Supporting Information

Fawcett et al. 10.1073/pnas.0900906106

SI Text

Estimating the Age of Gene Duplication Events: The Use of K_S Values. One of the most common methods used to study and visualize gene duplication events in eukaryotic genomes is to build age distributions of paralogs, where the number of duplicates is plotted against their age (Fig. S1). The age of a duplication event is usually inferred from the number of synonymous substitutions per synonymous site (K_S). Peaks in the K_S distribution reflect sudden bursts in the number of new genes and are therefore considered evidence for large-scale gene or entire genome duplications (Fig. S1). If the rate of synonymous substitutions is known, one can convert the K_S values to absolute ages. However, estimates of synonymous substitution rates can vary considerably. For instance, Koch et al. (1) obtained a synonymous substitution rate of 1.5×10^{-8} synonymous substitutions per year for *Arabidopsis* and related species based on the divergence of the loci for chalcone synthase and alcohol dehydrogenase loci. Jakobsson et al. (2), using many more nuclear markers, obtained a synonymous substitution rate of 6.0×10^{-9} for *Arabid opsis* which is much closer to earlier estimates. Lynch and Conery (3) assumed a rate of 6.1×10^{-9} synonymous substitutions per year which was the average of 2 surveys based on analyses of multiple genes in vascular plants. Using this latter substitution rate, they dated the youngest genome duplication in *Arabidopsis* at 65 mya, whereas Simillion et al. (4) arrived at ≈75 mya - the discrepancy is due to the difference in the peak K_S values used to date the whole genome duplication (WGD) event (0.8 vs. 0.91). Blanc and Wolfe (5), assuming the much faster synonymous rate of 1.5×10^{-9} synonymous substitutions per year for *Arabidopsis*, dated the youngest WGD in *Arabidopsis* much younger, at 25–26.7 mya. Other authors have proposed K_S rates for other plants as well. For instance, Gaut et al. (6) proposed a synonymous rate of 6.5×10^{-9} synonymous substitutions per year for the grasses, and Lescot et al. (7) proposed an average rate of 4.3×10^{-9} synonymous substitutions per year for the Musaceae. The synonymous rate for actin genes in Solanaceae has been estimated at 6.96×10^{-9} substitutions per site per year (8). Although many synonymous rates seem to be quite similar for different species, one has to be cautious in applying such rates for dating purposes. For example, in the case of the perennial species *Populus* (poplar), a rate of 6.0×10^{-9} synonymous substitutions per year suggests 13 mya for the age of the WGD (9). It was later suggested that the substitution rate in *Populus* is much slower (≈6 times) compared to other species such as *Arabidopsis*, and that the WGD event probably shortly predates the split of *Populus* and *Salix*, which is estimated to be around 60 mya (10) (see main manuscript and below).

Estimating the Age of Gene Duplication Events: The Use of Phylogenetics. We sought to provide more accurate estimates of the absolute ages of WGDs in plants by estimating the divergence dates of all WGD-derived paralogs through phylogenetic means. Estimating the divergence or duplication time of sequences in a phylogenetic tree has been an important topic for many years and various methods have been developed to account for the rate variations across branches. Here, we used the penalized likelihood (PL) method (11), implemented in the r8s package (12). This method accounts for rate variation between lineages by using a semiparametric smoothing approach that penalizes rates that vary too much across a phylogeny, based on an optimal smoothing value that can be obtained by a cross-validation procedure (see below). This method was chosen first because it is one of the most commonly used methods in phylogenetic dating (13, 14), and second because it seemed to be the most suitable for processing a large number of trees in a script-wise manner. For instance, various dating methods based on Bayesian models have been developed (15–17), and these have been suggested to be pretty robust to rate variation as they do not assume an autocorrelated rate of molecular evolution. Although we also considered these approaches, the major obstacle was that it turned out to be difficult to automate these procedures to analyze a large number of datasets [i.e., BEAST (17) is based on a graphical user interface, and the outputs of mcmctree (18) or multidivtime (19) are not so easy to parse]. Further development in this area might allow us to compare the dates based on the PL methods with those calculated with Bayesian methods.

Sequence Data Sets Used in the Current Study. Whole-genome sequences and the annotation of protein-coding genes for the following genomes were used: *Arabidopsis thaliana* from the TAIR7 release (http://www.arabidopsis.org), *Populus trichocarpa* assembly from JGI (http://genome.jgi-psf.org/Poptr1/Poptr1.home.html), the Mt1.0 release from the Medicago Genome Sequencing Consortium (http://mips.gsf.de/proj/plant/jsf/medi/index.jsp) for *Medicago truncatula*, the *Vitis vinifera* assembly from Velasco et al. (20) (http://genomics.research.iasma.it), *Oryza sativa* subsp. *japonica* chromosome pseudomolecule version 4 from TIGR (http://rice.plantbiology.msu.edu), *Physcomitrella patens* version 1.1 assembly from JGI (http://genome.jgi-psf.org/Phypa1_1/Phypa1_1.home.html), and *Chlamydomonas reinhardtii* version 3.1 assembly from the JGI (http://genome.jgi-psf.org/Chlre3/Chlre3.home.html). For plants with no publicly available genome assembly, which were *Gossypium hirsutum*, *Solanum lycopersicum*, *Manihot esculenta*, *Lactuca sativa*, *Eschscholzia californica*, and *Acorus americanus*, EST clusters were downloaded from the TIGR Plant Transcript Assemblies (http://plantta.tigr.org).

Identification of Paralogs Created by Large-Scale Duplication Events. First, to identify paralogous gene pairs in each species that were most likely created by the WGD, an all-against-all BLASTP was performed by using the whole protein dataset of each species. Paralogous gene pairs were retained if the 2 sequences were alignable over a length of more than 150 amino acids with an identity score of >30% (21). For those species with an available genome assembly and that have undergone a recent WGD (*A. thaliana*, *P. trichocarpa*, *M. truncatula*, and *O. sativa*), paralogs were used to detect duplicated segments by running i-ADHoRe version 2.0 (22) with the gap size set to 40 genes, the minimum number of paralogs (anchors) to define a duplicated segment to 4, and the *P* value cutoff to 0.001. Because duplicated segments reported by i-ADHoRe include segments that are not derived from the most recent WGD, such as segments from older WGDs (23) or more recent small-scale segmental duplications, the mean K_S of each duplicated segment was calculated to filter out segments that were not created by the WGD of interest. The K_S value was calculated for each paralogous gene pair (anchor) reported by i-ADHoRe by using CODEML from the PAML package (24). The K_S with the highest likelihood score from 10 runs was used to infer the mean K_S of each duplicated segment (4). Paralogs lying in duplicated segments with a mean K_S of 0.6–0.9 for *A. thaliana*, 0.1–0.4 for *P. trichocarpa*, and 0.6–1.1 for *M. truncatula* and *O. sativa* were retained for the dating proce-

dure as pairs that are likely to be created by the most recent WGD.

