0.025 was applied to model the rate. An exponential prior was applied to the tree height, with an offset 10 myr less than the minimum age of 250 Ma and a mean of 312.2 Ma. For each analysis, two independent runs were performed using four chains and 1 000 000 MCMC generations. The sampling frequency was every 200 generations, and the initial 25% of samples were discarded as burn-in. The accuracy of divergence-time estimates was determined using both the error of the estimated ages, and coverage measured as the proportion of 95% HPDs that included the true fossil age. In addition, the precision of divergence-time estimates was assessed in terms of accuracy (difference between the mean of the posterior distribution and the true age) and precision (the relative interval width which is the width of the 95% HPD range).
Tip-calibration of empirical data

For the empirical tip-calibration experiments, a molecular and morphological dataset of living bears (Ursidae) and extant relatives was analysed. This clade was selected because of the availability of molecular, morphological and fossil occurrence data (Montoya et al. 2001; Baryshnikov 2002; Jin et al. 2007; Abella et al. 2012; Heath et al.)
The outgroup was the grey wolf (Canis lupus), and its molecular information was obtained from GenBank (JX013645.1, AY525044.1). The morphological information from fossil and living bears was obtained from Heath et al. (2014) and the fossil stratigraphical occurrence data were obtained from the Paleobiology Database (PBDB; https://paleobiodb.org; Püschel et al. 2019, appendix 5). Since singletons are invariant to the calibration strategies we evaluate, only extinct taxa with at least two fossil occurrences were considered. The Mk+Γ model (Lewis 2001) was used for morphological data, and the GTR+I+Γ model was used for molecular data. Branch rate variation was modelled with the independent gamma rate (IGR) relaxed-clock model (Lepage et al. 2007). The clock rate prior was lognormally distributed, with a mean of 4 and standard deviation of 0.1, the prior for the variance was an exponential distribution with rate 126.887, similar to that used for the Canidae phylogeny (Matzke & Wright 2016). The choice of priors was mainly based on those used in Krause et al. (2008) and their observation of an approximately clock-like rate. The root of the tree was calibrated using an exponential distribution, offset with the age of Hesperocyon gregarious with a minimal value of 37.2 Ma and a mean of 38.6 Ma, which informs the fossil node calibration for crown Carnivora (Benton et al. 2015).

The topology was fixed to be consistent with the molecular phylogenies of Krause et al. (2008) and Heath et al. (2014), as well as with the morphological phylogeny of fossil taxa from (Abella et al. 2012). Constraining the analyses in this way resulted in the estimation of divergence times alone, allowing quantification of the effect of the different tip calibration approaches on these age estimates. Each analysis consisted of two runs of four chains and ten million MCMC generations with the first 25% of samples discarded. The program Tracer v1.7.1 (Rambaut et al. 2018) was used to assess whether the two independent runs achieved convergence and stationarity.

**RESULTS**

**Simulation analyses**

*Effect of calibration distance from the root.* The divergence-time estimates in the simulation analyses tended to recover node ages close to the generating ages (especially for nodes close to the tips) if the posterior age estimates are interpreted as the median or the mean of the posterior age distribution (Fig. 3). However, accuracy and precision showed a tendency to decrease under all calibration approaches in the trees where the tip-calibrations were shallow (topologically remote from the root), in comparison to the tree in which the tip-calibrations were positioned closer to the root showing a minor but consistently

![Diagram](https://example.com/diagram.png)
**FIG. 3.** Boxplots of divergence-time estimates of the nodes of the tree under different calibration approaches over 100 replicates. The two columns separate the tip-calibrations using fossils in the stem group (A–E) from those using fossils in the crown group (F–J); the rows show the five tip-calibration approaches described in Figure 2A–E for each column. Triangles represent the true ages for each node and the dots the mean of the divergence-time estimates. Outliers were removed for clarity. CI, confidence interval. Colour online.
higher error (Fig. 3; Püschel et al. 2019, appendices 1–3). However, the percent error showed an opposite trend (Fig. 4; Püschel et al. 2019, appendices 2, 3) and although the absolute error was lower in the most recent nodes of both trees (Fig. 3; Püschel et al. 2019, appendices 1–3), the percent error showed the opposite trend with lower error in the deepest nodes (Fig. 4; Püschel et al. 2019, appendices 2, 3).

