Supplemental Data. Li and Defoort et al. (2016). Plant Cell 10.1105/tpc.16.00877

Gene duplicability of core genes is highly consistent across all angiosperms

- Supplemental Material -

Zhen Li[1,2,3*], Jonas Defoort[1,2,3*], Setareh Tasdighian[1,2,3], Steven Maere[1,2,3], Yves Van de Peer[1,2,3,4], Riet De Smet[1,2,3]

[1]Department of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Ghent, Belgium

[2]Department of Plant Biotechnology and Bioinformatics, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

[3]Bioinformatics Institute Ghent, Ghent University, Technologiepark 927, B-9052 Gent, Belgium

[4]Genomics Research Institute, University of Pretoria, Hatfield Campus, Pretoria 0028, South Africa

[*]Equal contribution

Correspondence should be sent to:

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

Riet De Smet

VIB / Ghent University

Technologiepark 927

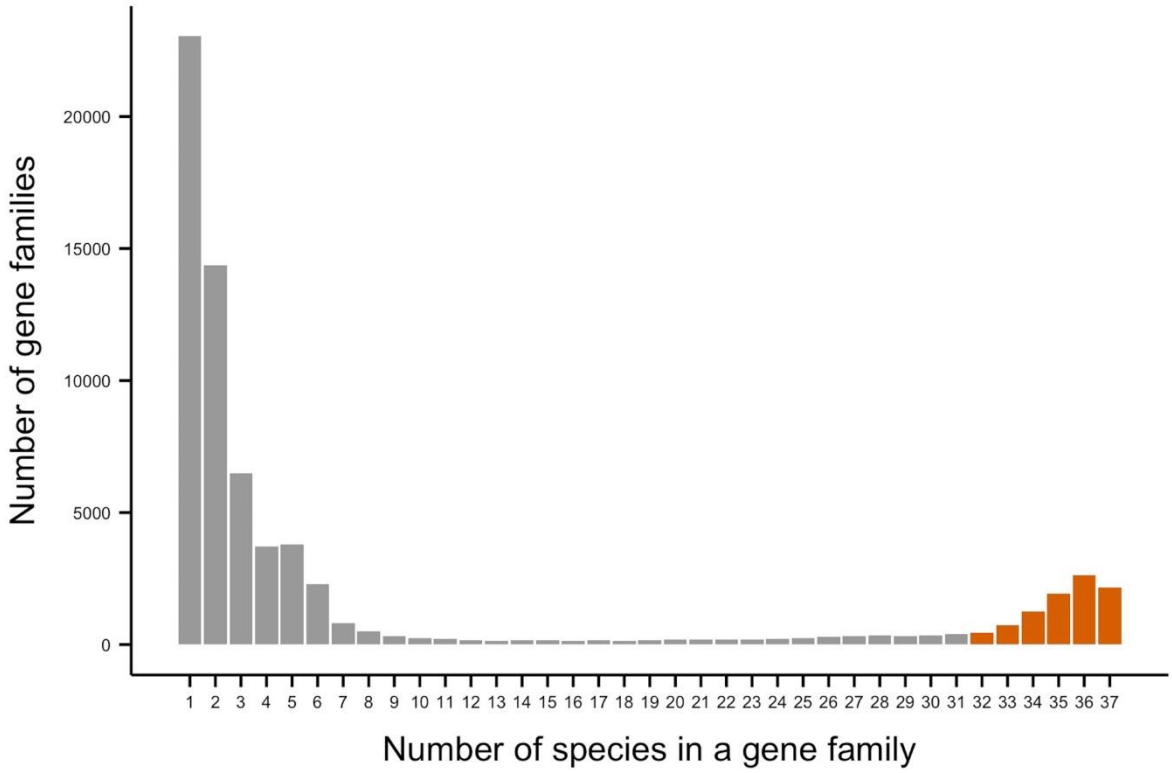
Gent (9052), Belgium

Tel: +32 (0)9 331 35 36

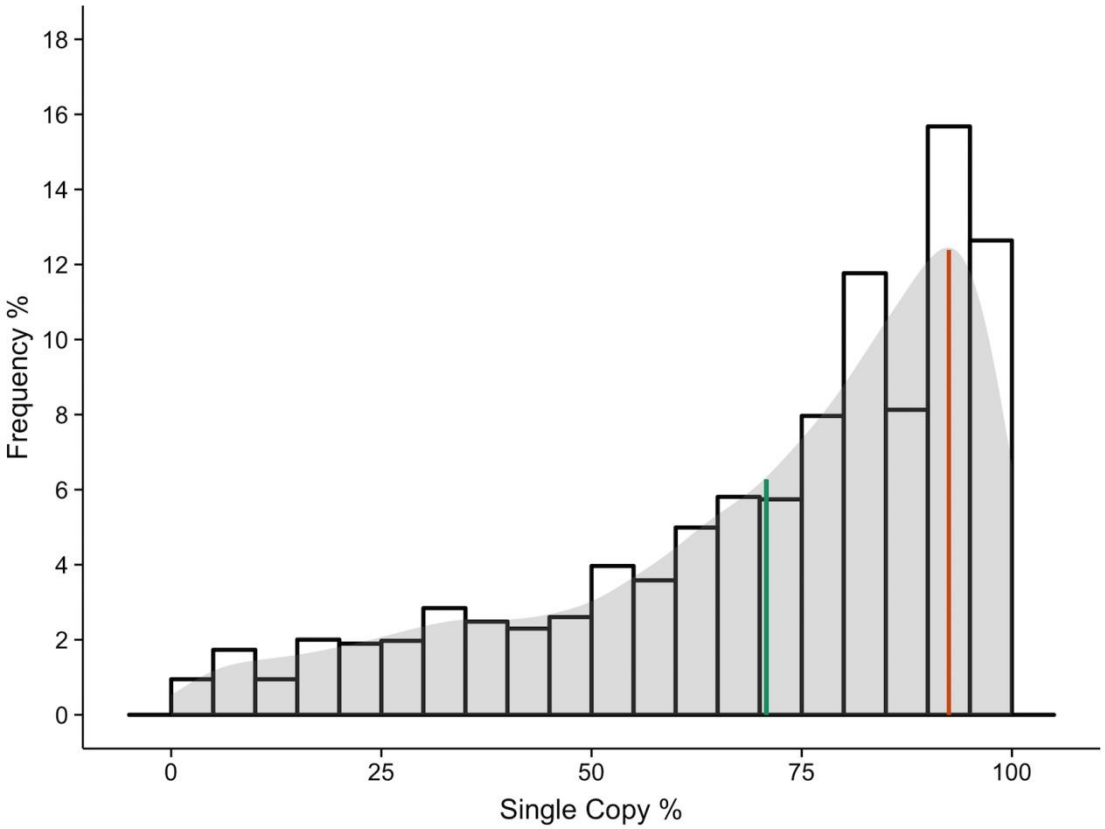
Fax: +32 (0)9 331 3809

E-mail: riet.desmet@psb.vib-ugent.be

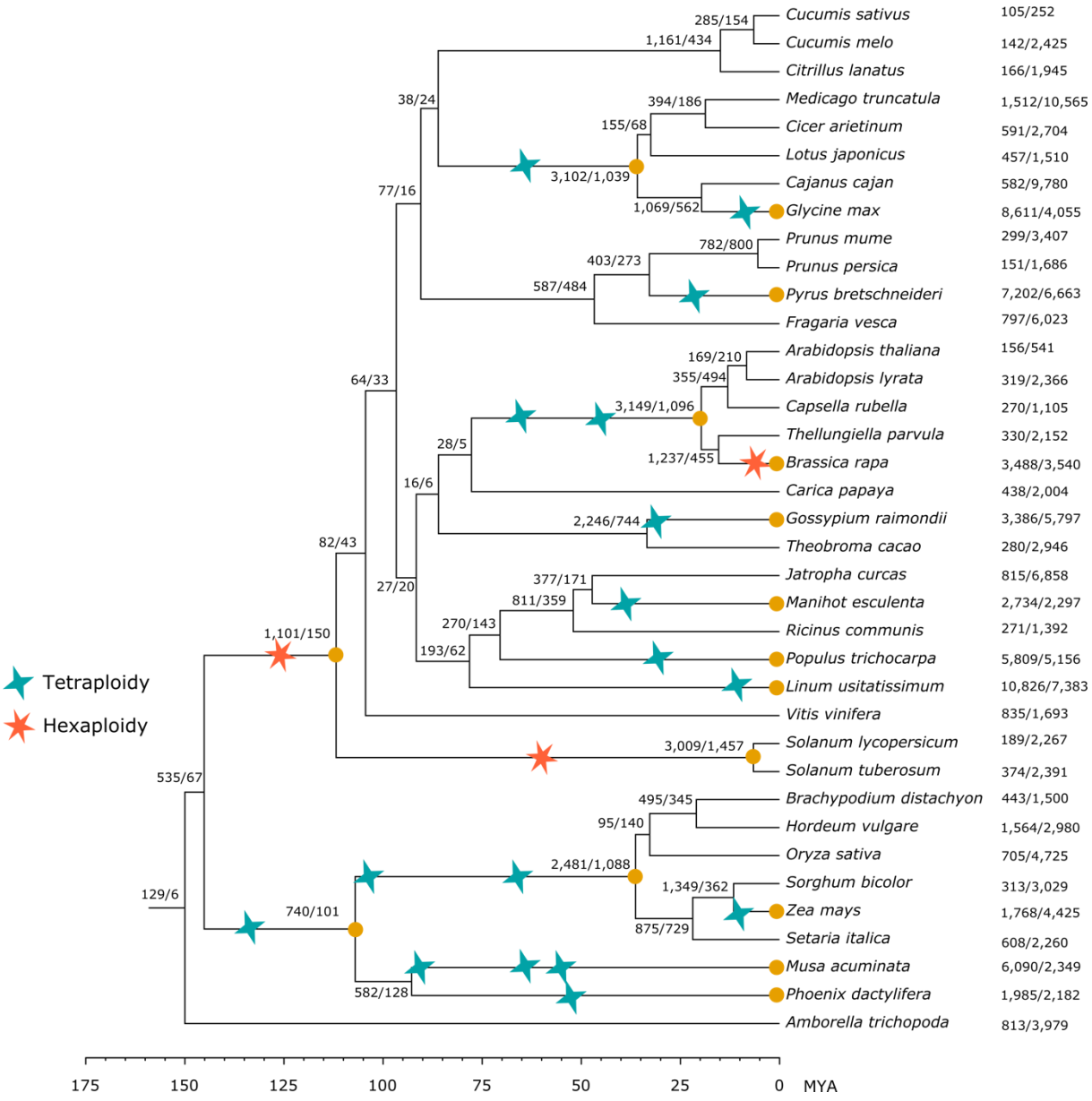
SUPPLEMENTAL FIGURES



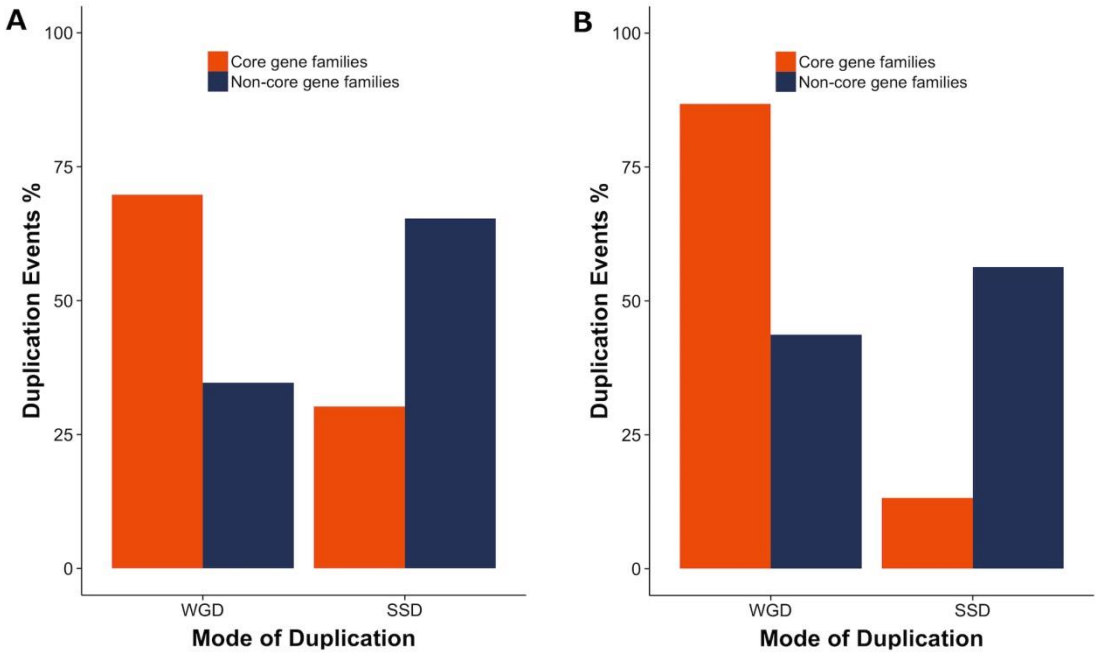
Supplemental Figure 1. Motivation for the 32 out of 37 species cut-off to define core gene families. To distinguish core from non-core gene families we assessed the distribution of the number of species in each gene family based on all 69,542 gene families obtained by reconciliation. This distribution is U-shaped, suggesting