A large-scale species level dated angiosperm phylogeny for evolutionary and ecological analyses

Steven B. Janssens[‡,§], Thomas L.P. Couvreur[|], Arne Mertens[‡], Gilles Dauby[¶], Leo-Paul M. J. Dagallier[|], Samuel Vanden Abeele[‡], Filip Vandelook[‡], Maurizio Mascarello[‡], Hans Beeckman[#], Marc Sosef[‡], Vincent Droissart[¶], Michelle van der Bank[¤], Olivier Maurin[«], William Hawthorne[»], Cicely Marshall[^], Maxime Réjou-Méchain[¶], Denis Beina[˘], Fidele Baya[¦], Vincent Merckx[ʔ,ʕ], Brecht Verstraete[¢], Olivier Hardy[ℓ]

‡ Botanic Garden Meise, Meise, Belgium
§ Laboratory for Plant Conservation and Population Biology, KULeuven, Leuven, Belgium
| DIADE, IRD, Univ. Montpellier, Montpellier, France
¶ AMAP Lab, IRD, CIRAD, CNRS, INRA, Univ Montpellier, Montpellier, France
# RMCA, Tervuren, Belgium
¤ University of Johannesburg, Johannesburg, South Africa
« Royal Botanic Gardens, Kew, United Kingdom
» Department of Plant Sciences, University of Oxford, Oxford, United Kingdom
^ Department of Plant Sciences, University of Cambridge, Cambridge, United Kingdom
˘ Université de Bangui – Cerphameta, Bangui, Central African Republic
¦ Ministère des Eaux, Forêts, Chasse et Pêche, Bangui, Central African Republic
ʔ Understanding Evolution Group, Naturalis Biodiversity Center, Leiden, Netherlands
ʕ Department of Evolutionary and Population Biology, University of Amsterdam, Amsterdam, Netherlands
¢ Natural History Museum, University of Oslo, Oslo, Norway
ℓ Universite Libre de Bruxelles, Brussels, Belgium

Corresponding author: Steven B. Janssens (steven.janssens@plantentuinmeise.be)

Academic editor: Stephen Boatwright

Abstract

Phylogenies are a central and indispensable tool for evolutionary and ecological research. Even though most angiosperm families are well investigated from a phylogenetic point of view, there are far less possibilities to carry out large-scale meta-analyses at order level or higher. Here, we reconstructed a large-scale dated phylogeny including nearly 1/8th of all angiosperm species, based on two plastid barcoding genes, *matK* (incl. *trnK*) and *rbcL*. Novel sequences were generated for several species, while the rest of the data were mined from GenBank. The resulting tree was dated using 56 angiosperm fossils as calibration points. The resulting megaphylogeny is one of the largest dated phylogenetic tree of angiosperms yet, consisting of 36,101 sampled species, representing 8,399 genera, 426 families and all orders. This novel framework will be useful for investigating different broad scale research questions in ecological and evolutionary biology.

Keywords

phylogeny, angiosperms, large-scale dating analyses, evolution, ecology

Introduction

During the past two decades, awareness has grown that ecological and evolutionary studies benefit from incorporating phylogenetic information (Wanntorp et al. 1990, Webb et al. 2002). In some ecological disciplines, it has even become almost unimaginable that a spatiotemporal context is not considered when specific hypotheses are tested. For example, in the fields of community ecology, trait-based ecology and macroecology, macroevolutionary and historical biogeography research hypotheses cannot be properly tested without the incorporation of a phylogenetic framework (e.g. Graham and Fine 2008 , Hardy 2008, Kissling 2017, Vandelook et al. 2012, Vandelook et al. 2018, Couvreur et al. 2011, Janssens et al. 2009,Janssens et al. 2016). Likewise, phylogenetic diversity is considered an important element in conservation biology and related biodiversity assessment studies (Chave et al. 2007). Even though the importance of phylogenetics in ecology and evolution is recognised, it remains somehow strenuous to combine ecological research with evolutionary biology and integrate it in a phylogenetic scenario. This discrepancy is sometimes caused by a lack of awareness and knowledge about the other disciplines, whereby researchers could be reluctant to reach out to such expertise and combine their results into new disciplines. Additionally, differences in methodologies and techniques applied by ecologists and evolutionary biologists can sometimes cause a certain hesitation to go for a complementary approach with blending disciplines. In addition, there is a nearly continuous development of new insights and techniques in the fields of ecology and evolution (e.g. Bouckaert et al. 2019, Revell et al. 2008, Revell 2012 , Suchard et al. 2018), making it rather challenging to keep up to date with the latest novelties. Furthermore, not all organisms investigated from an ecological perspective are present in molecular databases, which make it difficult to construct a perfectly matching phylogenetic hypothesis for further analysis. For scientists who focus on resolving specific evolutionary or ecological queries, building a phylogenetic framework from novel gene sequence data is often a heavy burden as it takes a lot of time, money and effort, even apart from the specific expertise needed. The construction of a purpose-built phylogeny can be considered as rather costly and labour-intensive and requires more elaborate expertise on novel techniques than when sequences are merely mined from GenBank in order to make a tree, based on already existing sequences. Whereas the former strategy allows the user to make a tailor-made phylogeny that can be used for further ecological or evolutionary purposes, the latter is less proficient, as one can only use the sequences that are available in genetic databases. Nevertheless, in the case of large-scale meta-analyses, it becomes almost impossible to obtain sequence data from all species investigated. When there is a need to examine evolutionary and ecological trends in an historical context, a large-scale phylogenetic hypothesis, that is optimised in a spatiotemporal context, provides an optimal solution.

There is currently an ongoing quest to optimise the methodology for constructing large-scale mega-phylogenies that can be used for further ecological and evolutionary studies. This is done by either mining and analysing publicly available DNA sequences (Zanne et al. 2014), amalgamating published phylograms (Hinchliff et al. 2015) or the combination of both (Smith and Brown 2018). For example, Zanne et al. (2014) constructed their own large supermatrix-based phylogeny that was used to gain more insights into the evolution of cold-tolerant angiosperm lineages. However, the study of Qian and Jin (2016) showed that the phylogeny of Zanne et al. (2014) contained several taxonomic errors. The approaches of Smith and Brown (2018) and Hinchliff et al. (2015) also do not always provide the most optimal phylogenetic framework for further analyses as both studies use a (partially) synthetic approach, based on already published phylograms that can putatively contain inconsistencies in their estimated node ages. The main goal of the present study is, therefore, to provide a large-scale dated phylogeny - encompassing nearly 1/8th of all angiosperms - that can be used for further ecological and evolutionary analyses. In order to construct this angiosperm phylogeny, a comprehensive approach was applied in which sequence data were both mined and generated, subsequently aligned, phylogenetically analysed and dated using over 50 fossil calibration points. With the applied methodology, we aimed to create sufficient overlap in molecular markers without having too much missing sequence data in the datamatrix. In addition, phylogenetic analyses, as well as the age estimation assessment, were performed as a single analysis on the whole datamatrix in order to create a dated angiosperm mega-phylogeny that is characterised by a low degree of synthesis.

Material and methods

Marker choice

In 2009, the Consortium for the Barcode of Life working group (CBOL) advised sequencing of the two plastid markers *matK* (incl. *trnK*) and *rbcL* for identifying plant species, resulting in a massive amount of data available on GenBank. *rbcL* is a conservative locus with low level of variation across flowering plants and therefore useful for reconstructing higher level divergence. In contrast, *matK* contains rapidly evolving regions that are useful for studying interspecific divergence (Hilu et al. 2003, Kress et al. 2005). Thus, the combination of *matK* (incl. *trnK*) and *rbcL* has the advantage of combining different evolutionary rates, making it possible to infer relationships at different taxonomic levels. In addition, we sampled only *matK* (incl. *trnK*) and *rbcL* markers in order to reduce missing data to a minimum, as this impacts the phylogenetic inference between species. These supermatrix approaches - which generally contain a substantial amount of missing data – can suffer from imbalance in presence/absence for each taxon per locus, resulting in low resolution and support or even wrongly inferred relationships (Sanderson and Shaffer 2002, Roure et al. 2013).

