LETTER

doi:10.1038/nature09916

Ancestral polyploidy in seed plants and angiosperms

Yuannian Jiao[1,2], Norman J. Wickett[2], Saravanaraj Ayyampalayam[3], André S. Chanderbali[4], Lena Landherr[2], Paula E. Ralph[2], Lynn P. Tomsho[5], Yi Hu[2], Haiying Liang[6], Pamela S. Soltis[7], Douglas E. Soltis[4], Sandra W. Clifton[8], Scott E. Schlarbaum[9], Stephan C. Schuster[5], Hong Ma[1,2,10,11], Jim Leebens-Mack[3] & Claude W. dePamphilis[1,2]

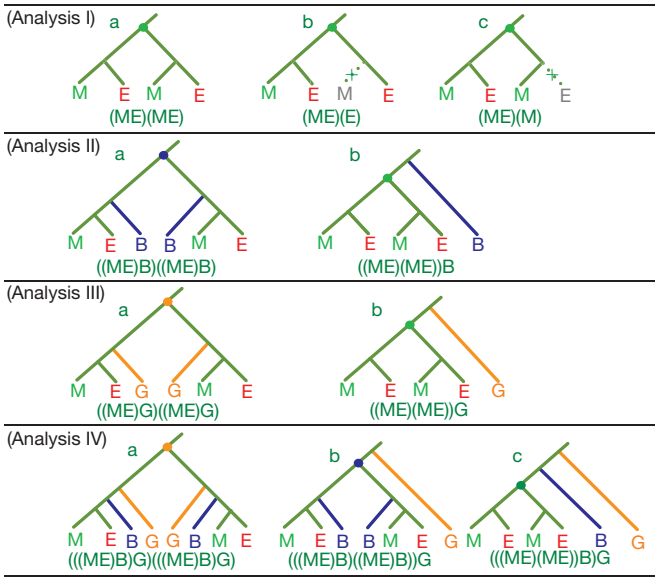
Whole-genome duplication (WGD), or polyploidy, followed by gene loss and diploidization has long been recognized as an important evolutionary force in animals, fungi and other organisms[1–3], especially plants. The success of angiosperms has been attributed, in part, to innovations associated with gene or whole-genome duplications[4–6], but evidence for proposed ancient genome duplications pre-dating the divergence of monocots and eudicots remains equivocal in analyses of conserved gene order. Here we use comprehensive phylogenomic analyses of sequenced plant genomes and more than 12.6 million new expressed-sequence-tag sequences from phylogenetically pivotal lineages to elucidate two groups of ancient gene duplications—one in the common ancestor of extant seed plants and the other in the common ancestor of extant angiosperms. Gene duplication events were intensely concentrated around 319 and 192 million years ago, implicating two WGDs in ancestral lineages shortly before the diversification of extant seed plants and extant angiosperms, respectively. Significantly, these ancestral WGDs resulted in the diversification of regulatory genes important to seed and flower development, suggesting that they were involved in major innovations that ultimately contributed to the rise and eventual dominance of seed plants and angiosperms.

Angiosperms are by far the largest group of land plants, with more than 300,000 living species. Significantly, most flowering plant lineages reflect one or more rounds of ancient polyploidy. For example, extensive analyses of the complete genome sequence of *Arabidopsis thaliana* support two recent WGDs (named α and β) within the crucifer (Brassicaceae) lineage and one triplication event (γ) that is probably shared by all core eudicots[7–13]. The *Populus trichocarpa* genome shows evidence of the core eudicot triplication as well as a more recent WGD[14]. Two polyploidy events in monocots (ρ and σ) have been inferred to have pre-dated the diversification of cereal grains and other grasses[15] (Poaceae). Several studies have hinted that an ancient WGD event occurred even earlier in angiosperm evolution[4,5,10,16]. However, the existence and timing of these ancient events, and their long-term impact, remain uncertain.

Here we use a rigorous phylogenomic approach (Supplementary Fig. 1; details in Supplementary Methods) to test the hypothesis that one or more ancient genome duplications occurred before the divergence of monocots and eudicots. By mapping the duplication events onto phylogenetic trees, we determine whether the paralogues were duplicated before or after a given speciation event[8,17] (Fig. 1a). Although individual genes might be lost in some phylogenies, a broad picture can be drawn from simultaneous consideration of many or all gene families.

We used species with completely sequenced genomes (Supplementary Table 1; two monocots (*Oryza sativa* and *Sorghum bicolor*) and five eudicots (*A. thaliana*, *Carica papaya*, *P. trichocarpa*, *Cucumis sativus*

a



b

	Total	Orthogroups with duplication*	Type a		Type b		Type c	
			BS ≥ 80%	BS ≥ 50%	BS ≥ 80%	BS ≥ 50%	BS ≥ 80%	BS ≥ 50%
Analysis I	7,470	799	234	285	142	312	91	219
Analysis II	540	345	262 (99%)	343 (99%)	1 (1%)	5 (1%)		
Analysis III	338	229	62 (51%)	147 (57%)	59 (49%)	110 (43%)		
Analysis IV	322	198	65 (53%)	130 (58%)	54 (44%)	88 (39%)	4 (3%)	6 (3%)

*Some orthogroups contain more than one type of ancient duplication as shown in Supplementary Fig. 4.

