The Rise of Polyploids During Environmental Catastrophes

Hengchi Chen[1,2†], Fabricio Almeida-Silva[1,2], Garben Logghe[3,4], Dries Bonte[3], Yves Van de Peer[1,2,5,6†]

[1] Department of Plant Biotechnology and Bioinformatics, Ghent University, 9052 Ghent, Belgium

[2] VIB Center for Plant Systems Biology, VIB, 9052 Ghent, Belgium

[3] Department of Biology, Terrestrial Ecology Unit, Ghent University, 9000 Ghent, Belgium

[4] Research Institute for Nature and Forest (INBO), Herman Teirlinckgebouw, Havenlaan 88 Bus 73, 1000 Brussels, Belgium

[5] Centre for Microbial Ecology and Genomics, Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria 0028, South Africa.

[6] College of Horticulture, Academy for Advanced Interdisciplinary Studies, Nanjing Agricultural University, Nanjing 210095, China.

†Correspondence should be addressed to H.C. (hengchi.chen@psb.vib-ugent.be) and/or Y.V.d.P. (yves.vandepeer@psb.ugent.be).

ABSTRACT

Polyploidy, or whole-genome duplication (WGD), both a strong evolutionary and ecological force and a potential pitfall, occurs extensively across the tree of life, particularly in angiosperms. While polyploid organisms are prevalent, ancient polyploidy (paleopolyploidy) is decisively rare. Numerous studies have revealed adaptive changes at the genomic, cellular, and physiological levels that confer evolutionary advantages to polyploids, but the factors contributing to their long-term success remain poorly understood. Here, we assembled a large angiosperm genome dataset encompassing 470 different angiosperm species. We reconstructed a highly concordant evolutionary timescale of angiosperms and conducted a large-scale study to identify and date WGDs across this set of species. We unveiled 132 independent ancient WGD events that do not seem randomly distributed in time, but clustered around periods of great environmental upheaval. We observed that a majority of paleopolyploidizations occurred between 50 and 75 mya, coinciding with the Paleocene–Eocene Thermal Maximum and the Cretaceous–Paleogene (K–Pg) boundary. A more recent wave of ancient WGDs can be observed between 25 and 40 mya, coinciding with periods of glaciation expansion, rapid shifts in sea surface temperatures, and a marked decrease in global mean surface temperature. Integrating different lines of evidence, we argue that polyploids have an increased chance of survival during times of great environmental turmoil.

Keywords: WGD, Polyploid, Paleopolyploidization, Angiosperm, Environmental turmoil

Introduction

The realization that the evolutionary trajectory of most, if not all, flowering plants has been profoundly influenced by multiple rounds of polyploidization - the result of whole genome duplication (WGD) - has been a cornerstone of early plant genomics research (Wendel 2015). These "remarkable cycles of polyploidy" had been predicted by Haldane (1933) in the early days of polyploidy research, but were later often considered as “evolutionary noise” (Wagner Jr. 1970) or even as evolutionary “dead ends” (Stebbins 1950; Stebbins 1971). Nevertheless, such skepticism seemed partially justified. Although extant polyploid organisms significantly enrich botanical diversity (Rice et al. 2019), and polyploidization is acknowledged as a major driver of plant speciation (Landis et al. 2018), the long-term persistence of polyploids appears to be a rare and infrequent phenomenon (Van de Peer et al. 2017). This infrequency suggests that substantial numbers of polyploid lineages may falter, failing to endure through evolutionary timescales. The presence of numerous instances of recurring polyploidy and recently originated polyploids, against a backdrop of scarce evidence for ancient polyploidy presents an interesting evolutionary paradox (Van de Peer et al. 2017; Carretero-Paulet and Van de Peer 2020).

Whole genome duplication carries transformative implications as a significant mutagenic occurrence. WGD can impose immediate fitness detriments, including complications in cell cycling, stunted growth rates, and reproductive incompatibility with ancestral (diploid) forms. Simultaneously, WGDs elicit an instantaneous increase in cell dimensions, various genomic and epigenomic alterations, along with an array of downstream effects that directly influence fitness (Mortier et al. 2024). Both disadvantages and potential benefits of WGDs have been extensively documented (Comai 2005; Bomblies 2020; Van de Peer et al. 2021). Although the long-term consequences of WGD are relatively easy to understand (i.e., extra genetic material created by WGD can lead to novel gene functions and more robust regulatory networks (Almeida-Silva and Van de Peer 2023), the immediate or short-term advantages following WGD are less apparent. Several recent studies have meticulously described the immediate effects of polyploidy (Baduel et al. 2018; Doyle and Coate 2019; Bomblies 2020). Among the most consistent outcomes of WGD is amplified cell size, coupled with physiological modifications that can lead to increased resilience in stress conditions. For instance, neo-autotetraploid cultivars of *Arabidopsis thaliana* display increased tolerance to salt and drought stresses

compared to their diploid progenitors (Chao et al. 2013; Pozo and Ramirez-Parra 2014). Similarly, tetraploid cultivars of rice (*Oryza sativa*) and citrange (*Citrus sinensis × Poncirus trifoliata*) exhibit increased tolerance to salt and drought stress due to WGD's modulation of genes involved in stress response and phytohormone signaling pathways (Yang et al. 2014; Ruiz et al. 2016). In watermelon (*Citrullus lanatus*), plants grafted onto tetraploid rootstock display increased salt stress tolerance compared to their diploid counterparts (Zhu et al. 2018). A recent study on nascent autopolyploid spoonworts (*Cochlearia* genus) showed that the adaptive ion homeostasis might underpin the high salinity and dehydration tolerance of autopolyploid (Bray et al. 2024). Yang and co-authors (Yang et al. 2024) showed that polyploid kiwifruits (*Actinidia chinensis)* bear a larger number of negative turgor loss points and larger cell elastic modulus, which could mitigate damages from freezing-induced cell dehydration, underpinning their superior tolerance to cold stress over diploids. Coate and co-authors (Coate et al. 2013) showed that, compared to the diploid progenitors, polyploid soybean (*Glycine dolichocarpa*) had higher quantum yield of photosystem II in the absence of CO_2, suggesting a greater capacity of non-photosynthetic electron transport, likely contributed by the increase in xanthophyll pigments coupled with enhanced electron flow. As these examples of physiological and cellular responses to stress are frequently observed in cultivated polyploids, they have led to the hypothesis that polyploidisation generally aligns with an increased resilience to stress (Comai 2005; Van de Peer et al. 2021).

An instinctive corollary of this hypothesis is that the occurrence, and more importantly the persistence of WGD events during angiosperm evolution might not be coincidental but likely correlated with periods of environmental upheaval. For instance, Fawcett and co-authors (Fawcett et al. 2009) and Vanneste and co-authors (Vanneste et al. 2014) suggested a clustering of WGDs across several independent plant lineages around the Cretaceous–Paleogene (K–Pg) boundary, approximately 66 million years ago (mya) (Renne et al. 2013). Additional 'waves' of WGDs were suggested around the Paleocene–Eocene transition, around 56-54 Mya (Cai et al. 2019), or during the last glaciations (Novikova et al. 2018). The potential link between the ‘fixation’ of WGDs and periods of environmental disturbance or periods of extinction is fascinating, yet the underlying mechanisms remain elusive. Also, previous correlations between WGDs and periods of environmental turmoil were based on very limited numbers of genomes, unbalanced – from an evolutionary ‘sampling’

perspective - datasets, inflated secondary calibrations, and/or oversimplified molecular dating methods that failed to properly accommodate fossil and branch length uncertainty, rendering the validity of such correlations debatable.

Here, we conducted an extensive analysis of a dataset encompassing 470 angiosperm genomes to identify and date 132 independent ancient polyploidization events (i.e., excluding species that are currently still polyploid) using the Bayesian relaxed molecular clock model and 44 fossil calibrations for angiosperms that were highly curated following best practices (Parham et al. 2011). To investigate patterns of paleopolyploidization across lineages in conjunction with paleoclimate dynamics, we first meticulously assembled a robust phylogenetic framework and evolutionary timeline for the angiosperms. Our findings show that paleopolyploidization occurs across each major clade of angiosperms, including the ANA grade, Ceratophyllales, Chloranthales, eudicots, monocots, and magnoliids, and spans an evolutionary timeline from the Lower Cretaceous to the Neogene. Although the frequency of paleopolyploidization across different lineages and time varies dramatically, polyploidization seems to correlate with periods of environmental turmoil or mass extinction. To facilitate data reuse, we built a web application (available at https://bioinformatics.psb.ugent.be/AngioWGD/, Supplementary Note 1) where users can explore WGDs across the angiosperm phylogeny, download inferred dates for each WGD, and access original genomic data included in this work.

Results

Genome assessment based on BUSCO and OMArk

Because the quality of genome assemblies is essential for reliable downstream comparative analyses, we assessed all genome assemblies, i.e., 470 angiosperm genomes and the genome of the outgroup species *Cycas panzhihuaensis*. Complete BUSCO (Simão et al. 2015) scores are generally high (median, Q1, and Q3 = 92%, 85.15%, and 96.45%, respectively), and so are OMArk (Nevers et al. 2024) scores (median, Q1, and Q3 = 95.45%, 90.41%, and 96.52%, respectively) (Supplementary Note 2, Figure S1 and S2). The percentages of missing BUSCOs are low (mean and median = 7.93% and 4.10%, respectively), and so are percentages of OMArk missing families (mean and median = 7.85% and 5.55%, respectively). Taken together, these statistics indicate the high quality of our datasets.

Reconstruction of an evolutionary timescale of angiosperms

Using 32 gene tree datasets, each with 1614 reference single-copy gene families obtained from the *embryophyta_odb10* dataset (Kriventseva et al. 2018), we inferred species trees and assessed their overall accuracy and consistency, and ultimately reconstructed a 'consensus tree' (see Materials and Methods for details). The resulting consensus tree (hereafter referred to as 'main tree'), covering 45 orders, 131 families, and all major clades of Mesangiospermae and ANA grade, displays high dataset- and branch-wise concordance (Supplementary Note 3, Figure S2 and S3-7). High congruence was achieved across alternative Multiple Sequence Alignment (MSA) datasets, while two discordances pertaining to Chloranthales and Crossosomatales were further tested, concluding that the phylogenetic relationships that Chloranthales is sister to magnoliids, and Crossosomatales being sister to Fabales and Myrtales, were most supported (Supplementary Note 3, Figure S3 and S5). We also show that our phylogenetic inference is robust to random sampling bias of gene trees (Supplementary Note 3, Figure S7).