For species with no available genome assembly but for which a considerable number of EST sequences exist (*G. hirsutum*, *S. lycopersicum*, *L. sativa*, *E. californica*, and *A. americanus*), amino acid sequences were obtained from each EST cluster using FrameD (25). Only those of more than 50 amino acids were retained. Paralogous gene pairs were identified as described above and their K_S were calculated. All of these species show a recent peak in their K_S distribution, consistent with the occurrence of a WGD (26, 27). Gene pairs with a K_S between 0.2 and 1.0 were considered candidates for having a WGD origin. However, multiple pairs with $0.2 \leq K_S \leq 1.0$ may have originated from the same WGD pair through the occurrence of subsequent duplications. To correct for this redundancy, paralogs were clustered with an average linkage hierarchical clustering algorithm using K_S as a distance measure (28). For the inferred duplication events with an average K_S between 0.2 and 1.0 (0.4–1.0 for *L. sativa* as a K_S of 0.2–0.4 still contained a fair amount of background duplication), a random pair of descendant sequences was chosen as a representative (Fig. S2). Only these representative non-redundant pairs were used for further analysis. The same approach was taken for *P. patens* even though the genome assembly was available, because starting from all paralogous gene pairs identified within duplicated blocks rendered too few pairs (anchors) to obtain a reliable age distribution.

Construction of Orthologous Gene Families and Phylogenetic Trees. After identification of paralogs that were likely created by WGD events, genes from other plant species that were orthologous to each paralog were collected to build phylogenetic trees to date the duplication events. For each species that we wanted to estimate the age of the WGD, Inparanoid (29) was run against each of the following species (apart from the species being dated): *A. thaliana*, *P. trichocarpa*, *M. truncatula*, *V. vinifera*, and *O. sativa*, by using Inparanoid (29). One orthologous gene from each species was added to each paralogous pair. *P. patens* (moss) was included when an extra outgroup was required. An orthologous gene family was created for each paralogous pair. The amino acid sequences were aligned for each family using CLUSTALW (30). The alignments were cleaned up by removing ambiguously aligned sites as previously outlined (31), and only those gene families with an alignment of more than 100 aa were retained for the construction of phylogenetic trees and dating. A 100-bootstrapped sequence alignments were created by using the SEQBOOT program of the PHYLIP package (http://evolution.genetics.washington.edu/phylip.html) for each remaining family. The maximum-likelihood branch lengths were calculated by using PhyML (32) for each of the 100 inferred trees, thereby generating 100 replicates with (slightly) different branch lengths for each family. The tree topology was fixed according to the commonly accepted species phylogeny as shown in Fig. 3 (33, 34). The paralogous genes of the query species were fixed to cluster together, as each WGD investigated here is thought to have occurred independently in each lineage (10, 26, 27, 35, 36). *Oryza* was used as outgroup when dating the WGD of a eudicot species. *Physcomitrella* was included as outgroup when dating the WGD of monocots. For each gene family, the branch lengths in the phylogenetic trees were computed by PhyML (32), and the age of the node connecting the 2 paralogs was estimated based on the branch lengths using the PL method (11) implemented in the r8s package version 1.7 (12).

In this study, we chose to build gene families and phylogenetic trees first by identifying orthologs of each species by using Inparanoid (29). We fixed the topology according to the accepted species phylogeny assuming that a correct ortholog had been identified by Inparanoid. If, for instance, for a paralogous pair of gene A and gene B, the ortholog (for any of the species) of gene A was different from the ortholog of gene B, such pairs were not retained. If gene A and gene B had the same but more than 1 ortholog in the same species, the "main ortholog" as reported by Inparanoid was selected. Only duplicates for which their orthologs could be identified in all of the species were retained for the next step. Thus, each family contained 1 gene each of *A. thaliana*, *P. trichocarpa*, *M. truncatula*, *V. vinifera*, *O. sativa*, and *P. patens* when required as an outgroup, plus the paralog of the species duplications that were to be dated. This rather stringent approach was adopted to minimize possible errors in tree topologies or inclusion of paralogs rather than orthologs, and to ensure that all duplicates from the same species were dated under the same conditions in terms of taxon sampling in order to avoid any potential bias (i.e., duplicates in families with more/fewer taxa or with/without a certain taxon might have younger/older ages).

It should be noted that the phylogenetic position of *Vitis* has received attention recently as Jaillon et al. (37) assumed a common ancestry of *Vitis* and *Populus*, and *Arabidopsis* showing a sister group relationship to *Vitis* and *Populus*. However, this position of *Vitis* is likely to reflect the slower evolutionary rate of *Populus* and *Vitis* compared to *Arabidopsis*, rather than their true evolutionary relationship. To our knowledge, almost all recent large-scale phylogenetic studies (33, 34, 38, 39) have placed *Vitis* as an early-diverging rosid. As an outgroup, we used *Oryza* when dating the WGD of a eudicot species, and *Physcomitrella* was included as an outgroup when dating the WGD of monocots. This was because we fixed the age of a parental node of the paralogs and used this as a calibration point when estimating the age of the paralogs. The node uniting the outgroup and the rest cannot be used as a calibration point as there are no means of accurately dissecting the single branch of the outgroup into 2 branches. Because an outgroup is required to compute PL values when using r8s (12), we overcame this problem by arbitrarily splitting the branch of the outgroup into a branch length of 0.01 and the remaining branch length. However, this node could not be used as a calibration point, and, thus, an extra taxon that is sister to the taxa whose most recent common ancestor was used as a calibration point was always required. For the same reason, *Chlamydomonas* was included and was used as an outgroup when dating the WGD of *Physcomitrella*. Use of *Physcomitrella*, instead of *Oryza*, as an outgroup when dating the WGDs of eudicots was also an option. We opted not to do so because of its larger phylogenetic distance (and its effect on estimating branch lengths), and this would largely reduce the number of paralogs that could be dated because *Physcomitrella* has fewer orthologs of eudicot genes than *Oryza*.