Figure 1 | Hypothetical tree topologies and summary of orthogroups consistent with ancient gene duplications before the split of monocots and eudicots. a, Analysis I: three examples of phylogenetic trees showing the patterns of retention or loss of paralogues: (a) both of the paralogues are retained in monocots and eudicots; (b) one of the paralogues was lost in monocots; (c) one of the paralogues was lost in eudicots. Analysis II: orthologues from basal angiosperms were added to core-orthogroups to refine the timing of ancient gene duplications in angiosperms: (a) gene duplication shared across all angiosperms; (b) gene duplication shared only by monocots and eudicots. Analysis III: orthologues from gymnosperms were added to core-orthogroups to place shared gene duplications before (a) and/or after (b) the split of extant gymnosperms and angiosperms. Analysis IV: three different topologies consistent with the timing of duplications shared by seed plants (a), angiosperms (b) and monocots + eudicots (c) when we expanded core-orthogroups with additional orthologues from both basal angiosperms and gymnosperms. M, monocots; E, eudicots; B, basal angiosperms; G, gymnosperms. Exemplar trees in analyses II, III and IV illustrate expected patterns with all branches retained. Observed topologies typically had partial gene losses similar to analyses Ib and Ic. **b**, Summary of orthogroups showing different types of duplications corresponding to proposed topologies inferred from orthogroup trees.

[1]Intercollege Graduate Degree Program in Plant Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA. [2]Department of Biology, Institute of Molecular Evolutionary Genetics, and the Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, Pennsylvania 16802, USA. [3]Department of Plant Biology, University of Georgia, Athens, Georgia 30602, USA. [4]Department of Biology, University of Florida, Gainesville, Florida 32611, USA. [5]Center for Comparative Genomics, Center for Infectious Disease Dynamics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA. [6]Department of Genetics and Biochemistry, Clemson University, Clemson, South Carolina 29634, USA. [7]Florida Museum of Natural History, University of Florida, Gainesville, Florida 32611, USA. [8]The Genome Center at Washington University, Saint Louis, Missouri 63108, USA. [9]Department of Forestry, Wildlife & Fisheries, Institute of Agriculture, The University of Tennessee, Knoxville, Tennessee 37996, USA. [10]State Key Laboratory of Genetic Engineering, School of Life Sciences, Institute of Plant Biology, Center for Evolutionary Biology, Fudan University, Shanghai 200433, China. [11]Institute of Biomedical Sciences, Fudan University, Shanghai 200433, China.

and *Vitis vinifera*)) to construct gene families or subfamilies. One lycophyte (*Selaginella moellendorffii*) and one moss (*Physcomitrella patens*) were used as outgroups when dating gene duplications and potential WGDs that occurred before the monocot–eudicot divergence. In total, 77.03% of all protein-coding genes in the sequenced genomes were grouped in 31,433 multigene 'core-orthogroups'. We define orthogroups as clusters of homologous genes that derive from a single gene in the common ancestor of the focal taxa, and refer to orthogroups for the nine sequenced genomes as core-orthogroups. Of these, 7,470 core-orthogroups contain at least one monocot, one eudicot, and one *Selaginella* and/or *Physcomitrella* sequence. These core-orthogroups were used in our investigation of duplication events predating the divergence of monocots and eudicots.

We queried maximum-likelihood trees (MLTs) for each core-orthogroup for topologies indicative of shared duplications (Fig. 1a, analysis I). We filtered our gene trees (Supplementary Methods), requiring that at least one of the seven core species retained both paralogues following the inferred gene duplication event in a common monocot–eudicot ancestor (see Supplementary Data 1 for a list of orthogroups). For example, the MLT for orthogroup 1711 (DEAD-box RNA helicase) contained duplicate genes in both monocots and eudicots whereas the MLTs for orthogroup 2312 (spermidine synthase) and orthogroup 396 (function unknown) showed that either one of the monocot or eudicot paralogues was lost after the divergence of monocots and eudicots (see exemplar trees in Supplementary Figs 2a, 3a and 4). On the basis of this conservative criterion, we identified a large number of core-orthogroups with shared duplication of monocots and eudicots (829 duplications in 799 core-orthogroups with bootstrap support (BS) greater than or equal to 50%; 474 duplications in 451 core-orthogroups with BS $\geq$ 80%; Supplementary Data 2). These duplications occurred before the γ triplication[9,13] (which may be restricted to eudicots). As expected[9,13], many younger duplications within the sampled eudicot lineages were also observed on these trees (1,146 orthogroups surviving at least one eudicot-wide triplication (γ)), but for this study we focused on ancient duplications that occurred before the divergence of monocots and eudicots.

Additional homologues from basal angiosperms (*Aristolochia*, *Liriodendron*, *Nuphar* and *Amborella*; Supplementary Table 2) and gymnosperms (*Pinus*, *Picea*, *Zamia*, *Cryptomeria* and others; Supplementary Table 2) were added to 799 core-orthogroups to form expanded orthogroups[18]. These phylogenetically critical lineages increase gene sampling and provide better resolution of the timing of ancient duplications. By 'basal angiosperms' we mean the earliest-branching lineages of flowering plants that arose before the separation of monocots and eudicots. Before re-estimating gene trees for the expanded orthogroups, we added another quality control step to remove short or highly divergent unigenes (sequences produced from assembly of expressed-sequence-tag data sets; Supplementary Methods). After filtering, there remained 540 and 338 orthogroups with unigenes sampled from basal angiosperms and gymnosperms, respectively. Among these, 322 orthogroups contained unigenes from both basal angiosperms and gymnosperms (Fig. 1b).

