nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-020-01387-8



# Fossil data support a pre-Cretaceous origin of flowering plants

Daniele Silvestro[1,2,3,4] ✉, Christine D. Bacon[3,4], Wenna Ding[5], Qiuyue Zhang[1,5,6], Philip C. J. Donoghue[7], Alexandre Antonelli[3,4,8,9] and Yaowu Xing[5]

**Flowering plants (angiosperms) are the most diverse of all land plants, becoming abundant in the Cretaceous and achieving dominance in the Cenozoic. However, the exact timing of their origin remains a controversial topic, with molecular clocks generally placing their origin much further back in time than the oldest unequivocal fossils. To resolve this discrepancy, we developed a Bayesian method to estimate the ages of angiosperm families on the basis of the fossil record (a newly compiled dataset of ~15,000 occurrences in 198 families) and their living diversity. Our results indicate that several families originated in the Jurassic, strongly rejecting a Cretaceous origin for the group. We report a marked increase in lineage accumulation from 125 to 72 million years ago, supporting Darwin's hypothesis of a rapid Cretaceous angiosperm diversification. Our results demonstrate that a pre-Cretaceous origin of angiosperms is supported not only by molecular clock approaches but also by analyses of the fossil record that explicitly correct for incomplete sampling.**

Ubiquitous across terrestrial and aquatic systems, flowering plants (angiosperms) are the most diverse group of land plants on Earth today. Fossil evidence indicates that angiosperms and gymnosperms had already diverged by the late Carboniferous (306.2 million years ago (Ma))[1]. The earliest unequivocal fossil evidence of crown angiosperms dates to the Early Cretaceous (Valanginian stage; ~135 Ma)[2], but the true time of origin of the living clade remains debated[3,4]. The sudden stratigraphic appearance of crown angiosperm fossils, apparently without forebears displaying evidence of the gradual assembly of the angiosperm body plan, was considered "an abominable mystery" by Darwin and his contemporaries[5]. Angiosperms have been ecologically dominant since the Late Cretaceous and have subsequently increased in diversity and complexity[6]. The sudden appearance of a high level of diversity shortly after their origin underlies Darwin's perplexity, leading him to hypothesize a long, undiscovered pre-Cretaceous angiosperm history and to search for drivers of rapid plant diversification, such as coevolution with pollinators[7,8].

Darwin's abominable mystery may have a modern analogy. Molecular data increasingly inform our understanding of the tree of life, but these data often seem to contradict palaeontological evidence[9]. While some of the discrepancies between molecular clock and palaeontological estimates of macroevolutionary dynamics can be reconciled through the integration of fossil and phylogenetic data[10–13], contrasting estimates of the origins of major clades in the tree of life remain an open challenge[14].

The discrepancy between the fossil record and crown age estimates in angiosperms has been long debated. The Early Cretaceous angiosperm fossil record comprises lineages that were species-rich and morphologically diverse by ~130–100 Ma (ref. [2]), suggesting that the ancestor of all angiosperms should be considerably older. Reports of putative angiosperm pollen from the Triassic and a leaf from the Jurassic[15–20] hint towards a significantly older origin of flowering plants[21], but the discrepancy remains, due to the lack of undisputed pre-Cretaceous angiosperm fossils[3,6]. A recent review suggests that Early Cretaceous pollen records might be compatible with a latest Jurassic origin of the clade, but not earlier[3]. Meanwhile, large phylogenomic studies continue to point to a substantially older origin of the clade, perhaps as early as the Permian[22]. This 'Jurassic gap', indicating the discrepancy between molecular and fossil age estimates[4,22], has been attributed to the rarity or small size of early angiosperms[23], lower fossil preservation rates[24], heterogeneity in the rock record[25] or some combination of these factors.

The fossil record alone provides only minimum constraints based on clade ages; evolutionary timescales require further inference. However, molecular clocks are not the only methods available for estimating clade age, and alternative approaches exist that eschew molecular data and phylogenetic methods altogether. These methods estimate the age ranges that may include the true times of origination (and extinction) of taxa on the basis of the observed stratigraphic range of taxa and the number of fossiliferous horizons[26–28], although we are not aware of applications of these methods to the angiosperm fossil record. More recent advances have used Bayesian inference to model fossil occurrences, while accounting for the underlying preservation processes and dating uncertainties[13,29,30]. These methods have been used to infer diversification dynamics of vascular plants[31,32], but they are limited in their ability to analyse clades with a scarce fossil record (such as most angiosperm families) and are not explicitly designed to estimate clade age.

Significant methodological advances have been made in recent years in inferring phylogenetic trees with extinct and extant taxa[13,30,33–35]. However, these advances have not closed the gap between fossil and phylogenetic estimates in relation to the age of clades, leading some to suggest the presence of systematic biases in phylogenetic estimates of the origination times of clades[14,36].

[1]Department of Biology, University of Fribourg, Fribourg, Switzerland. [2]Swiss Institute of Bioinformatics, Fribourg, Switzerland. [3]Department of Biological and Environmental Sciences, University of Gothenburg, Gothenburg, Sweden. [4]Gothenburg Global Biodiversity Centre, Gothenburg, Sweden. [5]CAS Key Laboratory of Tropical Forest Ecology, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla, China. [6]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland. [7]School of Earth Sciences, University of Bristol, Bristol, UK. [8]Royal Botanic Gardens, Kew, Richmond, UK. [9]Department of Plant Sciences, University of Oxford, Oxford, UK. ✉e-mail: daniele.silvestro@unifr.ch

Here, we revisit the Jurassic gap controversy through the analysis of a newly compiled, extensive dataset of over 15,000 angiosperm meso- and macrofossils spanning the Cretaceous and the Cenozoic. To resolve the discrepancy in previous estimates, we develop a new, phylogeny-free model to infer the age of origin of clades on the basis of present diversity and the known fossil record. Our Bayesian method is explicitly designed for clades that are (or have been in the past) highly diverse but present a patchy and severely incomplete fossil record, such as angiosperms. After validating our approach through extensive simulations, we infer the range of plausible ages of origin of 198 angiosperm families. We then test whether an analysis of the fossil record, accounting for incomplete sampling, supports a pre-Cretaceous angiosperm origin, as speculated by Darwin.

## Results

**A new method to infer clade age.** We developed a model, which we term Bayesian Brownian bridge (BBB), to infer the age of origin of a clade on the basis of its present diversity and the known fossil record. The method is specifically designed to accommodate not only fossil-rich groups but also clades with extremely poor sampling (such as groups of organisms in which the great majority of species that have existed did not leave a fossil record).

The estimation is implemented within a Bayesian framework and uses the following input data:

- Present species richness, $N > 0$ (note that the current implementation is designed for extant clades)
- Sampled species richness through time, $\mathbf{x} = \{x_0, \ldots, x_T\}$

The vector $\mathbf{x}$ includes the number of sampled species within time bins of a predefined size, in our analyses set to 2.5 Myr. Since the assumption is that the fossil record can be extremely incomplete, most of the bins are likely to have a number of sampled species equal to 0.

We assume that the diversity of a clade through time follows a Brownian bridge (Fig. 1)—that is, a random walk process constrained at the two endpoints to have a value of $d_T = 1$ at its origin (one ancestral species at time $T$) and to have a value of $d_0 = N$ in the present (time 0). We denote the vector of (unknown) diversity for each time bin as $\mathbf{d} = \{d_1, \ldots, d_{T-1}\}$. The Brownian bridge is further conditioned such that $d_i \geq \max(1, x_i)$; thus, the diversity trajectory for a sampled Brownian bridge is $\mathbf{d} \sim B_0^T(\sigma^2, 1, N)$. This condition implies that the clade cannot go extinct between time $T$ and time 0, even if there are no fossils in a time bin, and that the true diversity cannot be lower than the sampled diversity.

*Likelihood and data augmentation.* We implemented data augmentation to compute the likelihood of the fossil data and present diversity given an average sampling rate ($q$) while accounting for multiple diversity trajectories. In particular, given the two parameters of the Brownian bridge (time of origin $T$ and variance $\sigma^2$), we sampled a large number ($K$) of Brownian bridges and averaged the likelihood of the data across them following the algorithm described by Tanner and Wing[37]. The likelihood of the data is thus approximated as

$$P(\mathbf{x}, N|q, T, \sigma^2) \approx \frac{1}{K} \sum_{k=1}^{K} P(\mathbf{x}, N|q, \mathbf{d}_k) \quad (1)$$

where the $k$th diversity trajectory $\mathbf{d}_k \sim B_0^T(\sigma^2, 1, N)$ is sampled from a Brownian bridge conditioned as described above, and $P(\mathbf{x},N|q,\mathbf{d}_k)$ is the likelihood of the sampled species richness through time and the present diversity. The likelihood of the fossil count in time bin $i$ under each conditioned Brownian bridge is computed on the basis of the probability mass function of a binomial distribution:

$$P(x_i, d_i|q) = \binom{d_i}{x_i} q^{x_i} (1-q)^{d_i - x_i}. \quad (2)$$

In our simulations and empirical analyses, we set $K = 1{,}000$. We note that increasingly large values of $K$ yield improved convergence of the analysis, although at the cost of more expensive computation[37]. We sampled all model parameters from their posterior distributions using a Markov chain Monte Carlo (MCMC) algorithm (Methods).

*Time-increasing sampling rate.* Empirical studies of the fossil record indicate that there is a general trend for sampling rates to increase towards the recent[38,39]. Additionally, sampling rates for individual clades might be low at their time of origin and later increase as the clade diversifies and expands geographically[40,41]. To accommodate these potential heterogeneities in the rock and fossil records, we implemented a model in which the sampling rate at time $t$ is equal to

$$q_t = q_T \times \exp[a(T - t)] \quad (3)$$

where $a \geq 0$ is the parameter determining the rate of exponential increase in the sampling rate as a function of time, and $q_T$ is the minimum sampling rate at the clade origin. While this remains a rough approximation of how sampling rates might vary over time, it accounts for some degree of rate variation while adding only a single parameter to the model. Although more complex alternatives (such as models with rate shifts[29]) are possible in principle, they would not be readily applicable to clades with a scarce fossil record. To assess the effect of accounting for rate increase through time, we performed all analyses of simulated and empirical datasets under both models, where $a$ was either set equal to 0 (constant rate) or inferred from the data.

**Performance of the BBB model.** The BBB model infers the age of origin of a clade on the basis of its present diversity and on its sampled fossil record (Fig. 1). All results presented here, unless otherwise specified, refer to analyses carried out using the time-increasing-rate model, which our simulations showed to be the most flexible and robust, as shown below. The results obtained under the constant-rate model ($a = 0$) are available in the Supplementary Information. The times of origin estimated from datasets simulated under randomly varying sampling rates were unbiased (Fig. 2a,b) and accurately estimated with a mean absolute relative error of 0.16 (standard deviation across simulations, 0.15). As expected, the relative error was generally lower for datasets with a higher number of fossil occurrences (Fig. 2b), and the size of the 95% credible intervals (a measure of the precision of the estimates) was larger for datasets with a lower number of fossil occurrences (Fig. 2c). The accuracy and precision of the estimates did not vary as a function of the age of the simulated clade (Fig. 2a). The coverage in the estimation of $T$ (the frequency at which the true time of origin was included in the 95% credible interval of the estimated one) was 0.97. All cases where the true time of origin was not encompassed in the estimated 95% credible interval were due to a significant underestimation of the parameter (that is, the estimated age was significantly younger than the true age; Supplementary Table 1). The log variance of the Brownian bridge was slightly underestimated, although the mean absolute relative error remained small at 0.13 (s.d., 0.07). This consistent underestimation might be linked to the fact that the model assumptions (constant sampling rates through time) are strongly violated in the simulated datasets. We note, however, that the underestimation does not have a biasing impact on the estimation of the time of origin (Extended Data Fig. 1). The estimated sampling rates ($q_T$) ranged between $3.4 \times 10^{-5}$ and $1.3 \times 10^{-3}$ (median, $2.6 \times 10^{-4}$), and the trend parameter ($a$) was small (median, 0.68; s.d., 2.76; Fig. 2e,f).