a large number of gene families that are species- or lineage-specific (left side of the distribution) and also an excess of gene families present in the large majority of angiosperm species (right side of the distribution). Based on this distribution we decided to consider all gene families containing genes from at least 32 species as being 'core gene families'. As such we account for a limited number of putative missing orthologs from core gene families due to for instance errors in genome annotation, gene family construction errors or the presence of incomplete genomes.



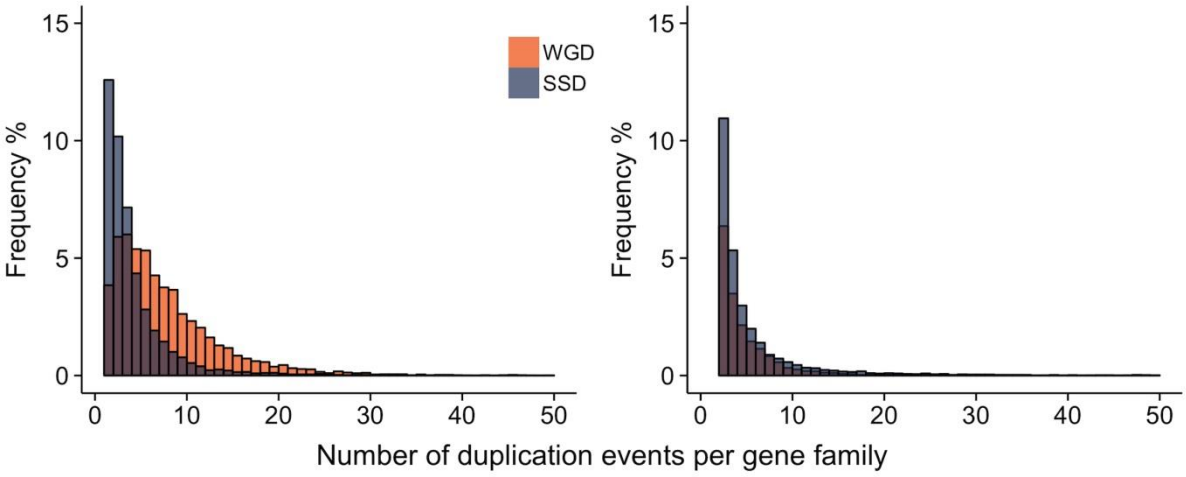
Supplemental Figure 2. The distribution of Single-Copy Percentages (SCPs) for all core gene families, with SCPs calculated upon removing the highly duplicated genomes of *Glycine max*, *Linum usitatissimum*, *Brassica rapa*, and *Zea mays*. This distribution has a mode of 92% and a mean of 70.8%.



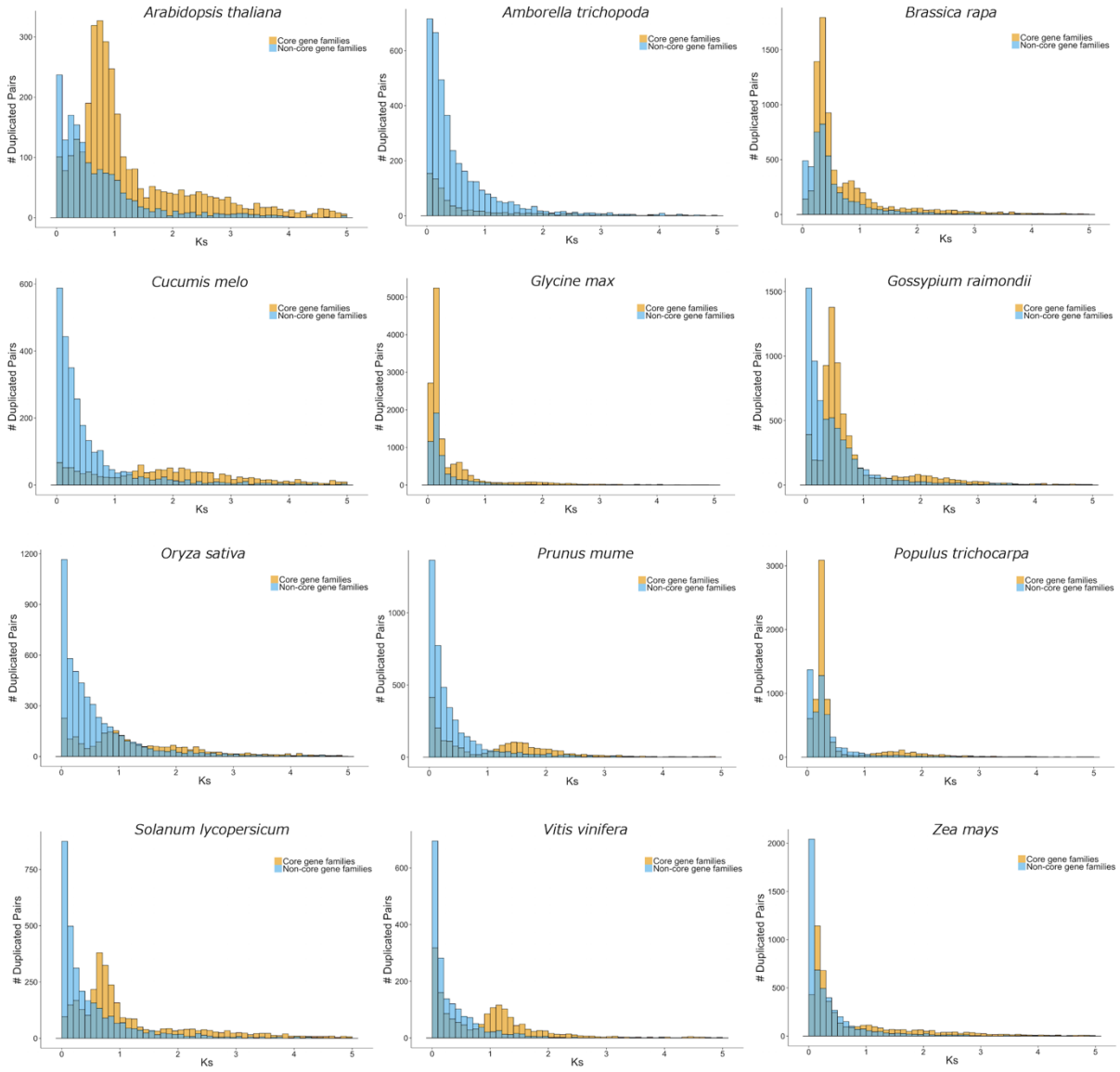
Supplemental Figure 3. Classification of species tree nodes as SSD or WGD. On the species tree, nodes with WGDs on their parent branches were considered as WGD nodes (orange dots), while the rest of the nodes were considered as SSD nodes. Next to each node are the number of duplication events predicted by gene tree-species tree reconciliation for both core and non-core gene families (core/non-core). There are in total 93,942 predicted duplication events in core gene families and 140,786 duplication events in non-core gene families.