Taxon sampling

We extracted angiosperm sequence data of *rbcL* and *matK* (incl. *trnK*) from GenBank (15 February 2015) using the 'NCBI Nucleotide extraction' tool in Geneious v11.0 (Auckland, New Zealand). Five gymnosperm genera were chosen as outgroup (Suppl. material 1). This large dataset was supplemented with 468 specimens of African tree species obtained via multiple barcoding projects (available at the Barcode of Life Data Systems (BOLD)), as well as via additional lab work (see paragraph on molecular protocols below). In total, 820 newly obtained sequences are submitted to GenBank (Suppl. material 1).

Molecular protocols

A modified CTAB protocol was used for total genomic DNA isolation (Tel-Zur et al. 1999). Secondary metabolites were removed by washing ground leaf material with extraction buffer (100 mM Tris pH 8, 5mM EDTA pH 8, 0.35 M sorbitol). After the addition of 575 µl CTAB lysis buffer with addition of 3% PVP-40, the samples were incubated for 1.5 hours (60°C). Chloroform-isoamylalcohol (24/1 v/v) extraction was done twice, followed by an ethanol-salt precipitation (absolute ethanol, sodium acetate 3 M). After centrifugation, the pellet was washed twice (70% ethanol), air-dried and dissolved in 100 µl TE buffer (10 mM Tris pH 8, 1 mM EDTA pH 8).

Amplification reactions of *matK* (incl. *trnK*) and *rbcL* were carried out with a 25 µl reaction mix containing 1 µl DNA, 2 x 1 µl oligonucleotide primer (100 ng/µl), 2.5 µl of 10 mM dNTPs, 2.5 µl Taq Buffer, 0.2 µl KAPA Taq DNA polymerase and 16.8 µl MilliQ water. Reactions commenced with a 3 minute heating at 95°C, followed by 30 cycles consisting of 95°C denaturation for 30 s, primer annealing for 60 s and extension at 72°C for 60 s. Reactions ended with a 3 minute incubation at 72°C. Annealing temperatures for *matK* (incl. *trnK*) and *rbcL* were set at 50°C and 55°C, respectively. Primers designed by Kim J. (unpublished) were used to sequence *matK* (incl. *trnK*), whereas *rbcL* primers were adopted from Fay et al. (1997) and Little and Barrington (2003). PCR products were cleaned using an ExoSap purification protocol. Purified amplification products were sequenced by the Macrogen sequencing facilities (Macrogen, Seoul, South Korea). Raw sequences were assembled using Geneious v11.0 (Biomatters, New Zealand).

Sequence alignment and phylogenetic analyses

We are aware that the publicly available database, GenBank, contains a large amount of erroneous data (Ashelford et al. 2005, Yao et al. 2004, Shen et al. 2013). Retrieving the sequence data was, therefore, subjected to a quality control procedure. All downloaded sequences were blasted (Megablast option) against the GenBank database, thereby discarding all sequences with anomalies against their original identification. Minimum similarity in BLAST was set at 0.0005, whereas word size (W) was reduced to 8 for greater sensitivity of the local pairwise alignment and the maximum hits was set at 250. A

single sequence of each fragment was retained for each taxon name or non-canonical NCBI taxon identifier given in GenBank. In the case where multiple accessions per species were available on GenBank, we chose the accession with the highest sequence length, the best quality and the highest sequence similarity compared to the other accessions of the same species in the GenBank database. Additionally, sequences with multiple ambiguities were discarded, as well as sequences with similar taxon names, but different nucleotide sequences. In addition, sequences with erroneous taxonomic names (checked in R using the "Taxize" and "Taxonstand" packages (R Development Core Team 2009, Cayuela et al. 2012, Chamberlain et al. 2016)) were removed from further analyses. Importantly, Taxize uses the Taxonomic Name Resolution Service (TNRS; Boyle et al. 2013) function to match taxonomic names, whereas Taxonstand is linked with 'The Plant List' database. As such, we also checked the validity of the taxonomic names in our dataset using both databases. Only those taxa which had names that were considered valid for both databases were kept for further analyses.

For sequence fragments that are protein-encoded, comparison of amino acid (AA) sequences, based on the associated triplet codons between taxa, was applied. As a result, taxa with a sudden shift in AA or frame shift were discarded from the dataset.

Alignment was carried out in multiple stages. Due to our large angiosperm-wide dataset, an initial alignment (automatically and manually) was conducted for each order included in the dataset. Subsequently, the different alignments were combined using the Profile alignment algorithm (Geneious v11.0, Auckland, New Zealand). The initial automatic alignment was conducted with MAFFT (Katoh et al. 2002) using an E-INS-i algorithm, a 100PAM/k = 2 scoring matrix, a gap open penalty of 1.3, and an offset value of 0.123. Manual fine-tuning of the aligned dataset was performed in Geneious v11.0 (Auckland, New Zealand). During the manual alignment of the different datasets, we carefully assessed the homology of every nucleotide at each position in the alignment (Phillips et al. 2009). The large amount of angiosperm taxa included in the analyses often provided a good view on the evolution of the nucleotides at certain positions, in which some taxa functioned as transition lineages between differing nucleotides and their exact position in the alignment. The importance of a well-designed homology assessment for a complex sequence dataset has been proven successful here for the phylogenetic inference of the angiosperms.

The best-fit nucleotide substitution model for both *rbcL* and *matK* (incl. *trnK*) was selected using jModelTest 2.1.4. (Posada 2008) out of 88 possible models under the Akaike Information Criterion (AIC). The GTR+G model was determined as the best substitution model for each locus and, as such, both markers were jointly analysed under this model. Maximum Likelihood (ML) tree inference was conducted using the Randomized Axelerated Maximum Likelihood (RAxML) software version 7.4.2 (Stamatakis 2006) under the general time-reversible (GTR) substitution model with gamma rate heterogeneity and lewis correction. Although the phylogeny, based on the plastid dataset, generated relationships that corresponded well with currently known angiosperm phylogenies (e.g. Wikström et al. 2001, Soltis et al. 2002, Moore et al. 2007, Magallón and Castillo 2009, Magallón 2014, Magallón et al. 2015, Bell et al. 2005, Bell et

al. 2010), we decided to use a constraint (Suppl. material 2) in order to make sure that possible unrecognised mismatches for certain puzzling lineages were significantly reduced. The constraint tree follows the phylogenetic framework of APG4 (Angiosperm Phylogeny Group 2016) at order level. At the lower phylogenetic level, families were only constrained as polytomy in their specific angiosperm order. Genera and species were not constrained.

Support values for the large angiosperm dataset were obtained via the rapid bootstrapping algorithm as implemented in RAxML 7.4.2 (Stamatakis 2006), examining 1000 pseudo-replicates under the same parameters as for the heuristic ML analyses. Bootstrap values were visualised using the Consensus Tree Builder algorithm as implemented in Geneious v11.0.

Divergence time analysis

Evaluation of fossil calibration points was carried out following the specimen-based approach for assessing paleontological data by Parham et al. (2012). As such, 56 angiosperm fossils were used as calibration points in our molecular dating analysis. Detailed information about the fossils, including (1) citation of museum specimens, (2) locality and stratigraphy of fossils, (3) referenced stratigraphic age and (4) crown/stem node position is provided in Table 1. Fossils are placed at both early and recently diversified lineages within the angiosperms. Due to the large size of the dataset, we applied the penalised likelihood algorithm as implemented in treePL (Smith and O'Meara 2012), which utilises hard minimum and maximum age constraints. In order to estimate these hard minimum and maximum age constraints, we calculated the log normal distribution of each fossil calibration point using BEAUti v.1.10 (Suchard et al. 2018). Maximum age constraints for each fossil correspond to the 95.0% upper boundary of the computed log normal distribution, in which the offset equals the age of the fossil calibration point, the mean is set at 1.0 and the standard deviation at 1.0. This methodology resulted in a minimum 15 million year broad interval for each angiosperm calibration point (Table 1). Due to recently published studies in which both old and young age estimates were retrieved for the crown node of the angiosperms (e.g. Bell et al. 2005, Bell et al. 2010, Magallón et al. 2015, Magallón 2014, Magallón and Castillo 2009, Moore et al. 2007, Smith et al. 2010, Wikström et al. 2001, Soltis et al. 2002), we opted to set the hard maximum and minimum calibration of the angiosperms at 220 and 180 million years, respectively. As for the overall calibration, we followed the strategy of Smith et al. (2010), in which all fossils were considered as a minimum-age constraint. Smith et al. (2010) applied this approach since earlier studies on angiosperm evolution had treated tricolpate fossil pollen as maximum-age constraint, thereby maybe artificially pushing the root age of the angiosperms towards more recent times (e.g. Soltis et al. 2002, Magallón et al. 2015, Magallón 2014, Magallón and Castillo 2009, Moore et al. 2007, Bell et al. 2010, Bell et al. 2005).