For the 540 orthogroups with unigenes from basal angiosperms, the number of trees in which we identified an ancestral duplication before the origin of angiosperms[19] (Fig. 1a, analysis IIa) greatly exceeded the number in which we identified a shared duplication after the origin of angiosperms (Fig. 1a, analysis IIb). Inference of a duplication pre-dating the diversification of basal angiosperms (ancestral angiosperm duplication) was supported by 262 (BS $\geq$ 80%) or 343 (BS $\geq$ 50%) orthogroups, whereas only one (BS $\geq$ 80%) or five (BS $\geq$ 50%) orthogroups supported inference of a gene duplication just after the origin of the angiosperm crown group (Fig. 1b, analysis II). We also found only five orthogroups with a surviving duplication shared with some, but not all, sampled basal angiosperms. Although basal angiosperms are a grade (and not a clade), we represent them with a single line in Fig. 1a because the duplication signal is inclusive of all basal angiosperms.

Additional analyses of 338 orthogroups populated with unigenes of gymnosperms identified 62 (BS $\geq$ 80%) or 147 (BS $\geq$ 50%) trees containing a seed-plant-wide gene duplication and 59 (BS $\geq$ 80%) or 110 (BS $\geq$ 50%) trees with a later duplication shared only by angiosperms (Fig. 1b, analysis III). In addition, analyses of the 322 orthogroups expanded with orthologues from both basal angiosperms and gymnosperms also detected similar signals of the two ancient shared duplications: 65 (BS $\geq$ 80%) or 130 (BS $\geq$ 50%) trees showing an ancestral seed plant duplication (see exemplar tree in Supplementary Fig. 2b), and 54 (BS $\geq$ 80%) or 88 (BS $\geq$ 50%) trees supporting an ancestral angiosperm duplication (Supplementary Fig. 3b and Fig. 1b, analysis IV).

In summary, our conservative filtering procedure identified 799 trees with topologies suitable for testing hypotheses concerning the presence of ancient duplications. These trees provided overwhelming support for the presence of two groups of duplications, one in the common ancestor of all angiosperms and the other in the common ancestor of all seed plants. Several mechanisms could explain the concerted patterns of gene duplication revealed in the gene trees, including WGD or multiple segmental or chromosomal duplications. The most parsimonious interpretation of the existing data is ancient WGD. We performed divergence time analyses to investigate this hypothesis further.

If the proposed WGDs were real, the estimated dates for gene duplication events in independent gene trees would be expected to be similar. Alternatively, if the duplications were unrelated (that is, a collection of independent events), a uniform distribution of duplication times within the intervals between the origins of gymnosperms and angiosperms would be expected for the ancestral angiosperm duplicates or on the branch leading to seed plants for the ancestral seed plant duplicates. We calibrated 799 core-orthogroups supporting (BS $\geq$ 50%) ancient duplications before the separation of monocots and eudicots from analysis I and estimated the divergence times of 860 nodes in 774 core-orthogroups using the program R8S (Supplementary Methods).

We then analysed the distribution of the inferred duplication times using a Bayesian method that assigned divergence time estimates to classes specified by a mixture model[20]. The distribution of duplication times was bimodal, with peaks 192 $\pm$ 2 (95% confidence interval) and 319 $\pm$ 3 million years (Myr) ago. Dates were clustered in two relatively short time intervals, suggesting that these duplications were not uniformly distributed (Fig. 2a). Furthermore, we also analysed the 499

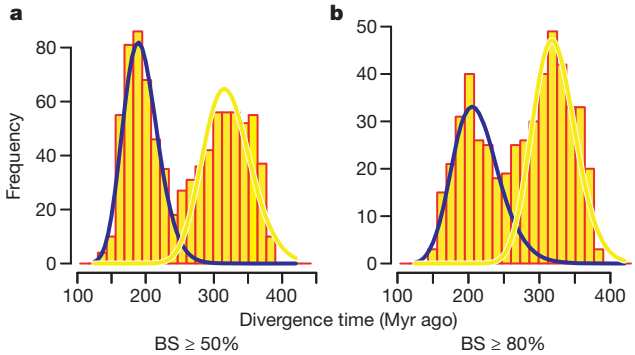


Figure 2 | Age distribution of ancient duplications shared by monocots and eudicots. **a**, The inferred divergence times for 866 ancestral duplication nodes in 779 core-orthogroups (BS $\geq$ 50%) were analysed by EMMIX to determine whether these duplications occurred randomly over time or within some small time frame. Each component is written as 'colour/mean molecular timing/proportion' where 'colour' is the component (curve) colour and 'proportion' is the percentage of duplication nodes assigned to the identified component. There are two statistically significant components: blue/192 (Myr ago)/0.48 and yellow/319/0.52. **b**, When we required the bootstrap support of the monocot + eudicot duplication to be greater than or equal to 80%, there were 504 nodes in 439 core-orthogroups for analysis of the inferred divergence times by EMMIX. Two statistically significant components were identified: blue/210/0.43 and yellow/321/0.57.

nodes with ancient duplications in 435 orthogroups with BS ≥ 80% (Fig. 2b) and found a similar distribution pattern (two components: 210 ± 4 and 321 ± 4 Myr ago).