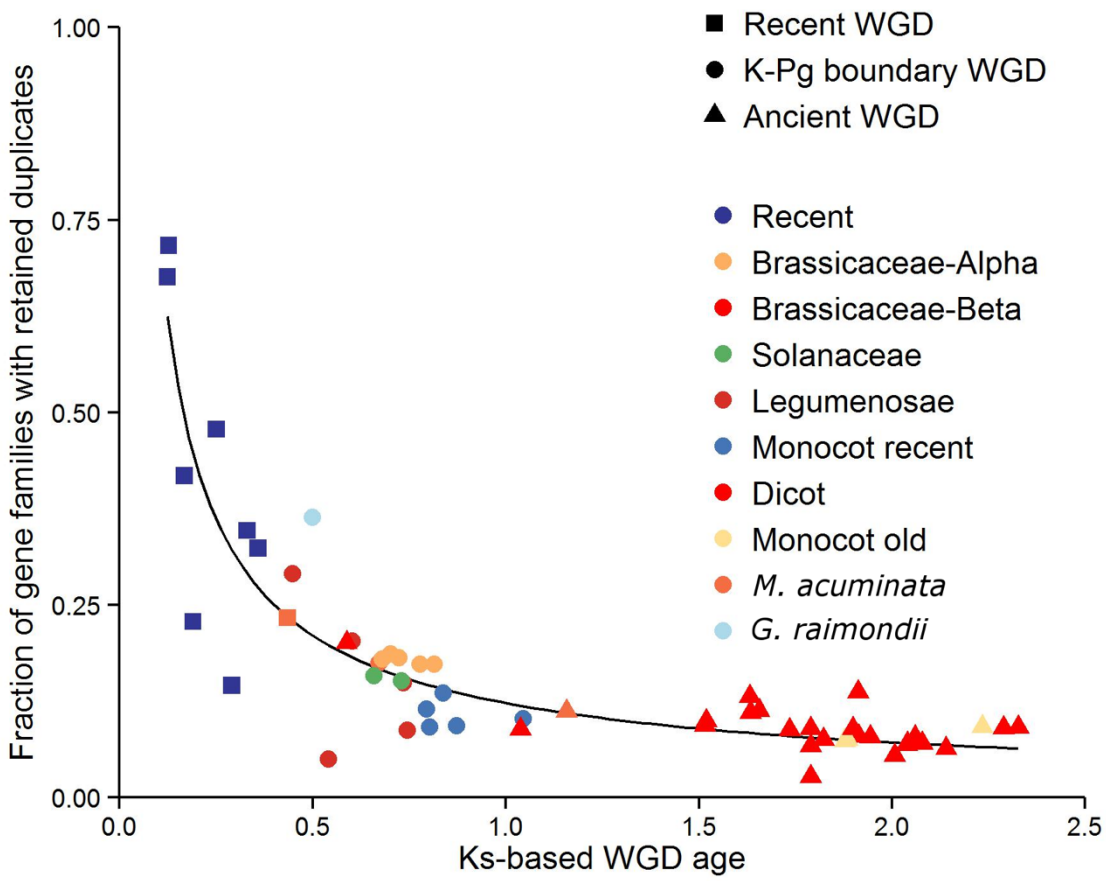
Supplemental Figure 4. Core gene families mainly duplicate through WGD. Bar plots represent the fraction of duplication events, summed over all gene families, attributed to WGD or SSD in core and non-core gene families. Panel (A) represents results obtained from all nodes in the species tree in (Supplemental Figure 2) and shows that for core genes families, as compared to non-core gene families, the presence of duplicates seems to be biased towards WGD-associated gene duplication ($p < 2.2e-16$, Fisher's exact test). In panel (B) we assessed the possibility that these observations might be caused by an overrepresentation of WGD-associated nodes in the species tree for core gene families as opposed to non-core gene families: since core gene families cover by definition a larger number of species, some of the more ancient WGD events that are shared by many species will only be represented by core gene families. Hence, we repeated this analysis by only considering nodes from the species tree that are also ubiquitously present in non-core gene families (top 10 of the nodes) and came to the same conclusion ($p < 2.2e-16$, Fisher's exact test).



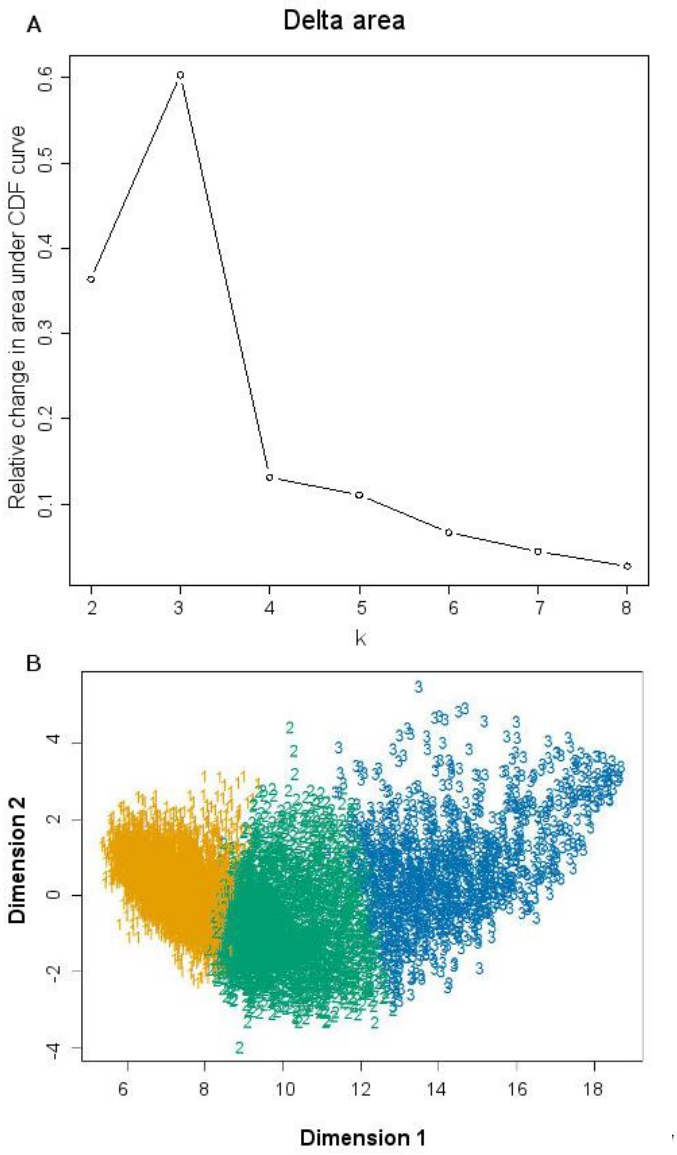
Supplemental Figure 5. Comparison of the number of duplications for core and non-core gene families at WGD and SSD nodes on a gene family base (only illustrating gene families with no more than 50 duplications). (A) The number of WGD and SSD duplications per gene family for core gene families. There are significantly more nodes associated with WGD derived duplications than SSD derived duplications ($p < 2.2e-16$, Wilcoxon-rank-sum test). (B) The number of WGD and SSD duplication per gene family for non-core gene families. Here the number of WGD derived duplications is not significantly larger than those of SSD derived duplications ($p = 1$, Wilcoxon-rank-sum test). Predicted duplication events were obtained by gene tree - species tree reconciliation (see Materials and Methods).



