www.sciencemag.org/content/342/6165/1241089/suppl/DC1

Supplementary Materials for

The *Amborella* Genome and the Evolution of Flowering Plants

Amborella Genome Project

Address for general correspondence: cwd3@psu.edu

Published 20 December 2013, *Science* **342**, 1241089 (2013)
DOI: 10.1126/science.1241089

The PDF file includes:

Text
Figs. S1 to S42
Tables S1 to S46
Additional Acknowledgment
References

The *Amborella* Genome and the Evolution of Flowering Plants

Amborella* Genome Project‡

‡All authors and their affiliations and contributions are listed at the end of the paper.

*Address for general correspondence: cwd3@psu.edu. Contact information for working groups is provided in the authorship details.

Authorship information

Authorship of this paper should be cited as "*Amborella* Genome Project." Participants are arranged by working group and then are listed in alphabetical order. Major contributions (†) and the author for correspondence (*) are indicated within each working group. Joshua P. Der, Srikar Chamala, Andre S. Chanderbali, and James C. Estill made major and equal contributions to this project.

Research leadership: Victor A. Albert,[1] W. Bradley Barbazuk,[2,3] Claude W. dePamphilis[4,5,6,*] (cwd3@psu.edu), Joshua P. Der,[4,6,†] James Leebens-Mack,[7] Hong Ma,[4,8] Jeffrey D. Palmer,[9] Steve Rounsley,[10,11] David Sankoff,[12] Stephan C. Schuster,[6,13,14] Douglas E. Soltis,[2,3,15] Pamela S. Soltis,[3,15] Susan R. Wessler,[16] Rod A. Wing[10,17]

Genome sequencing and assembly: Victor A. Albert,[1] Jetty S. S. Ammiraju,[10,17] W. Bradley Barbazuk[2,3,*] (bbarbazuk@ufl.edu), Srikar Chamala,[2,†] Andre S. Chanderbali,[2] Claude W. dePamphilis,[4,5,6] Joshua P. Der,[4,6] Ronald Determann,[18] James Leebens-Mack,[7] Hong Ma,[4,8] Paula Ralph,[4] Steve Rounsley,[10,11] Stephan C. Schuster,[6,13,14] Douglas E. Soltis,[2,3,15] Pamela S. Soltis,[3,15] Jason Talag,[10,17] Lynn Tomsho,[13] Brandon Walts,[2] Stefan Wanke,[20] Rod A. Wing[10,17]

Cytogenetics: Victor A. Albert,[1] W. Bradley Barbazuk,[2,3] Srikar Chamala,[2] Andre S. Chanderbali,[2,†] Tien-Hao Chang,[1] Ronald Determann,[18] Tianying Lan,[1,19] Douglas E. Soltis[2,3,15,*] (dsoltis@ufl.edu), Pamela S. Soltis[3,15]

Genome annotation and database development: Siwaret Arikit,[21] Michael J. Axtell,[4,5] Saravanaraj Ayyampalayam,[7] W. Bradley Barbazuk,[2,3] James M. Burnette, III,[16] Srikar Chamala,[2] Emanuele De Paoli,[22] Claude W. dePamphilis,[4,5,6] Joshua P. Der,[4,6] James C. Estill,[7,†] Nina P. Farrell,[4] Alex Harkess,[7] Yuannian Jiao,[4,23] James Leebens-Mack[7,*] (jleebensmack@plantbio.uga.edu), Kun Liu,[16] Wenbin Mei,[2] Blake C. Meyers,[21] Saima Shahid,[5] Eric Wafula,[4] Brandon Walts,[2] Susan R. Wessler,[16] Jixian Zhai,[21] Xiaoyu Zhang[7]

Synteny analysis: Victor A. Albert[1,*] (vaalbert@buffalo.edu), Lorenzo Carretero-Paulet,[1] Claude W. dePamphilis,[4,5,6] Joshua P. Der,[4,6] Yuannian Jiao,[4,22] James Leebens-Mack,[7] Eric Lyons,[10,24] David Sankoff,[12,†] Haibao Tang,[25] Eric Wafula,[4] Chunfang Zheng[12]

Global gene family analysis: Victor A. Albert,[1] Naomi S. Altman,[26] W. Bradley Barbazuk,[2,3] Lorenzo Carretero-Paulet,[1] Claude W. dePamphilis[4,5,6,*] (cwd3@psu.edu), Joshua P. Der,[4,6,†] James C. Estill,[7] Yuannian Jiao,[4,23,†] James Leebens-Mack,[7] Kun Liu,[16] Wenbin Mei,[2] Eric Wafula[4]

Targeted gene family curation and analysis: Naomi S. Altman,[26] Siwaret Arikit,[21] Michael J. Axtell,[4,5] Srikar Chamala,[2] Andre S. Chanderbali,[2] Feng Chen,[27] Jian-Qun Chen,[28] Vincent Chiang,[29] Emanuele De Paoli,[22] Claude W. dePamphilis,[4,5,6] Joshua P. Der[4,6,*] (jpd18@psu.edu), Ronald Determann,[18] Bruno Fogliani,[30,31] Chunce Guo,[32] Jesper Harholt,[33] Alex Harkess,[7] Claudette Job,[34] Dominique Job,[34] Sangtae Kim,[35] Hongzhi Kong,[32] James Leebens-Mack,[7] Guanglin Li,[27] Lin Li,[32] Jie Liu,[29] Hong Ma,[4,8] Blake C. Meyers,[21] Jongsun Park,[35] Xinshuai Qi,[2,36] Loïc Rajjou,[37] Valérie Burtet-Sarramegna,[30] Ron Sederoff,[29] Saima Shahid,[5] Douglas E. Soltis,[2,3,15] Pamela S. Soltis,[3,15] Ying-Hsuan Sun,[38] Peter Ulvskov,[33] Matthieu Villegente,[30] Jia-Yu Xue,[28] Ting-Feng Yeh,[39] Xianxian Yu,[32] Jixian Zhai[21]

Population genomics: Juan J. Acosta,[40] Victor A. Albert,[1] W. Bradley Barbazuk,[2,3] Riva A. Bruenn,[4,41] Srikar Chamala,[2] Alexandre de Kochko,[42] Claude W. dePamphilis,[4,5,6] Joshua P. Der,[4,6] Luis R. Herrera-Estrella,[43] Enrique Ibarra-Laclette,[43] Matias Kirst,[40,3] James Leebens-Mack,[7] Solon P. Pissis,[15,44] Valérie Poncet,[42] Stephan C. Schuster,[6,13,14] Douglas E. Soltis,[2,3,15] Pamela S. Soltis[3,15,*] (psoltis@flmnh.ufl.edu), Lynn Tomsho[13]

Affiliations:
[1]Department of Biological Sciences, University at Buffalo, Buffalo, NY 14260, USA
[2]Department of Biology, University of Florida, Gainesville, FL 32611, USA
[3]University of Florida Genetics Institute, University of Florida, Gainesville, FL 32610, USA
[4]Department of Biology and Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA
[5]Intercollege Plant Biology Graduate Program, The Pennsylvania State University, University Park, PA 16802, USA
[6]Center for Comparative Genomics and Bioinformatics, The Pennsylvania State University, University Park, PA 16802 USA
[7]Department of Plant Biology, University of Georgia, Athens, GA 30602, USA
[8]State Key Laboratory of Genetic Engineering and Institute of Genetics, Institute of Plant Biology, Center for Evolutionary Biology, Institutes of Biomedical Sciences, School of Life Sciences, Fudan University, Shanghai 200433, CHINA
[9]Department of Biology, Indiana University, Bloomington, IN 47405, USA
[10]School of Plant Science & BIO5 Institute for Collaborative Research, University of Arizona, Tucson, AZ 85721, USA
[11]Dow Agrosciences, Indianapolis, IN 46268, USA
[12]Department of Mathematics and Statistics, University of Ottawa, Ottawa, ONT, K1N 6N5, CANADA
[13]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA 16802, USA
[14]Singapore Centre on Environmental Life Sciences Engineering, SINGAPORE
[15]Florida Museum of Natural History, University of Florida, Gainesville, FL 32611, USA
[16]Department of Botany and Plant Sciences, University of California, Riverside, Riverside, CA 92521, USA
[17]Arizona Genomics Institute, University of Arizona, Tucson, AZ 85721, USA
[18]Atlanta Botanic Garden, Atlanta, GA 30309, USA
[19]Department of Biology, Chongqing University of Science and Technology, Chongqing 4000042, CHINA
[20]Technische Universität Dresden, Institut für Botanik, 01062 Dresden, GERMANY
[21]Delaware Biotechnology Institute, University of Delaware, Newark, DE 19711, USA
[22]Dipartimento di Scienze Agrarie ed Ambientali, Università degli Studi di Udine, via delle Scienze 206, 33100 Udine, ITALY
[23]Plant Genome Mapping Laboratory, University of Georgia, Athens, GA 30602, USA
[24]iPlant Collaborative, University of Arizona, Tucson, AZ 85721, USA
[25]J. Craig Venter Institute, Rockville, MD 20850, USA
[26]Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA
[27]Department of Plant Sciences, University of Tennessee, Knoxville, TN 37996, USA
[28]School of Life Sciences, Nanjing University, Nanjing 210093, CHINA

[29]Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, NC 27695, USA
[30]Laboratoire Insulaire du Vivant et de l'Environnement, University of New Caledonia, BP R4, 98851 Noumea, NEW CALEDONIA
[31]Institut Agronomique néo-Calédonien (IAC), Diversités Biologique et Fonctionnelle des Ecosystèmes Terrestres, BP 73, 98890 Païta, NEW CALEDONIA
[32]State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, The Chinese Academy of Sciences, Beijing 100093, CHINA
[33]Department of Plant and Environmental Sciences, University of Copenhagen, Copenhagen, DENMARK
[34]CNRS-Université Claude Bernard Lyon-Institut National des Sciences Appliquées-Bayer CropScience Joint Laboratory (UMR5240), Bayer CropScience, F–69263 Lyon cedex 9, FRANCE
[35]School of Biological Sciences and Chemistry, and Basic Science Research Institute, Sungshin Women's University, Seoul 142-732, REPUBLIC OF KOREA
[36]Key Laboratory of Conservation Biology for Endangered Wildlife of the Ministry of Education, and Laboratory of Systematic & Evolutionary Botany and Biodiversity, College of Life Sciences, Zhejiang University, Hangzhou 310058, CHINA
[37]INRA-AgroParisTech, Jean-Pierre Bourgin Institute (IJPB, UMR1318), Laboratory of Excellence "Saclay Plant Sciences" (LabEx SPS), F-78026 Versailles, FRANCE
[38]Department of Forestry, National Chung Hsing University, Taichung 40227, TAIWAN
[39]School of Forestry and Resource Conservation, National Taiwan University, Taipei 10617, TAIWAN
[40]School of Forest Resources and Conservation, University of Florida, Gainesville, FL 32611, USA
[41]Department of Plant and Microbial Biology, University of California, Berkeley, Berkeley, CA 94720, USA
[42]Institut de Recherche pour le Développement (IRD), Unité Mixte de Recherche–Diversité, Adaptation et Développement des Plantes (UMR DIADE), BP 64501, F-34394 Montpellier cedex 5, FRANCE
[43]Laboratorio Nacional de Genomica para la Biodiversidad, Irapuato, MEXICO
[44]Scientific Computing Group, Heidelberg Institute for Theoretical Studies, Heidelberg 69118, GERMANY

Contents: *Amborella* Genome Supplementary Online Materials

Section Page

Authorship Information 1

Supplementary Text

1: DNA Sequencing, *de novo* Assembly, and Validation 5

2: Genome Annotation 7

3: Genome Structure and Synteny 14

4: Gene Order in the Core Eudicot Ancestor 15

5: Global Gene Family Phylogenomics and Ancestral Gene Content 18

6: MADS-Box Genes 24

7: Glycogen Synthase Kinase 3 Genes 27

8: Seed Proteomics and 11S Globulin Proteins 29

9: Terpene Synthase Gene Family 33

10: Cell Wall and Lignin Biosynthesis Genes 35

11: Transposable Element Annotation 37

12: Small RNA Annotation and Analysis 42

13: Methods for Population Genomic Analyses 47

Supplementary Figures 51

Supplementary Tables 97

Additional Acknowledgements 124

Supplementary References 124

1: DNA Sequencing, *de novo* Assembly, and Validation

DNA sequencing – We sequenced *Amborella* using a next-generation sequencing based whole-genome shotgun approach. Single-end 454 FLX Titanium and 454 FLX+, along with 11 Kb mate pair 454 FLX Titanium, and 3 Kb mate pair Illumina HiSeq reads, totaling 48 Gb, were generated for assembly; all of these were combined with Sanger BAC end sequences from a 5.5X library with an average insert size of 123 Kb (*20*) (Table S1). Reads originating from contaminants (insect, bacterial, fungal, and human), as well as chloroplast (Genbank ID NC_005086) and mitochondrial genomes, were removed from all sequence data prior to assembly. Additionally, reads were screened to remove artificial duplicates originating from PCR errors (*90*, *91*) and chimeric reads (see *18* for detail s) generated during library construction or sequencing. We removed short and low quality reads, and trimmed the Illumina reads to include only regions with a PHRED score > 20 (Table S1). Screening reduced the total size of the data set by >51%, to 23 Gb (Table S1). Sequence reads were assembled with Newbler version 2.6.b20110517_1502 (*92*).

