doi:10.1038/nature09916

Supplementary Discussion

K_s analysis

The topological evidence overwhelmingly supports a shared genome-scale duplication by all angiosperms. This finding of ancient gene duplication contrasts with earlier studies based on K_s distributions of duplicated genes in basal angiosperms using much smaller numbers of ESTs[16,55]. The previous analyses had detected evidence of an ancient WGD event in several basal angiosperms, but not in *Amborella*, the sister to all other extant angiosperms[56]. However, K_s analysis with the greatly expanded set of ESTs (2,592,984) from *Amborella* can now detect two significant ancient duplication peaks: 1.97 and 2.76 (Supplementary Fig. 7). Furthermore, forty-three *Amborella* unigene pairs were identified from transcriptome K_s analysis where both genes mapped to a phylogenetic tree. Gene pairs with small K_s values (<1.5) were duplicated after the divergence of *Amborella* and the rest of the angiosperms, while gene pairs with large K_s values were generated by ancient duplications in the phylogenetic trees (Supplementary Table 5). The results are generally consistent between the K_s and phylogenomic timing analyses. Therefore, both phylogenetic and K_s analyses of very large EST datasets provide strong evidence for a concentration of duplication events prior to the origin of angiosperms.

Synteny analysis of ancient gene duplications

A graph-based analysis of synteny[57] in the *Vitis* genome was performed in order to further test the hypothesis that two ancient WGDs occurred before the divergence of monocots and eudicots and the more recent paleohexaploidy event[9], γ[12], that has been characterized in *Vitis* and other available eudicot genomes. A total of 2322 sets of *Vitis* genes in 571 orthogroups showing evidence of gene duplication before the monocot-eudicot divergence (Analysis I described above; Supplementary Data 4) were used to test for the existence of syntenic blocks (i.e. collinear redundancy) in addition to those that that have already been described as biproducts of the γ triplication[9,12,58]. Over time, rearrangements and gene loss following WGD are expected to degrade synteny between duplicated blocks in an ancient paleopolyploid genome[7,9,12,15], but using the approach of Dehal and Boore[57] we did find suggestive patterns of loose synteny among multiple segments that are hypothesized to have been derived from pre-γ duplication events

(Supplementary Data 4). Gene phylogenies were used to define genes along each *Vitis* chromosome representing the pre-γ ancestral genome (Supplementary Fig. 8) and matches between these genes were used to anchor searches for 2 or more shared genes within 200 gene windows (100 genes on either side of anchor) across the *Vitis* genome. Whereas a single pre-γ WGD would be diagnosable with up to 4-fold collinear redundancy along each chromosome, two pre-γ WGDs could be evidenced with a maximum of 10-fold collinear redundancy (Supplementary Fig. 8a). However, these levels of collinearity are expected to be rare given the processes of gene fractionation following WGDs[7,9,15,58] and structural mutations independent of WGDs. Inspection of the synteny graphs reveals that 5-fold collinear redundancy is the most common pattern in the *Vitis* genome (Supplementary Fig. 8b), meaning that the largest fraction of genes shared another 4 paralogous regions. This result supports the phylogenomic inference that at least two pre-γ WGDs contributed to the complexity of angiosperm genomes.

MADs-box transcription factors

Many MADS-box transcription factors are important regulators of plant development, particularly as regulators of floral organ identity. Previous phylogenetic analyses of the *AGAMOUS* (*AG*), *APETALA3* (*AP3*)/*PISTILLATA* (*PI*), and *SEPALLATA* (*SEP*) MADS-box subfamilies indicated that these gene families experienced duplication prior to the eudicot-monocot divergence[59-62]. The placement of basal angiosperm genes indicates that the duplication events in the *AG*, *AP3/PI*, and *SEP* subfamilies predate the diversification of extant angiosperms[60-63]. These duplications are therefore consistent with WGD before the origin of the angiosperms. Furthermore, the *SEP* and *AGL6* subfamilies are sister clades formed by a duplication event that likely occurred before the split of angiosperms and gymnosperms[61], a duplication that is possibly the same as the seed plant-wide WGD documented here. The duplication events and subsequent evolution of expression patterns and functions of these MADS-box components of the ABCE model have likely contributed to the wide spectrum of morphological diversification of flowers[61,64].

Supplementary References

55 Soltis, D. E. *et al.* Polyploidy and angiosperm diversification. *Am. J. Bot.* **96**, 336-348 (2009).
56 Soltis, P. S., Soltis, D. E. & Chase, M. W. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature* **402**, 402-404 (1999).
57 Dehal, P. & Boore, J. L. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* **3**, e314 (2005).
58 Sankoff, D. *et al.* Towards improved reconstruction of ancestral gene order in angiosperm phylogeny. *J. Comput. Biol.* **16**, 1353-1367 (2009).
59 Kramer, E. M., Dorit, R. L. & Irish, V. F. Molecular evolution of genes controlling petal and stamen development: duplication and divergence within the *APETALA3* and *PISTILLATA* MADS-box gene lineages. *Genetics* **149**, 765-783 (1998).
60 Kramer, E. M., Jaramillo, M. A. & Di Stilio, V. S. Patterns of gene duplication and functional evolution during the diversification of the *AGAMOUS* subfamily of MADS box genes in angiosperms. *Genetics* **166**, 1011-1023 (2004).
61 Zahn, L. M. *et al.* The evolution of the *SEPALLATA* subfamily of MADS-box genes: a preangiosperm origin with multiple duplications throughout angiosperm history. *Genetics* **169**, 2209-2223 (2005).
62 Zahn, L. M. *et al.* Conservation and divergence in the *AGAMOUS* subfamily of MADS-box genes: evidence of independent sub- and neofunctionalization events. *Evol. Dev.* **8**, 30-45 (2006).
63 Stellari, G. M., Jaramillo, M. A. & Kramer, E. M. Evolution of the *APETALA3* and *PISTILLATA* lineages of MADS-box-containing genes in the basal angiosperms. *Mol. Biol. Evol.* **21**, 506-519 (2004).
64 Ma, H. & dePamphilis, C. The ABCs of floral evolution. *Cell* **101**, 5-8 (2000).

Supplementary Table 1: Summary of datasets for nine sequenced plant genomes included in this study. Analyzed gene number is the number of genes contained in the core-orthogroups having at least one monocot and one eudicot, and one *Selaginella* and/or *Physcomitrella* sequence, which is the minimum requirement for the detection of a possible ancient duplication prior to the divergence of monocots and eudicots by a phylogenetic approach.