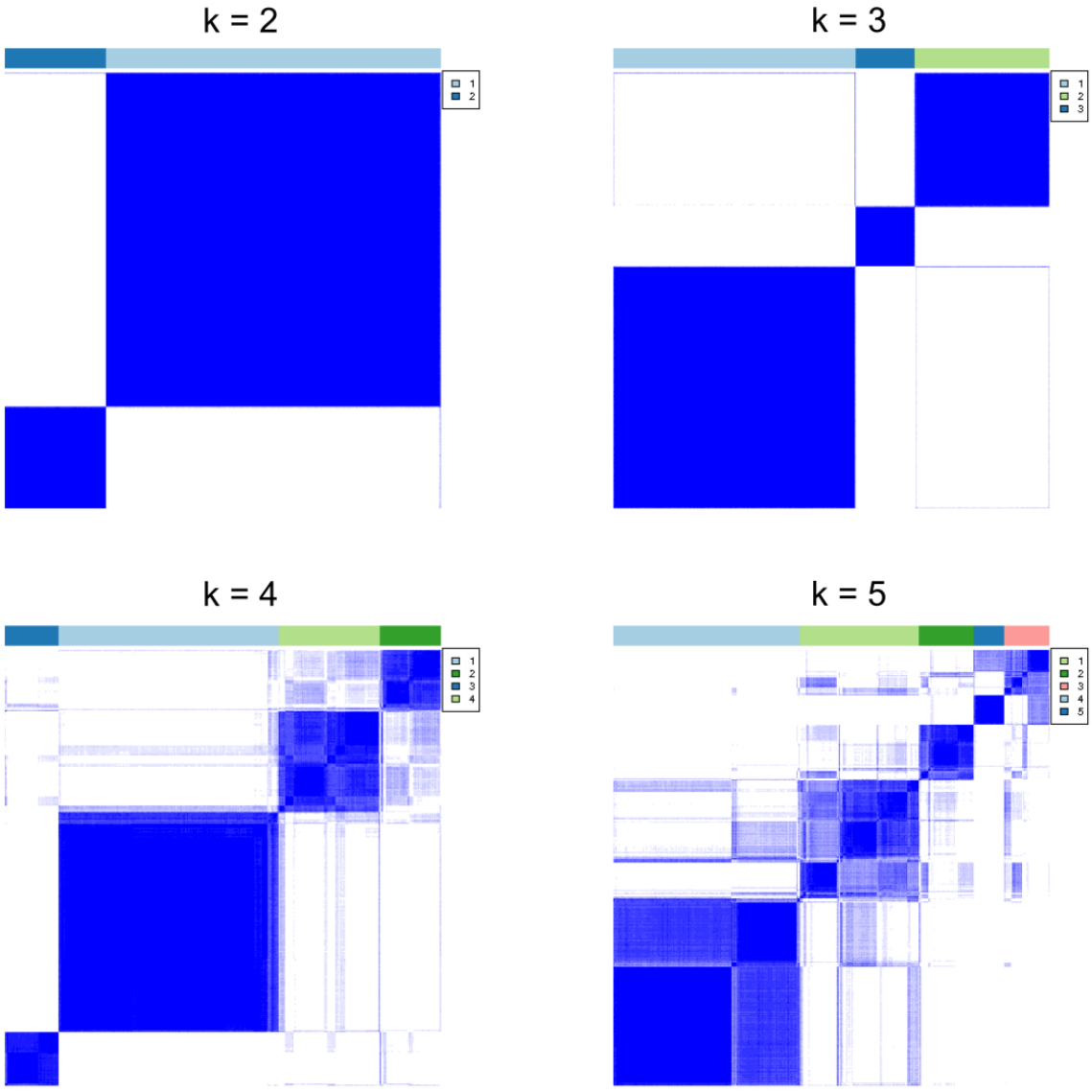
Supplemental Figure 6. K_S-distributions of duplicated pairs from core and non-core gene families in 12 species, i.e. *Arabidopsis thaliana*, *Amborella trichopoda*, *Brassica rapa*, *Cucumis melo*, *Glycine max*, *Gossypium raimondii*, *Oryza sativa*, *Prunus mume*, *Populus trichocarpa*, *Solanum lycopersicum*, *Vitis vinifera*, and *Zea mays*.



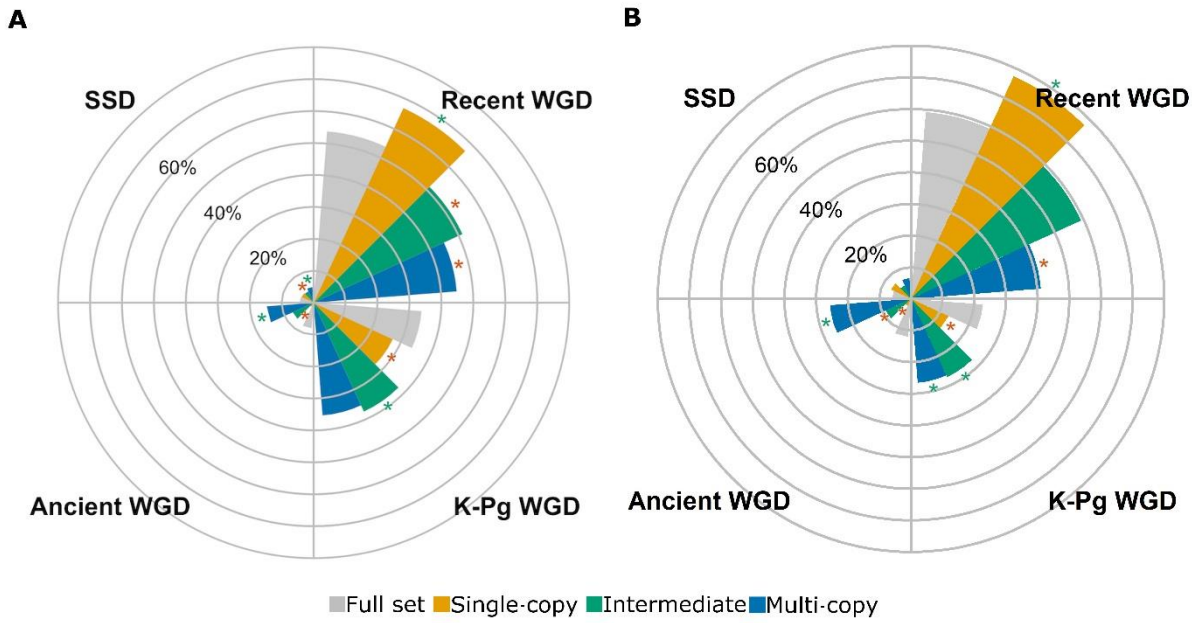
Supplemental Figure 7. Duplicate gene retention in function of time since WGD. Each dot represents the fraction of core gene families with retained duplicates following a specific WGD (y-axis), as a function of WGD age, expressed in K_S-units (x-axis). The timing of the WGD events and the particular gene families that retained duplicates following a specific WGD event were inferred by fitting Gaussian mixture models to K_S-age distributions for all 37 species separately (see Materials and Methods). This figure is related to Figure 3, but here all WGD peak callings were included. Since the Dicot and Brassicaceae-Beta peaks can not be distinguished from each other they are denoted by the same color. Additional information on all the peaks is provided in the Supplemental Table 2.



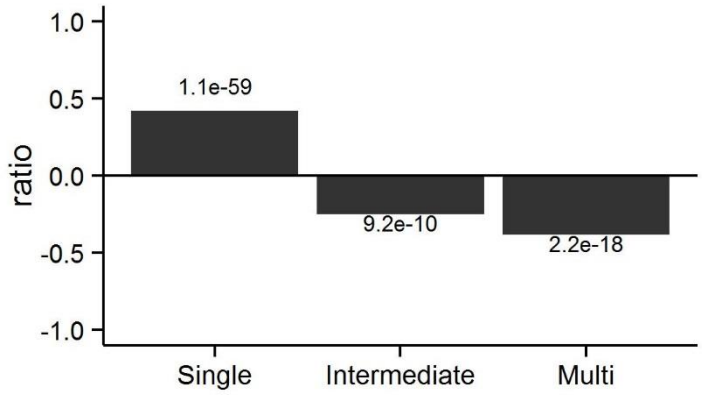
Supplemental Figure 8. Criteria that we used to choose the optimal number of clusters for k-means clustering of the copy-number matrix. (A) We used the Delta Area Plot from the ConsensusClusterPlus R-package to select the optimal number of clusters. The results of 1000 clustering runs, each time on subsampled matrices, are summarized into a consensus matrix, whose values represent the proportion of clustering runs in which two items (i.e. gene families) are grouped together. Hence, values in this matrix are between 0 and 1 (= always clustered together). The Delta Area Plot assesses the 'cleanness' of this consensus matrix: if all clustering runs agree on the same solution than this matrix only consists of 0's and 1's (bimodal distribution). To determine the optimal numbers of clusters the largest changes in these consensus values are detected by calculating the change in the area under the Cumulative Distribution of consensus values for increasing cluster number (Monti et al., 2003). The 'Delta area' represents this change, with *k* corresponding to cluster number. (B) Corresponding multidimensional scaling plot of the copy-number matrix, with data points colored according to cluster membership.



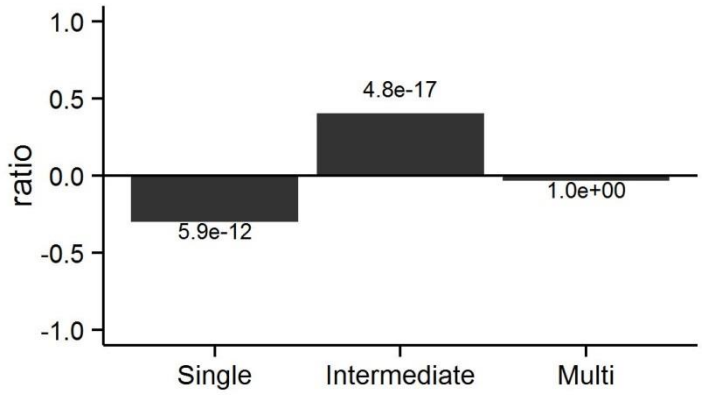
Supplemental Figure 9. Consensus matrices obtained for different number of clusters *k*. The consensus matrix represents the number of times that two gene families belonged to the same cluster over 1,000 clustering runs of the subsampled copy-number matrix. The values within this matrix range from 0 (gene families were never grouped into the same cluster; white in this figure) to 1 (gene families were always grouped into the same cluster; blue in this figure). Here results are shown for *k* = 2-5 clusters. Color bars on top of the visualized consensus matrix indicate cluster assignments.