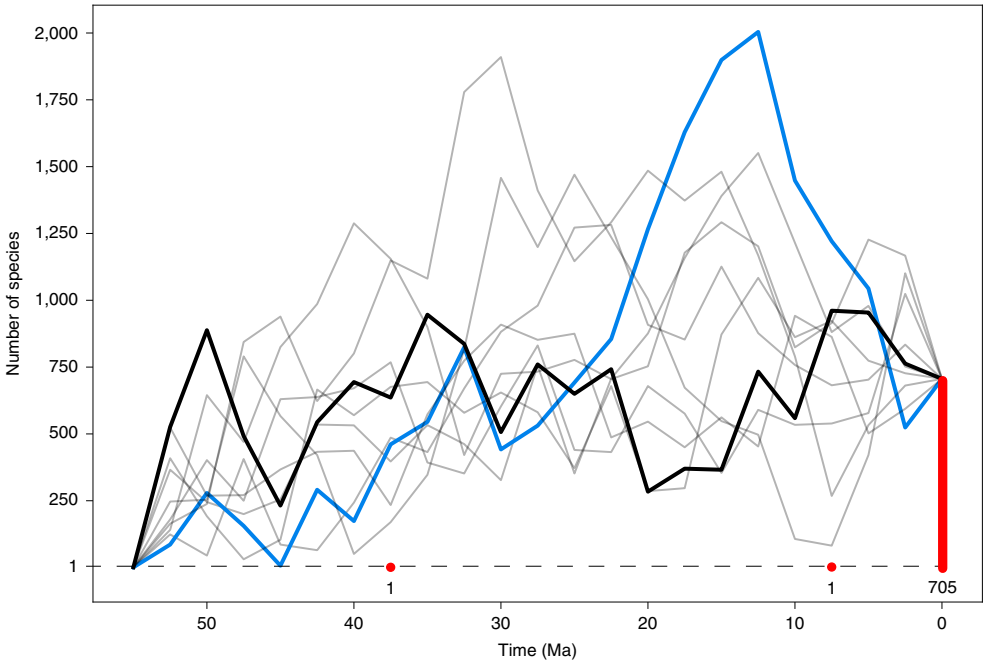



**Fig. 1 | Examples of diversity trajectories simulated using a conditioned Brownian bridge.** The thicker lines highlight two of the simulated trajectories. The Brownian bridges are constrained to a minimum diversity of one species; this threshold is highlighted by a dashed line. They are further constrained by fossil diversity, here represented by the red circles, which indicate the temporal placements of the simulated fossils (with the number of occurrences), and by present diversity, represented by the red bar. These simulations are based on a Brownian bridge starting at the diversity of one species at the time of origin $T=55$, with a variance $\sigma^2=10^{4.4}$ and a present diversity $N=705$ and on a sampling process with an average sampling rate $q=10^{-3.6}$.

Analyses performed without using the MCMC approximation described in the Methods showed that parameter estimation is virtually unaltered by this procedure (Extended Data Fig. 2). However, non-approximated MCMCs yielded 4.7% of simulations with poor convergence (effective sample size (ESS) < 100) compared with 1.5% using the approximation. Furthermore, in 2% of the simulations without approximation, the ESS was lower than 25, whereas the overall lowest ESS obtained using the approximate MCMC was 78. Our approximated MCMC thus provided more efficient sampling without visibly altering the parameter estimates.

A re-analysis of the same datasets under the constant-rate model resulted in parameter estimates largely consistent with those from the time-increasing-rate model (Extended Data Fig. 3). However, the coverage decreased to 0.93, with 6.5% of the datasets resulting in a significant underestimation of the time of origin, while overestimation remained rare (Supplementary Table 1).

Simulations with moderately and strongly increasing sampling rates through time had the effect of reducing the accuracy of the estimated times of origin, with mean absolute relative errors increasing to 0.20 and 0.36, respectively (Supplementary Table 1 and Extended Data Figs. 4 and 5). The coverage decreased to 0.74 and 0.38, respectively, and the decrease is entirely due to instances of underestimated times of origin. The estimated trend parameter ($a$) reflected the increasing sampling rates through time, with a median value of 1.20 (s.d., 2.67) for simulations with moderately increasing rates and 2.52 (s.d., 3.71) for simulations with strongly increasing rates (Extended Data Figs. 4 and 5). In these simulations, the use of a constant-sampling-rate model ($a=0$) resulted in considerably higher relative errors and lower coverage (Supplementary Table 1 and Extended Data Figs. 6 and 7).

**Origin of angiosperm families.** Analyses of the angiosperm fossil record carried out under temporal bin sizes of 1, 2.5 and 5 Myr produced highly consistent results (Extended Data Fig. 8), indicating that the discretization of the time axis had a negligible impact on the analyses. We therefore report the results based on 2.5 Myr bins to match the setting used in the simulations, while detailed results from all analyses are available in the Supplementary Information (Supplementary Table 2 and Extended Data Fig. 8).

The estimated times of origin across 198 angiosperm clades (Angiosperm Phylogeny Group IV families[42]) were spread across the Cenozoic (64 families, 32%), Cretaceous (131 families, 66%) and Jurassic (3 families: Lardizabalaceae, Papaveraceae and Triuridaceae). The detailed results for each family are provided in Supplementary Table 2. The credible intervals for several families (20%) extended well into the Jurassic and, in fewer instances (8%), into the Triassic (Fig. 3a and Supplementary Table 2). As observed with the simulated data, the size of the credible intervals was largest in clades with few fossil occurrences (Extended Data Fig. 9a). The log variances of the Brownian bridge scaled by the number of extant species ranged between 0.48 and 14.29 (median, 7.79; Extended Data Fig. 9b). The estimated sampling rates at the time of origin ($q_T$) ranged between $4.7\times10^{-6}$ and 0.29 (median, 0.0014), and the trend parameter ($a$) ranged between 0.33 and 14.99 (median, 1.87), which indicates a moderate rate increase through time, on the basis of the values observed in our simulations (Extended Data Figs. 4, 5 and 9c,d).

By combining the posterior estimates of the times of origin of all families to obtain an indirect estimate of the crown age of angiosperms, we calculated that the probability that at least one family originated before the Cretaceous is $P(\max(T)>145\,\text{Ma})=0.998$ (Fig. 3b and Extended Data Fig. 10a). The estimated 95% credible interval for the time of origin across all families was 254.8 to 153.7 Ma, matching almost exactly the estimated range of the crown age of angiosperms obtained from a recent molecular clock study[25] (256–149 Ma; Fig. 3b). Selected pollen data were included for four families; this inclusion did not change the estimated age of Arecaceae, but it pushed the origination times of Aponogetonaceae,

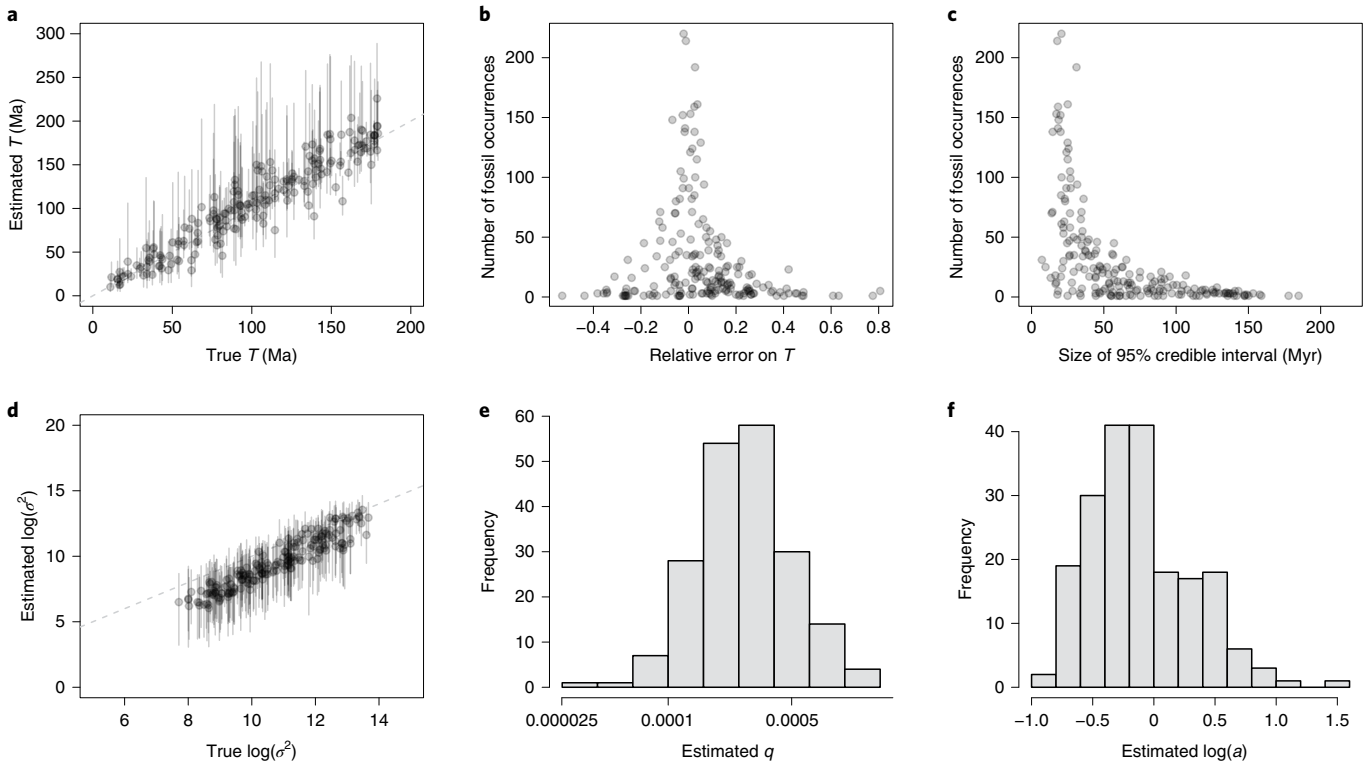



**Fig. 2 | Performance of the BBB method assessed through 200 simulations with randomly varying sampling rates through time. a**, The times of origin were accurately estimated. The circles and bars indicate the posterior estimates and 95% credible intervals. **b**, The relative errors on the time of origin were generally small and unbiased (that is, centred on 0), and they were smaller in datasets with richer simulated fossil records. **c**, The size of the 95% credible intervals around the times of origin decreased with an increasing number of fossils. **d**, The log variances were slightly underestimated. **e**,**f**, The estimated sampling rates (**e**) and sampling trends (**f**; the *x* axis is $\log_{10}$-transformed) cannot be plotted against the true values because the underlying simulations were based on time-heterogeneous sampling with different rates in each time bin to more closely reflect the biases in the fossil record.

Araliaceae and Asteraceae to older dates, although the credible intervals with and without pollen data overlapped (Supplementary Table 2). This suggests that the inclusion of additional pollen data in the analysis could increase the estimated age of angiosperm clades. Analyses performed on the meso- and macrofossil record only, however, showed that these pollen records did not change the overall pattern of accumulation of family-level diversity in angiosperms (Extended Data Fig. 10b). Similarly, analyses based on a model assuming constant sampling rates ($a=0$) inferred a substantially similar pattern of lineage accumulation, with an estimated 95% credible interval for the time of origin across all families spanning from 253.5 to 152.6 Ma (Extended Data Fig. 10c).

Family diversity accumulated most rapidly throughout the Cretaceous (Fig. 3b), followed by a slow-down in the Cenozoic. Diversification rates were low until the Early Cretaceous (Fig. 4), during which they underwent a 1.7-fold increase in the Aptian (125 Ma), followed by a gradual rate decline. The family-level diversification rate peaked again in the Campanian (83.6–72.1 Ma), after which it dropped fourfold at the onset of the Cenozoic.

We compared BBB estimates of the times of origin of families with their crown ages inferred in a molecular clock study, in which the age of crown angiosperms was constrained to the Cretaceous[43]. In 129 families (71.7% of the families found in both datasets), the credible intervals of the estimated root ages overlapped, indicating that our inferred ages are compatible with molecular clock estimates (Fig. 5a, grey circles). Our estimates were significantly older than molecular phylogenetic estimates in 24 families (13.3%; blue circles) and significantly younger in 27 families (15%; red circles). These results show that, while there remain several discrepancies between molecular phylogenetic and fossil-based age estimates across angiosperm clades, there are no consistent differences between them.

A re-analysis of the fossil data with the stratigraphic confidence interval method[26] provided age estimates that are highly consistent with our Bayesian inferences (Fig. 5b). However, for a few families, the inferred range of plausible ages was significantly larger under this method, spanning well beyond the Triassic (Supplementary Table 2).

## Discussion

We present a Bayesian model to infer the time of origin of clades, while integrating all plausible diversification histories (Brownian bridges) via data augmentation. Our method uses the temporal distribution of sampled fossil diversity and the modern diversity of a clade to jointly estimate the time of origination of the clade, the amount of heterogeneity in the diversification process and an overall sampling rate, which approximates the probability of sampling a species in the fossil record per unit of time. Using simulations, we have shown that clade ages inferred by our model are accurate (Fig. 2) even when the fossil record is extremely poor, with only one in several thousand species expected to leave a fossil record. While the BBB model makes a number of simplifying assumptions (for example, using a constant or exponentially increasing sampling rate through time), it produced accurate results even in the presence of strong violations of such assumptions (elevated rate heterogeneity and preservation gaps). The accuracy and precision of the estimated times of origin were, as expected, functions of the number and density of fossil occurrences, with fossil-rich datasets producing the most reliable results. Strong temporal trends in the sampling rates,

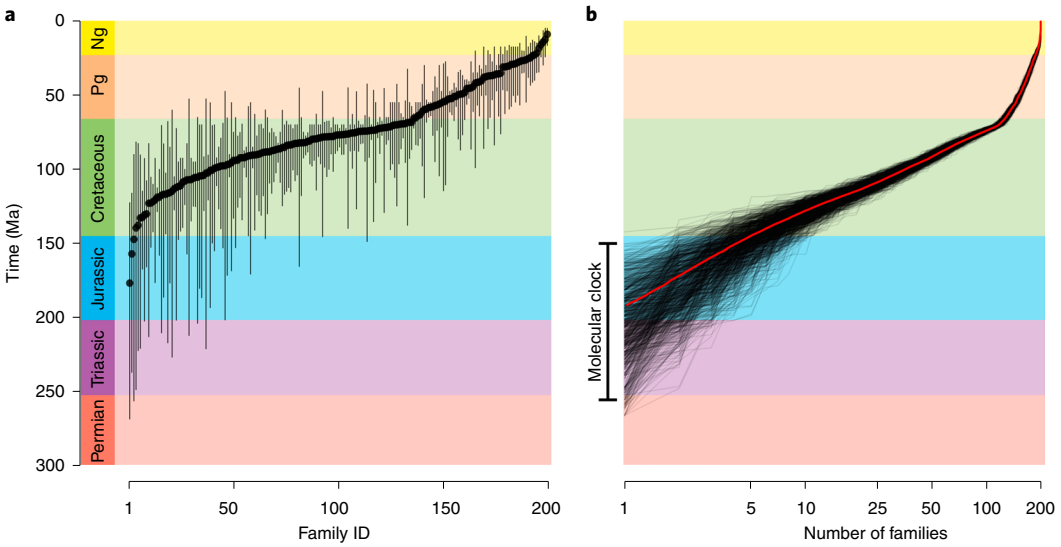



**Fig. 3 | Estimated times of origin of angiosperm families and cumulative family diversity plot. a**, Estimated times of origin (circles) of 198 sampled families of angiosperms with 95% credible intervals (vertical lines). The coloured bars on the *y* axis show the boundaries of the geological periods. Pg, Palaeogene; Ng, Neogene. **b**, The estimated times of origin were used to produce a cumulative family diversity plot (the *x* axis is $\log_{10}$-transformed). The black lines show diversity trajectories based on 1,000 posterior samples; the red line shows the mean. The left bar shows the estimated plausible range for the crown age of angiosperms, inferred in a recent molecular phylogenetic study[25].