In comparison to a recent large-scale angiosperm phylogeny based on 353 nuclear orthologues (Zuntini et al. 2024), our main tree shows concordance in most of the ordinal-level relationships (Supplementary Note 3, Figure S8), and in particular the results of our statistic topological tests of Chloranthales and Crossosomatales are supported. However, the phylogenetic positions of magnoliids, Ceratophyllales, and Caryophyllales remain uncertain. Although the position of magnoliids has long been disputed, several recent large-scale phylogenetic studies (Leebens-Mack et al. 2019; Yang et al. 2020; Hu et al. 2023) are in agreement with our main tree, supporting monocots being sister to magnoliids and eudicots. Phylogenetic inference from 30 out of 32 datasets supports our main tree, while the only two datasets that did not (the codon-level and TAPER-masked datasets at the branch filtering cut-off of 70%) also fail to recover monophyly of several orders and denoted numerous branches as highly uncertain (Supplementary Note 3), probably due to an insufficient number of remaining branches after stringent filtering. The phylogenetic location of Ceratophyllales in Zuntini et al. (2024) is highly uncertain (local posterior probability (LPP) below 0.5), and it was claimed by the authors themselves as inconclusive, while our main tree fully supports Ceratophyllales as sister to the eudicots (LPP = 1). Our tree also supports Caryophyllales as sister to the Lamiids and Asterids (LPP = 0.83) in 31 out of 32 datasets (except for the codon-level dataset at the branch filtering cut-off of 70%), in

line with previous studies (APG 2016; Leebens-Mack et al. 2019; Yang et al. 2020; Hu et al. 2023), but also disagreeing with Zuntini and co-authors (Zuntini et al. 2024), who showed high gene tree conflict at the node pertaining to Caryophyllales.

Using the topology of our main tree, we reconstructed a time-calibrated phylogenetic tree based on our angiosperm fossil calibration dataset (Supplementary Note 4). Given that only 1 Single-Copy Orthogroup (SOG, see Materials and Methods for details) was present in all 470 angiosperm species, we slightly relaxed the species coverage to 99%, resulting in 12 SOGs encompassing 466 species. The uncorrelated lognormal (UCLN) relaxed molecular clock model was applied with soft fossil calibration bounds to consider both rate uncertainty and time uncertainty (Supplementary Note 4), resulting in an evolutionary timetree of 466 angiosperm species (Figure 1). The stem order and family ages in our timetree (Supplementary Table 3.2 and 3.3) are highly correlated with those obtained by Ramírez-Barahona and co-authors (Ramírez-Barahona et al. 2020) (Pearson's correlation coefficients = 0.8157 and 0.7542; P <0.0001; Supplementary Note 5, Figure S9), suggesting comparable divergence time estimates.

Dating WGDs reveals temporal and phylogenetic dynamics of paleopolyploidization

We identified 132 independent ancient WGD events including both previously reported and novel ones, and conducted standardized WGD dating analysis on each one of them (Figure 1; Supplementary Note 6). Eudicots display the largest number of ancient WGD events (*N* = 95), followed by monocots (*N* = *25*), magnoliids (*N* = 5), ANA grade (*N* = 4), Ceratophyllales (*N* = 2), and Chloranthales (*N* = 1). WGD frequency among different orders and families varies dramatically. For instance, Lamiales harbors 10 independent ancient WGD events, which is greater than the 95th percentile of all orders with ancient WGDs (Figure S1b; Table 1). Likewise, the plant families Poaceae, Fabaceae, and Brassicaceae contain 8, 5 and 5 independent ancient WGD events, respectively, which is greater than the 95th percentile of all families with ancient WGDs (Figure S1c; Table 1).

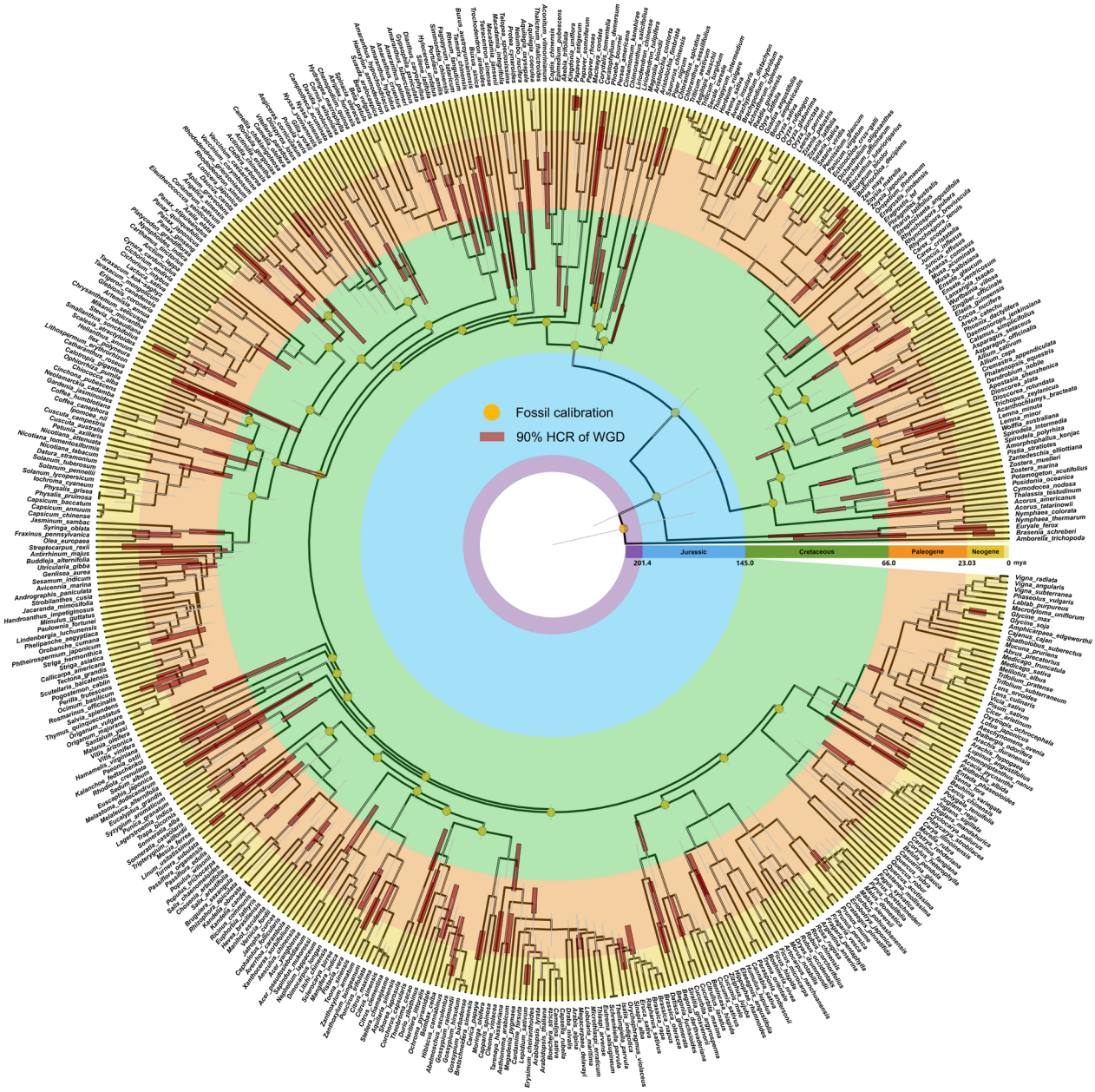

Figure 1. Reconstructed evolutionary timetree of 466 angiosperms. The 95% Highest Posterior Density (HPD) of divergence time estimates are represented as light gray lines on the internodes, while the 90% Highest Convergence Region (HCR) of WGD date estimates are represented as red rectangles. Applied fossil calibrations are represented as yellow dots. Geological epochs are marked in distinct colors.

The most ancient WGD event dated in monocots is the well-known τ event (Jiao et al. 2014), shared by all non-Alismatales monocots (excluding Acorales), dated at 121.31 mya (90% HCR = 112.87-129.35 mya) (Figure S1d; Supplementary Note 6). In eudicots, the oldest dated WGD events are the 'RANU' event (WGD nomenclature in Supplementary Note 6), shared by Ranunculales (Figure S1e), dated at 113.92 mya (90% HCR = 108.24-122.32 mya), and the γ event (Jiao et al. 2012), shared by core

eudicots (Figure S1d), dated at 107.58 mya (90% HCR = 94.96-119.90 mya). In magnoliids, the oldest WGD event is the 'LAUR β' event, shared by Laurales and Magnoliales (Figure S1d), dated at 109.98 mya (90% HCR = 98.56-120.51 mya). There are 10 independent ancient WGD events shared by entire plant orders, and 10 ancient WGD events shared by 'partial' orders (i.e., shared by multiple families, but not the entire order; Figure S1e and S1f). We also identified 31 independent ancient WGD events shared by entire plant families (Figure S1g), and 20 ancient WGD events shared by 'partial' families (i.e., shared by multiple genera, but not the entire family; Figure S1h).

The remaining 58 independent paleopolyploidy events are shared by entire genera or unique to specific species (Figure S1i). At first sight, the occurrence of ancient WGD events seems to present a seemingly random pattern spanning the Lower Cretaceous to the Neogene (Figure 1). However, by applying a Bayesian change point analysis (see Materials and Methods for details), we detected the critical time points at which the mean cumulative rate of WGD occurrences significantly shifted. The number of ancient WGD events per species, suggestive of the phylogenetic propensity of WGD incidence, shows an oscillating upward trend from approximately 124 mya until 50 mya and then declines from 50 mya until now (Figure 2a), while the Maximum A Posteriori (MAP) of shift time points is estimated at 74.1 mya (Figure 2b), suggesting an accelerated rate of WGD occurrence up to 74.1 mya. Altogether, these results suggest a heterogeneous incidence of WGDs across times and phylogeny.

Model-based analyses unveil a correlation between WGDs and environmental turmoil.