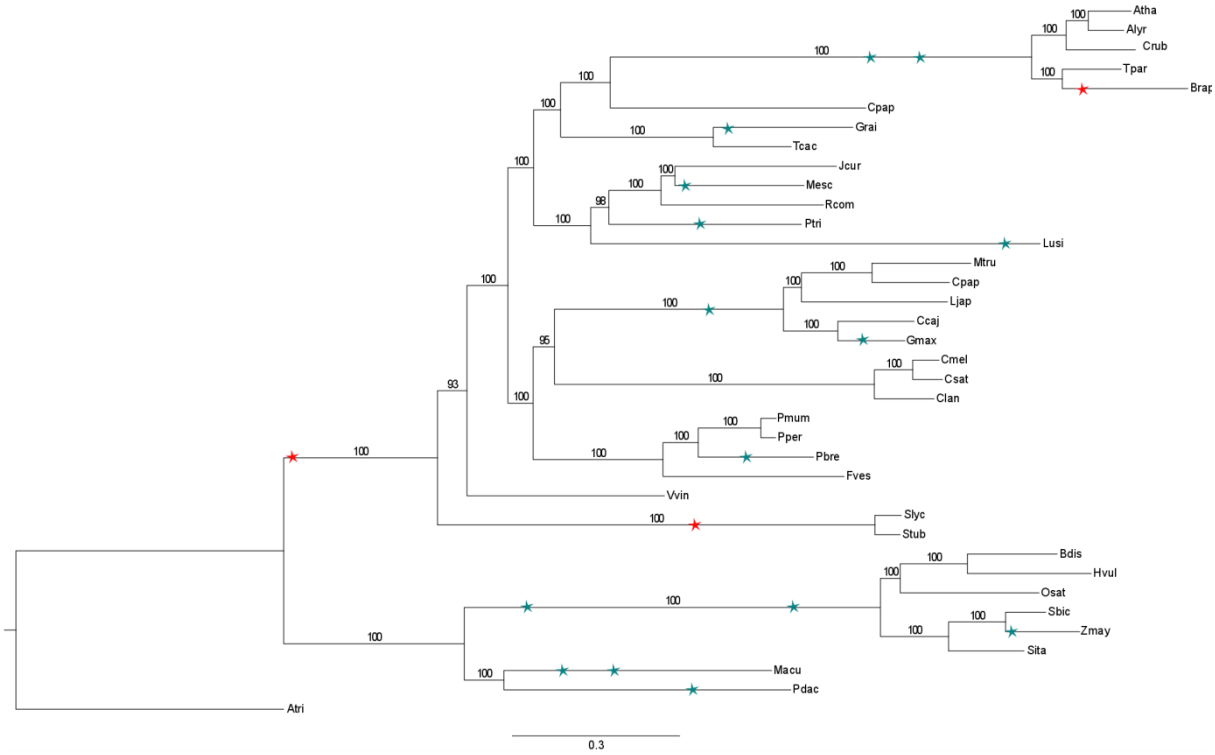
Supplemental Figure 10. Polar diagrams depicting the fraction of duplication events in each gene family group belonging to either the 'Recent', 'K-Pg Boundary', 'Ancient' or 'SSD' duplication classes. (A) Represents predictions of duplication timing for all core gene families, obtained by using gene tree – species tree reconciliation. This Figure is the same as Figure 5B. In contrast to GMM (see panel B), which provides estimates of the ages of the duplication events for each species separately, here estimates of the duplication age is based on a gene family basis and hence no averaging over species is necessary. To obtain the bar plots we normalised the absolute counts of duplication events for each node in the species tree with the number of nodes in the species tree of that duplication class, correcting for the fact that there are for instance more nodes associated to the 'SSD' duplication class. Significance values are indicated by asterisks (green = overrepresentation, red = underrepresentation) and were calculated based on the absolute counts of predicted duplications of each class, using the Fisher's exact test with Bonferroni multiple-testing correction. (B) Represents predictions of duplication timing for all core gene families based on GMM of K_S-based species-specific age distributions. We classified each duplicate pair to a certain duplication class depending on the K_S-peak it belonged to (see Supplemental Table 2). The bars in the Figures represent averages, obtained from averaging over the number of duplications assigned to a certain class for all species. Statistical significant over- and underrepresentations were calculated based on the Wilcoxon-rank-sum test and are denoted by asterisks.



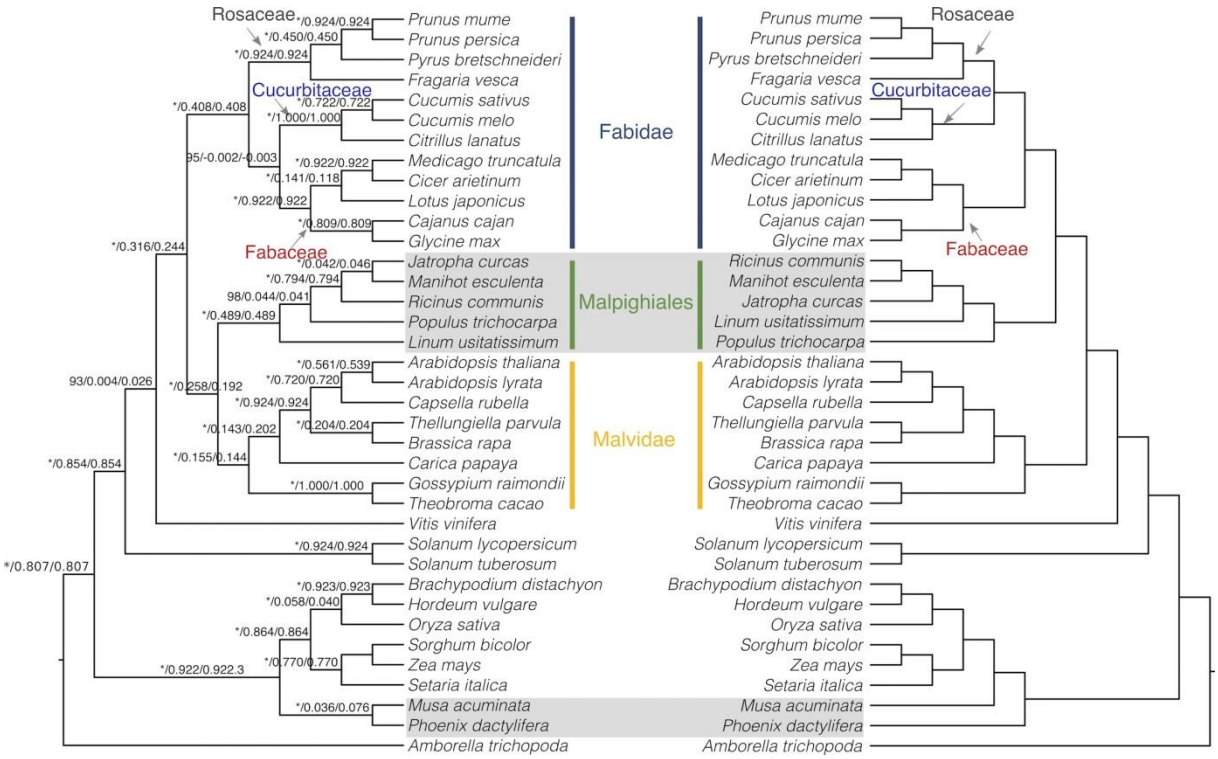
Supplemental Figure 11. Over- and underrepresentation of an independent set of 2,090 nuclear-encoded chloroplast-targeted genes obtained from The Chloroplast Function Database (Myouga et al., 2013). The y-axis represents over- (positive values) or under- (negative values) representation of these chloroplast genes in the three different functional groups as compared to the full set. In specific, to obtain the values on the y-axis we calculated the ratio of the proportion of group genes (i.e. 'Single', 'Intermediate' or 'Multi') that are chloroplast genes to the proportion of genes in the full set that are chloroplast genes. Positive values for overrepresentation (ratio > 1) and negative values for underrepresentation (ratio < 1) were obtained by subtracting one from the above described ratio. P-values as obtained by Fisher's exact test with Bonferroni multiple-testing correction are indicated on the bars.



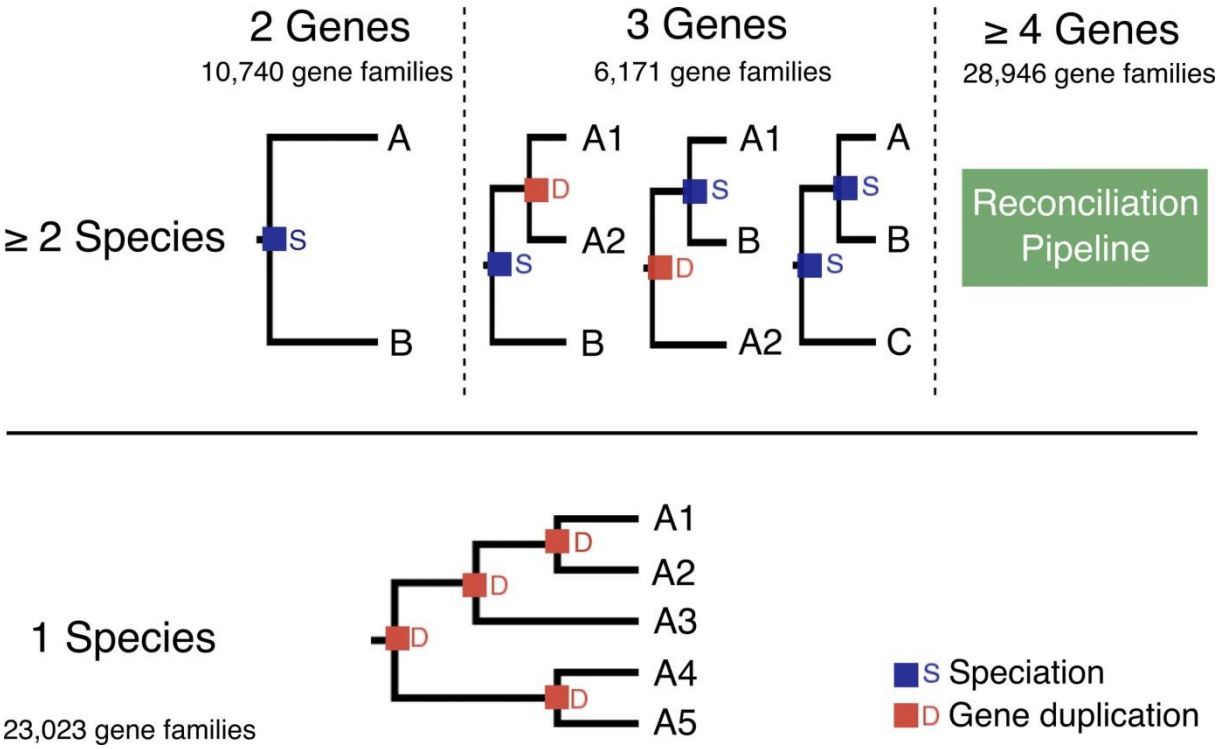
Supplemental Figure 12. Over- and underrepresentation of an independent set of 1,795 putative transcription factors, obtained from (Perez-Rodriguez et al., 2010). The y-axis represents over- (positive values) or under- (negative values) representations for transcription factor genes in the three different functional groups as compared to the full set. In specific, to obtain the values on the y-axis we calculated the ratio of the proportion of group genes (i.e. 'Single', 'Intermediate' or 'Multi') that are transcription factors to the proportion of genes in the full set that are transcription factors. Positive values for overrepresentation (ratio > 1) and negative values for underrepresentation (ratio < 1) were obtained by subtracting one from the above described ratio. P-values as obtained by Fisher's exact test with Bonferroni multiple-testing correction are indicated on the bars.