Assembly – Our finished assembly (version 1.0) resulted in 5,745 scaffolds with a mean scaffold length of 123 Kb, and an N50 size of 4.9 Mb (in 50 scaffolds), totaling 706 Mb of assembled sequence (Table S2). Our longest scaffold is 16 Mb and 159 scaffolds are longer than 1 Mb. We monitored contig growth as additional reads were included in the assembly to determine when we reached maximum coverage and contiguity, stopping when the average contig length curve and genome coverage curve became flat. This analysis identified points at which the inclusion of additional single end sequence data no longer had an effect on contig growth. A similar analysis was done to monitor the impact of paired reads on scaffold size and contiguity (see *18* for additional details).

Genome size estimate – Initial estimates from flow cytometry suggested that the genome of *Amborella* is approximately 870 Mb (*19*). Given this, our current assembly was predicted to represent 82% of the *Amborella* genome. However, three additional size estimates on the basis of genome sequence content suggest that the flow cytometry results overestimate genome size. Analysis of k-mer frequencies (*93*, *94*) suggests a genome size of 793 Mb, while estimates based on expanding putative repeat regions and sequence coverage across finished regions of the *Amborella* genome suggest sizes of 713 Mb and 736 Mb, respectively (*18*). These sequence-based methods for genome size suggest that the *Amborella* genome size is closer to 748 Mb, which represents an average of these three estimation methods. On the basis of the input data size and a 748 Mb genome, the high-quality sequence represents an average depth of coverage of approximately 25X, and the scaffolds cover >94% of the genome.

Physical map validation – We checked the incorporated BAC end sequences against a physical map previously generated from those BACs (*20*) to search for un-joined pairs of scaffolds and to assess accuracy of the assembly. Of the 3,106 FPC contigs that make up the physical map (*20*), 2,941 contained BACs with paired end sequences that were incorporated into 291 of our scaffolds; all 155 scaffolds comprising 90% of the assembly (N90) incorporate BAC end sequences. Based on the order of the end sequenced BACs within the FPC contigs and our scaffolds, 99% of the FPC contigs in the physical map are congruent with our assembly. Of the remaining 35 FPC contigs, 33 of them map to two scaffolds, one maps to three, and one contig (ctg4047) maps to over 50 scaffolds (*18*), corresponding to a highly repetitive sequence in the

genome. Whether or not those fingerprinted BAC contigs that map to two scaffolds reveal evidence of a physical join between these scaffold pairs remains to be examined.

We also aligned contig sequences to two high quality *Amborella* BAC sequences (*20*) using Mummer (*95*) to assess the completeness of our assembled contigs. Over 93% of the bases in the reference contigs were covered by our assembly, with nearly 100% identity.

To further improve the assembly, whole-genome (optical) mapping technology was used in collaboration with OpGen, Inc. (Gaithersburg, MD, USA). The OpGen Genome-Builder™ software was used to query assembled scaffold sequences against single molecule restriction maps to produce super-scaffolds. Briefly, the software uses single molecule restriction maps generated by the ARGUS instrument to iteratively extend *in silico* restriction maps predicted using the *de novo* assembly scaffolds. It then identifies overlapping regions between adjacent scaffolds based on pair-wise alignments between them and constructs super scaffolds using the highest confidence pairwise alignment path (details are provided in *18*).

We employed cytogenetic analyses using florescence *in situ* hybridization (FISH) to test the long-range integrity of the *Amborella* genome assembly. We selected 205 BACs assigned to 82 of the larger scaffolds to determine whether pairs of BACs from the individual scaffold generated FISH signals in close proximity on specific chromosomes. An example is shown in Fig. S1, where BAC FISH joined 110 contigs that are 1,298,028 bp apart on scaffold 49. The largest number of contigs joined by BAC FISH is 865; these are 6,172,508 bp apart on the 7th largest scaffold. Only one instance of BAC FISH did not agree with a scaffold; this was during an initial draft of the *Amborella* genome assembly and was corrected in a later version (*18*).

FISH probes were produced from BAC DNA extracted from overnight *Escherichia coli* culture, amplified by rolling circle amplification (RCA), and labeled by nick translation as described by Berr & Schubert (*96*). To test fluorescent markers for each chromosome pair, BAC probes generating minimal background signal through cross-hybridization were divided in pools of four to six BACs and sequentially hybridized to a single chromosome squash preparation. The identity and location of individual BACs was tracked across hybridizations, allowing deduction of their chromosomal assignment in the final karyotype, and physical placement of their corresponding scaffolds.

An *Amborella* karyotype has been achieved (Fig. S2), with 24 landmarks designating all 12 chromosome pairs (one pair is currently recognized by the absence of a BAC marker). Significantly, the hybridization patterns of 71 BACs are able to anchor 35 of the larger scaffolds on chromosomes (Fig. S2). This represents 190 Mb of the assembly—20% of the genome (*18*).

2: Genome Annotation

Protein-coding gene annotation overview – The automated annotation of protein coding genes used a combination of *ab-initio* predictions, alignments to *Amborella* expressed sequence data and alignments to protein coding genes from published plant genome annotations. These computations were managed using DAWGPAWS (*21*) and results were integrated to produce gene models with EVidenceModeler (*22*) using PASA (Program to Assemble Spliced Alignments) (*22*) terminal exon support and including masked transposable elements as input into gene prediction. Individual families of interest were selected for further manual curation by gene family experts.

***De novo* assembly of RNA-seq data –** The Trinity v. 2011-07-13 (*97*) *de novo* assembly pipeline was used to assemble 302,324,984 paired-end reads of Illumina GAIIx transcriptome data derived from *Amborella* apical meristems, premeiotic female buds or whole plant normalized RNA pools (Table S3). This assembly process assumed a paired fragment length of 300 bp and reported minimum contig lengths of more than 200 bp. The resulting assembly included 334,172 contigs with an average length of 1,052 bp and median length of 470 bp (Table S4).

Training *ab initio* gene model software – To identify gene models suitable for use in the training of *ab initio* gene finding programs, publicly available collections of 38,147 Sanger and 2,943,273 Roche 454 *Amborella* expressed sequence tags (EST) sequences (http://ancangio.uga.edu/content/amborella-trichopoda) were combined with 334,172 *de novo* RNA-seq assemblies generated by Trinity (*97*). These sequences were vector trimmed (*98*) and cleaned with seqclean (http://sourceforge.net/projects/seqclean/) before being aligned against the *Amborella* genome scaffolds using PASA (*22*). Assembled gene models with predicted transcripts that encode proteins less than 100 amino acids in length were removed, as were models that did not contain predicted coding sequences (CDS) start and stop signals. The resulting transcripts were verified as putatively full-length by alignments against rice (*99*, *100*) and *Arabidopsis* (*101*) protein databases. Only those *Amborella* protein sequences with a best hit alignment that was at least 50% identical, and shared a minimum of a 90% match length were retained as possible full-length *Amborella* proteins. This collection of proteins was clustered using custom PERL scripts that utilized an all-by-all BLAST approach to retain only a single representative sequence from each cluster of sequences that shared 90% identity over 90% of their lengths. These models were further improved where possible by manual curation using input from Cufflinks v 1.2.1 (*102*) alignments, and EVidence Modeler (*22*) derived gene models. To remove potential transposable element contamination in the training data, the transcripts were masked by searching against a database of repeats identified by a RepeatScout analysis of the *Amborella* scaffolds. Models were excluded from training if the transposable element (TE)-like sequences from RepeatScout matched more than 20% of the CDS length. This process resulted in a set of 1,723 high-quality full-length transcripts (Table S5) that were used for training of *ab initio* gene finding programs.

Training Augustus – The Augustus *ab initio* gene prediction tool (*103*) is based on Generalized Hidden Markov Models (HMM) that can be trained for individual species using alignment data as input. The *Amborella* HMM training for Augustus used Augustus v 2.5.5 to train with the 1,723 PASA gene models aligned against the *Amborella* genome scaffolds following standard

methods (http://augustus.gobics.de/binaries/retraining.html). The PASA alignments were converted to GenBank format with 1000 bp of flanking genomic DNA using the *gff2gbSmallDNA.pl* translation tool. The *Amborella* metaparameters and species-specific parameters were trained using the *optimize_augustus.pl* script run with default options. Augustus was then trained with these parameters using the *etraining* tool using the entire set of 1,723 PASA gene models as input. This training achieved a nucleotide level sensitivity and specificity of 0.961 and 0.925 respectively. The exon level sensitivity was 0.807 with an exon specificity of 0.766. Gene level sensitivity and specificity were 0.395 and 0.356.

Training SNAP – Amborella specific SNAP (*104*) parameterizations were trained with SNAP version 2006-07-28 using the 1,723 high-quality PASA gene models aligned against the *Amborella* genome scaffolds as training inputs. The PASA GFF3 files were converted to SNAP ZFF format using the *gff3_to_SNAP_train.pl* EvidenceModeler script (*22*) and gene models were validated using the validate option in the SNAP *fathom* program. The input gene alignments were then obtained from the genome with 1000 bp flanking sequence and converted to plus strand models using the SNAP *fathom* program. The *forge* program estimated *Amborella* parameters using these files as input, and a SNAP HMM model was constructed using the *hmm-assembler.pl* program from SNAP with default options.

Training TWINSCAN_EST – TWINSCAN_EST (*105, 106*) is based on TWINSCAN (*107-109*) and incorporates available EST alignments from the genome of interest in addition to integrating a traditional HMM based gene-prediction probability model with alignment information from closely related species. TWINSCAN_EST (Version 4.1.2) training was performed as detailed in the TWINSCAN documentation (http://mblab.wustl.edu/software/twinscan/) and prediction accuracy was assessed by fourfold cross-validation (*110*) using custom PERL scripts to automate the process. Genomic segments containing each of the 1,723 predicted models with up to 2 kb of flanking sequence were first identified from the *Amborella* draft genome assembly. Each of these genomic sequences were WU-BLAST aligned to the TAIR9 *Arabidopsis* genomic sequence (http://www.arabidopsis.org) and the alignments converted to conseq format (*107, 108*). The *Amborella* genomic segments were BLAT aligned (*111*) to a comprehensive set of *Amborella* expressed sequences that included publicly available collections of 38,147 Sanger and 2,943,273 Roche 454 *Amborella* EST sequences as well as 334,172 *de novo* RNA-seq assemblies generated by Trinity (*97*), and the resulting alignments were converted to estseq format (*106*). Training and cross-validation were performed by splitting the 1,723 gene models into four approximately equal groups consisting of the *Amborella* genomic segment, a GTF representation of the PASA aligned gene structure and the corresponding conseq and estseq format alignments. After four-fold cross validation of the TWINSCAN predictions (Table S6), the final parameter set was generated by training with the complete collection of 1,723 models.

***Ab initio* gene predictions –** Before annotation with *ab initio* gene finding programs, the *Amborella* genomic scaffolds (AmTr_v1.0) were masked by RepeatMasker v. 3.3.0 (http://www.repeatmasker.org) with default options using the WU-Blast engine (http://blast.wustl.edu) and a custom database of transposable elements identified in the *Amborella* genome. The SNAP search against the masked scaffolds using SNAP v 2006-07-28 (*104*) with the *Amborella* specific parameterization described above predicted 194,023 exons in 49,598 gene models. An Augustus v 2.5.5 (*103*) analysis identified 134,383 exons in 31,098 gene models using only the *Amborella* specific parameterizations as input.

TWINSCAN_HMM v. 4.1.2 (*106*) with the *Amborella* specific parameterization described above was used to predict genes in the masked *Amborella* genome scaffolds (AmTr_v1.0). The informant database was the TAIR9 *Arabidopsis* chromosome sequence (http://www.arabidopsis.org). The EST database for estseq generation was a collection of 2,943,273 Roche 454 and 38,147 Sanger *Amborella* EST sequences combined with 334,172 *de novo* RNA-seq assemblies from Trinity (*97*). The TWINSCAN gene predictions included 142,608 exons in 50,435 gene models.

The *ab initio* gene predictions derived from SNAP, Augustus and TWINSCAN_EST can be downloaded from the *Amborella* Genome Database (http://www.amborella.org/) and are viewable through the GBROWSE genome browser that is also accessible through the project website.