In principle, an alternative approach could have been adopted based on building gene families starting from a dataset including all proteins from all of the species under consideration, followed by the construction of phylogenetic trees for each gene family, and estimating the divergence date of each node corresponding to a duplication event. However, this would rely on the construction of phylogenetic trees to identify orthologs. This approach turned out to be problematic (and difficult to automate) because incorrect species tree topologies can be inferred, making it especially difficult to assign age calibrations or constraints to certain nodes. For instance, we observed that sequences of *Populus* and *Vitis* often clustered together with high bootstrap values when phylogenetic trees were constructed without fixing the topology, even though *Vitis* is most probably sister to other rosids such as *Medicago*, *Arabidopsis*, and *Populus*, as discussed earlier. One possibility is that this topology is due to erroneous assumptions of orthologs and paralogs, which is possible as these species share a hexaploidization event followed by gene loss. However, suppose that despite the correct species phylogeny being (((*Populus*, *Medicago*), *Arabidopsis*), *Vitis*), we get a tree of (((*Vitis*, *Populus*), *Medicago*), *Arabidopsis*), which is what we

most frequently obtained when building trees with these species without any topology constraint, even with gene families including all orthologs and paralogs. To assume that this tree reflects a correct evolutionary scenario would require the true orthologs of *Arabidopsis* and *Medicago* to *Vitis*/*Populus* to have been lost. Although there may be such cases, it is unlikely that the majority of the genes in *Arabidopsis* and *Medicago* that are orthologous to *Vitis* and *Populus* are lost. Tang et al. (40) showed that the sequences of *Vitis* and *Populus* evolve much slower than sequences of *Arabidopsis*, and thus we believe that it is more likely that most of these topologies are due to genes of *Vitis* and *Populus* evolving slower than genes of weeds such as *Arabidopsis*, causing artificial clustering of both species (see main text and ref. 4). This would result in either having to calculate the dates based on an incorrect species topology, which will make it difficult to assign calibrations and constraints to the nodes, or remove trees/nodes with incorrect species topology, which would result in removing a lot of trees and/or having many trees with very few species. Therefore, we opted to fix the topology with the orthologs collected by running Inparanoid (29). It must be noted that wrongly assigning paralogs as orthologs may lead to the overestimation of the ages of some divergence points, and consequently overestimation of the ages of the duplication nodes. However, as we took the mode of the distribution as the age of the WGD (see *Methods*), which is less sensitive to bias caused by the outliers than the mean, this is unlikely to have a large effect on the estimated ages of the WGDs.

Estimating the Duplication Date Using the PL Method. For each gene family, the age of the node connecting the 2 paralogs in the phylogenetic tree was estimated based on the branch lengths, computed by PhyML (32), using the PL method (11) implemented in the r8s package ver. 1.7 (12). PERL scripts from Torsten Eriksson's software package (http://www.bergianska.se/index_forskning.php) as recommended in the r8s manual, with some modifications, were used to assist the process of inferring duplication dates. First, the cross validation in the r8s package was performed for each of the 100-bootstrapped replicates for each family to obtain the optimum smoothing values. This procedure removes each terminal branch, estimates the remaining parameters without that branch, and predicts the length of the removed branch. Smoothing values ranging from 10^{-3} to $10^{3.5}$ in increments of $10^{0.5}$ were tested and the value giving the best score was used as the smoothing value for that replicate. To minimize the chance that the optimal smoothing value falls out of the tested range, if the reported optimal smoothing value for more than 30 out of the 100 replicates were 10^{-3} or $10^{-2.5}$, a second run testing ranges from 10^{-6} to $10^{3.5}$ in increments of $10^{0.5}$ was performed, and if more than 30 were $10^{3.5}$ or 10^{3}, a second run testing ranges from 10^{-3} to 10^{6} in increments of $10^{0.5}$ was performed. The cross validation procedure failed for some replicates, because of zero-length terminal branches or too extreme rate variations. For each family where a smoothing value could not be obtained for more than 30 replicates, a second run was performed testing ranges from 10^{-4} to $10^{4.5}$ in increments of $10^{0.5}$. Families where a smoothing value could not be obtained for more than 30 replicates in both runs were not retained for further analyses.

Calibrations and Constraints. A combination of different calibration points (dates of divergences or speciation events) and minimum age constraints primarily based on fossil data were used to estimate the age of the duplicated genes (Table S1). We always chose a parental node of the paralogs as a calibration point with a fixed age. This was the node uniting *Vitis* and the remaining rosids or the node uniting *Arabidopsis* and *Populus*/*Medicago* for dating the WGDs of rosids, asterids (*Solanum* or *Lactuca*) and rosids when dating the WGD of asterids, *Eschscholzia* and the remaining eudicots when dating the WGD of *Eschscholzia*, and monocots and eudicots when dating the WGDs of monocots. The age of the eudicots (about 125 mya) is considered to be one of the most reliable fossil dates, and has been used as a fixed calibration point for many molecular dating analyses (41, 42). In particular, this is based on several reports of tricolpate pollen fossils, a unique trait of the eudicot clade, that had not been identified before this time. It is also suggested that many of the major lineages of eudicots, including those of *Arabidopsis*, *Populus*, *Medicago*, *Vitis*, *Solanum*, and *Lactuca* all diverged within a short time frame, before 90 mya (43, 44). When dating the WGDs of eudicots, the calibration points were fixed at ages that were within this range and were consistent with the range of ages suggested by various molecular dating analyses (38, 41, 45–47). These were: 115 mya for the node uniting *Vitis* and the remaining rosids, 120 mya for the node uniting asterids and rosids, and 125 mya for the node uniting *Eschscholzia* and the remaining eudicots. Although many early studies, especially when using molecular clocks without taking into account the rate variation, estimated very old ages for the origins of angiosperms and the divergence of monocots and eudicots, recent molecular estimates for the dates of these events have converged on 140–190 mya (13, 47). Moore et al. (33) estimated that the lineages of eudicots, monocots, and magnoliids diverged at around 140–145 mya, and the divergence of the lineages of monocots and eudicots were estimated around 145 mya by others (41, 46). We therefore fixed the node uniting monocots and eudicots at 145 mya when these nodes were used as a calibration point. The ages of the duplicated genes were estimated using different calibration points or the same calibration points with slightly different ages, with and without constraining certain nodes with a minimum age based on fossil records. The constraints applied were minimum ages of 95 mya on the node uniting *Populus* and *Medicago*, 90 mya on the node uniting *Populus* and *Manihot* (49), 80 mya on the node uniting *Arabidopsis* and *Gossypium* (50), and 120 mya on the node uniting *Oryza* and *Acorus* (13).