We then examined the age distribution of ancient duplications restricted only to orthogroups in analysis III that had been populated with nearly full-length gymnosperm unigenes. Among the 338 orthogroups with inferred absolute dates, there are 110 (BS ≥ 50%; 59 with BS ≥ 80%) that place a duplication on the angiosperm branch after divergence from gymnosperms. The distribution of duplication times inferred from these orthogroups showed one significant peak (234 ± 9 or 236 ± 9 Myr ago; Supplementary Fig. 5a, b). The most recent common ancestor of extant angiosperms existed has been dated to 130–190 Myr ago[19,21]. Therefore, the identified duplication event occurred before the radiation of extant angiosperms, which agrees with the results from phylogenetic analysis (Fig. 1b, analysis II). An additional analysis was restricted to those 147 (BS ≥ 50%) or 62 (BS ≥ 80%) orthogroups (Fig. 1b, analysis IIIa) that contained a seed-plant-wide duplication based on phylogenetic analysis. The mixture model analysis identified only one significant component for the distribution of duplication times (349 ± 3 or 347 ± 4 Myr ago; Supplementary Fig. 5c, d), which was older than the ancestral node for extant seed plants[22] (~310 Myr ago). Thus, both molecular dating and phylogenetic analyses support another ancient genome-wide duplication shared by all extant seed plants (Fig. 3). Distributions of synonymous site divergence for duplicated genes and synteny analyses also support this conclusion (Supplementary Discussion).

Gene duplication provides raw genetic material for the evolution of functional novelty. WGD in ancient seed plants would have generated duplicate copies of every gene, some of which could have had crucial roles in the origin of phenotypic novelty and, ultimately, in the origin and rapid diversification of the angiosperms. Although those genes retained as duplicates from the ancestral WGDs represent all functional categories, there is an overabundance of retained duplicate genes from several functional categories, including transferases and binding proteins, transcription factors and protein kinases (Supplementary Fig. 6 and Supplementary Data 3). These categories are significantly enriched for orthogroups surviving the monocot–eudicot duplication described in analysis I and for orthogroups surviving pre-angiosperm and/or pre-seed-plant duplications in analysis III. These results are consistent with patterns of gene retention following the more recent WGDs in the *Arabidopsis* lineage (ref. 23 and references therein), and WGD in vertebrates[24], supporting the interpretation that the concurrent duplications observed here are products of WGD. Taken together, these patterns suggest that the tendency for some types of gene duplicates to be retained following polyploidy has been a common feature of the post-WGD diploidization process throughout the evolutionary history of plants.

One subset of duplicated genes that could have contributed to ancient seed plant and angiosperm innovations includes those that have special roles in reproduction and flower development. In this study, we identified 35 orthogroups involved in flower developmental pathways with at least one ancient duplication event before the divergence of monocots and eudicots (Supplementary Table 3). For example, orthogroup 361 (containing *Arabidopsis PHYTOCHROME* genes), which includes regulators of flowering time[25] and seed germination[26], retained duplicate genes following two putative WGDs pre-dating the origin of angiosperms and seed plants, respectively, consistent with a published phylogeny for the *PHYTOCHROME* gene family[27]. Other published gene family phylogenies also suggested common patterns of gene duplication, hinting at the genome-scale duplications seen here. For example, *TIR1/AFB* has experienced an ancient duplication before the diversification of extant angiosperms[28]. Phylogenetic analyses of the *ZINC FINGER HOMEOBOX* (*ZHD*) family[29], the *HD-ZIP III* gene family[30], and MADS-box genes (Supplementary Discussion) show duplication patterns consistent with WGDs pre-dating the origin of angiosperms and seed plants. Hence, these previous studies of individual genes or gene families bolster our conclusions based on a genome-wide survey of thousands of genes, and identify some of the many genes derived from these duplications that could potentially have had important roles in seed plant and angiosperm evolution.