**Effect of calibration approach.** The most accurate and precise approach used the oldest stratigraphical occurrence of a lineage to inform point age tip-calibrations (Fig. 3A, F; Püschel et al. 2019, appendices 1–3). This result holds irrespective of whether the tip-calibrations are located topologically close to, or remote from, the root. In contrast, using the youngest occurrence of a lineage recovered the most inaccurate results, at least 6 myr older than the estimates obtained when using other interpretations of stratigraphic range data to inform the age of tip-calibrations (Fig. 3B, G; Püschel et al. 2019, appendices 1–3). However, considering only the percent error, the most accurate approach is using a uniform distribution spanning the oldest and youngest occurrences (Fig. 4C, H; Püschel et al. 2019, appendices 2, 3).

**Coverage and width of 95% HPD intervals.** When accuracy was measured using the 95% HPD coverage (i.e. the generating age is encompassed by the 95% HPD), there are no significant differences between estimates derived from competing approaches to tip-calibration, or the topology of tip-calibrations approaches. Almost all approaches yield estimates that approximate 100% coverage, the exception being the youngest occurrence of a lineage in the tree with calibrations near to the root, which performed worst (Püschel et al. 2019, appendices 2–4). However, the span of the 95% HPD interval differed between tip-calibration approaches and topologies (Fig. 5; Püschel et al. 2019, appendices 2, 3). The tree with the most precise timescale (narrowest 95% HPDs on node ages) had the tip-calibrations located topologically close to the root, informed by the oldest stratigraphical occurrences (Fig. 5A). The tree with the least precise divergence times (broadest 95% HPDs on node ages) had tip-calibrations located remote from the root and informed by a uniform distribution spanning the oldest and youngest occurrences plus a 95% confidence interval added to the maximum age (Fig. 5I); this was followed very closely by tip calibrations informed by the youngest stratigraphical occurrences (Fig. 5G).

**Empirical analyses.** The empirical analyses of the bear (Ursidae) dataset recovered clade age estimates that are generally more consistent than those seen in the simulations (Figs 6, 7). The most precise method used the youngest fossil occurrences as errorless tip calibrations, and tip calibrations based on the oldest occurrences recovered the oldest clade age estimates. However, these differences become more noticeable the closer the position of the node relative to the root. Nodes within crown-Ursidae showed only minor differences, especially the most recent nodes of subfamily Ursinae (Fig. 7; Table 1).

**DISCUSSION**

**Simulation analyses**

No evidence of deep root attraction. The simulation results (Fig. 3; Püschel et al. 2019, appendices 1–3) do not support previous inferences of a deep root attraction phenomenon and the tendency of tip-calibrated analyses to recover inaccurately old divergence-time estimates (Ronquist et al. 2012a, 2016; Schrago et al. 2013; Slater 2013, 2015; Wood et al. 2013; Beck & Lee 2014; Sharma & Giribet 2014; Tseng et al. 2014; Arcila et al. 2015; Dornburg et al. 2015; O’Reilly et al. 2015; Winterton & Ware 2015; Bapst et al. 2016; Matzke & Wright 2016; Puttick et al. 2016; Saladin et al. 2017). Even if all the calibration approaches tested here showed different degrees of error, this error did not exhibit a general trend towards older divergence-time estimates, with the possible exception of using the oldest stratigraphical occurrences for calibration; these exhibited a small but consistent overestimation of the true ages in all nodes regardless of the position of the tip-calibrations (Fig. 3A, F; Püschel et al. 2019, appendices 1–3). In other words, the results of our simulation experiments suggest that the tendency of tip-calibrated analyses to recover erroneously ancient clade age estimates is not an intrinsic property of this approach. The deep root attraction phenomenon observed in previous studies could be due to inappropriate priors in the model and the inability of the only currently available morphological model (the Mk model) to include correlations or dependencies between morphological traits (Ronquist et al. 2016). As our morphological data was simulated using the Mk model, there is no mismatch between the data and the model used for analysing the data, so logically it should not present the problems associated with analysing empirical data. In that sense, our results are compatible with the notion that deep root attraction is a consequence of an inadequate model of morphological evolution.

**Optimizing results.** The best results are achieved when mismatch of the simulation model and the specified model is minimized. Considering different measurements of accuracy and precision, tips calibrated from the oldest occurrences...
FIG. 4. Boxplots of the percent error in divergence-time estimates of the nodes of the tree under different calibration approaches over 100 replicates. The two columns separate the tip-calibrations using fossils in the stem group (A–E) from those using fossils in the crown group (F–J); the rows show the five tip-calibration approaches described in Figure 2A–E for each column. Outliers were removed for clarity. CI, confidence interval.
FIG. 5. Boxplots of the 95% HPD widths of each node of the tree under different calibration approaches over 100 replicates. The two columns separate the tip-calibrations using fossils in the stem group (A–E) from those using fossils in the crown group (F–J); the rows show the five tip-calibration approaches described in Figure 2A–E for each column. Outliers were removed for clarity. CI, confidence interval; HPD, highest posterior density.
FIG. 6. Dated phylogenies of bears (Ursidae) under five different calibration approaches: A, maximum; B, minimum; C, minimum–maximum; D, minimum–maximum 95% confidence interval in maximum; E, minimum–maximum 95% confidence interval in minimum. More details of the calibration approaches are described in Figure 2 in the same order. Node bars represent the 95% highest posterior density (HPD) for the estimated node ages. Colour online.
TABLE 1. Posterior estimates of divergence times and 95% HPD widths of the Ursidae phylogeny in the five calibrations approaches described in Figure 2.