One might argue that the uncertainty in the ages of the calibration points will affect the ages of the WGDs. As explained above, it is unlikely that the ages of calibration points are highly unrealistic, and the fossil constraints we used should buffer the uncertainty in the ages of the calibration points. We tested alternative calibrations, and although the estimated ages of the WGDs do slightly change (Table S1), the clustering of WGD events in time remained significant when combinations of different calibration ages were used (see below). Thus, although there will be some uncertainty related to the calibrations in the estimated ages of each WGD, we believe that this is unlikely to change the significant clustering of WGD events in time.

Effects of Taxon/Gene Sampling. One factor that might influence the estimated ages of the duplication events is taxon or gene sampling. Although it has been shown that the PL method is quite robust to undersampling compared to other methods (14), it is possible that our limited sampling might not be sufficient to account for the rate variation across branches, and could lead to under- or overestimation of the inferred duplication ages. In our approach, to better automate the process, we only used 1 orthologous gene for each species. It can be argued that using only one might reduce the ability to account for rate variation across genes compared to when more genes (paralogs) from each species are included, especially because the evolutionary rate can change rapidly for some genes after duplication due to different selective constraints. This possibility was tested for the WGD in *Arabidopsis* by taking gene families where either *Medicago*, *Populus*, or *Vitis* had 2 orthologous genes (in-paralogs to each other), and included both genes and obtained an age distribution. This age distribution differed very little from the age

distribution obtained with only 1 gene per species. Although it is possible that the age of some paralogs within a genome will be over- or underestimated due to local rate variation in some genes, the age of the WGD is based on a distribution of the dates of a large number of duplicates, and thus we find it unlikely that such cases will have a large effect on the age of the WGD.

Another concern was the rate variation across species. We tested this possible effect by including additional taxa when dating the duplications in *Arabidopsis* and *Populus*. First, *G. hirsutum* was added when dating the duplication in *Arabidopsis*. The lineages of *Arabidopsis* (Brassicales) and *Gossypium* (Malvales) diverged likely around 83–87 mya (49), and share a more recent common ancestry than *Arabidopsis* (eurosids II) and *Poplar*/*Medicago* (eurosids I). It is well accepted that the youngest WGD event in the ancestor of *Arabidopsis* is not shared with *Gossypium* (49–51). The inclusion of *Gossypium*, therefore breaks up the branch leading from the split of *Arabidopsis* and *Poplar*/*Medicago* to the paralogs of *Arabidopsis*, which may allow a better smoothing of the evolutionary rates and lead to a more accurate estimate of the ages of the duplications by the PL method. Although the ages of the individual paralogs did change in some cases, the overall estimated range of the WGD event changed very little (43.0 mya without *Gossypium* and 40.5 mya with *Gossypium*) (Table S1; AV115 cons). This would suggest that in the case of the *Arabidopsis* WGD event, the effect of undersampling is likely to be small, and might have been attenuated by obtaining an age distribution with a large number of duplicates.

It is well acknowledged that the substitution rate in the tree *Populus* is much lower than in weeds such as *Arabidopsis*, leading to underestimation of the age of duplicated genes in poplar (see above) (10). Our initial estimated age of ≈30 mya is much older than the age of 8–13 mya, which was calculated by converting K_S values into ages using substitution rates of *Arabidopsis* or grasses, but still much younger than the >60 mya reported in (10). This latter age was based on observations that the duplication event in poplar is likely to be shared with *Salix* species (as inferred from phylogenetic trees), thought to share a most recent common ancestry with *Populus* at around 60 mya. Our results show that our dating method has accounted for the much slower evolutionary rate in *Populus* to a certain extent but perhaps not enough. To investigate this in more detail we added sequences of *Manihot esculenta* (cassava). *Manihot* and *Populus* are both members of Malpighiales and share a more common recent ancestry than *Populus* and *Medicago*. The lineage of *Manihot* and *Populus* diverged early in the evolution of Malpighiales, which the fossil record indicates before 90 mya (48). When no age constraints were applied apart from the fixed nodes (*Vitis* – other rosids: 115 mya, or eurosids I - eurosids II: 105 mya), the estimated age of the *Populus* WGD did not change much. However, when a minimum age of 90 mya was assigned to the node uniting *Populus* and *Manihot*, an older age of about 45–50 mya was obtained for the WGD in *Populus* (Fig. S3). We observed that the age of the divergence of *Populus* and *Manihot* was constantly being estimated younger than what the fossil records suggest (>90 mya) when this node was not constrained, and therefore constraining this node resulted in older estimates for the ages of the paralogs. This shows that the effect of undersampling may be considerable when the species of interest has a very different evolutionary rate, which is likely to be the case with *Populus*. We, therefore, feel that the ages of paralogs in *Populus* are still underestimated and that additional taxon sampling (which was currently not possible) might result in yet older estimates for the age of the WGD.