The molecular clock hypothesis was tested using a chi^2 likelihood ratio test (Felsenstein 1988) and demonstrated that the substitution rates in the combined dataset are not clock-

like (P < 0.001 for all markers). The most optimal maximum likelihood tree obtained via RAxML was used as input for the penalised likelihood dating analysis in treePL (Smith and O'Meara 2012). Due to the large size dataset, treePL was preferred over other age estimation software packages such as BEAST 1.10 (Suchard et al. 2018), BEAST 2.5 (Bouckaert et al. 2019) or MrBayes 3.2 (Ronquist et al. 2012). The best-fit smoothing parameter of 0.0033 was specified empirically using an adaptation of the cross-validation test as implemented in treePL (Sanderson 2003, Smith and O'Meara 2012). An adapted methodology was set up as the original tree of over 35,000 taxa was too large for correctly calculating the best-fit smoothing parameter. In order to accurately carry out the cross-validation test, 500 replicates were made of the original dataset in which 90% of the original species were randomly pruned. Each of the replicates was then subjected to a cross-validation test under the following parameters: cvstart = 10; cvstop = 0.0001; cvmultstep = 0.9; randomcv. The best-fit smoothing parameter was selected as the variable with the highest proportion (0.0033; 12%), with the second best-fit smoothing parameter being situated at 0.0036 (11%). Smoothing parameters calculated per replicate followed a normal distribution with its optimum around 0.0033 and 0.0036 (Suppl. material 3). This strategy of calculating the smoothing parameter of very large datasets seemed effective and robust for estimating node ages of our angiosperm phylogeny using treePL. Furthermore, since there is a large amount of rate heterogeneity amongst angiosperm lineages that could likely infringe the treePL model, it is considered that a low smoothing parameter will provide a more robust analysis. So, by applying a lower penalty, potential issues that could be caused by strongly contrasting evolutionary rates within distant angiosperm clades will putatively be avoided (Stephen Smith, pers. comm.). In order to generate 95% confidence intervals for the dated nodes, we generated 1,000 bootstrap pseudo-replicates using the ML topology of the earlier heuristic analysis as constraint. Each ML bootstrap tree was then individually dated using treePL under the same parameters as for the single age estimation analysis, described above. Subsequently, the 1,000 dated bootstrap trees were imported into TreeAnnotator v1.10 in order to calculate and visualise the 95% confidence intervals for each node (Suchard et al. 2018).

Results and Discussion

The final aligned data matrix consists of 36,101 angiosperm species. *matK* (incl. *trnK*) sequences were mined for 31,391 species (87%), whereas *rbcL* sequences were obtained for 26,811 (74%) species (Suppl. material 1). The sequence dataset has an aligned length of 4,968 basepairs (bp) of which 4,285 (86%) belong to *matK* (incl. *trnK*) and 683 (14%) to *rbcL*. Within *rbcL,* all characters were variable (100%), whereas for *matK* (incl. *trnK*) 3,921 characters (91.5%) were variable. Support value analyses indicate that approximately 26% of the branches have a bootstrap value > 75 (Suppl. material 4 Suppl. material 3). Based on the different studies that estimated the total number of flowering plants currently described (between 260,000 and 450,000 species) (Crane et al. 1995, Christenhusz and Byng 2016, Cronquist 1981, Lupia et al. 1999, Pimm and Joppa 2015, Prance et al. 2000, Thorne 2002), the presented phylogeny represents

between 14% and 8% of the known flowering plants, respectively. In addition, the phylogenetic tree contains 54.6% (8,399) of all currently accepted angiosperm genera and 94.5% (426) of all families of flowering plants are included, as well as all currently known angiosperm orders. As such, the current angiosperm tree can be regarded as the largest dated angiosperm phylogenetic framework that is generated by combining genuine sequence data and fossil calibration points and will be useful for large-scale ecological and biogeographical studies. Compared to the species-level-based tree of Zanne et al. (2014) and its updated version by Qian and Jin (2016), the current phylogeny is larger in size, containing more species (+4,797 species) and genera (+468). However, the phylogeny of Zanne et al. (2014) included more families and an equal number of orders. Additionally, Zanne et al. (2014)'s updated phylogeny (Qian and Jin 2016) also included 1,190 taxa of bryophytes, pteridophytes and gymnosperms, whereas the current phylogeny only contains 5 outgroup gymnosperm species. As a result, when comparing the differences in species number between both angiosperm mega-phylogenies, the current tree contains nearly 20% more flowering plant lineages (+5,987 species).

Age estimation of the large-scale angiosperm tree resulted in a dated phylogeny (Fig. 1; Suppl. material 5) that largely corresponds to the different recent angiosperm-wide dating analyses (e.g.Bell et al. 2010, Magallón et al. 2015, Smith et al. 2010, Wikström et al. 2001, Zanne et al. 2014). Even though small dissimilarities are present concerning the age of the most early diversified angiosperm lineages (see Table 1), the overall age of the different families corresponds rather well to what is known from these other studies. Differences in stem node age of large clades such as superasterids, superrosids, eudicots, monocots or magnoliids are probably due to the use of a slightly different and larger set of fossil calibration points, as well as not using tricolpate fossil pollen as maximum-age for eudicots. Compared to the angiosperm phylogeny of Zanne et al. (2014), where time-scaling was carried out with 39 fossil calibrations, the current tree contains 56 fossils in total. Although some fossils are the same between both Zanne's study and ours (e.g. *Pseudosalix handleyi, Fraxinus wilcoxiana, Spirematospermum chandlerae*), several fossils that have been used to optimise the age estimation of the current megaphylogeny are carefully chosen from other dating analyses (Bell et al. 2010, Magallón et al. 2015, Smith et al. 2010).

Recently, Qian and Jin (2016) developed a novel tool (S.PhyloMaker package as implemented in the R environment) to generate artificially enriched species trees, based on an updated version of the original angiosperm mega-phylogeny of Zanne et al. (2014) . According to the study of Qian and Jin (2016), the software package produces phylogenies for every species that one needs to assess in a community ecological environment. S.PhyloMaker grafts species of interest, either as a basal polytomy (regular or Phylomatic/BLADJ approach; Webb et al. 2008), or randomly branched within the existing parental clades that are found in the mega-phylogeny. Likewise, branch lengths or time-calibrated node splits of newly added taxa are also artificially estimated according to their relative position in the original mega-phylogeny. Even though the software package of Qian and Jin (2016) provides a good alternative for the lack of decent

sampling of angiosperm taxa in mega-phylogenies for some ecological studies, not all ecological or evolutionary disciplines that are in need of a phylogenetic framework can rely on this methodology, as it is not based on the inclusion of original sequence data. Additionally, the current, more densely sampled phylogenetic framework could be used in the S.Phylomaker system in order to reduce the variance that is related to the random addition of new lineages, as the placement of new taxa can be more precisely carried out due to the presence of more nodes with known heights. The use of only chloroplast data for the construction of this large-scale angiosperm mega-phylogeny has, indeed, some disadvantages as chloroplasts constitute a single, linked locus that is mainly maternally inherited within angiosperms and processes such as hybridisation and subsequent introgression, as well as reticulate evolution and incomplete lineage sorting, are difficult to detect with only data from one genome (Soltis and Soltis 2009, Lee et al. 2011). This, in combination with the fact that only two gene markers were used for phylogeny reconstruction, results in making this phylogeny to be regarded as an angiosperm gene tree rather than a species tree. Despite these putative issues, the large-scale phylogenetic hypothesis, that has been constructed here, has proven to be useful for resolving large-scale evolutionary questions at angiosperm level (e.g. Dagallier et al. in press). To date, it remains a continuous challenge to increase the size of large-scale angiosperm phylogenies with new species and gene markers to create a reliable platform, in which ecological and evolutionary research can be combined with phylogenetics. The current phylogeny is a further step towards an all-encompassing angiosperm phylogeny that can be used to resolve large-scale ecological and evolutionary queries.