<table>
<thead>
<tr>
<th>Calibration</th>
<th>Maximum</th>
<th>Minimum</th>
<th>Minimum–Max</th>
<th>Min–Max 95% CI Max</th>
<th>Min–Max 95% CI Min</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>95% HPDw</td>
<td>Mean</td>
<td>95% HPDw</td>
<td>Mean</td>
</tr>
<tr>
<td>Node 1</td>
<td>40.35</td>
<td>10.07</td>
<td>38.71</td>
<td>5.87</td>
<td>38.89</td>
</tr>
<tr>
<td>Node 2</td>
<td>24.9</td>
<td>9.34</td>
<td>14.24</td>
<td>2.84</td>
<td>16.5</td>
</tr>
<tr>
<td>Node 3</td>
<td>22.08</td>
<td>4.63</td>
<td>10.24</td>
<td>1.73</td>
<td>13.55</td>
</tr>
<tr>
<td>Node 4</td>
<td>21.07</td>
<td>2.71</td>
<td>9.91</td>
<td>0.97</td>
<td>12.91</td>
</tr>
<tr>
<td>Node 5</td>
<td>13.3</td>
<td>3.55</td>
<td>8.31</td>
<td>2.88</td>
<td>8.82</td>
</tr>
<tr>
<td>Node 6</td>
<td>11.97</td>
<td>1.44</td>
<td>5.62</td>
<td>1.21</td>
<td>6.15</td>
</tr>
<tr>
<td>Node 7</td>
<td>5.98</td>
<td>3.63</td>
<td>4.35</td>
<td>2.41</td>
<td>4.48</td>
</tr>
<tr>
<td>Node 8</td>
<td>2.15</td>
<td>1.58</td>
<td>1.82</td>
<td>1.17</td>
<td>1.84</td>
</tr>
<tr>
<td>Node 9</td>
<td>1.95</td>
<td>1.43</td>
<td>1.63</td>
<td>1.05</td>
<td>1.65</td>
</tr>
<tr>
<td>Node 10</td>
<td>1.72</td>
<td>1.3</td>
<td>1.44</td>
<td>0.96</td>
<td>1.45</td>
</tr>
<tr>
<td>Node 11</td>
<td>1.46</td>
<td>1.21</td>
<td>1.22</td>
<td>0.88</td>
<td>1.23</td>
</tr>
<tr>
<td>Node 12</td>
<td>0.36</td>
<td>0.48</td>
<td>0.4</td>
<td>0.41</td>
<td>0.4</td>
</tr>
<tr>
<td>Rate</td>
<td>0.64</td>
<td>0.25</td>
<td>0.65</td>
<td>0.26</td>
<td>0.65</td>
</tr>
<tr>
<td>Igrvar</td>
<td>0.08</td>
<td>0.04</td>
<td>0.01</td>
<td>0.11</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Divergence times are millions of years ago (Ma). CI, confidence interval; HPD, highest posterior density; Igrvar, variation in the rate of evolution across branches from the independent branch rate (IGR) relaxed clock model; max, maximum; min, minimum; Rate, evolutionary rate (substitutions per site per million years); w, width.