1. Koch MA, Haubold B, Mitchell-Olds T (2000) Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). *Mol Biol Evol* 17:1483–1498.
2. Jakobsson M, et al. (2006) A unique recent origin of the allotetraploid species *Arabidopsis suecica*: Evidence from nuclear DNA markers. *Mol Biol Evol* 23:1217–1231.
3. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290:1151–1155.
4. Simillion C, et al. (2002) The hidden duplication past of *Arabidopsis thaliana*. *Proc Natl Acad Sci USA* 99:13627–13632.
5. Blanc G, Wolfe KH (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* 16:1667–1678.
6. Gaut BS, Doebley JF (1997) DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc Natl Acad Sci USA* 94:6809–6814.
7. Lescot M, et al. (2008) Insights into the *Musa* genome: Syntenic relationships to rice and between *Musa* species. *BMC Genomics* 9:58.
8. Moniz de Sá M, Drouin G (1996) Phylogeny and substitution rates of angiosperm actin genes. *Mol Biol Evol* 13:1198–1212.
9. Sterck L, et al. (2005) EST data suggest that poplar is an ancient polyploid. *New Phytol* 167:165–170.
10. Tuskan GA, et al. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313:1596–1604.
11. Sanderson MJ (2002) Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. *Mol Biol Evol* 19:101–109.
12. Sanderson MJ (2003) r8s: Inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19:301–302.
13. Bell CD, Soltis DE, Soltis PS (2005) The age of the angiosperms: A molecular timescale without a clock. *Evolution* 59:1245–1258.
14. Linder HP, Hardy CR, Rutschmann F (2005) Taxon sampling effects in molecular clock dating: An example from the African Restionaceae. *Mol Phyl Evol* 35:569–582.
15. Thorne JL, Kishino H (2002) Divergence time and evolutionary rate estimation with multilocus data. *Syst Biol* 51:689–702.
16. Rannala B, Yang Z (2007) Inferring speciation times under an episodic molecular clock. *Syst Biol* 56:453–466.
17. Drummond AJ, Ho SY, Phillips MJ, Rambaut (2006) A relaxed phylogenetics and dating with confidence. *PLoS Biol* 4:e88.
18. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591.
19. Kishino H, Thorne JL, Bruno WJ (2001) Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol Biol Evol* 18:352–361.
20. Velasco R, et al. (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* 2:e1326.
21. Li WH, Gu Z, Wang H, Nekrutenko A (2001) Evolutionary analyses of the human genome. *Nature* 409:847–849.
22. Simillion C, Janssens K, Sterck L, Van de Peer Y (2008) i-ADHoRe 2.0: An improved tool to detect degenerated genomic homology using genomic profiles. *Bioinformatics* 24:127–128.
23. De Bodt S, Maere S, Van de Peer Y (2005) Genome duplication and the origin of angiosperms. *Trends Ecol Evol* 20:591–597.
24. Yang Z (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13:555–556.
25. Schiex T, Gouzy J, Moisan A, de Oliveira Y (2003) FrameD: A flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences. *Nucleic Acids Res* 31:3738–3741.
26. Cui L, et al. (2006) Widespread genome duplications throughout the history of flowering plants. *Genome Res* 16:738–749.
27. Barker MS, et al. (2008) Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol Biol Evol* 25:2445–2455.
28. Maere S, et al. (2005) Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci USA* 102:5454–5459.
29. Remm M, Storm CE, Sonnhammer EL (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* 314:1041–1052.
30. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680.
31. Vandepoele K, et al. (2004) Major events in the genome evolution of vertebrates: Paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc Natl Acad Sci USA* 101:1638–1643.
32. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696–704.
33. Jansen RK, et al. (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc Natl Acad Sci USA* 104:19369–19374.
34. Moore MJ, Bell CD, Soltis PS, Soltis DE (2007) Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proc Natl Acad Sci USA* 104:19363–19368.
35. Cannon SB, et al. (2006) Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes. *Proc Natl Acad Sci USA* 103:18026.
36. Paterson AH, Bowers JE, Chapman BA (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci USA* 101:9903–9908.
37. Jaillon O, et al. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449:463–467.
38. Wikström N, Savolainen V, Chase MW (2001) Evolution of the angiosperms: Calibrating the family tree. *Proc Biol Sci* 268:2211–2220.

39. Jansen RK, et al. (2006) Phylogenetic analyses of *Vitis* (Vitaceae) based on complete chloroplast genome sequences: Effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. *BMC Evol Biol* 6:32.
40. Tang H, et al. (2008) Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res* 18:1944–1954.
41. Leebens-Mack J, et al. (2005) Identifying the basal angiosperm node in chloroplast genome phylogenies: Sampling one's way out of the Felsenstein zone. *Mol Biol Evol* 22:1948–1963.
42. Rutschmann F, Eriksson T, Salim KA, Conti E (2007) Assessing calibration uncertainty in molecular dating: The assignment of fossils to alternative calibration points. *Syst Biol* 56:591–608.
43. Friis EM, Pedersen KR, Crane PR (2001) Fossil evidence of water lilies (Nymphaeales) in the Early Cretaceous. *Nature* 410:357–360.
44. Zhou Z-k, Crepet WL, Nixon KC (2001) The earliest fossil evidence of the Hamamelidaceae: Late Cretaceous (Turonian) inflorescences and fruits of Altingioideae. *Am J Bot* 88:753–766.
45. Schneider H, et al. (2004) Ferns diversified in the shadow of angiosperms. *Nature* 428:553–557.
46. Chaw S-M, Chang C-C, Chen H-L, Li W-H (2004) Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. *J Mol Evol* 58:424–441.
47. Soltis DE, Bell CD, Kim S, Soltis PS (2008) Origin and early evolution of angiosperms. *Ann N Y Acad Sci* 1133:3–25.
48. Crepet WL, Nixon KC, Gandolfo MA (2004) Fossil evidence and phylogeny: The age of major angiosperm clades based on mesofossil and macrofossil evidence from cretaceous deposits. *Am J Bot* 91:1666–1682.
49. Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422:433–438.
50. Blanc G, Hokamp K, Wolfe KH (2003) A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res* 13: 137–144.
51. Rong J, et al. (2005) Comparative genomics of *Gossypium* and *Arabidopsis*: Unraveling the consequences of both ancient and recent polyploidy. *Genome Res* 15:1198–1210.