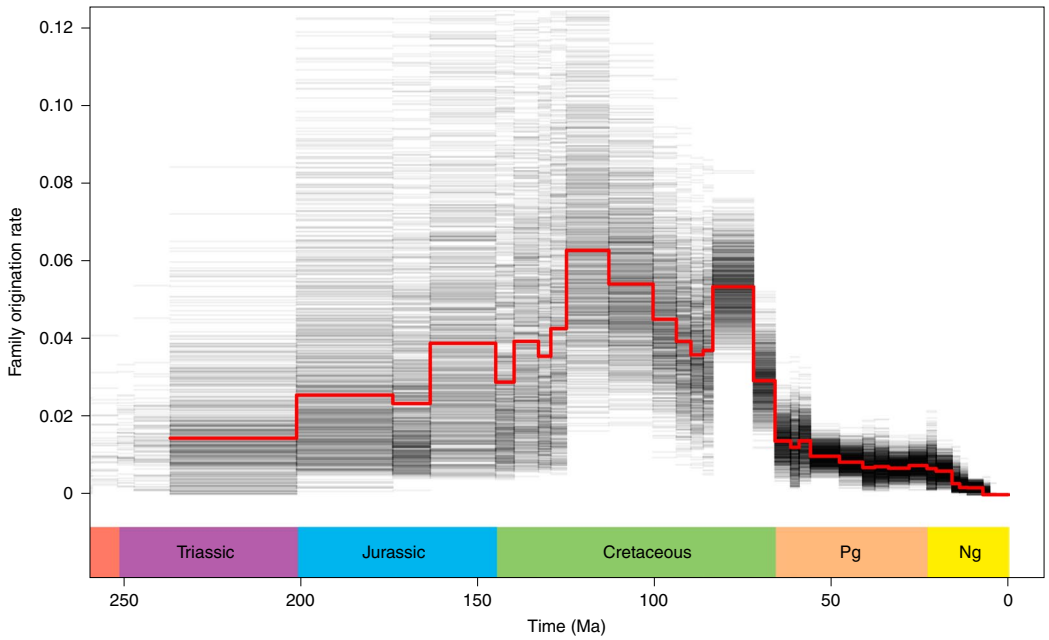



**Fig. 4 | Family-level origination rates inferred from the estimated diversity trajectories of the sampled families (Fig. 3b).** The black lines represent 1,000 posterior samples, and the red line shows their median.

whereby early rates are much lower than those close to the recent, result in an increased age gap between the true time of origin and the oldest sampled fossil[41]. Under these settings, our model showed decreased accuracy, often leading to an underestimation of the age of origin. Crucially for our angiosperm analysis, our simulations showed that the BBB model is robust to overestimating the time of origin of clades, regardless of the dynamics of the sampling process.

The BBB model does not make explicit assumptions about the allocation of fossils to the stem or crown of the clade. Instead, it estimates the age of the most recent common ancestor of all species included in the dataset (modern and fossil) by estimating the time at which the diversity of the clade was a single species. Whether our estimates represent the age of the crown group or the age of the total group thus depends on whether the fossil species attributed to a family are descendants of the crown ancestor alone, or whether they include members of the stem as well[41]. Since our dataset was limited to fossil taxa that had been assigned to extant families, our results can be interpreted conservatively as estimates for the age of

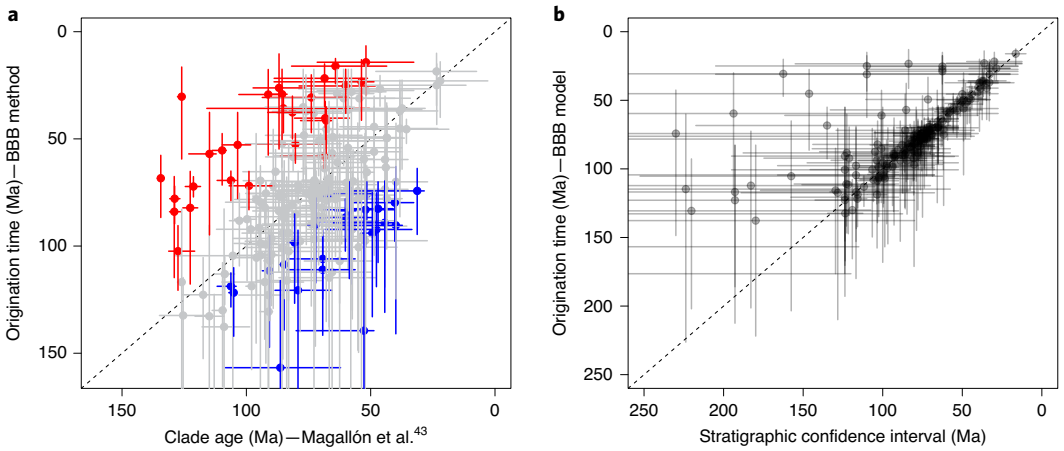



**Fig. 5 | Comparison between our estimates of the age of origin of angiosperm families and estimates based on a molecular clock and the stratigraphic confidence interval. a**, Comparison between fossil (this study) and molecular clock estimates[43] of the age of origin across the 180 angiosperm families found in both analyses. Instances in which the 95% credible intervals overlapped between the two estimates are shown in grey. Families inferred to be significantly younger in the fossil record compared with molecular clock estimates are shown in red, whereas older fossil estimates are shown in blue. **b**, Comparison between estimates of the times of origin of angiosperm families based on the BBB model and estimates based on the stratigraphic confidence interval[26]. The circles in **b** represent BBB posterior age estimates and are plotted against the midpoint of the stratigraphic confidence intervals.

the extant family-level total groups—that is, the clades encompassing all known extant and extinct species in the family. There are no recognized extinct angiosperm families, although it is clear from the literature that there are many fossil species that cannot be accommodated within even total-group definitions of extant families[44–48]. These could not be recognized in our analysis because they cannot be accommodated by our current model. The exclusion of these Early Cretaceous records could lead our analysis to underestimate early angiosperm familial diversity and potentially bias the estimated diversification rates within this time frame.

Our model is based exclusively on the fossil record and on the modern diversity of clades; it therefore provides age estimates that are independent of the assumptions of molecular clocks. The BBB model is not based on an explicit phylogenetic framework and should therefore be less subject to potential biases associated with birth–death processes[14,49,50]. It does, however, assume an equivalency between living and extinct species, which future developments should aim to correct for in light of the differences between species concepts[12,51].

The model developed here offers the opportunity to estimate Bayesian credible intervals for the time of origin even of clades with very scarce fossil records. These estimates can be used to define objective and data-driven priors on clade ages (for instance, setting normal or gamma prior distributions with 2.5 and 97.5 percentiles matching the 95% credible intervals inferred under the BBB model). This can be easily applied to inform molecular phylogenetic analyses using node calibration[52] or total evidence dating[35,53], where a prior on the root age must be specified, even in the presence of fossil tips.

**The fossil record and the origin of flowering plants.** There has been intense debate about the time of origin of flowering plants[3,4,54], with most palaeontological studies firmly placing the crown age of angiosperms in the Cretaceous[6,55], while molecular clock analyses indicate a much earlier origination of the group in the Jurassic and possibly even extending to the Permian[22,56,57]. This apparent discrepancy—the Jurassic gap—has been attributed to biases in molecular dating[14] or gaps in the rock and fossil records[22]. Ultimately, absolute divergence times inferred from molecular clocks are necessarily dependent on the integration of fossil data through node calibration or the inclusion of extinct tips in the phylogeny[58,59], but inadequate molecular evolutionary models can lead to spurious results[49]. However, a reading of the fossil record that does not explicitly attempt to correct for missing data and heterogeneous sampling is insufficient to understand the time of origin of large ancient clades, due to their inevitable incompleteness[27].

Our analysis of the angiosperm fossil record indicates that palaeontological evidence, when interpreted in the light of incomplete preservation, does not reject a pre-Cretaceous origin of flowering plants. In fact, our findings indicate that several families with living descendants originated in the Jurassic, thus placing strong statistical support on an early origin of crown angiosperms, with a probability of a Cretaceous crown age for angiosperms as low as $P=0.002$ (Extended Data Fig. 10a). The range of estimated times of origin across 198 sampled families spans the Triassic and the Jurassic, matching remarkably well with recent molecular clock estimates of the crown age of angiosperms[25]. Like these molecular clock studies, our fossil-based analysis cannot discriminate between an early or late origin of crown angiosperms within this broad range. Yet, we have shown that literal interpretations of the fossil record can be rejected and that the palaeobotanical quest for the "mythical Jurassic angiosperm" (sensu Bateman[55]), is supported by the currently known and accepted fossil record; it is not just a product of molecular phylogenetics.

Many hypotheses have been invoked to explain the discord between molecular estimates for the timing of the origin of crown angiosperms and their appearance in the fossil record. These include the possibilities that early crown angiosperms were ecologically or geographically restricted[60,61], that they lived in environments with low preservation potential and that their fossil record is subject to heterogeneities in the rock record[25]. Certainly, most claims of pre-Cretaceous crown angiosperms have been robustly refuted[3,6,55,62], though there remain outstanding records of Late Triassic crown-angiosperm-like pollen[16,19,20] and a Middle Jurassic crown-angiosperm-like leaf[15,18]. Another possibility is that molecular clock estimates are simply wrong and the fossil record presents an accurate account of crown angiosperm evolutionary history. In any instance, the fossil record requires interpretation, and ours indicates that the fossil record supports a pre-Cretaceous origin of crown angiosperms compatible with some recent molecular clock studies[25].

The estimated family-level diversification rates through time suggest a pre-Cretaceous phase of slow diversification of flowering plants, which is consistent with the hypotheses that early angiosperms were rare and slowly evolving[25,63]. This phase was followed by a rapid radiation of lineages between 125 Ma and 72 Ma, as shown by a strong increase in diversification rates, resulting in the increasing levels of taxonomic diversity observed during the Cretaceous[2,6]. This is in line with recent estimates based on molecular clocks[61] and supports Darwin's assertion that angiosperms underwent a rapid diversification at that time. Finally, family-level diversification levels off towards the recent, as expected for higher taxonomic clades.

## Conclusions

Inferences about ancient events shaping the tree of life remain a challenge in evolutionary biology, and future fossil discoveries and methodological advances might change the plausible range of hypotheses regarding the origin of angiosperms and diversification of their many families. Yet, our results indicate that an early, pre-Cretaceous origin of angiosperms is supported not only by molecular phylogenetic hypotheses but also by an analysis of the fossil record that accounts for incomplete sampling, thus reconciling palaeontological and molecular clock estimates of the evolutionary history of flowering plants.

## Methods

**Bayesian parameter estimation.** We used an MCMC algorithm to estimate the model parameters $q$ (or $q_T$ and $a$ under the time-increasing-rate model), $T$ and $\sigma^2$. We used an arbitrarily large uniform prior on the time of origin ($T \sim \mathcal{U}[\max(\mathbf{x}), 300]$), which we deemed appropriate for angiosperm clades as it goes back to the Carboniferous–Permian boundary (more than twice the age of the oldest unequivocal crown angiosperm fossil). We set an exponential prior on the ratio between the variance of the Brownian bridge and the number of species at the present ($\frac{\sigma^2}{N} \sim$ Exp (0.1)) and a gamma prior on the average sampling rate ($q \sim \Gamma[1.1, 1]$). We used a normal kernel proposal for $T$ with reflection at the boundaries determined by the prior and multiplier proposals for $q$ and $\sigma^2$. When running with a time-increasing-rate model, we additionally set a vague exponential prior on the parameter $a \sim$ Exp(0.01) and used normal proposals with reflection at the 0 boundary.