Alignments to *Amborella* expressed sequences – Transcribed regions of the *Amborella* genome were identified using both reference genome-guided and reference genome-independent transcript reconstruction methods (*112*).

PASA alignments - A dataset of 334,172 *de novo* RNA-seq assemblies generated by Trinity (*97*) was cleaned with seqclean (http://sourceforge.net/projects/seqclean/) using the UniVec vector database (http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html), yielding 334,129 cleaned sequences that included 66,292 trimmed sequences. These cleaned assemblies were aligned against the masked *Amborella* genome assembly using PASA version 2011_05_20 (*22*) with default options. The resulting PASA alignments included 213,182 matches of assembled RNA-seq sequences against the *Amborella* genome. Terminal exons were identified from these results using the *retrieve_terminal_CDS_exons.pl* program from EVidenceModeler (*22*).

Cufflinks alignments – The Cufflinks (*102*) RNA-seq alignment analysis started with 328,618,794 paired-end Illumina GAIIx transcriptome reads derived from *Amborella* tissues that were either leaf shoots, premeiotic female buds, apical meristems with young leaves, or whole plant normalized libraries (Table S3). Of these 327,951,474 reads were used after quality filtering. These sequences were aligned with TopHat v 1.3.1 (*113*) against the *Amborella* genome scaffolds (AmTr_v1.0) producing 231,135,710 reads with at least a single reported alignment across a total of 248,811,352 reported alignments. These alignments identified 220,107 high quality splice junctions.

The RNA-seq data were assembled into transcript models using a Cufflinks v 1.2.1 (*102*) analysis of the TopHat accepted_hits.bam file mapped to the *Amborella* genome scaffolds (AmTr_v1.0). The resulting GTF format file was converted to a GFF3 format using the Cufflinks gtf2gff3 program with default options. The Cufflinks models included 36,294 gene predictions with 201,624 exons in 47,445 transcripts. The Cufflinks GFF3 result was converted to an EVidenceModeler (*22*) compliant alignment format using the *cnv_cufflinks2evm.pl* program of DAWGPAWS (*21*). Of the 27,313 unfiltered EVM 27 gene models, 20,627 (~75%) gene models with cufflinks support.

Exonerate alignments – The dataset of 334,172 *de novo* RNA-seq assemblies generated by Trinity (*97*) was masked by RepeatMasker v. 3.3.0 (http://www.repeatmasker.org/) with default options using the WU-Blast engine (http://blast.wustl.edu) and a custom database of transposable

elements identified in the *Amborella* genome. The resulting soft-masked fasta files were aligned to the masked *Amborella* genome assembly using the est2genome function of Exonerate v. 2.2.0 (*114*). This exonerate search returned only the best match, and required a minimum match of 96% of the maximum possible score for each query, a minimum gene-seed score of 250, and a minimum alignment score of 500. The search returned 765,197 exon matches from 227,341 individual Trinity models.

Alignments to published gene models – Alignments to published gene models used Exonerate v. 2.2.0 (*114*) to search masked *Amborella* genomic DNA against translated CDS protein models from *Arabidopsis thaliana* (*101*), *Oryza sativa ssp. japonica* (*99*, *100*), and *Vitis vinifera* (*27*) and returned the single best alignment for each gene model using default exonerate parameters. The *Arabidopsis thaliana* search included 35,386 translated CDS protein models (*101*) downloaded from Phytozome (http://www.phytozome.net/) and resulted in 135,565 predicted exon matches among 30,400 *Arabidopsis* genes. The rice alignment search included 42,124 translated CDS protein models from *Oryza sativa ssp japonica* (*99*, *100*) downloaded from the International Rice Genome Sequencing Project (IRGSP, http://rapdb.dna.affrc.go.jp/download/irgsp1.html). These alignments resulted in 31,497 gene matches comprised of 113,271 exon models. The *Vitis vinifera* search included 26,346 translated CDS protein models (*27*) downloaded from Phytozome (http://www.phytozome.net/) that had 22,234 gene matches comprising 101,356 exons.

Automated computational combined evidence gene annotation – The automated combination of gene evidences to produce gene models used the EVidenceModeler (EVM) (*22*) program version 2012-06-25. *Ab initio* gene model outputs from Augustus (*103*), SNAP (*104*) and TWINSCAN_EST (*106*) were translated to EVM compliant format using DAWGPAWS tools, and were checked for compatibility with the *gff3_gene_prediction_file_validator.pl* program. Exonerate protein alignments against *Arabidopsis thaliana* (*101*), *Oryza sativa ssp. japonica* (*99*, *100*), and *Vitis vinifera* (*27*) were also included as protein evidence files. Expressed sequence evidence used the *Amborella* expressed sequence alignments from Cufflinks (*102*), Exonerate (*114*) and PASA (*22*) as described above. The PASA generated terminal exons were also used to improve the EVM predictions. The RepeatMasker TE annotation results using the *Amborella* structurally identified TE database was also used as input to exclude transposable element models from gene predictions.

The resulting EVM predictions produced 27,313 gene models with 110,895 exons. These models were BLAST screened against the database of *Amborella* transposable elements and potential TEs were removed to yield a TE filtered database of 26,846 genes with 109,783 exons. All of the EVM input files and resulting gene model predictions (EVM release version 27) are available from the *Amborella* Genome Database (http://amborella.org/).

Manual gene model curation – To evaluate and improve the quality of the genome annotation, both random gene model sampling and gene family focused surveys were conducted.

Random gene model surveys – In order to assess the quality of the EVM 27 annotation, over 800 randomly chosen RNA-seq based Cufflinks models lacking corresponding EVM models in an early release annotation (EVM 15) were visually inspected within the *Amborella* genome browser (www.amborella.org). Cufflinks models were compared with all other lines of evidence

used to infer the EVM 27 gene models including *ab initio* and homology-based evidence. As expected, the Cufflinks models included upstream- and downstream-untranslated regions (UTRs) that were not included in the EVM 27 models. In about half of the instances where RNA-seq based Cufflinks models were not associated with an EVM 27 gene prediction, the Cufflinks model was found to correspond to an apparently expressed transposable element. In some instances (approximately 10%), EVM 27 gene predictions were not made for *bona fide* protein-coding genes corresponding to Cufflinks models. Most of these genes included high copy domains. Highly abundant gene families with conserved domains such as the nucleotide-binding site (NBS)-leucine-rich repeat (LRR) super family may have been over-filtered in our efforts to remove transposable element repeats from the gene annotations or were under-represented in alignment evidences due to filtering of the exonerate alignments for the best match. This represents a tradeoff between sensitivity and specificity in automated genome annotation that can be supplemented by manual annotation curation of high-copy-number gene families.

Manual annotation of NBS-encoding genes – To compensate for overmasking, we invested additional effort in annotating the NBS-encoding genes that may have been missed in the automated genome annotation. This included directed BLAST and hidden Markov models (HMM) searches, as described in Zhou *et al.* (*115*). First, possible homologs encoded in plant genomes were searched using BLASTp with the amino acid sequence of the NB-ARC domain (Pfam: PF00931) as a query. The threshold expectation value was set to 1.0, a value determined empirically to filter out most of the spurious hits. Second, the nucleotide sequences of candidate NBS-encoding genes were used as queries to find homologs in the genome by a BLASTn search. This step was crucial to finding the maximum number of candidate genes. All new BLAST hits in the genome, together with flanking regions of 10,000 bp on both sides, were annotated using the gene-finding programs FGENESH (http://www.softberry. com/) and GENSCAN with the *Arabidopsis thaliana* training set (http://genes.mit.edu/GENSCAN.html) to obtain information on complete open reading frames (ORFs). To exclude potentially redundant candidate NBS-encoding genes, all sequences were orientated by BLASTn, and sequences identified in the same location were eliminated. All non-redundant candidate NBS genes were surveyed to further verify whether they encoded NBS or LRR motifs using the Pfam database v23.0 (E value cut-off of 10^{-4}) (http://pfam.janelia.org/), SMART protein motif analyses (http://smart.embl-heidelberg. de/), and Multiple Expectation Maximization for Motif Elicitation (MEME) (*116*). CC motifs were detected using COILS with a threshold of 0.9 (*117*). A total of 126 NBS genes are present in *Amborella*, of which 44 are full-length genes with intact NBS domains. These gene models have been incorporated into the stable gene annotations.

Alternative splicing – 6,172 of the *Amborella* gene models were predicted to undergo alternative splicing during transcript processing, and many splice junctions were determined to be tissue specific (Fig. S3). Of the 20,235 transcript isoforms predicted assuming 2X coverage across junctions and/or retained introns, 5,356 (37.4%) contain an intron retention event, while isoforms containing an alternative acceptor, alternative donor, alternative exon, or skipped-exon represented 2,134, 5,296, 3,975, 3,564 and 2,134 respectively. This same trend is also observed when considering the relative abundance of unique splicing event types when taken out of their isoform context (Fig. S4).

RNA-seq data cleaning – Five lanes of Illumina RNA-seq data (including whole plant normalization, female bud, apical meristem and leaves) were cleaned using the following

procedure. The Flexible Adapter Remover (FAR) tool (http://sourceforge.net/apps/mediawiki/theflexibleadap) was used to identify the adapter sequence for each run, then we used cutadapt (http://code.google.com/p/cutadapt/) to remove those adapter sequences with these parameters: --error-rate=0.1 --times=1 --overlap=5 --minimum-length=0 --quality-base=64. After removing the adapter sequence, we used Trimmomatic (*118*) to qualify filter the reads -phred64 -threads 3 HEADCROP:0 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:40. Duplicate RNA-seq reads that were considered artifacts were identified and removed with the Picard (http://picard.sourceforge.net) MarkDuplicates software package.

RNA-seq alignment – RNA-seq reads were aligned to the unmasked *Amborella* v1.0 Scaffolds with GSNAP (*119*) using the following parameters: -D gmapdb -d Amborella_v1.0_unmasked --nthreads=16 -B 5 --suboptimal-levels=0 -m 0.05 -n 5 -Q -N 1 -w 20000 --nofails --format=sam --sam-multiple-primaries --pairmax-rna=20000 --split-output. Only the concordant unique reads, concordant multiple mapped reads mapping to five or less locations, half mapping unique mapped reads and unpaired unique mapping reads were used.

Transcriptome assemblies – Artificial duplicate reads were identified and removed from 454 RNA-Seq reads with the cd-hit-454 (*91*) software utility, using the following parameters: -c 0.96 -aS 0 -M 0 -g 1 -T 2. After removing duplicates, 454 and Sanger ESTs were combined and assembled with Newbler version 2.7 with –cdna and –urt options based on previous comparisons (*120*). Illumina RNA-Seq reads were assembled with a beta version of the Trinity software package that used a genome-guided approach (http://trinityrnaseq.sourceforge.net). Default settings were used with Trinity, and the maximum allowable intron size was set to equal the largest intron identified in the EVM 27 *Amborella* annotation.

Identification and classification of alternative splicing – Hybrid 454-Sanger Newbler assembled sequences and Trinity RNA-Seq assemblies longer than 300 bp were aligned to the *Amborella* v1.0 genomic sequence assemblies using the PASA (*79*) annotation pipeline. All PASA identified isoforms, and the associated GTF were provided as reference inputs to the Cufflinks (*121*) packages, which were used to assess isoform expression levels based on RNA-Seq GSNAP alignment describe previously. All PASA assembled isoforms with an FPKM < 1 were removed from further consideration. Additionally, all retained isoforms required each splice junction to be confirmed by a minimum of two independent, uniquely aligning, reads from RNA-Seq, or 454, or Sanger EST sequences. In the case of a retained intron, we further required that every base of the retained intron was covered by at least two independent reads. Isoforms that were retained were re-aligned to the *Amborella* v1.0 genomic sequence assemblies with PASA, which classifies splice isoforms based on their splice events (exon skip, intron retention etc.). As stated above, every junction in an isoform is expected to be confirmed by a minimum of two reads. While increasing junction support requirements results in a decrease in isoform detection (Table S7), a reduction in isoform count of less than 10% results from a 5-fold increase in junction read support. This suggests that the requirements for detection are robust, and the majority of isoforms lost as a result of increased depth requirements likely reflect the loss of low abundance isoforms rather than a reduction in noise. Additionally, the type of isoform lost as a result of increasing sequence depth requirements is not restricted to one class (Fig. S4).