Species	Annotation version	Annotated genes	Analyzed genes
Arabidopsis thaliana Thale cress	TAIR version 9	27379	11669
Carica papaya Papaya	ASGPB release	25536	8713
Cucumis sativus Cucumber	BGI release	21635	9985
Populus trichocarpa Black cottonwood	JGI version 2.0	41377	15050
Vitis vinifera Grape vine	Genoscope release	30434	11020
Oryza sativa Rice	RGAP release 6.1	56979	14483
Sorghum bicolor	JGI version 1.4	34496	11258
Selaginella moellendorffii	JGI version 1.0	34697	16711
Physcomitrella patens	JGI version 1.1	35938	12551

Supplementary Table 2: Summary of unigene sequences of basal angiosperm and gymnosperm ESTs and unigenes included in phylogenetic study. Gymnosperm data (except *Zamia vazquezii*) are from TIGR PTA database (http://plantta.jcvi.org/). Sequences and assemblies for basal angiosperms and *Zamia vazquezii* data (12,660,332 previously unreported ESTs) are available at the AAGP project website (Ancestral Angiosperm Genome Project) (http://ancangio.uga.edu/); data for these species will be described in detail in additional papers. Legend: # EST = total number of ESTs in the database; # Unigenes = total number of unigenes; # Included = total number of unigenes assembled in the core-orthogroups with one or more monocot + eudicot duplications.

	SPECIES (COMMON NAME)	# EST	# Unigenes	# Included
Gymnosperms	*Chamaecyparis obtusa* (Hinoki false cypress)	5830	4061	583
	Cryptomeria japonica (Japanese cedar)	16187	9098	1121
	Cycas rumphii (Cycad)	7899	4335	616
	Ginkgo biloba (Ginkgo)	5940	4178	478
	Gnetum gnemon (Melinjo)	3920	2859	195
	Picea abies (Norway spruce)	10030	5204	608
	Picea engelmannii x Picea glauca	28160	14201	1831
	Picea glauca (White spruce)	132151	49412	7782
	Picea sitchensis (Sitka spruce)	98987	25425	3047
	Pinus pinaster (Maritime pine)	13067	9166	2336
	Pinus taeda (Loblolly pine)	326641	78873	11006
	Pseudotsuga menziesii (Douglas fir)	18100	12074	291
	Taiwania cryptomerioides (Coffin tree)	1407	778	66
	Welwitschia mirabilis (Tree tumbo)	10122	6680	1408
	"Zamia fischeri"	8248	7374	345
	Zamia vazquezii	603139	50336	4067
Basal Angiosperms	*Aristolochia fimbriata (Dutchman's pipe)*	3828275	155371	5154
	Liriodendron tulipifera (Yellow-poplar)	2012281	141494	11582
	Nuphar advena (Yellow pond lily)	3623653	289773	27588
	Amborella trichopoda	2592984	208394	11760
	Total number of sequences	**13347021**	**1079086**	**91864**

Supplementary Table 3: Floral gene regulators surviving ancient duplications. In this study we identified 35 orthogroups that included genes known to regulate aspects of reproductive development in plants and containing at least one ancient gene duplication. “ME DUP” shows the number of duplications identified before the divergence of monocots and eudicots from 9-genome phylogenies. “Angio DUP” means number of angiosperm-wide duplications identified from phylogenetic trees that include basal angiosperms and gymnosperms. “Seed DUP” shows the number of seed plant-wide duplications indicated from phylogenetic trees that include basal angiosperms and gymnosperms. Numbers missing for both columns Angio DUP and Seed DUP mean the orthogroups have not been populated with unigenes of basal angiosperms and gymnosperms.

Ortho ID	Representative Gene	Annotation	ME DUP	Angio DUP	Seed DUP
34	AT1G75820	CLV1, controls shoot and floral meristem size, and contributes to establish and maintain floral meristem identity	2		
58	AT5G41170	PPR, Pentatricopeptide repeat, expressed during petal differentiation and expansion stage	1		
87	AT1G68530	CUT1, required for cuticular wax biosynthesis and pollen fertility	1		
112	AT4G04890	PDF2, encodes a homeodomain protein that is expressed in the LI layer of the vegetative, floral and inflorescence meristems	1	1	1
126	AT5G13930	CHS, Chalcone synthase family, a key enzyme involved in the biosynthesis of flavonoids	1	1	
163	AT5G43810	ARGONAUTE, along with WUS and CLV genes, controls the relative organization of central zone and peripheral zone cells in meristems	1		

166	AT3G61160	Shaggy- elated protein kinase beta / ASK-beta (ASK2), involved in protein amino acid phosphorylation	1	1	1
242	AT1G07920	EF-1-alpha, response to cadmium ion	1		
245	AT2G34710	PHB, involved in adaxial/abaxial pattern formation, determination of bilateral symmetry, integument development, meristem initiation, polarity specification of adaxial/abaxial axis, primary shoot apical meristem specification, regulation of transcription, DNA-dependent	1	1	1
309	AT1G01040	CAF, mutants convert the floral meristems to an indeterminate state, others yet show defects in ovule development	1		
361	AT2G18790	PHYTOCHROME, regulates the time of flowering and seed germination	1	1	1
423	AT1G30330	ARFs, Auxin response factors, act redundantly with ARF8 to control stamen elongation and flower maturation	1		
454	AT4G32551	LEUNIG, regulates floral organ identity,gynoecium and ovule development. Negatively regulates AGAMOUS	1		1
576	AT1G55680	WD40 repeat, other	1		1
595	AT4G37750	ANT, required for control of cell proliferation and encodes a putative transcriptional regulator similar to AP2	1		
643	AT1G66340	ETR1, Ethylene receptor, similar to prokaryote sensory transduction proteins	1		
651	AT3G58510	DEAD box RNA helicase	1		
700	AT2G42830	SHP2, SHATTERPROOF 2 (AGL5), AG, MADS box protein	1		1
752	AT1G59750	AUXIN RESPONSE FACTOR 1	1		
876	AT5G57050	ABI2, ABA INSENSITIVE	1		
1088	AT4G08920	HY4, ELONGATED HYPOCOTYL 4 (CRY1)	1		
1141	AT5G08390	WD40 repeat	1		
1168	AT2G38630	WD40 repeat-like	1		

1172	AT3G61240	DEAD/DEAH box helicase	1		
1412	AT1G68050	FKF1, FLAVIN-BINDING KELCH DOMAIN F BOX PROTEIN, is clock-controlled and regulates transition to flowering	1		
1429	AT3G01540	DEAD Box RNA helicases, CAF-like	1		
1538	AT1G53230	CYC3, CYCLOIDEA, TCP, involved in heterchronic regulation of leaf differentiation			
1676	AT2G23380	ICU1, INCURVATA 1, required for stable repression of AG and AP3	1		
1711	AT5G63120	DEAD box RNA helicase, putative (RH20)	1		1
1848	AT1G69310	WRKY	1		
2833	AT3G23350	ENTH domain-containing protein/clathrin assembly protein-related, expressed in leaf whorl, petal, flower; EXPRESSED DURING: 4 anthesis, petal differentiation and expansion stage	1	1	
2920	AT5G10630	EF-1-alpha, putative, GTP binding, translation elongation factor activity, GTPase activity, zinc ion binding	1		
3676	AT3G56400	WRKY	1		
4072	AT2G44745	WRKY	1		
4129	AT2G37630	AS1, Asymmetric leaves1, encodes a MYB-domain protein involved in specification of the leaf proximodistal axis	1		

Supplementary Table 4: List of plant photographs and photo credits used in Fig. 3.