All analyses were carried out on the basis of 250,000 MCMC iterations, sampling every 500 iterations. When summarizing the results, we discarded the first 10% of the samples as burn-in. We used the same MCMC settings and priors in the analyses of all simulated and empirical datasets described below. To improve the mixing of the MCMC, we introduced a small fraction of iterations in which the parameters were not updated but a new set of conditional Brownian bridges were drawn. These draws were performed randomly with a frequency of 5% and treated as an approximation of samples from the posterior (as they do not involve changes in $q_T$, $a$, $T$ and $\sigma^2$), thus accepted in the MCMC. While acknowledging that this results in an approximation of the posterior, we found through the analyses of 200 simulated datasets that these moves have a negligible effect on the estimated times of origin but can substantially improve the mixing and convergence of the MCMC, which we quantified as the ESS of the sampled posterior probability (Results).

We summarized the parameters as posterior mean and 95% credible intervals and computed relative errors to assess the accuracy of the method. Since the parameters $q$, $q_T$ and $a$ do not have a corresponding single value in the simulation settings (where we use instead a vector of preservation rates varying through time), we did not compute the relative errors for these parameters. We also computed the coverage for $T$—that is, the frequency at which the true time of origin was found within the 95% credible interval of the estimated $T$.

**Simulations.** We simulated 200 datasets and, for each, sampled the true root age from a uniform distribution $\mathcal{U}[10, 180]$. The bin size was set to 2.5 Myr, and different sampling rates were drawn randomly for each time bin, a setting that explicitly violates the assumptions of the BBB model, but which we think better reflects empirical observations of the fossil record. We obtained the vector of true sampling rates from $\mathbf{q} \sim \exp(\mathcal{N}(-8.52, 1))$, which generates rates with a median equal to 0.0002 (that is, 1 in 5,000 lineages is expected to leave a fossil record in a time bin) with a 95% confidence interval from $3 \times 10^{-5}$ to 0.001. The distribution was further truncated at 0.1 to avoid unrealistically high sampling probabilities. Finally, we added random gaps in preservation by setting the sampling probability to 0 in 10% of the bins. The simulated sampling rates in these simulations thus vary stochastically through time. The number of species in the present was randomly sampled as $N \sim \exp(\mathcal{U}(\log(100), \log(20{,}000)))$, and the variance of the Brownian bridge was sampled from $\sigma^2 \sim \mathcal{U}(10, 50) \times N$.

We additionally generated and analysed two sets of simulations (200 datasets each) in which we introduced a moderate and a strong trend towards increasing sampling rates through time. To simulate a moderate rate increase, after obtaining the vector of sampling rates as described above, we resampled them without replacement with a probability proportional to their value. Under this setting, the probability of choosing a rate equal to 0.1 as the first one is twice as large as the probability of choosing a rate of 0.05. Higher rates were thus more likely to appear among the first values in the re-ordered vector, while lower rates tended to be placed at the other end of it. The re-ordered rates were then assigned to each time bin starting from the most recent to the oldest one. To simulate strongly increasing rates through time, we repeated the procedure but set the resampling probability proportional to their value raised to the power of 5. Under this setting, the probability of choosing a rate equal to 0.1 as the first one is 32 times larger than the probability of choosing a rate of 0.05 ($0.1^5/0.05^5 = 32$), thus resulting in a stronger trend towards increasing rates through time.

**Empirical data from the angiosperm fossil record.** We compiled and analysed a database of 15,570 meso- and macrofossils of angiosperms spanning the Cretaceous and Cenozoic. We eschewed pollen records, which can be problematic to assign to extant families and require different sampling assumptions. However, we included six well-identified pollen records belonging to the families Aponogetonaceae, Araliaceae, Arecaceae and Asteraceae (Supplementary Table 3), as they provided early and reliable records for these clades. We repeated the analyses with and without these pollen data to identify their impact on the results.

The Cenozoic and Cretaceous data were obtained from the Cenozoic Angiosperm Database[64], and additional data were compiled from >700 publications (for detailed information, see Supplementary Table 3). As we were unable to evaluate in detail every fossil record and our analysis is sensitive to the earliest fossil record of each lineage, we cleaned the dataset in three steps. First, we carefully evaluated the earliest fossil record for each family on the basis of previous reviews[65,66] and removed unreliable records as well as putative angiosperm pollen from the Triassic and Jurassic[16,17,19,20]. Second, we removed occurrences that had not been identified to the species level or assigned to a family. Finally, we discarded extremely imprecisely dated fossil occurrences (those with an assigned age range larger than 20 Myr). The cleaned dataset, including the six pollen records, encompassed 14,571 occurrences of 5,780 unique species representative of 198 families, all of which are still extant. We compiled the modern diversity of the families (indicated by $N$ in equation (1)) on the basis of the Angiosperm Phylogeny Group IV classification[42] and a recent assessment of the number of known plant species[67].

We performed the analysis at the family level and estimated their time of origin. We chose to use families as the unit of our analyses because they represent clades[42] and as a way of accommodating taxon-specific heterogeneity in fossilization potential. We did not estimate the time of origin of angiosperms from the entire record because this would imply the same preservation potential for all flowering plants, and densely sampled groups would bias the overall sampling rates of angiosperms under our model parameterization. Instead, we used the age of the oldest family as an indirect estimate of the crown age of all angiosperms. We then binned the records using time bins of sizes 1, 2.5 and 5 Myr to assess the robustness of our results to different temporal resolutions.

On the basis of the estimated times of origin of the sampled families, we plotted the number of lineages through time. We quantified the rate of family-level diversification as the change in family diversity relative to the standing diversity standardized by time unit (1 Myr):

$$\left(\frac{d(t_{i+1})}{d(t_i)} - 1\right) \times (t_i - t_{i+1})^{-1}. \quad (4)$$

We computed diversification rates for each stage of the Cenozoic and Cretaceous. For earlier time intervals (Jurassic and Late Triassic), we computed the rates at a coarser temporal resolution (epochs), since these estimates are based on a limited number of lineages.

We compared the results obtained from the BBB method with confidence intervals on stratigraphic range data to infer the maximum plausible ages of origin of a lineage[26,27]. We compared these confidence intervals for the time of origin of a lineage with the 95% credible intervals obtained from the BBB posterior samples. Only families with more than one fossil occurrence could be analysed for stratigraphic confidence intervals ($N = 179$).

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

All data analysed in this study are available in Supplementary Table 3 and in a permanent Zenodo (zenodo.org) repository with doi: 10.5281/zenodo.4290423.

## Code availability

We implemented the BBB method in Python v.3. The code and input files are available in Supplementary Table 3 and in a permanent Zenodo (zenodo.org) repository with doi: 10.5281/zenodo.4290423. The code and input files and

any future updates of the program are additionally available as an open access repository: https://github.com/dsilvestro/rootBBB.

Received: 6 July 2020; Accepted: 17 December 2020;
Published online: 28 January 2021

## References


1. Clarke, J. T., Warnock, R. C. M. & Donoghue, P. C. J. Establishing a time-scale for plant evolution. *New Phytol.* **192**, 266–301 (2011).
2. Friis, E. M., Pedersen, K. R. & Crane, P. R. Diversity in obscurity: fossil flowers and the early history of angiosperms. *Phil. Trans. R. Soc. B* **365**, 369–382 (2010).
3. Coiro, M., Doyle, J. A. & Hilton, J. How deep is the conflict between molecular and fossil evidence on the age of angiosperms? *New Phytol.* **223**, 83–99 (2019).
4. Donoghue, P. Evolution: the flowering of land plant evolution. *Curr. Biol.* **29**, R738–R761 (2019).
5. Buggs, R. J. A. The deepening of Darwin's abominable mystery. *Nat. Ecol. Evol.* **1**, 0169 (2017).
6. Herendeen, P. S., Friis, E. M., Pedersen, K. R. & Crane, P. R. Palaeobotanical redux: revisiting the age of the angiosperms. *Nat. Plants* **3**, 17015 (2017).
7. Darwin, C. *More Letters of Charles Darwin* Vol. 2 (John Murray, 1903).
8. Friedman, W. E. The meaning of Darwin's 'abominable mystery'.*Am. J. Bot.* **96**, 5–21 (2009).
9. Marshall, C. R. Five paleobiological laws needed to understand the evolution of the living biota. *Nat. Ecol. Evol.* **1**, 0165 (2017).
10. Slater, G. & Harmon, L. J. Unifying fossils and phylogenies for comparative analyses of diversification and trait evolution. *Methods Ecol. Evol.* **4**, 699–702 (2013).
11. Didier, G., Royer-Carenzi, M. & Laurin, M. The reconstructed evolutionary process with the fossil record. *J. Theor. Biol.* **315**, 26–37 (2012).
12. Silvestro, D., Warnock, R. C. M., Gavryushkina, A. & Stadler, T. Closing the gap between palaeontological and neontological speciation and extinction rate estimates. *Nat. Commun.* **9**, 5237 (2018).
13. Stadler, T., Gavryushkina, A., Warnock, R. C. M., Drummond, A. J. & Heath, T. A. The fossilized birth–death model for the analysis of stratigraphic range data under different speciation concepts. *J. Theor. Biol.* **447**, 41–55 (2018).
14. Budd, G. E. & Mann, R. P. The dynamics of stem and crown groups. *Sci. Adv.* **6**, 1626 (2020).
15. Seward, A. C. The Jurassic flora II. Liassic and Oolitic floras of England. In *Catalogue of the Mesozoic plants in the Department of Geology, British Museum (National History)* (British Museum, 1904).
16. Cornet, B. Late Triassic angiosperm-like pollen from the Richmond Rift Basin of Virginia, U.S.A. *Palaeontogr. Abt. B* **213**, 37–87 (1989).
17. Ren, D. Flower-associated Brachycera flies as fossil evidence for Jurassic angiosperm origins. *Science* **280**, 85–88 (1998).
18. Cleal, C. J. & Rees, P. M. The Middle Jurassic flora from Stonesfield, Oxfordshire, UK. *Palaeontology* **46**, 739–801 (2003).
19. Hochuli, P. A. & FeistBurkhardt, S. A boreal early cradle of angiosperms? Angiosperm-like pollen from the Middle Triassic of the Barents Sea (Norway). *J. Micropalaeontol.* **23**, 97–104 (2004).
20. Hochuli, P. A. & FeistBurkhardt, S. Angiosperm-like pollen and *Afropollis* from the Middle Triassic (Anisian) of the Germanic Bascin (northern Switzerland). *Front. Plant Sci.* **4**, e344 (2013).
21. Bell, C. D., Soltis, D. E. & Soltis, P. S. The age and diversification of the angiosperms re-visited. *Am. J. Bot.* **97**, 1296–1303 (2010).
22. Li, H.-T. et al. Origin of angiosperms and the puzzle of the Jurassic gap. *Nat. Plants* **5**, 461–470 (2019).
23. Friis, E. M., Crane, P. R., Pedersen, K. R., Stampanoni, M. & Marone, F. Exceptional preservation of tiny embryos documents seed dormancy in early angiosperms. *Nature* **528**, 551–554 (2018).
24. Doyle, J. A. Molecular and fossil evidence on the origin of angiosperms. *Annu. Rev. Earth Planet. Sci.* **40**, 301–326 (2012).
25. Barba-Montoya, J., dosReis, M., Schneider, H., Donoghue, P. C. J. & Yang, Z. Constraining uncertainty in the timescale of angiosperm evolution and the veracity of a Cretaceous terrestrial revolution. *New Phytol.* **218**, 819–834 (2018).
26. Strauss, D. & Sadler, P. M. Classical confidence-intervals and Bayesian probability estimates for ends of local taxon ranges. *Math. Geol.* **21**, 411–421 (1989).
27. Marshall, C. R. Confidence-intervals on stratigraphic ranges. *Paleobiology* **16**, 1–10 (1990).
28. Marshall, C. R. Confidence intervals on stratigraphic ranges with nonrandom distributions of fossil horizons. *Paleobiology* **23**, 165–173 (1997).
29. Silvestro, D., Salamin, N., Antonelli, A. & Meyer, X. Improved estimation of macroevolutionary rates from fossil data using a Bayesian framework. *Paleobiology* **45**, 546–570 (2019).
30. Warnock, R. C., Heath, T. A. & Stadler, T. Assessing the impact of incomplete species sampling on estimates of speciation and extinction rates. *Paleobiology* **46**, 137–157 (2020).
31. Silvestro, D., Cascales-Miñana, B., Bacon, C. D. & Antonelli, A. Revisiting the origin and diversification of vascular plants through a comprehensive Bayesian analysis of the fossil record. *New Phytol.* **207**, 425–436 (2015).
32. Nowak, H., Schneebeli-Hermann, E. & Kustatscher, E. No mass extinction for land plants at the Permian–Triassic transition. *Nat. Commun.* **10**, 384 (2019).
33. Hedman, M. M. Constraints on clade ages from fossil outgroups. *Paleobiology* **36**, 16–31 (2010).
34. Lloyd, G. T., Bapst, D. W., Friedman, M. & Davis, K. E. Probabilistic divergence time estimation without branch lengths: dating the origins of dinosaurs, avian flight and crown birds. *Biol. Lett.* **12**, 20160609 (2016).
35. Gavryushkina, A. et al. Bayesian total-evidence dating reveals the recent crown radiation of penguins. *Syst. Biol.* **66**, 57–73 (2017).
36. Budd, G. E. & Mann, R. P. History is written by the victors: the effect of the push of the past on the fossil record. *Evolution* **72**, 2276–2291 (2018).
37. Tanner, M. & Wing, H. The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* **82**, 528–540 (1987).
38. Holland, S. M. The non-uniformity of fossil preservation. *Phil. Trans. R. Soc. B* **371**, 20150130 (2016).
39. Pimiento, C. et al. The Pliocene marine megafauna extinction and its impact on functional diversity. *Nat. Ecol. Evol.* **1**, 1100–1106 (2017).
40. Brocklehurst, N., Upchurch, P., Mannion, P. D. & O'Connor, J. The completeness of the fossil record of Mesozoic birds: implications for early avian evolution. *PLoS ONE* **7**, e39056 (2012).
41. Marshall, C. R. Using the fossil record to evaluate timetree timescales. *Front. Genet.* **10**, 449 (2019).
42. Angiosperm Phylogeny Group et al. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linn. Soc.* **181**, 1–20 (2016).
43. Magallón, S., Gòmez-Acevedo, S., Sánchez-Reyes, L. L. & Hernández-Hernández, T. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytol.* **207**, 437–453 (2015).
44. Müller, J. Significance of fossil pollen for angiosperm history. *Ann. Mo. Bot. Gard.* **71**, 419–443 (1984).
45. Collinson, M. E., Boulter, M. C. & Holmes, P. L. *The Fossil Record 2* (ed. Benton, M. J.) 809–841 (Chapman and Hall, 1993).
46. Doyle, J. A. & Endress, P. K. Integrating Early Cretaceous fossils into the phylogeny of living angiosperms: ANITA lines and relatives of Chloranthaceae. *Int. J. Plant Sci.* **175**, 555–600 (2014).
47. Doyle, J. A. Recognising angiosperm clades in the Early Cretaceous fossil record. *Hist. Biol.* **27**, 414–429 (2015).
48. Coiro, M., Martínez, L. C. A., Upchurch, G. R. & Doyle, J. A. Evidence for an extinct lineage of angiosperms from the Early Cretaceous of Patagonia and implications for the early radiation of flowering plants. *New Phytol.* **228**, 344–360 (2020).
49. Beaulieu, J. M., O'Meara, B. C., Crane, P. & Donoghue, M. J. Heterogeneous rates of molecular evolution and diversification could explain the Triassic age estimate for angiosperms. *Syst. Biol.* **64**, 869–878 (2015).
50. Louca, S. & Pennell, M. W. Extant timetrees are consistent with a myriad of diversification histories. *Nature* **580**, 502–505 (2020).
51. Foote, M. On the probability of ancestors in the fossil record. *Paleobiology* **22**, 141–151 (1996).
52. Drummond, A. J., Ho, S., Phillips, M. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88 (2006).
53. Ronquist, F. et al. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. *Syst. Biol.* **61**, 973–999 (2012).
54. van der Kooi, C. J. & Ollerton, J. The origins of flowering plants and pollinators. *Science* **368**, 1306–1308 (2020).
55. Bateman, R. M. Hunting the Snark: the flawed search for mythical Jurassic angiosperms. *J. Exp. Bot.* **71**, 22–35 (2020).
56. Smith, S. A., Beaulieu, J. M. & Donoghue, M. J. An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants. *Proc. Natl Acad. Sci. USA* **107**, 5897–5902 (2010).
57. Zhang, L. et al. The water lily genome and the early evolution of flowering plants. *Nature* **577**, 79–84 (2020).
58. Warnock, R. C. M., Parham, J. F., Joyce, W. G., Tyler, R. L. & Donoghue, P. C. J. Calibration uncertainty in molecular dating analyses: there is no substitute for the prior evaluation of time priors. *Proc. R. Soc. B* **282**, 20141013 (2015).
59. Ronquist, F., Lartillot, N. & Phillips, M. J. Closing the gap between rocks and clocks using total-evidence dating. *Phil. Trans. R. Soc. B* **371**, 20150136 (2016).
60. Feild, T. S., Arens, N. C., Doyle, J. A., Dawson, T. E. & Donoghue, M. J. Dark and disturbed: a new image of early angiosperm ecology. *Paleobiology* **30**, 82–107 (2004).