Given the above observation, we further conducted more elaborate model-based analyses to explore whether WGDs might have clustered around specific times. To this end, we first conducted Gaussian mixture modeling (GMM) on the observed WGDs using log-transformed (natural log) WGD dates with 1, 2, 3 or 4 components as previously done by Vanneste and co-authors (Vanneste et al. 2014) and Cai and co-authors (Cai et al. 2019). The mixture model with 2 components is the model fitting the data best according to the Akaike Information Criterion (AIC), while the mixture model with 3 components is the model fitting the data best according to the silhouette score (see Materials and Methods for details) (Figure 3a-d). The AIC-best mixture model supports 2 components with modes 20.56 and 53.54 mya (Figure 3a), and

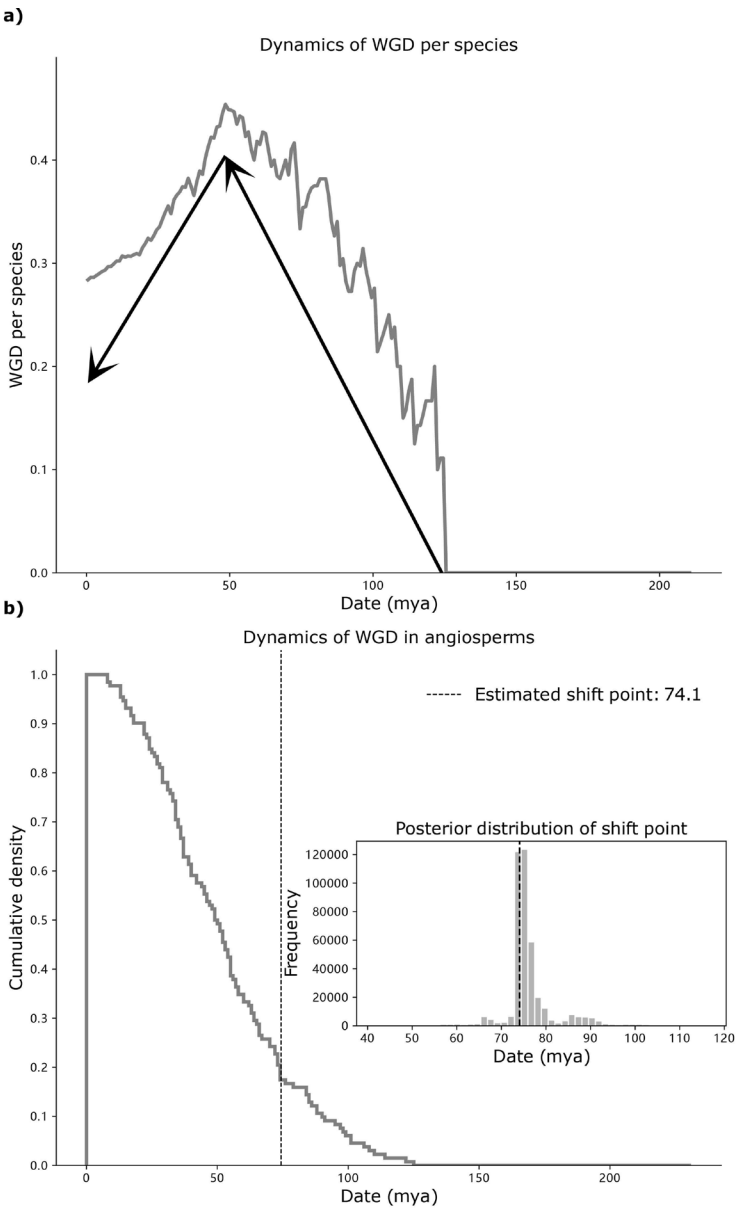

Figure 2. **Temporal dynamics of paleopolyploidizations.** Panel a) shows the temporal dynamics of the number of ancient WGD events per species. Panel b) shows the temporal cumulative dynamics of ancient WGD events in which the subplot shows the posterior distribution of shift point at which the mean cumulative rate is significantly changed from the Bayesian change point analysis. The Maximum A Posteriori (MAP) estimate of the posterior distribution is represented by the identified shift point (dashed line).

standard deviations (std) 21.09 and 24.13 mya, respectively. The silhouette score-best mixture model supports 3 components with modes 12.47, 32.14 and 62.46 mya (Figure 3b), and std 5.48, 9.81 and 21.23 mya, respectively. Although with different numbers of components, both models seem to suggest at least one relatively recent 'cluster' of WGDs, within a 0-40 mya time range, and one older cluster of WGDs within a 40-80 mya time range (using peak ± one std).

Evidently, this inference relies on the assumption of no phylogenetic signal in WGD events, which has been shown to be false (Pagel's lambda = 0.99, see Materials and Methods for details). We therefore conducted a Bayesian phylogenetic clustering – on a more limited dataset, because of computational constraints (see Materials and Methods) - using a phylogenetic multilevel model GMM to account for the phylogenetic

signal underlying the inferred WGD dates. We fitted phylogenetic multilevel regressions with one, two, and three Gaussian components, and compared model fits

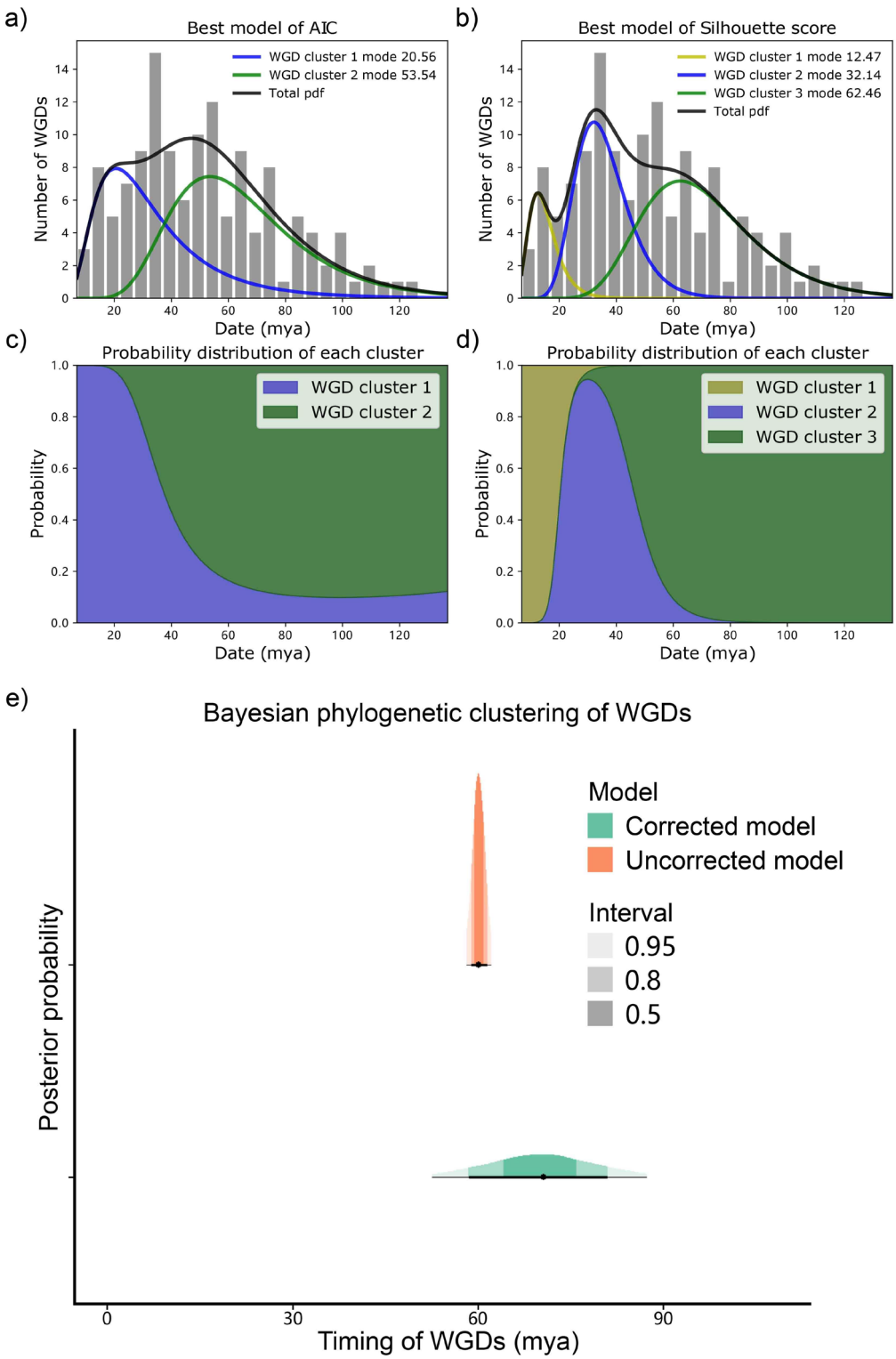

Figure 3. Clustering analysis of paleopolyploidizations. Panels a-d show the results of the log-scale Gaussian mixture modeling (GMM) analysis on observed WGD dates, in which panels a) and b) show the best mixture model based on AIC and silhouette score, respectively, with corresponding panels c) and d) showing the probability distribution of each cluster across time range. Panel e) shows the posterior probability of the mean timing of WGDs in the optimal single Gaussian model of the Bayesian phylogenetic clustering analysis before and after phylogenetic correction. See text for details.

with leave-one-out (LOO) cross-validation. The results show that, when considering

the underlying evolutionary relationships (i.e., the tree topology), the model with two Gaussian components failed to converge because of chain mixing and was therefore discarded. The model with three Gaussian components had poor Pareto k-values, of which 66.8% were within (0.7, 1), and 16.9% were among (1, Inf), suggesting low model validation because of the strong phylogenetic signal observed. Finally, the single Gaussian model showed a reasonable model validation (all Pareto k-values <0.7, with 93.2% being < 0.5) (Figure 3e). The single Gaussian model also outperformed the non-corrected (i.e., not considering the tree topology) single Gaussian model, with lOO criterion -101.0 (standard error (SE): 18.7). These results suggest that the single Gaussian model provides a reliable estimate for a clustering of WGDs, with a 95% posterior credible interval between 53.70 and 88.26 mya, and a mean of 70.62 mya (SE: 8.76) (Figure 3e). The strong phylogenetic signal in WGD together with convergence failure of the two Gaussian model is thus likely the result from the clade-specific timing of a second peak either later or earlier than the consistent peak.

We further conducted a complementary analysis and developed a simulation model which models WGD events as independent Poisson processes with clade-specific rates throughout the angiosperm tree (see Materials and Methods for details). The distribution of the inferred WGD dates differs significantly from the null distribution (two-sample Kolmogorov-Smirnov test, P <0.0001), suggesting a strong non-random distribution of WGDs across time (Figure 4d). To test whether WGDs indeed cluster in time, as already suggested by the previous analyses, we compared the median pairwise distance of the observed WGD dates to the median pairwise distances of simulated WGD dates bounded by three time intervals informed by the preceding clustering analysis, i.e., 0-40 mya, 40-80 mya, and 80-120 mya. Only in the time intervals 0-40 mya and 40-80 mya, the median pairwise distance of the observed WGD dates is significantly smaller than that of simulated WGD dates based on 10,000 iterations (Wilcoxon signed-rank test and one-sample t-test, P <0.0001; Figure 4a-c), again suggesting clustering of WGDs. We further devised a so-called ‘episodical pulled rate model’ test to explore whether the WGD occurrence rate was significantly higher than the phylogenetic background at specific time ranges (see Materials and Methods for details). Starting from 205 mya, around which the crown group of angiosperms started to diversify, we set time windows of 5 million years, approximately one std of a moderate Gaussian component. The ‘pulled WGD occurrence rate’ is