Acknowledgements

This study is part of the HERBAXYLAREDD project (BR/143/A3/HERBAXYLAREDD), funded by the Belgian Belspo-BRAIN program axis 4. This project is supported by Plant.ID, which has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement N° 765000. This study is also supported by the BRAIN.be BELSPO research program AFRIFORD and by the French Foundation for Research on Biodiversity (FRB) and the Provence-Alpes-Côte d'Azur region (PACA) region through the Centre for Synthesis and Analysis of Biodiversity data (CESAB) programme, as part of the RAINBIO research project (http://rainbio.cesab.org). The authors thank Kenneth Oberlander and an anonymous reviewer for improving the manuscript.

Conflicts of interest

References

- Angiosperm Phylogeny Group (2016) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Botanical Journal of the Linnean Society 181: 1-20. https://doi.org/10.1111/boj.12385
- Ashelford KE, Chuzhanova NA, Fry JC, Jones AJ, Weightman AJ (2005) At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. Applied and Environmental Microbioogly 71: 7724-7736. https://doi.org/10.1128/AEM.71.12.7724-7736.2005
- Basinger JF, Christophel DC (1985) Fossil flowers and leaves of the Ebenaceae from the Eocene of southern Australia. Canadian Journal of Botany 63: 1825-1843. https://doi.org/10.1139/b85-258
- Basinger JF, Greenwood DR, Wilson PG, Chistophel DC (2007) Fossil flowers and fruits of capsular Myrtaceae from the Eocene of South Australia. Canadian Journal of Botany 85: 204-215. https://doi.org/10.1139/B07-001
- Batten DJ (1981) Stratigraphy, palaeogeography and evolutionary significance of Late Cretaceous and Early Tertiary Normapolles pollen. Review of Palaeobotany and Palynology 35: 125-137. https://doi.org/10.1016/0034-6667(81)90104-4
- Bell CD, Soltis DE, Soltis PS (2005) The age of the angiosperms: A molecular timescale without a clock. Evolution 59: 1245-1258. https://doi.org/10.1554/05-005
- Bell CD, Soltis DE, Soltis PS (2010) The age and diversification of the angiosperms re-visited. American Journal of Botany 97: 1296-1303. https://doi.org/10.3732/ajb.0900346
- Boucher LD, Manchester S, Judd W (2003) An extinct genus of Salicaceae based on twigs with attached flowers, fruits, and foliage from the Eocene Green River Formation of Utah and Colorado, USA. American Journal of Botany 90: 1389-1399. https://doi.org/10.3732/ajb.90.9.1389
- Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, De Maio N, Matschiner M, Mendes FK, Müller NF, Ogilvie H, Plessis L, Popinga A, Rambaut A, Rasmussen D, Siveroni I, Suchard MA, Wu CH, Xie D, Zhang C, Stadler T, Drummond AJ (2019) BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS Computational Biology 15: 1006650. https://doi.org/10.1371/journal.pcbi.1006650
- Boyle B, Hopkins N, Lu Z, Raygoza Garay J, Mozzherin D, Rees T, Matasci N, Narro ML, Piel WH, McKay SJ, Lowry S, Freeland C, Peet R, Enquist BJ (2013) The taxonomic name resolution service: an online tool for automated standardization of plant names. BMC bioinformatics 14: 16. https://doi.org/10.1186/1471-2105-14-16
- Call VB, Dilcher DL (1992) Investigations of angiosperms from the Eocene of southeastern North America: Samaras of *Fraxinus wilcoxiana* Berry. Review of Palaeobotany and Palynology 74: 249-266. https://doi.org/10.1016/0034-6667(92)90010-E

- Cayuela L, Granzow-de la Cerda I, Albuquerque FS, Golicher DJ (2012) taxonstand: An r package for species names standardisation in vegetation databases. Methods in Ecology and Evolution 3: 1078-1083. https://doi.org/10.1111/j.2041-210X.2012.00232.x
- Cevallos-Ferriz SR, Estrada-Ruiz E, Perez-Hernandez BR (2008) Phytolaccaceae infructescence from Cerro del Pueblo Formation, Upper Cretaceous (Late Campanian). American Journal of Botany 95: 77-83. https://doi.org/10.3732/ajb.95.1.77
- Chamberlain S, Szocs E, Boettiger C, Ram K, Bartomeus I, Baumgartner J, Foster Z, O'Donnell J (2016) Taxize: taxonomic information from around the web. Version 0.7.8. URL: https://github.com/ropensci/taxize.
- Chave J, Chust G, Thébaud C (2007) The importance of phylogenetic structure in biodiversity studies. In: Storch D, Marquet P, Braun J (Eds) Scaling Biodiversity. Cambridge University Press, Cambridge, 16 pp.
- Chen I, Manchester S (2007) Seed morphology of modern and fossil *Ampelocissus* (Vitaceae) and implications for phytogeography. American Journal of Botany 94: 1534-1553. https://doi.org/10.3732/ajb.94.9.1534
- Christenhusz MJ, Byng JW (2016) The number of known plants species in the world and its annual increase. Phytotaxa 261: 201-217. https://doi.org/10.11646/phytotaxa.261.3.1
- Christopher RA (1979) Normapolles and triporate pollen assemblages from the Raritan and Magothy formations (upper Cretaceous) of New Jersey. Palynology 3: 73-122. https://doi.org/10.1080/01916122.1979.9989185
- Couvreur TL, Pirie MD, Chatrou LW, Saunders RM, Su YC, Richardson JE, Erkens RH (2011) Early evolutionary history of the flowering plant family Annonaceae: Steady diversification and boreotropical geodispersal. Journal of Biogeography 38: 664-680. https://doi.org/10.1111/j.1365-2699.2010.02434.x
- Crane PR, Manchester SR, Dilcher DL (1990) A preliminary survey of fossil leaves and well-preserved reproductive structures from the Sentinel Butte Formation (Paleocene) near Almont, North Dakota. Fieldiana Geology 20: 1-63.
- Crane PR, Pedersen KR, Friis EM, Drinnan AN (1993) Early Cretaceous (early to middle Albian) platanoid inflorescences associated with *Sapindopsis* leaves from the Potomac Group of North America. Systematic Botany 18: 328-344. https://doi.org/10.2307/2419407
- Crane PR, Friis EM, Pedersen KR (1994) Paleobotanical evidence on the early radiation of magnoliid angiosperms. Plant Systematics and Evolution 8: 51-72.
- Crane PR, Friis EM, Pedersen KR (1995) The origin and early diversification of angiosperms. Nature 374: 27-33. https://doi.org/10.1038/374027a0
- Crepet W, Nixon K (1996) The fossil history of stamens. In: D'Arcy W, Keating R (Eds) The anther: form, function and phylogeny. Cambridge University Press, Cambridge, UK, 25-27 pp.
- Crepet WL, Nixon KC (1998) Fossil Clusiaceae from the Late Cretaceous (Turonian) of New Jersey and implications regarding the history of bee pollination. American Journal of Botany 85: 1122-1133. https://doi.org/10.2307/2446345
- Cronquist A (1981) The evolution and classification of flowering plants. Columbia University Press, New York.
- Dagallier L, Janssens S, Dauby G, Blach-Overgaard A, Mackinder B, Droissart V, Svenning J, Sosef M, Stévart T, Harris D, Sonké B, Wieringa J, Hardy O, Couvreur T (in press) Cradles and museums of generic plant diversity across tropical Africa. Journal of Biogeography https://doi.org/10.1111/nph.16293