Bisulfite sequencing and gene body methylation analysis – Bisulfite sequencing library construction was carried out as described by Lister *et al.* (*122*). Briefly, 5 µg of *Amborella* DNA was extracted from young, expanding leaf tissue using the DNeasy Plant Mini kit (Qiagen) and fragmented to ~350 bp by sonication. After end repair, DNA fragments were ligated to adaptors, treated with the Chemicon CpGenome DNA Modification kit (Millipore), and amplified using Illumina PCR primers. Approximately 115 million 50 bp single-ended reads were generated from the library. No significant contamination of Illumina adapter sequences was detected with FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). Reads were quality filtered using FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) in a two-step process. First, 3' ends with a Phred score <20 were clipped and the read was removed if greater than 10nt were clipped, then at least 80% of the remaining read must be >20 Phred score or it was removed.

In total, 63 million cleaned reads mapped to single loci across 668 Mb of contig sequences for the assembled reference genome (4.72X coverage) using *Bismark* version 0.7.7 (*123*) with default settings except for using Bowtie2 and allowing 1 mismatch per default 20 nt seed alignment. The methylation signal of every mapped reference cytosine in CpG, CHG, and CHH contexts was calculated in *Bismark* using *bismark_methylation_extractor*. Methylation signal is calculated as the ratio of methylated cytosines to the total number of methylated and unmethylated cytosines at a reference cytosine position. Custom Perl scripts then report the relative position, methylation signal and context of every cytosine in a range around the start and stop positions of a given genomic element on both strands, requiring at least two aligned bisulfite reads for a cytosine position to be considered. Meta-plots were constructed to display average methylation levels inside and outside of annotated protein-coding gene models (Fig. S5) and LTR-retrotransposons (see supplementary text section 11).

Structural RNA annotation –

Transfer RNA annotation – Transfer RNAs were predicted in the *Amborella* genome assembly (AmTr_v1.0) with Aragorn v.1.2.34 (*124*) and tRNAScan-SE v 1.23 (*125*) using default parameters for scanning eukaryotic genomes. Aragorn identified 617 putative eukaryotic tRNAs, include 590 decoding the standard amino acids, and three potential selenocysteine tRNAs. The tRNAScan-SE analysis identified 471 tRNAs decoding the standard amino acids as well as two selenocysteine tRNAs, two suppressor tRNAs, 85 pseudogenes, and seven tRNAs with undetermined isotypes. The predicted tRNA models from the two programs were manually reconciled by combining spatially overlapping tRNA predictions and taking the longest prediction as correct when there was disagreement in start and end locations between the two programs. The combined predictions included 651 eukaryotic tRNAs and four selenocysteine predictions (Table S8). These combined models included 22 tRNAs with introns.

Codon usage analysis of the EVM 27 *Amborella* gene predictions used the GCUA v 1.2 (*126*) program to analyze the CDS sequences of the 26,846 gene predictions that were filtered to remove transposable elements. Resulting codon usage counts along with combined tRNA prediction counts are summarized in Table S9.

Ribosomal RNA annotation – Ribosomal RNAs were identified using RNAmmer v1.2 (*127*) with default options for eukaryotic ribosomal RNA identification.

3: Genome Structure and Synteny

Structural syntenic analyses were performed with the SynMap tool of the CoGe comparative genomics platform (*128*). The genome of *Vitis vinifera* (grape) was compared to the genome of *Amborella* with the chaining algorithm DAGChainer (*129*). We specified a maximum of 40 genes between identified gene pairs and required a minimum of 5 genes to seed a syntenic region (Fig. S6). Contigs of *Amborella* were ordered and oriented based on their syntenic path to grape (*130*); the order of *Vitis* chromosomes was customized to maximize the visualization of the 3:1 relationship of syntenic regions to *Amborella* using the syntenic path algorithm implemented in CoGe (*130*). The genome of *Amborella* is almost entirely covered by three syntenic *Vitis* regions; these have been highlighted with colored blocks in Fig. S7. The analysis can be regenerated in SynMap, albeit without the custom ordering of grape chromosomes (http://genomevolution.org/r/5if3). For a detailed description of how to compare genomes using SynMap to identify and screen for polyploidy events, see (*131*).

High-resolution analysis of microsyntenic regions was performed with CoGe's GEvo application (*29*), which permits dynamic comparison of two or more genomic regions. Comparison of a single *Amborella* region to three syntenic regions of *Vitis* shows that each grape region is syntenic through the pattern of collinear homologous gene arrangement (Fig. S8). In addition, this analysis shows fractionation (gene loss of homeologous genes) among the grape regions, as expected following polyploidy (*132*).

We experimented with different parameters, to find parameters that worked to identify duplicates with relatively little noise. This is an informal process with (as yet) no clear theoretical basis for predicting the optimum settings for maximizing signal over noise. However, although more or less stringent parameter choices do yield different numbers and sizes of syntenic blocks, the observed patterns and conclusions are fundamentally the same.

The self-self comparison of the *Amborella* genome also employed SynMap (http://genomevolution.org/r/4s30) using DAGChainer (Max gene distance 80; minimum of four genes to seed a syntenic regions; Fig. S9A). Each syntenic gene pair is colored based on its synonymous mutation value (K_s) as calculated by CODEML (*105*). The histogram of K_s values shows a peak around a log10 transformed value of 0.4 and is indicated in green on Fig. S9B. These values correspond to a population of syntenic gene pairs that are inferred to have approximately the same age and thus appear to have been contemporaneously duplicated. Given that the genes cover a significant portion of the genome, this is evidence that they were derived from a whole genome duplication event (Table S10). Since this is an old event and many duplicated genes were likely lost due to fractionation, relatively loose parameters were used to identify syntenic regions. Such loose parameters can increase noise in the analysis due to incorrectly identifying gene pairs as being syntenic. The K_s values for these gene pairs are much larger and manifest as a peak on the far right of the K_s value histogram. Microsynteny analysis of the six largest inferred duplicated regions of *Amborella* shows the expected pattern of collinear homeologous genes, albeit much weaker than for the *Amborella-Vitis* comparisons (Fig. 2 of the main text; Fig. S10). Phylogenetic analysis of curated syntenic blocks also provides evidence for an angiosperm-wide genome duplication (Table S11, see also supplementary text section 5, below).

4: Gene Order in the Core Eudicot Ancestor

The CoGe SynMap tool (*29*) was used to identify gene homology relationships between sequenced rosid genomes and *Amborella* by comparing the *Amborella* genome (id19515) to grapevine (*Vitis vinifera*, id9050), cacao (*Theobroma cacao*, id10997), and peach (*Prunus persica*, id8400) (*27*, *28*, *30*). Because the gamma hexaploidization event preceded the radiation of the rosids, it is expected that each subgenome in a rosid should resemble its homologs in other rosids more than it resembles the other two other homeologous subgenomes. Further, because Amborella lacks the gamma hexaploidization, we can use gene order in Amborella as an outgroup in the ancestral gene order reconstruction. We therefore expect that SynMap would correctly identify the individual orthologous gene from the set of two or three paralogs when aligning orthologs between genomes. The SynMap algorithm was applied with a chaining distance of 50 genes and a minimum of five aligned gene pairs to identify likely orthologous genes in all six pairwise-comparisons of the four genomes, as well as to determine the paralogs identified by self-comparisons of each of the three rosids. We then processed the combined set of orthologs and paralogs with the 'OMG!' program (*31*) to produce 12,833 homology sets including at most three paralogous versions of each gene in each rosid, and at most one copy in *Amborella*. We also required at least one gene from two different genomes in each set. These sets, each representing one "candidate gene" in the reconstructed ancestral genome, contained a total of 47,526 extant genes from the four input genomes with an average of almost four homologs per set. Eighty-nine homology sets contained the maximum of ten genes, one from *Amborella* and three paralogs from each rosid, while 2,772 sets only contained one pair of genes. The above set of genes with identified homology relations were used to reconstruct the ancestral gene order using an efficient algorithm for Maximum Weight Matching (MWM) (*32*) first at the level of gene adjacencies followed by a reconstruction at the level of contig adjacencies.

First we identified all the gene adjacencies in *Amborella* and in the three rosids considering only the genes within the set of 47,526 genes with identified gene homology relationships. Each adjacency was then weighted according to how many times homologs of the two genes involved were adjacent with somewhat more weight on *Amborella* adjacencies (1.6 vs. 1.0 for adjacencies in the rosids). Note that the 5'-3' orientations of the adjacencies were specified, so that the MWM, where the weights were defined by the number of homologous copies of each adjacency in the data, produced an optimal set of 1,007 contigs containing all 12,833 putative ancestral genes, or 11,826 distinct adjacencies out of a total of 25,885 adjacencies in the data set. The remaining adjacencies were not consistent with the optimal set.

For each of the three rosids, which are all structurally conservative, clear patterns of paralogy enable us to assign most genes to one of seven groups of three homeologous chromosomes or segments. Paralogies in these rosid genomes are almost entirely between segments within homeologous groups but not within individual segments or chromosomes themselves (*27*, *28*, *30*, *133-136*). We assigned all the rosid genes in all the homology sets to subgenomes by first establishing the chromosomal coordinates of the endpoints of each region making up a subgenome. For cacao, this information was directly provided in the original genome publication (*28*). For grape, where these regions are largely co-extensive with single chromosomes, we relied on the dotplot of a SynMap self-comparison to pick out the coordinates. For peach, we combined dotplot examination of self-syntenies with coordinate information on genes with orthologs in grape or cacao to estimate endpoints of regions, some of which are made up of several non-

contiguous fragments. Once the endpoints were available, we assigned each gene lying between two corresponding endpoints to the appropriate subgenome, whether or not it had a homolog in that subgenome in the other two rosids. Fig. S11 illustrates the decomposition of the three genomes into subgenomes.

We used the contigs in the output of the MWM to reconstruct each of the seven ancestral chromosomes one at a time. Only 49 contigs contained large proportions of genes originating in two or more of the ancient chromosomes and were discarded so that the remaining contigs could all be associated unambiguously with one of the seven pre-hexaploid chromosomes. In addition, to ensure a level of syntenic robustness, if a contig did not have more than three genes in at least one subgenome in some rosid, we discarded it. Losing these 360 contigs plus the 49 multicolored ones, constituting 40% of the total number of contigs, only deprived the analysis of 11% of the genes involved in the MWM analysis. As seen in Fig. S12, these were largely in homology sets of size 2.

For each of the 598 remaining contigs, we calculated an average chromosomal position on each subgenome for which it contained more than three genes, and/or on a single *Amborella* scaffold, and we associated a polarity, namely that of the majority of these genes, to the position. Then on each of the 9 = 3 x 3 subgenomes and on *Amborella*, we could identify adjacencies among the contigs, up to 20 adjacencies per contig, taking into account positions and polarities.

A second iteration of MWM was then applied not to gene adjacencies, but to all the contig adjacencies pertinent to one of the seven ancestral chromosomes to be reconstructed. We defined three “subgenome sets” each containing the three descendant chromosomes (peach, grape and cacao) from the single chromosome in the ancestral hexaploid. The MWM was set up between four sets, the three subgenome sets corresponding to the three ancestral homeologs of the chromosome in the hexaploid, and the *Amborella* scaffolds involved in all of the contigs within these subgenome sets. An adjacency with two or three copies within a subgenome set was weighted somewhat more (peach and cacao: weight 1.1, grape and one of peach or cacao: 1.25, all three rosids: 1.5) than a single-copy adjacency, but not as much as the weights that an adjacency in two (weight 2) or three (weight 3) different subgenome sets would automatically receive during the MWM. This weighting is consistent with the hexaploidization event preceding the radiation of the rosids, so that an adjacency in two different homeologs, even in a single species, is more revealing of original configurations than an adjacency in two homeologous chromosomes in different species.

The second MWM gives the optimal ordering of the contigs. Since we already have the gene content, gene order and gene orientation within each contig, this determines the positioning and polarity of all the genes on the ancestral chromosome. The reconstruction of the seven chromosomes results in a total of 61 scaffolds, containing 10,456 genes. Fig. S13 shows that the modal size of the scaffolds ranges from 129 to 256 genes. Indeed, 90% of the genes are in scaffolds containing more than 100 genes each.

Orthologous gene alignments between one of the ancestral chromosomes and the four input genomes is shown in Fig. 2C (main text). Most striking is the high degree of interleaving of the orthologies from the three rosids covering each of the three subgenomes. Reconstructed scaffolds within a chromosome are ordered to minimize the number of crossing lines.

An overall assessment of the reconstruction is provided by a measure of genomic change between the ancestor and the four genomes. Traditional measures of genome rearrangement cannot be applied in the context of polyploid genomes because fractionation greatly and erroneously elevates this distance in a complex manner (*135*), and because there is no consensus on how to treat duplicate genes in ancient polyploid comparisons. Instead we simply count the total number of different gene adjacencies and normalize by the number of adjacencies in one of the genomes. This number can vary between 1 for identical genomes to 2 for completely scrambled diploid genomes and higher for polyploids. Fig. S14 shows how these normalized measures are affected as we increase the weight on *Amborella*. Clearly, for large values of the weight parameter, the reconstruction simply tends to resemble *Amborella* more and more, and deviates increasingly from the rosids. But for weights up to 1.5 or 1.6, the measures are almost stable for the rosids, while the *Amborella* values improve. What is happening is that there is a degree of uncertainty in the reconstruction so that different reconstructions will have approximately the same cost. Increasing the weight on *Amborella* first resolves some of these ambiguities, with little extra cost to the rosid comparisons, but bringing the reconstruction more in line with *Amborella* wherever possible.