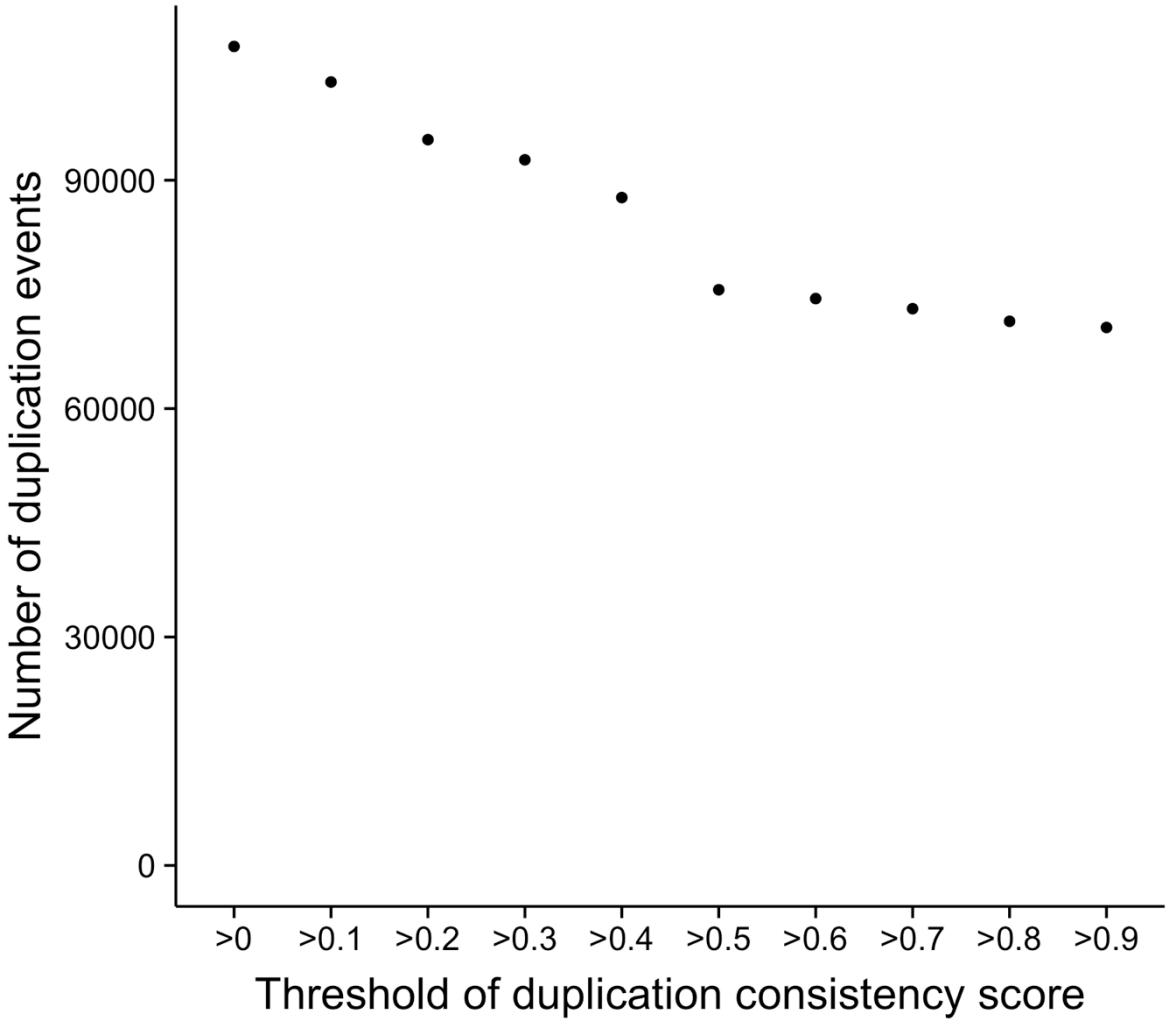
Supplemental Figure 13. Mapping of the whole-genome duplications and triplications on the species tree as obtained by the approach outlined in 'Dating whole-genome duplications' and as used for the simulations of gene family evolution according to the stochastic gene birth-death null model.



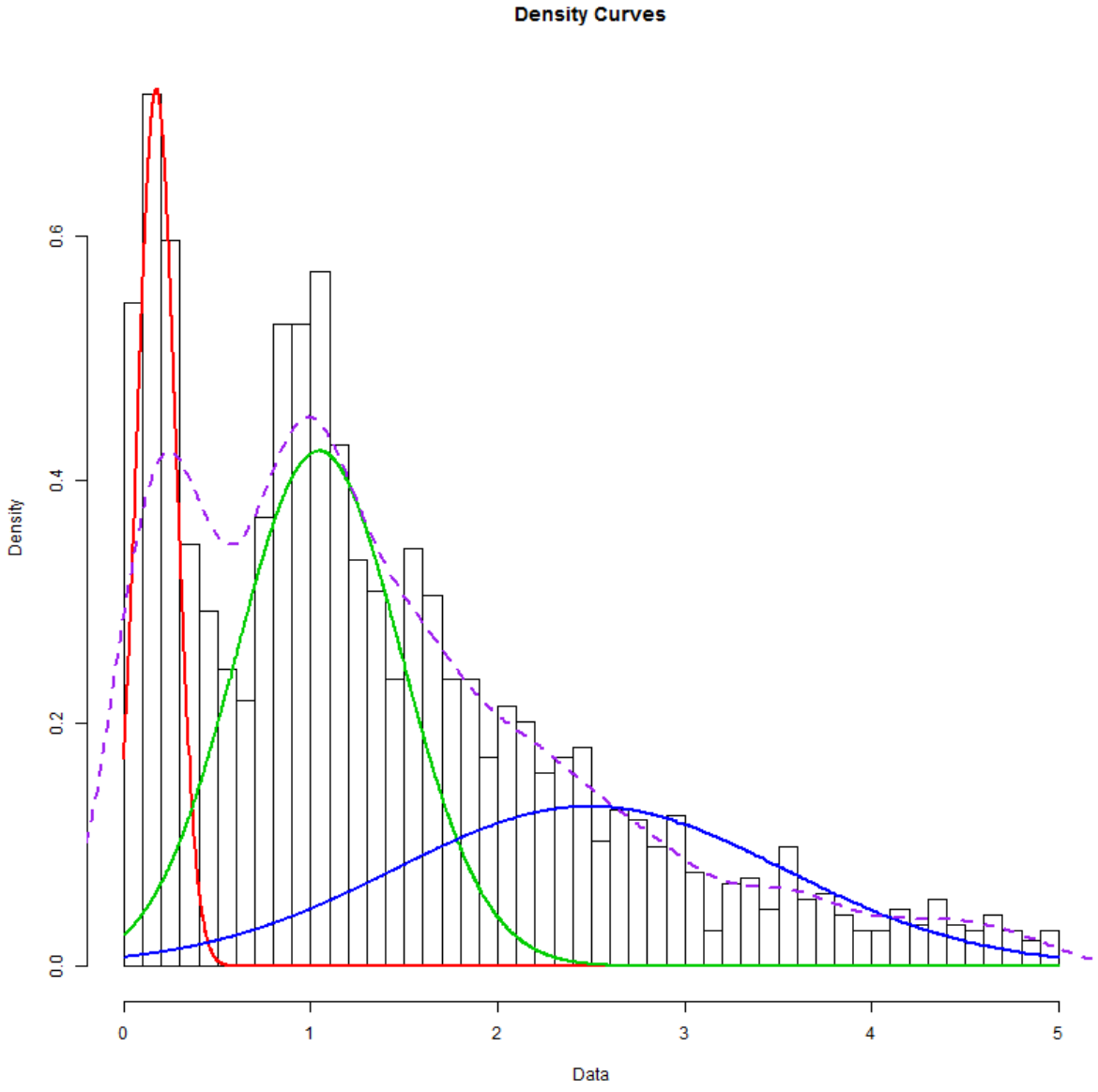
Supplemental Figure 14. Conflicting clades between the species tree used in this paper and which we inferred from 107 core gene families (left) and the APGIII tree (right). The here obtained species tree is largely consistent with the APGIII tree (Bremer et al., 2009), yet there are some conflicts. The incongruence between the positions of the Malpighiales clade in trees constructed from nuclear genes versus chloroplast genes have long been recognized, and is thought to be caused by introgressive hybridization in the ancestral lineages of Fabidae and Malvidae (Sun et al., 2015). Moreover, due to rapid diversification at the mid-Cretaceous, the relationships within Malpighiales are hard to determine (Xi et al., 2012). The close to zero values of IC and ICA suggest incongruence of the gene trees and the species tree on the branch leading to *Populus trichocarpa* and on the branch leading to *Jartropha curcas* and *Manihot esculenta*. Similarly, the monophyletic group consisting of Cucurbitaceae and Fabaceae is also only supported by half of the gene families used to reconstruct the species tree.