61. Ramírez-Barahona, S., Sauquet, H. & Magallón, S. The delayed and geographically heterogeneous diversification of flowering plant families. *Nat. Ecol. Evol.* **4**, 1232–1238 (2020).
62. Sokoloff, D. D., Remizowa, M. V., El, E. S., Rudall, P. J. & Bateman, R. M. Supposed Jurassic angiosperms lack pentamery, an important angiosperm-specific feature. *New Phytol.* **228**, 420–426 (2020).
63. Cascales-Miñana, B., Cleal, C. J. & Gerrienne, P. Is Darwin's 'abominable mystery' still a mystery today? *Cretac. Res.* **61**, 256–262 (2016).
64. Xing, Y. et al. Testing the biases in the rich Cenozoic angiosperm macrofossil record. *Int. J. Plant Sci.* **177**, 371–388 (2016).
65. Friis, E. M., Crane, P. R. & Pedersen, K. R. *Early Flowers and Angiosperm Evolution* (Cambridge Univ. Press, 2011).
66. Manchester, S. R., Grímsson, F. & Zetter, R. Assessing the fossil record of asterids in the context of our current phylogenetic framework. *Ann. Mo. Bot. Gard.* **100**, 329–363 (2015).
67. Christenhusz, M. J. M. & Byng, J. W. The number of known plants species in the world and its annual increase. *Phytotaxa* **261**, 201–217 (2016).

## Acknowledgements

We thank R. C. M. Warnock, T. Stadler's lab and E. Carlisle for feedback on the methods and models presented here. We also thank P. R. Crane for constructive feedback on the manuscript. D.S. received funding from the Swiss National Science Foundation (grant no. PCEFP3_187012) and from the Swedish Research Council (grant no. 2019-04739). A.A. acknowledges financial support from the Swedish Research Council (grant no. 2019-05191), the Swedish Foundation for Strategic Research (grant no. FFL15-0196), the Knut and Alice Wallenberg Foundation (grant no. KAW 2014.0216) and the Royal Botanic Gardens, Kew. Y.X. received funding from the National Natural Science Foundation of China (grant nos 31770226 and U1802242) and the Strategic Priority Research Program of the Chinese Academy of Sciences (grant no. XDB31000000).

## Author contributions

D.S., C.D.B. and Y.X. conceived the study. W.D., Q.Z. and Y.X. compiled the fossil data. D.S. developed and implemented the methods and analysed the data. D.S. wrote the manuscript with contributions from C.D.B., W.D., Q.Z., P.C.J.D., A.A. and Y.X.

## Competing interests

The authors declare no competing interests.

## Additional information

**Extended data** is available for this paper at https://doi.org/10.1038/s41559-020-01387-8.

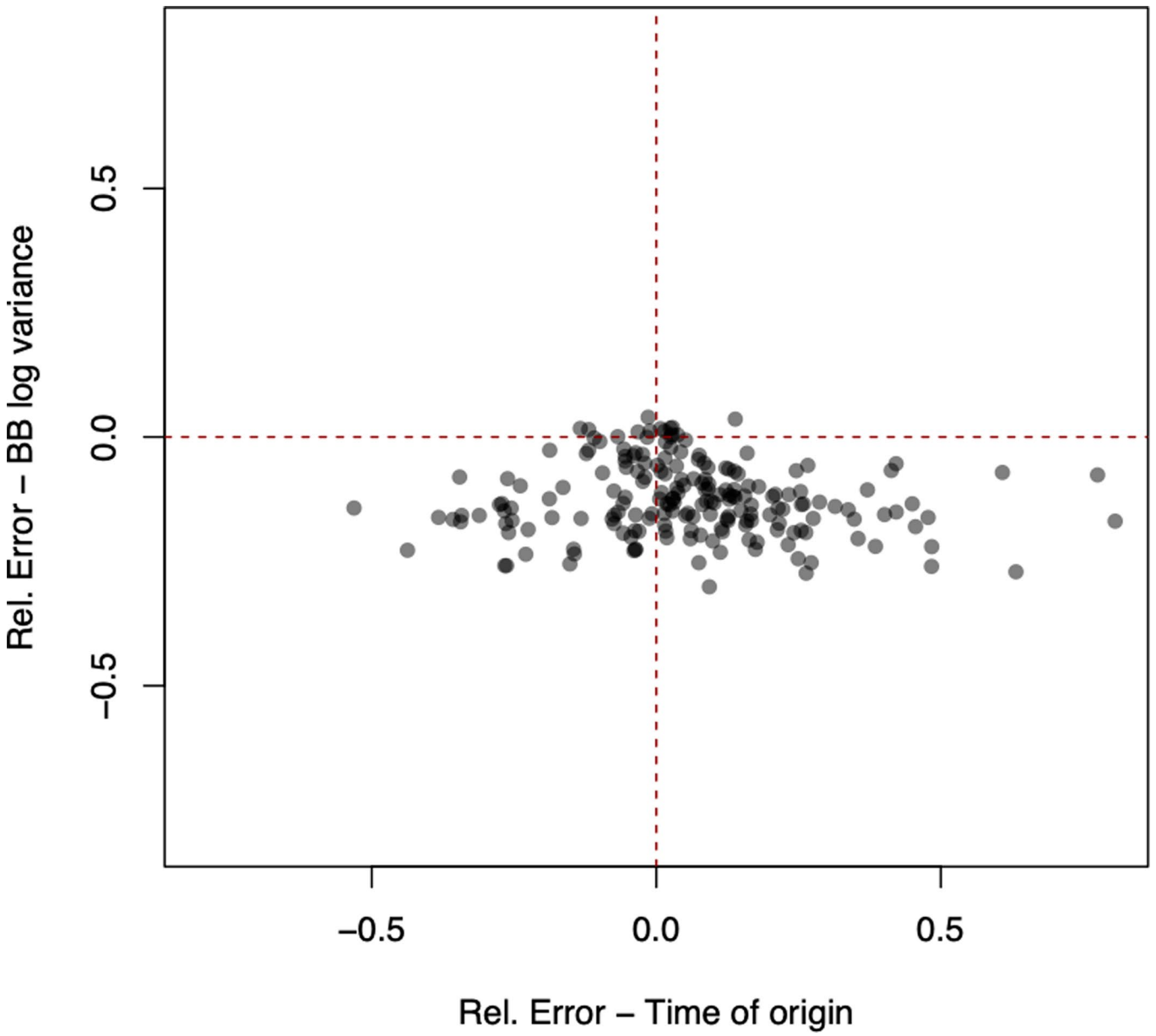
**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41559-020-01387-8.

**Correspondence and requests for materials** should be addressed to D.S.

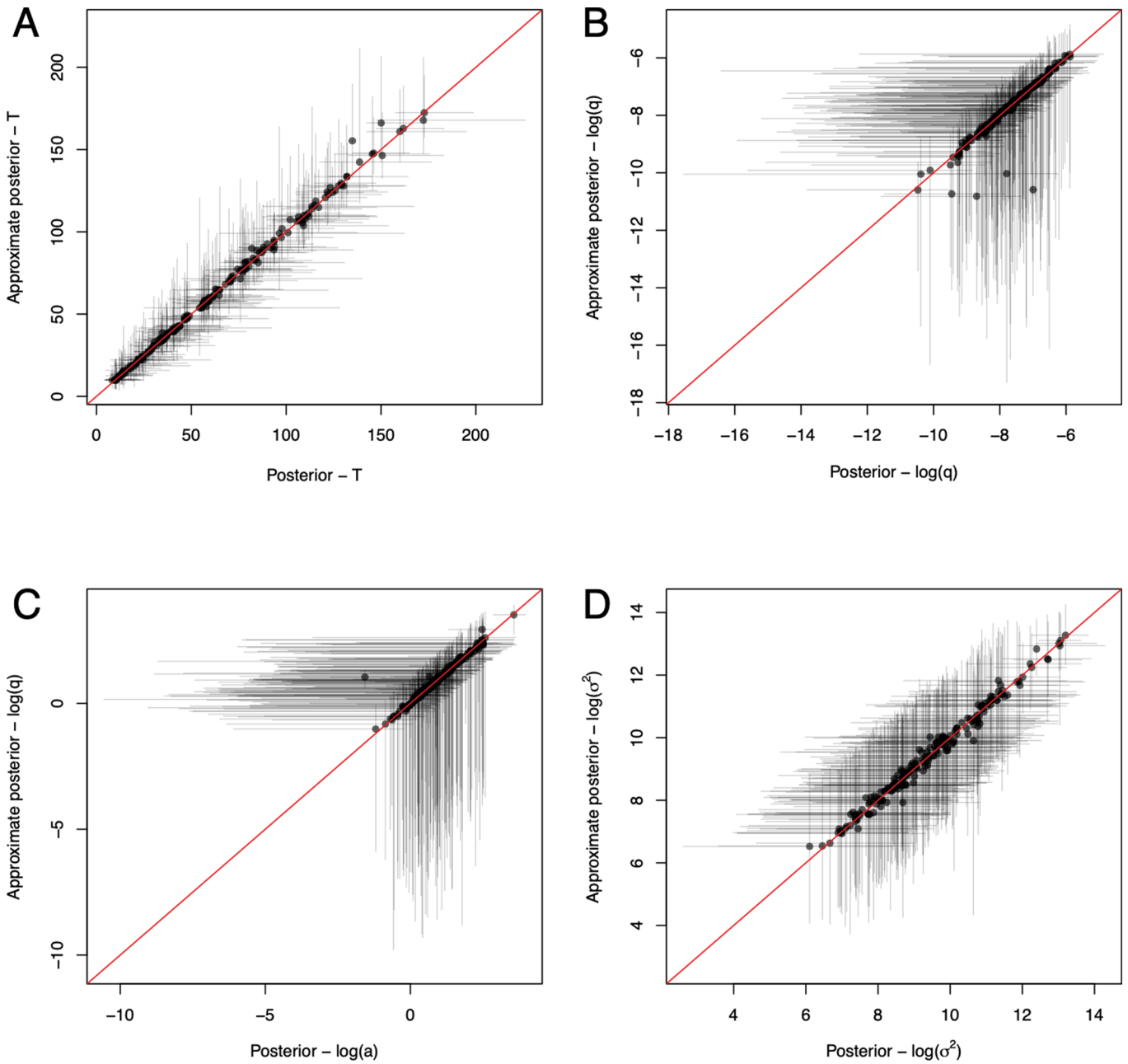
**Peer review information** *Nature Ecology & Evolution* thanks Pamela Soltis, David Cerny and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at www.nature.com/reprints.