The difference in gene complements among the core eudicots is partially accounted for by the different patterns of fractionation from the original hexaploid ancestor. For each of the seven reconstructed pre-hexaploid ancestral chromosomes, the regions in each of peach, cacao and grape descended from that chromosome in one of the three original subgenomes has more genes than the other two, the same subgenome in all three rosids (Fig. S15). There is only one exception out of all 21 comparisons.

For these genomes, we do not have the kind of chromosomal data examined by Schnable *et al.* (*137*) for post-tetraploid maize and the outgroup (sorghum), so we cannot be sure that the largest descendant of all the seven chromosomes originate in the same subgenome, but this is likely, especially in the light of the fractionation results from maize.

The pattern in Fig. S15 is also consistent with the idea that the hexaploidy resulted from an initial tetraploidization, followed by an initial fractionation of the two subgenomes involved (red and green in the figure), before the incorporation of a third genome (in blue). The latter would then have had less time to lose genes, depending on the interval between the two polyploidization events.

It is important to note that the clear pattern of dominance by a single daughter region in Fig. S15 only emerges when we examine the genes syntenically orthologous with genes in the reconstructed ancestor, a model dependent on knowledge of *Amborella* gene order. If we examine all the genes in the triplicated regions in the extant genomes, where the large majority cannot be linked syntenically to the reconstructed ancestor, i.e., where their current location results from gene family expansion, rearrangements such as transposition, or other process, the pattern is violated for four of the seven chromosomes.

5: Global Gene Family Phylogenomics and Ancestral Gene Content

Constructing a global gene family classification – The OrthoMCL method (*81*) was used to classify the complete set of protein coding genes from 22 plant genomes into narrowly-defined gene lineages (i.e. orthogroups). We selected taxa to represent all of the major land plant lineages for which genome sequence data are available, including nine rosid genomes (*Arabidopsis thaliana, Carica papaya, Fragaria vesca, Glycine max, Medicago truncatula, Populus trichocarpa, Thellungiella parvula, Theobroma cacao, Vitis vinifera*), three asterids (*Solanum lycopersicum, Solanum tuberosum, Mimulus guttatus*), two basal eudicots (*Nelumbo nucifera, Aquilegia coerulea*), five monocots (*Oryza sativa, Brachypodium distachyon, Sorghum bicolor, Phoenix dactylifera, Musa acuminata*), one basal angiosperm (*Amborella trichopoda*), one lycophyte (*Selaginella moellendorffii*), and one moss (*Physcomitrella patens*). A total of 53,136 orthogroups containing at least two genes were identified, 12,540 of which contain at least one gene from *Amborella* (Table S12). Of the 26,846 annotated genes in *Amborella*, 18,448 (69%) were classified in an orthogroup and the remaining 8,398 (31%) genes are treated as singletons, a classification rate that is also typical for the other taxa (Table S12). Subsequent analyses based on this classification exclude *Mimulus guttatus, Nelumbo nucifera*, and *Aquilegia coerulea* to respect the pre-publication data usage agreements for these genome projects. Complete details for each orthogroup, including gene counts and functional annotations are reported in Table S13. The complete orthogroup classification and sequence data are available at the *Amborella* Genome Database (www.amborella.org). We further clustered orthogroups into a larger hierarchical gene family classification (i.e., super-orthogroups) by using the minimum E-value between all pairs of orthogroups as the input similarity matrix for a second round of MCL clustering using an inflation value of 3.5 that balances sensitivity and specificity in delineating clusters across a wide range of gene family sizes. We found that this inflation value often yielded a stable classification consistent with many traditionally defined gene families characterized by functional domains. A total of 6,054 super-orthogroups were defined, of which 4,580 contain genes from *Amborella* (Table S13).

Sorting basal angiosperm and gymnosperm genes into orthogroups – To determine whether orthogroups appearing to be angiosperm-specific may have members from other seed plant lineages, and to expand phylogenomic analyses to include taxa that do not have genome sequences, we then sorted gene models from the Norway spruce genome and a large collection of EST assemblies (*7*) from four basal angiosperms (*Aristolochia*, *Liriodendron*, *Nuphar*, and *Persea*) and 16 gymnosperms (including *Pinus*, *Picea*, *Zamia*, and *Cryptomeria*) into orthogroups. Assembled transcripts were translated and candidate orthogroups were identified using BLASTP by retaining all hits in the orthogroup classification with E-values <= 1e-5. Profile hidden Markov models were constructed for each orthogroup using the hmmbuild program of HMMER (*138*, *139*) and transcripts were assigned to the orthogroup yielding the best HMM bitscore after screening with hmmsearch. The number of gymnosperm and basal angiosperm transcripts assigned to each orthogroup is reported in Table S13.

To evaluate the adequacy of the EST datasets for detecting the presence of an orthogroup in either basal angiosperms or gymnosperms, we performed an analysis of orthogroup saturation by randomly sampling successively larger fractions of the collected EST assembly for gymnosperms and basal angiosperms recording the total number of orthogroups detected. Transcript sequencing coverage for basal angiosperms was much higher (274,678 unigenes) than

for gymnosperms (146,704 unigenes) but for both datasets the rate of new orthogroup detection approached saturation (Fig. S16). In gymnosperms, 12,524 orthogroups were observed and 16,205 orthogroups were observed in basal angiosperms. For comparison, just over 9,500 orthogroups are present in *Selaginella* or *Physcomitrella* and 12,540 orthogroups are present in *Amborella*.

Ancestral gene content reconstruction – We reconstructed the ancestral gene content at key nodes in land plant phylogeny represented by 19 sequenced plant genomes (Fig. 3 of the main text) and modeled the changes occurring along each branch using both a Wagner parsimony framework with a weighted 1.2 gain penalty (Table S14) and a birth-death likelihood model (Table S15), as implemented in the program Count (*140*). The pattern of change in ancestral orthogroup content reconstructed between the two methods is similar; so most results are discussed in the context of the parsimony analysis. The lack of fern and gymnosperm genomes limits the resolution of this reconstruction at the origin of angiosperms relative to their presence in the MRCA of seed plants.

To gain further insight into the origin of novel angiosperm traits, we evaluated the presence/absence of transcripts from gymnosperms and other basal angiosperms to resolve the presence of orthogroups in the MRCA of seed plants and the gain of new orthogroups in the MRCA of angiosperms using a simplified parsimony-based scoring algorithm. We first identified orthogroups that were absent in the ancestral tracheophyte, but present in the MRCA of angiosperms based on the Wagner parsimony reconstruction using genomic data. If gymnosperm transcripts were detected in any of these orthogroups, the origin of these orthogroups was inferred to have occurred prior to the origin of seed plants. Otherwise, the orthogroup was inferred to have been gained in the ancestral angiosperm. For those orthogroups not reconstructed in the ancestral angiosperm based on genome data alone, but are present in *Amborella* and basal angiosperms, we revised the origin of the orthogroups to have been gained in the MRCA of angiosperms (and is inferred to be lost in some angiosperm lineages). For those orthogroups not reconstructed in the MRCA of angiosperms but were present in gymnosperms, if they were also detected in *Selaginella*, the orthogroup was inferred to be ancestral in tracheophytes. For those orthogroups not inferred to be ancestral in angiosperms, but present in gymnosperms and absent in *Selaginella*, but present in *Amborella*, or other basal angiosperms, or the MRCA of monocots plus eudicots, the orthogroup was inferred to have been present in the MRCA of seed plants, and may have been lost in some angiosperm lineages. Orthogroups not meeting any of these criteria were identified as lineage specific or derived after the origin of angiosperms. In a few ambiguous cases, orthogroup origins were left unscored if there were equally parsimonious reconstructions (gains/losses) surrounding the MRCA of angiosperms or the MRCA of seed plants. The results and details of this reclassification of orthogroup origins are reported in Table S13. This revised ancestral orthogroup reconstruction was then transferred to the super-orthogroup level if other members of the super-orthogroup were not detected in earlier lineages.

With the inclusion of transcript sequences in gymnosperms and basal angiosperms, we were able to assign 5,704 orthogroups as having been gained in the MRCA of gymnosperms or the MRCA of angiosperms, compared to the 3,285 orthogroups inferred as gained along the branch between angiosperms and the tracheophyte ancestor using genomes alone.

Functional analysis of new angiosperm orthogroups – To evaluate the functional roles that new genes may have played in the evolution of plants, we extracted the *Arabidopsis thaliana* genes in these new angiosperm and new seed plant orthogroups and super-orthogroups and examined the enrichment of Gene Ontology (GO) slim terms associated with these genes relative to the whole *Arabidopsis* genome in agriGO using the hypergeometric statistical test and the Yekutieli (FDR under dependency) multi-test correction method (*141*). GO slim enrichment for new angiosperm and new seed plant orthogroups is presented in Table S16.

Evolutionary analysis of genes with known roles in flowering – We also examined the origin and duplication history of orthogroups containing genes known to play roles in floral timing and development (Table S17), or genes identified based on expression patterns in functional screens in *Arabidopsis* flowers (Table S18).

Timing and duplication history of known floral genes – We used the orthogroup classification and gene family phylogenetic results (Table S13) to investigate the evolutionary history of 232 *Arabidopsis* genes with floral functions (*142-148*). Phylogenetic analysis identified 21 of these floral genes as members of orthogroups with duplicated lineages that were retained from the *epsilon* genome duplication (Tables S13, S17), including clavata (*CLV1*), several phytochromes, several MADS-box genes (shatterproof, *SHP*; seedstick, *STK*), auxin response factor (*ARF*) and Argonaut (*AGO*) genes, among others. Nine orthogroups with seed plant wide gene duplications contained floral genes including *CLV1*, cryptochromes (*CRY*), curly-leaf (*CLF*), and clock-associated genes *ADO3* and zeitlupe (*ZTL*). Orthogroups identified as “new in angiosperms”, “new in seed plants” or “ancestral in land plants” contained 19, 41, and 22 floral genes, respectively (see main text).

Enrichment for timing of origin of floral gene classes – Having identified the overall timing for the origin of orthogroups containing 232 genes with known functions in flower development, we then compared the distribution of floral orthogroup origins to the genome background of all other orthogroups in the *Arabidopsis* genome. Compared to the genome background, floral genes of known function are overrepresented in orthogroups that were present in the ancestral tracheophyte, originating with seed plants, and originating with angiosperms, but they are underrepresented in orthogroups that originated in derived angiosperm lineages (Table S18). Similar patterns were observed for a collection of 2,146 B-class target genes (*41*), and also for a set of 573 floral genes identified through massively parallel signature sequencing (MPSS), except that the latter set had lower representation by genes originating with angiosperms. Interestingly, 1162 stamen genes (*39*) and 260 carpel genes (*39*) identified through expression analysis of specified mutant plants had lower frequencies of “old” orthogroups originating with tracheophytes, and were overrepresented by genes in “new” orthogroups that originated within derived angiosperm lineages. Origin patterns for genes known or inferred to have roles in ovule development (*40*) were not significantly different from the background genome, although they were similar in pattern to stamen and carpel. These results suggest that although much of the floral development genetic toolkit is ancient, gene recruitment has been an ongoing process in floral evolution, and that different components of the flower have recruited genes at different rates through history, with a larger proportion of “new orthogroups” being represented by genes functioning in the inner floral whorls than by genes functioning in the perianth.

Global gene family phylogenetic analysis – Amino acid sequence alignments for each orthogroup were generated with MAFFT (*149*) using default parameters. Corresponding DNA sequences were then forced onto the amino acid alignments using custom Perl scripts. DNA alignments were trimmed using trimAl (*150*) using the heuristic automated1 method, and the trimmed DNA alignments were then used in subsequent phylogenetic analysis. Maximum likelihood (ML) analyses were conducted using RAxML version 7.2.1 (*151*), searching for the best ML tree with the GTRGAMMA model and conducting 100 bootstrap replicates. In total, we constructed 11,519 phylogenetics trees containing *Amborella* genes. The remaining 1,021 orthogroups with *Amborella* genes contain fewer than four total genes and were not suitable for phylogenetic analysis.