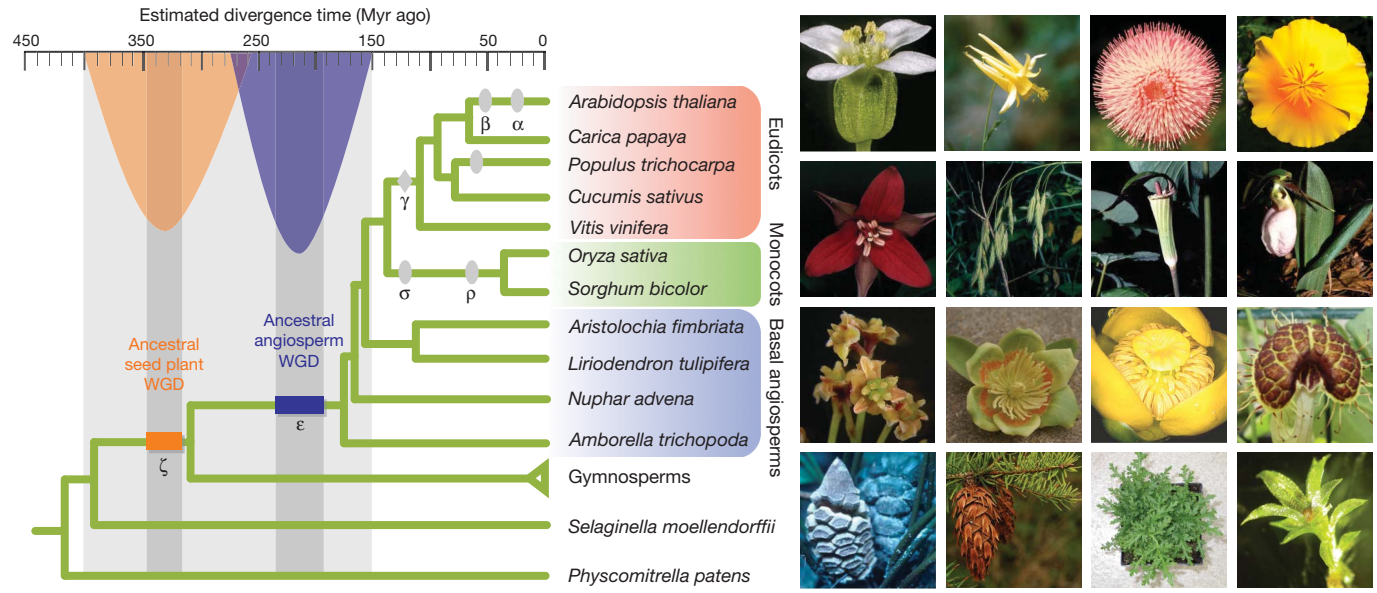


Figure 3 | Ancestral polyploidy events in seed plants and angiosperms. Two ancestral duplications identified by integration of phylogenomic evidence and molecular time clock for land plant evolution. Ovals indicate the generally accepted genome duplications identified in sequenced genomes (see text). The diamond refers to the triplication event probably shared by all core eudicots. Horizontal bars denote confidence regions for ancestral seed plant WGD and ancestral angiosperm WGD, and are drawn to reflect upper and lower bounds of mean estimates from Fig. 2 (more orthogroups) and Supplementary Fig. 5 (more taxa). The photographs provide examples of the reproductive diversity of eudicots (top row, left to right: *Arabidopsis thaliana*, *Aquilegia chrysantha*, *Cirsium pumilum*, *Eschscholzia californica*), monocots (second row, left to right: *Trillium erectum*, *Bromus kalmii*, *Arisaema triphyllum*, *Cypripedium acaule*), basal angiosperms (third row, left to right: *Amborella trichopoda*, *Liriodendron tulipifera*, *Nuphar advena*, *Aristolochia fimbriata*), gymnosperms (fourth row, first and second from left: *Zamia vazquezii*, *Pseudotsuga menziesii*) and the outgroups *Selaginella moellendorfii* (vegetative; fourth row, third from left) and *Physcomitrella patens* (fourth row, right). See Supplementary Table 4 for photo credits.

METHODS SUMMARY

Phylogenetic analysis. We used the OrthoMCL method to construct a set of core-orthogroups. All orthogroup amino-acid alignments were generated with MUSCLE and then trimmed by removing poorly aligned regions using TRIMAL 1.2. Additional sorted unigene sequences for the core-orthogroups (retrieved with HaMStR) were aligned at the amino-acid level into the existing nine species' full alignments (before trimming) using CLUSTALX 1.8. After trimming, each unigene sequence was checked and removed from the alignment if the sequence contained less than 70% alignment length. Corresponding DNA sequences were then forced onto the amino-acid alignment using custom Perl scripts and used for subsequent phylogenetic analysis. Maximum-likelihood analyses were conducted using RAXML, version 7.2.1, searching for the best MLT with the GTRGAMMA model, which represents an acceptable trade-off between speed and accuracy (RAXML 7.0.4 manual).

Molecular dating analyses and 95% confidence intervals. The divergence time of the two paralogous clades derived from each duplication was estimated from the best maximum-likelihood topologies under the assumption of a relaxed molecular clock by applying a semi-parametric penalized likelihood approach using a truncated Newton optimization algorithm as implemented in the program R8S. The smoothing parameter was determined by cross-validation. Dating constraints are described in Methods. The EMMIX software package was used to fit a mixture model of multivariate normal or *t*-distributed components to a given data set. For each significant component identified by EMMIX, the 95% confidence interval of the mean date estimate was then calculated.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 29 August 2010; accepted 10 February 2011.
Published online 10 April 2011.