FIG. 7. Divergence-time ages measured in nodes of extant taxa of bears (Ursidae) under five different calibration approaches. The bars and dots represent the 95% highest posterior density (HPD) and the mean for the estimated node ages respectively. The calibration methods described in Figure 2 are shown in colour. Results from previous analysis (Krause et al. 2008) are shown in black. Silhouettes of ursids from http://www.phylopic.org/ are reproduced under a Public Domain Dedication License 1.0.
recovery the most accurate and precise estimates of clade age (Figs 3, 5; Püschel et al. 2019, appendices 1–3). Although the percent error does not show the same (Fig. 4; Püschel et al. 2019, appendices 2, 3), this is probably related to the fact that in very shallow nodes small differences (i.e. 1–2 myr) produce a considerable percent error. As we are interested in the actual divergence ages of different lineages, it can be argued that the absolute error is more relevant. Nevertheless, all results agree that the worst performing approach was using the youngest stratigraphic occurrence to inform tip-calibration. Evidently, the best results are obtained when the stratigraphic fossil occurrences used to simulate the morphological data are used to analyse the data. This is because the morphological data were simulated according to the branch lengths of the input tree. In consequence, as the data were simulated using trees in which branch lengths were trimmed to the oldest fossil occurrence, the phylogenetic signal in the morphological data will tend to generate best results when using the oldest fossil occurrence to calibrate the tree. In some sense, these results are linked to the simplicity of our simulation framework, however, they are designed to reflect stasis: the expectation that morphospecies are morphologically consistent. Regardless, these results highlight the importance of calibrating fossil tips based on the age of the fossil specimen from which the cladistic morphological data were obtained, avoiding using older or younger specimens referred to that species that could cause a mismatch in the model. This is important because often the taxonomic rank of species does not encompass the full intraspecific variation within it, and the distinction between morphospecies is often arbitrary (Simpson 1951; Mallet 1995; Manceau & Lambert 2018).

It is important to emphasize that in terms of coverage (the 95% HPD encompassing the true node age), tip-calibration proved accurate for almost all of the calibration approaches that we explored, with the exception of using the youngest occurrence of a lineage in the tree with calibrations near to the root (Püschel et al. 2019, appendix 4). However, this high level of accuracy is achieved because the imprecision of the most inaccurate calibration strategies is larger.

Proximity to the root. Tip-calibration better integrates the uncertainty of fossil calibrations leading to more accurate and precise results when the calibrating fossils are located topologically close to, rather than remote from, the root (Figs 3, 5; Püschel et al. 2019, appendices 1–3). These results are probably related to the effect of the position of the calibrating fossils in relation to the internal nodes of tree. It has been observed that the larger the distance in time between the nodes and the calibrating fossils, the greater the error of the divergence-time estimates (Conroy & van Tuinen 2003). There are examples of this pattern in node calibration (Conroy & van Tuinen 2003; Linder et al. 2005; Rutschmann et al. 2007) and more recently in tip-calibration (Arcila et al. 2015). Accordingly, as the tree with the fossil tip-calibrations remote from the root does not establish any constraint on the deepest nodes of the tree, estimates of the age of these nodes tend to be older than the true node ages, resulting in a magnification of the error approaching the root. The reverse does not obtain because nodes ages are strongly constrained by the extant tip ages, whereas there are no such constraints towards the root. As node ages increase with proximity to the root, the uncertainty associated with their age estimates increases concomitantly (Yang & Rannala 2006; Rannala & Yang 2007). Therefore, in order to increase precision and accuracy of divergence-time estimates in the construction of deep-time phylogenies, it is important to obtain calibrations closer to the root than to the tips. This pattern has also been observed in node-calibration studies (Duchêne et al. 2014; Mello & Schrago 2014). The percent error (Fig. 4; Püschel et al. 2019, appendices 2, 3) showed a different pattern probably because this error is strongly sensitive to small differences in shallow nodes.

Empirical analyses: bear divergence time estimates

Considering the phylogeny of bears (Ursidae) under the different calibration approaches, it is clear that the results (Fig. 6A–E) are consistent with the patterns exhibited in the simulation analyses (Fig. 3), viz differing formulations of tip-calibrations based on the same stratigraphical data result in different estimates of clade age. A similar observation has been made in an attempt to estimate the evolutionary timescale of Pinus (Saladin et al. 2017). This could explain why the most precise approach in the simulations (Fig. 5A) proved to be the least precise in the empirical analysis (Fig. 6A). It is very likely that similarity in divergence-time estimates across calibration approaches within crown-Ursidae (and specifically within subfamily Ursinae; Fig. 7) is a consequence of their very recent cladogenesis. Shallow node age estimates are heavily constrained by the recent age of living lineages and, hence, they converge readily on similar ages, in contrast to deep nodes which are much more unconstrained. The same pattern was noted above for the most recent nodes in the simulations (Fig. 3).