Identification and timing of pre-angiosperm gene duplications – Among the 11,519 gene family trees, 473 orthogroups showed at least one pre-angiosperm gene duplication with bootstrap support greater than 50% and 341 with bootstrap support greater than 80% following the criteria of Jiao *et al.* (*7*). To evaluate the timing of these ancient gene duplications, we applied a relaxed molecular clock to the 473 orthogroups supporting gene duplications before the separation of *Amborella* and the rest of the angiosperms (BS>=50%). We estimated the divergence times of 414 nodes separating the two paralogous clades in r8s (*152*), using semi-parametric penalized likelihood (PL) approach with the truncated Newton (TN) optimization algorithm. The smoothing parameter was determined by cross-validation. We used the following dates in our estimation procedure, following Jiao *et al.* (*7*): minimum age of 400 mya and maximum age of 450 mya for the divergence of *Physcomitrella patens* (*153*), a fixed constraint age of 400 mya for the divergence of *Selaginella moellendorffii* (*154*), a minimum age of 125 mya and maximum age of 150 mya for the split of monocots and eudicots (*155*, *156*), and minimum age of 125 mya and maximum age of 190 mya for the split of *Amborella* and the rest of the angiosperms (*157*). The distribution of inferred duplication times was then analyzed using EMMIX (*158*) to separate the components in a mixture model. Two statistically significant components were identified with mean divergence times at 244 and 341 mya (Fig. S17). These peaks correspond to the two peaks previously identified by Jiao *et al.* (*7*, *159*), 192 mya for *epsilon* and 319 mya for *zeta*, but these should be more accurate estimates because of the much denser sampling of taxa and genes in this analysis, particularly for nodes subtending branches with *Amborella* genes.

To further refine the phylogenetic placement of these ancient gene duplications, EST assemblies from basal angiosperms and gymnosperms were added to the orthogroup alignments and short or highly divergent unigenes were removed prior re-estimating gene trees. The resulting phylogenies were screened for gene duplications shared in seed plants and angiosperms. Angiosperm-wide (*epsilon*) duplication was supported in 115 (BS>=80%) or 187 (BS>=50%) orthogroups, and seed-plant-wide (*zeta*) duplication was supported in 104 (BS>=80%) or 207 (BS>=50%) orthogroups.

Terminal duplications – Gene family trees were also examined for terminal duplications in *Amborella*, e.g., gene duplications in outgroup rooted trees where two (or more) annotated *Amborella* genes were resolved on an "*Amborella* only" branch. 498 terminal duplications (95 sets with >2 genes) in 485 orthogroup trees were found with BS>=50%, of which, 466 terminal duplications (92 sets with >2 genes) in 452 trees were supported by BS>=80%, including tandem duplications (as defined above). Of the 403 resolved gene pairs with K_s value calculated, 121 pairs are tandem duplications, and 282 pairs are non-tandem duplications.

Duplicated gene divergence (K_s analysis) – We sought evidence for ancient genome duplication in the divergence patterns for duplicated genes in the *Amborella* genome (*7*, *159*).

Rate of synonymous substitution (K_s) – Paralogous pairs of sequences were identified from best reciprocal matches in all-by-all BLASTP searches of the annotated gene set in *Amborella*. We matched the best reciprocal hits before and after removing tandem duplicated pairs from the BLAST output, and calculated the two sets of pairs independently. Protein sequences were aligned using MUSCLE 3.6 (*160*). Nucleotide sequences were then forced to fit the amino acid alignments using PAL2NAL (*161*). The K_s (or D_s) values were calculated using a simplified version of the model of the Goldman and Yang maximum likelihood method implemented in the codeml package of PAML (*105*). The Ks frequency in each interval size of 0.05 within the range of 0 to 3.0 was plotted (Fig. S18).

The K_s distribution of reciprocal best BLAST hits in the genes in the *Amborella* annotation identifies two ancient peaks (Fig. S18a), consistent with Fig. S17 and the phylogenomic analysis of Jiao *et al.* (*7*).

Phylogenetic analysis of gene pairs identified on syntenic blocks – From the intra-genomic synteny analysis in *Amborella* (see supplementary text section 3), 466 gene pairs located in 47 syntenic blocks putatively resulting from the pre-angiosperm *epsilon* WGD, were identified. Of these gene pairs, just 44 pairs were shown to have phylogenetic signal of angiosperm-wide gene duplication, and 11 were shown to be *Amborella* terminal duplicates (Table S10). The remaining pairs either were not contained in the same orthogroup (292), did not have outgroup taxa to root the tree, had too few in group sequences to validly score the tree, or the tree topology was weakly supported.

To see if more focused analysis of gene pairs from manually validated syntenic blocks could resolve additional phylogenetic signal for ancient duplication, we took a more detailed look at the 155 gene pairs that were identified on the six large syntenic blocks that had been selected for manual curation (Fig. S10). For these analyses the regions had already been identified as syntenic blocks within the specified K_s window, so we were able to score angiosperm-wide duplications even when the phylogenies were unrooted (Table S11). In cases where the *Amborella* duplicate genes had been assigned to two different orthogroups or one of the *Amborella* genes was divergent enough to be left as a singleton in the orthogroup classification, we merged the related orthogroups and singletons identified in the synteny analysis, added basal angiosperm and gymnosperm transcript sequences, and re-computed the phylogeny, providing additional opportunities to identify angiosperm-wide duplications.

Phylogenetic analyses of these expanded orthogroup sets resolved angiosperm-wide duplications in 80 (BS50) or 77 (BS80) of the 155 gene pairs tested. These results illustrate the complementary nature of the phylogenetic and structural syntenic data, and confirm that both are accurate indicators of the same *epsilon* WGD event.

Combined with the global phylogenetic screen for angiosperm-wide duplicates (473 orthogroups at BS50), a total of 851 non-redundant, high-confidence gene pairs across the *Amborella* genome still retain a clear signal from the *epsilon* WGD that occurred some 200 million years ago. This estimate is likely to be an underestimate of the number of gene pairs retained from *epsilon*

because our assays to detect duplicated regions were conservative, requiring either (for phylogenies) specified minimal taxon participation in an orthogroup, including an outgroup to root trees, score-able topologies, and bootstrap support, or (for syntenic blocks) a minimum of five homologous genes with a maximum of 40 intervening genes were required to seed the syntenic regions plus a specified range of synonymous sequence divergence between gene pairs.

6: MADS-Box Genes

MADS-box genes encode a family of transcription factors and are involved in numerous developmental processes in plants, animals and fungi; their functions are correlated with the emergence of many key innovations. In plants, MADS-box transcription factors are associated with floral organ identities, inflorescence identities, floral timing, and other processes (*162-166*). The plant MADS-box genes are believed to be derived from at least two lineages: the SRF-like (type I) and MIKC-type (type II) (*163, 166*). The best known plant MADS-box transcription factors are encoded by Type II genes involved in floral organ identity determination, and previous studies indicated that duplications of these genes were associated with the origin and diversification of flowering plants (*163, 166-168*). Because the protein products of these genes usually bind to target DNA as dimers, attention has also been paid to the evolution of the protein-protein interactions among them (*50, 169, 170*).

Phylogenetic analysis – To identify MADS-box genes from the *Amborella* genome, we performed profile hidden Markov model (HMM) searches (*138, 139*) against the predicted proteome and found 12 type I and 23 type II MADS-box genes. Because the annotation was still not perfect, we also performed tBLASTn searches against the *Amborella* genome, and found one additional type I MADS-box gene. Taken together, the *Amborella* genome contains 36 MADS-box genes (Tables S19 and S20). The relatively small number of MADS-box genes compared with members of eudicots and monocots is consistent with the lack of lineage-specific whole genome duplications in *Amborella* since its divergence from all other extant angiosperms.

To perform phylogenetic analysis, we retrieved the protein sequences of MADS-box homologs from the predicted proteomes of *Arabidopsis thaliana* (TAIR10, http://www.arabidopsis.org/), *Oryza sativa* (GRAP6, http://rice.plantbiology.msu.edu/) and *Populus trichocarpa* (JGI version 1.1, http://www.phytozome.net/poplar). The resulting dataset was screened using the Pfam website (http://pfam.sanger.ac.uk) to remove sequences without the MADS domain. To avoid possible long-branch attraction in phylogenetic estimates, MADS-box genes from *Ananas comosus*, *Antirrhinum majus*, *Asparagus virgatus*, *Dendrobium thyrsiflorum*, *Houttuynia cordata*, *Magnolia praecocissima* and *Persea americana* were also added for the analysis of the *AG/STK* and *AGL2/9* lineages. Protein sequences were initially aligned using ClustalX 1.83 (*171*) and the alignment was manually adjusted in GeneDoc 3.2 (*172*). A DNA version of alignment was also generated using aa2dna (http://www.personal.psu.edu/nxm2/software.htm) with the protein sequence alignment as a guide. Phylogenetic analyses were carried out using the maximum likelihood (ML) method in PhyML 2.4 (*173*). The most appropriate model, HKY, was used, with proportion of invariable sites and gamma distribution parameter optimized automatically and a BIONJ tree applied as a starting point. Support for each branch was assessed with 100 bootstrap replicates.

Phylogenetic analysis showed that these genes could represent a total of 21 clades (three for type I genes and 18 for type II genes) of MADS-box genes in the MRCA of angiosperms (Fig. S19 and S20). Good phylogenetic resolution and bootstrap support for most clades containing *Amborella* MADS-box genes suggest that the *Amborella* genome annotation is of high quality.

The inclusion of *Amborella* genes has allowed a better estimate of the origin of various angiosperm MADS-box gene lineages. Known floral organ identity genes are members of eight

major lineages (i.e., *AP1*/*SQUA*, *AP3*/*DEF*, *PI*/*GLO*, *AG*, *STK*, *AGL2*/*SEP1*, *AGL9*/*SEP3*, and *AGL6*, also known as A, B1, B2, C, D, E1, E2, and G lineages, respectively; Fig. 4A of the main text), each with one or two copies from Amborella, indicating that they existed in the MRCA of extant angiosperms but were likely derived from three ancestral lineages in the MRCA of extant seed plants. It is very likely that the genetic network for floral organ identity existed in the MRCA of angiosperms or even earlier (*174-176*).

In addition, the newly identified *Amborella* genes have provided crucial information for the early angiosperm history of other lineages, including the *AGL12*, *AGL15* (embryo development), *ANR1* (root), *StMADS11* (including the flowering time regulators *SVP* and *AGL24*), MIKC*-P and MIKC*-S. In particular, it is noteworthy that the *OsMADS32* lineage, which was previously believed to be monocot-specific (*45*), actually has an ortholog in *Amborella*, indicating that this lineage is rather ancient (Fig. S20). A brand new lineage, TM8/*PtMADS35*, whose members have been found in *Populus* and *Vitis* (*46, 47*), but not in *Arabidopsis* or rice, was also identified in *Amborella*. This suggests that *TM8/PtMADS35*-like genes have been lost independently in different plant groups (Fig. S20). Similarly, the discovery of two homologs of the gymnosperm *GMM13* gene in *Amborella* suggests that the *GMM13* lineage may have two ancient subgroups in angiosperms, possibly resulting from the pre-angiosperm WGD, although the support is low. These results clearly indicate that the *Amborella* genome not only acts as a reference, but also provides new insights into the evolution of MADS-box genes.

Protein-protein interactions between floral MADS-box proteins – MADS-box transcription factors involved in floral development usually form dimers or higher-level complexes before they can bind to the promoter region of their downstream targets. We conducted yeast-two-hybrid (Y2H) assays to investigate the protein-protein interactions (PPIs) between the *Amborella* counterparts of these genes using the GAL4-based MATCHMAKER Two-Hybrid System (Clontech, Mountain View, CA). Full-length coding sequences of ten *Amborella* MADS-box genes were introduced into pGADT7 and pGBKT7, and confirmed by sequencing. All constructs were transformed into yeast (*Saccharomyces cerevisiae* strain AH109). Because the single transformants containing either the pGBKT7–Am.tr.AGL9 or pGBKT7–Am.tr.AGL6 construct showed expression of the *HIS3* reporter gene without an interactive partner construct, truncated constructs of these genes lacking the C-terminal ends were used. Confirmation of the transformants and interaction analyses were conducted as described in the manufacturer's instructions. Experiments were repeated three times, using the empty-vector transformants (pGADT7, pGBKT7) as negative controls.

We observed interactions between Am.tr.AP1 and each of Am.tr.AP1, Am.tr.AP3.1, Am.tr.AG, Am.tr.STK, Am.tr.AGL2/9, and Am.tr.AGL6; between Am.tr.AP3.1 and Am.tr.PI.1/2 or Am.tr.STK; between Am.tr.AG and each of Am.tr.STK, Am.tr.AGL2/9, and Am.tr.AGL6, between Am.tr.STK and Am.tr.AGL2/9 or Am.tr.AGL6; and between Am.tr.AGL6 and Am.tr.AGL6 or Am.tr.AGL2/9 (Fig. S21). Apparently, Am.tr.AP1 and Am.tr.AGL6 each were capable of forming homodimers, whereas all others can only form heterodimers. The PPI patterns in *Amborella* are generally consistent with those in other angiosperms but show clear differences from those in gymnosperms (Fig. 4B of the main text and Fig. S21).

