Insights from empirical analyses and simulations on using multiple fossil calibrations with relaxed clocks to estimate divergence times

Tom Carruthers[1] and Robert W. Scotland[1*]

Affiliations:

[1] Department of Plant Sciences, University of Oxford, South Parks Road, Oxford, OX1 3RB, United Kingdom

[*] Correspondence to be sent to: Department of Plant Sciences, University of Oxford, South Parks Road, Oxford, OX1 3RB, United Kingdom; E-mail: robert.scotland@plants.ox.ac.uk

Abstract

Relaxed clock methods account for among-branch-rate-variation when estimating divergence times by inferring different rates for individual branches. In order to infer different rates for individual branches, important assumptions are required. This is because molecular sequence data does not provide direct information about rates, but instead provides direct information about the total number of substitutions along any branch, which is a product of the rate and time for that branch. Often, the assumptions required for estimating rates for individual branches depend heavily on the implementation of multiple fossil calibrations in a single phylogeny. Here, we show that the basis of these assumptions is often critically undermined. First, we highlight that the temporal distribution of the fossil record often violates key assumptions of methods that use multiple fossil calibrations with relaxed clocks. With respect to "node calibration" methods, this conclusion is based on our inference that different fossil calibrations are unlikely to reflect the relative ages of different clades. With respect to the fossilised-birth-death-process, this conclusion is based on our inference that the fossil recovery rate is often highly heterogeneous. We then demonstrate that methods of divergence time estimation that use multiple fossil calibrations are highly sensitive to assumptions about the fossil record and among-branch-rate-variation. Given the problems associated with these assumptions, our results highlight that using multiple fossil calibrations with relaxed clocks often does little to improve the accuracy of divergence time estimates.

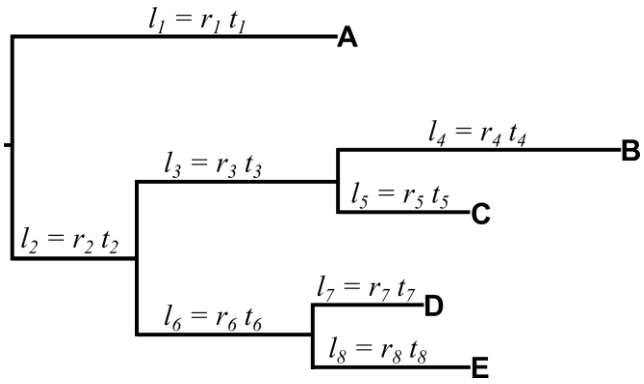
INTRODUCTION

In phylogenetics, molecular sequence data provides direct information about the number of substitutions that have occurred since homologous sequences diverged. The number of substitutions is a product of the rate that the sequences are evolving, and the time since they diverged (Zuckerkandl and Pauling 1962, 1965; Margoliash 1963). Additional evidence is therefore required to estimate rates and times in molecular phylogenies. Often, this additional evidence is a fossil calibration, where the age of a fossil is used as a basis for estimating the age of a particular clade (Zuckerkandl and Pauling 1962; Benton 1995; Sanderson 1997; Donoghue et al. 2001; Renner 2004; Donoghue and Benton 2007; Gandolfo et al. 2008; Donoghue and Yang 2016). This provides a timeframe over which molecular evolution within the clade has occurred, enabling the estimation of a rate. In the simplest case of the strict molecular clock, the rate is the same for every branch in a phylogeny (Zuckerkandl and Pauling 1962, 1965; Margoliash 1963; Miyata 1980; Baldwin and Sanderson 1997). A single well-placed calibration can therefore enable the inference of accurate divergence times throughout the phylogeny.

Often, rates vary among branches (Fig. 1) (Langley and Fitch 1974; Britten 1984; Gillespie 1989, 1991; Bromham et al. 1996). Relaxed clock methods that allow among-branch-rate-variation are therefore often used (Sanderson 1997, 2002; Thorne et al. 1998; Kishino et al. 2001; Drummond et al. 2006; Drummond and Suchard 2011; Lartillot et al. 2016). These methods rely heavily on

assumptions about rates and times of individual branches. Often, multiple fossil calibrations are used in a single phylogeny as a basis for making these important assumptions, providing "landmarks" to calibrate the relaxed clock (Sanderson 1997; Donoghue and Benton 2007; Magallón et al. 2015; Donoghue and Yang 2016). The use of multiple fossil calibrations to provide "landmarks" to calibrate the relaxed clock – and facilitate the estimation of rates and times for individual branches, and therefore the relative ages of different clades within a single phylogeny – represents a fundamental change from a traditional view of the fossil record, where fossil ages are interpreted purely as minimum age estimates for their respective clades (Donoghue and Benton 2007).

Figure 1.



Previous studies have evaluated the robustness of methods that use multiple fossil calibrations and relaxed clocks (Sanderson 1997, 2002; Thorne et al. 1998; Kishino et al. 2001; Hedges and Kumar 2004; Near and Sanderson 2004; Britton 2005; Near et al. 2005; Yang and Rannala 2006; Drummond et al. 2006; Marshall 2008; Drummond and Suchard 2011; Magallón et al. 2013; Gavryushkina et al. 2014, 2017; Heath et al. 2014; Warnock et al. 2015, 2017; Zhu et al. 2015; Zhang et al. 2016; Lartillot et al. 2016; Barba-Montoya et al. 2017). These studies have provided critical insights into the performance of different methods, and the implications of different assumptions about the nature of the fossil record and molecular evolution. Such studies typically consider one of either the fossil record (Hedges and Kumar 2004; Near and Sanderson 2004; Near et al. 2005; Yang and Rannala 2006; Marshall 2008; Heath 2012; Gavryushkina et al. 2014, 2017; Heath et al. 2014; Warnock et al. 2011, 2015, 2017; Zhang et al. 2016; Barba-Montoya et al. 2017) or molecular evolution (Thorne et al. 1998; Kishino et al. 2001; Drummond et al. 2006; Drummond and Suchard 2011; Zhu et al. 2015; Lartillot et al. 2016), and rarely consider the interactions between these two types of evidence (Magallón et al. 2013). This is despite the fact that such interactions are likely to be important in divergence time estimation – assumptions about the fossil record affect assumptions about times, assumptions about molecular evolution affect assumptions about rates, and the product of

the inferred times and rates is the number of substitutions, the parameter directly inferred from molecular sequence data.

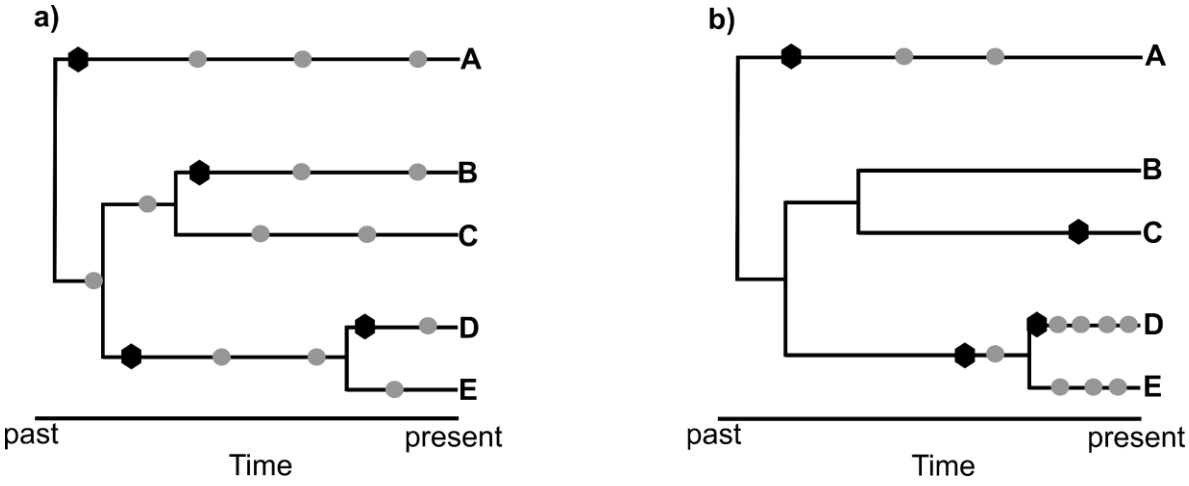
Further, the majority of methodological studies of fossil calibrations are performed in groups with well-preserved fossil records, or where fossil calibrations can be implemented in numerous clades throughout the phylogeny (Near and Sanderson 2004; Near et al. 2005; Warnock et al. 2011, 2015, 2017; Heath 2012; Magallón et al. 2013; Gavryushkina et al. 2014, 2017; Heath et al. 2014; Zhang et al. 2016). However, in many empirical studies that include divergence time estimates, the fossil record of the relevant group is considerably more fragmentary. This is especially the case for vascular plants, and means that the fossil record provides an uncertain temporal basis upon which to calibrate relaxed clocks, and that different interpretations of the temporal signal within the fossil record can lead to markedly different divergence time estimates (Särkinen et al. 2013; Magallón et al. 2015; Lagomarsino et al. 2016; Mitchell et al. 2016; Cardillo et al. 2017; Barba-Montoya et al. 2018; Morris et al. 2018; Folk et al. 2019; Muñoz-Rodríguez et al. 2019).

Here, we explore the validity and implications of different assumptions about the fossil record and molecular evolution, and also the interaction between these two types of evidence, in Osmundaceae, and Convolvulaceae and Solanaceae (CS). Time-calibrated phylogenies have recently been constructed for these two clades that follow "best practice" approaches (Parham et al. 2012; Särkinen et al. 2013; Grimm et al. 2015; Mitchell et al. 2016). Our analyses can therefore be interpreted in the context of commonly used methodologies. Further, these clades have markedly contrasting fossil records with Osmundaceae possessing a far better preserved fossil record than CS (Särkinen et al. 2013; Bomfleur et al. 2015; Grimm et al. 2015). We can therefore compare our findings in different contexts, unlike the majority of previous studies that are carried out in groups with well-preserved fossil records (Near and Sanderson 2004; Near et al. 2005; Warnock et al. 2011, 2015, 2017; Heath 2012; Magallón et al. 2013; Gavryushkina et al. 2014, 2017; Heath et al. 2014; Zhang et al. 2016). As well as these two clades, we also perform a subset of analyses across the entire Spermatophyta. A time-calibrated phylogeny has recently been constructed for this clade that incorporates over 100 fossil calibrations (Magallón et al. 2015). We can therefore also interpret our results in the context of a study of much broader phylogenetic scale that incorporates a very large number of fossil calibrations.

In our analyses of these clades, we first investigate whether the temporal distribution of their fossil records is consistent with the assumptions of methods that use multiple fossil calibrations with relaxed clocks. Specifically, we explore whether the relationship between the ages of sets of fossils used as calibrations throughout a phylogeny, and the actual ages of the respective clades of these fossils, provides a framework for estimating rates and times on individual branches, and thus the relative ages of different clades. We explore the nature of this relationship in the context of the assumptions of node calibration methods and the fossilised-birth-death-process (FBDP).

To investigate whether the temporal distribution of the fossil record is consistent with the assumptions of methods that use multiple node calibrations, we estimate whether the ages of sets of fossils previously used together as part of a set of fossil calibrations in a single phylogeny – subsequently referred to as *node-calibration-fossils* – reflect the relative ages of different clades (Fig. 2a), or instead have no such relationship (Fig. 2b). We focus on this, because to provide relevant information about times and rates for individual branches, sets of node-calibration-fossils must provide information about relative clade ages. In contrast to this, we consider the highly variable time-lags between a set of node-calibration-fossils and the actual ages of their respective clades may mean that a set of node-calibration-fossils does not reflect the relative ages of clades throughout a phylogeny.

Figure 2.



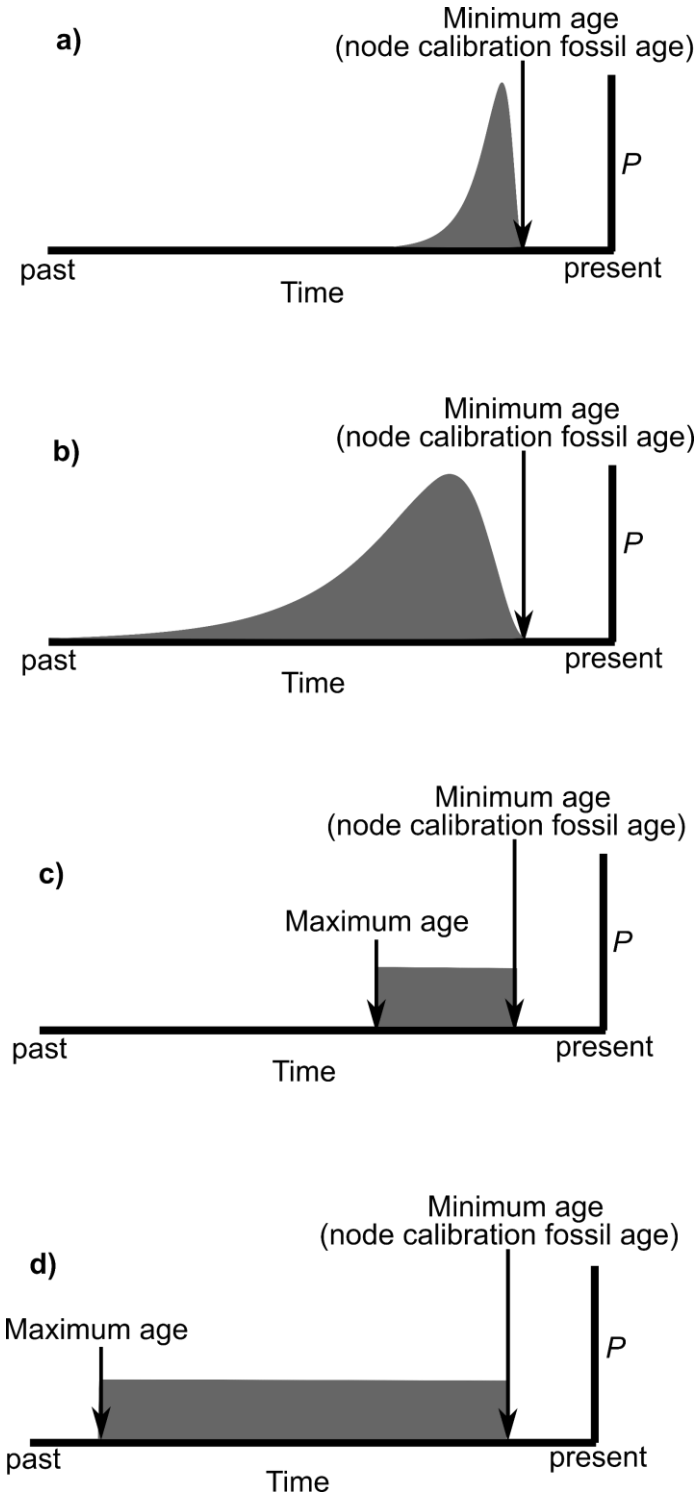
We investigate whether sets of node-calibration-fossils meet this criteria in two different ways. First, we analyse the temporal distribution of all known fossils within their respective clades, and use methods derived from palaeontology for estimating taxon age ranges (Marshall 1990, 2008; Springer 1995) to calculate confidence intervals for the ages of these clades. This provides a basis to quantify the probable relationship between the ages of node-calibration-fossils and the actual ages of their respective clades. Second, we calculate the empirical scaling factor (ESF) for node-calibration-fossils following the method of Marshall (2008). The ESF provides an alternative approximation of how the age of a node-calibration-fossil reflects the age of its respective clade by taking into account the age of a fossil and the phylogenetic depth of the node that it calibrates (see methods). By quantifying how the ages of node-calibration-fossils reflect actual clade ages, both of these measures (confidence intervals and ESFs) provide a basis to determine how sets of node-calibration-fossils are likely to reflect the relative ages of different clades.