Top row, left to right: eudiocts

Arabidopsis thaliana: Yi Hu

Aquilegia chrysantha: G.A. Cooper (http://persoon.si.edu/PlantImages)

Cirsium pumilum: West Virginia University Herbarium (http://persoon.si.edu/PlantImages)

Eschscholzia californica: Yi Hu

Second row, left to right: monocots

Trilium erectum: Joel McNeal

Bromus kalmii: Joel McNeal

Arisaema triphyllum: Joel McNeal

Cypripedium acaule: Joel. McNeal

Third row, left to right: basal angiosperms

Amborella trichopoda: Sangtae Kim

Liriodendron tulipifera: Haiying Liang

Nuphar advena: Yi Hu

Aristolochia fimbriata: Stefan Wanke

Fourth row, left to right

Zamia vazquezii (a cycad, gymnosperm): Dennis Stevenson

Pseudotsuga menziesii (Douglas fir, a gymnosperm): B. Legler (http://www.biodiversity.wa.gov)

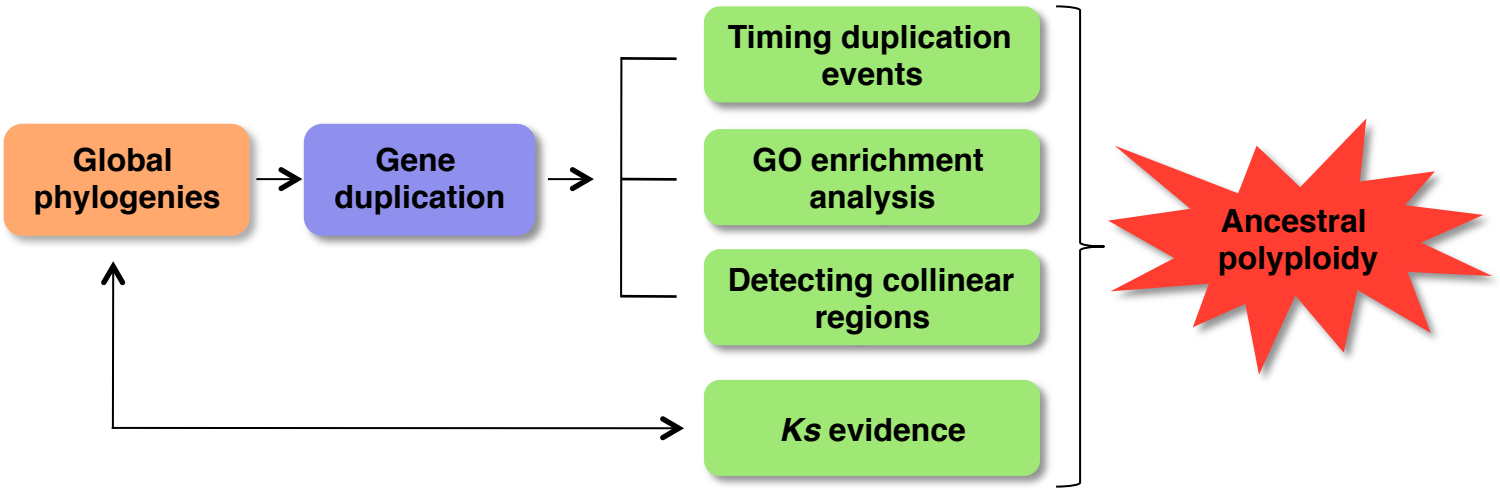
Selaginella mollendorfii (lycophyte, vegetative): Mike Axtell