Members of the *AP3*/*DEF* and *PI/GLO* lineages (B class genes), which are associated with the identity of petals and stamens, form heterodimers in angiosperms. Their counterparts in

gymnosperms, however, can only form homodimers and heterodimers between similar proteins encoded by recently duplicated genes (Fig. 4B of the main text). We therefore constructed the deduced ancestral protein sequences of these two lineages in the MRCA of angiosperms and investigated the PPI patterns. These two ancestral proteins formed homodimers and heterodimers, suggesting that a shift in PPI patterns occurred after the gene duplication has probably contributed to the innovations in the regulatory network and the origin of the flower.

Evolution of floral MADS-box genes was tightly associated with the origin of the flower – We have investigated the evolutionary history of the MADS-box gene family and interactions among the floral MADS-box proteins in *Amborella* to infer the MADS box gene content in the ancestral angiosperm and establish whether PPIs that distinguish angiosperms from gymnosperms are also found in *Amborella*. We found that the floral MADS-box genes increased by gene duplication from at least three in the MRCA of extant seed plants to at least eight in the MRCA of extant angiosperms. Correspondingly, the PPI patterns among these proteins have become more complex as a consequence of gene duplication. Phylogenetic analyses suggest that the ancestor of seed plants had fewer reproductive MADS-box genes. Therefore, the seed plant ancestral MADS-box proteins probably formed homodimers, which are maintained in gymnosperms that have been studied thus far, even when they have recent duplicates. On the other hand, following the duplication and diversification of MADS-box genes in angiosperms, their encoded proteins have evolved to form heterodimers, greatly expanding the repertoire of functional dimers. Previous studies have indicated that the diversification of duplicate genes/proteins is very important for plant evolution because it can increase the complexity in PPIs and lead to innovations in phenotype. Our results further support the hypothesis that duplication, diversification and expansion of the types of PPI patterns of floral MADS-box proteins were associated with the origin of the flower.

7: Glycogen Synthase Kinase 3 Genes

Glycogen synthase kinase 3 (*GSK3*) proteins, also known as SHAGGY-like kinases, have important roles in a wide range of cellular processes throughout eukaryotes (*51*). In animal development, products of *GSK3* homologs participate in the critically important Wnt signaling pathway that regulates cellular differentiation, patterning, and growth in perhaps all metazoans (*177*). The *GSK3* homolog in the protozoan *Dictyostelium discoideum* is also involved in the regulation of development (*178*). Recognition of possible roles of GSK3 in human disease has prompted recent interest in these genes in medicine (*51*, *179*). Compared to animals, *GSK3* genes in land plants have radiated into a relatively large multi-gene family (*180-182*). For example, five *GSK3* genes have been reported from the moss *Physcomitrella patens* (*183*), and 10 *GSK3* genes are present in the genome sequence of the flowering plant *Arabidopsis thaliana* (*181*). Conceivably, therefore, *GSK3* genes have had a dynamic history of gene duplication during the course of land plant evolution and have acquired roles in plant-specific processes. For example, different *Arabidopsis GSK3* genes function in hormonal signaling, osmotic stress responses (*51*), and flower development (*180*, *184*).

Previous phylogenetic analyses suggest that four major lineages evolved in the land plant branch of the *GSK3* gene family (*182*), but their origins and relationships are currently unclear. *Physcomitrella* GSK3 genes occupy different positions relative to groups of angiosperm genes in various analyses (*183*, *185*), and the positions of fern and gymnosperm *GSK3* sequences have been similarly fluid (Figures 5-7 in *185*). We use newly available genomic resources, including the *Amborella* genome sequence, to reconstruct the phylogenetic history of *GSK3* genes in land plants (embryophytes).

Phylogenetic analysis - To retrieve members of the *GSK3* gene family in *Amborella*, *Arabidopsis* and rice sequences were used to seed tblastx searches of EVM 27-predicted *Amborella* gene models. Additional land plant sequences were obtained by similar searches of the Phytozome (http://www.phytozome.net/); AAGP (http://ancangio.uga.edu/); TIGR Plant Transcript Assemblies (http:// http://plantta.jcvi.org/), 1KP (http://www.onekp.com/), and NCBI sequence databases. We obtained a total of 444 *GSK3* genes from 67 species representing all major green plant lineages, including six from *Amborella*. Nucleotide sequences were translation-aligned using MAFFT and phylogenetic analyses were conducted using RAxML 7.3.0 (*151*). All phylogenetic analyses were performed on the University of Florida High Performance Computing cluster (http://hpc.ufl.edu/).

The Maximum Likelihood (ML) phylogeny of land plant *GSK3* genes (Fig. S22) indicates a grade of “bryophyte” sequences, above which ancient gene duplications occurred along the branches leading to extant tracheophytes (vascular plants), euphyllophytes (ferns and seed plants), and spermatophytes (seed plants), respectively. A subsequent duplication, possibly a WGD, along the branch leading directly to extant angiosperms produced a total of five angiosperm-wide gene lineages and a sixth represented by a single *Amborella* gene.

The *Amborella* genome alone retains all *GSK3* genes duplicated on the ancestral branch to angiosperms.

Sister angiosperm-wide gene lineages (angiosperm *GSK3* Group I-1 and I-2) imply duplication of the ancestral Group I locus along the branch to extant flowering plants with retention of both duplicate copies. This duplication coincides with a proposed WGD event that occurred ca. 192 mya along the ancestral lineage of angiosperms (*7*). Synchronous "angiosperm" duplications are not immediately obvious for the other angiosperm *GSK3* gene lineages, but the two *Amborella* genes in the GSK3 Group II gene lineage may be noteworthy. One (scaffold00170.9) occupies the expected position at the base of a pan-angiosperm gene lineage (angiosperm *GSK3* Group II-1), while the other (scaffold00019.82) is placed within a paraphyletic group of gymnosperm genes. The existence of two *Amborella GSK3* Group II genes indicates a duplication event, and their phylogenetic distance suggests that it was ancient. We propose that two angiosperm Group II lineages (angiosperm *GSK3* Groups II-1 and II-2) originated through a gene duplication event along the branch to extant angiosperms, followed by loss of the angiosperm *GSK3* Group II-2 lineage after the separation of *Amborella* from other flowering plants. The placement of the surviving *Amborella* gene among gymnosperm genes, instead of sister to angiosperm *GSK* Group II-1, may be an artifact of phylogeny reconstruction, rather than a reflection of true relationship. Thus, among flowering plants, *Amborella* alone may contain all the *GSK3* gene lineages descended from duplications along the ancestral branch to extant angiosperms.

8: Seed Proteomics and 11S Globulin Proteins: Understanding Seed Proteins in the Ancestral Angiosperms While Validating the *Amborella* Genome Annotation

Seed germination processes are diverse but can be characterized in part by studying seed storage proteins, because the remobilization of these proteins during germination and initial seedling growth is crucial for the establishment of vigorous seedlings (*186*). We used a proteomic approach to study the seed storage proteins of *Amborella trichopoda*, with the goal of identifying proteome changes that could be associated with the origin and early diversification of angiosperms. We focused most of our attention on globulins. Whereas 7S vicilin type globulin genes sampled from monocots and eudicots cannot be aligned (*187*), 11S legumin type globulins have been characterized and compared across seed plants in evolutionary analyses of these important protein-coding genes (*55, 56, 188*). Here we refine understanding of this gene family with the characterization of three *Amborella* 11S legumin type globulin homologs. These new data allow us to determine the timing of key changes in the genes that distinguish previously characterized angiosperm and gymnosperm members of this gene family.

Preparation of protein extracts – Fruits of *A. trichopoda* (Fig. S23) were collected at “plateau de Dogny-Sarraméa” (New Caledonia; 21°37'0" N, 165°52'59" E). Protein extracts were prepared from 25 mature *Amborella* seeds (approximately 350 mg) after removing the dried pericarp tissue. Following grinding of the seeds in liquid nitrogen with a mortar and pestle, the protein characterization protocols of Osborne (*189*) were applied. Albumins were first extracted by placing the powder in 1.5 ml of distilled water containing a cocktail of protease inhibitors (Roche Molecular Biochemicals). After incubation for 20 min at 4°C, the suspension was centrifuged (15 min, 4°C, 40,000 *g*), and the supernatant containing albumins was recovered. Following three washes, the pellet was resuspended in 1.5 ml 0.8 M NaCl containing the cocktail of protease inhibitors for 20 min at 4°C. After centrifugation (4°C, 40,000 *g*, 15 min), the supernatant containing the globulins was collected. Following precipitation of the proteins in trichloroacetic (TCA) solution (10% TCA and 0.07% 2-mercaptoethanol in acetone) overnight at -20°C, the pellet was suspended in 1.5 ml of the buffer used for the isoelectrofocusing (IEF) step (see below) and containing 8 M urea, 2% [w/v] 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate (CHAPS), 0.5% [v/v] immobilized pH gradient (IPG) buffer 4-7, 20 mM dithiothreitol (DTT), and 0.01% [w/v] bromophenol blue (*190*). Protein concentration in the various extracts was measured according to Bradford assay (*191*) using bovine serum albumin as a standard.

Two-dimensional gel electrophoresis and staining – Isoelectric focusing (IEF) of protein extracts in IEF buffer was conducted using 18-cm (3-10 nonlinear) IPG strips on a Multiphor II system (Amersham-Pharmacia Biotech). Strips were rehydrated for 24 h at 22°C with the thiourea/urea lysis buffer containing 2% (v/v) Triton X-100, 20 mM DTT, and the protein extracts. IEF was performed at 22°C for 1 min at 300 V, 1 h 30 at 500 V and for 17 h at 3,500 V. IEF strips were equilibrated for 2 x 20 min in 2 x 100 ml of equilibration solution containing 6 M urea, 30% (v/v) glycerol, 2.5% (w/v) sodium dodecyl sulfate (SDS), 0.15 M bis-Tris, and 0.1M HCl (9-10). DTT (50 mM) was added to the first equilibration solution, and iodoacetamide (4% [w/v]) was added to the second (*190*). Equilibrated gel strips were placed on top of vertical polyacrylamide gels (10% [v/v] acrylamide, 0.33% [w/v] piperazine diacrylamide, 0.18 M Trizma base, 0.166 M HCl, 0.07% [w/v] ammonium persulfate, and 0.035% [v/v] tetramethylethylenediamine [TEMED]). A denaturing solution (1% [w/v] low-melting agarose,

0.4% [w/v] SDS, 0.15 M bis-Tris, and 0.1 M HCl) was loaded on gel strips. After agarose solidification, electrophoresis was performed at 10°C in a buffer (pH 8.3) containing 25 mM Trizma Base, 200 mM taurine, and 0.1% (w/v) SDS, for 1 h at 35 V and 110 V overnight. Ten gels were run in parallel (Isodalt system from Amersham Pharmacia Biotech). All 2D gel assays were run at least in triplicate on three independent protein isolations.

Gels were stained with silver nitrate either as described by Blum *et al.* (*192*) for the analytical gels or by Shevchenko *et al.* (*193*) for the preparative gels using the Hoefer Automated Gel stainer from Amersham Pharmacia Biotech. Silver-stained gels were scanned with the Sharp JX-330 scanner equipped with the Labscan version 3.00 from Amersham Pharmacia Biotech.

Peptide identification – Protein spots were analyzed by liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) on the PAPPSO platform (B. Valot, T. Balliau, M. Zivy, INRA Moulon, France; http://pappso.inra.fr). Based on the spectrum generated, proteins were identified using the X-Tandem software (*194*, *195*). All peptides with an E-value match lower than 0.05 were parsed with our X-Tandem pipeline v. 3.1.2 (http://pappso.inra.fr/bioinfo/xtandempipeline). Proteins identified with at least two unique peptides and an E-value lower than 10E-8 were validated. Peptide sequences and protein identifications are listed in Table S21.

We found 210 albumin spots corresponding to 162 *Amborella* genes and 129 globulin spots corresponding to 24 *Amborella* genes, of which 99 are globulin spots. A reference map for *Amborella* 11S globulins shows the localization of the 99 protein spots identified by LC-MS/MS analysis (Fig. S24). These peptides were mapped to three distinct 11S legumin type globulins found in the *Amborella* annotation (evm_27.model.AmTr_v1.0_scaffold00067.7, evm_27.model.AmTr_v1.0_scaffold00067.11, and evm_27.model.AmTr_v1.0_scaffold00067.12) (Table S21).