1. Ohno, S. *Evolution by Gene Duplication* (Springer, 1970).
2. Lynch, M. *The Origins of Genome Architecture* (Sinauer, 2007).
3. Edger, P. P. & Pires, J. C. Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res.* **17,** 699–717 (2009).
4. De Bodt, S., Maere, S. & Van de Peer, Y. Genome duplication and the origin of angiosperms. *Trends Ecol. Evol.* **20,** 591–597 (2005).
5. Soltis, D. E., Bell, C. D., Kim, S. & Soltis, P. S. Origin and early evolution of angiosperms. *Ann. NY Acad. Sci.* **1133,** 3–25 (2008).
6. Fawcett, J. A., Maere, S. & Van de Peer, Y. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc. Natl Acad. Sci. USA* **106,** 5737–5742 (2009).
7. Lyons, E. *et al.* Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol.* **148,** 1772–1781 (2008).
8. Bowers, J. E., Chapman, B. A., Rong, J. & Paterson, A. H. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422,** 433–438 (2003).
9. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449,** 463–467 (2007).
10. Vision, T. J., Brown, D. G. & Tanksley, S. D. The origins of genomic duplications in *Arabidopsis. Science* **290,** 2114–2117 (2000).
11. Barker, M. S., Vogel, H. & Schranz, M. E. Paleopolyploidy in the Brassicales: analyses of the *Cleome* transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales. *Genome Biol. Evol.* **1,** 391–399 (2009).
12. Tang, H. *et al.* Synteny and collinearity in plant genomes. *Science* **320,** 486–488 (2008).
13. Tang, H. *et al.* Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* **18,** 1944–1954 (2008).
14. Tuskan, G. A. *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313,** 1596–1604 (2006).
15. Tang, H., Bowers, J. E., Wang, X. & Paterson, A. H. Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proc. Natl Acad. Sci. USA* **107,** 472–477 (2010).
16. Cui, L. *et al.* Widespread genome duplications throughout the history of flowering plants. *Genome Res.* **16,** 738–749 (2006).
17. Blomme, T. *et al.* The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol.* **7,** R43 (2006).
18. Ebersberger, I., Strauss, S. & von Haeseler, A. HaMStR: profile hidden Markov model based search for orthologs in ESTs. *BMC Evol. Biol.* **9,** 157 (2009).
19. Moore, M. J., Bell, C. D., Soltis, P. S. & Soltis, D. E. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proc. Natl Acad. Sci. USA* **104,** 19363–19368 (2007).
20. McLachlan, G., Peel, D., Basford, K. E. & Adams, P. The EMMIX algorithm for the fitting of normal and t-components. *J. Stat. Softw.* **4,** i02 (1999).
21. Bell, C. D., Soltis, D. E. & Soltis, P. S. The age of the angiosperms: a molecular timescale without a clock. *Evolution* **59,** 1245–1258 (2005).
22. Schneider, H. *et al.* Ferns diversified in the shadow of angiosperms. *Nature* **428,** 553–557 (2004).
23. Freeling, M. Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annu. Rev. Plant Biol.* **60,** 433–453 (2009).
24. Kassahn, K. S., Dang, V. T., Wilkins, S. J., Perkins, A. C. & Ragan, M. A. Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. *Genome Res.* **19,** 1404–1418 (2009).
25. Devlin, P. F., Patel, S. R. & Whitelam, G. C. Phytochrome E influences internode elongation and flowering time in *Arabidopsis. Plant Cell* **10,** 1479–1487 (1998).
26. Dechaine, J. M., Gardner, G. & Weinig, C. Phytochromes differentially regulate seed germination responses to light quality and temperature cues during seed maturation. *Plant Cell Environ.* **32,** 1297–1309 (2009).
27. Mathews, S., Burleigh, J. G. & Donoghue, M. J. Adaptive evolution in the photosensory domain of phytochrome A in early angiosperms. *Mol. Biol. Evol.* **20,** 1087–1097 (2003).
28. Parry, G. *et al.* Complex regulation of the *TIR1/AFB* family of auxin receptors. *Proc. Natl Acad. Sci. USA* **106,** 22540–22545 (2009).
29. Hu, W., dePamphilis, C. W. & Ma, H. Phylogenetic analysis of the plant-specific zinc finger-homeobox and mini zinc finger gene families. *J. Integr. Plant Biol.* **50,** 1031–1045 (2008).
30. Prigge, M. J. & Clark, S. E. Evolution of the class III HD-Zip gene family in land plants. *Evol. Dev.* **8,** 350–361 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported primarily by NSF Plant Genome Research Program (DEB 0638595, The Ancestral Angiosperm Genome Project) and in part by the Department of Biology and by the Huck Institutes of Life Sciences of the Pennsylvania State University. H.M. was also supported by funds from Fudan University. We thank J. Carlson, M. Frohlich, S. DiLoretto, L. Warg, S. Crutchfield, C. Johnson, N. Naznin, X. Zhou, J. Duarte, B. J. Bliss, J. Der and E. Wafula for help and discussion, D. Stevenson and C. Schultz for *Zamia* samples, J. McNeal, S. Kim and M. Axtell for photographs, and all the members of The Genome Center at Washington University production team, especially L. Fulton, K. Delehaunty and C. Fronick.

Author Contributions Y.J. and C.W.d. designed the study and Y.J. performed the principal data analyses. A.S.C., L.L., P.E.R., Y.H., S.E.S. and H.L. prepared tissues, RNAs, and/or libraries. S.W.C., L.P.T. and S.C.S. generated sequence data. S.A. and J.L.-M. performed the Ancestral Angiosperm Genome Project transcriptome assemblies and MAGIC database construction. Y.J. and C.W.d. drafted the manuscript, and N.J.W., A.S.C., L.L. P.E.R., P.S.S., D.E.S., H.M. and J.L-M. contributed to the planning and discussion of the research and the editing of the manuscript. All authors contributed to and approved the final manuscript.

Author Information Alignments and phylogenetic trees have been deposited in Dryad with package identifier doi:10.5061/dryad.8546. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to C.W.d. (cwd3@psu.edu).