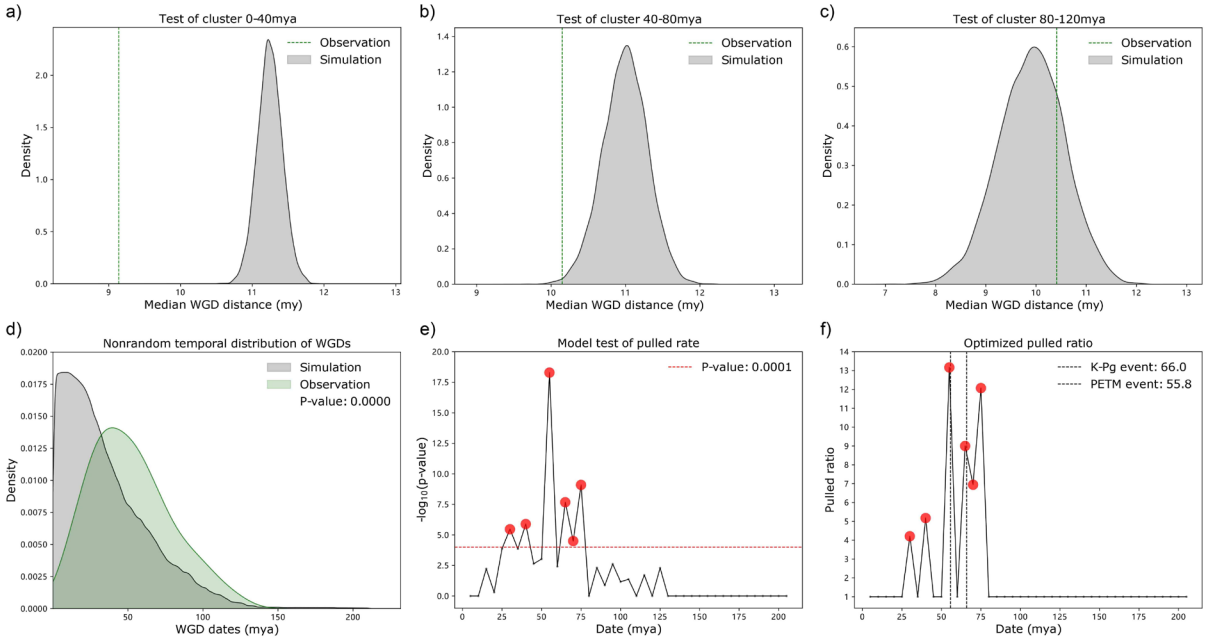

Figure 4. Model-based test of paleopolyploidizations. Panel a-c show the 'cluster test' of WGDs in the time intervals 0-40 mya, 40-80 mya and 80-120 mya, respectively. As can be observed, only in the 0-40 mya and 40-80 mya time intervals, the observed median pairwise WGD distance is significantly shorter than the simulated median pairwise WGD distance ($P < 0.0001$ of Wilcoxon signed-rank test and one sample t-test). Panel d) shows the significantly different distributions of simulated and observed WGD dates ($P < 0.0001$ of a two-sample Kolmogorov-Smirnov test). Panel e) shows the negative logarithmically transformed p-value by base 10 from the likelihood ratio test of the 'pulled WGD occurrence rate model' in comparison to the null model with a phylogenetic background rate at each time window of 5 million years, starting from 205 mya. The older bound of which is represented as the x-coordinate and the windows with p-value lower than 0.0001 are highlighted in red dots. Panel f) shows the corresponding optimized pulled ratio of each time window, with the time of the K–Pg boundary and PETM event marked as dashed lines. See text for details.

significantly favored in the time windows 70-75 mya, 65-70 mya, 60-65 mya, 50-55 mya, 35-40 mya and 25-30 mya ($P <0.0001$; Figure 4e). Among these time windows, the most remarkable 'pulled ratio' occurs at the 50-55 mya window, with an approximately 12-fold increase compared to the phylogenetic background (Figure 4f). The time windows 70-75 mya, 65-70 mya, and 60-65 mya display fold changes of over six-fold increase, while smaller fold changes are observed for time windows 35-40 and 25-30 mya (Figure 4f).

Discussion

Waves of paleopolyploidizations during angiosperm evolution

Analyzing an extensive dataset containing 470 complete angiosperm genomes, we found evidence for 132 ancient polyploidization events. This seems a lot but if we

consider independent evolutionary lineages, most plant lineages seem to contain remnants of only one, sometimes two, and in rare cases, three, four, or five WGD events (Table 1). The contrast between numerous instances of recurring polyploidy and recently originated polyploids, and the few events of ancient polyploid species' survival over time periods of 140 to 150 million years (Figure 1), is striking and provides a "polyploidy paradox" (Van de Peer et al. 2017; Carretero-Paulet and Van de Peer 2020). This paradox can be understood when the short-term 'high rates' of WGD are countered by a reduced survival rate at longer evolutionary temporal scales. If one sees so many parallel, mostly single-occurrence events, one can wonder about the odds that these independent WGDs have occurred and got established at 'random times', or whether these WGDs survived because they occurred at decisive moments in evolution and therefore cluster in time. Furthermore, if species birth and death rates are not randomly distributed in time, some deterministic processes may be at play.

As we have shown here, different model-based analyses all seem to support some 'waves' of WGD events, i.e., a clustering of inferred WGDs in time. For instance, there is a clustering of WGDs 50-55 mya as supported by the WGD cluster peak at 53.54 mya in the AIC-best mixture model, and the 'episodical pulled rate' model test. Another period of increased established polyploidizations is apparent 60 to 75 mya, supported by the WGD cluster peaking at 62.46 mya in the silhouette score-best mixture model, the single Gaussian model peaking at 70.62 mya favored by the Bayesian phylogenetic clustering analysis, and the 'episodical pulled rate' model test. Finally, there is some support for a potentially more recent clustering of WGDs, as for instance implied by the WGD cluster peak at 20.56 mya in the AIC-best mixture model, and 32.14 mya in the silhouette score-best mixture model, respectively (see above).

Based on simulation, we identified two periods of an excess of WGDs, one in the 0-40 mya and one in the 40-80 mya time intervals. While Gaussian mixture modeling (GMM) provides information on single time points for each WGD cluster, they fail to consider variations in the number of lineages at different time points, potentially leading to false positives. Therefore, instead of relying on single time points, by integrating all information obtained by modeling and simulation, we propose two somewhat extended periods, namely 50-75 mya and 25-40 mya, respectively, during which the number of paleopolyploidizations seems to have significantly increased and WGDs cluster in time.

The single Gaussian model peaking at 70.62 mya, as favored by the Bayesian phylogenetic clustering analysis, seems to lend strong support for the paleopolyploid wave during 50-75 mya. The accelerated cumulative rate of established WGDs at the shift point 74.1 mya, as indicated by the Bayesian change point analysis, together with a strong drop in the number of WGDs per species since 50 mya, reinforces the paleopolyploid wave 50-75 mya.

Ancient polyploidy correlates with the Paleocene–Eocene Thermal Maximum (PETM) and the Cretaceous–Paleogene (K–Pg) extinction

Within the 50-75 mya period, possibly two different waves of WGDs can be discerned (see above). A first cluster of WGDs seems to correlate with the PETM event ~55.8 mya, a period characterized by a 4-5 °C increase in global mean surface temperature (Zachos et al. 2008; Dunkley Jones et al. 2013), and a massive release of carbon to the ocean–atmosphere system (Zeebe et al. 2009; Zeebe et al. 2016; Gutjahr et al. 2017). Cai and coauthors (Cai et al. 2019) also reported a clustering of ancient WGD events around the Paleocene–Eocene transition, based on the dating of 22 independent ancient WGD events from genomic and transcriptomic assemblies of 36 Malpighiales species. A second cluster of WGDs seems to overlap with the K–Pg boundary, the result of the last of the five most recent mass extinctions, ~66 mya (Renne et al. 2013). Previously, Fawcett and co-authors (Fawcett et al. 2009) and Vanneste and coauthors (Vanneste et al. 2014) also suggested a potential clustering of WGDs around the K–Pg boundary. Here, using a genomic dataset twelve times larger (N = 470, with 132 ancient WGDs), we find stronger support for this observation. It is striking to see that, at a time all non-avian dinosaurs went extinct, as well as more than 70% of all other animal and plant life (Nichols and Johnson 2008; Longrich et al. 2012; Longrich et al. 2016), well-known plant lineages and all of their descendants, such as Orchidaceae, Papilionoideae, Oleaceae, Polygonaceae, Cactaceae, Portulacaceae, Proteaceae, Convolvulaceae, Santalaceae, Nelumbonaceae, Apiaceae, Thymelaeaceae, Zingiberales, Nyssaceae, Brassicaceae, Aquifoliaceae, Actinidiaceae, Theaceae, and Ericaceae, show evidence for WGDs that got established around that time (Supplementary Table 4).

We tested for significant increases in oscillations in atmospheric CO_2 concentration (Foster et al. 2017), low-latitude Sea Surface Temperature (SST) (Song et al. 2019), and Global Mean Surface Temperature (GMST) (Judd et al. 2024) in the

period between 50 and 75 mya relative to background oscillation levels. We sampled 10,000 random time periods with duration of 25 million years and compared their standard deviations to that of the time period between 50 and 75 mya using Wilcoxon signed-rank test and one-sample t-test. We observed significantly increased oscillations in CO_2 concentration and GMST, but not low-latitude SST ($P < 0.0001$). The clustering of WGDs in a period marked by increased oscillations in CO_2 concentration and GMST suggests a potential association between environmental turmoil and WGD occurrence, supporting the hypothesis that WGD can confer a selective advantage in stressful times.

Support for a wave of paleopolyploidizations, 25-40 mya

As stated above, there is some support for a potentially more recent clustering of WGDs. The 25-40 mya period experienced a prominent glaciation paleolatitude expansion (Cather et al. 2009) and rapid shifts of low-latitude SST and a marked drop of GMST (Figure 5b-d). All these paleoclimate proxies show significantly higher oscillation levels during the 25-40 mya period compared to the background in the past 145 million years ($P < 0.0001$).