Physcomitrella patens (moss): Mike Axtell

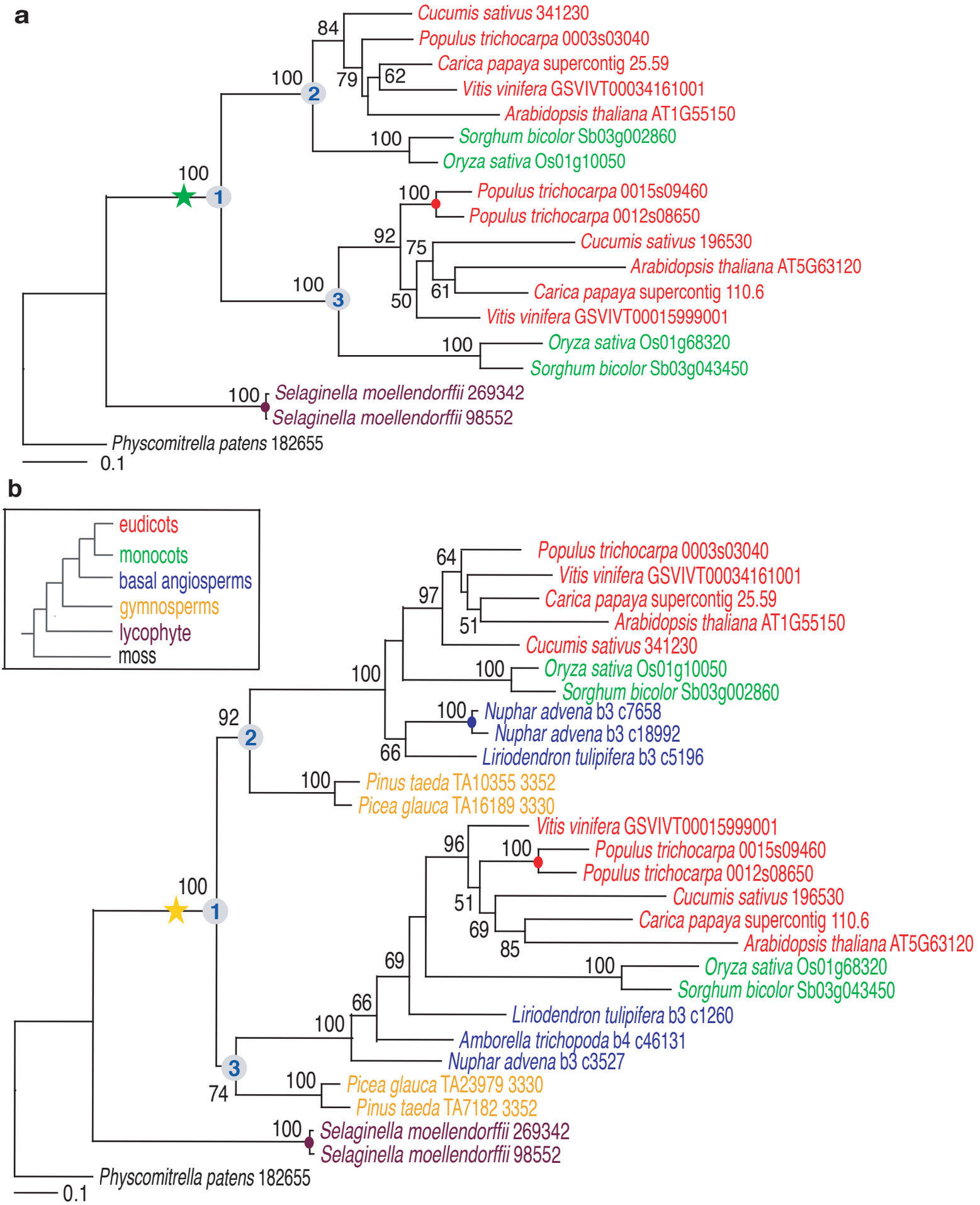
Supplementary Table 5: Consistency of inferred duplications from independent K_s analysis of *Amborella* transcriptome and phylogenomic analysis. 43 unigene gene pairs were identified from transcriptome K_s analysis where both genes mapped to a phylogenetic tree in Analysis IV. "Duplication Time" shows the duplication pattern inferred from phylogenetic tree (Seed dup = Seed plant-wide duplication; Angiosperm dup = Angiosperm-wide duplication, with gymnosperm outgroup root; "Ancient dup" refers to gene families with angiosperm-wide duplication, but without a gymnosperm outgroup root; "Recent dup" indicates that the duplication occurred after the divergence between *Amborella* and rest of the angiosperms). The results are generally consistent between the K_s and phylogenomic timing analyses.

Gene 1	Gene 2	K_s	ORTHO	Duplication Time
Amborella_b4_c10793	Amborella_b4_c18179	2.9327	1616	Seed dup
Amborella_b4_c304	Amborella_b4_c6763	2.9201	1104	Angiosperm dup
Amborella_b4_c24813	Amborella_b4_c903	2.8264	384	Seed dup
Amborella_b4_c13924	Amborella_b4_c5891	2.773	997	Angiosperm dup
Amborella_b4_c52686	Amborella_b4_c642	2.7691	1051	Seed dup
Amborella_b4_c2641	Amborella_b4_c4013	2.7257	385	Ancient dup
Amborella_b4_c3510	Amborella_b4_c416	2.6857	1045	Angiosperm dup
Amborella_b4_c45137	Amborella_b4_c8336	2.6479	632	Seed dup
Amborella_b4_c182	Amborella_b4_c2514	2.6434	849	Seed dup
Amborella_b4_c1485	Amborella_b4_c45698	2.3394	2313	Ancient dup
Amborella_b4_c45	Amborella_b4_c650	2.2687	166	Angiosperm dup
Amborella_b4_c1979	Amborella_b4_c569	2.1382	2384	Ancient dup
Amborella_b4_c464	Amborella_b4_c662	2.0983	513	Angiosperm dup
Amborella_b4_c12254	Amborella_b4_c1873	1.9299	50	Angiosperm dup
Amborella_b4_c173	Amborella_b4_c5177	1.9165	932	Angiosperm dup
Amborella_b4_c7837	Amborella_b4_c8759	1.9157	174	Angiosperm dup
Amborella_b4_c5282	Amborella_b4_c6224	1.8744	1648	Angiosperm dup
Amborella_b4_c1711	Amborella_b4_c54947	1.8423	606	Angiosperm dup
Amborella_b4_c15841	Amborella_b4_c5262	1.7058	231	Angiosperm dup
Amborella_b4_c2852	Amborella_b4_c7861	1.529	468	Angiosperm dup
Amborella_b4_c13559	Amborella_b4_c7335	1.4656	611	Angiosperm dup
Amborella_b4_c13082	Amborella_b4_c37	1.45	7000	Ancient dup