- Daghlian CP (1981) A review of the fossil record of monocotyledons. Botanical Review 47: 517-555. https://doi.org/10.1007/BF02860540
- Dilcher D, Crane P (1984) Archaeanthus: an early angiosperm from the Cenomanian of the Western Interior of North America. Annals of the Missouri Botanical Garden 71: 351-383. https://doi.org/10.2307/2399030
- Doyle JA, Hotton CL, Ward JV (1990) Early Cretaceous tetrads, zonasulculate pollen, and Winteraceae. 1. Taxonomy, morphology, and ultrastructure. American Journal of Botany 77: 1544-1557. https://doi.org/10.1002/j.1537-2197.1990.tb11395.x
- Doyle JA, Manchester S, Souquet H (2008) A seed related to Myristicaceae in the Early Eocene of South England. Systematic Botany 33: 636-646. https://doi.org/10.1600/036364408786500217
- Edelman D (1975) The Eocene Germer Basin flora of south-central Idaho. MSc thesis, University of Idaho, Moscow, ID, USA.
- Estrada-Ruiz E, Calvillo-Canadell L, Cevallos-Ferriz SR (2009) Upper Cretaceous aquatic plants from Northern Mexico. Aquatic Botany 90: 283-288.
- Fay MF, Swensen SM, Chase MW (1997) Taxonomic affinities of *Medusagyne oppositifolia* (Medusagynaceae). Kew Bulletin 52: 111-120. https://doi.org/10.2307/4117844
- Felsenstein J (1988) Phylogenies and quantitative characters. Annual Review of Ecology and Systematics 19: 445-471. https://doi.org/10.1146/annurev.es.19.110188.002305
- Feng X, Tang B, Kodrul TM, Jin J (2013) Winged fruits and associated leaves of *Shorea* (Dipterocarpaceae) from the Late Eocene of South China and their phytogeographic and paleoclimatic implications. American Journal of Botany 100: 574-581. https://doi.org/10.3732/ajb.1200397
- Friis E, Pedersen KR, Crane P (2001) Fossil evidence of water lilies (Nymphaeales) in the Early Cretaceous. Nature 410: 357-360. https://doi.org/10.1038/35066557
- Friis E, Pedersen K, Crane P (2004) Araceae from the Early Cretaceous of Portugal: evidence on the emergence of monocotyledons. Proceedings of the National Academy of Sciences of the USA 101: 16565-16570. https://doi.org/10.1073/pnas.0407174101
- Friis EM (1988) *Spirematospermum chandlerae* sp. nov., an extinct species of Zingiberaceae from the North American Cretaceous. Tertiary Research 9: 7-12.
- Friis EM, Pedersen KR, Schönenberger J (2003) Endressianthus, a new Normapolles-producing plant genus of Fagalean affinity from the Late Cretaceous of Portugal. International Journal of Plant Sciences 164: 201-223. https://doi.org/10.1086/376875
- Graham A (1977) New records of *Pelliciera* (Theaceae/Pelliceriaceae) in the Tertiary of the Caribbean. Biotropica 9: 48-52. https://doi.org/10.2307/2387858
- Graham C, Fine P (2008) Phylogenetic beta diversity: linking ecological and evolutionary processes across space and time. Ecology Letters 11: 1265-1277. https://doi.org/10.1111/j.1461-0248.2008.01256.x
- Hardy OJ (2008) Testing the spatial phylogenetic structure of local communities: statistical performances of different null models and test statistics on a locally neutral community. Journal of Ecology 96: 914-926. https://doi.org/10.1111/j.1365-2745.2008.01421.x
- Herendeen P, Crepet W, Dilcher D (1992) The fossil history of the Leguminosae: phylogenetic and biogeographic implications. In: Herendeen P, Dilcher D (Eds) Advances in Legume systematics, Part 4: The Fossil Record. Royal Botanic Gardens, Kew, UK, 303–316. pp.

- Hermsen E, Gandolfo MA, Nixon KC, Crepet WL (2003) *Divisestylus* gen. nov. (aff. *Iteaceae*), a fossil saxifrage from the Late Cretaceous of. American Journal of Botany 90: 1373-1388. https://doi.org/10.3732/ajb.90.9.1373
- Hickey LJ, Peterson AK (1978) Zingiberopsis, a fossil genus of the ginger family from the Late Cretaceous to Early Eocene sediments of western interior North America. Canadian Journal of Botany 56: 1136-1152. https://doi.org/10.1139/b78-128
- Hilu KW, Borsch T, Müller K, Soltis DE, Soltis PS, Savolainen V, Chase MW, Powell MP, Alice LA, Evans R, Sauquet H, Neinhuis C, Slotta TA, Rohwer JG, Campbell CS, Chatrou LW (2003) Angiosperm phylogeny based on matK sequence information. American Journal of Botany 90: 1758-1776. https://doi.org/10.3732/ajb.90.12.1758
- Hinchliff CE, Smith SA, Allman JF, Burleigh JG, Chaudhary R, Coghill LM, Crandall KA, Deng J, Drew BT, Gazis R, Gude K, Hibbett DS, Katz LA, Laughinghouse HD, McTavish EJ, Midford PE, Owen CL, Ree RH, Rees JA, Soltis DE, Williams T, Cranston KA (2015) Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proceedings of the National Academy of Sciences, USA 112: 12764-12769. https://doi.org/10.1073/pnas.1423041112
- Janssens SB, Knox EB, Huysmans S, Smets EF, Merckx VS (2009) Rapid radiation of Impatiens: Result of a global climate change. Molecular Phylogenetics and Evolution 52: 806-824. https://doi.org/10.1016/j.ympev.2009.04.013
- Janssens SB, Vandelook F, De Langhe E, Verstraete B, Smets E, Vandenhouwe I, Swennen R (2016) Evolutionary dynamics and biogeography of Musaceae reveal a correlation between the diversification of the banana family and the geological and climatic history of Southeast Asia. New Phytologist 210: 1453-65. https://doi.org/10.1111/nph.13856
- Jardiné S, Magloire H (1965) Palynologie et stratigraphie du Crétacé des Bassin du Sénégal et Cote d'Ivoire. Mémoires du Bureau de Recherches Géologiques et Miniéres 32: 187-245.
- Jarzen DM (1978) Some Maastrichtian palynomorphs and their phytogeographical and paleoecological implications. Palynology 2: 29-38. https://doi.org/10.1080/01916122.1978.9989163
- Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on Fourier transform. Nucleic Acids Research 30: 3059-3066. https://doi.org/10.1093/nar/gkf436
- Kedves M (1989) Evolution of the Normapolles complex. In: Crane P, Blackmore S (Eds) Evolution, Systematics, and Fossil History of the Hamamelidae, 1-7. Systematics Association Special Volume, 40b. Clarendon Press, Oxford, 1-7 pp.
- Kissling WD (2017) Has frugivory influenced the macroecology and diversification of a tropical keystone plant family? Research Ideas and Outcomes 3: e14944. https://doi.org/10.3897/rio.3.e14944
- Knobloch E, Mai DH (1986) Monographie der Fruchte and Samen in der Kreide von Mitteleuropa. Rozpravy Ustredniho Ustavu Geologickeho Praha 47: 1-219.
- Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. Proceedings of the National Academy of Sciences USA 102: 8369-8374. https://doi.org/10.1073/pnas.0503123102
- Lee E, Cibrian-Jaramillo A, Kolokotronis S, Katari M, Stamatakis A, Ott M, Chiu J, Little D, Stevenson D, McCombie WR, Martienssen R, Coruzzi G, DeSalle R (2011) A