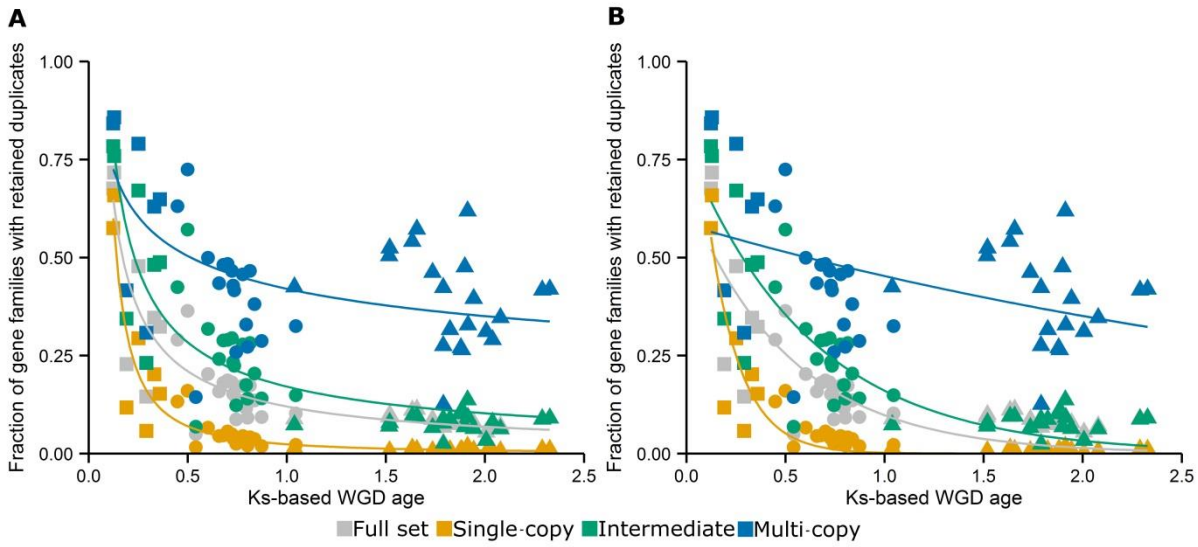
Supplemental Figure 15. Explanation of how duplications were inferred for gene families with at least two species but no more than three genes or gene families that are only present in one species. For gene families with two genes in two species (10,740 gene families), the node connecting both genes is assumed to be a speciation node. For gene families with three genes (6,171 gene families), we mid-point rerooted the gene tree and distinguished between three possible scenarios. If the three genes come from two species, the duplication occurred either in one species or in the common ancestor of the two species, depending on the topology of the gene tree. If the three genes come from three species, we assume that no duplications have occurred in the history of the gene family (most parsimonious scenario). For gene families that only cover one species (23,023) but with two genes or more, e.g. five genes in the figure, we mid-point rerooted the gene tree and considered all nodes in the tree to be duplication nodes. For the remaining 28,946 gene families with at least four genes (including all core gene families) duplications were inferred using the reconciliation pipelines as described in Materials and Methods.



Supplemental Figure 16. The change in the total number of predicted duplication events in core gene families in function of the threshold on the duplication consistency score. The predicted number of duplication events stays relatively stable for duplication consistency score thresholds up until 0.5, yet shows a drop for duplication consistency scores larger than 0.5. The large reduction at 0.5 can be explained by the large number of nodes in the species tree that only encompass two species and hence the large effect of an increase in the duplication consistency score threshold from 0.4 to 0.5 on the total number of duplication events: e.g. ((ath,aly)ath) will not make the cut of a duplication consistency score > 0.5.



Supplemental Figure 17. Gaussian mixture models were fit to the K_S-distribution of each species. Peaks were considered solid if they had a good visual fit with the density line (dashed purple line) and the K_S-histogram and had a μ lower than 3. Flat peaks, e.g. peaks which span the whole K_S - distribution, where also removed. The annotation of the peaks was done using known literature (Vanneste et al., 2014). The figure shows the K_S -distribution for *Sorghum bicolor*. The red and green peaks have a good fit to the density line whereas the flat blue peak shows no correspondence to density line and spans the whole K_S -distribution.



Supplemental Figure 18. Comparison of (A) power-law fit and (B) exponential fit to the data obtained from the Gaussian Mixture Modeling of K_S-based age distributions. The power-law shows consistently a better fit than the exponential, as assessed by Chi-squared Goodness-Of-Fit test (see Supplemental Table 3).

SUPPLEMENTAL TABLES

Supplemental Table 1. Comparison of the numbers of interacting protein pairs in each group to those obtained from randomized networks.

	Number of PPIs within group	Average number of PPIs within group for 1000 randomized networks	Z-score	P-value enrichment of PPI vs random (one-sided test)	P-value with multiple-testing correction (Bonferroni)
Full	15949	15949			
Single-copy	2550	2813.012	-1.005	0.84	1
Intermediate	2277	1740.331	2.710	0.0034	0.010
Multi-copy	1034	990.558	0.322	0.374	1

Supplemental Table 2. Description of all identified peaks inferred from the K_S-based age distributions.

Species	*k*	μ	σ	λ	L_bound	H_bound	Annotation	WGD type	Included
Alyr1	4	0.095	0.086	0.131	0.000	0.289	SSD	SSD	NO
Alyr2	4	0.723	0.258	0.579	0.289	1.199	BRAalpha	KT	YES
Alyr3	4	2.038	0.720	0.227	1.199	2.970	BRABeta	OLD	NO
Alyr4	4	3.848	0.631	0.063	2.970	5.000	HighKS	HighKS	NO
Atha1	4	0.178	0.122	0.088	0.000	0.411	SSD	SSD	NO
Atha2	4	0.778	0.243	0.574	0.411	1.231	BRAalpha	KT	YES
Atha3	4	2.059	0.783	0.286	1.231	3.185	BRABeta	OLD	NO
Atha4	4	4.083	0.533	0.052	3.185	5.000	HighKS	HighKS	NO
Bdis1	4	0.182	0.108	0.144	0.000	0.400	SSD	SSD	NO
Bdis2	4	0.802	0.263	0.374	0.400	1.240	MON1	KT	YES
Bdis3	4	1.878	0.613	0.383	1.240	2.762	MON2	OLD	YES
Bdis4	4	3.688	0.671	0.100	2.762	5.000	HighKS	HighKS	NO
Brap1	3	0.331	0.082	0.513	0.000	0.479	REC	REC	YES
Brap2	3	0.701	0.340	0.334	0.479	1.292	BRAalpha	KT	YES
Brap3	3	2.220	1.025	0.153	1.292	5.000	BRABeta	OLD	NO
Cari1	4	0.047	0.039	0.118	0.000	0.155	SSD	SSD	NO
Cari2	4	0.735	0.316	0.543	0.155	1.273	LEG	KT	YES
Cari3	4	2.078	0.725	0.277	1.273	3.064	DIC	OLD	YES
Cari4	4	3.945	0.581	0.063	3.064	5.000	HighKS	HighKS	NO
Ccaj1	4	0.032	0.037	0.100	0.000	0.138	SSD	SSD	NO
Ccaj2	4	0.602	0.214	0.569	0.138	1.009	LEG	KT	YES
Ccaj3	4	1.789	0.679	0.279	1.009	2.794	DIC	OLD	YES
Ccaj4	4	3.746	0.617	0.052	2.794	5.000	HighKS	HighKS	NO
Clan1	3	0.2643	0.1755	0.2239	0.0000	0.6731	SSD	SSD	NO
Clan2	3	1.8231	0.7083	0.6317	0.6731	2.7961	DIC	OLD	YES
Clan3	3	3.7459	0.6738	0.1444	2.7961	5.0000	HighKS	HighKS	NO
Cmel1	3	0.2786	0.2019	0.1872	0.0000	0.7310	SSD	SSD	NO
Cmel2	3	1.9139	0.7743	0.6712	0.7310	2.9552	DIC	OLD	YES
Cmel3	3	3.8984	0.6355	0.1416	2.9552	5.0000	HighKS	HighKS	NO
Cpap1	3	0.249	0.202	0.306	0.000	0.765	SSD	SSD	NO
Cpap2	3	2.006	0.595	0.602	0.765	2.995	DIC	OLD	YES
Cpap3	3	3.897	0.517	0.092	2.995	5.000	HighKS	HighKS	NO
Crub1	4	0.124	0.075	0.070	0.000	0.308	SSD	SSD	NO
Crub2	4	0.814	0.273	0.593	0.308	1.289	BRAalpha	KT	YES
Crub3	4	2.039	0.724	0.263	1.289	3.027	BRABeta	OLD	NO
Crub4	4	3.907	0.580	0.075	3.027	5.000	HighKS	HighKS	NO
Csat1	3	0.318	0.216	0.192	0.000	0.777	SSD	SSD	NO
Csat2	3	1.789	0.680	0.580	0.777	2.596	DIC	OLD	YES
Csat3	3	3.425	0.773	0.228	2.596	5.000	HighKS	HighKS	NO
Fves1	3	0.334	0.222	0.365	0.000	0.791	SSD	SSD	NO
Fves2	3	1.735	0.658	0.552	0.791	2.631	DIC	OLD	YES
Fves3	3	3.543	0.685	0.083	2.631	5.000	HighKS	HighKS	NO
Gmax1	3	0.124	0.044	0.622	0.000	0.216	REC	REC	YES
Gmax2	3	0.448	0.208	0.261	0.216	0.872	LEG	KT	YES
Gmax3	3	1.868	0.967	0.117	0.872	5.000	DIC	OLD	NO

Supplemental Table 2. Description of all identified peaks inferred from the K_S-based age distributions.