METHODS

Detection of ancient WGD events. Several methodologies have been proposed and widely used to detect the signature of genome duplication. Identification of large syntenic blocks of genes within genomes provides strong evidence to support genome duplication[7,8]. The timing of WGDs is inferred through cross-species genome comparisons, but extensive genome rearrangements and gene loss reduce the size of syntenic blocks over time and obscure identification of ancient pre-γ WGD[31,32]. Another approach is to estimate the age distribution of paralogous gene pairs, where synonymous site divergence (K_s) or non-synonymous site divergence (K_a) is used as a proxy for the age of the duplication event[4,10,16,33]. However, this method may be confounded by excessive gene loss, concentration of duplicate pair estimates on more recent nodes, saturation of K_s between older paralogue pairs, and molecular rate heterogeneity among lineages, gene families or even genes. For example, the β and γ GWDs inferred in analyses of syntenic blocks were not evident in a K_s plot for *Arabidopsis* paralogue pairs[13,33,34]. Therefore, both methods present challenges to inferring ancient genome duplications that may have occurred close to or well before the origin of angiosperms. For this reason, we used phylogenomic analyses to identify ancient gene duplications that occurred before monocots and dicots, and evaluated their phylogenetic timing and estimated age to identify whether there were temporal concentrations of gene duplications (Supplementary Fig. 1).

Phylogenetic analysis. The OrthoMCL method[35] was used to construct a set of core-orthogroups based on protein similarity graphs. This approach has been shown to yield fewer false positives than other methods[36], which is critical for this study. If genes from outside the core-orthogroup in question (false positives) are included in the analysis, the core-orthogroup could be incorrectly scored as retaining ancient duplicates. All orthogroup amino-acid alignments were generated with MUSCLE using default parameters[37]. The multiple sequence alignments were trimmed by removing poorly aligned regions using TRIMAL 1.2 with the option 'automated1'[38]. Additional sorted unigene sequences for the core-orthogroups (retrieved with HaMStR) were aligned at the amino-acid level into the existing 11 species' full alignments (before trimming) using CLUSTALX 1.8[39]. After trimming, each unigene sequence was checked and removed from the alignment if the sequence contained less than 70% alignment coverage. Corresponding DNA sequences were then forced onto the amino-acid alignment using custom Perl scripts and used for subsequent phylogenetic analysis. Maximum-likelihood analyses were conducted using RAXML, version 7.2.1[40,41], invoking a rapid bootstrap (100 replicates) analysis and search for the best-scoring MLT with the general time-reversible model of DNA sequence evolution with gamma-distributed rate heterogeneity (the GTRGAMMA model, which represents an acceptable trade-off between speed and accuracy; RAXML 7.0.4 manual) in a single program run. Alignments and phylogenetic trees are deposited at http://dx.doi.org/10.5061/dryad.8546, and Perl scripts are available on request from C.W.d.

Scoring gene duplications. By carefully interpreting all of the trees, duplication events were identified in rooted trees using *Physcomitrella* genes (or *Selaginella* if there were no *Physcomitrella* genes in the orthogroup) as outgroup sequences. Three relevant bootstrap values were taken into account when evaluating support for a particular duplication. For example, given a topology of (((M1E1)bootstrap1,(M2E2)bootstrap2)bootstrap3), bootstrap1 and bootstrap2 are the bootstrap values supporting the M1E1 clade and the M2E2 clade, respectively, and bootstrap3 is the bootstrap value supporting the large clade including M1E1 and M2E2. A monocot–eudicot duplication supported by 50% (or 80%) means that bootstrap3 and at least one of the bootstrap1 and bootstrap2 values are greater than or equal to 50% (or 80%). When basal angiosperm and/or gymnosperm genes were added, bootstrap1 and bootstrap2 were evaluated for nodes subtending ME + B (Fig. 1a), whereas bootstrap3 was evaluated for the node subtending the large clade including the angiosperm-wide or seed-plant-wide duplications.

Gene tree estimation may be susceptible to long-branch attraction, particularly with sparse taxon sampling (that is, sparse gene sampling in the gene tree context) or when there is mis-specification of the model of molecular evolution used for phylogenetic reconstruction[42,43], leading to erroneous conclusions of topology. For example, an orthogroup with the phylogenetic pattern ((*Oryza*, *Populus*)(*Arabidopsis*)) is consistent with a gene duplication shared by monocots and eudicots, with subsequent paralogue losses in both monocot and eudicot lineages (Fig. 1a, analysis Ib). Alternatively, it is possible that the *Arabidopsis* gene was especially divergent and therefore was placed as sister to the *Oryza–Populus* pair owing to long-branch attraction. Distinguishing between these alternative explanations can be facilitated by increased gene sampling to split long branches[43]. Moreover, inference of gene duplication may be ambiguous if all taxa are represented by a single gene in a given gene tree (as in the example above). With these considerations in mind, we filtered our gene trees, requiring that at least one of the seven core species has retained both paralogues following the inferred gene duplication event in a common monocot–eudicot ancestor. Therefore, an example of the smallest possible gene tree with a monocot–eudicot duplication would be (((*Oryza*, *Vitis*)(*Vitis*))*Selaginella*). On the basis of these criteria, we scored each orthogroup with or without ancient duplications, and counted the total number of orthogroups supporting each hypothesis illustrated in Fig. 1a. Supplementary Data 2 details the number of duplication of each type scored for every orthogroup.

Finite mixture models of genome duplications. To explore the timing of genome duplication events, the inferred distribution of divergence times was fitted to a mixture model comprising several component distributions in various proportions. The EMMIX software[20] can be used to fit a mixture model of multivariate normal or *t*-distributed components to a given data set (http://www.maths.uq.edu.au/~gjm/emmix/emmix.html). The mixed populations were modelled with one to four components. The EM algorithm was repeated 100 times with random starting values, as well as ten times with *k*-mean starting values. The best mixture model was identified using the Bayesian information criterion.