Functional Phylogenomic View of the Seed Plants. PlosONE https://doi.org/10.1371/journal.pgen.1002411
- Little DP, Barrington DS (2003) Major evolutionary events in the origin and diversification of the fern genus Polystichum (Dryopteridaceae). American Journal of Botany 90: 508-514. https://doi.org/10.3732/ajb.90.3.508
- Li Y, Smith T, Liu CJ, Awasthi N, Yang J, Wang YF, Li CS (2011) Endocarps of *Prunus* (Rosaceae: Prunoideae) from the early Eocene of Wutu, Shandong Province, China. Taxon 60: 555-564. https://doi.org/10.1002/tax.602021
- Lupia R, Lidgard S, Crane PR (1999) Comparing palynological abundance and diversity: Implications for biotic replacement during the Cretaceous angiosperm radiation. Paleobiology 25: 305-340. https://doi.org/10.1017/S009483730002131X
- Magallón S, Castillo A (2009) Angiosperm diversification through time. American Journal of Botany 96: 349-365. https://doi.org/10.3732/ajb.0800060
- Magallón S (2014) A review of the effect of relaxed clock method, long branches, genes, and calibrations in the estimation of angiosperm age. Botanical Sciences 92: 1-22. https://doi.org/10.17129/botsci.37
- Magallón S, Gomez-Acevedo S, Sanches-Reyes LL (2015) A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. New Phytologist 207: 437-453. https://doi.org/10.1111/nph.13264
- Mai DH (1985) Entwicklung der Wasser- und Sumpfpflanzen-Gesellschaften Europas von der Kreide bis ins Quartär. Flora 176: 449-511. https://doi.org/10.1016/S0367-2530(17)30141-X
- Mai DH (1987) Neue Fruchte und Samen aus Palaozanen Ablagerungen Mitteleuropas. Feddes Repertorium 98: 197-229.
- Manchester S, Kress J (1993) Fossil bananas (Musaceae): *Ensete oregonense* sp. nov. from the Eocene of western North America and its phytogeographic significance. American Journal of Botany 80: 1264-1272. https://doi.org/10.1002/j.1537-2197.1993.tb15363.x
- Manchester S (1999) Biogeographical relationships of North American Tertiary floras. Annals of the Missouri Botanical Garden 86: 472-522. https://doi.org/10.2307/2666183
- Manchester S, Kapgate D, Wen J (2013) Oldest fruits of the grape family (Vitaceae) from the Late Cretaceous Deccan Cherts of India. American Journal of Botany 100: 1849-1859. https://doi.org/10.3732/ajb.1300008
- Martinez-Millan M, Crepet WL, Nixon KC (2009) *Pentapetalum trifasciculandricus* gen. et sp. nov., a thealean fossil flower from the Raritan Formation, New Jersey,USA (Turonian, Late Cretaceous). American Journal of Botany 96: 933-949. https://doi.org/10.3732/ajb.0800347
- Mohr BA, Bernardes-De-Oliveira ME (2004) *Endressinia brasiliana*, a magnolialean angiosperm from the Lower Cretaceous Crato Formation (Brazil). International Journal of Plant Sciences 165: 1121-1133. https://doi.org/10.1086/423879
- Moore M, Bell C, Soltis P, Soltis D (2007) Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proceedings of the National Academy of Sciences USA. 104: 19363-19368. https://doi.org/10.1073/pnas.0708072104
- Muller J (1981) Fossil pollen records of extant angiosperms. The Botanical Review 47: 1-142. https://doi.org/10.1007/BF02860537

- Nichols DJ, Ott HL (1978) Biostratigraphy and evolution of the Momipites-Caryapollenites lineage in the Early Terti-ary in the Wind River Basin, Wyoming. Palynology 2: 93-112. https://doi.org/10.1080/01916122.1978.9989167
- Pacltova B (1966) Pollen grains of angiosperms in the Cenomanian Peruc Formation in Bohemia. Palaeobotanist 15: 52-54.
- Pan AD (2010) Rutaceae leaf fossils from the Late Oligocene (27.23 Ma) Guang River flora of northwestern Ethiopia. Review of Palaeobotany and Palynology 159: 188-194. https://doi.org/10.1016/j.revpalbo.2009.12.005
- Parham JF, Donoghue PC, Bell CJ, Calway TD, Head JJ, Holroyd PA, Inoue JG, Irmis RB, Joyce WG, Ksepka DT, Patane JS, Smith ND, Tarver JE, Tuinen M, Yang Z, Angielczyk KD, Greenwood JM, Hipsley CA, Jacobs L, Makovicky PJ, Müller J, Smith KT, Theodor JM, Warnock RC, Benton MJ (2012) Best practices for justifying fossil calibrations. Systematic Biology 61: 346-359. https://doi.org/10.1093/sysbio/syr107
- Phillips SJ, Dudik M, Elith J, Graham CH, Lehmann A, Leathwick J, Ferrier S (2009) Sample selection bias and presence-only distribution models: Implications for background and pseudo-absence data. Ecological Applications 19: 181-197. https://doi.org/10.1890/07-2153.1
- Piel KM (1971) Palynology of Oligocene sediments from central British Columbia. Canadian Journal of Botany 49: 1885-1920. https://doi.org/10.1139/b71-266
- Pimm SL, Joppa LN (2015) How many plant species are there, where are they and at what rate are they going extinct? Annals of the Missouri Botanical Garden 100: 170-176. https://doi.org/10.3417/2012018
- Pole M (1996) Plant macrofossils from the Foulden Hills Diatomite (Miocene), Central Otago, New Zealand. Journal of The Royal Society of New Zealand 26: 1-39. https://doi.org/10.1080/03014223.1996.9517503
- Posada D (2008) jModelTest: phylogenetic model averaging. Molecular Biology and Evolution 25: 1253-1256. https://doi.org/10.1093/molbev/msn083
- Prance G, Beentje H, Dransfield J, Johns R (2000) The tropical flora remains undercollected. Annals of the Missouri Botanical Garden 87: 67-71. https://doi.org/10.2307/2666209
- Qian H, Jin Y (2016) An updated megaphylogeny of plants, a tool for generating plant phylogenies and an analysis of phylogenetic community structure. The Plant Journal 9: 233-239.
- R Development Core Team (2009) R: A language and environment for statistical computing. Vienna, Austria:. R Foundation for Statistical Computing..
- Reid E, Chandler M (1926) Catalogue of Cainozoic plants in the department of Geology. Vol I. The Bembridge flora. British Museum (Natural History), London. https://doi.org/10.5962/bhl.title.110151
- Revell LJ, Harmon LJ, Collar DC (2008) Phylogenetic signal, evolutionary process, and rate. Systematic Biology 57: 591-601. https://doi.org/10.1080/10635150802302427
- Revell LJ (2012) phytools: An R package for phylogenetic comparative biology (and other things). Methods in Ecology and. Evolution 3: 217-223.
- Ronquist F, Teslenko M, Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology 61: 539-542. https://doi.org/10.1093/sysbio/sys029