Revisiting the polyploid paradox

Nature is teeming with polyploid species that have extra sets of chromosomes. Yet, ancient WGDs that survive over the long term are surprisingly scarce (Table 1). The apparent paucity of (established) ancient genome duplications and the existence of so many species that are currently polyploid provides a fascinating but puzzling paradox (Soltis et al. 2015; Van de Peer et al. 2017). The rarity of polyploidy events that get 'fixed' on the long run seems to suggest that polyploidy is usually disadvantageous, and an evolutionary dead end. However, there is growing evidence that polyploidization can sometimes help overcome environmental change (stress), or even periods of extinction (Carruthers et al. 2024; Van de Peer et al. 2021; Li et al. 2024). At decisive moments in evolution, as again shown here, polyploids seem to have had an adaptive advantage and likely outcompeted many of their diploid progenitors. Despite much research, the mechanistic underpinnings of why and how polyploids might be able to outcompete or outlive non-polyploids remain elusive. In the longer term, WGD might increase both mutational and environmental robustness due to redundancy and increased genetic variation. Apart from the overall observation

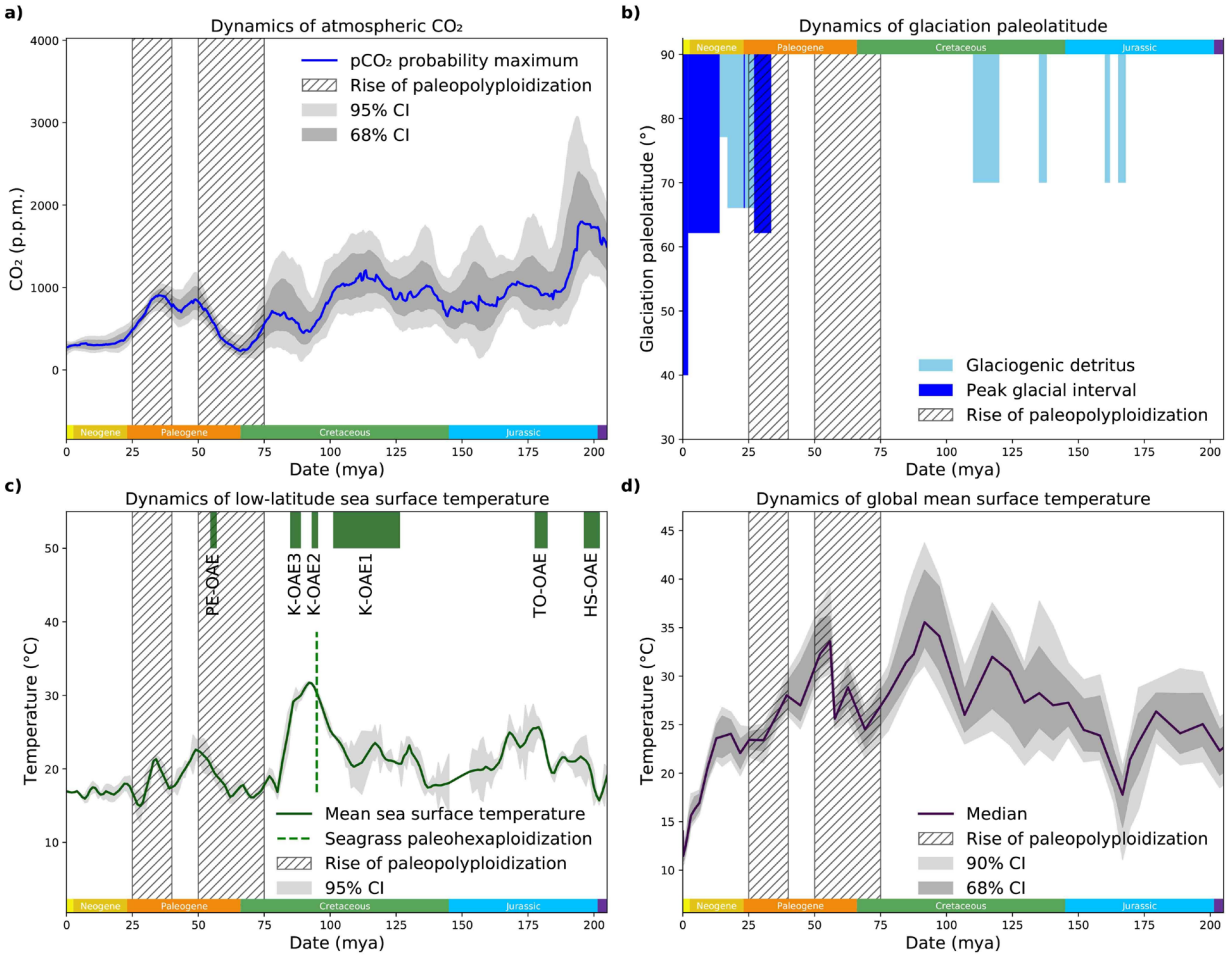

Figure 5. Dynamics of paleoclimates and paleopolyploidizations. The time periods 25-40 and 50-75 mya are indicated in shading in all panels. Temporal dynamics of different paleoclimate proxies during the last 205 million years are presented in different panels with chronostratigraphic periods annotated. Panel a) shows the temporal dynamics of the atmospheric CO_2 concentration in which the dark blue solid curve represents the probability maximum while the dark and light gray bands represent the 68 and 95% confidence intervals, respectively (Foster et al. 2017). Panel b) shows the temporal paleolatitudinal distribution of the glaciogenic detritus and peak glacial interval (after Cather et al. (2009)), in blue and dark blue bars, respectively. Panel c) shows the temporal dynamics of the low-latitude sea surface temperature (Song et al. 2019), in which the green solid curve represents the mean sea surface temperature, while the grey band represents the 95% confidence interval. The date of seagrass paleohexaploidization is denoted with dashed green line. Oceanic anoxic events (OAEs) are denoted as green bars (after Song et al. (2019) in which the abbreviation HS represents Hettangian–Sinemurian, TO represents Toarcian, K represents Cretaceous and PE represents Paleocene–Eocene. Panel d) shows the temporal dynamics of the global mean surface temperature in which the purple solid curve represents the median while the dark and light gray bands represent the 68 and 90% confidence intervals, respectively (Judd et al. 2024).

that the ratio of polyploids over diploids is higher in more extreme environments (Rice et al. 2019), there are anecdotal studies where it has been shown that specific genes have been retained that could be linked to increased chances for survival under extreme conditions. For instance, it has been shown that duplicates of RNA-binding

proteins involved in cold responses exhibited convergent retention across diverse angiosperm lineages following independent WGD events occurring around the K–Pg boundary (Guo et al. 2024). Another study reported that gene duplicates functioning in low temperature and darkness were commonly retained in different plant lineages, after independently occurring WGDs around the K–Pg boundary, when global cooling and darkness were the two main stresses (Wu et al. 2020). However, first the polyploid lineage must survive and on the short - or even immediate - term, selective advantages of WGDs are harder to explain.

Recently, some of us (Ebadi et al. 2023) showed that, by duplicating artificially generated Gene Regulatory Networks (GRNs), duplicated GRNs - and thus duplicated genomes - show higher signal output variation than nonduplicated GRNs. This increased variation leads, at least in theory, to niche expansion and can promote phenotypic variation in polyploids. In contrast, GRNs might also be maladaptive to changes under the same environmental conditions in which they emerge, a phenomenon that is exacerbated in duplicated GRNs (Yao et al. 2019). Together with other known disadvantages of WGD (Comai 2005; Bomblies 2020), this could explain why WGD might be a dead-end under relatively stable conditions (hence their scarcity on the evolutionary tree) but could confer a selective advantage under times of considerable environmental upheaval. Polyploids with more extreme phenotypes emerging shortly before or after periods of environmental disturbance such as the PETM event or cataclysmic events (such as the one causing the K–Pg extinction) might have had higher chances of survival compared to their diploid progenitors (Ebadi et al. 2023). In the aftermath of environmental turmoil, when a considerable proportion of existing species might have perished and new niches became available, polyploids might have been exempted from the frequency-dependent minority cytotype exclusion caused by factors such as ineffective pollination and inviable zygotes (Levin 1975), and might show novel beneficial traits by which they could quickly occupy those newly opened niches (Thompson and Lumaret 1992; Rodriguez 1996; te Beest et al. 2012; Blaine Marchant et al. 2016; Baniaga et al. 2020; Castro et al. 2024).

Conclusions

The advent of genomics made it clear that the significance of polyploidy extends across all eukaryotes, and that most, if not all, extant species (including our own) carry the signature of at least one ancient whole genome duplication (WGD) in their history

(Van de Peer et al. 2009; Wendel 2015; Soltis and Soltis 2016; Van de Peer et al. 2017; Fox et al. 2020). Nevertheless, while finding evidence for ancient WGD is relatively straightforward (Chen and Zwaenepoel 2023), dating these events, going back tens to hundreds of millions of years ago, is notoriously difficult. Here, we assembled and assessed a large angiosperm genome dataset comprising 470 different angiosperm species. We first reconstructed a highly concordant evolutionary timescale of angiosperms which agrees well with other previously published large-scale phylogenetic studies. We then used state-of-the-art methods to identify and date 132 independent, ancient WGDs. Model-based analyses and simulations reveal that the inferred WGD events are not randomly distributed but rather cluster around a few periods of environmental upheaval, in particular the K–Pg boundary and the PETM event, indicative of a major paleopolyploidization wave during 50-75 mya, while there is also some evidence for a minor one, 25-40 mya. These findings support the view that polyploidy often is an evolutionary dead end but can be an evolutionary and ecological force in unstable times. In this respect it is interesting to speculate that, in the current Anthropocene, a time during which humans are having a substantial impact on our planet, the number of polyploidy events might be on the rise again.

Materials and Methods

Collection and quality assessment of genome assemblies

We searched for all published angiosperm genome assemblies on the platform PlabiPD (January 5th, 2024). We downloaded at least one representative genome for most genera via the link provided in the data availability section of the corresponding article, leading to 470 genomes in total. To assess the quality of the genome assemblies and annotations, we used BUSCO (Simão et al. 2015) (v5.4.3) ('protein' mode, embryophyta_odb10 database) and OMArk (Nevers et al. 2024) (Viridiplantae.h5 database).

Phylogenetic inference of angiosperms

Reconstructing species trees from gene trees involves uncertainties and biases such as incomplete lineage sorting (ILS) and sampling, horizontal gene transfer (HGT), gene flow, etc. Inferring orthologues de novo from a sequence similarity graph, as implemented in some ortholog inference algorithms (Li et al. 2003; Emms and Kelly 2019), is sensitive to parameters such as inflation factor (De Smet et al. 2013),

especially for single-copy orthogroups (SOGs). To avoid such biases, we used the "reference orthologues" from the embryophyta_odb10 BUSCO dataset (Simão et al. 2015) to retrieve the orthologues in each one of the 470 angiosperm genomes, which were used to build individual gene trees, and summarize the species tree under the multispecies coalescence (MSC) model. Briefly, for each SOG in the *embryophyta_odb10* database, we first used HMMER (Potter et al. 2018) (v3.1b2) to find corresponding orthologues using the pre-built profile hidden Markov model and the cut-off table from BUSCO. When multiple hits occurred, only the best hit was used for subsequent analysis. Secondly, a multiple sequence alignment (MSA) of nucleotides was inferred for each SOG using MAFFT (Katoh and Standley 2013) (v7.310). IQTREE (Nguyen et al. 2014) (v1.6.5) was then applied to infer the maximum likelihood (ML) gene tree for each MSA with the automatic detection of the best-fit substitution model via ModelFinder (Kalyaanamoorthy et al. 2017). The uncertainties of branches were measured by the SH-aLRT (Guindon et al. 2010) and aBayes test with 1000 ultrafast bootstrap replicates, which were further optimized by a hill-climbing nearest neighbor interchange (NNI) search.