To investigate whether the temporal distribution of the fossil record is consistent with the assumptions of methods that use the FBDP, we analyse the fossil-recovery-rate (ψ). This parameter describes the rate that fossils are sampled from a branching process. We are specifically interested in estimating whether ψ is constant (Fig. 2a), or instead whether it is heterogeneous (Fig. 2b). This is because in current implementations of the FBDP, ψ is *the* fundamental assumption that the FBDP makes of the temporal distribution of the fossil record, and in groups for which morphological matrices are not available for relevant fossils – as is often the case in plants – ψ is the single parameter through which fossil ages are incorporated into divergence time analyses (Stadler 2010; Gavyryushkina et al. 2014; Heath et al. 2014).

Following our investigation of the temporal distribution of the fossil record, we infer divergence times for Osmundaceae and CS in different methodological contexts. We can therefore determine the sensitivity of different methods to the assumptions they make about molecular evolution and the temporal distribution of the fossil record. For this second set of analyses, we compare methods for implementing fossil calibrations – subsequently referred to as *calibration strategies*, and methods for analysing molecular sequence data – subsequently referred to as *molecular clock strategies*.

We compare four different node calibration strategies. *Non-Uniform Precise* (NUP) and *Non-Uniform Imprecise* (NUI) use non-uniform probability densities (calibration densities) that assume the clade age is very similar to the age of the node-calibration-fossil (NUP) (Fig. 3a), or make less precise assumptions about the clade age in relation to the node-calibration-fossil (NUI) (Fig. 3b). *Uniform Precise* (UP) and *Uniform Imprecise* (UI) use uniform calibration densities in which the maximum is either young (UP) (Fig. 3c) or old (UI) (Fig. 3d). For calibration strategies, we also implement the FBDP in three different ways. For *FBDP Low* (FBDPL) we use a prior for ψ that is set to a low value, for *FBDP Intermediate* (FBDPI) we use a prior for ψ that is set to an intermediate value, and for *FBDP High* (FBDPH) we use a prior for ψ that is set to a high value.

Figure 3.



For molecular clock strategies, we compare; a *strict clock* – where the same rate is inferred for every branch; a *relaxed clock* – where rates are inferred independently for each branch; and no *molecular data* – where no molecular data is analysed when inferring divergence times. When no molecular data is analysed, inferred divergence times will be equivalent to the effective time prior.

It is important to note that the purpose of this second set of analyses is not to estimate the most accurate timescale for the evolution of these two groups. Instead, the purpose is to provide a comparative framework to assess the implications of different methodological assumptions. In order to provide such a comparative framework, we inevitably implement some methodologies, such as the strict clock, that are no longer widely used. As well as implementing methodologies that are no longer widely used, we also use terminology that best reflects the comparative framework that is central to this study. For example, we refer to no molecular data as a molecular clock strategy, rather than the effective time prior. This is because the purpose of different molecular clock strategies is to compare the implications of different methods for analysing molecular sequence data. We consider no molecular data as an extreme case, and a baseline from which evaluate alternative approaches for analysing molecular sequence data.

When comparing inferred divergence times with these different methods, we refer to mean posterior age estimates (MPEs) and 95% highest posterior density (HPD) intervals. We also compare the difference between the age of the oldest fossil for a clade (the node-calibration-fossil for node calibration methods), and the MPE for that clade. This quantifies the relationship between fossil ages and inferred divergence times. Taken together, the comparisons between these different methods enable a comprehensive analysis of the interaction between different assumptions about the fossil record and molecular evolution, and the implications of this for divergence time estimation.

RESULTS

Phylogenetic distribution of fossils

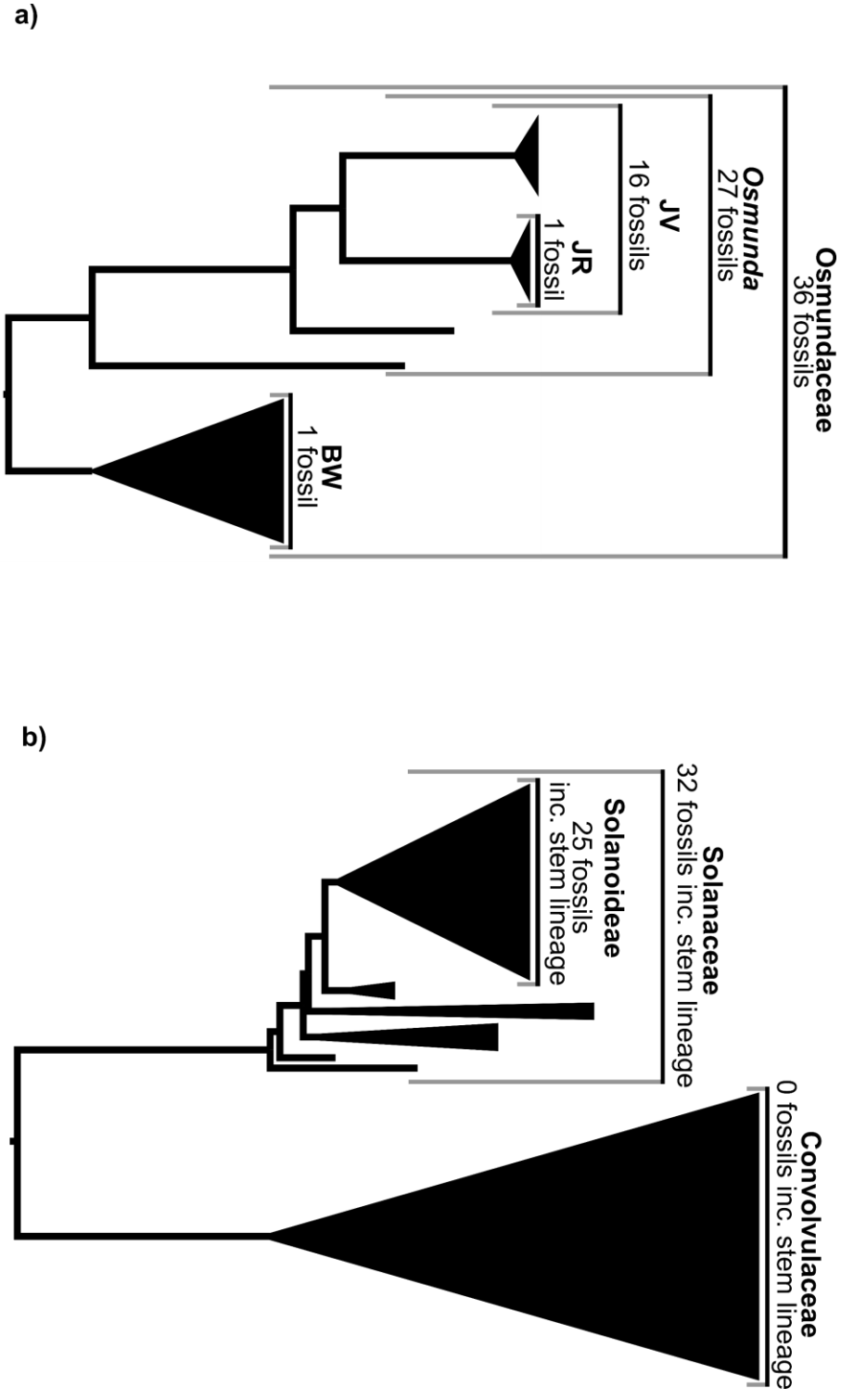
We inferred phylogenies for Osmundaceae, CS, and Spermatophyta as a basis for subsequent analyses. These phylogenies are congruent with previous studies (Fig. S1-3) (Särkinen et al. 2013; Grimm et al. 2015; Magallón et al. 2015). We therefore assign the fossils in this study to the same clades as previous studies (Table S1-3) (Särkinen et al. 2013; Grimm et al. 2015; Magallón et al. 2015).

The fossils analysed in Osmundaceae and CS are the entire known fossil records for these groups (Särkinen et al. 2013; Grimm et al. 2015). Figure 4a summarises the phylogenetic distribution of fossils within Osmundaceae. 36 fossils belong to Osmundaceae, 27 fossils belong to *Osmunda*, 16 fossils belong to the clade of *O. japonica – O. vachellii* (JV), 1 fossil belongs to the clade of *O. japonica – O. regalis* (JR), and 1 fossil belongs to the clade of *T. barbara – L. wilkesiana* (BW). Given that 27 fossils belong to *Osmunda* (crown group) and 1 fossil belongs to BW (crown group), 8 fossils belong to the stem lineages of *Osmunda* and/or BW (thus giving the total of 36 fossils in

Osmundaceae). Figure 4b summarises the phylogenetic distribution of fossils within CS. 32 fossils belong to Solanaceae (including the stem lineage), 25 fossils belong to Solanoideae (including the stem lineage), and no fossils belong to Convolvulaceae (including the stem lineage).

For Spermatophyta, we only analysed the oldest known fossil within each clade. These fossils are distributed throughout Spermatophyta (Table S3).

Figure 4.



Confidence intervals for clade ages

Osmundaceae

The fossil record for Osmundaceae extends into the Upper Triassic (237 Ma), for *Osmunda* it extends into the Middle Jurassic (165 Ma), for JV it extends into the Upper Cretaceous (86 Ma), for JR it extends into the Miocene (14 Ma), and for BW it extends into the Eocene (52 Ma) (Fig. 5a). For all clades with more than one fossil (Osmundaceae, *Osmunda* and JV), Kolmogrov-Smirnov tests indicate that fossils are sampled at an exponentially decreasing frequency with time from present (Table S4). Based on these distributions, the upper limit of a 95% confidence interval that includes the actual age of Osmundaceae is in the Lower Orodovician (471 Ma), for *Osmunda* it is in the Upper Carboniferous (318 Ma), and for JV it is in the Lower Jurassic (197 Ma) (Fig. 5a, Table S5). These ages are far older than the 95th percentiles of calibration densities used in a recent time-calibrated phylogeny (Grimm et al. 2015) (Fig. 5a).

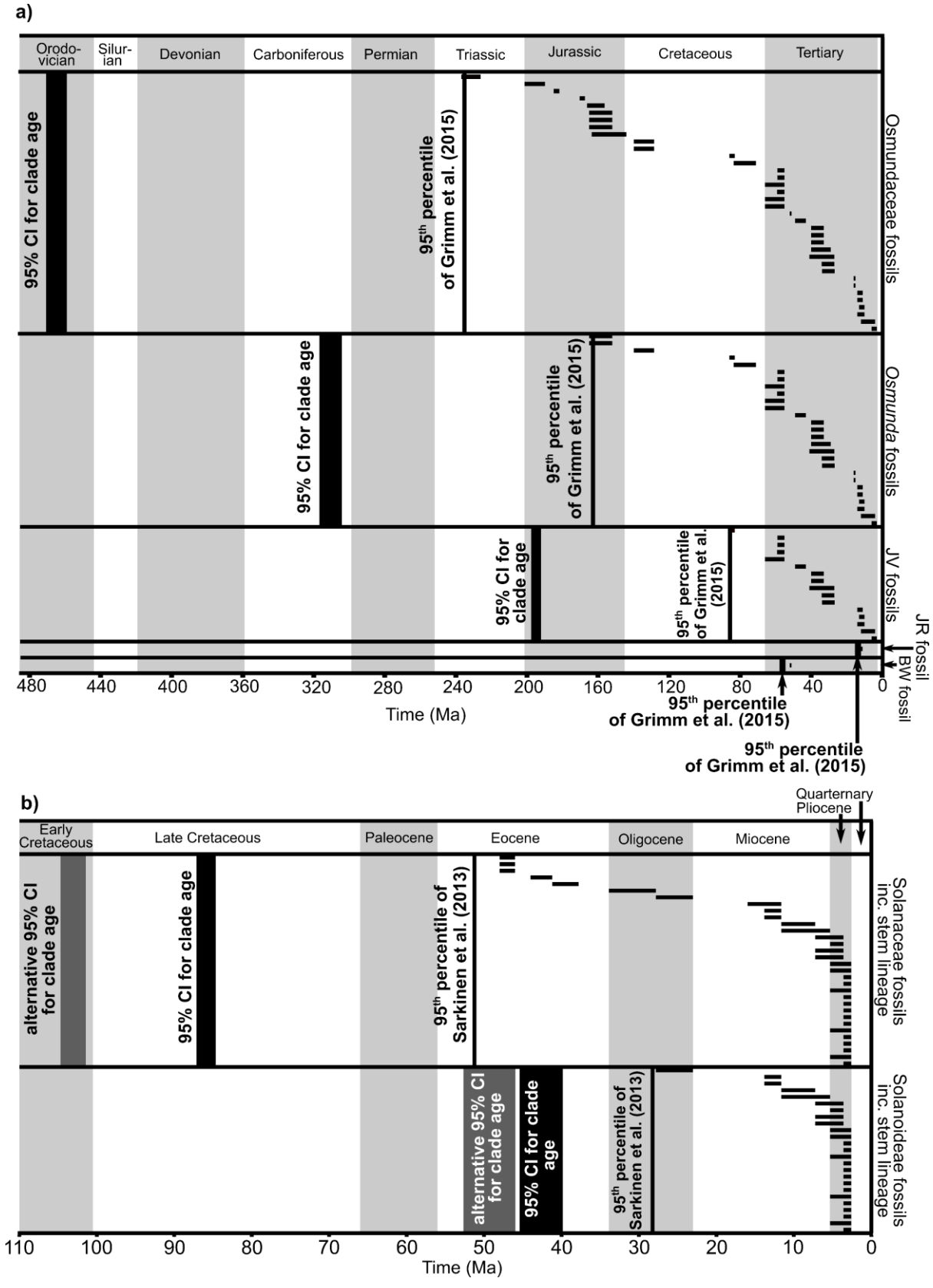
CS

The Solanaceae fossil record (including the stem lineage) extends into Eocene (48.0 Ma), and the Solanoideae fossil record (including the stem lineage) extends into the Oligocene (28.4 Ma) (Fig. 5b). For both clades, Kolmogrov-Smirnov tests indicate that fossils are more likely to be sampled at an exponentially decreasing frequency with time from present, rather than a uniform frequency (Table S6). However, the high frequency of Pliocene fossils made fitting an exponential distribution problematic. We therefore analysed a subsample of the fossil record in which 13 Pliocene fossils were removed. In this case, the exponential distribution was a far better fit for both clades (Table S6). When using this subsample of fossils, the upper limit of a 95% confidence interval that includes the actual age of Solanaceae (and its stem lineage) is in the Lower Cretaceous (104.9 Ma), and for Solanoideae (and its stem lineage) is in the Eocene (52.5 Ma) (Fig. 5b, Table S7). When Pliocene fossils were not removed, the upper limits of 95% confidence intervals were in the Upper Cretaceous (87 Ma) for Solanaceae (and its stem lineage) and still in the Eocene (45 Ma) for Solanoideae (and its stem lineage) (Fig. 5b). Both sets of ages are far older than the 95th percentile of calibration densities used in a recent time-calibrated phylogeny (Särkinen et al. 2013) (Fig. 5b).