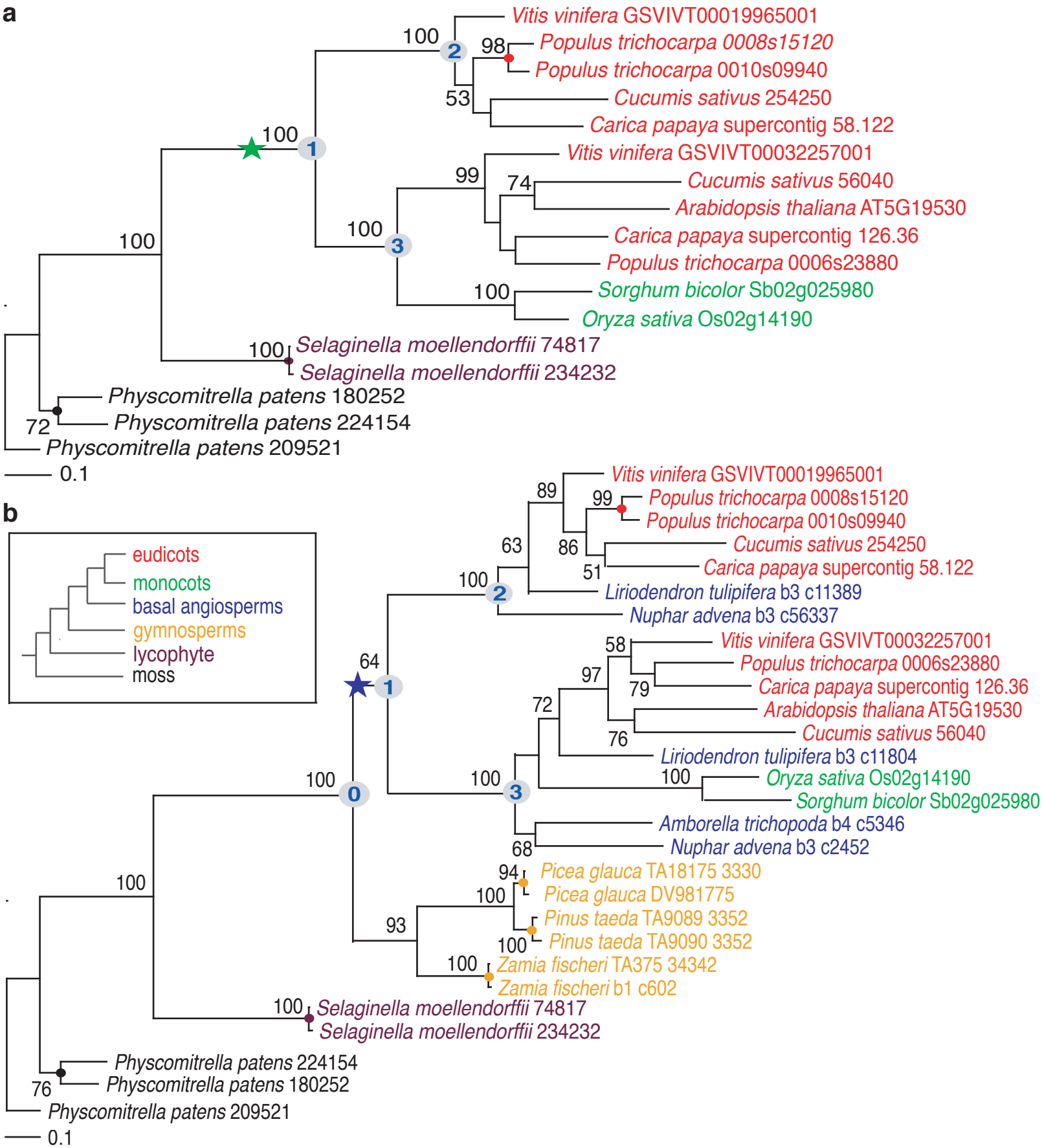
Amborella_b4_c10026	Amborella_b4_c1753	1.533	2638	Recent dup
Amborella_b4_c7113	Amborella_b4_c771	1.5069	611	Recent dup
Amborella_b4_c5525	Amborella_b4_c772	1.1129	2381	Recent dup
Amborella_b4_c8201	Amborella_b4_c940	0.9651	32	Recent dup
Amborella_b4_c4164	Amborella_b4_c584	0.8319	1222	Recent dup
Amborella_b4_c2722	Amborella_b4_c4376	0.7287	642	Recent dup
Amborella_b4_c19277	Amborella_b4_c4233	0.3018	1736	Recent dup
Amborella_b4_c17	Amborella_b4_c37511	0.2599	1638	Recent dup
Amborella_b4_c24344	Amborella_b4_c694	0.2535	1346	Recent dup
Amborella_b4_c1484	Amborella_b4_rep_c89517	0.2235	4175	Recent dup
Amborella_b4_c3486	Amborella_b4_c7007	0.2127	2443	Recent dup
Amborella_b4_c13428	Amborella_b4_c24767	0.1584	441	Recent dup
Amborella_b4_rep_c43092	Amborella_b4_rep_c44167	0.1514	1041	Recent dup
Amborella_b4_rep_c43163	Amborella_b4_rep_c91576	0.1418	2903	Recent dup
Amborella_b4_c25165	Amborella_b4_c8936	0.1393	1289	Recent dup
Amborella_b4_c20850	Amborella_b4_c7029	0.1377	707	Recent dup
Amborella_b4_rep_c48144	Amborella_b4_rep_c73838	0.1325	1651	Recent dup
Amborella_b4_c4904	Amborella_b4_rep_c43438	0.1263	2263	Recent dup
Amborella_b4_rep_c43168	Amborella_b4_rep_c45186	0.1227	3450	Recent dup
Amborella_b4_c427	Amborella_b4_rep_c43972	0.1168	407	Recent dup
Amborella_b4_c13545	Amborella_b4_rep_c42788	0.116	4233	Recent dup



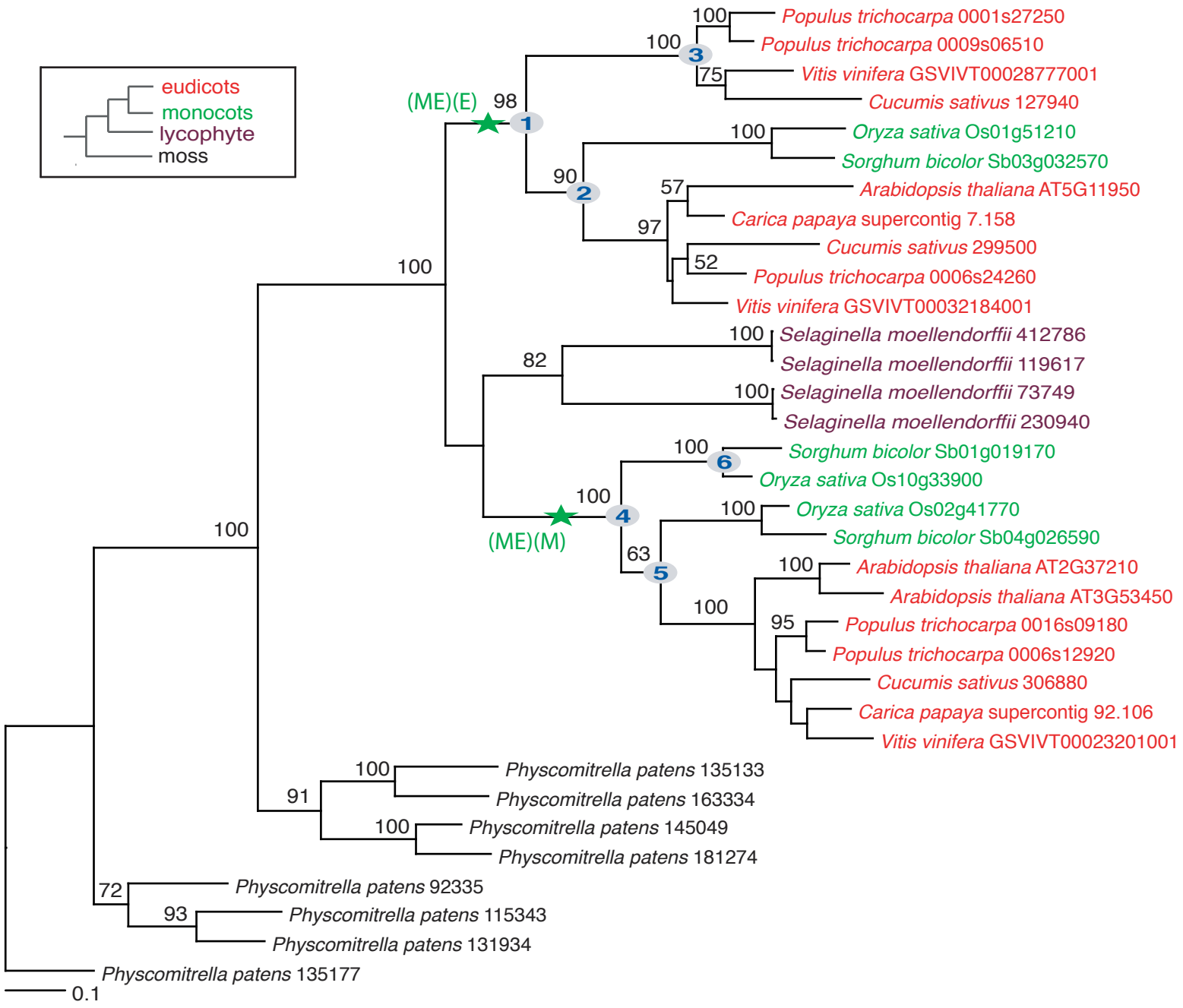
Supplementary Figure 1. Schematic diagram detailing the main flows of data analysis. In this study, we used a phylogenomic approach, along with supporting evidence from *Ks* analysis and collinear investigation, to unravel ancestral polyploidy events (WGD) that occurred before the split of monocots and eudicots.