To find the optimal branch support cut-off to use for filtering branches, we tested cut-offs of 0%, 10%, 20%, 30%, 40%, 50%, 60% and 70%. To account for the effect of different MSA data types on the branch length and topology of inferred gene tree (Wickett et al. 2014), we conducted the gene tree inference using i) codon-level nucleotide; ii) 1st- and 2nd-position codon nucleotide; iii) TAPER (Zhang et al. 2021)-masked nucleotide; and iv) raw nucleotide datasets. Thus, we generated 32 gene tree datasets (eight branch support cut-offs x four MSA data types), each containing gene trees inferred from 1614 SOGs. The 1614 individual gene trees of each dataset were then supplied to ASTRAL-Pro (Zhang et al. 2020) to summarize the species tree under default parameters. We adopted the phylogenetic inference from the raw nucleotide dataset with the branch support cut-off of 50% based on its concordance with the statistical topological test and the optimized trade-off between high branch support of gene trees and high statistical support of the inferred phylogeny in terms of local posterior probability (LPP) (Supplementary Note 3).

Divergence time tree inference of angiosperms

Using the species tree obtained with ASTRAL-Pro, we conducted molecular dating analyses based on the concatenated amino acid MSA of 12 SOGs that covered

at least 466 species (species coverage 99%) to obtain a time-calibrated tree. The time tree was inferred using the uncorrelated lognormal (UCLN) relaxed molecular clock model implemented in mcmctree (Yang 2007) (v4.9j), with soft fossil calibration bounds. The model parameters in mcmctree (v4.9j) were set as follows: the clock model was set to 'independent rates model'; a log-normal distribution of evolutionary rates across branches was set with a LG amino acid substitution matrix and a gamma model with five rate categories and $\alpha = 0.5$; parameters for the birth-death process were set as 1 1 0.1 to generate uniform age priors on nodes that didn't have a fossil calibration; gamma priors for the transition/transversion rate ratio and the shape parameter for variable rates among sites were set as 6 2 and 1 1; a Dirichlet-gamma prior was set with the parameters controlling the mean rate across loci and the logarithmic variance as 2 20 1 and 1 10 1. The first 2000 generations were discarded as burn-in after which 2,000,000 generations were pulled with sampling per 100 generations. To ensure accuracy and consistency, we assembled an angiosperm fossil calibration dataset (Supplementary Note 4) and the full age constraints implemented in our molecular clock analysis are shown in the calibration map (Supplementary Materials). The effective sample size (ESS) was higher than 200 for all parameters, suggesting good convergence. The inferred time tree, i.e., our main tree, is provided in Supplementary Materials.

Identification and dating of putative WGD events

Intragenomic collinear segments and a burst of genes that seemed duplicated at the same time can be indicative of a putative WGD event. Here, we used i-ADHoRe (Proost et al. 2011) (v3.0.01) and wgd v2 (Chen et al. 2024) to search for genomic collinearity and for bursts of gene duplication events in K_S-based age distributions, where the number of duplicated genes is plotted against their age (Chen and Zwaenepoel 2023). In short, the program "wgd dmd" was applied with default parameters to construct the whole paranome (collection of all duplicated genes in a genome) and the program "wgd syn" was applied with default parameters to infer genomic collinearity using i-ADHoRe. The whole paranome and 'anchor pair' (pairs of duplicated genes derived from large-scale gene duplication events and residing in duplicated segments) K_S distributions were then constructed using the program "wgd ksd" with default parameters (Chen et al. 2024). We applied the exponential-lognormal mixture model (ELMM) and the log-scale Gaussian mixture model (GMM), as

implemented in the program “wgd viz”, to delineate potential WGD components from the whole paranome and anchor pair K_S distributions, respectively. The phylogenetic location of putative WGD events can be inferred by ranking the relative timing of WGD events and speciation events within a K_S timeframe after proper correction of substitution rate variation. The former is represented as peaks of the delineated log-normal components in the whole paranome and anchor pair K_S distributions while the latter is represented as peaks of the kernel density estimate (KDE) of the K_S distributions for orthologous genes, identified as the reciprocal best hit (RBH) across all species.

For each putative WGD event, we compared its relative timing to a sequence of speciation events based on prior investigations of earlier studies to delineate its phylogenetic location (detailed species composition for each putative WGD event in Supplementary Note 6). The absolute dating of the identified WGD event was conducted in a standardized way. In short, given an identified WGD event for a group of species belonging to a certain order, we assembled an orthologous gene family dataset consisting of the anchor pairs and their orthologues from species of other orders (detailed order composition and species composition for each WGD event in Supplementary Note 6), in total 16-18 species in the orthologous gene family following Chen et al. (2024). For the WGD event shared by multiple orders, the orders that shared this WGD event were not further included in the order composition. Based on the constructed orthologous gene family and our main tree topology, we conducted the phylogenetic WGD dating on the concatenated amino acid MSA using mcmctree (v4.9j) with the same composition of ordinal age constraints and the same parameters as implemented in the divergence time tree inference section, except for the increased generations as 20,000,000 and the sampling frequency as 1,000. For each WGD event, we dated at least one species, and we adopted the consensus mean, calculated as the overall mean of all WGD date estimates for a WGD event, and the 90% Highest Convergence Region (HCR) calculated as the intersection of the 90% Highest Posterior Density (HPD) of the WGD date estimates from each species for a WGD event, to represent the age and the associated uncertainty of a WGD event. The estimated K_S peak and WGD date for each independent WGD event are summarized in Supplementary Table 4.

Bayesian change point analysis

We set up a Bayesian model to identify the time point at which the mean cumulative rate of WGD occurrence significantly changed. In short, an uniform prior distribution for the location of change point was set from 0 to 125 mya, around which the WGD occurrence within angiosperms has just started to cumulate, a half-normal prior distribution with sigma of 1 was set for the mean and the std of the cumulative rate of WGD occurrence before and after the change point. We sampled 4 chains for 100,000 tune and 100,000 iterations, in total 800,000 iterations. The Monte Carlo Standard Error of the Standard Deviation (MCSE_SD) to SD ratio was less than 0.005, the R-hat was smaller than 1.01 and the effective sample size (ESS) was higher than 200, suggesting good convergence. The Maximum A Posteriori (MAP) estimate of the posterior distribution of change point was identified as the shift point.

WGD clustering

To investigate whether there could be WGD clusters around specific times, we applied two clustering methods, namely log-scale (natural log) Gaussian mixture modeling (GMM) and the Bayesian phylogenetic clustering. For the GMM analysis, we fit the observed WGD dates to the models with one, two, three and four components and evaluated their performance according to the Akaike Information Criterion (AIC) and silhouette score. AIC is a statistical measure used to select the best-fitting model by balancing the trade-off between goodness of fit and model complexity. The silhouette score is commonly used to evaluate the quality of different clustering results. For the Bayesian phylogenetic clustering analysis, we treated WGD dates as trait values and we conducted phylogenetic multilevel regression using the brms package (Bürkner 2021) to take the tree topology into account. In short, the WGD dates were modeled using an intercept-only Bayesian generalized linear model based on prior expectations to obtain posterior estimates of the WGD dates, assuming either a single Gaussian distribution or a mixture of two to three Gaussian distributions. Models were phylogenetically corrected by calculating Bray-Curtis distances between species and adding the resulting phylogenetic distances as blocking factor to the model. Models were initially run using 4 chains and 4,000 iterations using non-informative (uniform) priors. We assessed model convergence by R-hat and ensured all values <1.02 for all chains. If convergence was not reached, the number of iterations was increased to

10,000 with priors more narrowly defined according to information from phylogenetic uncorrected models. After convergence, a leave-one-out cross-validation approach based on the posterior probability was conducted for model evaluation using the loo (Vehtari et al. 2015) package. Different models were compared using loo-model criterion in terms of absolute values. Sensitivity analysis was conducted for the used priors, including changes towards broadly and narrowly defined Gaussian priors, which showed robust estimates for all models. Pagel's lambda, a measure of phylogenetic signal, was calculated using the phylosignal (Pagel 1999; Keck et al. 2016) package. Due to computational infeasibility, the Bayesian phylogenetic clustering analysis was conducted on a reduced dataset of a 416-species timetree with 126 ancient WGD events (Supplementary Materials).

WGD simulation model and episodic pulled rate model test

Inspired by Vanneste and co-authors (Vanneste et al. 2014), we devised a WGD simulation model whereby WGD events along each branch of the time-calibrated phylogenetic tree were modeled as an independent Poisson process with the time interval between events exponentially distributed and controlled by the WGD occurrence rate parameter. To accommodate the variation of WGD occurrence rates across different clades, we allowed clade-specific rate parameters for the ANA grade, Ceratophyllales, Chloranthales, eudicots, monocots and magnoliids, respectively. We calculated the clade-specific rate parameters that maximized the likelihood given the independent Poisson process model and the observed WGD events as the phylogenetic background rate parameters for each clade (Supplementary Table 5). We also validated that this model with such variable rates fit the data significantly better than the alternative simpler model with a constant rate across all clades in our prior model design (likelihood ratio test, $P < 0.0001$). Under this independent Poisson process model and the clade-specific background rate parameters, we simulated WGDs throughout the main tree and documented the randomly generated WGD dates for 10,000 generations to construct the null distribution. We conducted two-sample Kolmogorov-Smirnov test between the null distribution and the distribution of the observed WGD dates to determine whether the distribution of the observed WGD dates departed significantly from that of the randomly generated WGDs. We also conducted Wilcoxon signed-rank test and one-sample t-test between the median pairwise distance of the observed WGD dates and that of the simulated WGD dates

of 10,000 generations bounded by different time intervals to determine whether the observed median pairwise distance between WGDs was significantly shorter than the expected by chance given the null distribution, as described in Vanneste et al. (2014), so as to evaluate the significance of a potential WGD cluster at certain time intervals.

To test the hypothesis that the WGD occurrence rate might vary, or more specifically, increase, at specific time ranges, we further devise an episodic pulled rate model test. In short, we pulled the WGD occurrence rate at a specific time window to a ratio relative to the phylogenetic background rate ranging from 1 to 100 that maximized the likelihood of observing the WGD dates given the independent Poisson process model and the episodically pulled rate at that specific time window. A likelihood ratio test was then conducted between the independent Poisson process model with the episodically pulled rate and the independent Poisson process model with the pulled ratio constrained to 1 (i.e., the same rate as phylogenetic background) to determine whether adding the complexity of the pulled rate at specific time windows could significantly improve model fit. A pulled rate is penalized by branches that bear no WGD occurrence and a significantly favored pulled rate model at specific time windows indicates a strong phylogenetic trend that a larger number of independent lineages convergently have WGD occurrences at these time windows than would be expected under the phylogenetic background. As such, we can evaluate the significance and magnitude of an increase in the WGD occurrence rate against the phylogenetic background at specific time windows.

Acknowledgements

H.C. acknowledges funding from the Research Foundation—Flanders (FWO) (No. 3G032219). Y.V.d.P. acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (No. 833522). Y.V.d.P. and F.A-S. acknowledge funding from Ghent University (Methusalem funding, BOF.MET.2021.0005.01).