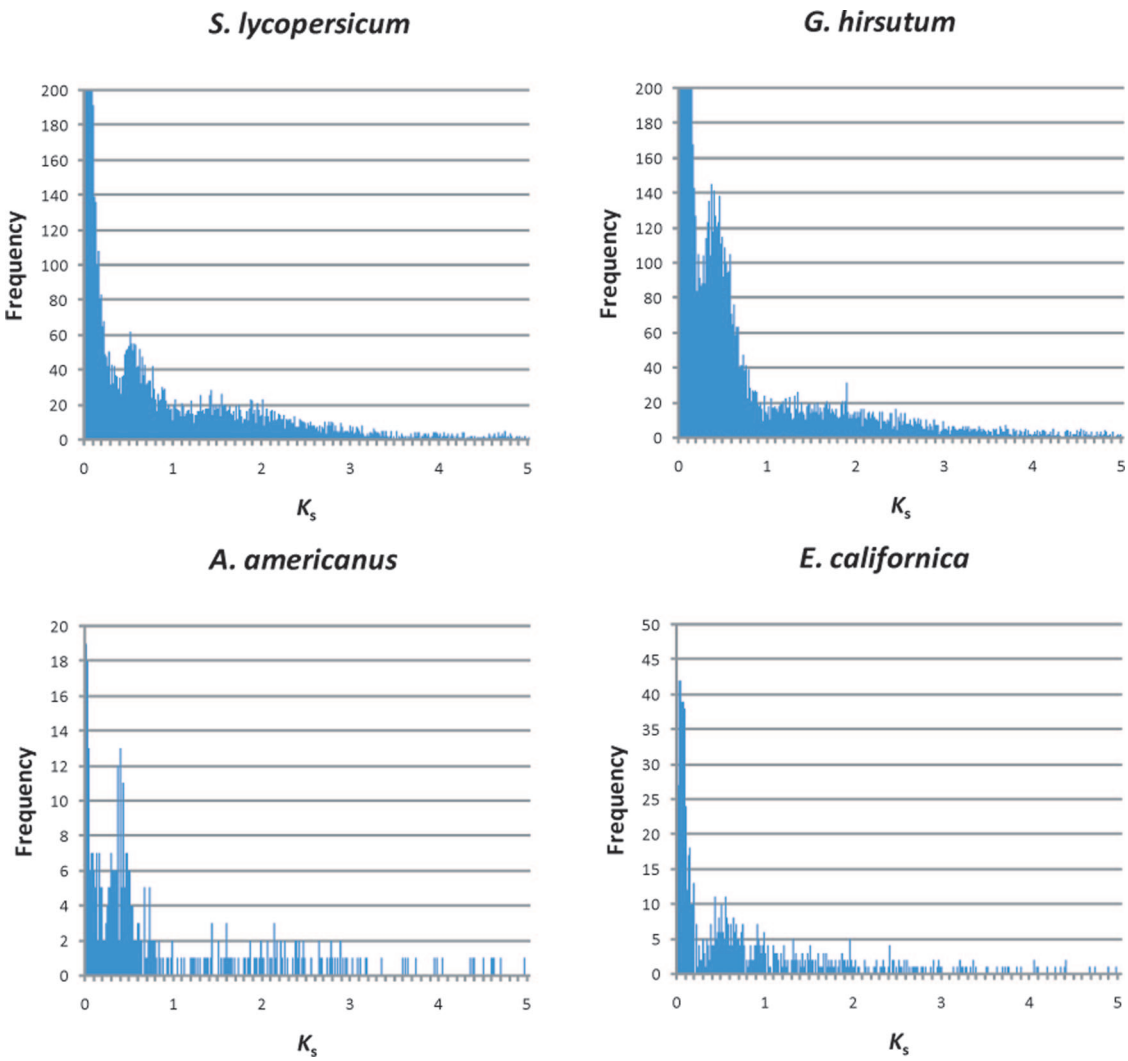

Fig. S1. K_S distributions of representative EST pairs for various non-sequenced plant species (the distributions for *S. lycopersicum* and *G. hirsutum* are truncated at frequency 200).

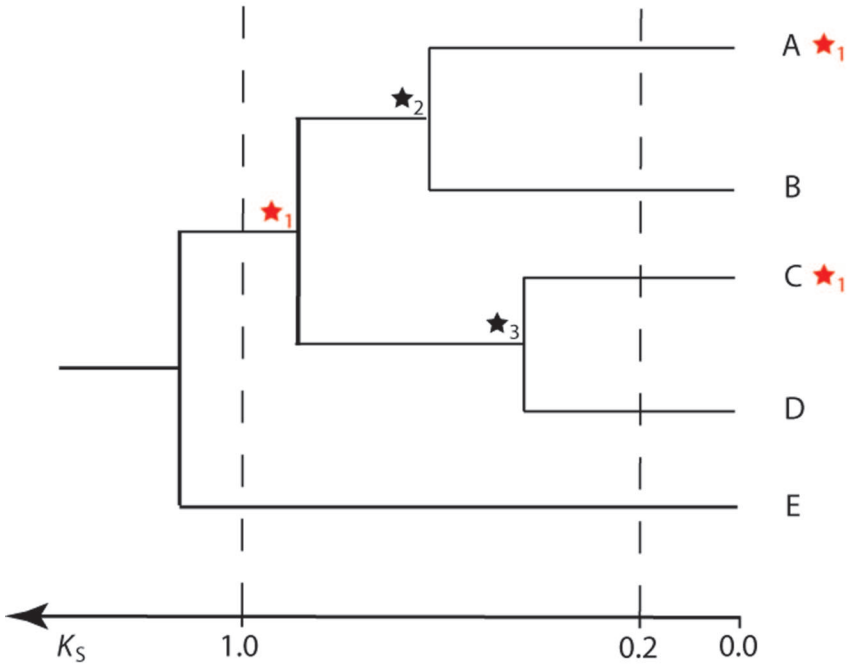

Fig. S2. K_S-based hierarchical clustering tree for a hypothetical EST-derived gene family with 5 members (A–E). Three duplication events are inferred in the K_S interval [0.2–1.0] (marked with stars). For events 2 and 3, there is only 1 possible representative pair, whereas there are 4 possibilities for event 1 (AC, AD, BC and BD). One of these, in this case the pair AC, was randomly chosen.