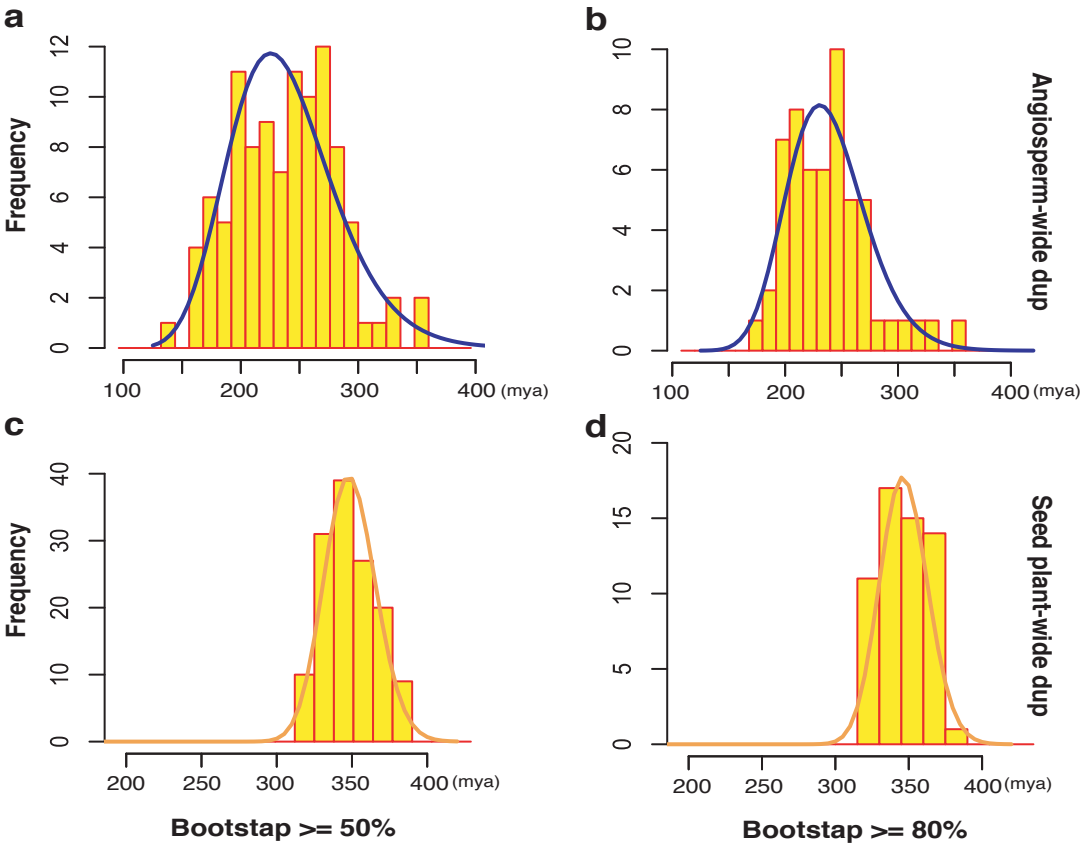
Supplementary Figure 2. Exemplar ML phylogenies consistent with seed plant-wide duplication. (a) RaxML topology of a core-orthogroup (Ortho 1711) where two major clades have survived the shared monocot and eudicot duplication. Since both of the duplicated clades (nodes #2 and #3) contained monocot and eudicot genes, we defined this duplication pattern as (ME)(ME). The scored BS value for this duplication is over 80%, because nodes #1 and #2 (and/or #3) have BS>80% (see "Scoring gene duplications" in Methods). (b) RaxML phylogeny of the core-orthogroup (Ortho 1711) with basal angiosperm and gymnosperm sequences added whose topology is consistent with seed plant-wide duplication. The scored BS value is over 80%, because nodes #1 and #2 have BS>80%. Legend: Green star = monocot+eudicot duplication; yellow star = seed plant duplication; colored circles = recent independent duplications; numbers = bootstrap support values.