Previous studies placed the radiation within the subfamily Ursinae close to the Miocene–Pliocene boundary at 5.33 Ma (Krause et al. 2008; Heath et al. 2014), in a period characterized by the expansion of C4 grass biomes (Cerling et al. 1998), faunal turnover and a generalized temperature drop (Van Valkenburgh 1999). In contrast, our analyses suggested that this radiation occurred in the Pleistocene, around 2 Ma (Fig. 7). These results probably
differ because of the position of the fossil calibrations. It has been established that even with multiple calibrations, the choice of a particular calibration arrangement can have significant effects on node ages estimation (Sauquet et al. 2012), and that precision tends to increase with the proximity of the calibration to the node estimated (Conroy & van Tuinen 2003; Linder et al. 2005; Rutschmann et al. 2007; Arcila et al. 2015). Previous studies used fossils within the subfamily Ursinae, such as Ursus minimus, to calibrate this part of the tree and this had an impact on the estimation of evolutionary rate (Krause et al. 2008; Heath et al. 2014). In our analyses, there were no fossil calibrations within this subfamily (because most species are known from a single stratigraphic occurrence) and so evolutionary rate change was not as well informed for that branch of the crown group. Indeed, just one fossil taxon (Indarctus arctoides) phylogenetically close to the panda (Ailuropoda melanoleuca) was used to calibrate crown-Ursidae.

Effective morphological stasis and tip-calibration

One of the prime motivations for our study was to determine how to calibrate fossil species as tip calibrations when they are known from multiple stratigraphic horizons. At least in terms of categorical characters such taxa exhibit morphological stasis, sometimes over protracted episodes of geological time, for which there are examples from both extinct (Eldredge et al. 2005) and extant (Lavoué et al. 2011) lineages. In such cases, molecular and morphological rates must be decoupled since, doubtless, molecular evolution will continue. Our results demonstrate that the choice of the calibration bounds will have a considerable effect on the inferred rate of evolution on the adjacent branches, leading to over or under-estimations of divergence ages (Ho 2009; O’Reilly et al. 2015). Morphological stasis has traditionally been attributed to either developmental constraints (an intrinsic factor) or to stabilizing selection (an extrinsic factor) (Maynard Smith et al. 1985), while recent theoretical and empirical studies suggest that morphological stasis is a product of stabilizing selection from ecological interactions (Beldade et al. 2002; Eldredge et al. 2005; Estes & Arnold 2007; Davis et al. 2014; Haller & Hendry 2014). Thus, from a theoretical perspective, it can be argued that stasis should not be considered in inference of evolutionary rates, and such lineages should be calibrated based on the oldest stratigraphic occurrence. The results of our analyses corroborate this view.

Tip-calibrations and the fossilized birth–death process. Our study was focused on comparing tip-calibration approaches within a total evidence dating framework, using a realistic simulation of fossil occurrence data; our results and their interpretations remain equally valid for morphological clock analyses. We anticipate that our results are also relevant to mechanistic models of cladogenesis like that applied to the prior on ages in the fossilized birth–death process (FBD). Certainly, comparison between mechanistic models and non-mechanistic uniform tree priors in the same dataset have revealed significant differences in divergence-time estimates, with mechanistic models giving more realistic results (Matzke & Wright 2016). Recently, the FBD model has shown attractive improvements in a new framework in which fossil occurrence data are considered explicitly in terms of their stratigraphical range (Stadler et al. 2018). Whether these improvements could lead to most precise and accurate results in tip-calibration remains to be assessed in future studies.

CONCLUSIONS

Overall, our study has shown for the first time the varying effects that different tip-calibration approaches have in divergence-time estimates. By performing simulation analyses, we have demonstrated that: (1) tip-calibration returns different divergence-time estimates depending on the calibration approach employed; (2) tip-calibration using the age of the fossil occurrence used to code the cladistic morphological data, recovers the most accurate and precise divergence time estimates; and (3) tip-calibrations located deep within phylogenetic trees tend to recover more accurate and precise results. Consequently, the best approach to calibrating the fossil tips is using fossils closer to the root and using the age of the specimen used to code the morphology to calibrate its tip. Empirical analyses based in bear phylogeny support these general conclusions, showing that the interpretation of fossil ages really matters in tip-calibrated molecular clock analyses.

Acknowledgements. This study was conducted by HPP under the guidance of JEO’R, DP and PCJD, in partial fulfilment of the requirements of the MSc Palaeobiology at the School of Earth Sciences, University of Bristol. We thank Mario dos Reis and an anonymous reviewer for helpful comments on an earlier version of the manuscript. HPP was funded by a Becas Chile scholarship (73180606, CONICYT-Chile). JEO’R, DP and PCJD are funded by NERC (NE/P013678/1; NE/N002067/1) and BBSRC (BB/N000919/1).

DATA ARCHIVING STATEMENT

Data for this study are available in the Dryad Digital Repository: https://doi.org/10.5061/dryad.7kh57t5

Editor. Imran Rahman


VEA, I. M. and GRIMALDI, D. A. 2016. Putting scales into evolutionary time: the divergence of major scale insect lineages (Hemiptera) predates the radiation of modern angiosperm hosts. Scientific Reports, 6, 23487.