Molecular dating analyses and 95% confidence intervals. The best maximum-likelihood topology for the core-orthogroups or orthogroups was used for divergence time analyses. The divergence time of the two paralogous clades was estimated under the assumption of a relaxed molecular clock by applying a semi-parametric penalized likelihood approach using a truncated Newton optimization algorithm as implemented in the program R8S[44]. The smoothing parameter was determined by cross-validation. We used the following dates in our estimation procedure: minimum age of 400 Myr and maximum age of 450 Myr for the divergence of *P. patens*[45], a fixed constraint age of 400 Myr for the divergence of *S. moellendorffii*[46], a minimum age of 309 Myr for crown-group seed plants[47] (this constraint was used only in analyses reported in Supplementary Fig. 5), a minimum age of 125 Myr for the divergence of monocots and eudicots[48], and a maximum age of 125 Myr for the origin of rosids[48]. We required that trees pass both the cross-validation procedure and provide estimates of the age of the duplication node. The inferred divergence times were then analysed by EMMIX. For each significant component identified by EMMIX, the 95% confidence interval of the mean was then calculated.

Calculation of K_s. Paralogous pairs of sequences were identified from best reciprocal matches in all-by-all BLASTN searches. Only protein sequences more than 200 base pairs in length were used for K_s calculations. Translated sequences of unigenes generated by ESTSCAN were aligned using MUSCLE 3.6[37]. Nucleotide sequences were then forced to fit the amino-acid alignments using PAL2NAL[49]. The K_s (also known as D_s) values were calculated using a simplified version of the Goldman–Yang maximum-likelihood method[50] implemented in the 'codeml' package of PAML[51]. The K_s frequency in each interval size of 0.05 within the range [0, 3.0] was plotted.

Gene ontology enrichment for orthogroups with ancient duplication. Gene ontology (GO) annotations of orthogroups with early ancient duplications were compared with orthogroups that did not have such duplications, to test for enrichment of GO terms[52]. *Arabidopsis* GO slim terms were downloaded and assigned to orthogroups directly if the orthogroup included *Arabidopsis* genes. Otherwise, we searched representative InterPro domains using INTERPROSCAN[53]. Then GO annotations were assigned to the orthogroups using InterPro2GO mapping. Subsequently, all GO annotations were mapped to GO slim categories using the 'map2slim' script. Finally, we evaluated statistical differences in enrichment of GO slim terms using agriGO by Fisher's exact test and the Yekutieli (false-discovery rate under dependency) multi-test adjustment method[54].

31. Buggs, R. J. *et al.* Gene loss and silencing in *Tragopogon miscellus* (Asteraceae): comparison of natural and synthetic allotetraploids. *Heredity* **103,** 73–81 (2009).
32. Vandepoele, K., Simillion, C. & Van de Peer, Y. Detecting the undetectable: uncovering duplicated segments in *Arabidopsis* by comparison with rice. *Trends Genet.* **18,** 606–608 (2002).
33. Blanc, G. & Wolfe, K. H. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16,** 1667–1678 (2004).
34. Van de Peer, Y., Fawcett, J. A., Proost, S., Sterck, L. & Vandepoele, K. The flowering world: a tale of duplications. *Trends Plant Sci.* **14,** 680–688 (2009).
35. Li, L., Stoeckert, C. J. Jr & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13,** 2178–2189 (2003).
36. Proost, S. *et al.* PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *Plant Cell* **21,** 3718–3731 (2009).
37. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
38. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25,** 1972–1973 (2009).
39. Thompson, J. D., Gibson, T. J. & Higgins, D. G. Multiple sequence alignment using ClustalW and ClustalX. *Curr. Protoc. Bioinf.* 2.3.1–2.3.22 (2002).
40. Stamatakis, A., Ludwig, T. & Meier, H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* **21,** 456–463 (2005).
41. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).
42. Felsenstein, J. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27,** 401–410 (1978).

43. Hendy, M. D. & Penny, D. A framework for the quantitative study of evolutionary trees. *Syst. Zool.* **38,** 297–309 (1989).
44. Sanderson, M. J. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19,** 301–302 (2003).
45. Rensing, S. A. *et al.* The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science* **319,** 64–69 (2008).
46. Kenrick, P. & Crane, P. R. The origin and early evolution of plants on land. *Nature* **389,** 33–39 (1997).
47. Miller, C. N. J. Implications of fossil conifers for the phylogenetic relationships of living families. *Bot. Rev.* **65,** 239–277 (1999).
48. Doyle, J. A. & Hotton, C. L. in *Pollen and Spores: Patterns of Diversification* (eds Blackmore, S. & Barnes, S. H.) 169–195 (Clarendon, 1991).
49. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34,** W609–W612 (2006).
50. Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11,** 725–736 (1994).
51. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13,** 555–556 (1997).
52. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.* **25,** 25–29 (2000).
53. Zdobnov, E. M. & Apweiler, R. InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17,** 847–848 (2001).
54. Du, Z., Zhou, X., Ling, Y., Zhang, Z. & Su, Z. agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Res.* **38,** W64–W70 (2010).