Author contributions

Y.V.d.P. and H.C. conceived and managed the project. H.C. collected the genomic data and conducted the assessment analysis. H.C. conducted the phylogenetic and molecular dating analysis. H.C. conducted the WGD inference and dating analysis. H.C. conducted the Bayesian change point analysis, GMM clustering,

WGD simulation and episodic pulled rate model test analysis. G.L., D.B. and F.A.S. conducted the Bayesian phylogenetic clustering analysis. F.A.S. constructed the website. H.C., F.A.S. and Y.V.d.P. wrote the manuscript. All authors read and approved the manuscript.

Conflict of interest

The authors declare no competing interests.

Data and code availability

The genomic data involved in this study are available at (Chen 2024b) and (Chen 2024c). The posterior samples of WGD dates for each independent WGD event are available at (Chen 2024a). The code for the Bayesian phylogenetic clustering and the independent Poisson process model analysis are available in Supplementary Materials. The code for other analyses is available at https://github.com/heche-psb/Angiosperm_WGD

Table 1. Number of WGDs in different clades.

Major clade	Order/Multi-Order	Family/Multi-Family	Number of WGDs
Eudicots (95)	Lamiales (10)	Lamiaceae	3
		Oleaceae	2
		Plantaginaceae, Orobanchaceae, Lentibulariaceae, Gesneriaceae	1
		Partial* Lamiales	1
	Brassicales (9)	Brassicaceae	5
		Cleomaceae, Capparaceae, Akaniaceae	1
		Partial Brassicales	1
	Malpighiales (7)	Linaceae	2
		Salicaceae, Rhizophoraceae, Passifloraceae, Euphorbiaceae, Calophyllaceae	1
	Fabales (7)	Fabaceae	5
		Polygalaceae	2
	Ranunculales (5)	Papaveraceae	2
		Ranunculaceae, Lardizabalaceae	1
		Ranunculales	1
	Myrtales (5)	Melastomataceae, Lythraceae	2
		Myrtales	1
	Malvales (5)	Malvaceae	3
		Thymelaeaceae, Dipterocarpaceae	1
	Caryophyllales (5)	Tamaricaceae, Polygonaceae, Caryophyllaceae, Amaranthaceae,	1
		Partial Caryophyllales	1
	Sapindales (4)	Sapindaceae, Rutaceae, Meliaceae Anacardiaceae	1
	Cucurbitales (4)	Begoniaceae	2
		Cucurbitaceae	1
		Cucurbitales	1
	Apiales (4)	Araliaceae	2
		Apiaceae	1
		Partial Apiales	1
	Rosales (3)	Elaeagnaceae	2
		Rosaceae	1
	Ericales (3)	Primulaceae, Actinidiaceae	1
		Partial Ericales	1
	Asterales (3)	Asteraceae	3
	Trochodendrales (2)	Trochodendraceae	2
	Solanales (2)	Solanaceae, Convolvulaceae	1
	Saxifragales (2)	Crassulaceae	2
	Proteales (2)	Proteaceae, Nelumbonaceae	1
	Cornales (2)	Nyssaceae, Cornaceae	1
	Boraginales (2)	Boraginaceae	2

	Santalales (1)	Santalaceae	1
	Gentianales (1)	Rubiaceae	1
	Fagales (1)	Juglandaceae	1
	Dipsacales (1)	Caprifoliaceae	1
	Crossosomatales (1)	Partial Crossosomatales	1
	Celastrales (1)	Celastraceae	1
	Buxales (1)	Buxaceae	1
	Aquifoliales (1)	Aquifoliaceae	1
	Core eudicots (1)	Core eudicots	1
Monocots (25)	Poales (9)	Poaceae	8
		Poales	1
	Alismatales (5)	Araceae	2
		Cymodoceaceae	1
		Partial Alismatales	2
	Zingiberales (3)	Zingiberaceae	1
		Musaceae	1
		Zingiberales	1
	Dioscoreales (2)	Dioscoreaceae	2
	Asparagales (2)	Orchidaceae	1
		Asparagaceae	1
	Pandanales (1)	Velloziaceae	1
	Arecales (1)	Arecaceae	1
	Acorales (1)	Acoraceae	1
	Non-Alismatales monocots (excluding Acorales) (1)	Non-Alismatales monocots (excluding Acorales)	1
Magnoliids (5)	Piperales (2)	Piperaceae	1
		Partial Piperales	1
	Laurales (2)	Lauraceae, Calycanthaceae	1
	Laurales+Magnoliales (1)	Laurales+Magnoliales	1
ANA grade (4)	Nymphaeales (4)	Cabombaceae	2
		Nymphaeaceae	1
		Partial Nymphaeales	1
Ceratophyllales (2)	Ceratophyllales (2)	Ceratophyllaceae	2
Chloranthales (1)	Chloranthales (1)	Chloranthaceae	1

Numbers in brackets denote the number of WGDs within that clade. *Partial (e.g., partial Lamiales) means that there is a WGD event shared by more than 1 family that we sampled within Lamiales (i.e., shared by Orobanchaceae, Bignoniaceae, Lamiaceae, Pedaliaceae, Scrophulariaceae and Acanthaceae).

References

Almeida-Silva F, Van de Peer Y. 2023. Whole-genome Duplications and the Long-term Evolution of Gene Regulatory Networks in Angiosperms. *Mol Biol Evol* **40**.

APG T. 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.

Baduel P, Bray S, Vallejo-Marin M, Kolář F, Yant L. 2018. The "Polyploid Hop": Shifting Challenges and Opportunities Over the Evolutionary Lifespan of Genome Duplications. *Front Ecol Evol* **6**: 117.

Baniaga AE, Marx HE, Arrigo N, Barker MS. 2020. Polyploid plants have faster rates of multivariate niche differentiation than their diploid relatives. *Ecology Letters* **23**: 68-78.

Blaine Marchant D, Soltis DE, Soltis PS. 2016. Patterns of abiotic niche shifts in allopolyploids relative to their progenitors. *New Phytologist* **212**: 708-718.

Bomblies K. 2020. When everything changes at once: finding a new normal after genome duplication. *Proceedings of the Royal Society B: Biological Sciences* **287**: 20202154.

Bray SM, Hämälä T, Zhou M, Busoms S, Fischer S, Desjardins SD, Mandáková T, Moore C, Mathers TC, Cowan L et al. 2024. Kinetochore and ionomic adaptation to whole-genome duplication in *Cochlearia* shows evolutionary convergence in three autopolyploids. *Cell Reports* **43**.

Bürkner P-C. 2021. Bayesian Item Response Modeling in R with brms and Stan. *Journal of Statistical Software* **100**: 1 - 54.

Cai L, Xi Z, Amorim AM, Sugumaran M, Rest JS, Liu L, Davis CC. 2019. Widespread ancient whole-genome duplications in Malpighiales coincide with Eocene global climatic upheaval. *New Phytologist* **221**: 565-576.

Carretero-Paulet L, Van de Peer Y. 2020. The evolutionary conundrum of whole-genome duplication. *American Journal of Botany* **107**: 1101-1105.

Carruthers T, Gonçalves DJP, Li P, Chanderbali AS, Dick CW, Fritsch PW, Larson DA, Soltis DE, Soltis PS, Weaver WN et al. 2024. Repeated shifts out of tropical climates preceded by whole genome duplication. *New Phytologist* **244**: 2561-2575.

Castro M, Celeste Dias M, Loureiro J, Husband BC, Castro S. 2024. Competitive ability, neopolyploid establishment and current distribution of a diploid–tetraploid plant complex. *Oikos* **2024**: e09949.

Cather SM, Dunbar NW, McDowell FW, McIntosh WC, Scholle PA. 2009. Climate forcing by iron fertilization from repeated ignimbrite eruptions: The icehouse–silicic large igneous province (SLIP) hypothesis. *Geosphere* **5**: 315-324.

Chao D-Y, Dilkes B, Luo H, Douglas A, Yakubova E, Lahner B, Salt DE. 2013. Polyploids Exhibit Higher Potassium Uptake and Salinity Tolerance in *Arabidopsis*. *Science* **341**: 658-659.

Chen H. 2024a. Posterior samples of WGD dates for 132 independent WGD events in angiosperms. doi:10.6084/m9.figshare.27596856.v1. figshare.

Chen H. 2024b. Supplementary cds files of 470 angiosperm species. doi:10.6084/m9.figshare.27011128.v1. figshare.

Chen H. 2024c. Supplementary gff3 files of 470 angiosperm species. doi:10.6084/m9.figshare.27011134.v1. figshare.

Chen H, Zwaenepoel A. 2023. Inference of Ancient Polyploidy from Genomic Data. In *Polyploidy: Methods and Protocols* (ed. Y Van de Peer), pp. 3-18. Springer US, New York, NY.

Chen H, Zwaenepoel A, Van de Peer Y. 2024. wgd v2: a suite of tools to uncover and date ancient polyploidy and whole-genome duplication. *Bioinformatics* **40**.

Coate JE, Powell AF, Owens TG, Doyle JJ. 2013. Transgressive physiological and transcriptomic responses to light stress in allopolyploid *Glycine dolichocarpa* (Leguminosae). *Heredity* **110**: 160-170.

Comai L. 2005. The advantages and disadvantages of being polyploid. *Nature Reviews Genetics* **6**: 836-846.

De Smet R, Adams KL, Vandepoele K, Van Montagu MCE, Maere S, Van de Peer Y. 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proceedings of the National Academy of Sciences* **110**: 2898-2903.

Doyle JJ, Coate JE. 2019. Polyploidy, the Nucleotype, and Novelty: The Impact of Genome Doubling on the Biology of the Cell. *International Journal of Plant Sciences* **180**: 1-52.

Dunkley Jones T, Lunt DJ, Schmidt DN, Ridgwell A, Sluijs A, Valdes PJ, Maslin M. 2013. Climate model and proxy data constraints on ocean warming across the Paleocene–Eocene Thermal Maximum. *Earth-Science Reviews* **125**: 123-145.

Ebadi M, Bafort Q, Mizrachi E, Audenaert P, Simoens P, Van Montagu M, Bonte D, Van de Peer Y. 2023. The duplication of genomes and genetic networks and its potential for evolutionary adaptation and survival during environmental turmoil. *Proceedings of the National Academy of Sciences* **120**: e2307289120.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**: 238.

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. *Proceedings of the National Academy of Sciences* **106**: 5737-5742.

Foster GL, Royer DL, Lunt DJ. 2017. Future climate forcing potentially without precedent in the last 420 million years. *Nature Communications* **8**: 14845.

Fox D, Soltis DE, Soltis PS, Ashman T-L, Van de Peer Y. 2020. Polyploidy: a biological force from cells to ecosystems. *Trends Cell Biology* **30**: 688-694.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* **59**: 307-321.

Guo L, Wang S, Jiao X, Ye X, Deng D, Liu H, Li Y, Van de Peer Y, Wu W. 2024. Convergent and/or parallel evolution of RNA-binding proteins in angiosperms after polyploidization. *New Phytol* **242**: 1377-1393.