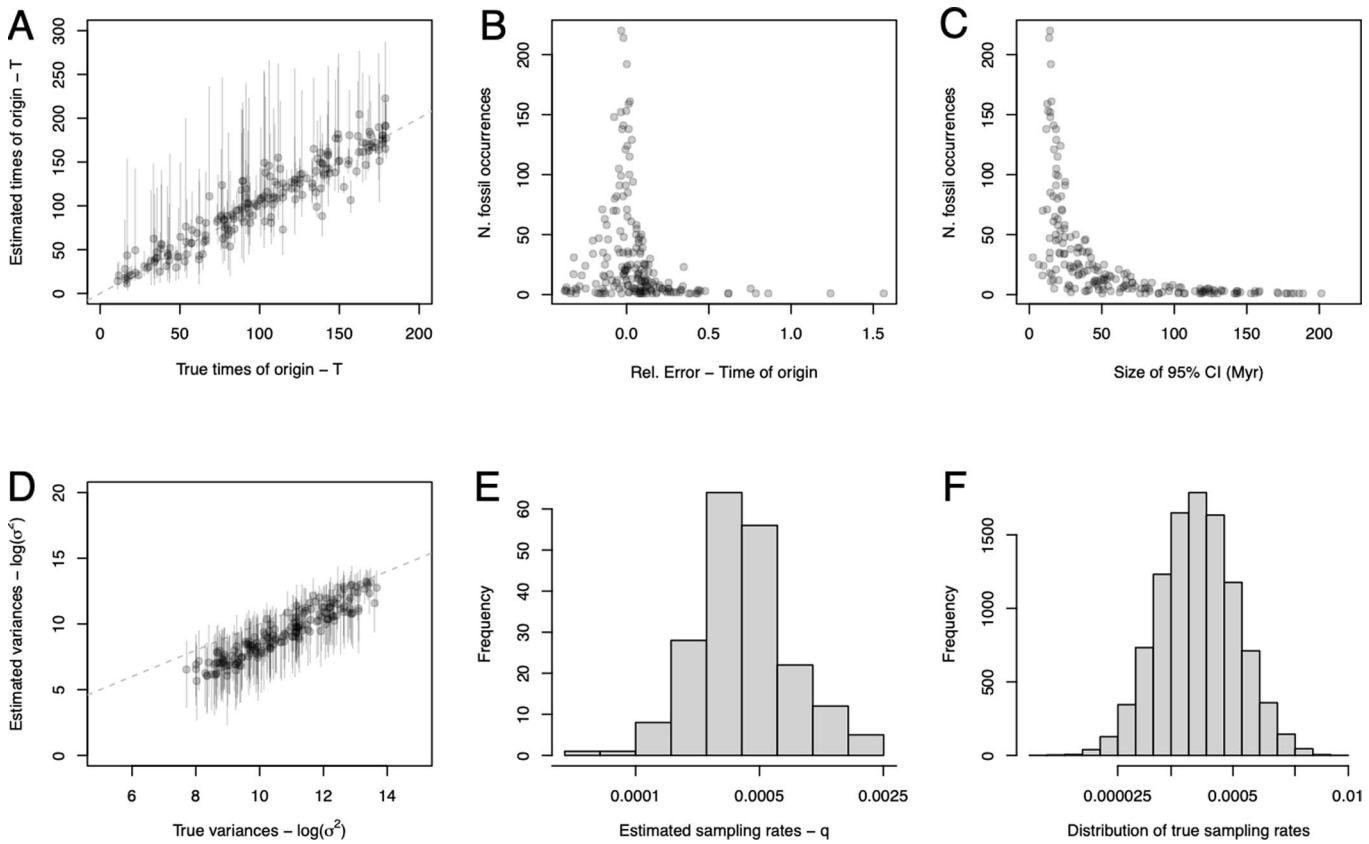
**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



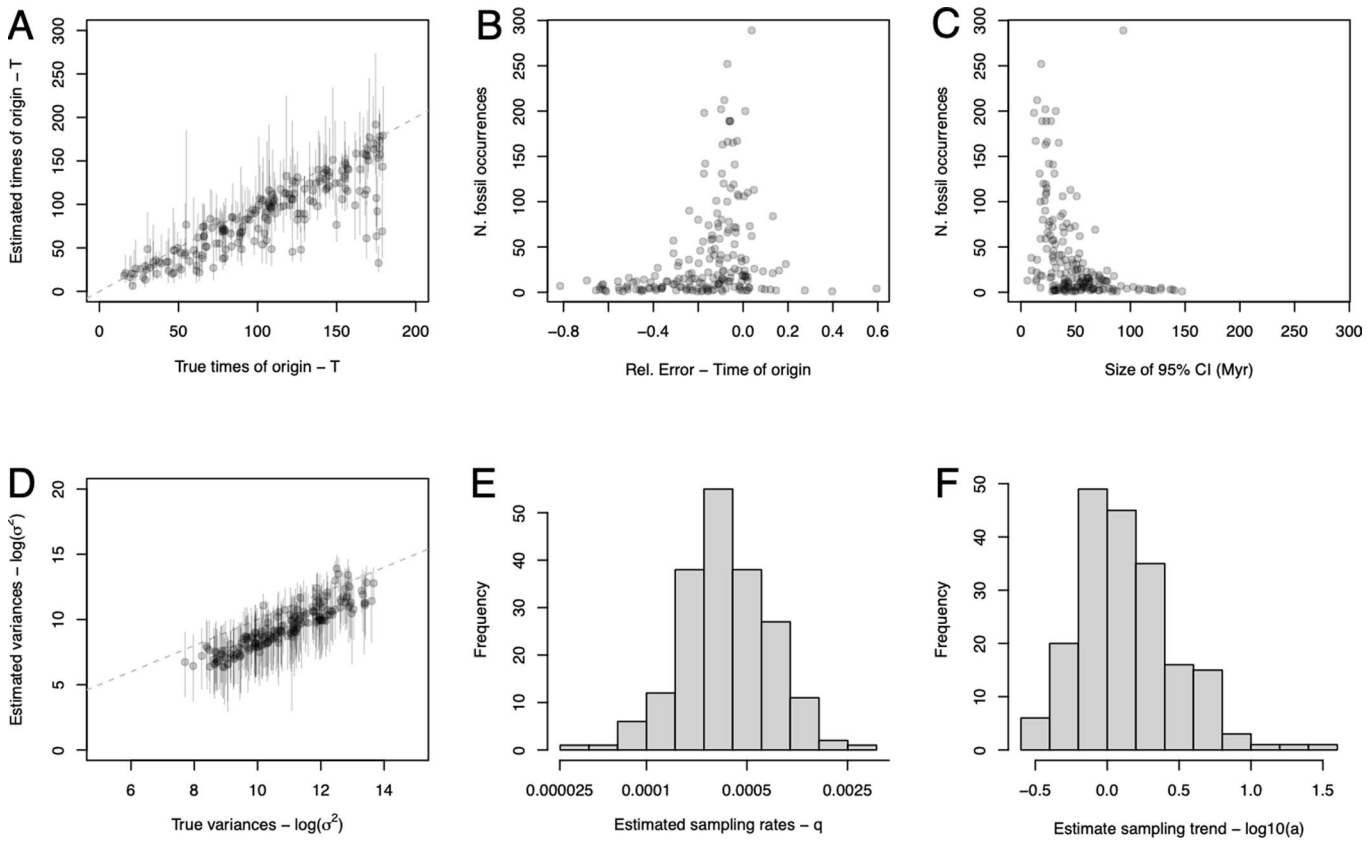

**Extended Data Fig. 1 | Relative errors of the estimated Brownian bridge log variances plotted against the relative error of the estimated time of origin based on 200 simulations.** While log variances tended to be slightly underestimated (mostly negative relative errors) they do not have a biasing effect on the estimated times of origin, which show an unbiased error around zero (see also Fig. 2, main text).



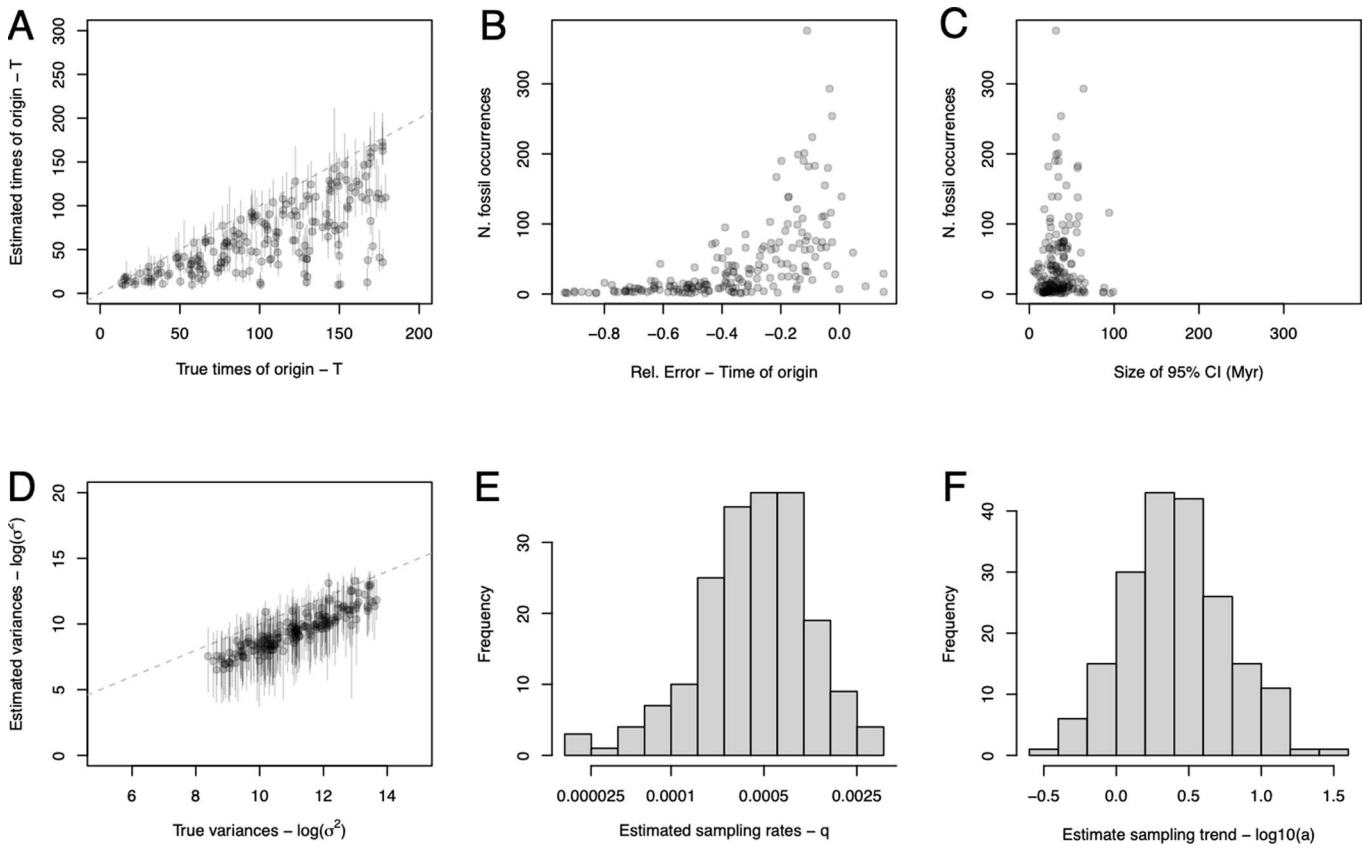

**Extended Data Fig. 2 | Parameter estimates from 200 simulated datasets obtained under MCMC and an approximated MCMC.** In the approximated MCMC, a fraction of the iterations involve no parameter updates (that is $q_7, a, T$, and $\sigma^2$ do not change), but a new set of conditional Brownian bridges are drawn and accepted as samples from the approximate posterior. This procedure was found to improve the convergence of the MCMC, while having negligible effect on the estimated time of origin **a**, and sampling rates **b**, rate trend **c**, and log variance **d**,.