Validating confidence intervals

Simulation experiments for validating confidence intervals highlight that the confidence intervals we calculated are unlikely to exaggerate uncertainty about the relationship between the age of node-calibration-fossils and the actual clade age (Supplementary Information 1, Table S8, Fig. S4). As with the empirical datasets, our simulation experiments highlighted that clades can be far older than node-calibration-fossils. By contrast, with simulation experiments for validating previously implemented calibration densities, it was shown that the 95th percentile of previously implemented calibration densities significantly underestimate uncertainty about the relationship between the age of the node-calibration-fossils and actual clade ages (Supplementary Information 1, Fig. S5).

Figure 5.



ESFs of node-calibration-fossils

Osmundaceae

When ESFs are calculated according to ultrametric phylogeny inferred in treePL, they range from 90.0 to 485.6 (Fig. 6a, Table S5). This indicates a more than five-fold difference in how the ages of node-calibration-fossils reflect the actual age of their respective node. For example, assuming the age of the JR node-calibration-fossil (ESF = 485.6) is equal to the actual age of its node (solely for the purposes of explanation), the age of the BW node-calibration-fossil (ESF = 90.0) is 18.5% of the age of its node. Differences in ESFs of a similar scale are recorded when they are calculated according to ultrametric phylogenies inferred with alternative methods (Fig. S6a and S7a).

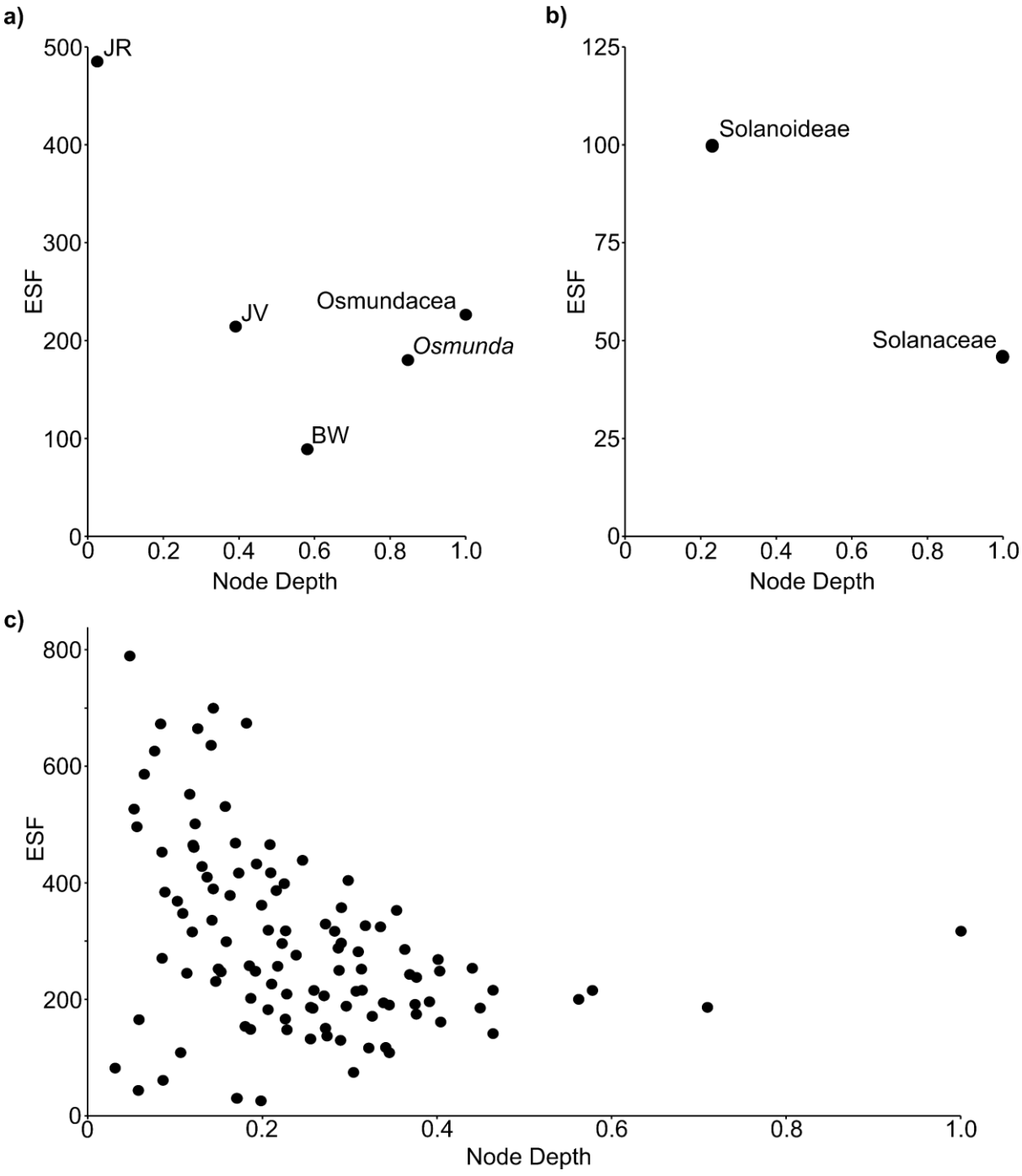
CS

When calculated according to the ultrametric phylogeny inferred in treePL, the ESF of the node-calibration-fossil at the Solanaceae stem node is 46.0, and the ESF of the node-calibration-fossil at the Solanoideae stem node is 99.9 (Fig. 6b, Table S7). This indicates a more than two-fold difference in how the ages of these two node-calibration-fossils reflect the actual age of their respective node. When ultrametric phylogenies are inferred according to alternative methods, differences in ESFs of a similar scale are recorded (Fig. S6b and S7b).

Spermatophyta

For Spermatophyta, when ESFs are calculated according to the ultrametric phylogeny inferred in treePL, the largest ESF is 790.1 and the smallest ESF is 26.9 (Fig. 6c). This indicates a more than 29-fold difference in how the ages of node-calibration-fossils reflect the actual age of their respective node. Further, node-calibration-fossils at deeper nodes tend to have lower ESFs. Differences in ESFs of a similar scale are recorded when ultrametric phylogenies are inferred with alternative methods (Fig. S6c and S7c).

Figure 6.



The fossil recovery rate (ψ)

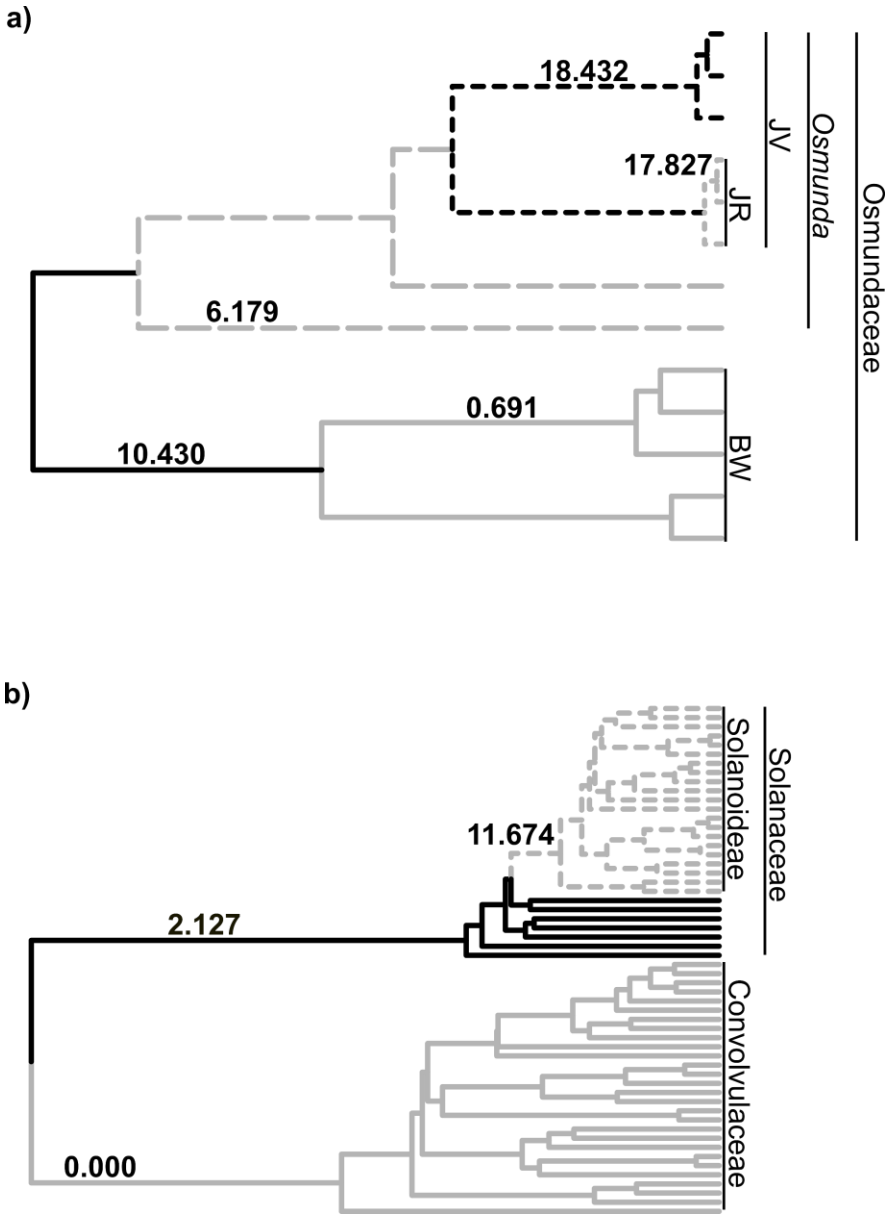
Osmundaceae

Large differences in ψ were estimated within Osmundaceae (Fig. 7a, Fig S8a, Fig S9a). When ψ was calculated according to the ultrametric phylogeny inferred in treePL, the highest estimate of ψ is for JV (excluding JR) with 18.4 fossils per-unit-of-time, and the lowest estimate of ψ is for BW with 0.7 fossils per-unit-of-time. Differences in ψ of a similar scale were recorded when alternative methods were used to infer the ultrametric phylogeny according to which ψ was calculated (Fig. S8a and S9a).

CS

Large differences in ψ were also estimated within CS (Fig. 7b, Fig. S8b, S9b). When ψ was calculated according to the ultrametric phylogeny inferred in treePL, the estimate of ψ for Solanaceae (including the stem lineage but excluding Solanoideae) is 2.1 fossils per-unit-of-time, the estimate of ψ for Solanoideae (including the stem lineage) is 11.7 fossils per-unit-of-time, and the estimate of ψ for Convolvulaceae (including the stem lineage) is 0.0 fossils per-unit-of-time. Differences in ψ of a similar scale were recorded when alternative methods were used to infer the ultrametric phylogeny according to which ψ was calculated (Fig. S8b and S9b).

Figure 7.



Inferring divergence times with different calibration strategies and molecular clock strategies

When discussing results in the context of node calibration strategies, we refer independently to NUP, NUI, UP, and UI, and their specific interactions with different molecular clock strategies. This is because inferred divergence times differ profoundly between these different calibration strategies. By contrast, for the FBDP, results are very similar for FBDPI, FBDPS, and FBDPH. This is likely to be because the posterior value for ψ was very similar, regardless of the prior (in all cases the posterior value was approximately $9*10^{-3}$ *fossils* Myr^{-1}). In all subsequent discussion of results for

the FBDP, we therefore refer collectively to the FBDP, rather than individually to FBDPI, FBDPS or FBDPH.

Mean posterior estimates (MPEs)

For node calibration strategies with non-uniform calibration densities, NUP leads to younger MPEs than NUI. With NUI, there is typically a 15-40% increase in MPEs relative to NUP (Fig. 8 and 9, Table S9 and S10). For node calibration strategies with uniform calibration densities, UP leads to younger MPEs than UI. With UI, there is typically a 15-40% increase in MPEs relative to UP (Fig. 8 and 9, Table S9 and S10). Inferred MPEs were also sensitive to the molecular clock strategy, with the precise impact of different molecular clock strategies being highly context specific. It was nonetheless notable that at shallower nodes, MPEs were markedly older when no molecular data was analysed (Fig. 10a-d and 11, Table S11 and S12). The ranked order of MPEs for different calibration strategies is also sensitive to the molecular clock strategy (Fig. 8 and 9, Table S9 and S10), and the ranked order at different nodes varies most with no molecular data, less with a relaxed clock, and least with a strict clock (Fig. 8 and 9).

For the FBDP, MPEs are intermediate between the oldest and youngest MPEs inferred with node calibration strategies (Fig. 8, Table S9). On average, there is a 2.2% decrease in MPEs with a relaxed clock relative to a strict clock, and the difference is larger at younger nodes (Fig 10e-f, Table S11).

Figure 8.

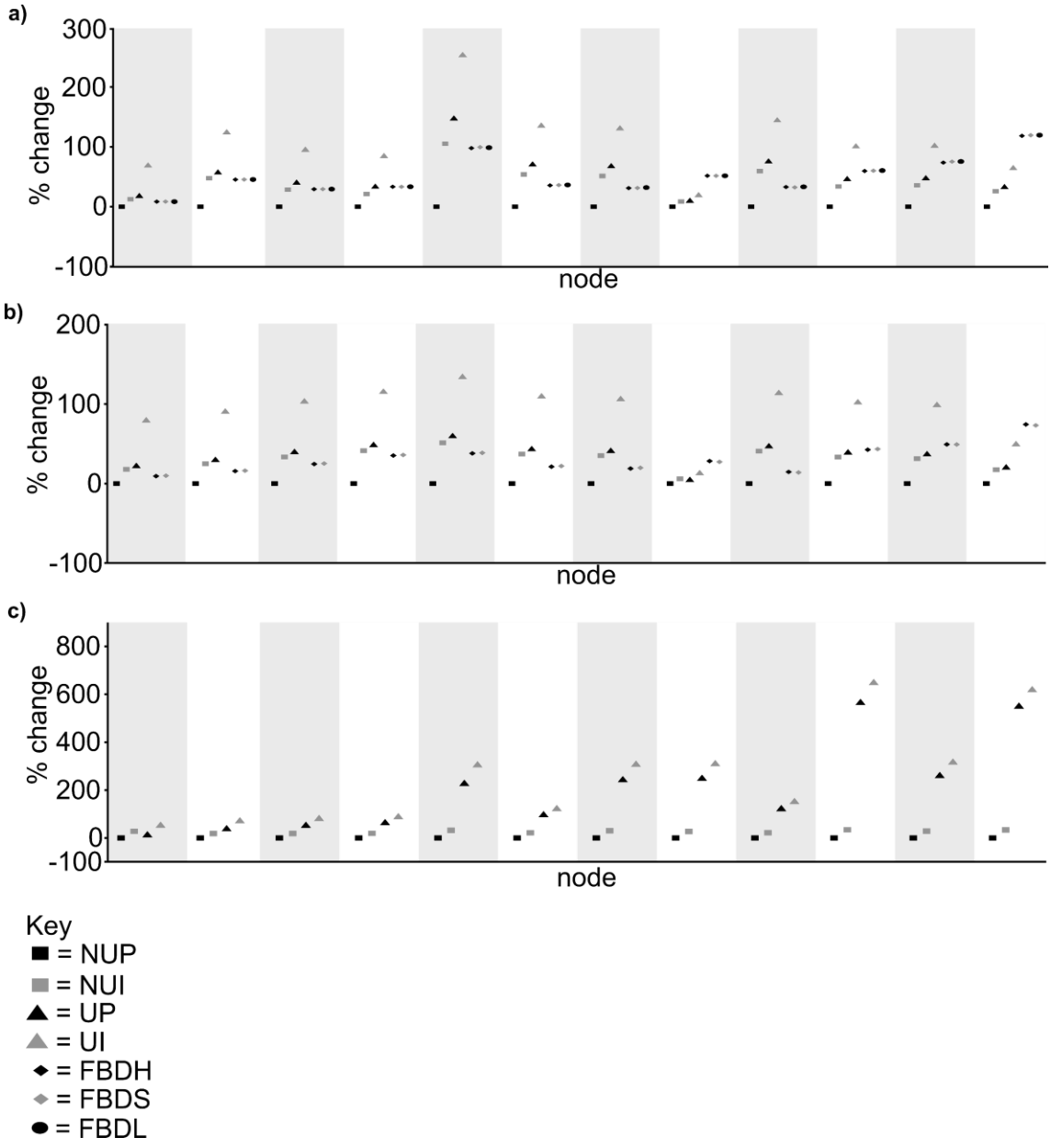


Figure 9.

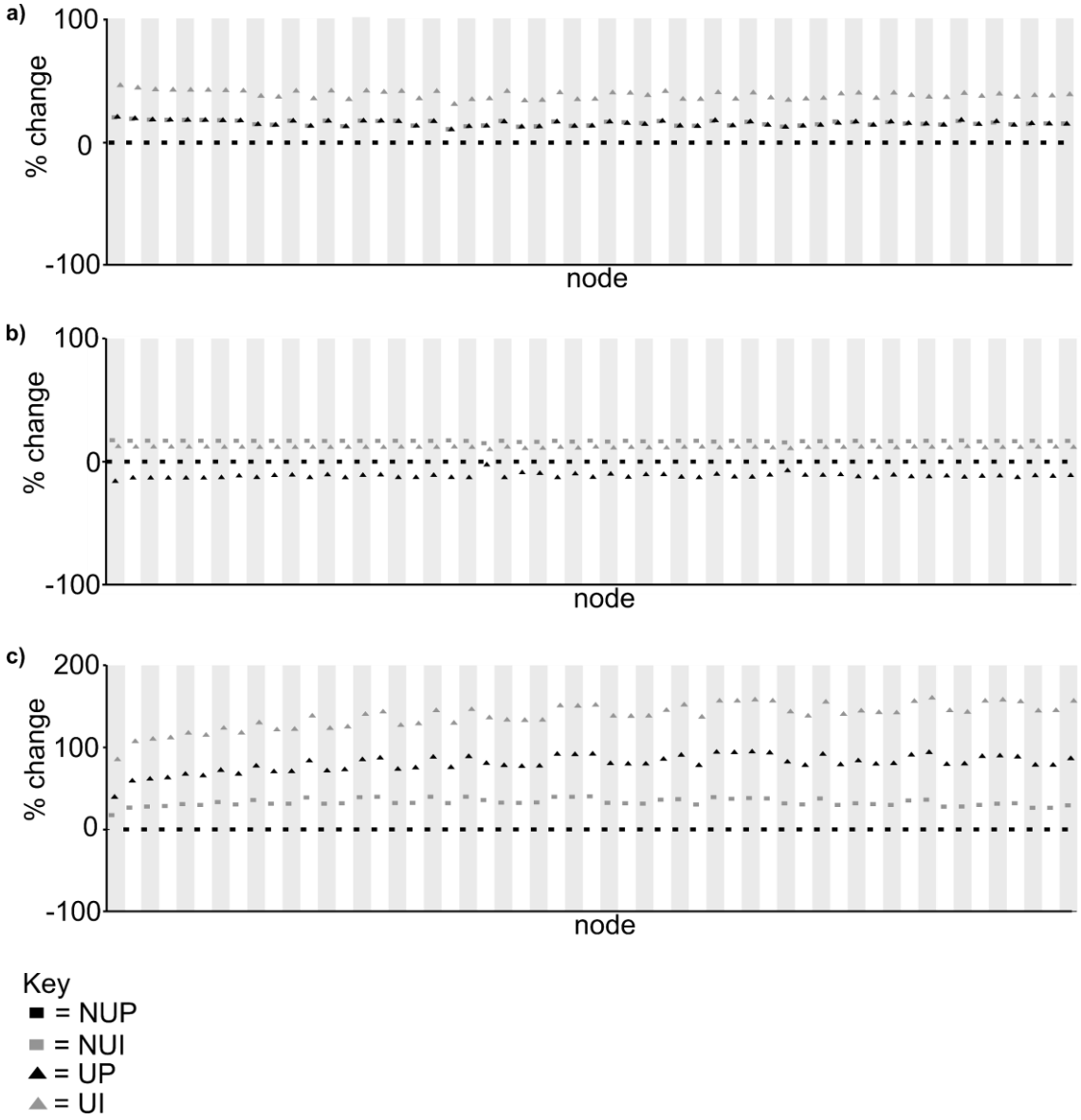


Figure 10.

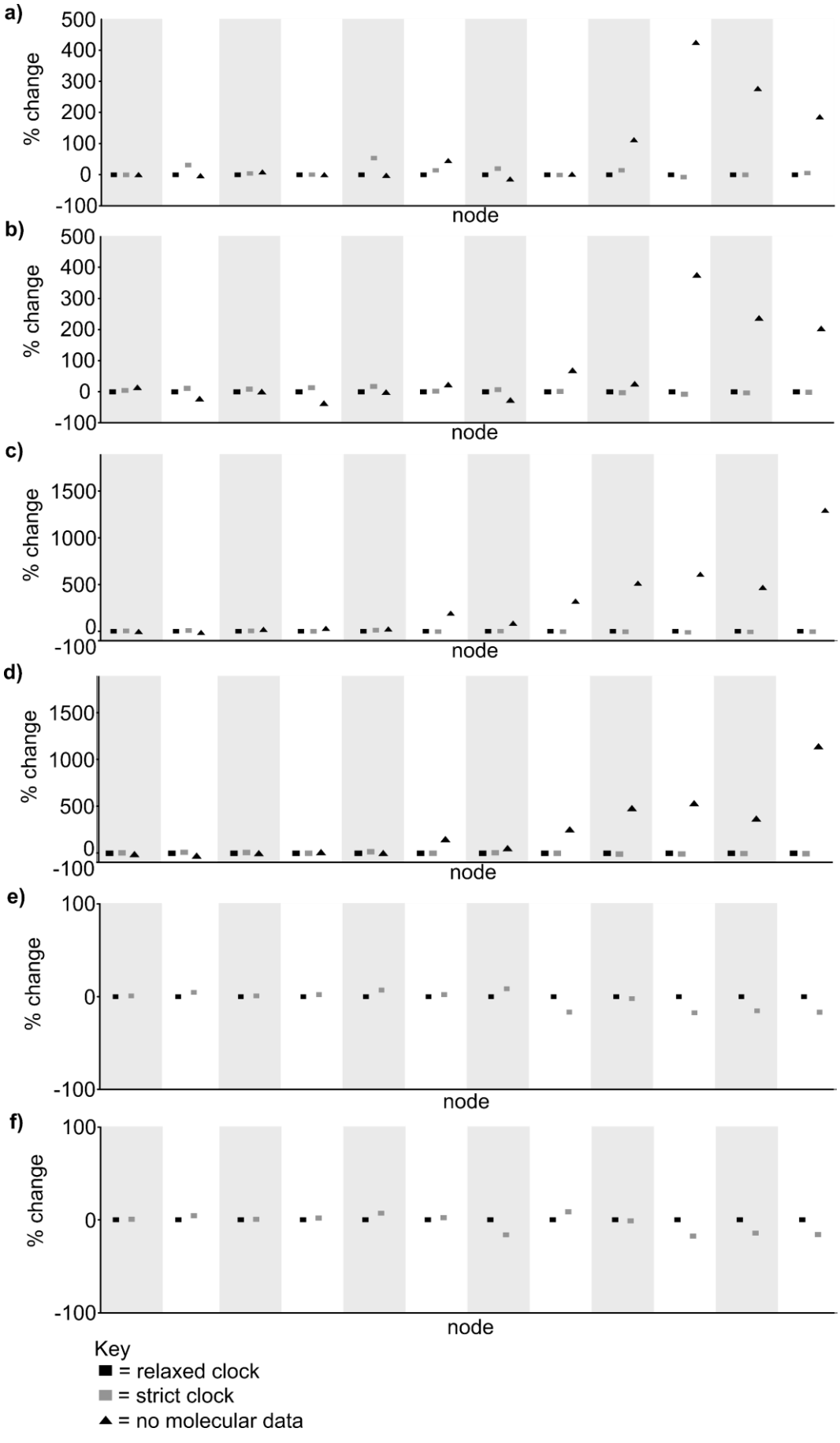
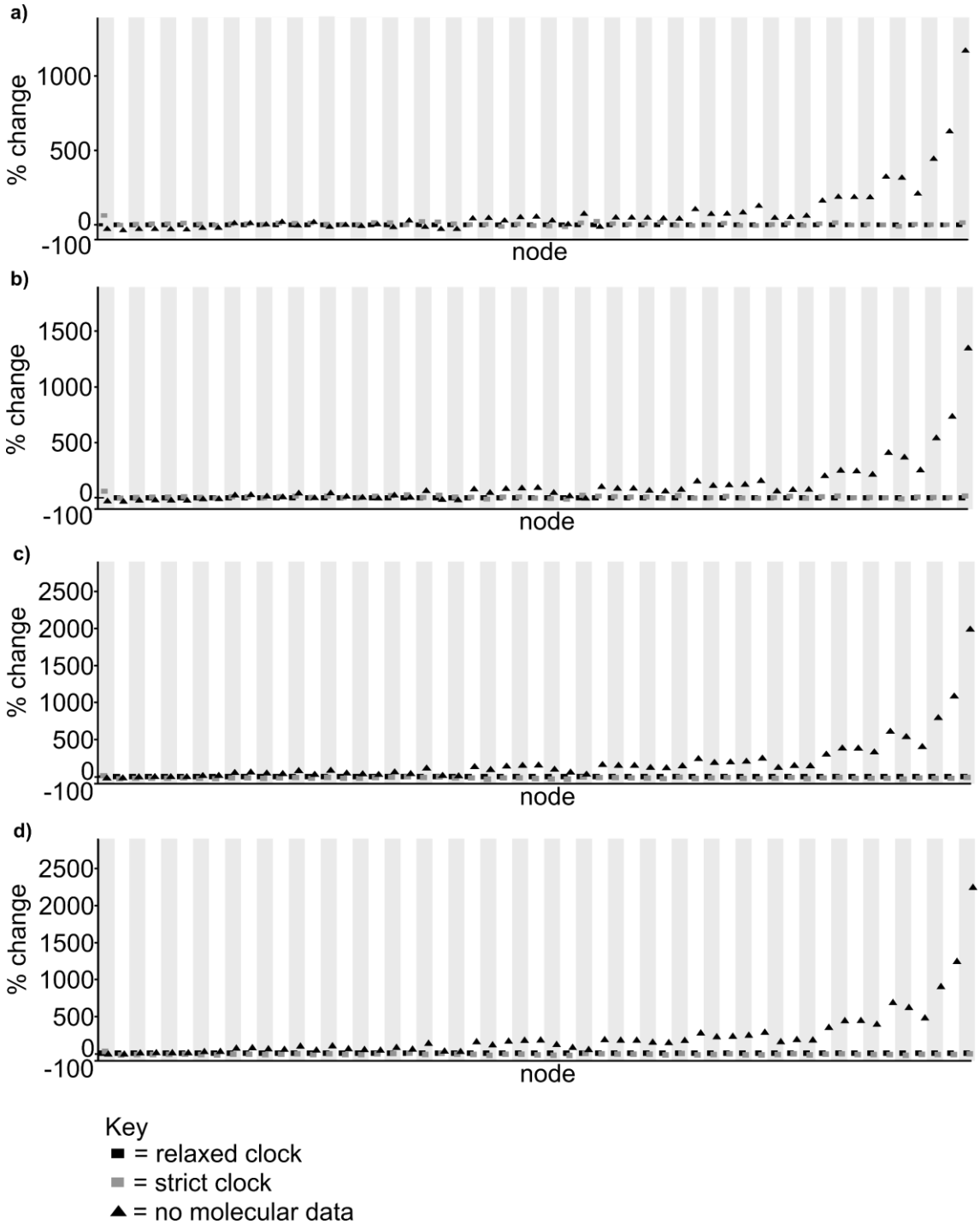


Figure 11.



Difference between MPE and the oldest fossil within a clade

For node calibration strategies with non-uniform calibration densities, NUP leads to smaller differences between the age of the oldest fossil for a clade and the MPE for the clade, compared to NUI. With NUI, there is typically a 100-1400% increase in the difference relative to NUP, although in some cases the increase is much greater (Table S13-S18). For node calibration strategies with uniform calibration densities, UP leads to smaller differences than UI. With UI, there is typically a 50-150% increase in the difference relative to UP (Table S13-S18). The difference is also sensitive to the molecular clock strategy. For non-uniform calibration densities (NUP and NUI), the difference becomes progressively smaller with a relaxed clock and no molecular data, compared to a strict clock (Table S13-S18). This is most clearly the case for NUP, and especially if the node-calibration-fossil has a low ESF, such as the Solanaceae stem node fossil (Table S5, S7, S13-S18). For uniform calibration densities (UP and UI), the effect of different molecular clock strategies is more complex. For older node-calibration-fossils (especially with low ESFs), the difference becomes progressively smaller with a relaxed clock and no molecular data compared to a strict clock (Table S5, S7, S13-S18). Further, the decrease in difference is greater for UI than UP, but the overall scale of the differences are larger for UI than UP, especially with a strict clock (Table S13-S18). By contrast, for younger node-calibration-fossils (especially with low ESFs), the difference becomes somewhat larger with a relaxed clock and no molecular data compared to a strict clock (Table S5, S7, S13-S18).

For the FBDP, the differences are intermediate between the smallest and largest differences inferred with node calibration strategies (Table S13 and S14). There were only moderate changes in the difference between the strict clock and the relaxed clock (Table S13 and S14). With a relaxed clock, the differences were smaller for clades/grades with a lower inferred ψ such as *Osmunda* (excluding JR and JV) and BW, and larger for clades/grades with a higher inferred ψ such as JR (Table S13 and S14, Fig. 7a).

95% HPD widths

For node calibration strategies with non-uniform calibration densities, NUP leads to narrower 95% HDPs than NUI. With NUI, there was typically a 10-50% increase in the 95% HPD width relative to NUP (Table S19 and S20). For node calibration strategies with uniform calibration densities, UP leads to narrower 95% HPDs than UI. With UI, there was typically a 15-35% increase in the 95% HPD width relative to UP (Table S19 and S20). In both datasets, 95% HPDs tend to be narrowest with the strict clock, wider with the relaxed clock (around 30-40% wider than the strict clock), and widest with no molecular data (often over 100% wider than the strict clock) (Table S19-20). For the FBDP, there was a 30% increase in the 95% HPD width with a relaxed clock relative to a strict clock (Table S19).

DISCUSSION

Multiple node-calibration-fossils are unlikely to reflect the relative ages of different clades

There is considerable uncertainty in how the ages of node-calibration-fossils reflect the actual ages of their respective clades in both Osmundaceae and CS. Multiple node-calibration-fossils are therefore unlikely to reflect the relative ages of different clades. This conclusion is based on the confidence intervals we calculated for the ages of clades within Osmundaceae and CS. The timespan covered by these confidence intervals differ profoundly, and cover timescales that are 10s to 100s of millions-of-years older than the node-calibration-fossil (Fig. 5a and Fig. 5b).

Given the upper limits of the confidence intervals are vastly older than the 95th percentiles of calibration densities from Grimm et al. (2015) and Särkinen et al. (2013), the confidence intervals also highlight that the assumptions of previous studies are likely to be unrealistic (Fig. 5a and 5b). This conclusion is supported by simulation experiments which indicate that previously implemented calibration densities significantly underestimate uncertainty about the relationship between the age of node-calibration-fossils and actual clade ages (Supplementary Information 1, Fig. S5). Even if the confidence intervals are ignored, Figures 5a and 5b illustrate problems with the assumptions of Grimm et al. (2015) and Särkinen et al. (2013). Gaps in the fossil records of Osmundaceae and CS are too extensive to make confident statements about clade ages in relation to the ages of node-calibration-fossils, as was done in these studies. Although it cannot be stated with certainty that the assumptions of Grimm et al. (2015) and Särkinen et al. (2013) are definitely wrong, available evidence does not provide a basis upon which to state with confidence that they are valid. In this context, the discovery of plant fossils that are vastly older than previously known fossils and previous divergence time estimates (Wilf and Escapa 2015) – including a 52 million-year-old Solanoideae fossil not included in this study (Wilf et al. 2017) – is unsurprising.

Our conclusion that multiple node-calibration-fossils are unlikely to reflect the relative ages of different clades is corroborated by the ESFs we calculated for node-calibration-fossils in Osmundaceae, CS, and Spermatophyta (Fig. 6a, b, and c). These suggest considerable differences in how the ages of node-calibration-fossils reflect the actual ages of their respective clades (more than 5-fold differences in Osmundaceae, more than 2-fold differences in CS, and more than 27-fold differences in Spermatophyta).

Analyses of ESFs are confounded by among-branch-rate-variation. The conclusions we make about Spermatophyta – which are based solely on ESFs – are therefore less robust. However, variation in ESFs within Spermatophyta is of a larger scale than Osmundaceae and CS. Given that confidence intervals within Osmundaceae and CS corroborate the argument that multiple node-calibration-fossils are unlikely to reflect relative clades ages, it cannot be assumed with confidence that this is not also the case in Spermatophyta.

Variation in ψ

Recent discussion has focussed on the value of the FBDP for integrating fossils with molecular phylogenies to estimate divergence times. It is often suggested that a strength of the FBDP

is that unlike node calibration methods, it explicitly models fossil occurrences and extracts temporal information from the entire fossil record (Heath 2014; Grimm et al. 2015; Lee and Palci 2015; Renner et al. 2016; Gavryushkina et al. 2017). In order to do this, ψ is inferred. However, an underlying assumption is that ψ is constant. Our analyses of Osmundaceae and CS indicate that contrary to this assumption, ψ is highly heterogeneous (Fig. 7a and b). This indicates that the manner by which the FBDP extracts temporal information from the fossil record is fundamentally undermined. This is likely to be a problem across much of the tree of life, given that analyses of empirical datasets often indicate that a constant ψ is an unreasonable assumption to make about the fossil record (Raup 1972; Signor and Lipps 1982; Smith 2001; Smith and McGowan 2005).

Similarly to ESFs, estimates of ψ are confounded by among-branch-rate-variation. Regardless, Figure 4 (which summarises the phylogenetic distribution of fossils) highlights that large differences in ψ are still highly likely. For example, within Osmundaceae there is a 27-fold difference between the number of fossils in two sister clades (Fig. 4a).

Comparing divergence time estimates with different methods

Comparing MPEs for node calibration strategies

When comparing node calibration strategies that use either non-uniform calibration densities (NUP and NUI) or uniform calibration densities (UP and UI), the strategies that assume actual clade ages are closer to the ages of node-calibration-fossils (NUP or UP) lead to younger MPEs than the strategies that assume actual clade ages may be considerably older (NUI or UI) (Fig. 8 and 9; Table S9 and S10). We observe this pattern regardless of the molecular clock strategy, thus highlighting the importance of the calibration strategy in all contexts.

Our results also highlight the sensitivity of MPEs to the molecular clock strategy (Fig. 10a-d, and Fig. 11, Table S11 and S12). These differences are highly context specific, and it is difficult to extract a general pattern when analysing general patterns of MPEs across all nodes. The context specific nature of these patterns is likely to reflect the complex nature of divergence time analyses, where assumptions are expressed through the calibration strategy, the molecular clock strategy, and the birth-death tree prior (which assumes the temporal distribution of branching events is controlled by a branching process with constant rates of speciation and extinction).

However, our results do highlight one consistent pattern with respect to the effect of the molecular clock strategy on MPEs. Specifically, the ranked order of MPEs with different calibration strategies at each node becomes progressively more varied with a relaxed clock and no molecular data, compared to a strict clock (Fig. 8 and 9). In this context, the relaxed clock is therefore intermediate between the strict clock and no molecular data. We explain this pattern by comparing the attributes of the three molecular clock strategies: the strict clock does not allow among-branch-rate-variation; the relaxed clock allows some among-branch-rate-variation but its magnitude and nature is constrained by the model used (in our experiments a UCLN relaxed clock); and analyses with no

molecular data are equivalent to assuming that rates can vary among branches by an infinite magnitude. Therefore, when multiple fossil calibrations and a birth-death tree prior are used with a strict clock, they do not influence divergence time estimates by affecting estimates of branch-specific rates. By contrast, when multiple fossil calibrations and a birth-death tree prior are used with a relaxed clock, they can influence divergence time estimates by affecting estimates of branch-specific rates, but in a manner that is constrained by the model of among-branch-rate-variation. Meanwhile, when multiple fossil calibrations and a birth-death tree prior are used with no molecular data, they can influence divergence time estimates in a manner that is completely unconstrained by assumptions about rates. In this context, it is unsurprising that in some respects, analyses with a relaxed clock are intermediate between a strict clock and no molecular data.

This explanation for the observation that the relaxed clock is intermediate between the strict clock and no molecular data is relevant for explanations of additional patterns that we discuss below. It also provides a clearer basis for understanding the context specificity of more general comparisons between MPEs when using different calibration strategies (discussed above). This is because with different molecular clock strategies, the relative importance of fossil calibrations and the birth-death tree prior will vary. Further, the manner by which the importance of the fossil calibrations and birth-death tree prior varies will depend on where the fossil calibrations are implemented, and the tree topology.

Difference between the MPE and age of the node-calibration-fossil (the oldest fossil within a clade)

When comparing node calibration strategies that use either non-uniform calibration densities (NUP and NUI) or uniform calibration densities (UP and UI), the strategies that assume actual clade ages are more similar to the ages of node-calibration-fossils (NUP or UP) lead to smaller differences between MPEs and the ages of the node-calibration-fossils than strategies that assume actual clade ages may be considerably older (NUI or UI) (Table S13-S18). We observe this pattern regardless of the molecular clock strategy, thus highlighting the importance of the calibration strategy in all contexts.

However, the relationship between MPEs and ages of node-calibration-fossils is strongly influenced by the molecular clock strategy, and interactions between specific calibration strategies and molecular clock strategies. For example, with NUP and NUI, the strict clock results in the largest differences (Table S14, S17), followed by the relaxed clock (Table S13 and S16), followed by no molecular data (Table S15 and S18).

We explain this pattern within the same framework as for the effects of different molecular clock strategies on MPEs. First, consider the strict clock – which infers the same rate for every branch, and NUP and NUI – which specify small but non-zero probabilities that clades are markedly older than node-calibration-fossils (Fig. 3a and b). In this context, MPEs that are vastly older than

node-calibration-fossils can be the most feasible divergence time estimates. By contrast, relaxed clocks infer different rates for each branch, enabling MPEs to correspond more closely to the specified calibration density. This results in smaller differences between MPEs and the ages of node-calibration-fossils, especially for NUP, which assumes the actual clade age is very similar to the age of the node-calibration-fossil. Meanwhile, with no molecular data, assumptions about rates confer no constraints on MPEs. As such, MPEs correspond exactly (or almost exactly after accounting for the effect of the tree prior) to the specified calibration density. This result is consistent with that of Brown and Smith (2018), who showed that with a relaxed clock, posterior age estimates at calibrated nodes are often very similar to age estimates inferred with no molecular data.

In relation to this pattern, in CS with a strict clock and NUP or NUI, the difference is far larger at the Solanaceae stem node than the Solanoideae stem node (Table S17). The difference at the Solanaceae stem node then becomes progressively smaller with a relaxed clock (Table S16) and no molecular data (Table S18). We suggest this reflects our analysis which shows the ESF for the node-calibration-fossil at the Solanoideae stem node is higher than at the Solanaceae stem node (Fig. 6b, Table S7). With a strict clock, the higher ESF at the Solanoideae stem node means the entire phylogeny is calibrated to a timescale that is far older than the node-calibration-fossil at the Solanaceae stem node. By contrast, with a relaxed clock or no molecular data there is no assumption of rate constancy, meaning the calibration at the Solanoideae stem node does not influence MPEs throughout the phylogeny in as consistent a manner. The MPE for the Solanaceae stem node therefore corresponds more closely to the calibration density specified by NUP or NUI.

This explanation for the difference in age between the node calibration fossil and the MPE for the Solanaceae stem node can also be interpreted in the context of inferred substitution rates. Where a strict clock was used with CS and NUP, the inferred substitution rate (MPE) for the entire phylogeny is $8.503*10^{-4}$ *subsitutions* Myr^{-1}. By contrast, when a relaxed clock was used with NUP, the inferred substitution rate for the Solanaceae stem branch is $3.222*10^{-3}$ *substitutions* Myr^{-1}. This considerably faster substitution rate will mean that the Solanaceae stem branch is shorter. This in turn causes the Solanaceae stem node to be younger, and closer in age to the fossil calibration at the Solanaceae stem node.

With UP and UI our findings are more complicated because of three characteristics of these calibration strategies. First, UP and UI do not make an explicit statement about the probable clade age in relation to the node-calibration-fossil, other than it lies somewhere between the minimum and maximum. Second, UP and UI implement a hard-maximum constraint, the clade can be no older than this maximum. Third, the same maximum is used for several different nodes because robust maximum constraints are difficult to justify (Marshall 2008; Warnock et al. 2015, Betts et al. 2018; Morris et al. 2018). For nodes with younger node-calibration-fossils, the range between the minimum and maximum is therefore often larger.

These differences have several implications. First, a relaxed clock or no molecular data does not cause MPEs to correspond to the ages of node-calibration-fossils in a predictable manner, as is the case for NUP and NUI (Table S13-18). This is because UP and UI do not make an explicit statement about probable clade ages in relation to node-calibration-fossils, and because there are different sized intervals between minimum and maximum constraints at each node.

With UI, and for node-calibration-fossils with lower ESFs, there are larger differences between the ages of node-calibration-fossils and MPEs (Table S5, S7, S13-S18). For older node-calibration-fossils, these differences become progressively smaller with a relaxed clock and no molecular data compared to a strict clock. We explain this according to the same framework as set out for NUP and NUI in CS. Specifically, relaxing the assumption of rate constancy means that node-calibration-fossils with higher ESFs do not calibrate the entire phylogeny to a timescale that is older than node-calibration-fossils with lower ESFs. However, for younger node-calibration-fossils, these differences do not become progressively smaller with a relaxed clock and no molecular data compared to a strict clock. This is because of the larger range between the minimum and maximum constraints at the nodes that these fossils calibrate, and the ambiguity about the clade age within this range.

With UP and for older node-calibration-fossils with lower ESFs, there are less noticeable differences between molecular clock strategies (Table S5, S7, S13-18). The younger hard maximum prevents MPEs from being markedly older than the node-calibration-fossil, regardless of molecular clock strategy or ESF. For UP and younger node-calibration-fossils with lower ESFs, different molecular clock strategies also have a limited effect, as was also the case with UI. However, the magnitude of the differences is smaller than for UI.

95% HPD widths

When comparing node calibration strategies that use either non-uniform calibration densities (NUP and NUI) or uniform calibration densities (UP and UI), the strategies that assume actual clade ages are more similar to the ages of node-calibration-fossils (NUP or UP) lead to narrower 95% HPDs than strategies that assume actual clade ages may be considerably older (NUI or UI) (Table S19 and S20). We observe this pattern regardless of the molecular clock strategy, thus highlighting the importance of the calibration strategy in all contexts.

In all cases, 95% HPDs are narrowest with a strict clock, and progressively wider with a relaxed clock and no molecular data (Table S19 and S20). We explain this within the same framework as our previous discussion of why the relaxed clock is intermediate between the strict clock and no molecular data. With weaker assumptions about rates for specific branches (from the strict clock, to relaxed clock to no molecular data), there is a larger range of probable divergence time estimates, and hence wider 95% HPDs.

The FBDP

Analyses with the FBDP allow us to compare this newly developed method to conventional node calibration methods. Although these analyses are more limited in scope, we identify several similar characteristics between these two methods. First, 95% HPD widths and MPEs are highly sensitive to molecular clock strategy (Table S9 and S19, Fig. 10e and f). Although the details of this relationship are complex for MPEs, and depend on the phylogenetic depth at which a comparison is made (Fig. 10e and f), for 95% HPD widths the picture is simple. 95% HPDs are broader with a relaxed clock (Table S19). This is consistent with node calibration methods, and we propose the same explanation as previously, whereby weaker assumptions about rates lead to a larger range of probable divergence time estimates.

A second key similarity is that with a relaxed clock, inferred divergence times are more consistent with assumptions about the fossil record. With node calibration methods, we make this conclusion because with a relaxed clock, inferred ages for calibrated nodes correspond more closely to the specified calibration density (Table S13-S18). For the FBDP, we make this conclusion because with a relaxed clock, in clades with a lower inferred ψ, divergence time estimates tend to be closer to the age of the oldest fossil (and are therefore younger), and in clades with a higher inferred ψ, divergence time estimates tend to be less close to the age of the oldest fossil (and are therefore older) (Table S13 and S14, Fig. 7a). We suggest this observation reflects the assumption that ψ is constant. With a relaxed clock, the inference of different rates for each branch means that younger ages are inferred for clades with a lower inferred ψ, therefore "smoothing" differences in ψ among clades. By contrast, the opposite occurs in clades with a higher inferred ψ. There are nonetheless, some exceptions to this pattern. These are likely to be because the inferred ages of different clades in a phylogeny are not independent of each other.

Relaxed clocks and multiple fossil calibrations do not reliably lead to more accurate divergence time estimates

The development of methods for integrating multiple fossil calibrations with relaxed clocks has led to increased agreement between the ages of known fossils and inferred divergence times. This is often interpreted as scientific "progress", with methodological developments enabling integration of temporal evidence from the fossil record with molecular sequence data to construct a more realistic timeframe for the tree of life (Donoghue and Benton 2007; Grimm et al. 2015; Hughes et al. 2015; Magallón et al. 2015; Donoghue and Yang 2016; Betts et al. 2018; Morris et al. 2018). Here, we show that this positive characterisation is undermined in two ways. First, the nature of the fossil record is inconsistent with the assumptions of current methods. Second, methods that integrate multiple fossil calibrations with relaxed clocks are highly sensitive to these assumptions.

For node calibration methods, we conclude that the nature of the fossil record is inconsistent with the assumptions of current methods from our analyses which highlight that multiple node-

calibration-fossils are unlikely to reflect the relative ages of different clades. For the FBDP, we make this conclusion from our analyses which highlight that ψ is highly heterogeneous.

Subsequently, our comparison of inferred divergence times when using multiple fossil calibrations with different methods highlights the sensitivity of methods to underlying assumptions. We show that divergence time estimates are highly sensitive to the calibration strategy, regardless of molecular clock strategy. We also show that divergence time estimates are sensitive to the molecular clock strategy, with context specific interactions between calibration strategies and molecular clock strategies.

Further, an important general pattern is that analyses with relaxed clocks are intermediate between strict clocks and no molecular data. This pattern is important because it indicates that analyses which use multiple fossil calibrations with relaxed clocks are often unlikely to provide novel or robust insights. Instead, the molecular sequence data does not provide direct information about rate and time. Therefore, in analyses with relaxed clocks whereby different rates are inferred on individual branches, divergence time estimates will be sensitive to implemented fossil calibrations, the birth-death tree prior, and the constraints of the relaxed clock model. By contrast, when no molecular data is analysed, divergence time estimates will solely be influenced by implemented fossil calibrations and the birth-death tree prior.

Although the relaxed clock analyses presented here were performed with a UCLN relaxed clock, we suggest that the principles of our results are applicable when alternative models of among-branch-rate-variation are used to infer divergence times, such as the autocorrelated relaxed clock (Thorne et al. 1998; Kishino et al. 2001; Drummond et al. 2006; Drummond and Suchard 2011; Lartillot et al. 2016). Although these alternative models make different assumptions about among-branch-rate-variation, and as such may infer different divergence times, it remains the case that regardless of the model, molecular sequence data does not provide direct information about rates. The implementation of any given relaxed clock model will therefore remain sensitive to implemented fossil calibrations, and relaxed clock analyses will likely remain intermediate between a strict clock and no molecular data. The difference is that with alternative relaxed clock models, divergence time estimates will be sensitive to the alternative constraints that are specific to a given relaxed clock model.

A further point concerning alternative relaxed clock models, is that even though different models may be more appropriate for different datasets, it remains difficult to determine the most appropriate model. Model selection methods are extremely computationally intensive, and as such may be conducted on a small subset of the original dataset (for example Barba-Montoya et al. 2018). Further, given that the rate is not extracted directly from molecular sequence data, model selection methods will also likely remain sensitive to the methodological context such as the implemented fossil calibrations and the birth-death tree prior.

Some aspects of the findings reported here are consistent with previous studies that highlight the importance of either the calibration strategy (Hedges and Kumar 2004; Near and Sanderson 2004; Near et al. 2005; Yang and Rannala 2006; Marshall 2008; Heath 2012; Gavryushkina et al. 2014, 2017; Heath et al. 2014; Warnock et al. 2011, 2015, 2017; Zhang et al. 2016; Barba-Montoya et al. 2017) or molecular clock strategy for divergence time estimates (Thorne et al. 1998; Kishino et al. 2001; Drummond et al. 2006; Drummond and Suchard 2011; Zhu et al. 2015; Lartillot et al. 2016) . However, the comprehensive nature of this study means the findings we present are fundamentally distinct from previous studies that consider specific aspects of divergence time estimation in relative isolation. Here, we critically evaluate the validity of assumptions that are made about the fossil record when implementing relaxed clocks, and evaluate in detail the implications of different calibration strategies and molecular clock strategies. Crucially, our findings also provide an in-depth insight into the interactions between the calibration strategy and molecular clock strategy. Further, many previous methodological studies focus on specific groups with well-preserved fossil records, despite the fact that the majority of divergence time analyses – especially in plants – are performed in groups with considerably poorer fossil records. The findings presented here, that highlight the important implications of different assumptions in groups with poor fossil records, will therefore be uniquely pertinent to many future divergence time analyses.

The purpose of this study is not to directly challenge the results of previous studies. Instead, by showing that current methods are highly sensitive to a set of assumptions that are often violated, we suggest that methods that use multiple fossil calibrations with relaxed clocks often do not produce robust inferences. As such, it may often be the case that analyses that use multiple fossil calibrations with relaxed clocks do not produce meaningful inferences that add significantly to those that can be made by simply analysing the fossil record. We therefore challenge a common assertion that highly integrated analyses that use multiple fossil calibrations with relaxed clocks inherently lead to more meaningful, robust, and novel inferences (Heath et al. 2014; Grimm et al. 2015; Lee and Palci 2015; Magallón et al. 2015; Renner et al 2016; Yang and Donoghue 2016; Betts et al. 2018; Morris et al. 2018).

Implications for macroevolutionary studies

The findings presented in this paper have concerning implications for macroevolutionary studies, because many macroevolutionary studies are heavily dependent on accurate time-calibrated phylogenies. This raises important questions about how time-calibrated phylogenies should best be inferred and interpreted. One potential option for node calibration methods, could be to implement calibration densities that are markedly broader than in many previous studies, or maximum age constraints that are markedly older. This approach was followed in the NUI and UI calibration strategies discussed here. However, such an approach makes extremely weak assumptions about times such that it is difficult to meaningfully implement different molecular clock strategies. Further, 95%

HPDs (or equivalent confidence intervals) would be extremely broad. Using inferred time-calibrated phylogenies as a basis for hypothesis testing would therefore likely become highly problematic.

Alternatively, a potential option for the FBDP could be to relax the assumption that ψ is constant when inferring divergence times. Although work continues on developing extensions to the FBDP (for example Stadler et al. 2018), no method for estimating divergence times which enables ψ to vary has been implemented. Further, the prospects of developing a method of divergence time estimation which enables ψ to vary are likely to be limited because it would cause fundamental issues with model identifiability. Further, any such method would likely be undermined by the continued assumption that the speciation rate and extinction rate are constant in the branching process.

Rather than hoping for methodological developments which may be theoretically impossible, a useful approach for future studies may be to change the way in which time-calibrated phylogenies are interpreted. Instead of using time-calibrated phylogenies as a basis to construct general narratives for the evolution of a particular group (for example Lagomarsino et al. 2016; Cardillo et al. 2017; Folk et al. 2019), a more robust approach may be to test specific evolutionary hypotheses. With such an approach analyses can be constructed that account for the biases inherent in different methods and provide a robust test of a particular hypothesis. This approach was followed in a recent study of the sweet potato, where calibration strategies biased divergence time estimates to younger ages to provide the most robust test of whether the sweet potato evolved in pre-human times (Munoz-Rodriguez et al. 2018, 2019, Carruthers et al. 2020). In this context, no assumptions are required about whether or not a given time-calibrated phylogeny is most accurate, a question which is often impossible to answer.

MATERIALS AND METHODS

Molecular Data

The matrix for Osmundaceae is the same as that used by Grim et al. (2015) and originally compiled by Metzgar et al. (2008). It includes 13 taxa sampled for three chloroplast genes (*rbcL*, *atpA* and *rps4*) and five spacer regions (*rps4-trnS, trnG-trnR, trnL-trnF, rbcL-accD, atpB-rbc*L). The matrix for CS contains 56 taxa sampled for four chloroplast genes (*atpB*, *matK*, *rbcL* and *ndhF*), the chloroplast intergenic spacer *trnL-trnF*, and the nuclear marker *ITS* (Table S21). The matrix for Spermatophyta is based on that of Magallón et al. (2015). It consists of 303 taxa sampled for two nuclear genes (*18S* and *26S*) and three chloroplast genes (*matK*, *rbcL* and *atpB*) (Table S22).

Fossil Data

For Osmundaceae, we analysed a dataset of 19 rhizome fossils and 17 frond fossils. Grimm et al. (2015) originally compiled this dataset, and assigned these fossils to one of five clades: Osmundaceae, *Osmunda*, *Osmunda japonica* – *Osmunda vachellii* (JV), *Osmunda japonica* – *Osmunda regalis* (JR), and *Todea barbara* – *Leptopteris wilkesiana* (BW). The 19 rhizome fossils possess sufficient morphological characters to be included in a 33-character morphological matrix (Bomfleur et al. 2015).

For CS, we analysed a dataset of 32 seed fossils that was originally compiled by Särkinen et al. (2013). Särkinen et al. (2013) assigned these fossils to either Solanaceae (including the stem branch), or Solanoideae (including the stem branch). Four additional fossils have putatively been assigned to Convolvulaceae (MacGintie 1954; Martin 2000; Martin 2001; Mitchell et al. 2016; Srivastava et al. 2018), and a further fossil that is drastically older than previously known fossils has recently been assigned to Solanoideae (Wilf et al. 2017). These fossils were not included in this study for two reasons. First, they cannot be conclusively assigned to specific clades within CS. Second, the purpose of this study is to investigate the implications of fossil calibrations in a way that is directly relevant and comparable to methods implemented in previous studies. These fossils have not been used as calibrations in previous studies.

For Spermatophyta, we analysed a dataset of 112 fossils that was based on that of Magallón et al. (2015). This dataset contains fossils that belong to clades throughout Spermatophyta (Table S3).

Phylogenetic inference

We constructed molecular phylogenies for extant sampled taxa. These would serve as starting topologies for subsequent analyses. For Osmundaceae, the same alignment was used as Grimm et al. (2015). For CS and Spermatophyta, sequences were aligned with MAFFT v7.271 (Katoh 2002; Katoh and Standley 2013), using the L-INS-I setting and a gap opening penalty of 1.53. Ambiguously aligned regions were removed using default settings in Gblocks (Castresana 2000; Talavera and Castresana 2007) and aligned sequences for each marker were concatenated using Sequence Matrix v1.8 (Vaidya et al. 2011).

We used mrBayes v3.2.6 (Huelsenbeck and Ronquist, 2001; Ronquist et al. 2012) for phylogenetic analyses. A GTR + G + I model of sequence evolution was used. In all analyses two independent runs were performed. For Osmundaceae and CS, the analysis was run for 2,000,000 generations (sampling every 200 generations). The final split frequency between runs was < 0.01 in both cases. For Spermatophyta, the analysis was run for 4,500,000 generations (sampling every 2,000 generations). The final split frequency between runs was 0.0185.

Confidence intervals for clade ages

Constructing confidence intervals

For Osmundaceae and CS, the fossil datasets represent the entire known fossil record (Särkinen et al. 2013; Grimm et al. 2015). We could therefore calculate confidence intervals for the ages of clades within these two groups using approaches analogous to previously developed methods for estimating taxon ages (Marshall 1990 and 2008; Springer 1995). The primary purpose of these confidence intervals is not to provide robust estimates of the actual ages of different clades, or to provide a useful basis upon which to derive fossil calibrations in divergence time analyses. Instead, the purpose is to use available evidence (the temporal distribution of all fossils within a cade) as a basis for estimating the extent to which different clades are potentially older than their oldest fossils.

This provides a basis for determining the extent to which the fossil record enables confident statements to be made about the relationship between clade ages and fossil ages, and subsequently the implications of this for estimating divergence times.

For Osmundaceae, we calculated confidence intervals for: the entire of Osmundaceae, *Osmunda*, and JV. We could not calculate confidence intervals for JR or BW because only one fossil was sampled from these clades. For CS, we calculated confidence intervals for the entire of Solanaceae (including the stem branch) and Solanoideae (including the stem branch).

To calculate confidence intervals, we first performed Kolmogrov-Smirnov tests to determine the distribution that fossil occurrence times within these two clades conform to. We tested two distributions: a uniform distribution in which the minimum was the present and the maximum was the age of the oldest fossil, and an exponential distribution for which the mean was equal to the mean age of all fossils within the clade.

We interpret the best fitting distribution as a description of how fossils are sampled through time from a given clade. An interval that is older than the oldest fossil for a clade, and within which there would be a 95% probability (according to our inferred distribution) of sampling a fossil were the clade still in existence, represents a 95% confidence interval for the actual clade age.

For the uniform distribution, the upper limit of the 95% confidence interval for the actual clade age (CI_{max}) is given by: $CI_{max} = \frac{OF}{\sqrt[n]{1-p}}$. OF is the age of the oldest fossil, n is the number of fossils and p is the confidence level (in this case 0.95). For the exponential distribution, CI_{max} is given by: $CI_{max} = OF + Q(p)$. $Q(p)$ is the quantile function of the part of the exponential distribution that is older than OF. p is the confidence level (in this case 0.95).

When calculating confidence intervals, we took into account uncertainty in fossil ages. Using a custom R script, we performed 10,000 replicate calculations. In each replicate, the age of each fossil was a drawn randomly from its stratigraphic range. When plotting the upper bound of the 95% confidence interval, the thickness of the bar represents the interval between the 2.5th and 97.5th percentiles following 10,000 replicate calculations.

When analysing the fossil record of CS, the high proportion of Pliocene fossils made fitting a distribution problematic. We therefore performed additional analyses where a subset of the Pliocene fossils were sampled. We inferred appropriate distributions and calculated confidence intervals based on this subsample. We suggest this subsampling approach is robust for two reasons. First, these young Pliocene fossils are unlikely to provide useful information about the ages of clades within CS because they are far younger than many other fossils within their respective clades. Second, by removing these fossils, we could more accurately model the temporal distribution of older fossils within their respective clades. Regardless, to ensure our inferences were robust, we also inferred distributions and calculated confidence intervals with the entire fossil dataset.

Validating Confidence Intervals

We performed simulation experiments to validate the confidence intervals we calculated. Fossil occurrences were simulated for clades of different ages (from 25 Myr to 475 Myr, in intervals of 50 Myr) and according to the distributions that fossil occurrences conformed to in the empirical datasets discussed above (only exponential distributions) (Table S4 and S6). For a clade of a given age, different numbers of fossils were simulated (from 2 to 40, in intervals of 2), and the mean age of the distribution from which fossils were simulated was altered (from 5 to 85, in intervals of 10). The simulated dataset of fossils was then used as a basis to calculated confidence intervals for clades of different ages (with the true clade age now known). We could therefore determine whether the confidence interval contained the true clade age. For each clade age, number of fossils, and mean age of the distribution from which fossils were simulated, the simulation experiment was repeated 500 times.

To compare the validity of the confidence intervals we calculated to previously implemented calibration densities, we performed a further set of simulation experiments. In these experiments, fossils were simulated for different clades as described previously. The node-calibration-fossil for each clade was then used as an offset for a lognormal distribution with $\mu = 0.01$, and $\sigma = 1$, the same calibration density as implemented by Särkinen et al. (2013). We could therefore determine whether the 95th percentile of the calibration density contained the true clade age.

Calculating ESFs for node-calibration-fossils

For all three datasets, we calculated ESFs for node-calibration-fossils. This followed the approach of Marshall (2008). With FA being the absolute age of the fossil, and D being the depth of the node in the ultrametric phylogeny to which the node-calibration-fossil is assigned, $ESF = \frac{FA}{D}$. We inferred the ultrametric phylogenies necessary for calculating ESFs in the penalised likelihood framework treePL (Smith and O'Meara 2012), and in the Bayesian inference framework RevBayes (Höhna et al. 2016).

In treePL, we used the trees inferred in mrBayes v3.2.6 as input trees and cross-validation to determine the optimum smoothing value. The smoothing value describes the extent to which rate differences between ancestral and descendant branches are penalised when inferring the ultrametric phylogeny, with a higher smoothing value penalising rate differences more than a lower smoothing value. For cross-validation analyses, we tested smoothing values of 0.01, 1, 100, and 10,000. For Osmundaceae, the optimum smoothing value of 10,000 was used to infer the ultrametric phylogeny, for CS, the optimum smoothing value of 0.01 was used to infer the ultrametric phylogeny, and for Spermatophyta, the optimum smoothing value of 0.01 was used to infer the ultrametric phylogeny. In treePL, we also used the "thorough" option to perform more thorough optimisation analyses, and the "prime" option to determine the most appropriate optimisation parameters for each analysis.

In RevBayes, we used the trees inferred in mrBayes v3.2.6 as input topologies. A GTR + G + I model of sequence evolution was used for the molecular data, with either a strict molecular clock or

a UCLN relaxed clock. The mean of the UCLN relaxed clock was sampled from a uniform distribution spanning 7 orders of magnitude, whilst the standard deviation of the UCLN relaxed clock incorporated 0.5 orders of magnitude. The molecular sequence data was analysed as a single partition for the sequence evolution model and branch rate model. This is because the ultrametric phylogeny only provides a rough approximation of the relative ages of different clades (see below), with inferences sensitive to the inherent limitations of any given branch rate model and branching process (see below). Further, assigning partitions for branch rate models can be especially problematic and may do little to improve the robustness of inferences (eg. dos Reis et al. 2014; Carruthers et al. 2019). This is likely to result from the fact that molecular sequence data does not provide direct information about rates. Regardless of this, in preliminary analyses, we also found that different partitioning schemes had a negligible effect on inferred ultrametric phylogenies. A birth-death branching process was used as the tree prior. Sufficient mixing and convergence was assessed in Tracer v1.6.0. A 25% burn-in was used prior to calculating mean posterior estimates (MPEs) for node heights, and 95% highest posterior density intervals (HPDs) for node heights.

The rationale for comparing ESFs rests on the assumption that the relative *depths* of different nodes in an ultrametric phylogeny reflect the relative *ages* of different clades. If this is the case, a comparison of the age of a set of node-calibration-fossils relative to the depth of the node to which they are assigned ($\frac{FA}{D}$) provides a measure of variation in how different node-calibration-fossils reflect the actual age of their respective clade.

In order for the relative depths of different nodes to reflect the relative ages of different clades, among-branch-rate-variation must be estimated accurately when inferring the ultrametric phylogeny. Given molecular sequence data only provides direct information about the expected number of substitutions on each branch, this is a problematic requirement, and methods for constructing ultrametric phylogenies are likely to be sensitive to underlying assumptions.

In this study, we primarily focus on the ultrametric phylogenies inferred in treePL because treePL does not implement the constant rate birth-death tree prior (characteristic of Bayesian analyses), that may, through the assumptions it makes about times, significantly bias inferences of among-branch-rate-variation. We consider this problem may be especially important in the large Spermatophyte dataset. Regardless, the assumption of among-branch-rate-variation made by treePL, whereby rates are inherited between ancestral and descendant branches, is one of potentially many assumptions that could be made about among-branch-rate-variation. It is therefore important to treat our results with caution, and to compare findings derived from ultrametric phylogenies inferred in treePL with those inferred with other methods. Regardless, despite the limitations of analyses of ESFs, in the context of other analyses such as the calculation of confidence intervals for clade ages, they can provide important insights.

Calculating the fossil recovery rate ψ

We also used the ultrametric phylogenies as a basis to estimate ψ within Osmundaceae and CS. Depending on the phylogenetic placement of fossils, we inferred ψ within different clades within these two groups, but in some cases also within different grades. With n being the number of fossils assigned to a clade or grade, and BL being the total length of all branches within the clade or grade in the ultrametric phylogeny, $\psi = \frac{n}{BL}$.

Our calculation of ψ is directly relevant to how ψ is formalised in the FBDP – where it describes the rate that fossils are sampled along the branches of the macro-evolutionary process that generated sampled extant taxa and sampled fossils. In some formalisations of the FBDP, fossils are assigned to more specific parts of the tree than they are here. Further, we do not explicitly consider the effect of sampling fossils along extinct branches. However, we suggest the estimates presented here provide a useful approximation of the extent to which ψ is likely to vary.

Inferring divergence times with different molecular clock strategies and different calibration strategies

Node calibrations

For Osmundaceae, we used four methods for implementing node calibrations: NUP, NUI, UP and UI. NUP was the same as the strategy used by Grimm et al. 2015. Each of the five fossil calibrations was implemented as an exponential distribution. The offset was equal to the minimum age of the stratum containing the node-calibration-fossil, and the 97.5th percentile was equal to the maximum age of the stratum. In NUI, each of the five fossil calibrations was implemented as a lognormal distribution. The offset was equal to the minimum age of the stratum containing the node-calibration-fossil, the mean (μ) was parameterised such that the 95th percentile of the calibration density equals the upper limit of the 95% confidence interval that we calculated for the clade's age, and the standard deviation (σ) was 1. In UP, the fossil calibrations were implemented as uniform distributions. The minimum was equal to the minimum age of the stratum containing the node-calibration-fossil, and the maximum for all calibrations was 299 Myr. This corresponds to a previously estimated divergence time between osmundaceous ferns and other leptosporangiate ferns (Schuettpelz and Pryer 2009). In UI, the fossil calibrations were also implemented as uniform distributions. The minimum age was equal to the minimum age of the stratum containing the node-calibration-fossil, and the maximum for all calibrations was 472 Myr. This corresponds to the age of the oldest known cryptospores (Rubinstein et al. 2010).

For CS, we also used four methods for implementing node calibrations: NUP, NUI, UP, and UI. NUP was the same as the strategy used by Särkinen et al. (2013). The two fossil calibrations were implemented as lognormal distributions. The offset was equal to the minimum age of the stratum containing the node-calibration-fossil, μ = 0.01, and σ = 1. NUI was parameterised according to the same approach as Osmundaceae, but we used the confidence intervals for clade ages that were calculated when 13 pliocene fossils were removed. In UP, fossil calibrations were implemented as

uniform distributions. The minimum was equal to the minimum age of the stratum containing the node-calibration-fossil, and the maximum for all calibrations was 86.9 Myr. This corresponds to the upper limit of the 95% HPD for the divergence time between Convolvulaceae and Solanaceae in Magallón et al. (2015). In UI, fossil calibrations were also implemented as uniform distributions. The minimum was equal to the minimum age of the stratum containing the node-calibration-fossil, and the maximum for all calibrations was 130 Myr. This corresponds to the maximum age of the stratum in which tricolpate pollen appears in the fossil record (Doyle et al. 1977).

Time-calibrated phylogenies were inferred in RevBayes v1.0.4. The topology was fixed to that previously inferred in mrBayes v3.2.6. A GTR + G + I model of sequence evolution was used in all cases. Analyses were performed with a strict clock, UCLN relaxed clock, and no molecular data. The rate for the strict clock and mean of the UCLN relaxed clock were sampled from a uniform distribution spanning 7 orders of magnitude. The standard deviation of the UCLN relaxed clock incorporated 0.5 orders of magnitude. The molecular sequence data was analysed as a single partition for the sequence evolution model and branch rate model. More complex partitioning schemes may enable aspects of molecular evolution to be modelled more accurately. However, the purpose of the analyses presented here is to provide a general framework for discussing the implications of different molecular clock strategies and calibration strategies, rather than estimating the most accurate timescale for the evolution of Osmundaceae and CS. Further, as discussed with respect to inferring ultrametric phylogenies, assigning partitions for branch rate models can be problematic (eg. dos Reis et al. 2014; Carruthers et al. 2019). A constant rate birth-death branching process was used as the tree prior. For each analysis, two independent runs were performed. Sufficient mixing and convergence was assessed in Tracer v1.6.0 (Rambaut et al. 2014). Most ESS values exceeded 500, and all exceeded 200. A 25% burn-in was used prior to calculating mean posterior estimates (MPEs), and 95% highest posterior density intervals (HPDs).

FBDP

We performed FBDP analyses for Osmundaceae. These analyses included all 36 Osmundaceae fossils, and the 33 character morphological matrix for the 19 rhizome fossils and 13 extant species. We incorporated the frond fossils by using the unresolved FBDP (Heath et al. 2014). This approach takes into account uncertainty in the phylogenetic position of the frond fossils, whilst allowing the analysis to be constrained to reflect existing knowledge of their phylogenetic position. The FBDP analyses were performed in RevBayes v1.0.4. In all cases the prior for ψ was an exponential distribution (FBDL mean = 0.02, for FBDS mean = 0.1, FBDH mean = 1). Trials were tested with lower mean values for the ψ prior but these analyses failed to mix correctly. A GTR + G + I model of sequence evolution was used for the molecular data, with either a strict molecular clock (FBDS and FBDH) or a UCLN relaxed clock (FBDL, FBDS, and FBDH). These were parameterised as outlined for node calibrations. A strict clock was not used with FBDL because it did not mix

properly in trial analyses. A symmetrical Markov model was used for the morphological data, with a strict clock that assumes morphological characters evolve at the same rate on all branches. The rate for the morphological strict clock was also sampled from a uniform distribution spanning 7 orders of magnitude. For each analysis, two independent runs were performed. Sufficient mixing and convergence was assessed in Tracer v1.6.0. A 25% burn-in was used prior to calculating mean posterior estimates (MPEs), and 95% highest posterior density intervals (HPDs).

We did not consider that an analysis with no molecular data represented an appropriate or useful critique of the FBDP. This method is highly integrated, with the times and morphological characters of fossils, and molecular sequence data of extant taxa, being modelled in a single macro-evolutionary framework. As such, it would be difficult to analyse the precise effects of not including molecular sequence data. The highly integrated nature of this method is also likely to explain why, in initial experiments, we found that implementing it without molecular sequence data was highly problematic.

We did consider implementing the FBDP in CS. This would necessarily use the unresolved FBDP because there is no appropriate morphological matrix available for known fossils (Heath et al. 2014). However, in initial experiments, we found that implementing this method with this dataset was problematic. Future, developments of this new method may enable it to function more effectively when analysing large numbers of fossils with limited morphological characters.

DATA AVAILABILITY

All sequence alignments, and custom R and RevBayes scripts developed for this study are available on Dryad XXX.

FUNDING

This work was supported by a NERC scholarship granted through the Environmental Research DTP programme to TC.

REFERENCES

Baldwin BG and Sanderson MJ. 1998. Age and rate of diversification of the Hawaiian silversword alliance (Compositae). *Proc Natl Acad Sci USA.* 95:9402-9406.

Barba-Montoya J, dos Reis M, Yang Z. 2017. Comparison of different strategies for using fossil calibrations to generate the time prior in Bayesian molecular clock dating. *Mol Phylogenetics Evol.* 114:386-400.

Barba-Montoya J, dos Reis M, Schneider H, Donoghue PCJ, Yang Z. 2018. Constraining uncertainty in the timescale of angiosperm evolution and the veracity of a Cretaceous Terrestrial Revolution. *New Phytol.* 218:819-834.

Benton MJ. 1995. Testing the time axis of phylogenies. *Phil Trans R Soc Lond B.* 349:5-10.

Betts HC, Puttick MN, Clark JW, Williams TA, Donoghue PCJ, Pisani D. 2018. Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nat Ecol Evol.* 2:1556-1562.

Bomfleur B, Grimm GW, McLoughlin S. 2015. *Osmunda pulchella* sp. nov. from the Jurassic of Sweden - reconciling molecular and fossil evidence in the phylogeny of modern royal ferns (Osmundaceae). *BMC Evol Biol.* 15:126.

Britten RJ. 1984. Rates of DNA sequence evolution differ between taxonomic groups. *Science*. 39:1393–1398.

Britton T. 2005. Estimating Divergence Times in Phylogenetic Trees Without a Molecular Clock. *Syst Biol*. 54:500–507.

Bromham L. 2006. Molecular dates for the Cambrian Explosion: is the light at the end of the tunnel an oncoming train? *Palaeontol Electron.* 9:2004–2006.

Brown JW and Smith SA. 2018. The Past Sure is Tense: On Interpreting Phylogenetic Divergence Time Estimates. *Syst Biol*. 67:340-353.

Carruthers T, Sanderson MJ, Scotland RW. 2019. The implications of lineage-specific rates for divergence time estimation. *Syst Biol.* Advance access online.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 17:540–552.

Cardillo M, Weston PH, Reynolds ZKM, Olde PM, Mast AR, Lemmon EM, Lemmon AR, Bromham L. 2017. The phylogeny and biogeography of *Hakea* (Proteaceae) reveals the role of biome shifts in a continental plant radiation. *Evolution.* 71:1928-1943.

Donoghue MJ., Bell CD., Li J. 2001. Phylogenetic Patterns in Northern Hemisphere Plant Geography. *Int J Plant Sci.* 162:S41-52.

Donoghue PCJ and Benton MJ. 2007. Rocks and clocks: calibrating the Tree of Life using fossils and molecules. *Trends Ecol Evol.* 22:424-431.

Donoghue PCJ and Yang Z. 2016. The evolution of methods for establishing evolutionary timescales. *Phil Trans R Soc B.* 371:20160021.

Doyle J. A., Biens P., Dorenkamp A., Jardiné S. 1977. Angiosperm pollen from the pre-Albian Cretaceous of Equatorial Africa. *Bull. Cent. Rech. Explor.* 1:451-473.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed Phylogenetics and Dating with Confidence. *PLOS Biol.* 4:e88.

Drummond AJ and Suchard MA. 2010. Bayesian random local clocks, or one rate to rule them all. *BMC Biol.* 8:114.

Folk RA, Stubbs RL, Mort ME, Cellinese N, Allen J, Soltis PS, Soltis DE, Guralnick RP. 2019. Rates

of niche and phenotype evolution lag behind diversification in a temperate radiation. *Proc Natl Acad Sci USA.* 116: 10874-10882

Gandolfo MA, Nixon KC, Crepet WL. 2008. Selection of Fossils for Calibration of Molecular Dating Models. *Ann Missouri Bot Gard.* 95:34-42.

Gavryushkina A, Welch D, Stadler T, Drummond AJ. 2014. Bayesian Inference of Sampled Ancestor Trees for Epidemiology and Fossil Calibration. *PLOS Comput Biol.* 10:e1003919.

Gavryushkina A, Heath TA, Ksepka DT, Stadler T, Welch D, Drummond AJ. 2017. Bayesian Total-Evidence Dating Reveals the Recent Crown Radiation of Penguins. *Syst Biol.* 66:57-73.

Gillespie JH. 1989. Lineage effects and the index of dispersion of molecular evolution. *Mol. Biol Evol.* 6:636-647.

Gillespie JH. 1991. The Causes of Molecular Evolution. Oxford: Oxford University Press.

Grimm GW, Kapli P, Bomfleur B, McLoughlin S, Renner SS. 2015. Using More Than the Oldest Fossils: Dating Osmundaceae with Three Bayesian Clock Approaches. *Syst Biol.* 64:396-405.

Heath TA, Huelsenbeck JP, Stadler T. 2014. The fossilized birth-death process for coherent calibration of divergence-time estimates. *Proc Natl Acad Sci USA.* 111:E2957-E2966.

Höhna S, Landis MJ, Heath TA, Boussau B, Lartillot N, Moore BR, Huelsenbeck JP, Ronquist F. 2016. RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language. *Syst Biol.* 65:726–736.

Hughes CE, Nyler R, Linder HP. 2015. Evolutionary plant radiations: where, when, why and how? *New Phytol.* 207:247-253.

Huelsenbeck JP and Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics.* 17:754–755.

Katoh K. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30:3059–3066.

Katoh K and Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol.* 30:772–780.

Kishino H, Thorne JL, Bruno WJ. 2001. Performance of a Divergence Time Estimation Method under a Probabilistic Model of Rate Evolution. *Mol Biol Evol.* 18:352–361.

Hedges SB and Kumar S. 2004. Precision of molecular time estimates. *Trends Genet.* 20:242-247.

Lagomarsino LP, Condamine FL, Antonelli A, Mulch A, Davis CC. 2016. The abiotic and biotic drivers of rapid diversification in Andean bellflowers (Campanulaceae). *New Phytol.* 210:1430-1442.