- Roure B, Baurain D, Philippe H (2013) Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Molecular Biology and Evolution 30: 197-214. https://doi.org/10.1093/molbev/mss208
- Sanderson MJ, Shaffer HB (2002) Troubleshooting molecular phylogenetic analyses. Annual Review of Ecology and Systematics 33: 49-72. https://doi.org/10.1146/annurev.ecolsys.33.010802.150509
- Sanderson MJ (2003) r8s: Inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics 19: 301-302. https://doi.org/10.1093/bioinformatics/19.2.301
- Shen YY, Chen X, Murphy RW (2013) Assessing DNA barcoding as a tool for species identification and data quality control. PLoS ONE 8: e57125. https://doi.org/10.1371/journal.pone.0057125
- Sims H, Herendeen P, Crane P (1998) New fenus of fossil Fagaceae from the Santonian (Late Cretaceous) of Central Georgia, U.S.A. International Journal of Plant Sciences 159: 391-404. https://doi.org/10.1086/297559
- Smith SA, Beaulieu JM, Donoghue MJ (2010) An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants. Proceedings of the National Academy of Sciences USA 107: 5897-5902. https://doi.org/10.1073/pnas.1001225107
- Smith SA, O'Meara BC (2012) treePL: divergence time estimation using penalized likelihood for large phylogenies. Bioinformatics 28: 2689-2690. https://doi.org/10.1093/bioinformatics/bts492
- Smith SA, Brown JW (2018) Constructing a broadly inclusive seed plant phylogeny. American Journal of Botany 105: 302-314. https://doi.org/10.1002/ajb2.1019
- Soltis D, Soltis P (2009) The role of hybridization in plant speciation. Annual Review of Plant Biology 60: 561-588. https://doi.org/10.1146/annurev.arplant.043008.092039
- Soltis P, Soltis D, Savolainen V, Crane P, Barraclough T (2002) Rate heterogeneity among lineages of tracheophytes: integration of molecular and fossil data and evidence for molecular living fossils. Proceedings of the National Academy of Sciences USA 99: 4430-4435. https://doi.org/10.1073/pnas.032087199
- Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688-2690. https://doi.org/10.1093/bioinformatics/btl446
- Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A (2018) Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evolution 4 (16).
- Sun G, Dilcher DL, Wang H, Chen Z (2011) A eudicot from the Early Cretaceous of China. Nature 471: 625-628. https://doi.org/10.1038/nature09811
- Takahashi M, Crane PR, Manchester S (2002) *Hironoia fusiformis* gen. et sp. nov.: A cornalean fruit from the Kamikitaba locality (Upper Cretaceous, Lower Coniacian) in northeastern Japan. Journal of Plant Research 115: 463-473. https://doi.org/10.1007/s10265-002-0062-6
- Tel-Zur N, Abbo S, Myslabodski D, Mizrahi Y (1999) Modified CTAB procedure for DNA isolation from epiphytic cacti of the genera Hylocereus and Selenicereus (Cactaceae). Plant Molecular Biology Reporter 17: 249-254. https://doi.org/10.1023/A:1007656315275
- Thorne RF (2002) How many species of seed plants are there? Taxon 51: 511-522. https://doi.org/10.2307/1554864

- Vandelook F, Janssens SB, Probert RJ (2012) Relative embryo length as an adaptation to habitat and life cycle in Apiaceae. New Phytologist 195: 479-487. https://doi.org/10.1111/j.1469-8137.2012.04172.x
- Vandelook F, Janssens SB, Matthies D (2018) Ecological niche and phylogeny explain distribution of seed mass in the central European flora. Oikos 127: 1410-1421. https://doi.org/10.1111/oik.05239
- Vaudois-Miéja N (1983) Extension paléogéographique en Europe de l'actuel genre asiatique *Rehderodendron* Hu (Styracacées). Comptes-Rendus des Seances de l'Academie des Sciences, Série 2: Mecanique-Physique, Chimie, Sciences de l'Univers, Sciences de la Terre. 296. 125–130 pp.
- Wanntorp HE, Brooks DR, Nilson T, Nylin S, Ronquist F, Stearns SC, Wedell N (1990) Phylogenetic approach in ecology. Oikos 41: 119-132. https://doi.org/10.2307/3565745
- Webb CO, Ackerly DD, Mcpeek MA, M.J D (2002) Phylogenies and community ecology. Annual Review of Ecology and Systematics 33: 475-505. https://doi.org/10.1146/annurev.ecolsys.33.010802.150448
- Webb CO, Ackerly DD, Kembel SW (2008) Phylocom: software for the analysis of phylogenetic community structure and trait evolution. Bioinformatics 24: 2098-2100. https://doi.org/10.1093/bioinformatics/btn358
- Wehr WC, Manchester S (1996) Paleobotanical significance of Eocene flowers, fruits, and seeds from Republic, Washington. Washington Geology 24: 25-27.
- Wikström N, Savolainen V, Chase MW (2001) Evolution of the angiosperms: Calibrating the family tree. Proceedings of the Royal Society of London B Biological Sciences 268: 2211-2220. https://doi.org/10.1098/rspb.2001.1782
- Wilde V (1989) Untersuchungen zur Systematik der Blattreste aus dem Mitteleozan der Grube Messel bei Darmstadt (Hessen, Bundesrepublik Deutschland). Courier Forschungsinstitut Senckenberg 115: 1-213.
- Wilf P, Carvalho MR, Gandolfo MA, Cuneo NR (2017) Eocene lantern fruits from Godwanan Patagonia and the early origins of Solanaceae. Science 355: 71-75. https://doi.org/10.1126/science.aag2737
- Yao YG, Bravi CM, Bandelt HJ (2004) A call for mtDNA data quality control in forensic science. Forensic Science International 141: 1-6. https://doi.org/10.1016/j.forsciint.2003.12.004
- Zanne AE, Tank DC, Cornwell WK, Eastman JM, Smith SA, FitzJohn RG, McGlinn DJ, O'Meara BC, Moles AT, Reich PB, Royer DL, Soltis DE, Stevens PF, Westoby M, Wright IJ, Aarssen L, Bertin RI, Calaminus A, Govaerts R, Hemmings F, Leishman MR, Oleksyn J, Soltis PS, Swenson NG, Warman L, Beaulieu JM (2014) Three keys to the radiation of angiosperms into freezing environments. Nature 506: 89-92. https://doi.org/10.1038/nature12872
- Zhou Z, Crepet WL, Nixon KC (2001) The earliest fossil evidence of the Hamamelidaceae: Late Cretaceous (Turonian) inflorescences and fruits of Altingioideae. American Journal of Botany 88: 753-766. https://doi.org/10.2307/2657028

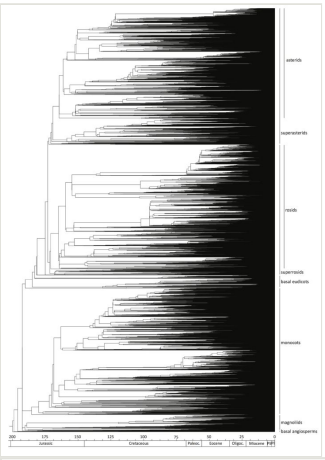

Figure 1.

Maximum Likelihood-based angiosperm phylogram based on the combined *rbcL* and *matK* (incl. *trnK*) dataset.

Table 1.

List of fossils used as calibration points, including their oldest stratigraphic occurrence, minimum and maximum ages, the calibrated clades and used references. cr.=crown, st.=stem.

Clade	Fossil	Reference	Period	Locality/Formation/Group
Ebenaceae	*Austrodiospyros cryptostoma* Basinger et Christophel	Basinger and Christophel 1985	Late Eocene	Anglesea formation (Victoria,
Apocynaceae	*Apocynophyllum helveticum* Heer	Wilde 1989	Middle Eocene	Messel formation (Darmstadt,
Cornaceae	*Hironoia fusiformis* Takahashi, Crane et Manchester	Takahashi et al. 2002	Early Conacian	Ashizawa formation, Futuba eastern Honshu, Japan)
Dipelta	*Dipelta europaea* Reid et Chandler	Reid and Chandler 1926	Late Eocene-Early Oligocene	Bembridge Flora (UK)
Oleaceae	*Fraxinus wilcoxiana* (Berry) Call et Dilcher	Call and Dilcher 1992	Middle Eocene	Claiborne formation (Tenness
Diervilla	*Diervilla echinata* Piel	Piel 1971	Oligocene	Fraser River system (British Canada)
Solanaceae (Physalinae)	*Physalis infinemundi* Wilf, Carvahlo, Gandolfo et Cuneo	Wilf et al. 2017	Early Eocene	Laguna del Hunco (Chubut, F Argentina)
Valeriana	*Valeriana sp.*	Mai 1985	Late Miocene	Europe
Emmenopterys	*Emmenopterys* Oliv.	Wehr and Manchester 1996	Middle Eocene	Middle Eocene Republic Flora (Washington, USA)
Pelliciera	*Pelliciera rhizophorae* Planch. et Triana	Graham 1977	Middle Eocene	Gatuncillo formation (Panama
Araliaceae	*Acanthopanax gigantocarpus* Knobloch et Mai	Knobloch and Mai 1986	Maastrichtian	Eisleben formation (Germany
Ilex	*Ilex hercynica* Mai	Mai 1987	Early Paleocene	Gonna formation (Sangerhaus Germany)
Actinidiaceae	*Saurauia antiqua* Knobloch et Mai	Knobloch and Mai 1986	Late Santonian	Klikov-Schichtenfolge (Germa
Nymphaeales	*unnamed Nymphaeales*	Friis et al. 2001	Late Aptian-Early Albian	Vale de Agua (Portugal)
Canellales	*Walkeripollis gabonensis* Doyle, Hotton et Ward	Doyle et al. 1990	Late Barremian-Early Aptian	Cocobeach (Gabon)
Magnoliaceae	*Archaeanthus linnenbergeri* Dilcher et Crane	Dilcher and Crane 1984	Early Cenomanian	Dakota formation (Kansas, U
Magnoliales	*Endressinia brasiliana* Mohr et Bernardes-de-Oliveira	Mohr and Bernardes-De-Oliveira 2004	Aptian-Albian	Crato formation (Brasil)
Lauraceae	*Potomacanthus lobatus* Crane, Friis et Pedersen	Crane et al. 1994	Early and Middle Albian	Puddledock locality (Virginia,
Arecaceae	unnamed palms	Christopher 1979, Daghlian 1981	Conacian-Santonian	Magothy formation (Marylanc
Musella-Ensete	*Ensete oregonense* Manchester et Kress	Manchester and Kress 1993	Middle Eocene	Clarno formation (Oregon, US
Zingiberaceae	*Zingiberopsis attenuata* Hickey et Peterson	Hickey and Peterson 1978	Middle to late Paleocene	Paskapoo formation (Alberta,

Zingiberales	*Spirematospermum chandlerae* Friis	Friis 1988	Santonian-Campanian	Neuse River formation (North USA)
Araceae	*Mayoa portugallica* Friis, Pedersen et Crane	Friis et al. 2004	Barremanian-Aptian	Almargem formation (Torres V Portugal)
Restionaceae	unnamed Restionaceae	Jarzen 1978	Maastrichtian	Morgan Creek (Saskatchewa
Poaceae	unnamed grasses	Jardiné and Magloire 1965	Maastrichtian	Senegal-Ivory Coast
Berberidaceae	*Mahonia* Nutt.	Manchester 1999	Middle Eocene	Green River formation (Color USA)
Platanaceae	*Platanocarpus brookensis Crane,* Pedersen, Friis et Drinnan	Crane et al. 1993	Early and Middle Albian	Patapsco formation (Virginia,
Sabiales	*Insitiocarpus moravicus* Knobloch et Mai	Knobloch and Mai 1986	Early Cenomanian	Peruc-schichten (Czeck Rep
Iteaceae	*Divisestylus brevistamineus*	Hermsen et al. 2003	Turonian	Raritan formation (New Jerse
Altingiaceae	*Microaltingia apocarpela*	Zhou et al. 2001	Turonian	Raritan formation (New Jerse
Tilia	*Tilia vescipites* Nichols et Ott	Nichols and Ott 1978	Middle Paleocene	Wind River basin (Wyoming,
Polygonaceae	*Persicaria* (L.) Mill.	Muller 1981	Paleocene	Europe
Clausena	*Clausena* Burm.f.	Pan 2010	Late Oligocene	Guang River Flora (Ethiopia)
Malpighiales	*Paleoclusia chevalieri* Crepet et Nixon	Crepet and Nixon 1998	Turonian	Raritan formation (New Jerse
Fagales	*Normapolles*	Batten 1981, Kedves 1989, Pacltova 1966	Late Cenomanian	Europa and USA
Phytolaccaceae	*Coahuilacarpon phytolaccoides* Cevallos-Ferriz, Estrada-Ruiz et Perez-Hernandez	Cevallos-Ferriz et al. 2008	Late Campanian	Cerro del Pueblo formation (M
Juglandaceae	*Cyclocarya brownii* Manchester et Dilcher	Crane et al. 1990	Late Paleocene	Almont and Beicegel Creek (I USA)
Rosales	*unnamed Rosidae*	Crepet and Nixon 1996	Turonian	Raritan formation (New Jerse
Betulaceae	*Endressianthus miraensis* Friis, Pedersen et Schoenenberger	Friis et al. 2003	Campanian-Maastrichtian	Mira (Portugal)
Fagaceae	*Antiquacupula sulcata* Sims, Herendeen et Crane	Sims et al. 1998	Late Santonian	Gaillard formation (Georgia, U
Salicaceae	*Pseudosalix handleyi* Boucher, Manchester et Judd	Boucher et al. 2003	Middle Eocene	Green River formation (Color USA)
Ranunculales	*Leefructus mirus* Sun, Dilcher, Wang et Chen	Sun et al. 2011	Barremanian-Aptian	Yixian formation (China)
Fabaceae	*Fabaceae sp.*	Herendeen et al. 1992	Early Eocene	Buchanan clay pit (Tenessee
Styracaceae	*Rehderodendron stonei* Vaudois-Mieja	Vaudois-Miéja 1983	Early Eocene	Sabals d'Anjou (France)
Dipterocarpaceae	*Shorea maomingensis* Feng, Kodrul et Jin	Feng et al. 2013	Late Eocene	Huangniuling formation (Maon China)
Lamiaceae	*Ajuginucula smithii* Reid et Chandler	Reid and Chandler 1926	Late Eocene-Early Oligocene	Bembridge Flora (UK)

Theaceae s.l.	*Pentapetalum trifasciculandricus* Martinez-Millan, Crepet et Nixon	Martinez-Millan et al. 2009	Turonian	Raritan formation (New Jerse
Myrsinaceae	*unnamed Myrsinaceae*	Pole 1996	Middle Miocene	Foulden Hills Diatomite (New
Myrtaceae	*Tristaniandra alleyi* Wilson et Basinger	Basinger et al. 2007	Middle Eocene	Golden Grove - East Yatala S (South Australia)
Lythraceae	*Decodon tiffneyi* Estrada-Ruiz, Calvillo-Canadell et Cevallos-Ferriz	Estrada-Ruiz et al. 2009	Late Campanian	Cerro del Pueblo formation (M
Ampelocissus s.l.	*Ampelocissus parvisemina* Chen et Manchester	Chen and Manchester 2007	Late Paleocene	Beicegal Creek (North Dakot
Vitaceae	*Indovitis chitaleyae* Manchester, Kapgate et Wen	Manchester et al. 2013	Maastrichtian	Mahurzari (India)
Rosa	*Rosa germerensis* Edelman	Edelman 1975	Early Eocene	Germer Basin Flora (Idaho, U
Prunus	*Prunus wutuensis* Li, Smith, Liu, Awasthi, Yang et Li	Li et al. 2011	Early Eocene	Wutu (China)
Myristicaceae	*Myristicacarpum chandlerae* Manchester, Doyle et Sauquet	Doyle et al. 2008	Early Eocene	London Clay (UK)

Supplementary materials

Suppl. material 1: Supplementary Table

Authors: Steven Janssens
Data type: Species list
Brief description: Table S1. Accession numbers of *rbcL* and *matK* (incl. *trnK*) sequences of the species included in the angiosperm phylogeny (including information on genera, family and order). Newly obtained accessions are indicated with an asterisk.
Download file (1.98 MB)

Suppl. material 2: Constraint input topology

Authors: Steven Janssens
Data type: Constraint topology
Brief description: Constraint input topology for RAxML analyses of all angiosperms analysed in this study (incl. outgroup taxa).
Download file (779.59 kb)

Suppl. material 3: Proportion of smoothing parameters

Authors: Steven Janssens
Data type: graph
Brief description: Proportion of smoothing parameters calculated for each of the 500 tree replicates
Download file (81.44 kb)

Suppl. material 4: Angiosperm phylogeny - ML bootstrap values

Authors: Steven Janssens
Data type: phylogeny
Brief description: Maximum Likelihood bootstrap consensus tree. Values above the branches indicate bootstrap support. Note that the support values above order level are all artificially set at 100 because of the use of a constraint backbone.
Download file (1.18 MB)

Suppl. material 5: Dated angiosperm phylogram

Authors: Steven Janssens
Data type: phylogeny
Brief description: Maximum Likelihood phylogram of 36101 angiosperm species (nexus file). Outgroup included. Blue bars indicate 95% confidence intervals.
Download file (23.66 MB)