Gutjahr M, Ridgwell A, Sexton PF, Anagnostou E, Pearson PN, Pälike H, Norris RD, Thomas E, Foster GL. 2017. Very large release of mostly volcanic carbon during the Palaeocene–Eocene Thermal Maximum. *Nature* **548**: 573-577.

Haldane J. 1933. The part played by recurrent mutation in evolution. *The American Naturalist* **67**: 5-19.

Hu H, Sun P, Yang Y, Ma J, Liu J. 2023. Genome-scale angiosperm phylogenies based on nuclear, plastome, and mitochondrial datasets. *Journal of Integrative Plant Biology* **65**: 1479-1489.

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 Biology* **13**: R3.

Jiao Y, Li J, Tang H, Paterson AH. 2014. Integrated Syntenic and Phylogenomic Analyses Reveal an Ancient Genome Duplication in Monocots. *The Plant Cell* **26**: 2792-2802.

Judd EJ, Tierney JE, Lunt DJ, Montañez IP, Huber BT, Wing SL, Valdes PJ. 2024. A 485-million-year history of Earth's surface temperature. *Science* **385**: eadk3705.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**: 587-589.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**: 772-780.

Keck F, Rimet F, Bouchez A, Franc A. 2016. phylosignal: an R package to measure, test, and explore the phylogenetic signal. *Ecology and Evolution* **6**: 2774-2780.

Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, Zdobnov EM. 2018. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research* **47**: D807-D811.

Landis JB, Soltis DE, Li Z, Marx HE, Barker MS, Tank DC, Soltis PS. 2018. Impact of whole-genome duplication events on diversification rates in angiosperms. *American Journal of Botany* **105**: 348-363.

Leebens-Mack JH Barker MS Carpenter EJ Deyholos MK Gitzendanner MA Graham SW Grosse I Li Z Melkonian M Mirarab S et al. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* **574**: 679-685.

Levin DA. 1975. Minority Cytotype Exclusion in Local Plant Populations. *Taxon* **24**: 35-43.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research* **13**: 2178-2189.

Li X, Wang M, Zou M, Guan X, Xu S, Chen W, Wang C, Chen Y, He S, Guo B. 2024. Recent and Recurrent Autopolyploidization Fueled Diversification of Snow Carp on the Tibetan Plateau. *Molecular Biology and Evolution* **41**.

Longrich NR, Bhullar B-AS, Gauthier JA. 2012. Mass extinction of lizards and snakes at the Cretaceous–Paleogene boundary. *Proceedings of the National Academy of Sciences* **109**: 21396-21401.

Longrich NR, Scriberas J, Wills MA. 2016. Severe extinction and rapid recovery of mammals across the Cretaceous–Palaeogene boundary, and the effects of rarity on patterns of extinction and recovery. *Journal of Evolutionary Biology* **29**: 1495-1512.

Mortier F, Bafort Q, Milosavljevic S, Kauai F, Prost Boxoen L, Van de Peer Y, Bonte D. 2024. Understanding polyploid establishment: temporary persistence or stable coexistence? *Oikos* **2024**: e09929.

Nevers Y, Warwick Vesztrocy A, Rossier V, Train C-M, Altenhoff A, Dessimoz C, Glover NM. 2024. Quality assessment of gene repertoire annotations with OMArk. *Nature Biotechnology*.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2014. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* **32**: 268-274.

Nichols DJ, Johnson KR. 2008. *Plants and the K–T Boundary*.

Novikova PY, Hohmann N, Van de Peer Y. 2018. Polyploid *Arabidopsis* species originated around recent glaciation maxima. *Current Opinion in Plant Biology* **42**: 8-15.

Pagel M. 1999. Inferring the historical patterns of biological evolution. *Nature* **401**: 877-884.
Parham JF, Donoghue PCJ, Bell CJ, Calway TD, Head JJ, Holroyd PA, Inoue JG, Irmis RB, Joyce WG, Ksepka DT et al. 2011. Best Practices for Justifying Fossil Calibrations. *Systematic Biology* **61**: 346-359.
Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. 2018. HMMER web server: 2018 update. *Nucleic Acids Research* **46**: W200-W204.
Pozo Jd, Ramirez-Parra E. 2014. Deciphering the molecular bases for drought tolerance in *Arabidopsis* autotetraploids. *Plant, Cell & Environment* **37**: 2722-2737.
Proost S, Fostier J, De Witte D, Dhoedt B, Demeester P, Van de Peer Y, Vandepoele K. 2011. i-ADHoRe 3.0—fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic Acids Research* **40**: e11-e11.
Ramírez-Barahona S, Sauquet H, Magallón S. 2020. The delayed and geographically heterogeneous diversification of flowering plant families. *Nature Ecology & Evolution* **4**: 1232-1238.
Renne PR, Deino AL, Hilgen FJ, Kuiper KF, Mark DF, Mitchell WS, Morgan LE, Mundil R, Smit J. 2013. Time Scales of Critical Events Around the Cretaceous–Paleogene Boundary. *Science* **339**: 684-687.
Rice A, Šmarda P, Novosolov M, Drori M, Glick L, Sabath N, Meiri S, Belmaker J, Mayrose I. 2019. The global biogeography of polyploid plants. *Nature Ecology & Evolution* **3**: 265-273.
Rodriguez DJ. 1996. A Model for the Establishment of Polyploidy in Plants. *The American Naturalist* **147**: 33-46.
Ruiz M, Quiñones A, Martínez-Cuenca MR, Aleza P, Morillon R, Navarro L, Primo-Millo E, Martínez-Alcántara B. 2016. Tetraploidy enhances the ability to exclude chloride from leaves in carrizo citrange seedlings. *Journal of Plant Physiology* **205**: 1-10.
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210-3212.
Soltis PS, Marchant DB, Van de Peer Y, Soltis DE. 2015. Polyploidy and genome evolution in plants. *Curr Opin Genet Dev* **35**: 119-125.
Soltis PS, Soltis DE. 2016. Ancient WGD events as drivers of key innovations in angiosperms. *Curr Opin Plant Biol* **30**: 159-165.
Song H, Wignall PB, Song H, Dai X, Chu D. 2019. Seawater Temperature and Dissolved Oxygen over the Past 500 Million Years. *Journal of Earth Science* **30**: 236-243.
Stebbins GL. 1950. Chapter VI. Isolation and the Origin of Species. In *Variation and Evolution in Plants*, pp. 189-250. Columbia University Press, New York Chichester, West Sussex.
Stebbins GL. 1971. Chromosomal evolution in higher plants.
te Beest M, Le Roux JJ, Richardson DM, Brysting AK, Suda J, Kubesova M, Pysek P. 2012. The more the better? The role of polyploidy in facilitating plant invasions. *Ann Bot* **109**: 19-45.
Thompson JD, Lumaret R. 1992. The evolutionary dynamics of polyploid plants: origins, establishment and persistence. *Trends in Ecology & Evolution* **7**: 302-307.
Van de Peer Y, Ashman T-L, Soltis PS, Soltis DE. 2021. Polyploidy: an evolutionary and ecological force in stressful times. *The Plant Cell* **33**: 11-26.

Van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics* **10**: 725-732.
Van de Peer Y, Mizrachi E, Marchal K. 2017. The evolutionary significance of polyploidy. *Nature Reviews Genetics* **18**: 411-424.
Vanneste K, Baele G, Maere S, Van de Peer Y. 2014. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. *Genome Research* **24**: 1334-1347.
Vehtari A, Gabry J, Magnusson M, Yao Y, Bürkner P-C, Paananen T, Gelman A. 2015. loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. *CRAN: Contributed Packages*.
Wagner Jr. WH. 1970. Biosystematics and Evolutionary Noise. *Taxon* **19**: 146-151.
Wendel JF. 2015. The wondrous cycles of polyploidy in plants. *American Journal of Botany* **102**: 1753-1756.
Wickett NJ, Mirarab S, Nguyen N, Warnow T, Carpenter E, Matasci N, Ayyampalayam S, Barker MS, Burleigh JG, Gitzendanner MA et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences* **111**: E4859-E4868.
Wu S, Han B, Jiao Y. 2020. Genetic Contribution of Paleopolyploidy to Adaptive Evolution in Angiosperms. *Mol Plant* **13**: 59-71.
Yang L, Su D, Chang X, Foster CSP, Sun L, Huang C-H, Zhou X, Zeng L, Ma H, Zhong B. 2020. Phylogenomic Insights into Deep Phylogeny of Angiosperms Based on Broad Nuclear Gene Sampling. *Plant Communications* **1**: 100027.
Yang PM, Huang QC, Qin GY, Zhao SP, Zhou JG. 2014. Different drought-stress responses in photosynthesis and reactive oxygen metabolism between autotetraploid and diploid rice. *Photosynthetica* **52**: 193-202.
Yang S-J, Zhang Y-J, Zhang Y, Peng J-C, Liu C-Y, Li D-W, Guo W. 2024. Divergence in cold tolerance promotes niche differentiation between diploid and polyploid kiwifruits along an altitudinal gradient in Southwest China. *Oikos* **2024**: e10181.
Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* **24**: 1586-1591.
Yao Y, Carretero-Paulet L, Van de Peer Y. 2019. Using digital organisms to study the evolutionary consequences of whole genome duplication and polyploidy. *PLoS One* **14**: e0220257.
Zachos JC, Dickens GR, Zeebe RE. 2008. An early Cenozoic perspective on greenhouse warming and carbon-cycle dynamics. *Nature* **451**: 279-283.
Zeebe RE, Ridgwell A, Zachos JC. 2016. Anthropogenic carbon release rate unprecedented during the past 66 million years. *Nature Geoscience* **9**: 325-329.
Zeebe RE, Zachos JC, Dickens GR. 2009. Carbon dioxide forcing alone insufficient to explain Palaeocene–Eocene Thermal Maximum warming. *Nature Geoscience* **2**: 576-580.
Zhang C, Scornavacca C, Molloy EK, Mirarab S. 2020. ASTRAL-Pro: Quartet-Based Species-Tree Inference despite Paralogy. *Molecular Biology and Evolution* **37**: 3292-3307.
Zhang C, Zhao Y, Braun EL, Mirarab S. 2021. TAPER: Pinpointing errors in multiple sequence alignments despite varying rates of evolution. *Methods in Ecology and Evolution* **12**: 2145-2158.
Zhu H, Zhao S, Lu X, He N, Gao L, Dou J, Bie Z, Liu W. 2018. Genome duplication improves the resistance of watermelon root to salt stress. *Plant Physiology and Biochemistry* **133**: 11-21.

Zuntini AR Carruthers T Maurin O Bailey PC Leempoel K Brewer GE Epitawalage N Françoso E Gallego-Paramo B McGinnie C et al. 2024. Phylogenomics and the rise of the angiosperms. *Nature*.