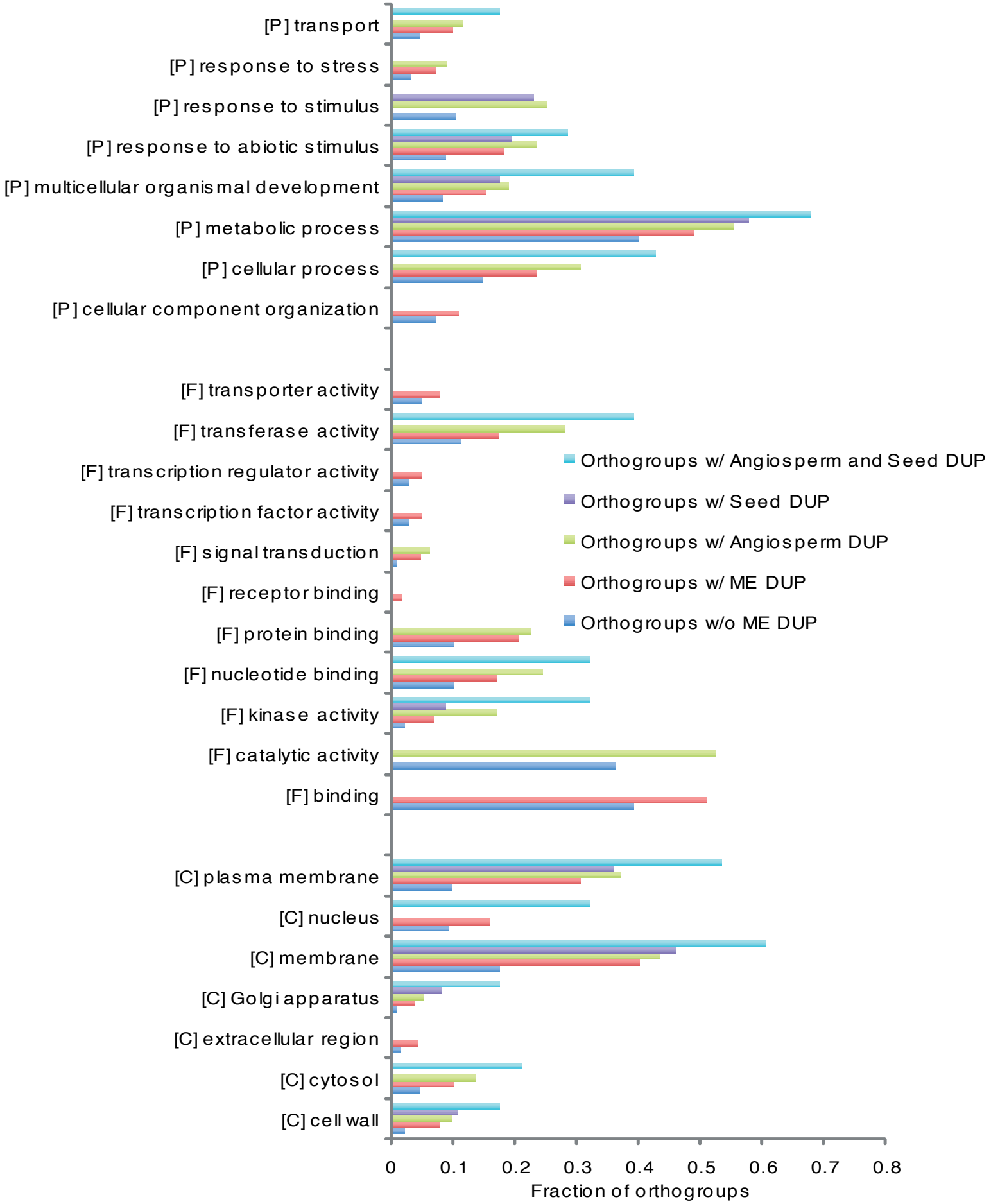
Supplementary Figure 3. Exemplar ML phylogenies consistent with angiosperm-wide duplication. (a) RaxML topology of a core-orthogroup (Ortho 2312) where two major clades have survived the shared monocot and eudicot duplication. The upper clade (node #2) only retains eudicot genes, while the lower clade (node #3) retains both monocot and eudicot genes. We defined this duplication pattern as (ME)(E). The scored BS value for this duplication is over 80% because nodes #1 and #2 (and/or #3) have BS values over 80%. (b) RaxML phylogeny of the Ortho 2312 with basal angiosperm and gymnosperm sequences added whose topology is consistent with an angiosperm-wide duplication not shared with gymnosperms. The scored BS is over 50%, because node #1 is over 50% and less than 80%. Symbols and colors same as for Supplementary Figure 2.



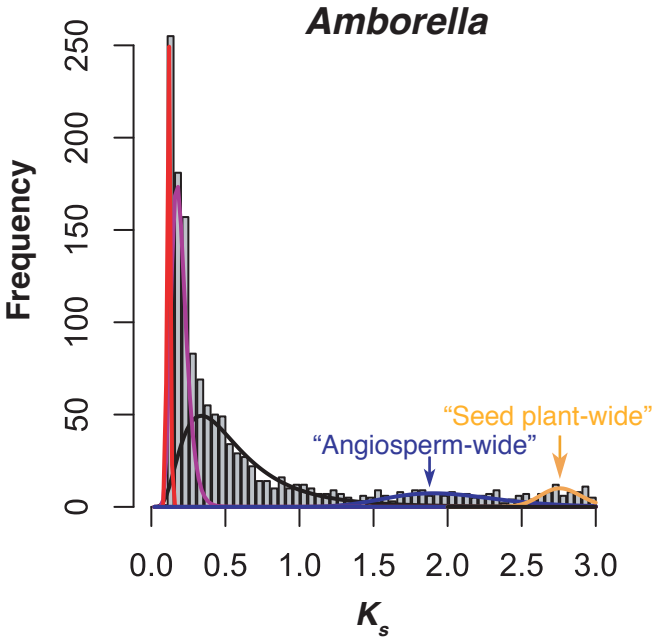
Supplementary Figure 4. Exemplar ML phylogeny contains two types of ME duplication.
RaxML topology of a core-orthogroup (Ortho 396) with two types of shared monocot and eudicot duplications surviving. The upper part of the tree was scored as (ME)(E) with bootstrap support over 80% (Type b), since both of the BS values of node #1 and #2 were over 80%. The lower part tree was scored as (ME)(M) with bootstrap support over 50% (Type c). If one of the paralogous clades had lost all monocot or eudicot genes, the BS value of the ME clade, together with the BS of the large clade, would have been used to determine the bootstrap support level of the duplication. For the lower part of the tree, the duplication was scored BS>50%, because the BS of node #5 is 63%, even though node #4 has BS>80%. This orthogroup was counted once as Type b and once as Type c of Analysis I. Symbols and colors same as for Supplementary Figure 2.