Species	*k*	μ	σ	λ	L_bound	H_bound	Annotation	WGD type	Included
Grai1	3	0.048	0.037	0.058	0.000	0.149	SSD	SSD	NO
Grai2	3	0.499	0.166	0.703	0.149	0.858	KT	KT	YES
Grai3	3	1.912	0.964	0.239	0.858	5.000	DIC	OLD	YES
Hvul1	3	0.011	0.010	0.115	0.000	0.042	SSD	SSD	NO
Hvul2	3	0.639	0.416	0.487	0.042	1.312	MON1	KT	NO
Hvul3	3	2.217	1.092	0.398	1.312	5.000	MON2	OLD	NO
Jcur1	3	0.120	0.116	0.274	0.000	0.432	SSD	SSD	NO
Jcur2	3	1.943	0.831	0.669	0.432	3.377	DIC	OLD	YES
Jcur3	3	4.271	0.432	0.057	3.377	5.000	HighKS	HighKS	NO
Ljap1	4	0.051	0.058	0.144	0.000	0.180	SSD	SSD	NO
Ljap2	4	0.541	0.268	0.490	0.180	1.018	LEG	KT	YES
Ljap3	4	1.790	0.655	0.252	1.018	2.634	DIC	OLD	YES
Ljap4	4	3.491	0.682	0.114	2.634	5.000	HighKS	HighKS	NO
Lusi1	3	0.128	0.056	0.726	0.000	0.249	REC	REC	YES
Lusi2	3	0.588	0.303	0.190	0.249	1.163	DIC	OLD	NO
Lusi3	3	2.265	1.025	0.084	1.163	5.000	HighKS	HighKS	NO
Macu1	5	0.075	0.039	0.021	0.000	0.198	SSD	SSD	NO
Macu2	5	0.435	0.081	0.326	0.198	0.556	MAC	KT	NO
Macu3	5	0.672	0.211	0.398	0.556	0.937	MAC	KT	NO
Macu4	5	1.158	0.398	0.220	0.937	1.782	MAC	OLD	NO
Macu5	5	2.538	1.049	0.036	1.782	5.000	HighKS	HighKS	NO
Mesc1	4	0.071	0.040	0.044	0.000	0.171	SSD	SSD	NO
Mesc2	4	0.359	0.086	0.671	0.171	0.580	REC	REC	YES
Mesc3	4	1.633	0.664	0.251	0.580	2.667	DIC	OLD	YES
Mesc4	4	3.717	0.681	0.034	2.667	5.000	HighKS	HighKS	NO
Mtru1	3	0.159	0.122	0.342	0.000	0.379	SSD	SSD	NO
Mtru2	3	0.744	0.324	0.414	0.379	1.330	LEG	KT	YES
Mtru3	3	2.338	1.063	0.244	1.330	5.000	DIC	OLD	NO
Osat1	4	0.143	0.114	0.197	0.000	0.396	SSD	SSD	NO
Osat2	4	0.873	0.266	0.356	0.396	1.300	MON1	KT	YES
Osat3	4	1.884	0.598	0.365	1.300	2.829	MON2	OLD	YES
Osat4	4	3.779	0.602	0.082	2.829	5.000	HighKS	HighKS	NO
Pbre1	3	0.010	0.010	0.290	0.000	0.038	SSD	SSD	NO
Pbre2	3	0.168	0.071	0.550	0.038	0.353	REC	REC	NO
Pbre3	3	1.564	0.950	0.160	0.353	5.000	DIC	OLD	NO
Pdac1	3	0.291	0.078	0.548	0.100	0.440	REC	KT	YES
Pdac2	3	0.706	0.375	0.394	0.440	1.354	?	?	NO
Pdac3	3	2.350	1.130	0.057	1.354	5.000	?	?	NO
Pmum1	3	0.167	0.150	0.418	0.000	0.534	SSD	SSD	NO
Pmum2	3	1.516	0.522	0.488	0.577	2.185	DIC	OLD	YES
Pmum3	3	2.813	0.957	0.094	2.162	5.000	HighKS	HighKS	NO
Pper1	3	0.194	0.153	0.391	0.000	0.571	SSD	SSD	NO
Pper2	3	1.519	0.488	0.519	0.571	2.189	DIC	OLD	YES
Pper3	3	2.894	0.946	0.089	2.189	5.000	HighKS	HighKS	NO
Ptri1	3	0.028	0.020	0.072	0.000	0.085	SSD	SSD	NO

Supplemental Table 2. Description of all identified peaks inferred from the K_S-based age distributions.

Species	*k*	μ	σ	λ	L_bound	H_bound	Annotation	WGD type	Included
Ptri2	3	0.251	0.067	0.719	0.085	0.428	REC	REC	YES
Ptri3	3	1.632	0.940	0.209	0.428	5.000	DIC	OLD	NO
Rcom1	3	0.278	0.197	0.186	0.000	0.736	SSD	SSD	NO
Rcom2	3	1.898	0.685	0.741	0.736	3.130	DIC	OLD	YES
Rcom3	3	4.087	0.483	0.073	3.130	5.000	HighKS	HighKS	NO
Sbic1	3	0.175	0.103	0.187	0.000	0.406	SSD	SSD	NO
Sbic2	3	1.045	0.442	0.469	0.406	1.711	MON1	KT	YES
Sbic3	3	2.490	1.045	0.344	1.711	5.000	MON2	OLD	NO
Sita1	3	0.079	0.062	0.126	0.000	0.231	SSD	SSD	NO
Sita2	3	0.837	0.398	0.490	0.231	1.461	MON1	KT	YES
Sita3	3	2.233	1.027	0.384	1.461	5.000	MON2	OLD	YESB
Slyc1	3	0.184	0.094	0.125	0.000	0.375	SSD	SSD	NO
Slyc2	3	0.729	0.228	0.541	0.375	1.197	SOL	KT	YES
Slyc3	3	2.327	1.068	0.334	1.197	5.000	DIC	OLD	YES
Stub1	3	0.118	0.085	0.212	0.000	0.300	SSD	SSD	NO
Stub2	3	0.658	0.223	0.501	0.300	1.121	SOL	KT	YES
Stub3	3	2.289	1.071	0.286	1.121	5.000	DIC	OLD	YES
Tcac1	3	0.128	0.061	0.142	0.000	0.311	SSD	SSD	NO
Tcac2	3	1.656	0.663	0.787	0.311	2.802	DIC	OLD	YES
Tcac3	3	3.874	0.600	0.071	2.802	5.000	HighKS	HighKS	NO
Tpar1	3	0.680	0.356	0.707	0.000	1.309	BRAalpha	KT	YES
Tpar2	3	2.140	0.555	0.211	1.309	2.959	BRABeta	OLD	NO
Tpar3	3	3.835	0.632	0.082	2.959	5.000	HighKS	HighKS	NO
Vvin1	3	0.088	0.067	0.292	0.000	0.258	SSD	SSD	NO
Vvin2	3	1.038	0.494	0.611	0.258	1.767	DIC	OLD	YES
Vvin3	3	2.608	1.089	0.097	1.767	5.000	HighKS	HighKS	NO
Zmay1	3	0.191	0.104	0.532	0.000	0.392	REC	REC	YES
Zmay2	3	0.795	0.394	0.226	0.392	1.426	MON1	KT	YES
Zmay3	3	2.248	1.036	0.242	1.426	5.000	MON2	OLD	NO

Each row in the table represents one peak: ***k*** denotes the number of components that were fitted; μ, σ and λ are the obtained parameters for fitted GMMs; L_bound and U_bound represent respectively the lower- and upperbound K_S-values associated with each peak; Annotation represents the annotation of the peak based on data from (Vanneste et al., 2014); WGD types is the classification of the peak as either 'SSD', 'Recent' (REC), 'K-Pg Boundary' (KT), 'Ancient' (OLD) or 'HighKS' if they had μ-values exceeding 3.5; 'Included' indicates whether we used the peak data to create Figure 3 and Figure 5B.

Supplemental Table 3. Comparison of the power-law and the exponential fit.

	χ^2-goodness-of-fit (p-value)	
	Power-law	Exponential
Full	0.76795 (p =1)	5.072 (p=1)
Single-copy	0.52465 (p = 1)	477.6 (p < 2.2e-16)
Intermediate	1.3838 (p = 1)	2.0733 (p = 1)
Multi-copy	1.8271 (p = 1)	2.1274 (p = 1)

REFERENCES

Bremer, B., Bremer, K., Chase, M.W., Fay, M.F., Reveal, J.L., Soltis, D.E., Soltis, P.S., Stevens, P.F., Anderberg, A.A., Moore, M.J., Olmstead, R.G., Rudall, P.J., Sytsma, K.J., Tank, D.C., Wurdack, K., Xiang, J.Q.Y., Zmarzty, S., and Grp, A.P. (2009). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Bot J Linn Soc **161,** 105-121.

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Mach Learn **52,** 91-118.

Myouga, F., Akiyama, K., Tomonaga, Y., Kato, A., Sato, Y., Kobayashi, M., Nagata, N., Sakurai, T., and Shinozaki, K. (2013). The Chloroplast Function Database II: A Comprehensive Collection of Homozygous Mutants and Their Phenotypic/Genotypic Traits for Nuclear-Encoded Chloroplast Proteins. Plant Cell Physiol **54,** E2-+.

Perez-Rodriguez, P., Riano-Pachon, D.M., Correa, L.G., Rensing, S.A., Kersten, B., and Mueller-Roeber, B. (2010). PlnTFDB: updated content and new features of the plant transcription factor database. Nucleic Acids Res **38,** D822-827.

Sun, M., Soltis, D.E., Soltis, P.S., Zhu, X., Burleigh, J.G., and Chen, Z. (2015). Deep phylogenetic incongruence in the angiosperm clade Rosidae. Molecular phylogenetics and evolution **83,** 156-166.

Vanneste, K., Baele, G., Maere, S., and 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 Res **24,** 1334-1347.

Xi, Z., Ruhfel, B.R., Schaefer, H., Amorim, A.M., Sugumaran, M., Wurdack, K.J., Endress, P.K., Matthews, M.L., Stevens, P.F., Mathews, S., and Davis, C.C. (2012). Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. Proc Natl Acad Sci U S A **109,** 17519-17524.