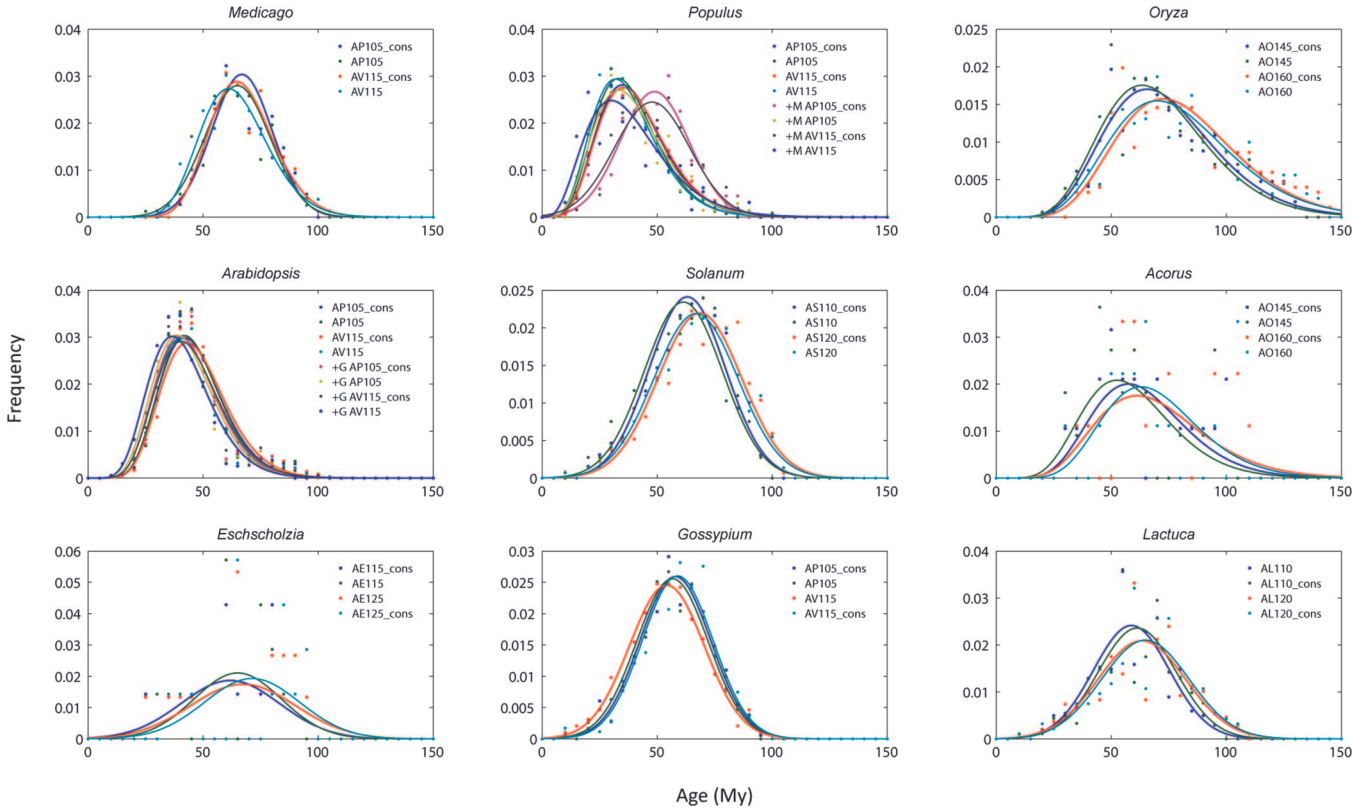

Fig. S3. Maximum likelihood fits for all species using different calibrations and constraints (Table S1). Note the remarkable shift of the distribution for *Populus* when adding *Manihot* and fixing the divergence between the 2 at minimum 90 my.

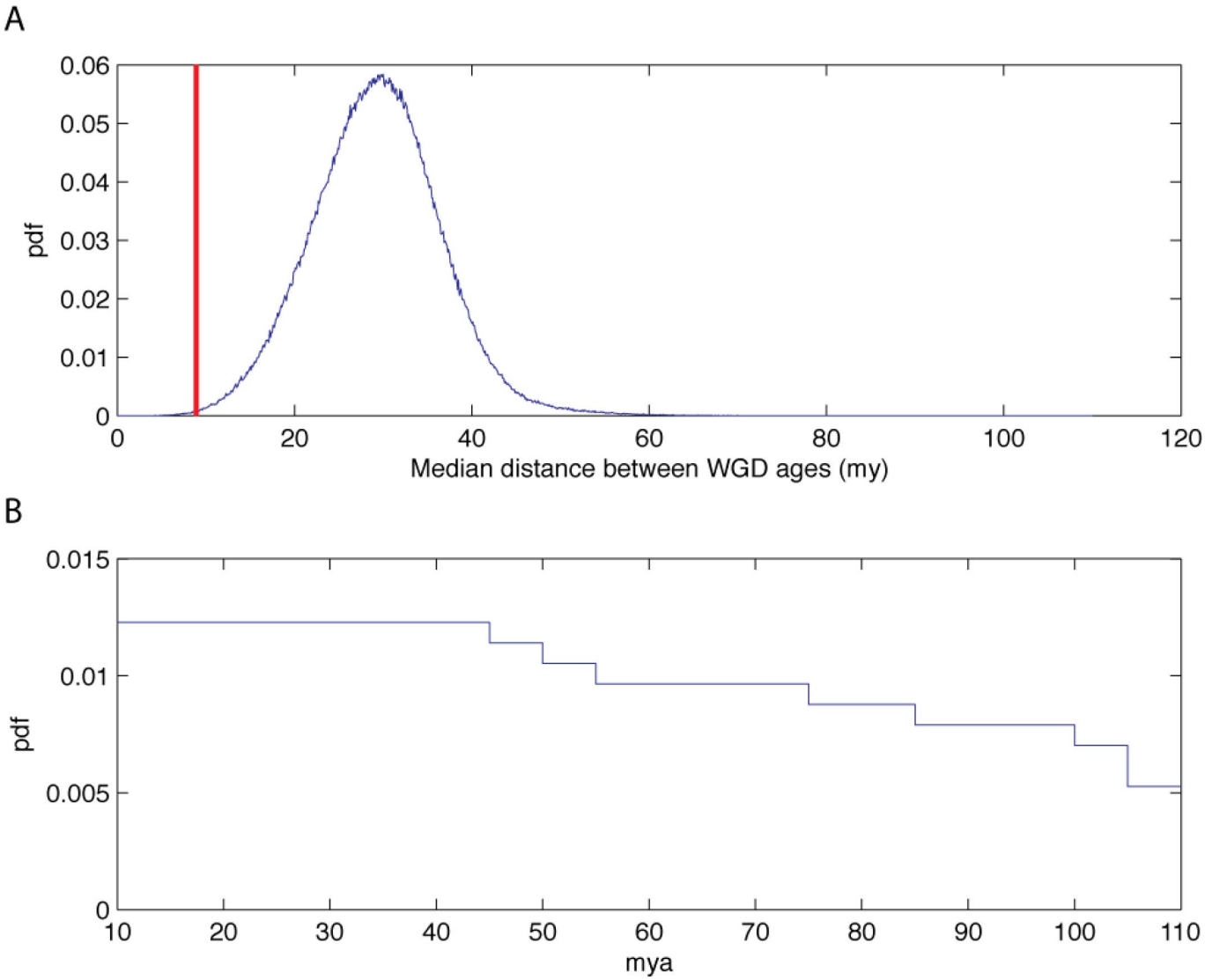

Fig. S4. (*A*) Assessment of the significance of grouping of WGD events in time. The sampled distribution of the median distance under the null hypothesis of random occurrence of WGD events is shown in blue; the median distance between the WGD ages estimated in this study (last column of Table 1) is indicated by the vertical red line. (*B*) Probability density function (pdf) from which the random WGD ages used in *A* were sampled. Each discontinuity in the pdf corresponds to a speciation event in Fig. 3.