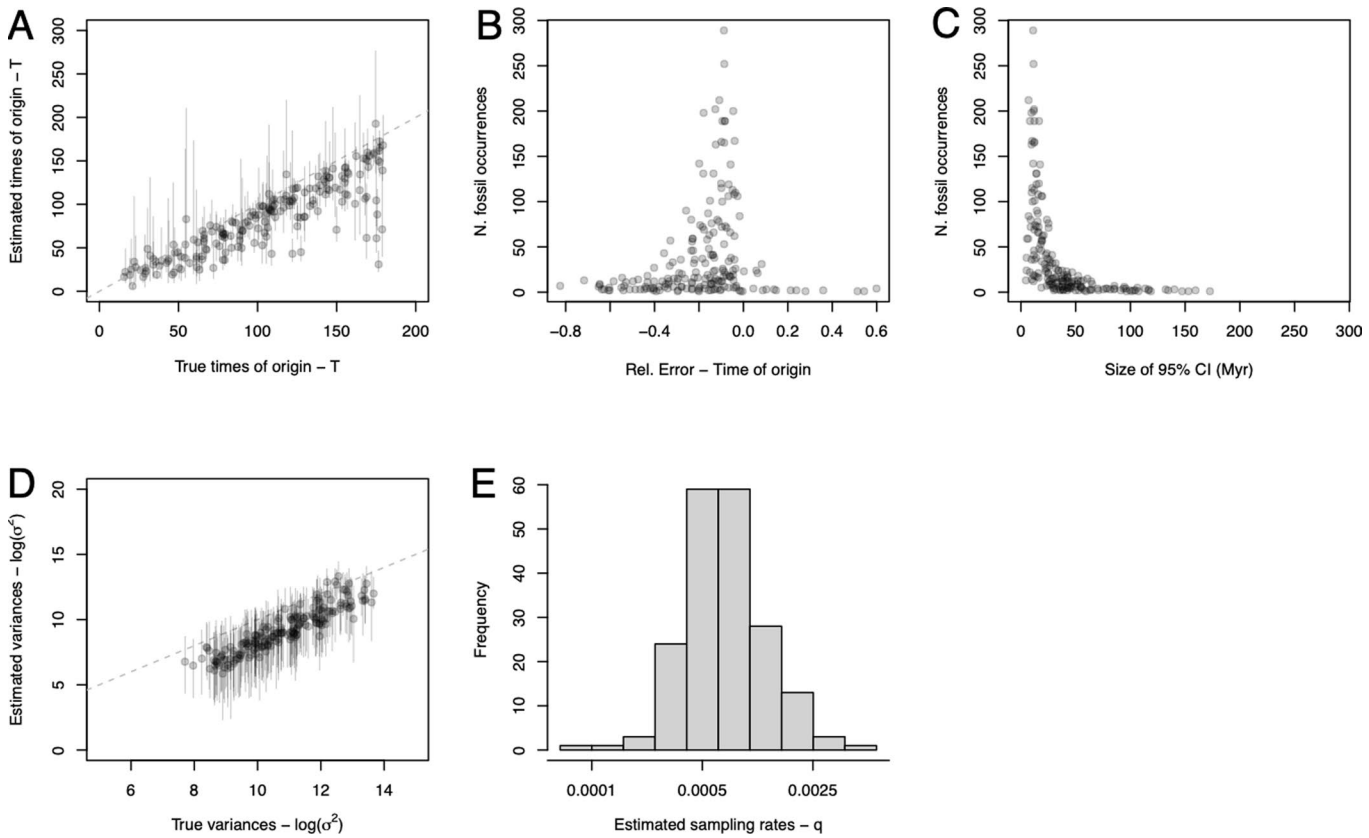

**Extended Data Fig. 3 | Analysis of 200 simulated datasets with random varying sampling rates through time using a BBB model with constant sampling rate ($a = 0$).** The times of origin were accurately estimated (**a**); circles and bars indicate posterior estimates and 95% credible intervals. The relative errors on the time of origin were smaller in datasets with richer simulated fossil record (**b**). The size of the 95% credible intervals around the times of origin decreased with increasing numb er of fossils (**c**). The log variances were slightly underestimated (**d**), while the estimated sampling rates (**e**; the X-axis is $\log_{10}$-transformed) cannot be plotted against true values because the underlying simulations were based on time-heterogeneous sampling with di?erent rates in each time bin. However, we plot for comparison the distribution from which sampling rates were sampled, randomly for each time bin (**f**; the X-axis is log 10 -transformed).



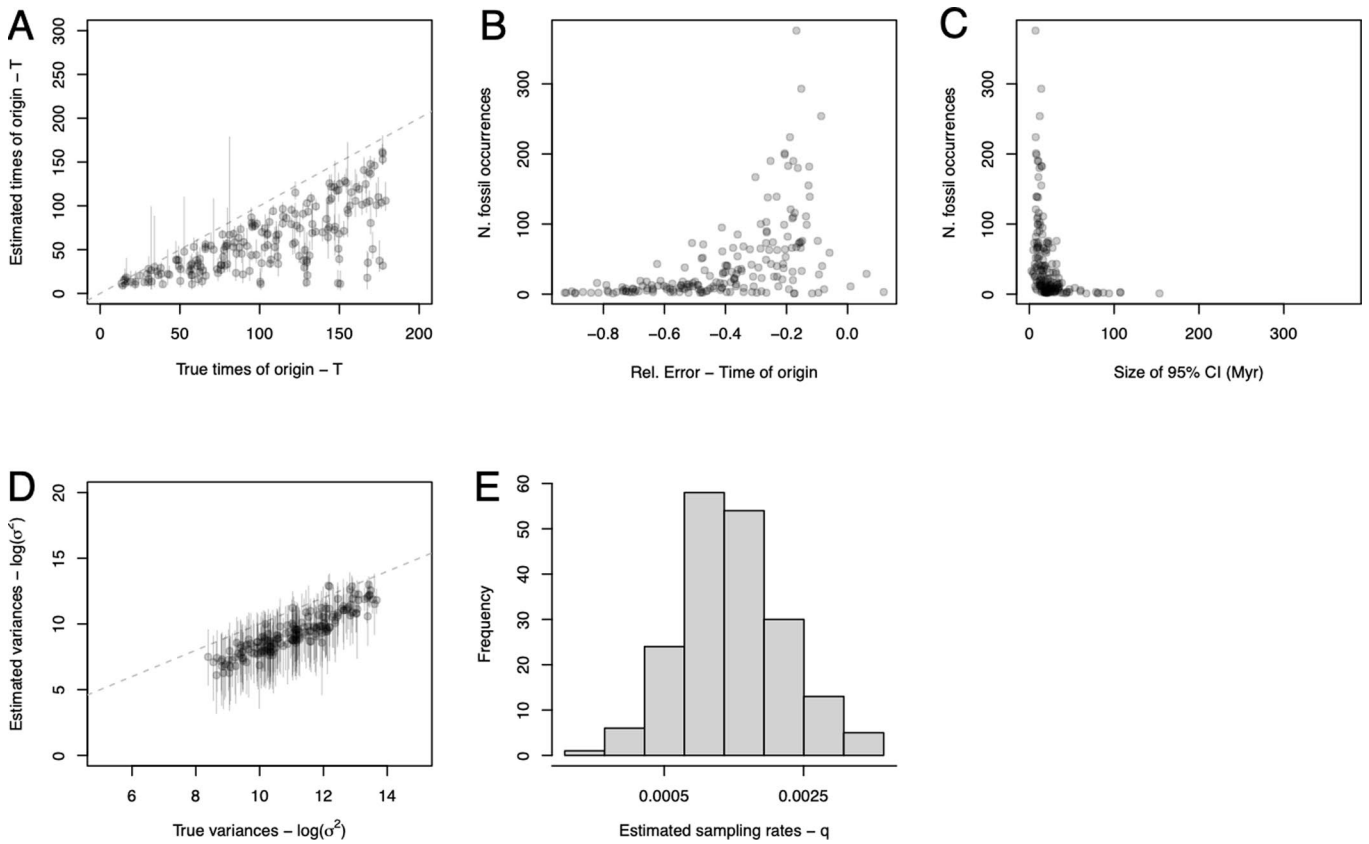

**Extended Data Fig. 4 | Analysis of 200 simulated datasets with sampling rates moderately increasing through time using a BBB model with time-varying sampling rates.** The times of origin were underestimated in some cases (**a**); circles and bars indicate posterior estimates and 95% credible intervals. The relative errors on the time of origin were smaller in datasets with richer simulated fossil record (**b**). The size of the 95% credible intervals around the times of origin decreased with increasing numb er of fossils (**c**). The log variances were slightly underestimated (**d**), while the estimated sampling rates at the time of origin and rate trends (**e** and **f**, respectively; the X-axis is $\log_{10}$-transformed) cannot be plotted against true values because they do not have a direct equivalent in the underlying simulations. The distribution from which sampling rates were sampled for each time bin is shown for reference in Extended Data Fig. 3f.



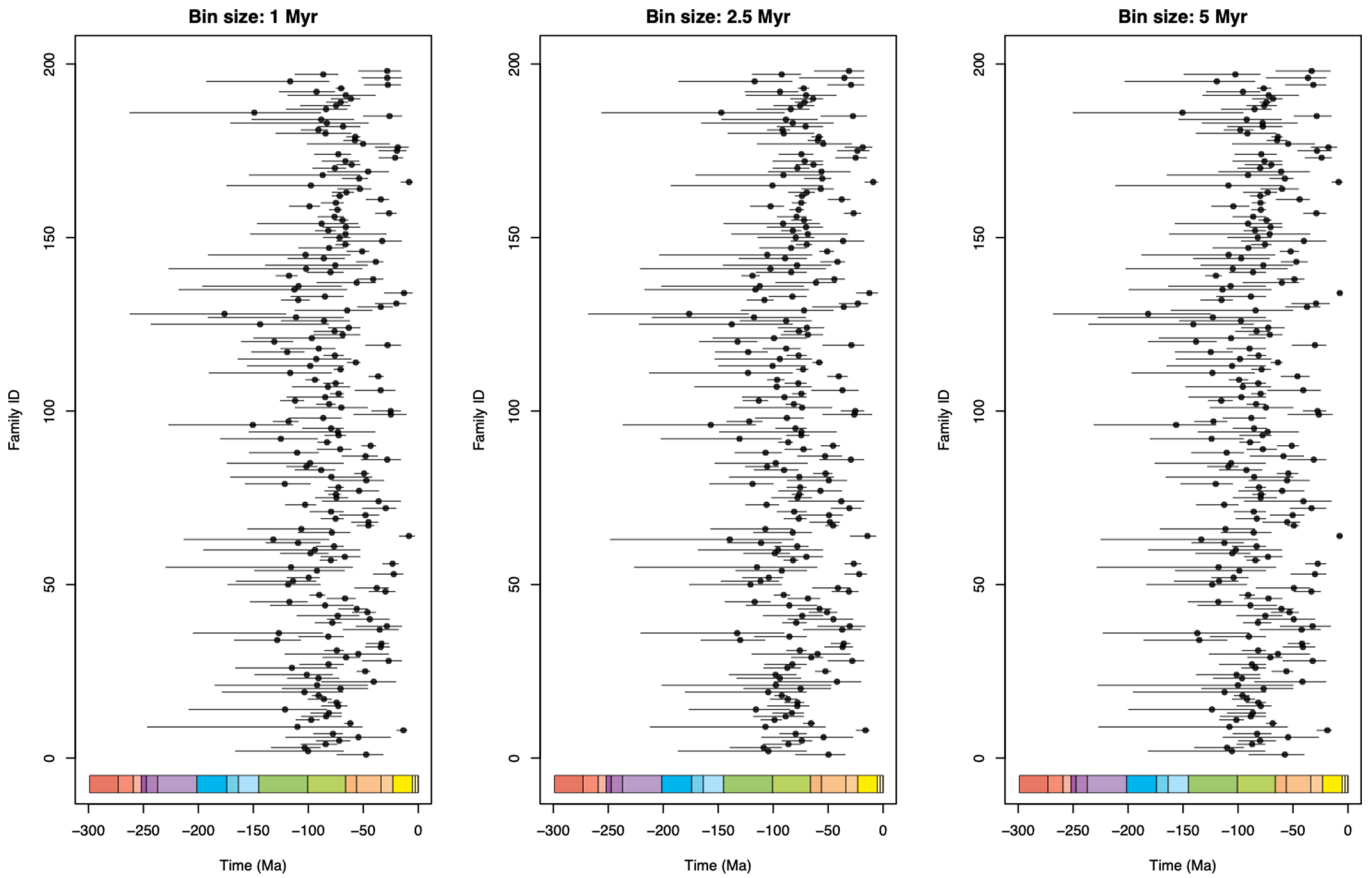

**Extended Data Fig. 5 | Analysis of 200 simulated datasets with sampling rates strongly increasing through time using a BBB model with time-varying sampling rates.** The times of origin were frequently underestimated (**a**); circles and bars indicate posterior estimates and 95% credible intervals. The relative errors on the time of origin were smaller in datasets with richer simulated fossil record (**b**). The size of the 95% credible intervals around the times of origin decreased with increasing numb er of fossils (**c**). The log variances were slightly underestimated (**d**), while the estimated sampling rates at the time of origin and rate trends (**e** and **f**, respectively; the X-axis is $\log_{10}$-transformed) cannot be plotted against true values because they do not have a direct equivalent in the underlying simulations. The distribution from which sampling rates were sampled for each time bin is shown for reference in Extended Data Fig. 3f.