Langley CH and Fitch WM. 1974. An examination of the constancy of the rate of molecular evolution. *J Mol Evol.* 3:161-177.

Lartillot N, Phillips MJ, Ronquist F. 2016. A mixed relaxed clock model. *Phil Trans R Soc B.* 371:20150132.

Lee MSY and Palci A. 2015. Morphological Phylogenetics in the Genomic Age. *Curr Biol.* 25:R922-R929.

MacGintie HD. 1953. Fossil plants of the Florissant beds, Colorado. Institution of Washington: Washington.

Magallón S, Gomez-Acevedo S, Sanchez-Reyes LL, Hernandez-Hernandez T. 2015. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytol.* 207:437–453.

Magallón S, Hilu KW, Quandt D. 2013. Land plant evolutionary timeline: Gene effects are secondary to fossil constraints in relaxed clock estimation of age and substitution rates. *Am. J. Bot.* 100:556-573.

Margoliash E. 1963. Primary Structure and Evolution of Cytochrome C. *Biochemistry.* 50:672-679.

Marshall C. 1990. Confidence intervals on stratigraphic ranges. *Paleobiology.* 16:1-10.

Marshall C. 2008. A Simple Method for Bracketing Absolute Divergence Times on Molecular Phylogenies Using Multiple Fossil Calibration Points. *Am Nat.* 171:726-742.

Martin HA. 2000. Re-assignment of the affinities of the fossil pollen type *Tricolpites trioblatus* Mildenhall and Pocknall to *Wilsonia* (Convolvulaceae) and a reassessment of the evcological interpretations. *Rev Palaeobot Palynol*. 111:237-251.

Martin HA. 2001. The family Convolvulaceae in the Tertiary of Australia: evidence from pollen. *Austral J Bot.* 49:221-234.

Metzgar JS, Skog JE, Zimmer EA, Pryer KM. 2008. The Paraphyly of *Osmunda* is Confirmed by Phylogenetic Analyses of Seven Plastid Loci. *Syst Bot.* 33:31-36.

Mitchell TC, Williams BRM, Wood JRI, Harris DJ, Scotland RW, Carine MA. 2016. How the temperate world was colonised by bindweeds: biogeography of the Convolvuleae (Convolvulaceae). *BMC Evol. Biol.* 16:16.

Miyata T, Yasunaga T, Nishida T. 1980. Nucleotide sequence divergence and functional constraint in mRNA evolution. *Proc Natl Acad Sci USA.* 77:7328-7332.

Morris JL, Puttick MN, Clark JW, Edwards D, Kenrick P, Pressel S, Wellman CH, Yang Z, Schneider H, Donoghue PCJ. 2018. The timescale of early land plant evolution. *Proc Natl Acad Sci USA.* 115:E2274-E2283.

Near TJ, Meylan PA, Shaffer HB. 2005. Assesing Concordance of Fossil Calibration Points in Molecular Clock Studies: An Example Using Turtles. *Am. Nat.* 165:137-146.

Near TJ and Sanderson MJ. 2004. Assessing the quality of molecular divergence time estimates by fossil calibrations and fossil-based model selection. *Phil Trans R Soc Lond B.* 359:1477-1483.

Parham JF, Donoghue PCJ, Bell CJ, Calway TD, Head JJ, Holroyd PA, Inoue JG, Irmis RB, Joyce WG, Ksepka DT, Patané JSL, Smith ND, Tarver JE, van Tuinen M, Yang Z, Angielczyk KD, Greenwood JM, Hipsley CA, Jacobs L, Makovicky PJ, Müller J, Smith KT, Theodor JM, Warnock RCM, Benton MJ. 2012. Best Practices for Justifying Fossil Calibrations. *Syst Biol.* 61:346-359.

Raup DM. 1972. Taxonomic Diversity during the Phanerozoic. *Science*. 177:1065-1071

Renner SS. 2004. Multiple Miocene Melastomataceae dispersal between Madagascar, Africa and India. *Phil Trans R Soc Lond B.* 359:1485-1494.

Renner SS, Grimm GW, Kapli P, Denk T. 2016. Species relationships and divergence times in beeches: new insights from the inclusion of 53 young and old fossils in a birth – death clock model. *Phil Trans R Soc B.* 371:20150135.

Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liang L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Syst Biol.* 61:539–542.

Rubinstein CV, Gerrienne P, de la Puente GS, Astini A, Steemans P. 2010. Early Middle Orodovician evidence for land plants in Argentina (eastern Gondwana). *New Phytol.* 188:365-369.

Sanderson MJ. 1997. A Nonparametric Approach to Estimating Divergence Times in the Absence of Rate Constancy. *Mol Biol Evol.* 14:1218–1231.

Sanderson MJ. 2002. Estimating Absolute Rates of Molecular Evolution and Divergence Times: A Penalized Likelihood Approach. *Mol Biol Evol.* 19:101–109.

Särkinen T, Bohs L, Olmstead RG, Knapp S. 2013. A phylogenetic framework for evoluitonary study of the nightshades (Solanaceae): a dated 1000 tip tree. *BMC Evol Biol.* 13:214.

Schuettpelz E and Pryer KM. 2009. Evidence for a Cenozoic radiation of ferns in an angiosperm-dominated canopy. *Proc Natl Acad Sci USA.* 106:11200-11205.

Signor PW and Lipps JH. 1982. Sampling bias, gradual extinction patterns and catastrophes in the fossil record. pp. 291-296. *in* Silver LT and Schultz PH., eds. Geological Implications of Impacts of Large Asteriods and Comets on the Earth. Geol Soc Am:Boulder.

Smith AB. 2001. Large-scale heterogeneity of the fossil record: implications for Phanerozoic biodiversity studies. *Phil Trans R Soc B.* 356:351-367.

Smith AB. and McGowan AJ. 2005. Cyclicity in the fossil record mirrors rock outcrop area. *Biol Lett.* 1:443–445.

Smith SA and O'Meara BC. 2012. treePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics.* 28:2689–2690.

Springer MS. 1995. Molecular Clocks and the Incompleteness of the Fossil Record. *J Mol Evol.* 41:531-538.

Srivastava G., Mehrotra RC, Dilcher DL. 2018. Paleocene Ipomoea (Convolvulaceae) from India with implications for an East Gondwana origin of Convolvulaceae. *Proc. Natl. Acad Sci USA.* 115:6028–6033.

Stadler T. 2010. Sampling-through-time in birth-death trees. *J Theor Biol.* 267:396-404.

Stadler T, Gavyryushkina A, Warnock RCM, Drummond AJ, Heath TA. 2018. The fossilized birth-death model for the analysis of stratigraphic range data under different speciation modes. *J Theor Biol*. 447:41-55.

Talavera G and Castresana J. 2007. Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. *Syst Biol.* 56:564–577.

Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol Biol Evol.* 15:1647–1657.

Vaidya G, Lohman DJ, Meier R. 2011. SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information. *Cladistics.* 27:171–180.

Warnock RCM, Parham JF, Joyce WG, Tyler R, Donoghue PCJ. 2015. Calibration uncertainty in molecular dating analyses: there is no substitute for the prior evaluation of time priors. *Proc R Soc B.* 282:20141013

Warnock RCM, Yang ZH, Donoghue PCJ. 2011. Exploring uncertainty in the calibration of the molecular clock. *Biol Lett.* 8:156 – 159.

Warnock RCM, Yang Z, Donoghue PCJ. 2017. Testing the molecular clock using mechanistic models of fossil preservation and molecular evolution. *Proc R Soc B.* 284:20170227.

Wilf P, Carvalho MR, Gandalfo MA, Ruben Cuneo N. 2017. Eocene lantern fruits from Gondwanan Patagonia and the early origins of Solanaceae. *Science.* 355:71–75.

Wilf P. and Escapa IH. 2015. Green Web or megabiased clock? Plant fossils from Gondwanan Patagonia speak on evolutionary radiations. *New Phytol.* 207:283–290.

Yang Z. and Rannala B. 2005. Bayesian Estimation of Species Divergence Times Under a Molecular Clock Using Multiple Fossil Calibrations with Soft Bounds. *Mol Biol Evol* 23:212–226.

Zhang C, Stadler T, Klopfstein S, Heath TA, Ronquist F. 2016. Total-evidence dating under the fossilised birth-death process. *Syst Biol.* 65:228-249.

Zhu T., dos Reis M., Yang Z. 2015. Characterization of the Uncertainty of Divergence Time Estimation under Relaxed Molecular Clock Models Using Multiple Loci. *Syst. Biol.* 64:267–280.

Zuckerkandl E. and Pauling LB. 1962. Molecular disease, evolution, and genetic heterogeneity. pp. 189–225. *in* Kasha M. and Pullman B., eds. Horizons in biochemistry. Academic Press: New York.

Zuckerkandl E. and Pauling L. 1965. Evolutionary Divergence and Convergence. pp. 97–166 *in* Bryson V. and Vogel HJ., eds. Evolving genes and proteins. Academic Press: New York.

FIGURE CAPTIONS

Figure 1. Branch lengths (l_{1-8}) are equal to the number of substitutions that have occurred on each branch. l_{1-8} are a product of the rate of molecular evolution for each branch (r_{1-8}) and the temporal duration of each branch (t_{1-8}). Only l_{1-8} can be directly inferred from molecular sequence data. In the case of the strict molecular clock, r_{1-8} are equal. In a more complex case, r_{1-8} are not equal.

Figure 2. A summary of two contrasting fossil records. Fossils are sampled along the branches of a macroevolutionary process that represents the "true" evolutionary history of the extant species A-E. The oldest fossil in each clade is a black hexagon. All other fossils are grey circles. In a) the oldest fossils in each clade (node-calibration-fossils) reflect the relative ages of different clades, and the fossil sampling rate (ψ) is constant. In b) the oldest fossils in each clade do not reflect the relative ages of different clades, and ψ is heterogeneous.

Figure 3. A summary of the different node calibration strategies that are compared in this study. a) Non-Uniform Precise – a non-uniform calibration density assumes the actual age of the clade is similar to the age of the node calibration fossil. b) Non-Uniform Imprecise – a non-uniform calibration density makes less precise assumptions about the age of the clade in relation to the age of the node calibration fossil. c) Uniform Precise – a uniform calibration density that assumes the age of

the clade is similar to the age of the node calibration fossil. d) Uniform Imprecise – a uniform calibration density that makes less precise assumptions about the age of the clade in relation to the age of the node calibration fossil. Note that in this study for NUP, we use exactly the same distribution as has been implemented in previous studies (Särkinen et al. 2013; Grimm et al. 2015). Therefore, for Osmundaceae we use an exponential distribution, and for CS we use a lognormal distribution (see methods). The purpose of this figure is to illustrate the general principles of different calibration strategies, not the methodological details of how they were implemented.

Figure 4. Summary of inferred phylogenies and number of sampled fossils within each clade, for the two datasets in which sampled fossils represent the entire known fossil record. a) Osmundaceae, b) CS.

Figure 5. The temporal distribution of sampled fossils in **a.** Osmundaceae, and **b.** CS. Black plotted points refer to individual fossil occurrences. Fossils are grouped into subclades – the same subclades as summarised in Figure 2. Where a clade has more than one sampled fossil, the upper bound of a 95% confidence interval (CI) for the actual clade age is indicated. This is labelled as the "95% CI for clade age". The width of this bar reflects uncertainty in fossil ages. For CS, two confidence intervals are calculated for each clade. Either the entire dataset of fossils is used when calculating the confidence interval, or a subsample is used in which some Pliocene fossils are removed. Where fossils have been removed, the relevant confidence interval is labelled as the "alternative 95% CI for clade age". The 95th percentile of calibration densities implemented in previous studies (Särkinen et al. 2013, Grimm et al. 2015) is also indicated. These are labelled as either "95th percentile of Grimm et al. (2015)" or "95th percentile of Särkinen et al. (2013).

Figure 6. The distribution of empirical scaling factors (ESFs) for sets of node-calibration-fossils previously implemented in a) Osmundaceae, b**)** CS, and c) Spermatophyta. The ESF provides an approximation of how node-calibration-fossils reflect the actual age of their respective clades. Comparison of ESFs for a set of node-calibration-fossils therefore provides a basis to determine the extent to which they reflect the relative ages of different clades. In all three clades, the range of ESFs suggests sets of node-calibration-fossils are unlikely to reflect the relative ages of different clades. The ultrametric phylogeny with which the ESFs are calculated was inferred in treePL with the optimum smoothing value.

Figure 7. Inferred ψ for different parts of the evolutionary history of a) Osmundaceae, and b) CS. Trees shown are ultrametric phylogenies inferred in treePL with the optimum smoothing value, and ψ is inferred according to these trees. The FBD assumes a constant ψ, whilst the analysis presented here suggests ψ is highly heterogeneous in Osmundaceae and CS.

Figure 8. Inferred divergence times in Osmundaceae with different calibration strategies and different molecular clock strategies. In a) divergence times are inferred with a relaxed clock, in b) divergence times are inferred with a strict clock, and in c) divergence times inferred with no molecular data. Plotted points refer to the percentage change of the MPE for a given node and calibration strategy relative to NUP implemented with the molecular clock strategy that is the subject of each subfigure. This figure enables comparisons between different calibration strategies within each molecular clock strategy. Different shaped and coloured points refer to different calibration strategies (see key). Each grey or white stripe refers to a different node – these are ordered by depth in an ultrametric phylogeny of the same taxa (see methods).

Figure 9. Inferred divergence times in CS with different calibration strategies and different molecular clock strategies. In a) divergence times are inferred with a relaxed clock, in b) divergence times are inferred with a strict clock, and in c) divergence times inferred with no molecular data. Plotted points refer to the percentage change of the MPE for a given node and calibration strategy relative to NUP implemented with the molecular clock strategy that is the subject of each subfigure. This figure enables comparisons between different calibration strategies within each molecular clock strategy.

Different shaped and coloured points refer to different calibration strategies (see key). Each grey or white stripe refers to a different node – these are ordered by depth in an ultrametric phylogeny of the same taxa (see methods).

Figure 10. Inferred divergence times in Osmundaceae with different calibration strategies and different molecular clock strategies. In a) divergence times are inferred with NUP, in b) divergence times are inferred with NUI, in c) divergence times are inferred with UP, in d) divergence times are inferred with UP, in d) divergence times are inferred with UI, in e) divergence times are inferred with FBDH, and in f) divergence times are inferred with FBDS. Plotted points refer to the percentage change of the MPE for a given node and inference method, relative to a relaxed clock implemented with the calibration strategy that is the subject of each subfigure. This figure enables comparisons between different molecular clock strategies within each calibration strategy. Different shaped and coloured points refer to different molecular clock strategies (see key). Each grey or white stripe refers to a different node – these are ordered by depth in an ultrametric phylogeny of the same taxa (see methods).

Figure 11. Inferred divergence times in CS with different calibration strategies and different molecular clock strategies. In a) divergence times are inferred with NUP, in b) divergence times are inferred with NUI, in c) divergence times are inferred with UP, in d) divergence times are inferred with UP, and in d) divergence times are inferred with UI. Plotted points refer to the percentage change of the MPE for a given node and inference method, relative to a relaxed clock implemented with the calibration strategy that is the subject of each subfigure. This figure enables comparisons between different molecular clock strategies within each calibration strategy. Different shaped and coloured points refer to different molecular clock strategies (see key). Each grey or white stripe refers to a different node – these are ordered by depth in an ultrametric phylogeny of the same taxa (see methods).