Table S1. Numbers of paralogs and gene families for different species used as well as inferred duplication dates (plus confidence intervals) under different constraints

Species	No. of paralogs created by WGD	No. of gene families	No. of gene families with divergence dates	Constraints	No. of duplicates in distribution	Estimated ages, my	95% CI, my	Fit
Arabidopsis	2,705	978	878	AV115	748	40.2	38.9–41.1	Gamma
				AV115_cons*	723	43.0	41.6–44.0	Gamma
				AP105	728	41.8	40.5–42.7	Gamma
				AP105_cons	721	42.0	40.7–42.9	Gamma
Arabidopsis (+ *Gossypium*)	—	572	467	AV115	390	36.6	34.7–37.8	Gamma
				AV115_cons	372	40.5	38.4–41.7	Gamma
				AP105	385	38.7	36.8–39.9	Gamma
				AP105_cons	343	39.9	37.8–41.2	Gamma
Gossypium	6,323	1,027	705	AV115	388	54.2	52.6–55.8	Normal
				AV115_cons*	348	58.9	57.3–60.6	Normal
				AP105	382	56.6	55.0–58.2	Normal
				AP105_cons	364	57.9	56.3–59.5	Normal
Populus	5,112	1,728	1,502	AV115	779	31.9	30.6–32.8	Gamma
				AV115_cons	732	35.3	33.9–36.3	Gamma
				AP105	811	33.1	31.8–34.0	Gamma
				AP105_cons	740	34.9	33.5–35.9	Gamma
Populus (+ *Manihot*)	—	436	297	AV115	128	30.1	25.5–32.7	Gamma
				AV115_cons*	126	47.8	44.9–50.7	Normal
				AP105	139	33.2	29.2–35.4	Gamma
				AP105_cons	133	48.8	46.3–51.4	Normal
Medicago	324	188	181	AV115	159	61.1	56.3–62.9	Gamma
				AV115_cons*	156	64.6	59.7–66.4	Gamma
				AP105	163	64.9	62.7–67.1	Normal
				AP105_cons	149	66.9	64.7–69.0	Normal
Solanum	3,088	599	414	AS120	292	66.9	64.8–69.0	Normal
				AS120_cons*	270	68.6	66.4–70.7	Normal
				AS110	292	61.4	59.4–63.3	Normal
				AS110_cons	259	63.1	61.1–65.2	Normal
L. sativa	1,871	386	295	AL110	202	58.8	56.5–61.1	Normal
				AL110_cons	183	61.3	58.9–63.8	Normal
				AL120	217	63.3	60.8–65.9	Normal
				AL120_cons*	187	64.9	62.1–67.6	Normal
Eschscholzia	358	59	28	AE125	15	67.4	54.7–80.0	Normal
				AE125_cons*	14	71.3	59.3–83.2	Normal
				AE115	14	65.0	54.1–76.0	Normal
				AE115_cons	14	61.8	49.5–74.2	Normal
Oryza	1,952	559	455	AO145	314	63.4	59.8–65.7	Gamma
				AO145_cons*	295	65.6	61.6–68.0	Gamma
				AO160	321	70.0	65.9–72.5	Gamma
				AO160_cons	302	72.6	68.3–75.2	Gamma
Acorus	235	86	40	AO145	22	52.5	25.4–58.4	Gamma
				AO145_cons*	19	57.4	24.5–64.1	Gamma
				AO160	18	63.0	25.6–70.2	Gamma
				AO160_cons	18	61.2	24.2–69.0	Gamma

*Age estimates calculated based on the constraints are the ages shown in Table 1 of the article:
–AV115: most recent common ancestor (mrca) *Arabidopsis Vitis*; fix 115
–AV115_cons: mrca *Arabidopsis Vitis*; fix 115, mrca *Populus Medicago*; min 95, (mrca *Arabidopsis Gossypium*; min 80, mrca *Populus Manihot*; min 90)
–AP105: mrca *Arabidopsis Populus*; fix 105
–AP105_cons: mrca *Arabidopsis Populus*; fix 105, mrca *Populus Medicago*; min 95, (mrca *Arabidopsis Gossypium*; min 80, mrca *Populus Manihot*; min 90)
–AS120: mrca *Arabidopsis Solanum*; fix 120
–AS120_cons: mrca *Arabidopsis Solanum*; fix 120, mrca *Populus Medicago*; min 95
–AS110: mrca *Arabidopsis Solanum*; fix 110
–AS110_cons: mrca *Arabidopsis Solanum*; fix 110, mrca *Populus Medicago*; min 95
–AL120: mrca *Arabidopsis Lactuca*; fix 120
–AL120_cons: mrca *Arabidopsis Lactuca*; fix 120, mrca *Populus Medicago*; min 95
–AL110: mrca *Arabidopsis Lactuca*; fix 110
–AL110_cons: mrca *Arabidopsis Lactuca*; fix 110, mrca *Populus Medicago*; min 95
–AE125: mrca *Arabidopsis Eschscholzia*; fix 125
–AE125_cons: mrca *Arabidopsis Eschscholzia*; fix 125, mrca *Populus Medicago*; min 95
–AE115: mrca *Arabidopsis Eschscholzia*; fix 115
–AE115_cons: mrca *Arabidopsis Eschscholzia*; fix 115, mrca *Populus Medicago*; min 95
–AO145: mrca *Arabidopsis Oryza*; fix 145
–AO145_cons: mrca *Arabidopsis Oryza*; fix 145, mrca *Populus Medicago*; min 95, (mrca *Oryza Acorus*; min 120)
–AO160: mrca *Arabidopsis Oryza*; fix 160
–AO160_cons: mrca *Arabidopsis Oryza*; fix 160, mrca *Populus Medicago*; min 95, (mrca *Oryza Acorus*; min 120)