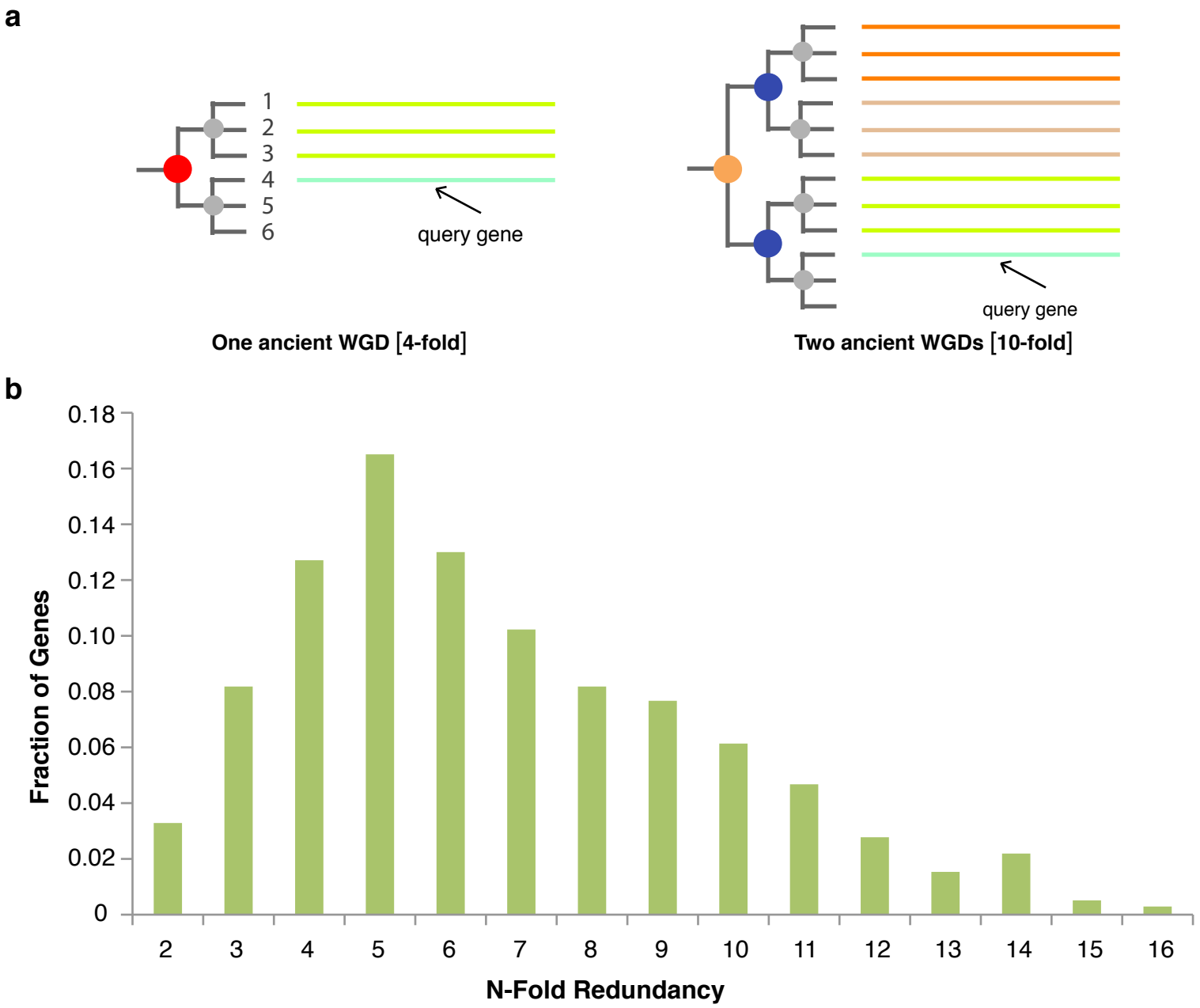
Supplementary Figure 5: Age distributions of angiosperm-wide and seed plant-wide duplications estimated from phylogenies with gymnosperms. In analysis III, both hypothetical topologies (type a and b) were supported by a large number of orthogroups, in which type a supports seed plant-wide duplication and type b support angiosperm-wide duplication. The divergence times of each type were analyzed by EMMIX. (a) and (b) Distributions of inferred divergence times from 110 and 59 orthogroups support angiosperm-wide duplication with bootstrap values over 50% and 80%, respectively. One significant component was identified by the mixture model in both panel (a): blue-234(mya)-1, and in panel (b): blue-236(mya)-1. (c) and (d) Distributions of inferred divergence times from 147 and 62 orthogroups support seed plant-wide duplication with bootstrap values over 50% and 80%, respectively. One significant component in (c): yellow-349(mya)-1, and one significant component in (d): yellow-347(mya)-1.



Supplementary Figure 6: Functional categorization of orthogroups by GO annotation. The orthogroups surviving ancient duplication (ME DUP , Angiosperm DUP, Seed plant DUP) and orthogroups without any of these ancient duplication were categorized by GO annotation. The results of statistical analysis were shown in Supplementary Data 3. The *X*-axis is the fraction of orthogroups mapped to the GO term and represents the abundance of the GO term. The fraction of orthogroups was calculated by the number of orthogroups mapped to the GO term divided by the number of all orthogroups in each category. [C] means GO cellular component categorization; [F] means GO functional categorization; [P] means GO biological process categorization.



Supplementary Figure 7: K_s distribution of 1365 paralogues in *Amborella* support ancient genome duplications. Pairwise K_s divergences for reciprocal 'best hit' genes in *Amborella* EST assembly (2,592,984 ESTs). Paralogous pairs of sequences were identified from best reciprocal matches in all-by-all BLASTN searches. Methods for sequence alignment and estimation of K_s were as reported (Cui, *et al.* 2006) except that only protein-coding sequences with inferred amino acid lengths >200bp were used for K_s calculations. Colored lines superimposed on K_s distribution represent significant duplication components identified by likelihood mixture model (see Methods). Graph shows "color-mean K_s-proportion" where color is the component (curve) color, and proportion is percentage of duplication nodes assigned to the identified component. Five statistically significant components: red-0.1164-0.10, purple-0.1868-0.32, black-0.4801-0.43, blue-1.9751-0.10, and yellow-2.7643-0.05.



Supplementary Figure 8: The estimate of *N*-fold redundancy. (a) The expected fold redundancy for hypotheses of one ancient WGD and two ancient WGDs in the history of the *Vitis* lineage. If one ancient WGD before the monocot-eudicot separation, six *Vitis* genes would be expected on the phylogenetic tree if there is no gene loss (a, left). In this case, we would identify 9 gene pairs supporting an ancient duplication before monocot-eudicot separation, which are (1,4), (1,5), (1,6), (2,4), (2,5), (2,6), (3,4), (3,5), (3,6). Each gene would have another 3 paralogous genes on the phylogenetic tree, not including younger duplicates generated by the γ triplication . For example, gene 4 (query gene) would detect gene 1, 2, 3 as homologous genes. Therefore, the genome region where gene 4 was located would be expected to find another three paralogous regions across the *Vitis* genome. Therefore, one ancient WGD would lead to 4-fold redundancy (including the query). Using the same logic, two ancient WGDs would lead to 10-fold redundancy. Red filled circle refers to one ancient WGD predating the monocot-eudicot split. Yellow filled circle indicates the seed plant-wide duplication (ζ). Blue filled circles refer to the angiosperm-wide duplication (ε). Gray filled circles denote the triplication γ event. (b) The histogram is generated by counting the redundancy across all *Vitis* chromosomes for each query gene as shown in the lower part of Supplementary Data 4. The peak at 5-fold coverage (including the query gene) means that the largest fraction of genes could detect another 4 paralogons in other regions of the genome. This is consistent with two ancient WGDs plus a more recent γ event.