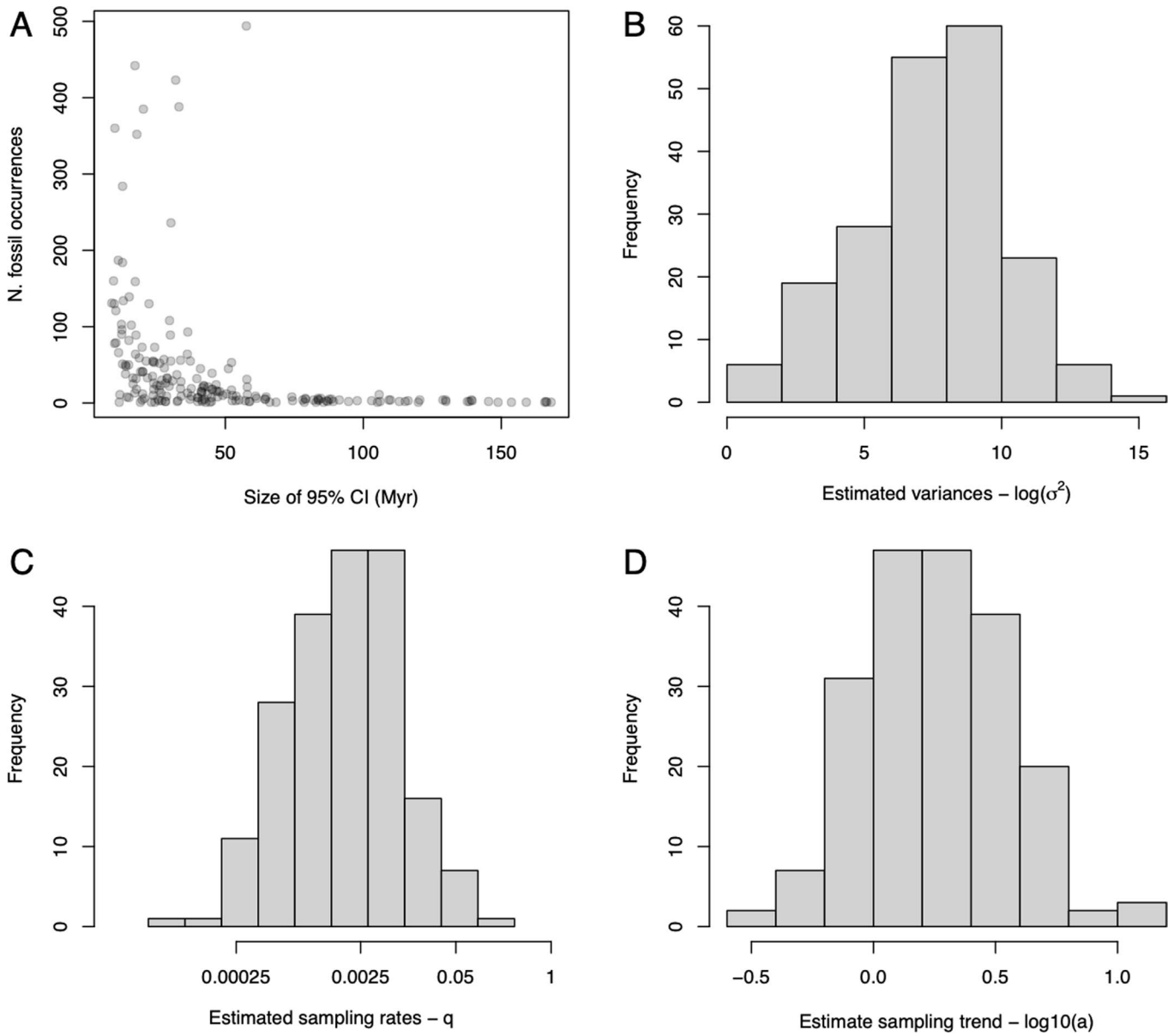

**Extended Data Fig. 6 | Analysis of 200 simulated datasets with sampling rates moderately increasing through time using a BBB model with constant sampling rate.** The times of origin were frequently underestimated (**a**); circles and bars indicate posterior estimates and 95% credible intervals. The relative errors on the time of origin were smaller in datasets with richer simulated fossil record (**b**). The size of the 95% credible intervals around the times of origin decreased with increasing number of fossils (**c**). The log variances were slightly underestimated (**d**), while the estimated sampling rate (**e**; the X-axis is $\log_{10}$-transformed) cannot be plotted against true values because it does not have a direct equivalent in the underlying simulations. The distribution from which sampling rates were sampled for each time bin is shown for reference in Extended Data Fig. 3f.



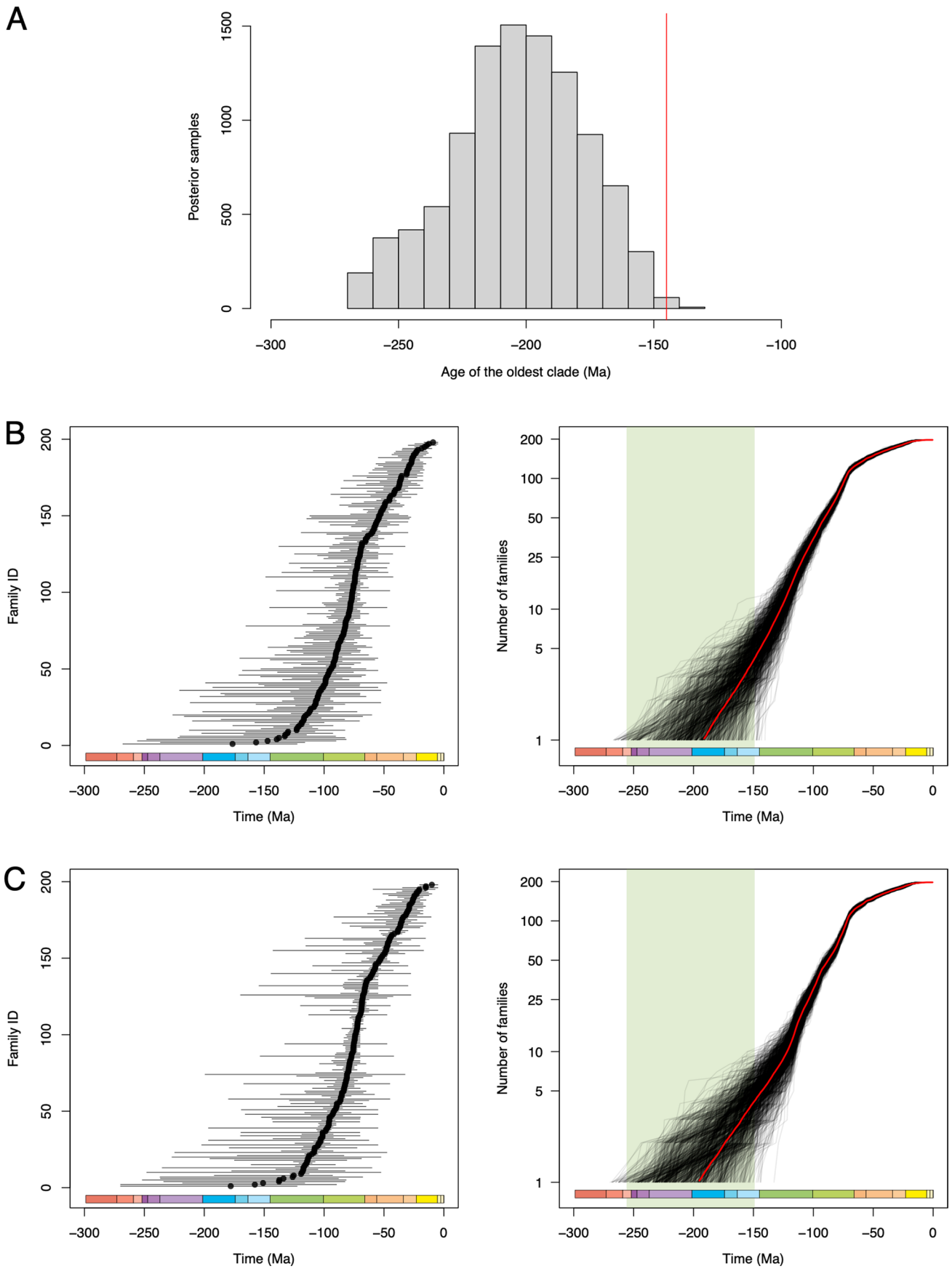

**Extended Data Fig. 7 | Analysis of 200 simulated datasets with sampling rates strongly increasing through time using a BBB model with constant sampling rate.** The times of origin were consistently underestimated (**a**); circles and bars indicate posterior estimates and 95% CI. The relative errors on the time of origin were smaller in datasets with richer simulated fossil record (**b**). The size of the 95% credible intervals around the times of origin decreased with increasing number of fossils (**c**). The log variances were slightly underestimated (**d**), while the estimated sampling rate (**e**; the X-axis is $\log_{10}$-transformed) cannot be plotted against true values because it does not have a direct equivalent in the underlying simulations. The distribution from which sampling rates were sampled for each time bin is shown for reference in Extended Data Fig. 3f.




**Extended Data Fig. 8 | Family-level origination times inferred using bin sizes equal to 1, 2.5, and 5 Myr.** The estimated times of origin and credible intervals were highly consistent across different settings.




**Extended Data Fig. 9 | Parameters estimated across angiosperm families. a**, Size of the 95% credible intervals for the estimated time of origin of angiosperm families plotted against the number of fossils available: the relationship reflects the observations based on simulated data. Increasing number of fossils results in substantially smaller credible intervals. **b**, Distributions of estimated variances of the Brownian bridge ($\sigma^2$; log-scale), **c**, sampling rates at the time of origin ($q_T$; log-scale), and **d**, sampling temporal trend (a; log-scale) as inferred across angiosperm families.




**Extended Data Fig. 10 | See next page for caption.**

**Extended Data Fig. 10 | Estimated origination times across angiosperm families. a**, Posterior samples of the oldest time of origin across all families obtained after combining the estimated ages of each. The red line indicates the boundary between the Jurassic and the Cretaceous. Only 0.2% of the samples fall within the Cretaceous providing strong statistical evidence for an earlier origin of crown angiosperm. **b**, Root age estimates of extant families of angiosperm with 95% credible intervals (left) as inferred from meso- and macrofossils only, excluding pollen data and cumulative family diversity (right) based on those estimates (Y-axis is log10 transformed). The analyses we run under a BBB model with time-increasing sampling rates. **c**, Root age estimates of extant families of angiosperm with 95% credible intervals (left) as inferred from a BBB model with sampling rate set to be constant (parameter $a = 0$) and cumulative family diversity (right) based on those estimates (Y-axis is $\log_{10}$-transformed).

nature research

Corresponding author(s): Daniele Silvestro

Last updated by author(s): Dec 7, 2020

# Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed | |
|---|---|---|
| ☐ | ☒ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☒ | ☐ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☒ | ☐ | The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☒ | ☐ | A description of all covariates tested |
| ☒ | ☐ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☒ | ☐ | For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☐ | ☒ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☐ | ☒ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | ☐ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | All the data and codes used in this paper are available in a permanent Zenodo repository (doi: 10.5281/zenodo.4290423) and on GitHub (https://github.com/dsilvestro/rootBBB). |
| Data analysis | All the data and codes used in this paper are available in a permanent Zenodo repository (doi: 10.5281/zenodo.4290423) and on GitHub (https://github.com/dsilvestro/rootBBB). |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All the data and codes used in this paper are available in a permanent Zenodo repository (doi: 10.5281/zenodo.4290423) and on GitHub (https://github.com/dsilvestro/rootBBB).

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Study description | Using a novel Bayesian method, we infer a pre-Cretaceous origin of multiple angiosperm families based on their fossil record and present day diversity |
| Research sample | Published fossil records of flowering plants |
| Sampling strategy | literature review |
| Data collection | NA |
| Timing and spatial scale | NA |
| Data exclusions | NA |
| Reproducibility | All the data and codes used in this paper are available in a permanent Zenodo repository (doi: 10.5281/zenodo.4290423) and on GitHub (https://github.com/dsilvestro/rootBBB). |
| Randomization | NA |
| Blinding | NA |

Did the study involve field work? ☐ Yes ☒ No

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology and archaeology |
| ☒ | ☐ Animals and other organisms |
| ☒ | ☐ Human research participants |
| ☒ | ☐ Clinical data |
| ☒ | ☐ Dual use research of concern |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |