SUPPORTING INFORMATION

Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary

Kevin Vanneste[1,2], Guy Baele[3], Steven Maere[1,2,*],

and Yves Van de Peer[1,2,4,*]

[1] Department of Plant Systems Biology, VIB, Ghent, Belgium

[2] Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium

[3] Department of Microbiology and Immunology, Rega Institute, KU Leuven, Leuven, Belgium

[4] Department of Genetics, Genomics Research Institute, University of Pretoria, Pretoria, South Africa

***Corresponding authors**

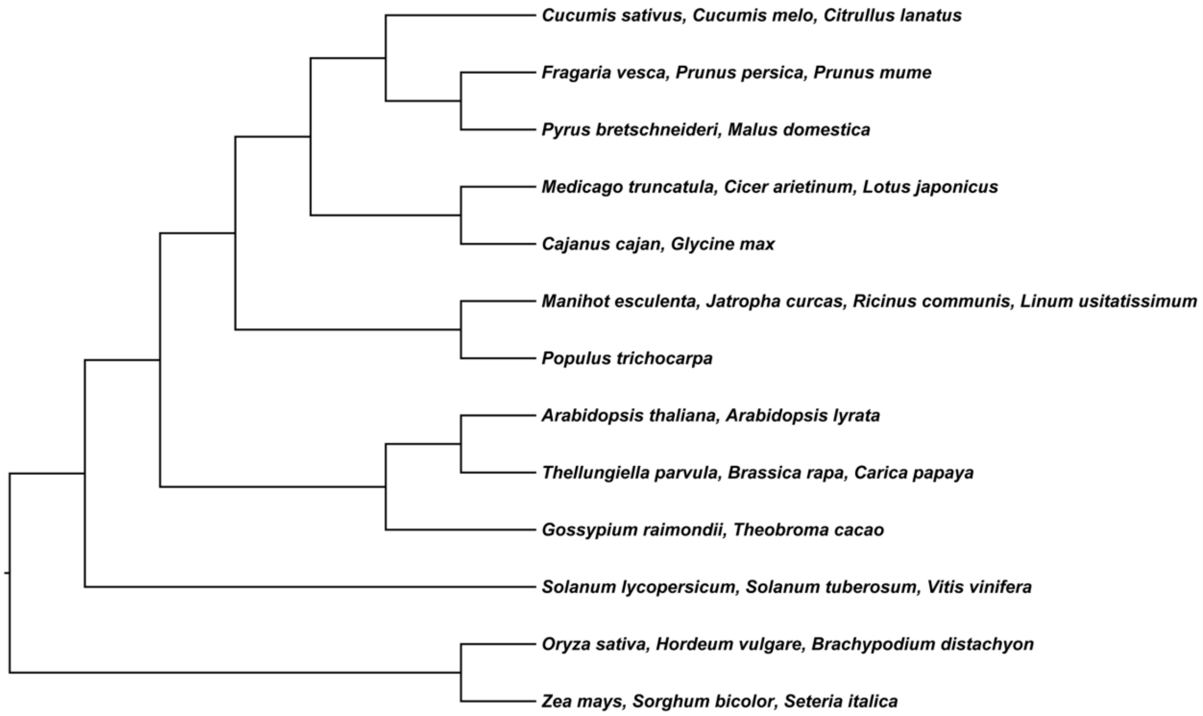
Yves Van de Peer
VIB / Ghent University
Technologiepark 927
Gent (9052), Belgium
Tel: +32 (0)9 331 3807
Fax: +32 (0)9 331 3809
E-mail: yves.vandepeer@psb.vib-ugent.be

Steven Maere
VIB / Ghent University
Technologiepark 927
Gent (9052), Belgium
Tel: +32 (0)9 331 3805
Fax: +32 (0)9 331 3809
E-mail: steven.maere@psb.vib-ugent.be

Overview

Species grouping topology

In order to date the node joining the homeologous pair, orthogroups were constructed consisting of both homeologs and orthologs from other plant species for which full genome sequence information was available. Different plant species were grouped into ‘species groups’ for which one ortholog was selected and added to the orthogroup, in order to keep the orthogroup topology fixed and to facilitate automation on the one hand, but also to allow enough orthogroups to be constructed on the other hand (see Material and methods). Supporting Figure S1 illustrates the employed species grouping topology.



Cucumis sativus, Cucumis melo, Citrullus lanatus
Fragaria vesca, Prunus persica, Prunus mume
Pyrus bretschneideri, Malus domestica
Medicago truncatula, Cicer arietinum, Lotus japonicus
Cajanus cajan, Glycine max
Manihot esculenta, Jatropha curcas, Ricinus communis, Linum usitatissimum
Populus trichocarpa
Arabidopsis thaliana, Arabidopsis lyrata
Thellungiella parvula, Brassica rapa, Carica papaya
Gossypium raimondii, Theobroma cacao
Solanum lycopersicum, Solanum tuberosum, Vitis vinifera
Oryza sativa, Hordeum vulgare, Brachypodium distachyon
Zea mays, Sorghum bicolor, Seteria italica

Supporting Figure S1 – Employed species grouping topology.

The topology presented in supporting Figure S1 is a trade-off between the total amount of sequence information within each individual orthogroup, and the total number of orthogroups that can be recovered. For instance, in case of the Brassicales, there is ample high-quality sequence information available from multiple genomes, so that splitting this order up in two different species groups (i.e., *A.*

thaliana and *A. lyrata* on the one hand, and *T. parvula*, *B. rapa*, and *C. papaya* on the other hand) instead of one single group entails that every orthogroup contains more sequence information (which increases the accuracy in the age estimate of the homeologous pair that is dated in the orthogroup), while the total number of recovered orthogroups also remains adequately high (which increases the total number of homeologous pairs that can be dated). Conversely, *Vitis* and *Solanum* were merged into one species group, because although splitting them would result in more sequence information per individual orthogroup, we found that in most cases not both a *Vitis* and *Solanum* ortholog could be found, drastically decreasing the total number of recovered orthogroups. The topology illustrated in supporting Figure S1 was the result of some 'trial-and-error', i.e., merging and splitting different groupings of species until we found a topology that maximized the total amount of sequence information per individual orthogroup, while still allowing a sufficiently large number of orthogroups to be recovered.

The topology presented in supporting Figure S1 also offers some additional advantages. First, it avoids any phylogenetic uncertainties, as the underlying topology between the different grouped species conforms to the well accepted current plant phylogeny (Jansen et al. 2007; Magallon and Castillo 2009; Wang et al. 2009; Bell et al. 2010; Smith et al. 2010; Smith et al. 2011; Soltis et al. 2011; Leitch et al. 2013), and is in accordance with the Angiosperm Phylogeny Group classification (APGIII) (Bremer et al. 2009). Second, because most often closely related species were grouped into species groups, the overall phylogenetic coverage remains high through including at least one ortholog for most major plant clades for which full genome sequence information is available. Third, WGDs in species not included in the topology could still be dated by introducing their homeologs at their respective phylogenetic location, after which one ortholog per species group (see supporting Figure S1) was added. This was the case for *L. sativa*, *A. formosa x pubescens*, and *N. advena*, because only a transcriptome assembly was available for these, for *P. heterocycla* because this genome only became available towards the end of this study when dating for the other species was finishing, for *P. patens* because of its very large phylogenetic distance from all the other species, and for *M. acuminata* and *P. dactylifera* because these were used only for dating WGDs in monocot species (see 'Calibrations and constraints'). The exact phylogenetic position of these species is indicated on Figure 3.

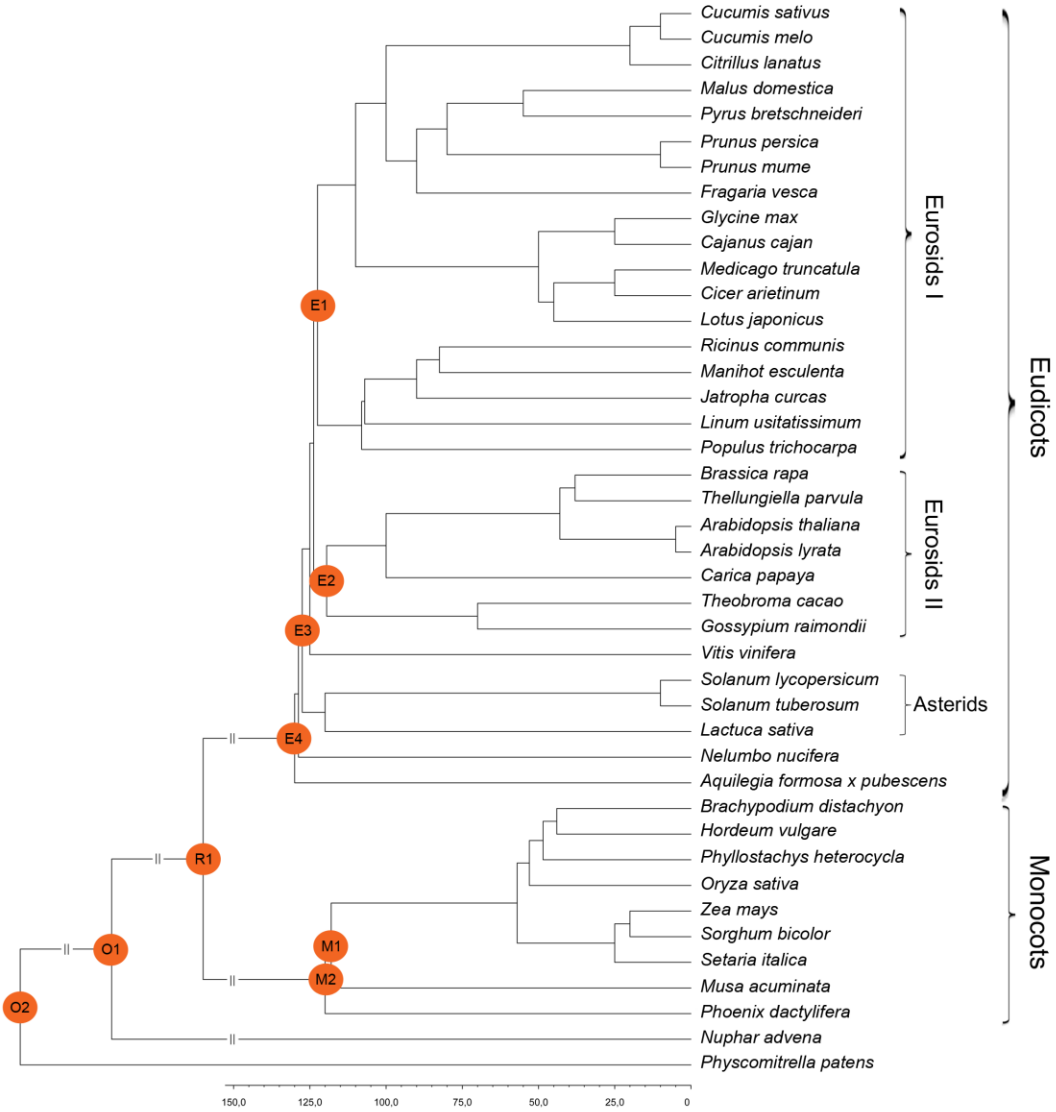
Calibrations and constraints

General

Recent molecular dating studies within the angiosperms benefit from a relatively wide array of fossil information that has become available, which typically allows implementing several high-quality primary fossil calibrations in large-scale dating studies where representatives from a large set of taxa are included based on a few high-quality sequenced marker genes (Bell et al. 2010; Magallon 2010; Smith et al. 2010; Clarke et al. 2011; Magallon et al. 2013). However, in our study, the value of any particular calibration is highly dependent on the species sampling in our trees, which is limited by the number of full plant genome sequences that are currently available. Only a small minority of the available fossils can in fact properly describe the divergence events within the species grouping topology (see supporting Figure S1). The majority of fossils routinely used in recent large-scale molecular dating studies cannot be used because no representative orthologs could be included in the orthogroups, due to the lack of a representative sequenced plant genome. For instance, there are several high-quality fossils available within the order Sapindales that could increase dating quality, but no representatives from this clade have been sequenced yet. Similarly, there are several high-quality fossils available within the order Arecales (Couvreur et al. 2011; Baker and Couvreur 2013), but only one representative genome sequence is currently available (*P. dactylifera*) so that all these fossils can only describe the same divergence event in the orthogroups (i.e., the divergence from a *P. dactylifera* ortholog from other monocot species orthologs) and are therefore redundant. In such cases, only the oldest available fossil can be used to describe the divergence event (Forest 2009).

A considerable body of literature has emerged in the last few years on the proper use of fossil data in molecular dating analysis. It is known that calibration priors in Bayesian time estimation can have a profound impact on posterior time estimates (Yang and Rannala 2006; Hug and Roger 2007; Inoue et al. 2010; Mulcahy et al. 2012; Magallon et al. 2013). Point calibrations result in illusionary precision of the posterior time estimate, so that flexible statistical distributions that describe the error associated with the fossil age more realistically are preferred (Ho and Phillips 2009). Early work focused on uniform distributions with hard minimum and maximum boundaries. These are however limited to clearly delineated fossil age boundaries, and can also lead to illusionary precision in the confidence intervals of the resulting posterior time estimate (Benton and Donoghue 2007). Such problems are mitigated by the introduction of soft maximum bounds that allow a certain small but nonzero part of the probability distribution, typically 2.5 to 5%, to be outside the maximum bound (Yang and Rannala 2006). The youngest possible age to which a fossil can reliably be attributed (based on radiometric dating, biostratigraphy etc.) still constitutes a hard minimum bound (Hug and Roger 2007). Soft maximum bounds eliminate the need for arbitrarily 'safe' high hard maximum bounds because they allow the sequence signal to overcome and correct poor calibrations by pulling the posterior past the maximum bound (Yang and Rannala 2006). Several flexible statistical distributions are commonly used but the lognormal distribution is particularly useful because of the way it mimics the error associated with estimating the divergence time of lineages from fossil information (Forest 2009; Magallon et al. 2013). It has a hard minimum bound but allows placing its peak mass probability anywhere between the minimum and maximum bound. This way, it can accommodate for the lag-phase between the first appearance of a particular fossil and the actual divergence event it documents, a discrepancy that has led to much controversy in the early days of molecular dating (Hedges and Kumar 2004). The lognormal distribution also accommodates for soft maximum bounds because it has an infinitely extending horizontal asymptote.

Recent research demonstrates that the use of arbitrary lognormal calibration priors without justification for their shape, perhaps not surprisingly, can however still have a profound impact on the resulting posterior time estimates (Warnock et al. 2012). Especially the position of the peak mass probability within the calibration boundaries has been demonstrated to pull the posterior time estimates towards its location (Clarke et al. 2011; Warnock et al. 2012). There is no reason to assume that the lag between lineage origin and first fossil occurrence will be consistent for all calibration points across the tree (Hugall et al. 2007). Guidelines about the magnitude of the parameters of the lognormal distribution are therefore currently assigned based on rough confidence around prior beliefs, see for instance Magallon et al. (Magallon et al. 2013). We calibrated any particular divergence by concentrating the prior peak mass probability on the most recent and accurate estimates found in literature (described below in detail for the individual calibrations). Although these literature-based estimates do not necessarily represent the true time of divergence, their effect on posterior time estimates should be less biased compared to a strategy where the peak mass probability is always arbitrarily placed at the beginning, middle, or end of a calibration interval. The proper placement of the calibration priors was always checked by performing a run without data (Drummond et al. 2006) because the marginal calibration prior does not necessarily correspond to the desired calibration density, since the former is combined with the tree prior (Heled and Drummond 2012). A starting tree with branch lengths satisfying all the fossil prior constraints was manually constructed. Supporting Figure S2 represents an overview of both the initial tree branch lengths and all fossil calibrations (initial branch lengths were implemented based on the specific ortholog selected for each species group).



Supporting Figure S2 - Tree with initial branch lengths and employed fossil calibrations. Branch lengths are truncated after 150 mya for improved clarity (the initial branch length for the divergence described by O2, O1, and R1, was put at 450 mya, 220 mya, and 170 mya, respectively).

Eudicot calibrations (E1, E2, E3, and E4)

E1 is based on the fossil *Paleoclusia chevalieri*, which is the oldest known fossil we found from the order Malpighiales (Crepet and Nixon 1998). This fossil originates from the South Amboy Fire Clay at Old Crossman Clay Pit (New Jersey, USA), with a minimum bound of 82.8 mya (Clarke et al. 2011). This fossil is a member of the Clusiaceae family, but there exists some uncertainty whether the Clusiaceae split off between the Salicaceae and Euphorbiaceae (Davis et al. 2005), or if they are rather sister to both of these (Xi et al. 2012). We therefore used this fossil to calibrate the divergence of the total group Malpighiales from their nearest sister group for which full genome sequence information was available, namely the remainder of the Eurosids I. The divergence between the former has been estimated at ~122.5 mya (Xi et al. 2012). The mode of the lognormal distribution is located at $e^{\mu-\sigma^2}$, with μ and σ the mean and standard deviation of the lognormal distribution, respectively. We therefore specified a lognormal calibration prior with μ = 3.9314, σ = 0.5, and a minimum bound of 82.8 mya (because the peak of the lognormal calibration prior is hence located at $82.8 + e^{3.9314-0.5^2} = 122.5$ mya).

E2 is based on the fossil *Dressiantha bicarpellata*, which is the oldest known fossil from the order Brassicales (Gandolfo et al. 1998), also originating from the South Amboy Fire Clay at Old Crossman Clay Pit (New Jersey, USA). We used this fossil to calibrate the divergence of the Brassicales from their nearest sister group for which full genome sequence information was available, namely the order Malvales. The divergence between the former has been estimated at ~119.5 mya (Beilstein et al. 2010). We therefore specified a lognormal calibration prior with μ = 3.8528, σ = 0.5, and a minimum bound of 82.8 mya.

E3 is based on the fossil *Icacinicarya budvarensis*, which is the oldest known fossil from the asterids (Pigg et al. 2008). This fossil originates from České Budějovice Budvar (Czech Republic), with a minimum bound of 89.3 mya (Bremer et al. 2004). We used this fossil to calibrate the divergence of the asterids from their nearest sister group for which full genome sequence information was available, namely the remainder of the rosids. The divergence between the former has been estimated at ~125 mya (Bremer et al. 2004; Bell et al. 2010). We therefore specified a lognormal calibration prior with μ = 3.8252, σ = 0.5, and a minimum bound of 89.3 mya.

E4 is based on the fossil *Leefructus mirus*, which is the oldest known fossil from the order Ranunculales (Sun et al. 2011). This fossil originates from the Daxinfangzi Bed at the Yixian Formation (China), with a minimum bound of 123.0 mya. We used this fossil to calibrate the divergence of the Ranunculales from their nearest sister group for which full genome sequence information was available, namely the total group of rosids and asterids. The divergence between the former has been estimated at ~130 mya (Anderson et al. 2005; Bell et al. 2010). We therefore specified a lognormal calibration prior with μ = 2.1959, σ = 0.5, and a minimum bound of 123.0 mya.

Performing a run without data (Drummond et al. 2006; Heled and Drummond 2012) indicated however that implementation of all these four calibrations resulted in a situation where the marginal prior calibration distributions did not correspond to their specified calibration densities anymore. Rather, the prior calibration distributions of E1 and E2 pushed away the prior calibration distributions of E3 and E4, most likely because they were located on consecutive nodes (see supporting Figure S2). Calibrations E3 and E4 was therefore only used when dating WGDs in the asterids (i.e., *S. lycopersicum*, *S. tuberosum*, and *L. sativa*), and Ranunculales (i.e., *A. formosa x pubescens*), respectively, while calibrations E1 and E2 were used for dating WGDs in all other species (including non-eudicots). This ensures that always at least one rate-correcting calibration was present between the homeologous pair and root for dating the WGDs in all eudicot species.

Monocot calibrations (M1 and M2)

M1 and M2 were used only when dating WGDs in monocot species (i.e., *O. sativa*, *B. distachyon*, *Z. mays*, *S. bicolor*, *M. acuminata*, *S. italica*, *P. heterocycla*, *H. vulgare*, and *P. dactylifera*). This is because monocot calibrations necessitated the inclusion of either *M. acuminata* or *P. dactylifera* into the orthogroups, which led to a drastic drop in orthogroup recovery. This was true especially when dating WGDs in non-monocot species, but also to a large extent for dating WGDs in monocot species themselves, which is why we considered *M. acuminata* and *P. dactylifera* as a single species group and required only one representative ortholog with its corresponding calibration to be present (i.e., there are two possible monocot calibrations that were only implemented when dating WGDs in monocot species to ensure at least one rate-correcting calibration between the root and homeologous pair, but for each orthogroup only one was implemented based on whether a *M. acuminata* or *P. dactylifera* ortholog was added to the orthogroup).

M1 is based on the fossil *Spirematospermum chandlerae*, which is the oldest known fossil from the order Zingiberales (Friis 1988). This fossil originates from the Black Creek Formation at Neuse River Cut-Off (North Carolina, USA), with a minimum bound of 83.5 mya. We used this fossil when a *M. acuminata* ortholog was included in the orthogroup to calibrate the divergence of the Zingiberales from their nearest sister group for which full genome sequence information was available, namely the order Poales. The divergence between the former has been estimated at ~118 mya (Janssen and Bremer 2004; Kress 2006). We therefore specified a lognormal calibration prior with μ = 3.7910, σ = 0.5, and a minimum bound of 83.5 mya.

M2 is based on the fossil *Sabalites carolinensis*, which is the oldest known fossil from the order Arecales (Berry 1914). This fossil originates from the Black Creek Formation near Langley (South Carolina, USA), with a minimum bound of 85.8 mya (Couvreur et al. 2011). We used this fossil when a *P. dactylifera* ortholog was included in the orthogroup to calibrate the divergence of the Arecales from their nearest sister group for which full genome sequence information was available, namely the order Poales. The divergence between the former has been estimated at ~120 mya (Janssen and Bremer 2004; Couvreur et al. 2011; Baker and Couvreur 2013). We therefore specified a lognormal calibration prior with μ = 3.7822, σ = 0.5, and a minimum bound of 85.8 mya.

Root calibration (R1)

R1 is based on the sudden abundant appearance of eudicot tricolpate pollen in the fossil record at ~125 mya at several separate geographical localities (Doyle 2005). An error of 1 million year based on magnetostratigraphic evaluation is associated with the above described estimate of 125 mya, placing its minimum bound effectively at 124.0 mya (Clarke et al. 2011). We used this fossil information to calibrate the divergence of the eudicots from the monocots, which constitutes the root of orthogroup phylogenies. Selecting an appropriate peak mass probability location for this divergence is however less straightforward because there exists considerable variation in its estimate, ranging from about 140 mya until as old as 200 mya (Wikstrom et al. 2001; Bell et al. 2010; Magallon 2010; Smith et al. 2010; Clarke et al. 2011). We consequently selected a peak mass probability at 170 mya (effectively the middle of these intervals), and therefore specified a lognormal calibration prior with μ = 4.0786, σ = 0.5, and a minimum bound of 124.0 mya. The more uncertain position of this split, in combination with placing a soft bound on the maximum root age, could place undue weight on the assumption of the age of the root (Clarke et al. 2011). The effects thereof on our results are however most likely small because for all species, with the exception of *N. advena* and *P. patens* (see below), at least one extra rate-correcting calibration was incorporated between the root and homeologous pair.

N. advena **and** ***P. patens*** **calibrations (O1 and O2)**

N. advena and *P. patens* were not part of the species grouping topology because of their isolated basal position in the plant phylogeny. Applying the same strategy as for other species not part of the species grouping topology, i.e., adding the homeologous pair at its respective phylogenetic location in the orthogroup topology, entails however that a new root is instituted. When dating the WGD in *N. advena* and *P. patens*, we therefore implemented O1 and O2 as new root calibrations, respectively.

O1 is based on the sudden abundant appearance of eudicot tricolpate pollen in the fossil record at 125 mya at several separate geographical localities (Doyle 2005), with a minimum bound of 124.0 mya (see before). We used this fossil information to calibrate the divergence of the *N. advena* homeologous pair from the eudicots and monocots, which constitutes the new root when the *N. advena* WGD was dated. This divergence has been estimated at ~220 mya (Smith et al. 2010; Clarke et al. 2011; Magallon et al. 2013). We therefore specified a lognormal calibration prior with μ = 4.8143, σ = 0.5, and a minimum bound of 124.0 mya.

O2 is based on the fossil *Cooksonia*, which is the oldest known fossil from the Lycopsida (Edwards and Feehan 1980). This fossil originates from the Cloncannon Formation of County Tipperary (Ireland), with a minimum bound of 420.4 mya (Clarke et al. 2011). We used this fossil to calibrate the divergence of the *P. patens* homeologous pair from the eudicots and monocots, which constitutes the new root when the *P. patens* WGD was dated. This divergence has been estimated at ~450 mya (Smith et al. 2010; Clarke et al. 2011; Magallon et al. 2013). We therefore specified a lognormal calibration prior with μ = 3.6378, σ = 0.5, and a minimum bound of 420.4 mya.

Alternative calibrations and constraints

General

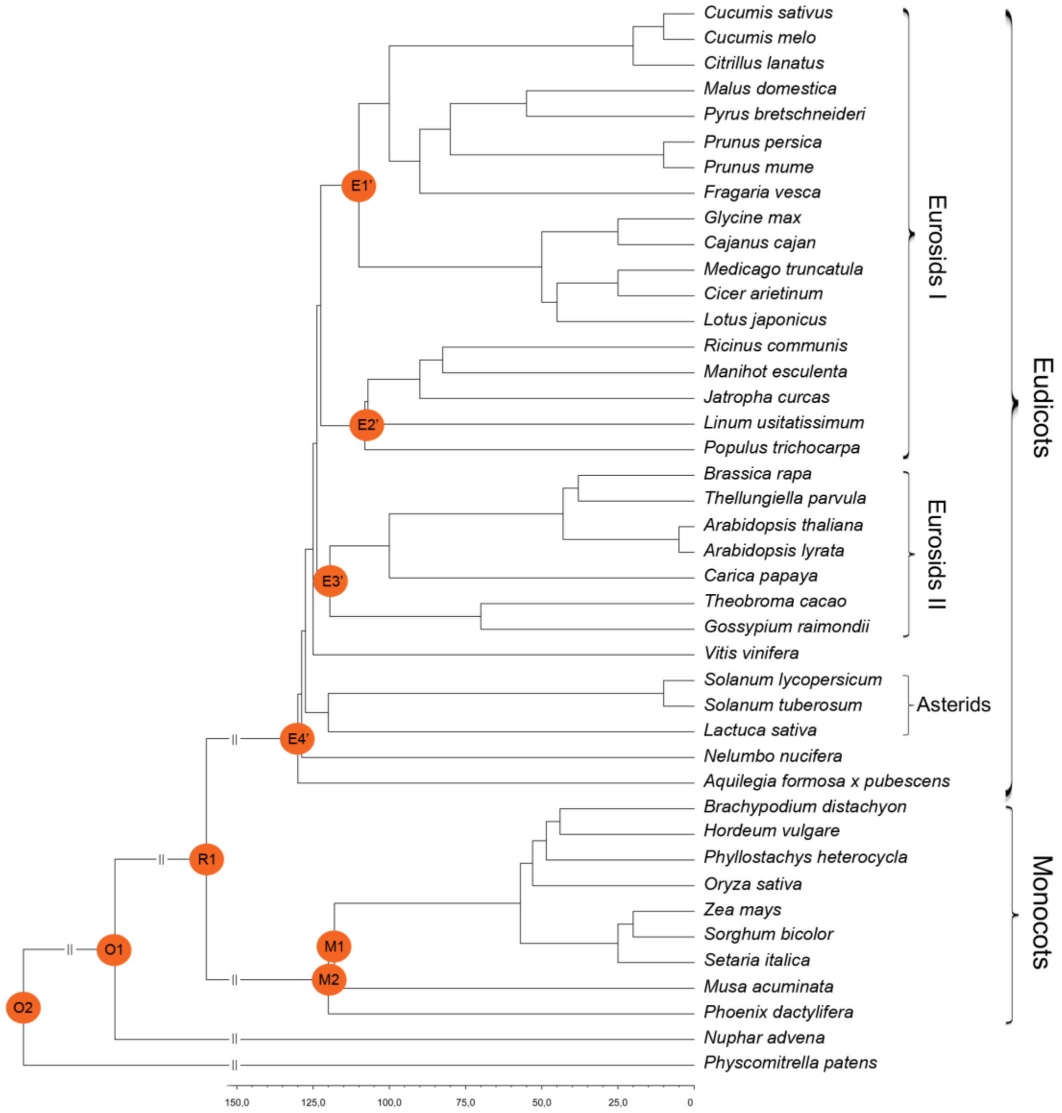
The set of calibrations used for the WGD age estimates presented in Table 1 (see main manuscript) are necessarily limited through the availability of full genome sequences and the species grouping topology (see 'Calibrations and constraints'). With regard to the remaining fossil calibration options, some of the choices we made may seem suboptimal at first sight. In particular, one may wonder why we did not adopt the eudicot crown node calibration based on eudicot tricolpate fossil pollen, in accordance with its sudden abundant appearance in the fossil record at ~125 mya (Doyle 2005). The latter has a long history of use in molecular dating studies to enforce a hard maximum bound of 125 mya on the eudicot crown node. The interpretation of this fossil information has however recently been called into question. The earliest tricolpate fossil pollen already displays considerable structural variety and can be found across widespread geographical localities, suggesting that they represent the rise to dominance, rather than the first origin of the eudicots (Smith et al. 2010). Additionally, the recent description of a fossil from the early-branching eudicot order Ranunculales estimated at 122.6-125.8 mya, argues that eudicots may have already been present some time before 125 mya (Sun et al. 2011). The latter is also supported by several recent clade-specific molecular dating studies that place key divergence events within the eudicots typically very close to 125 mya (Beilstein et al. 2010; Sauquet et al. 2012; Xi et al. 2012). Although it is difficult to explain why eudicots would remain hidden for so long if they had already diversified into clades that rose so rapidly in the mid-Cretaceous, angiosperms possibly originated in isolated freshwater lake-related wetlands from where they later quickly invaded other habitats, which would explain the discrepancy in the molecular record (Coiffard et al. 2012).

In light of this recent uncertainty, we preferred avoiding any controversy by not including this fossil calibration in our dating analysis. However, most recent large-scale molecular dating studies of the angiosperms converge mostly on the same age estimates for key divergence events within the eudicots, irrespective of whether this calibration was employed or not (Bell et al. 2010; Magallon 2010; Smith et al. 2010; Clarke et al. 2011; Magallon et al. 2013). Not surprisingly, studies that impose a hard maximum bound of ~125 mya on the eudicot crown typically find age estimates that are somewhat younger than studies that do not impose this constraint, but both nevertheless agree particularly well on most divergence time estimates within the eudicots, despite the fact that both disagree strongly on their estimates for the age of the eudicots themselves. We investigated the effects of including this eudicot crown calibration in our analysis by rerunning a substantial part of the calculations on our dataset with this particular calibration implemented (see below).

Simultaneously, we took advantage of the relatively rich fossil record of the eudicots to investigate how reliable our WGD age estimates are under an alternative calibration set. For instance, the fossil *Dressiantha bicarpellata* was used in our original calibration set to describe the divergence of the order Brassicales, in which it was originally placed based on morphological data (Gandolfo et al. 1998). This classification was later challenged by a combined molecular sequence + morphological character analysis (De Craene and Haston 2006), but afterwards placed firmly again within the Brassicales based on a more recent combined molecular sequence + morphological character analysis (Beilstein et al. 2010). This fossil has consequently been used in a series of recent molecular dating studies (Magallon and Castillo 2009; Beilstein et al. 2010; Couvreur et al. 2010; Clarke et al. 2011). Here, we studied the effect of omitting this fossil calibration in favor of other calibrations (see below).

The alternative calibration set

Re-dating all constructed orthogroups with an alternative calibration set was computationally prohibitive due to the immense computational resources required for running the MCMC component of the molecular sequence divergence estimation (Suchard and Rambaut 2009; Ayres et al. 2012). We therefore chose to re-date all orthogroups based on anchors, because these are based on actual duplicated segments, and we only employed orthogroups based on peak-based duplicates if the former were not available (i.e., for *L. sativa*, *A. formosa x pubescens*, *H. vulgare*, and *N. advena*). The analysis methods were exactly the same as described in the main manuscript (see Material and methods), with the exception that the original calibration set within the eudicots (i.e., E1, E2, E3, and E4 - see supporting Figure S2) was replaced in all orthogroups by a new alternative calibration set (i.e., E1’, E2’, E3’, and E4’ - see supporting Figure S3), as discussed in the next paragraphs.



Supporting Figure S3 - Tree with initial branch lengths and employed fossil calibrations for the alternative calibration set. Branch lengths are truncated after 150 mya for improved clarity (the initial branch length for the divergence described by O2, O1, and R1, was put at 450 mya, 220 mya, and 170 mya, respectively).

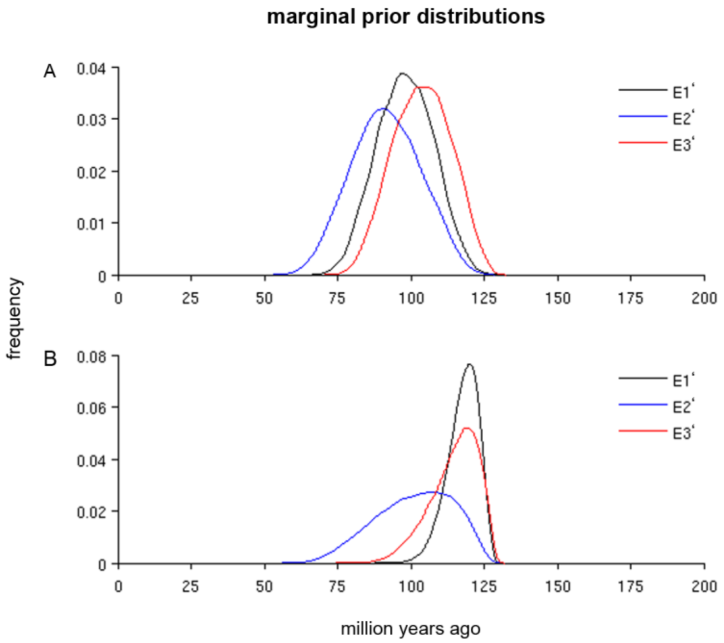
The alternative calibration E1’ is based on an unnamed fossil from the order Fabales (Herendeen and Crane 1992), which is the oldest known fossil we found for this order, with a minimum bound of 59.9 mya. We used this fossil to calibrate the divergence of the Fabales from their nearest sister group for which full genome sequence information was available, namely the total group Rosales + Cucurbitales. The divergence between the former has been estimated at ~120 mya (Sauquet et al. 2012). We therefore specified a lognormal calibration prior with μ = 4.3460 (but see below), σ = 0.5, and a minimum bound of 59.9 mya.

E2’ is based on the fossil *Pseudosalix*, which is the oldest known fossil from the family Salicaceae (Boucher et al. 2003), with a minimum bound of 48.0 mya. We used this fossil to calibrate the divergence of the Salicaceae from their nearest sister group for which full genome sequence information was available, namely all other representatives from the order Malpighiales. The divergence between the former has been estimated at ~108 mya (Xi et al. 2012). We therefore specified a lognormal calibration prior with μ = 4.3443 (but see below), σ = 0.5, and a minimum bound of 59.9 mya.

E3’ is based on the fossil *Parbombacaceoxylon*, which is the oldest known fossil from the order Malvales (Wheeler et al. 1987; Wheeler et al. 1994), with a minimum bound of 65.5 mya. We used this fossil to calibrate the divergence of the Malvales from their nearest sister group for which full genome sequence information was available, namely the Brassicales. The divergence between the former has been estimated at ~119.5 mya (Beilstein et al. 2010). We therefore specified a lognormal calibration prior with μ = 4.2390 (but see below), σ = 0.5, and a minimum bound of 65.5 mya.

E4’ is based on the aforementioned eudicot tricolpate fossil pollen at ~125 mya (Doyle 2005). We used this fossil information to constrain the crown group of the eudicots with a maximum age. To accommodate some small margin of error around this boundary, as suggested by recent findings of a fossil from the early-branching eudicot order Ranunculales estimated at 122.6-125.8 mya (Sun et al. 2011), we imposed a hard bound of 130 mya on the eudicots by implementing a uniform calibration prior between 0 and 130 mya.

We found that when imposing E4’ and running a scenario without data (Drummond et al. 2006), the marginal prior calibration distributions of E1’, E2’, and E3’ did not correspond to their specified calibration densities anymore. This type of behavior has been observed before, and has been ascribed to the fact that the marginal prior distribution is the combination of both the specified calibration density and the tree prior (Heled and Drummond 2012; Warnock et al. 2012). In fact, we experienced that implementing calibrations on nodes that were located very close to each other, in particular consecutive nodes, always resulted in a discrepancy between the specified calibration densities and effective marginal prior calibration distributions. We therefore increased parameter μ of calibrations E1', E2', and E3' until their marginal prior calibration distributions corresponded with their specified location at μ = 8.0978, μ = 4.5675, and μ = 5.0703, respectively, as also illustrated in supporting Figure S4.



Supporting Figure S4 - Marginal prior distributions for calibrations E1', E2', and E3' when E4' was also implemented with (**A**) μ = 4.3460, μ = 4.3443, and μ = 4.2390, respectively (**B**) μ = 8.0978, μ = 4.5675, and μ = 5.0703, respectively.

WGD age estimates under the alternative calibration set

Supporting Table S1 summarizes the WGD age estimates and their 90% CIs, as obtained using the alternative calibration set, while supporting Figure S5 illustrates the resulting absolute age distributions.

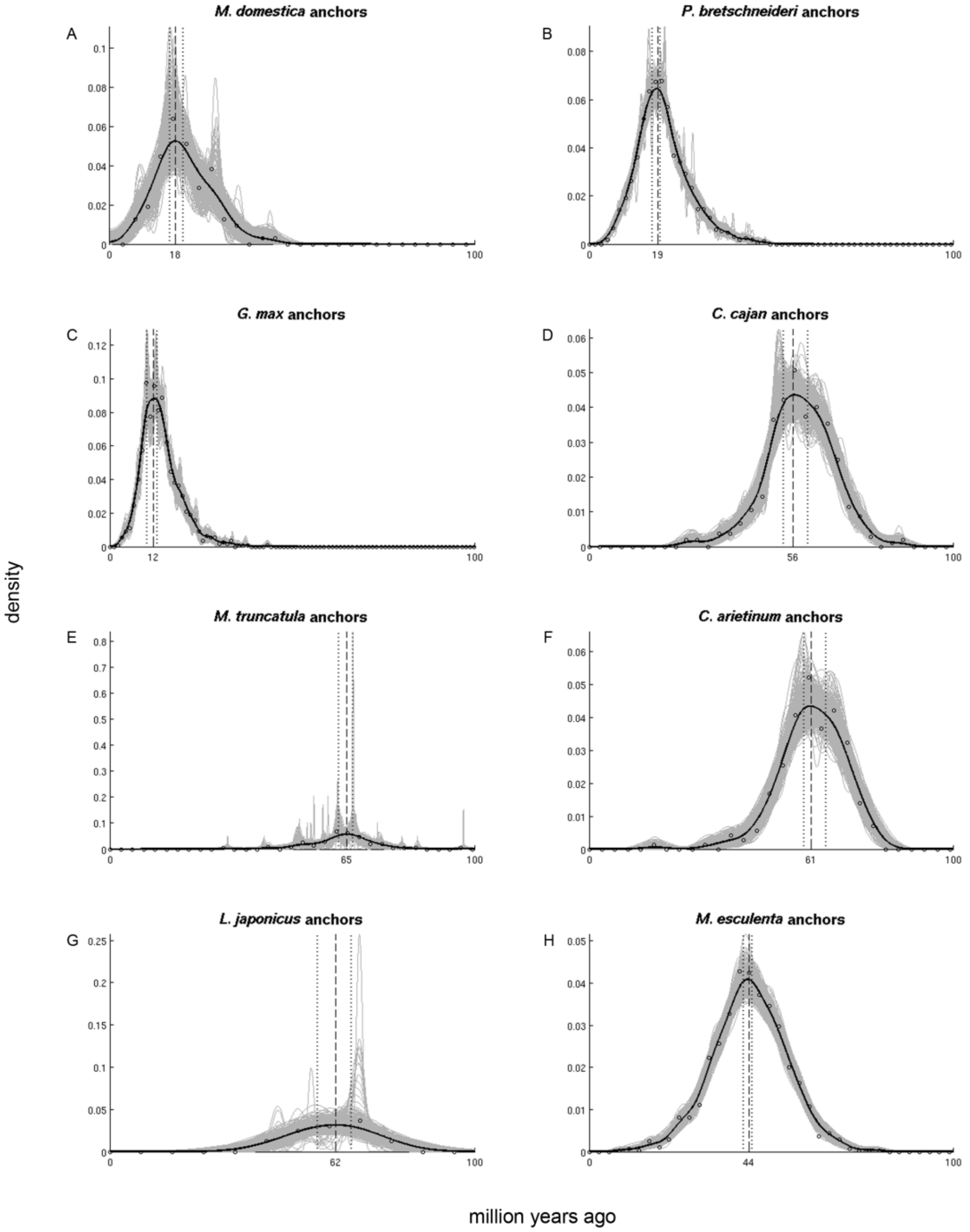
Supporting Table S1. Overview of the number of dated and accepted (ESS >200 for all statistics, see Material and methods) orthogroups per species, and their resulting WGD age estimates with 90% confidence intervals (CIs). All orthogroups are based on anchors, except if indicated otherwise.

Species	**# Dated (accepted) orthogroups**	**WGD age estimate (90% CI)**
Malus domestica	99 (90)	17.95 (16.48-20.07)
Pyrus bretschneideri	1,000 (986)	18.53 (17.47-19.45)
Glycine max	1,000 (987)	12.31 (10.33-13.08)
Cajanus cajan	361 (351)	56.41 (53.41-60.26)
Medicago truncatula	79 (77)	64.95 (62.78-66.67)
Cicer arietinum	210 (201)	60.73 (59.01-65.20)
Lotus japonicus	19 (19)	61.87 (56.96-66.26)
Manihot esculenta	1,000 (977)	43.52 (42.45-44.80)
Linum usitatissimum	1,000 (987)	9.67 (8.94-10.62)
Populus trichocarpa	1,000 (983)	35.38 (34.07-36.56)
Brassica rapa	1,000 (975)	24.95 (23.22-26.34)
Thellungiella parvula	779 (758)	46.01 (44.91-47.14)
Arabidopsis thaliana α*	754 (736)	47.58 (45.90-48.75)
Arabidopsis thaliana β*	9 (9)	55.86 (0-65.20)
Arabidopsis lyrata	706 (686)	46.37 (45.13-47.22)
Gossypium raimondii	1,000 (968)	54.36 (53.00-55.49)
Solanum lycopersicum	479 (466)	62.27 (61.01-63.63)
Solanum tuberosum	478 (462)	59.74 (57.77-62.67)
Lactuca sativa †	451 (422)	55.97 (53.70-57.80)
Aquilegia formosa x pubescens †	55 (49)	51.17 (45.82-60.55)
Brachypodium distachyon	319 (300)	66.04 (63.85-68.75)
Hordeum vulgare †	323 (303)	72.93 (70.26-74.49)
Phyllostachys heterocycla	503 (487)	18.53 (17.47-20.11)
Oryza sativa	334 (319)	62.75 (60.37-68.28)
Zea mays	948 (913)	19.30 (18.42-19.93)
Sorghum bicolor	170 (164)	66.08 (63.11-69.96)
Setaria italica	309 (296)	66.15 (64.10-68.75)
*Musa acuminata***	367 (346)	65.27 (61.54-67.73)
Phoenix dactylifera	32 (29)	53.11 (47.66-55.79)
Nuphar advena †	119 (115)	69.23 (63.74-73.15)
Physcomitrella patens	319 (255)	55.79 (51.83-65.79)

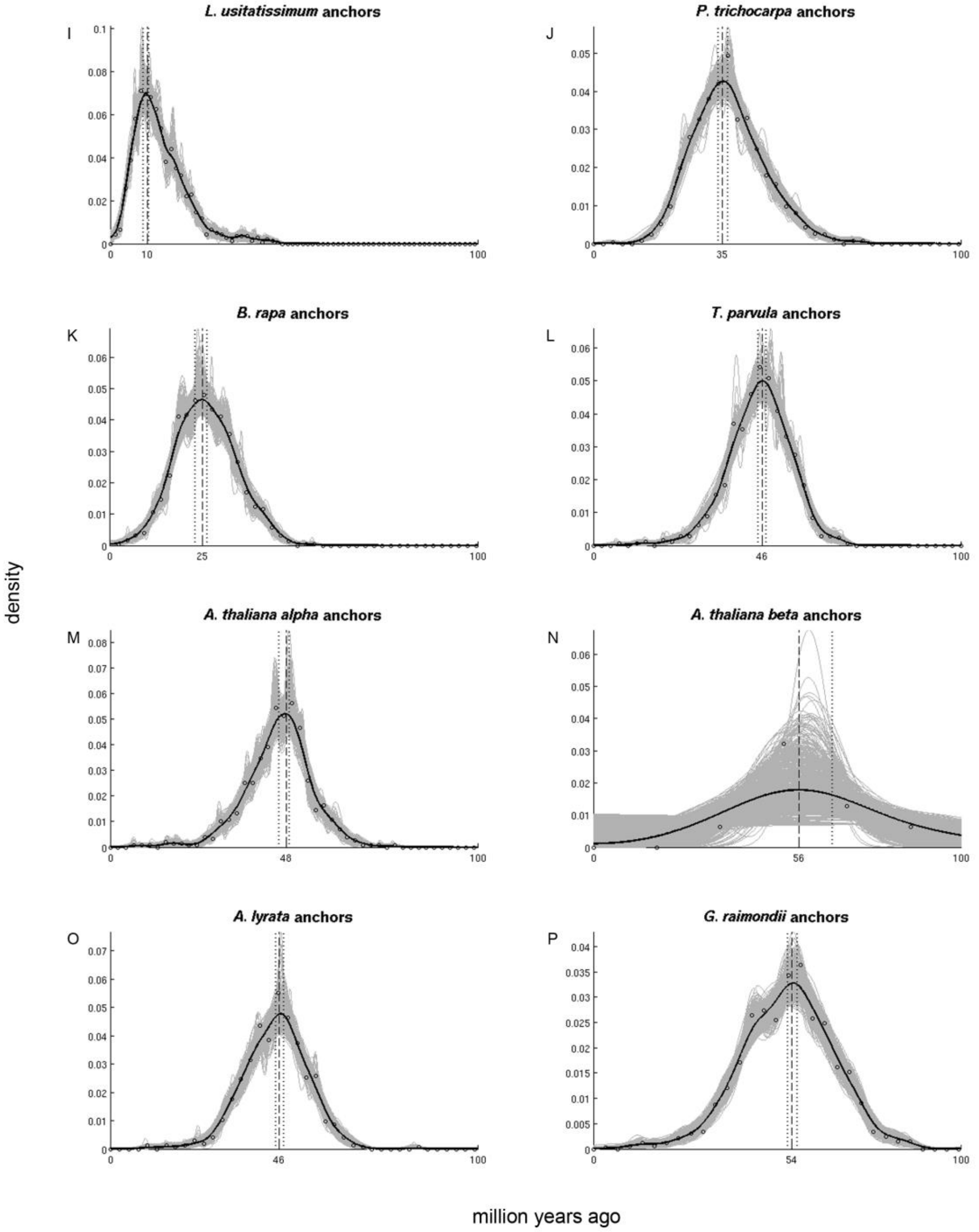
† Based on peak-based duplicates.

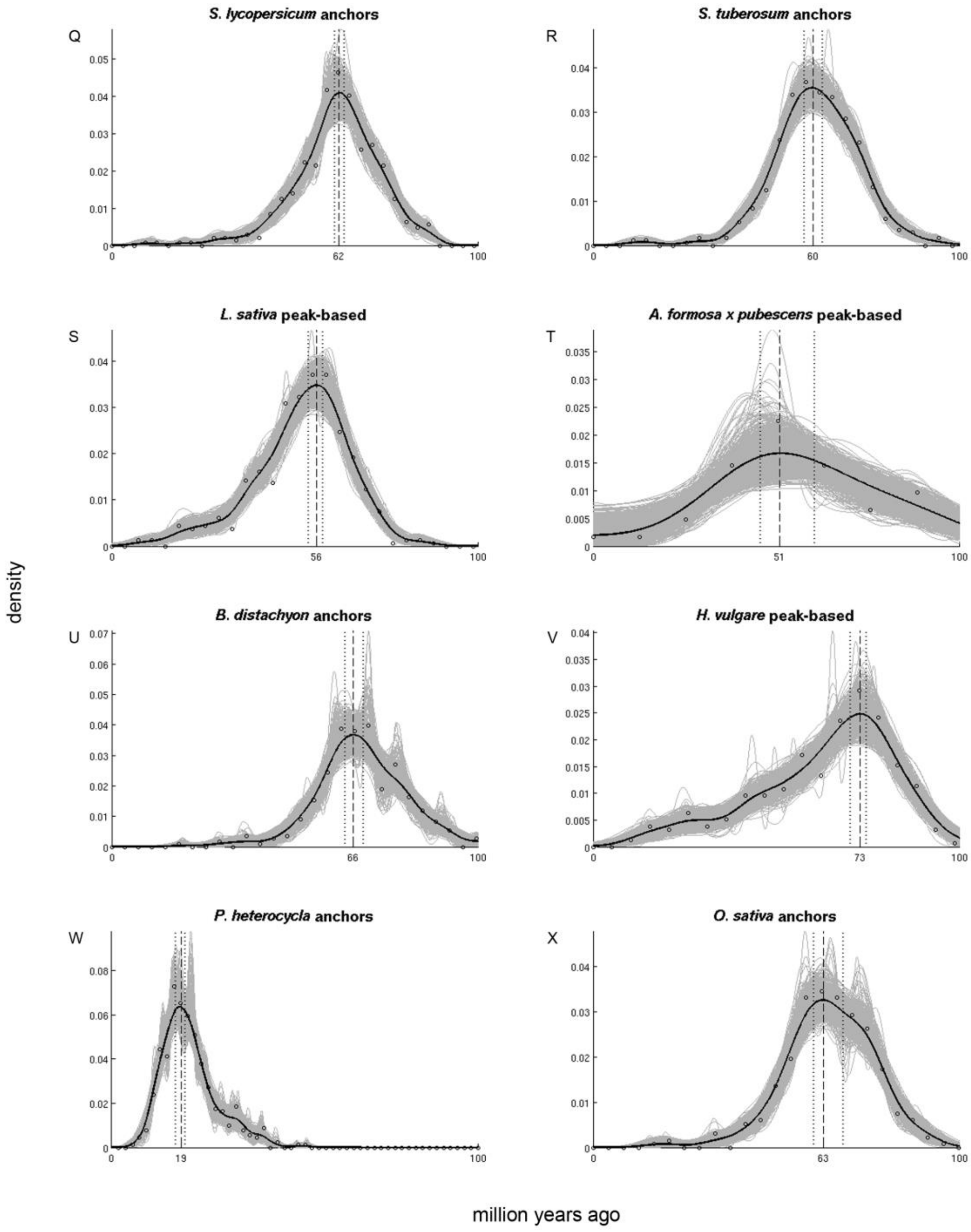
* α and β refer to the *A. thaliana alpha* and *beta* duplication, respectively (Bowers et al. 2003).

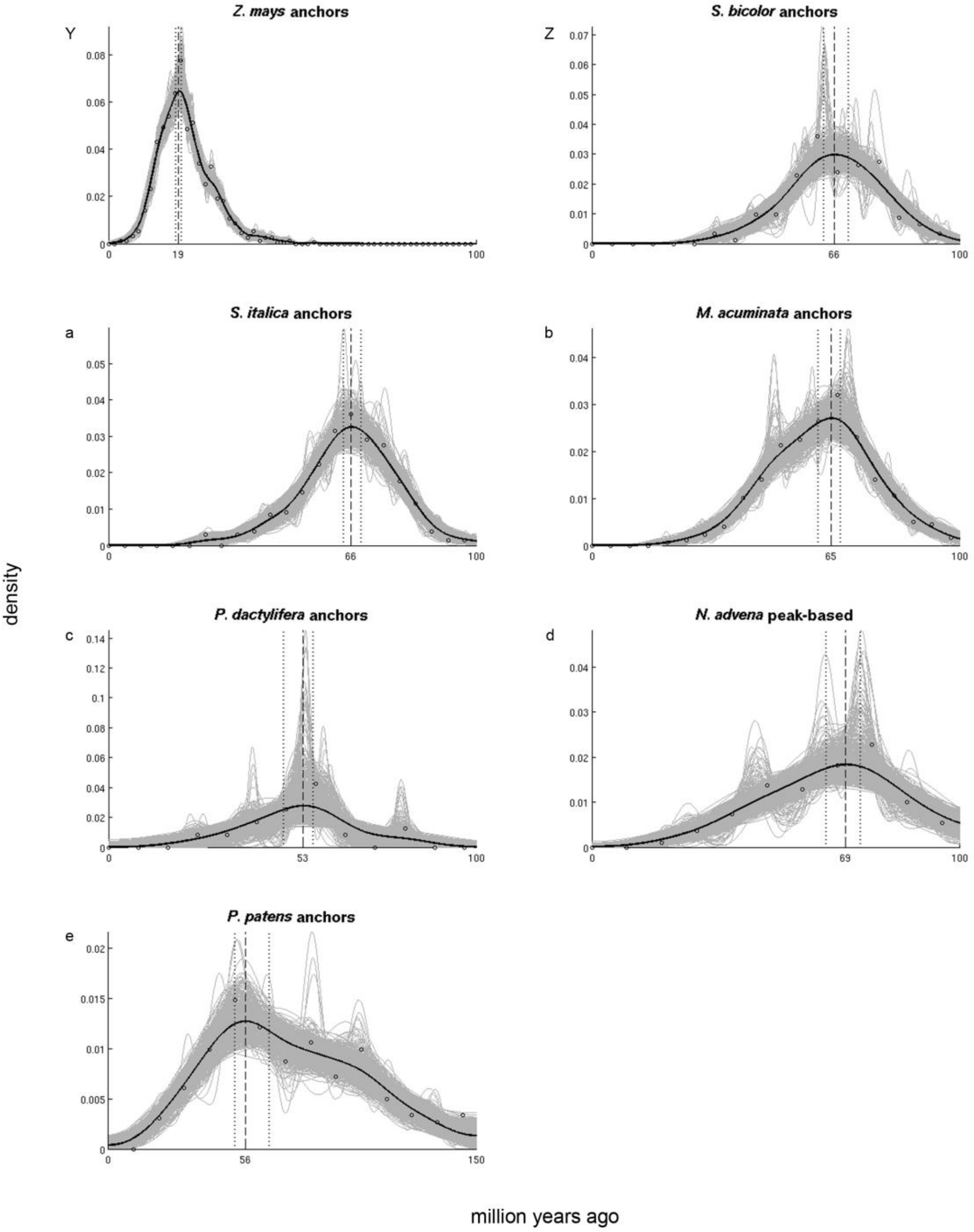
** This event most likely represents 2 separate WGDs in close succession (D'Hont et al. 2012).



M. domestica anchors
A
18
P. bretschneideri anchors
B
19
G. max anchors
C
12
C. cajan anchors
D
56
M. truncatula anchors
E
65
C. arietinum anchors
F
61
L. japonicus anchors
G
62
M. esculenta anchors
H
44
0
100
density
million years ago


I
L. usitatissimum anchors
0.1
0.08
0.06
0.04
0.02
0
0
10
100
J
P. trichocarpa anchors
0.05
0.04
0.03
0.02
0.01
0
0
35
100
K
B. rapa anchors
0.06
0.05
0.04
0.03
0.02
0.01
0
0
25
100
L
T. parvula anchors
0.06
0.05
0.04
0.03
0.02
0.01
0
0
46
100
M
A. thaliana alpha anchors
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
48
100
N
A. thaliana beta anchors
0.06
0.05
0.04
0.03
0.02
0.01
0
0
56
100
O
A. lyrata anchors
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
46
100
P
G. raimondii anchors
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
54
100
density
million years ago


S. lycopersicum anchors
Q
S. tuberosum anchors
R
L. sativa peak-based
S
A. formosa x pubescens peak-based
T
B. distachyon anchors
U
H. vulgare peak-based
V
P. heterocycla anchors
W
O. sativa anchors
X
62
60
56
51
66
73
19
63
0
100
density
million years ago


Z. mays anchors
Y
0.08
0.06
0.04
0.02
0
0
19
100
S. bicolor anchors
Z
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
66
100
S. italica anchors
a
0.05
0.04
0.03
0.02
0.01
0
0
66
100
M. acuminata anchors
b
0.04
0.03
0.02
0.01
0
0
65
100
density
P. dactylifera anchors
c
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
53
100
N. advena peak-based
d
0.04
0.03
0.02
0.01
0
0
69
100
P. patens anchors
e
0.02
0.015
0.01
0.005
0
0
56
150
million years ago

Supporting Figure S5 - Absolute age distributions obtained under the alternative calibration set for (**A**) *M. domestica*, (**B**) *P. bretschneideri*, (**C**) *G. max*, (**D**) *C. cajan*, (**E**) *M. truncatula*, (**F**) *C. arietinum*, (**G**) *L. japonicus*, (**H**) *M. esculenta*, (**I**) *L. usitatissimum*, (**J**) *P. trichocarpa*, (**K**) *B. rapa*, (**L**) *T. parvula*, (**M**) *A. thaliana alpha*, (**N**) *A. thaliana beta*, (**O**) *A. lyrata*, (**P**) *G. raimondii*, (**Q**) *S. lycopersicum*, (**R**) *S. tuberosum*, (**S**) *L. sativa*, (**T**) *A. formosa x pubescens*, (**U**) *B. distachyon*, (**V**) *H. vulgare*, (**W**) *P. heterocycla*, (**X**) *O. sativa*, (**Y**) *Z. mays*, (**Z**) *S. bicolor*, (**a**) *S. italica*, (**b**) *M. acuminata*, (**c**) *P. dactylifera*, (**d**) *N. advena*, and (**e**) *P. patens*. The black solid line represents the kernel density estimate of the dated homeologs, while the vertical dashed line represents its peak used as WGD age estimate. The grey lines represent the density estimates for the 1,000 bootstrap replicates, while the vertical dotted lines represent the corresponding 90% confidence intervals on the WGD age estimate. The original raw distribution of dated homeologs is also indicated on the individual plots by open dots. See supporting Table S1 for sample sizes and exact confidence interval boundaries.

The WGD age estimates obtained under the alternative calibration set presented in supporting Table S1 generally agree very well with the WGD age estimates obtained under the original calibration set presented in Table 1. Not surprisingly, implementation of a hard maximum bound on the eudicot crown node results in WGD age estimates and 90% CIs that are slightly younger. A similar shift is also apparent in other large-scale molecular dating studies within the angiosperms where this constraint was implemented (Magallon 2010), compared to studies where this was not the case (Smith et al. 2010). However, the 90% CIs obtained under the alternative calibration set overlap in all but two cases with the 90% CIs obtained under the original calibration set, and are on average only 1.57 million years younger for the complete set of all 31 species-specific WGD age estimates presented in supporting Table S1. The *G. raimondii* WGD, and the Brassicaceae *alpha* WGD shared by *A. thaliana*, *A. lyrata*, and *T. parvula*, constitute the only two WGDs where the 90% CIs of WGD age estimates obtained under the alternative calibration set do not overlap with those of the original calibration set. The *G. raimondii* WGD is 3.66 million years younger under the alternative calibration set, while the Brassicaceae *alpha* WGD is 2.53 million years younger (average of WGD age estimates of *A. thaliana*, *A. lyrata*, and *T. parvula*).

The WGD age estimates obtained under the original calibration set can arguably be considered more reliable for three reasons. First, with regard to the hard maximum bound used for the eudicot fossil pollen calibration, it needs to be remarked that a fossil can in fact only provide unequivocal evidence on a hard minimum bound, but not on a hard maximum bound. A hard minimum bound is provided by the earliest age to which the fossil can reliably be attributed to, whereas a maximum bound always needs to be inferred based on other types of evidence such as older fossils and stratigraphic information. The latter is therefore error-prone, which is exactly why soft maximum bounds were introduced (Yang and Rannala 2006). Recently, it was convincingly demonstrated that when the sequence signal is sufficiently strong and indicates an age different from the one suggested by the fossil calibration prior, soft maximum bounds can indeed allow to overcome a strong calibration prior (Magallon et al. 2013), whereas this evidently is not possible when a hard maximum bound has been imposed. Additionally, it has been suggested that hard maximum bounds result in narrower confidence intervals on the posterior divergence time estimates, which do not represent genuine high precision but rather the conflict between fossil and sequence information (Yang and Rannala 2006). Soft maximum bounds are therefore always preferred, and it was in fact argued that eudicot tricolpate fossil pollen constituted the only exception against these guidelines that was deemed acceptable (Forest 2009). In light of later scrutiny of the interpretation of eudicot tricolpate fossil pollen (Smith et al. 2010; Sun et al. 2011), a calibration strategy that strictly follows the conservative guidelines detailed above without allowing for any exceptions is preferable. Such a strategy does not question the value of eudicot tricolpate fossil pollen itself, but simply applies the same rules as enforced for all other fossil information.

Second, irrespective of this hard maximum constraint on the eudicot crown node, the fossils employed in the alternative calibration set may also be less optimal in the context of molecular sequence divergence estimation. The alternative calibration set contains calibrations with minimum bounds located more closely to the tips of the orthogroups compared to the original calibrations. It has been demonstrated that an abundance of constraints near the tips can bias the estimates for deeper nodes (Bell et al. 2005). Further, because the alternative calibrations have much younger minimum bounds, but necessarily still describe divergence events quite far from these minimum bounds due to a lack of genome sequences for intermediate taxa, the resulting marginal calibration priors are much wider, and hence more diffuse and uninformative (see supporting Figure S4). Informative calibration priors on these nodes are nevertheless important because they represent a period of angiosperm diversification that is characterized by “layer upon layer of rapid radiation” (Bell et al. 2010), for which informative calibration priors are most likely imperative to guide the posterior divergence time estimates. Simply combining all calibrations from both the original and alternative calibration set is not a viable option, because this would result in a scenario where the large majority of nodes within the orthogroup topology have a calibration prior imposed. This is problematic because calibrating the large majority of the available nodes can only lead to conclusions compatible with the prior assumptions, since even a very strong sequence signal will not be able to correct posterior divergence time estimates if the majority of the nodes situated close to the divergence of interest (i.e., the homeologous pair) carry a strong prior (Hugall et al. 2007). Additionally, the effective marginal prior distributions and specified calibration densities will always differ when specified priors on nested clades overlap temporally (Warnock et al. 2012), which is something we noticed in our own dataset as soon as calibration priors were specified on nodes located too close to each other.

Third, evaluation of the resulting absolute age distributions for all species-specific WGDs obtained under the alternative calibration set (see supporting Figure S5), indicates that they become less informative compared to the absolute age distributions obtained under the original calibration set (see Figure 2 and supplementary Figure S2). This is for instance particularly evident for the *A. thaliana beta* absolute age distribution. The original WGD age estimate and 90% CI of 61.21 mya and 54.58 to 69.38 mya, respectively, were necessarily based on only nine dated anchor pairs (see Table 1). Despite this very low number, we deemed this WGD age estimate fairly reliable because of the relatively strong unimodal pattern of its absolute age distribution (see supplementary Figure S2, panel I). Furthermore, this was re-affirmed by its peak-based absolute age distribution that was based on a much larger number of orthogroups, but still arrived at a very similar WGD age estimate and 90 % CI of 62.97 mya and 56.04 to 70.01 mya, respectively. Under the alternative calibration set however, a WGD age estimate and 90% CI of 55.86 mya and 0 to 65.20 mya, respectively, were obtained for this WGD (see supporting Table S1). The latter appears a particularly strong shift, but evaluation of the new absolute age distribution indicates that it exhibits a very uninformative shape (see supporting Figure S5, panel N). In particular, its kernel density estimate is very wide with only a poorly supported peak, as also indicated by the bootstrap replicates that reveal a mostly flat surface curve with a very diffuse peak. Consequently, the resulting 90% CI is over 65 million years wide. Although the uninformative shape of this absolute age distribution obtained under the alternative calibration set is not particularly striking, considering that it only consists of nine dated anchors, the drastic difference with the informative shape obtained under the original calibration set is remarkable. This most likely indicates that the new constraints imposed by the alternative calibration set conflict with the sequence signal to some extent.

In conclusion, using an alternative calibration set with in particular a hard maximum constraint on the eudicot crown node, we find that the resulting WGD age estimates are overall in good agreement with those obtained under the original calibration set, being on average only 1.57 mya younger and possessing overlapping 90% CIs for all but two independent WGDs, suggesting that our conclusions are robust against the particular choice of employed calibrations.

Relative rate tests

To obtain a measure for the relative rate at which species used in dating the WGDs evolve, we performed pairwise relative rate tests (RRTs) between the different WGDs. We used *P. patens* as an outgroup, since this allows consistent comparison of all other dated WGDs. Anchors and peak-based duplicates from different species used for dating WGDs were collected and grouped by plant order. Transcriptome assemblies were not considered because no positional information is available for these. Supporting Table S2 lists all employed species.

Supporting Table S2 – Overview of species employed for RRT comparisons

Plant order	Code	Species
Rosales	ROS	*P. bretschneideri, M. domestica*
Fabales	FAB	*M. truncatula, C. cajan, L. japonicus, C. arietinum*
Malpighiales	MAL	*M. esculenta, P. trichocarpa*
Brassicales	BRA	*A. thaliana, A. lyrata, T. parvula*
Malvales	MAV	*G. raimondii*
Solanales	SOL	*S. lycopersicum, S. tuberosum*
Poales	POA	*O. sativa, B. distachyon, S. italica, S. bicolor, H. vulgare*
Zingiberales	ZIN	*M. acuminata*
Arecales	ARE	*P. dactylifera*

The evolutionary rates between orthologs used in dating the WGDs, grouped by plant order, were then compared in a pairwise fashion. Orthogroups were constructed for each pairwise comparison based on Inparanoid data for *P. patens*, and always included the *P. patens* ortholog as outgroup and two orthologs representing the specific plant orders being compared. We performed the RRTs employing HyPhy (v2.0) (Pond et al. 2005), using a WAG model of evolution (Whelan and Goldman 2001) with gamma-distributed rate heterogeneity across sites using four rate categories (Yang 1996) for all orthogroups. Supporting Table S3 lists the fraction of all orthogroups evolving faster, and the total sample sizes, between all pairwise comparisons of orders. Supporting Table S4 does the same but only considers the orthogroups that were found to evolve significantly faster (*P*-value < 0.05).

Supporting Table S3 - Fraction of orthogroups evolving faster for the orders listed in the rows compared to the orders listed in the columns. The lower diagonal of the matrix lists the percentages, while the upper diagonal lists the sample sizes upon which these percentages are based.

from/to	ROS	FAB	MAL	BRA	MAV	SOL	POA	ZIN	ARE
ROS		438	1129	544	71	460	552	161	303
FAB	0.56		660	450	52	406	469	107	200
MAL	0.46	0.38		841	120	666	846	216	439
BRA	0.63	0.57	0.66		74	503	633	98	252
MAV	0.52	0.42	0.53	0.23		55	79	22	27
SOL	0.52	0.46	0.53	0.41	0.56		524	99	175
POA	0.62	0.63	0.68	0.51	0.70	0.58		120	249
ZIN	0.55	0.55	0.56	0.38	0.50	0.40	0.38		97
ARE	0.45	0.43	0.50	0.35	0.56	0.46	0.31	0.41	

Supporting Table S4 - Fraction of orthogroups evolving significantly faster (*P*-value < 0.05) for the order listed in the rows compared to the orders listed in the columns. The lower diagonal of the matrix lists the percentages, while the upper diagonal lists the sample sizes upon which these percentages are based.

from/to	ROS	FAB	MAL	BRA	MAV	SOL	POA	ZIN	ARE
ROS		49	94	71	4	43	95	14	36
FAB	0.65		89	73	n/a	47	83	13	36
MAL	0.44	0.27		115	7	77	143	25	57
BRA	0.77	0.67	0.83		12	42	73	9	58
MAV	0.25	n/a	0.43	0.17		6	10	3	3
SOL	0.58	0.36	0.51	0.31	0.50		74	8	21
POA	0.75	0.65	0.87	0.59	1.00	0.72		17	38
ZIN	0.79	0.54	0.72	0.33	0.33	0.63	0.35		9
ARE	0.58	0.36	0.65	0.28	1.00	0.52	0.08	0.44	

To facilitate evaluation, we scored each comparison binary as either evolving faster (1) or slower (0) depending on the fractions listed in supporting Table S4, using 50% as the cut-off. Since for the comparison between the Malvales and Fabales, no single statistically significant orthogroup was identified, this was scored as 1 based on the comparison of all their orthogroups in supporting Table S3. Similarly, since exactly half of all scored orthogroups evolved slower/faster for the comparison between the Solanales and Malvales, this was scored as 0 based on the comparison of all their orthogroups in supporting Table S3. The resulting binary matrix is listed in supporting Table S5.

Supporting Table S5 - Binary matrix representing the relationships between all considered plant orders. 0 and 1 represent an overall slower or faster evolutionary rate between the orders listed in the rows compared to the orders listed in the columns, respectively.

from/to	ROS	FAB	MAL	BRA	MAV	SOL	POA	ZIN	ARE
ROS		0	1	0	1	0	0	0	0
FAB	1		1	0	1	1	0	0	1
MAL	0	0		0	1	0	0	0	0
BRA	1	1	1		1	1	0	1	1
MAV	0	0	0	0		0	0	1	0
SOL	1	0	1	0	1		0	0	0
POA	1	1	1	1	1	1		1	1
ZIN	1	1	1	0	0	1	0		1
ARE	1	0	1	0	1	1	0	0	

Although our current approach is arguably very crude because different species belonging to the same plant order do not necessarily share the same evolutionary rates, similar trends based on similar life history traits are expected (Smith and Donoghue 2008). We tried an alternative strategy where individual species instead of plant orders were compared but this led to sample sizes that were too low for statistical evaluation. Despite the fact that our results should therefore be interpreted with due caution, our current approach allows for a rudimentary comparison between the different plant orders. This is supported by the fact that the resulting relationships between the different plant orders in the binary matrix are very consistent, ordered from slowest to fastest as follows:

MAV < MAL < ROS < SOL < ARE < FAB < ZIN < BRA < POA

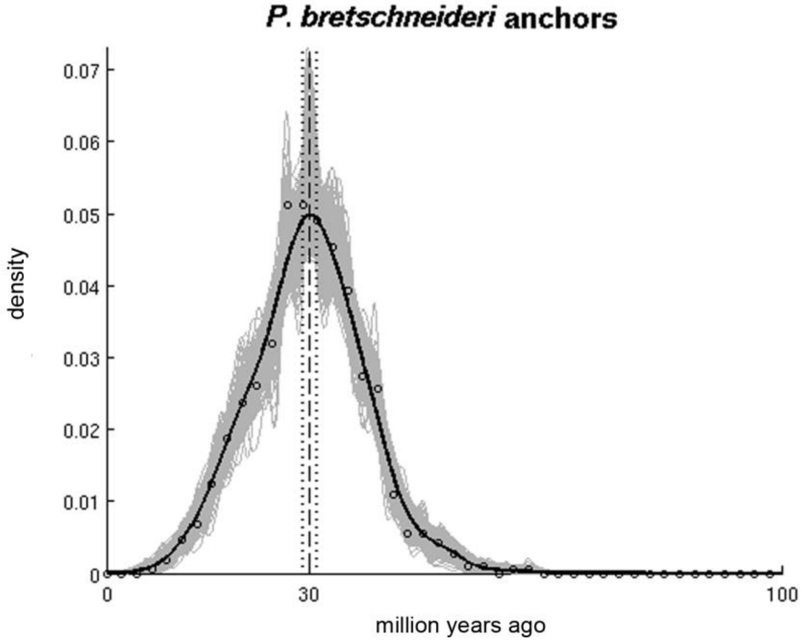
The above association represents the most parsimonious relationship between all plant orders. There was only one error in the binary matrix against this relationship, namely the comparison between the Zingiberales and Malvales, which was scored as 0 but should have been scored as 1. This is most likely because of a low sample size, as only three orthogroups were scored as statistically significant. All other comparisons in the binary matrix were consistent according to the relationships listed above.

Re-dating the *Pyrus bretschneideri* WGD

In the main manuscript, we presented fossil evidence that suggests that the ages of both the *P. trichocarpa* WGD and the WGD shared by *M. domestica* and *P. bretschneideri* are underestimated by our dating approach, most likely because of a drastic rate shift associated with their woody status that could not be completely corrected for (see Results and discussion). In case of the *P. trichocarpa* WGD, we quoted fossil information that establishes that the divergence between *Salix* and *Populus* is at least 47.4 million years old (Boucher et al. 2003). Although there is no genome sequence information available for *Salix*, it is well established that *Salix* and *Populus* shared the WGD in question (Tuskan et al. 2006; Berlin et al. 2010). A calibration on the node joining the *P. trichocarpa* homeologous pair enforcing a minimum age of 47.4 million years could therefore theoretically have been implemented. However, it remains very difficult to decide on a proper shape for the calibration prior that would not inadvertently bias the eventual WGD age estimate. Lognormal calibration priors are preferred (Magallon et al. 2013), but posterior time estimates are pulled to some extent towards their peak mass probability (Clarke et al. 2011). Incorporating prior information on the location of the peak mass, for which the current best estimate is in fact ~65 mya (Fawcett et al. 2009), would hence be highly undesirable because it entails placing a strong peak mass probability at 65 mya on the node joining the homeologous pair. Alternative shapes for this particular calibration are equally questionable. The most basic form, a uniform calibration prior, requires arbitrarily ‘safe’ high maximum bounds, since it is very difficult to distinguish proper upper boundaries based on the fossil record (Yang and Rannala 2006). The risk that the sequence signal is not strong enough to overcome poor calibration priors is inherent to all molecular dating (Yang and Rannala 2006). A strategy that avoids placing any *a priori* fossil evidence upon the node joining the homeologous pair is hence preferable because it ensures that the sequence signal of this node will yield the most unbiased age estimate possible, based upon other rate-correcting calibrations in the orthogroup topology.

The same applies to the WGD shared by *M. domestica* and *P. bretschneideri*. There is fossil evidence that indicates that their divergence should be at least 48.7 million years old (Wehr and Hopkins 1994), so that a calibration with this minimum bound could theoretically have been implemented on their homeologous pairs, which is nevertheless undesirable in light of the above. However, because there are more sequenced Rosaceae genomes available, we can break up the long branch leading to the homeologous pair by increasing the taxon sampling around this node, and also introduce a new calibration based on this fossil information closer to, but not on, the homeologous pair. Applying the same strategy for *P. trichocarpa* is impossible because the latter is the only genome available at the moment within the Salicaceae, while the most closely related available genome sequences are situated within other families of the Malpighiales, which all diverged about ~100 mya (Xi et al. 2012). We re-dated the *P. bretschneideri* WGD based on its anchors, because these are based on bona fide duplicated segments and many more anchors were available for this species compared to *M. domestica* (see Table 1). To break up the long branch leading the *P. bretschneideri* homeologous pairs, we included both one *Fragaria* and *Prunus* ortholog into the orthogroup topology, instead of grouping these together in one species group for which only one ortholog was required (see supporting Figure S1). We inserted a new primary fossil calibration, based on the aforementioned fossil evidence, to calibrate the divergence between the homeologous pair and the *Prunus* ortholog. The divergence between *Pyrus* and *Prunus* has been estimated at ~73 mya (Lo and Donoghue 2012). We therefore specified a lognormal calibration prior with μ = 3.4405, σ = 0.5, and a minimum bound of 48.7 mya. A run without data (Drummond et al. 2006; Heled and Drummond 2012) indicated however that the marginal prior calibration distribution did not correspond to its specified calibration density, and we had to increase μ to a value of 3.7851 so that the marginal prior calibration distribution was located at 73 mya. Apart from this new calibration, calibrations E2 and R1 were also implemented (see supporting Figure S2), while calibration E1 had to be removed because it overlapped temporally on a nested clade with the new calibration (Warnock et al. 2012). In total, 1,000 orthogroups were constructed and dated, of which 978 were accepted afterwards (ESS >200 for all statistics, see Material and methods). The resulting absolute age distribution is presented in supporting Figure S6.

A new WGD age estimate of 30.1 mya was obtained for the *P. bretschneideri* WGD. This constitutes an increase of more than 10 million years with respect to our original WGD age estimate of 19.85 mya, but is still 18.6 million years short of the previously described minimum fossil bound of 48.7 mya. This confirms that incomplete correction of rate deceleration led to an underestimation of the *P. bretschneideri* WGD, and that breaking up long branches in orthogroup phylogenies through better taxon sampling, in combination with new rate-correcting fossil calibrations, will help to correct for drastic rate shifts when more full plant genome sequences become available in the future.



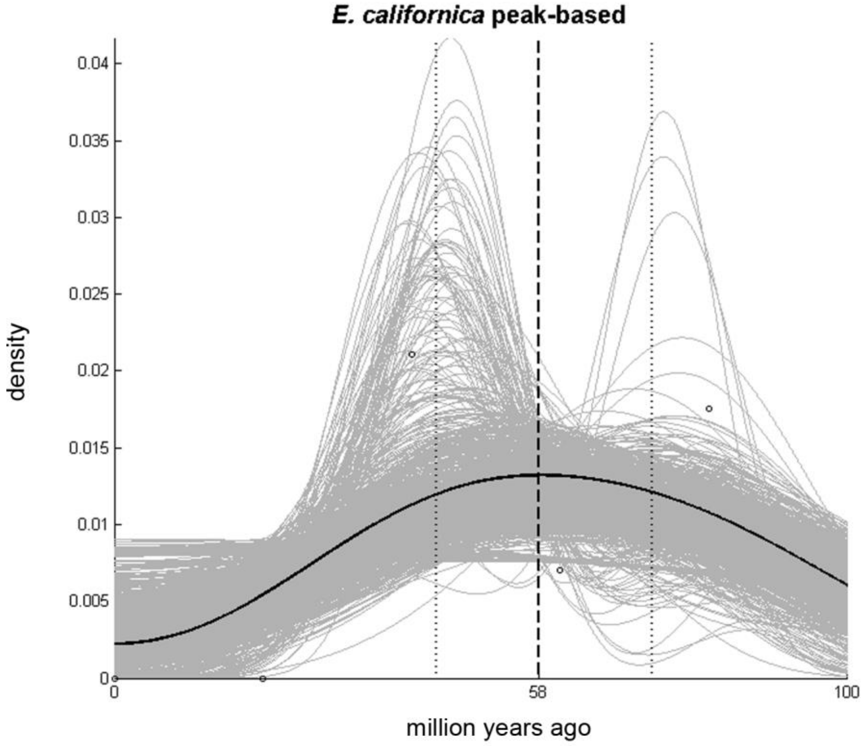
Supporting Figure S6 - Absolute age distribution of the dated anchors for *P. bretschneideri* with improved taxon sampling and a new primary fossil calibration closer to the homeologous pair. The black solid line represents the kernel density estimate of the dated homeologs, while the vertical dashed line represents its peak used as WGD age estimate. The grey lines represent the density estimates for the 1,000 bootstrap replicates, while the vertical dotted lines represent the corresponding 90% confidence intervals on the WGD age estimate. The original raw distribution of dated homeologs is also indicated by open dots. The WGD age is estimated at 30.15 mya, with a lower and upper 90% confidence interval boundary of 29.23 and 31.14 mya, respectively.

WGD age estimates from literature

The following WGD age estimates, corresponding to the black bars of Figure 3, were taken from literature. The *N. nucifera* WGD was estimated at 65 mya (Ming et al. 2013). The oldest WGD in *M. acuminata* was estimated at 96 mya (D'Hont et al. 2012). The core eudicot shared *gamma* hexaploidy was estimated somewhere between 117 and 133 mya (Jiao et al. 2012; Vekemans et al. 2012). The oldest shared WGD in the grasses, also referred to as *rho*, was estimated at 130 mya based on the median synonymous substitution rate, which was however close to saturation and therefore should be interpreted with caution (Paterson et al. 2004). Considering that both the Zingiberales and Arecales, which do not share this event, most likely branched off somewhere around 120 mya (Janssen and Bremer 2004; Kress 2006; Couvreur et al. 2011; Baker and Couvreur 2013), we placed this WGD right after the origin of the grasses, but its exact age remains unknown. The angiosperm- and seed plant-wide WGDs were estimated at 192 and 319 mya, respectively (Jiao et al. 2011).

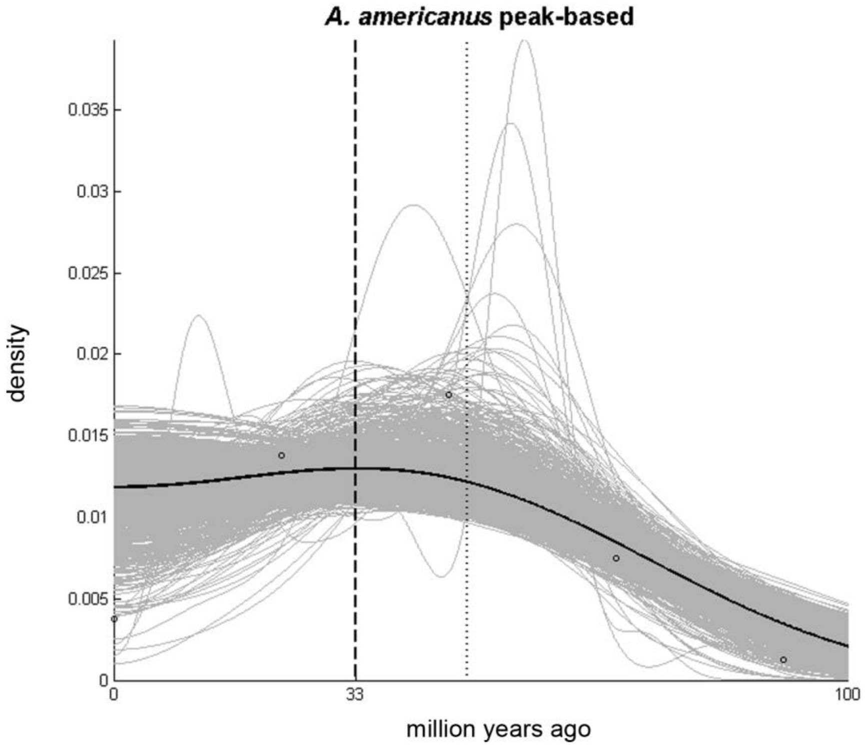
Eschscholzia californica* and *Acorus americanus

We originally included all transcriptome assemblies from a previous study (Fawcett et al. 2009), including *E. californica* and *A. americanus*, both of which were originally also dated close to the K-Pg boundary. However, in the current study, using the updated approaches, we were unable to obtain unambiguous WGD age estimates for both species. In the case of *E. californica*, only 15 orthogroups based on peak-based duplicates could be constructed, of which 14 were accepted (ESS >200 for all statistics, see Material and methods). Their resulting absolute age distribution is presented in supporting Figure S7. The mode of the underlying kernel density estimate was located at 58.23 mya, very close to the Gaussian component located at 60.05 mya in association with the K-Pg boundary (see supplementary Figure S4). However, our KDE bootstrapping procedure demonstrated the presence of a very strong bimodal underlying shape with one peak located at ~43 mya, and another peak at ~74 mya, as evidenced both by the open dots (representing the raw data) and grey curves (representing the bootstrap samples) on supporting Figure S7. Inclusion of this WGD in our results, represented by a very wide bar on Figure 3, would however be misleading, as its estimate of 58.23 mya would increase statistical support for the clustering of WGDs with the K-Pg boundary, whereas evaluation of its absolute age distribution demonstrates that this estimate clearly cannot be trusted. This is not necessarily due to the low number of dated homeologs, as other absolute age distributions, such as for instance the absolute age distribution of *L. japonicus* based on anchors (see Figure 2, panel C), are based on a similar small number of dated homeologs. The latter nevertheless shows strong support for a unimodal distribution, which is reinforced by its peak-based absolute age distribution that is based on a much larger number of homeologous pairs and displays a similar trend. The example of *E. californica* thus demonstrates the strengths of our bootstrapping KDE approach by allowing the exclusion of dubious WGD age estimates. In contrast, fitting a standard parametric distribution, such as a gamma or normal distribution, would forcibly fit a unimodal shape to a bimodal distribution and lead to the inclusion of erroneous data for statistical analysis of clustering.



Supporting Figure S7 - Absolute age distribution of the dated peak-based duplicates for *E. californica*. The black solid line represents the kernel density estimate of the dated homeologs, while the vertical dashed line represents its peak used as WGD age estimate. The grey lines represent the density estimates for the 1,000 bootstrap replicates, while the vertical dotted lines represent the corresponding 90% confidence intervals on the WGD age estimate. The original raw distribution of dated homeologs is also indicated by open dots. The WGD age is estimated at 58.28 mya, with a lower and upper 90% confidence interval boundary of 42.28 and 74.10 mya, respectively.

In the case of *A. americanus*, 35 orthogroups based on peak-based duplicates could be constructed, which were all accepted (ESS >200 for all statistics, see Material and methods). Their resulting absolute age distribution is presented in supporting Figure S8. The mode of the underlying kernel density estimate was located at 33.26 mya, very far from the K-Pg boundary. However, our bootstrapping KDE procedure demonstrated a very uninformative shape. In particular, the kernel density estimate is very wide with only a poorly supported peak that barely protrudes above the background, as also indicated by the bootstrap replicates themselves that reveal a mostly flat curve surface. In fact, the bootstrap replicates indicate the presence of a very diffuse peak centered on the 90% confidence interval upper boundary that is masked by the flat left flank, but still evident by the decreasing right flank. A trustworthy estimate for the *A. americanus* WGD, similarly to the *E. californica* WGD, hence remains elusive.



Supporting Figure S8 - Absolute age distribution of the dated peak-based duplicates for *A. americanus*. The black solid line represents the kernel density estimate of the dated homeologs, while the vertical dashed line represents its peak used as WGD age estimate. The grey lines represent the density estimates for the 1,000 bootstrap replicates, while the vertical dotted lines represent the corresponding 90% confidence intervals on the WGD age estimate. The original raw distribution of dated homeologs is also indicated by open dots. The WGD age is estimated at 33.26 mya, with a lower and upper 90% confidence interval boundary of 0.00 and 48.17 mya, respectively.

References

Anderson CL, Bremer K, Friis EM. 2005. Dating phylogenetically basal eudicots using rbcL sequences and multiple fossil reference points. *Am J Bot* **92**(10): 1737-1748.

Ayres DL, Darling A, Zwickl DJ, Beerli P, Holder MT, Lewis PO, Huelsenbeck JP, Ronquist F, Swofford DL, Cummings MP et al. 2012. BEAGLE: An Application Programming Interface and High-Performance Computing Library for Statistical Phylogenetics. *Syst Biol* **61**(1): 170-173.

Baker WJ, Couvreur TLP. 2013. Global biogeography and diversification of palms sheds light on the evolution of tropical lineages. I. Historical biogeography. *J Biogeogr* **40**(2): 274-285.

Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S. 2010. Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. *Proc Natl Acad Sci U S A* **107**(43): 18724-18728.

Bell CD, Soltis DE, Soltis PS. 2005. The age of the angiosperms: a molecular timescale without a clock. *Evolution* **59**(6): 1245-1258.

-. 2010. The age and diversification of the angiosperms re-revisited. *Am J Bot* **97**(8): 1296-1303.

Benton MJ, Donoghue PC. 2007. Paleontological evidence to date the tree of life. *Mol Biol Evol* **24**(1): 26-53.

Berlin S, Lagercrantz U, von Arnold S, Ost T, Ronnberg-Wastljung AC. 2010. High-density linkage mapping and evolution of paralogs and orthologs in Salix and Populus. *BMC Genomics* **11**.

Berry EW. 1914. The Upper Cretaceous and Eocene floras of South Carolina, Georgia. *US Geological Survey, Professional Paper* **84**: 1-200.

Boucher LD, Manchester SR, Judd WS. 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. *Am J Bot* **90**(9): 1389-1399.

Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**(6930): 433-438.

Bremer B, Bremer K, Chase MW, Fay MF, Reveal JL, Soltis DE, Soltis PS, Stevens PF, Anderberg AA, Moore MJ et al. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. *Bot J Linn Soc* **161**(2): 105-121.

Bremer K, Friis EM, Bremer B. 2004. Molecular phylogenetic dating of asterid flowering plants shows early Cretaceous diversification. *Syst Biol* **53**(3): 496-505.

Clarke JT, Warnock RC, Donoghue PC. 2011. Establishing a time-scale for plant evolution. *New Phytol* **192**(1): 266-301.

Coiffard C, Gomez B, Daviero-Gomez V, Dilcher DL. 2012. Rise to dominance of angiosperm pioneers in European Cretaceous environments. *Proc Natl Acad Sci U S A* **109**(51): 20955-20959.

Couvreur TL, Franzke A, Al-Shehbaz IA, Bakker FT, Koch MA, Mummenhoff K. 2010. Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae). *Mol Biol Evol* **27**(1): 55-71.

Couvreur TLP, Forest F, Baker WJ. 2011. Origin and global diversification patterns of tropical rain forests: inferences from a complete genus-level phylogeny of palms. *BMC Biol* **9**.

Crepet W, Nixon K. 1998. Fossil Clusiaceae from the late Cretaceous (Turonian) of New Jersey and implications regarding the history of bee pollination. *Am J Bot* **85**(8): 1122.

D'Hont A, Denoeud F, Aury JM, Baurens FC, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M et al. 2012. The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature* **488**(7410): 213-+.

Davis CC, Webb CO, Wurdack KJ, Jaramillo CA, Donoghue MJ. 2005. Explosive radiation of malpighiales supports a mid-Cretaceous origin of modern tropical rain forests. *Am Nat* **165**(3): E36-E65.

De Craene LPR, Haston E. 2006. The systematic relationships of glucosinolate-producing plants and related families: a cladistic investigation based on morphological and molecular characters. *Bot J Linn Soc* **151**(4): 453-494.

Doyle JA. 2005. Early evolution of angiosperm pollen as inferred from molecular and morphological phylogenetic analyses. *Grana* **44**(4): 227-251.

Drummond AJ, Ho SY, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol* **4**(5): e88.

Edwards D, Feehan J. 1980. Records of Cooksonia-Type Sporangia from Late Wenlock Strata in Ireland. *Nature* **287**(5777): 41-42.

Fawcett JA, Maere S, Van de Peer Y. 2009. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc Natl Acad Sci U S A* **106**(14): 5737-5742.

Forest F. 2009. Calibrating the Tree of Life: fossils, molecules and evolutionary timescales. *Ann Bot* **104**(5): 789-794.

Friis EM. 1988. *Spirematospermum chandlerae* sp. nov., an extinct species of Zingiberaceae from the North American Cretaceous. *Tert Res* **9**: 7-12.

Gandolfo MA, Nixon KC, Crepet WL. 1998. A new fossil flower from the Turonian of New Jersey: Dressiantha bicarpellata gen. et sp. nov. (Capparales). *Am J Bot* **85**(7): 964-974.

Hedges SB, Kumar S. 2004. Precision of molecular time estimates. *Trends Genet* **20**(5): 242-247.

Heled J, Drummond AJ. 2012. Calibrated Tree Priors for Relaxed Phylogenetics and Divergence Time Estimation. *Syst Biol* **61**(1): 138-149.

Herendeen PS, Crane PR. 1992. *Advances in Legume Systematics: Part 4 The Fossil Record*. Royal Botanical Gardens, Kew, UK.

Ho SY, Phillips MJ. 2009. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. *Syst Biol* **58**(3): 367-380.

Hug LA, Roger AJ. 2007. The impact of fossils and taxon sampling on ancient molecular dating analyses. *Mol Biol Evol* **24**(8): 1889-1897.

Hugall AF, Foster R, Lee MSY. 2007. Calibration choice, rate smoothing, and the pattern of tetrapod diversification according to the long nuclear gene RAG-1. *Syst Biol* **56**(4): 543-563.

Inoue J, Donoghue PC, Yang Z. 2010. The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. *Syst Biol* **59**(1): 74-89.

Jansen RK, Cai Z, Raubeson LA, Daniell H, Depamphilis CW, Leebens-Mack J, Muller KF, Guisinger-Bellian M, Haberle RC, Hansen AK et al. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc Natl Acad Sci U S A* **104**(49): 19369-19374.

Janssen T, Bremer K. 2004. The age of major monocot groups inferred from 800+rbcL sequences. *Bot J Linn Soc* **146**(4): 385-398.

Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, McNeal J, Rolf M, Ruzicka DR, Wafula E, Wickett NJ et al. 2012. A genome triplication associated with early diversification of the core eudicots. *Genome Biol* **13**(1): R3.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS et al. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**(7345): 97-100.

Kress WJ. 2006. The evolutionary and biogeographic origin and diversification of the tropical monocot order Zingiberales. *Aliso* **22**: 621-632.

Leitch IJ, Greilhuber J, Dolezel J, Wendel J. 2013. *Plant Genome Diversity Volume 2*. Springer-Verlag.

Lo EY, Donoghue MJ. 2012. Expanded phylogenetic and dating analyses of the apples and their relatives (Pyreae, Rosaceae). *Mol Phylogenet Evol* **63**(2): 230-243.

Magallon S. 2010. Using Fossils to Break Long Branches in Molecular Dating: A Comparison of Relaxed Clocks Applied to the Origin of Angiosperms. *Syst Biol* **59**(4): 384-399.

Magallon S, Castillo A. 2009. Angiosperm Diversification through Time. *Am J Bot* **96**(1): 349-365.

Magallon S, Hilu KW, Quandt D. 2013. Land Plant Evolutionary Timeline: Gene Effects Are Secondary to Fossil Constraints in Relaxed Clock Estimation of Age and Substitution Rates. *Am J Bot* **100**(3): 556-573.

Ming R, Vanburen R, Liu Y, Yang M, Han Y, Li LT, Zhang Q, Kim MJ, Schatz MC, Campbell M et al. 2013. Genome of the long-living sacred lotus (Nelumbo nucifera Gaertn.). *Genome Biol* **14**(5): R41.

Mulcahy DG, Noonan BP, Moss T, Townsend TM, Reeder TW, Sites JW, Jr., Wiens JJ. 2012. Estimating divergence dates and evaluating dating methods using phylogenomic and mitochondrial data in squamate reptiles. *Mol Phylogenet Evol* **65**(3): 974-991.

Paterson AH, Bowers JE, Chapman BA. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci U S A* **101**(26): 9903-9908.

Pigg KB, Manchester SR, Devore ML. 2008. Fruits of Icacinaceae (tribe Iodeae) from the Late Paleocene of western North America. *Am J Bot* **95**(7): 824-832.

Pond SLK, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**(5): 676-679.

Sauquet H, Ho SYW, Gandolfo MA, Jordan GJ, Wilf P, Cantrill DJ, Bayly MJ, Bromham L, Brown GK, Carpenter RJ et al. 2012. Testing the Impact of Calibration on Molecular Divergence Times Using a Fossil-Rich Group: The Case of Nothofagus (Fagales). *Syst Biol* **61**(2): 289-313.

Smith SA, Beaulieu JM, Donoghue MJ. 2010. An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants. *Proc Natl Acad Sci U S A* **107**(13): 5897-5902.

Smith SA, Beaulieu JM, Stamatakis A, Donoghue MJ. 2011. Understanding Angiosperm Diversification Using Small and Large Phylogenetic Trees. *Am J Bot* **98**(3): 404-414.

Smith SA, Donoghue MJ. 2008. Rates of molecular evolution are linked to life history in flowering plants. *Science* **322**(5898): 86-89.

Soltis DE, Smith SA, Cellinese N, Wurdack KJ, Tank DC, Brockington SF, Refulio-Rodriguez NF, Walker JB, Moore MJ, Carlsward BS et al. 2011. Angiosperm Phylogeny: 17 Genes, 640 Taxa. *Am J Bot* **98**(4): 704-730.

Suchard MA, Rambaut A. 2009. Many-core algorithms for statistical phylogenetics. *Bioinformatics* **25**(11): 1370-1376.

Sun G, Dilcher DL, Wang HS, Chen ZD. 2011. A eudicot from the Early Cretaceous of China. *Nature* **471**(7340): 625-628.

Tuskan GA DiFazio S Jansson S Bohlmann J Grigoriev I Hellsten U Putnam N Ralph S Rombauts S Salamov A et al. 2006. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* **313**(5793): 1596-1604.

Vekemans D, Proost S, Vanneste K, Coenen H, Viaene T, Ruelens P, Maere S, Van de Peer Y, Geuten K. 2012. Gamma paleohexaploidy in the stem lineage of core eudicots: significance for MADS-box gene and species diversification. *Mol Biol Evol* **29**(12): 3793-3806.

Wang HC, Moore MJ, Soltis PS, Bell CD, Brockington SF, Alexandre R, Davis CC, Latvis M, Manchester SR, Soltis DE. 2009. Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proc Natl Acad Sci U S A* **106**(10): 3853-3858.

Warnock RC, Yang Z, Donoghue PC. 2012. Exploring uncertainty in the calibration of the molecular clock. *Biol Lett* **8**(1): 156-159.

Wehr WC, Hopkins DQ. 1994. The Eocene Orchards and Gardens of Republic, Washington. *Washington Geology* **22**(3): 27-34.

Wheeler EF, Lee M, Matten LC. 1987. Dicotyledonous Woods from the Upper Cretaceous of Southern Illinois. *Bot J Linn Soc* **95**(2): 77-100.

Wheeler RJ, Lecroy SR, Whitlock CH, Purgold GC, Swanson JS. 1994. Surface Characteristics for the Alkali Flats and Dunes Regions at White-Sands-Missile-Range, New-Mexico. *RSEnv* **48**(2): 181-190.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* **18**(5): 691-699.

Wikstrom N, Savolainen V, Chase MW. 2001. Evolution of the angiosperms: calibrating the family tree. *Proceedings of the Royal Society B-Biological Sciences* **268**(1482): 2211-2220.

Xi Z, Ruhfel BR, Schaefer H, Amorim AM, Sugumaran M, Wurdack KJ, Endress PK, Matthews ML, Stevens PF, Mathews S et al. 2012. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. *Proc Natl Acad Sci U S A* **109**(43): 17519-17524.

Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol* **11**(9): 367-372.

Yang Z, Rannala B. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol Biol Evol* **23**(1): 212-226.