Peptide coverage of 11S globulin gene models – Peptide coverage of evm_27.model.AmTr_v1.0_scaffold00067.7 and evm_27.model.AmTr_v1.0_scaffold00067.12 is 84% and 73%, respectively (Fig. S25). Experimentally determined peptide sequences frequently overlap (Table S21), thereby increasing confidence in validation of the gene models. Furthermore, this high coverage validates the quality of genome sequencing data and of the exon/intron predictions. Coverage of evm_27.model.AmTr_v1.0_scaffold00067.11 was much lower (4.1%) (Fig. S25), presumably suggesting that this 11S globulin accumulates in lower concentrations in seeds.

Conserved gene structure of 11S globulins in *Amborella* – A multiple sequence alignment of the *Amborella* 11S globulin gene models to canonical seed plant 11S globulin genes (*Glycine max*, *Ginkgo biloba*, and *Arabidopsis thaliana*) revealed that two Cys residues involved in the disulfide bond linking the A- and B-subunits of 11S globulins are conserved in *Amborella* (red boxes in Fig. S26A).

Importantly, the conserved vacuolar processing enzyme/asparaginyl endopeptidase (VPE/Asn-endopeptidase) cut site is conserved in *Amborella* (Fig. S26A). Cutting at this site yields A- and B-subunits (red arrow in Fig. S26A, with the blue and green arrows designating the A and B subunits, respectively). This limited proteolysis is regularly directed to an Asn-Gly (N-G)

junction, resulting in an N-terminal glycyl residue of the B subunit formed. This proteolytic processing triggers protein conformational changes that enable dense packaging and long-term stable storage of reserves within the protein storage vacuoles (PSV). After transport to the PSV, proteolytic processing at the conserved N-G peptide bond by the VPE/Asn-endopeptidase converts the pro-form of 11S globulins to the two disulfide-linked mature polypeptides A- and B-subunits, which is accompanied by further assembly of the trimer precursor-protein complexes into mature hexamers that sediment as a 11S complex (*196*, *197*).

By conducting BLAST analyses we found that as in *Arabidopsis* (*196*), the *Amborella* VPE family is composed of four genes (corresponding to the αVPE, βVPE, γVPE, δVPE subunits), namely: evm_27.model.AmTr_v1.0_scaffold00036.100, scaffold00002.262, scaffold00002.263, and scaffold00002.265, thereby suggesting that at least the two *Amborella* 11S globulins encoded by the gene models evm_27.model.AmTr_v1.0_scaffold00067.11 and evm_27.model.AmTr_v1.0_scaffold00067.12, which do contain the N-G cutting site (Figs. S25 and S26A), are processed by an *Amborella* VPE system. In marked contrast, the *Amborella* 11S globulin encoded by the evm_27.model.AmTr_v1.0_scaffold00067.7 gene deviates conspicuously from the two above *Amborella* globulin precursors as the N-G junction is replaced by a N-V-I sequence (Figs. S25 and S26A). The alteration of this N-G junction motif is found for 11S globulins from gymnosperms, as observed in *Gingko biloba* (Fig. S26A) and *Metasequoia glyptostroboides* (*198*), thus highlighting an ancestral character of this 67.7 11S *Amborella* globulin.

The sequence alignment also revealed additional conserved motifs, including the amino acid residues involved in trimer formation and stabilization of 11S globulins (*57*) (yellow shading in Figure S26A), and those involved in the correct globular folding of 11S globulins (*57*) (blue shading in Figure S26A).

Furthermore, this analysis revealed protein regions that are homologous to PX14G (box a) and SX3G (box b) motifs (Fig. S26A) adjacent to introns III and IV in 11S globulin genes, respectively (*54*). The B-subunits of gymnosperm globulins exhibit a highly conserved intron/exon structure differing from those of angiosperms (*54*). In particular, the 11S globulin genes from gymnosperms display an additional intron (referred to as intron IV) that uniformly interrupts the sequence coding for the C-terminal part of the basic-subunits of 11S globulins (*54*). In contrast to the *Gingko biloba* 11S globulin homolog, intron IV is absent in *Amborella* 11S-globulin-encoding genes and other angiosperm homologs (Fig. S26B), consistent with the finding that this intron was lost in the MRCA of angiosperms (*54*).

We investigated the distribution between ordered and disordered regions in 11S globulin genes from *Amborella* (Fig. S26C). It appears that the gene model 67.11 is very different from other 11S globulins as it lacks a disordered region in the A subunit (Fig. S26C). During seed germination the initial mobilization of 11S globulins preferentially begins with proteolysis of the A chains (*199*). It is therefore possible that lack of this disordered region somehow impedes processing of the globulin encoded by the 67.11 gene model during seed imbibition. The function of this protein may have diverged from the other 11S globulins given its divergent structure and low concentration relative to the other 11S globulin homologs.

Phylogenetic analysis of 11S globulins in seed plants – Multiple sequence alignment of 65 available 11S globulin protein sequences from 55 diverse seed plants (Table S22) was performed using MAFFT and a gene tree was estimated using maximum likelihood (Fig. S27). Bootstrap analysis was performed with 100 pseudoreplicates.

The phylogenetic tree shows clades comparable to those described by APG III (*200*), confirming the ability of 11S globulins to be used for phylogenetic studies (*56*). The position of *Amborella* close to magnoliids (i.e., *Asarum europeaum* (Piperales), *Magnolia salicifolia* (Magnoliales)) generally agrees with broad molecular phylogenetic analyses (e.g. *15*), although the sister relationship of magnoliids and *Amborella* is not consistent with current understanding of species relationships. As suggested by Li *et al.* (*56*), most monocot 11S globulin homologs form a clade that is sister to a eudicot gene clade.

9: Terpene Synthase Gene Family

Identification and phylogenetic analysis of terpene synthase genes - The annotated proteome of *Amborella* was searched with two Pfam models, PF01397 and PF03936, which correspond to the conserved domains localized at the N and C termini of known terpene synthases (TPSs), respectively, using the hmmsearch command in the HMMER package (*201*). The significant hits were selected with an E value ≤ 1e-2 as the cutoff and further validated based on their significant sequence similarities to known plant TPSs in GenBank as identified through a BlastP search using the protein sequences of individual putative *Amborella* TPSs as queries. To understand their evolutionary relatedness, the TPSs identified from *Amborella* and the TPSs identified from other selected plants were subject to phylogenetic reconstruction. Multiple protein sequence alignments of TPSs were performed using MAFFT (*149*). Maximum likelihood trees were built using PhyML v2.4.4 (*173*) with the JTT model and 100 bootstrap replicates.

Terpene synthases in *Amborella* and gene evolution - Terpene syntheses (TPSs) are pivotal enzymes for the biosynthesis of terpenoids (*202*), which constitute the largest class of plant secondary metabolites and are involved in a multitude of plant interactions with the environment (*58*). A total of 39 gene models that encode proteins with significant sequence similarity to known plant TPSs were initially identified from the *Amborella* genome. Five pairs of these *TPS* genes were noted for their chromosomal locations and gene structures. The two genes in each of the five pairs are neighbor genes on the same scaffold, and their encoded proteins correspond to the N- and C-termini of known plant TPSs, respectively. By merging the two genes in each of the five pairs, we propose that the *TPS* gene family in *Amborella* contains 34 members (Table S23). A significant portion of the *Amborella TPS* genes are shorter than known plant *TPS*s. Manual annotation had only modest improvement on extending the coding sequences of these short *TPS* genes. It remains to be determined whether these short *TPS* genes are true partial genes or are due to incomplete sequence information. In our analysis, the 25 *Amborella* TPSs longer than 200 amino acids were treated as full-length genes.

TPS genes are ubiquitous in land plants and share a common evolutionary origin (*203*), with the exception of a new class of *TPS* genes recently found in *Selaginella moellendorffii* (*204*). Typical plant *TPS* genes have been recognized in seven subfamilies: *a*, *b*, *c*, *d*, *e/f*, *g*, and *h* (*59*). *TPS* genes from flowering plants belong to five of them, including TPS-c, e/f, a, b, and g. In contrast, gymnosperm *TPS* genes belong to *TPS-c*, *e/f*, and *d* subfamilies. While the *TPS-c* and *TPS-e/f* subfamilies are shared by gymnosperms and angiosperms, *TPS-d* appears to be specific to gymnosperms and contains diterpene synthases, monoterpene synthases, and sesquiterpene synthases (*205*). In contrast, *TPS-g*, *TPS-b*, and *TPS-a* subfamilies are specific to angiosperms, with the former two subfamilies consisting of monoterpene synthases and the latter consisting of sesquiterpene synthases. The 25 putative full-length *Amborella TPS* genes belong to four known subfamilies, including *TPS-c*, *e/f*, *b*, and *g* (Fig. S28). The most striking difference between the *Amborella TPS* genes and those from other flowering plants is the lack of *TPS-a* subfamily members in the *Amborella* genome. When all 34 *TPS* genes identified in the *Amborella* genome were analyzed phylogenetically instead of only the 25 putative full-length *Amborella TPS* sequences, the lack of *TPS-a* in *Amborella* was confirmed. This result suggests that the emergence of the *TPS-a* subfamily, which synthesizes the diverse C15 terpenoids, occurred after the divergence of *Amborella* from the stem lineage to the remaining extant angiosperms. Furthermore, an additional clade of *TPS* genes is confined to *Amborella*, based on currently

available data, and we refer tentatively to this set of *Amborella*-specific genes as the TPS-x subfamily.

10: Cell Wall and Lignin Biosynthesis

Chemical composition of wood - Extractive-free wood meals (40-60 mesh) from *Amborella* stem xylem were used for all analyses. Lignin content was determined according to the modified Klason lignin method (*206*), combining both the Klason lignin and acid-soluble lignin as the total lignin content. The extinction coefficient used was 110 l $g^{-1}cm^{-1}$ at 205 nm (*207*). Neutral sugar compositions were determined as alditol acetates from the acid-soluble fraction of the lignin analysis (*208*). A gas chromatograph equipped with a flame ionization detector (GC-FID, Agilent 7890A) and a DB-225 GC-column (Agilent) was used in the alditol acetate determination. The uronic acid content was determined using the sulfuric acid–carbazole method (*209*). Glucuronolactone was used as the standard, and the optical density at 530 nm was collected. Nitrobenzene oxidation analysis was conducted as per Katahira and Nakatsubo (*210*) with the following modifications. Approximately 100 mg of vacuum-dried extractive-free wood meals were reacted with 2N NaOH and nitrobenzene at 165°C for 2.75 hr. After the reaction, the mixtures were extracted with CH_2Cl_2, and the aqueous phases were acidified with 2N HCl. The solutions were further extracted with CH_2Cl_2, and the CH_2Cl_2 phases were trimethylsilylated with *N*-methyl-*N*-(trimethylsilyl)trifluoroacetamide (Sigma-Aldrich) and quantified by GC-FID (Agilent, 7890A) with a DB-5 GC-column (Agilent). Total chemical composition of wood is reported in Table S24 and monolignol subunit composition is reported in Table S25. A comparison of lignin subunit composition across plants is provided in Table S26.

Computational identification of *Amborella trichopoda* monolignol biosynthesis genes - The monolignol biosynthesis genes in *Amborella trichopoda* were identified from the EVM 27 version of the predicted *Amborella* gene models by reciprocal BLAST methods described in Shi *et al.* (*211*) using the 63 *Arabidopsis* (*212*) and 95 *Populus trichocarpa* (*211*) protein sequences of the 10 phenylpropanoid gene families. These phenylpropanoid genes include the monolignol biosynthesis genes and were first searched against all *Amborella* annotated gene models (EVM27) using blastp with the e-value cut-off set at 1-E03. The resulting protein sequences of the positive *Amborella* gene model hits were then queried against all 31,921 proteins in *Arabidopsis* (TAIR10, http://www.arabidopsis.org/) and 45,033 proteins in *P. trichocarpa* genome annotation V2.2. The resulting 97 *Amborella* gene models that had one of the above 63 *Arabidopsis* or 95 *P. trichocarpa* phenylpropanoid genes as the top hit were analyzed further for putative monolignol genes by phylogenetic analysis. Maximum likelihood phylogenetic trees with 100 bootstrap replicates were constructed using RAxML with nucleotide sequences forced onto the MAFFT multiple protein sequence alignment. The phylogenetic trees were plotted using FigTree (http://tree.bio.ed.ac.uk/software/figtree/) for all of the above phenylpropanoid genes of 10 enzyme families from *Arabidopsis*, *Populus*, and *Amborella* (Fig. S29A-H). The putative *Amborella* monolignol biosynthesis genes (Table S27) were identified by selecting the closest homologous *Amborella* gene models to each of the 12 *Arabidopsis* (*212*) and 21 *P. trichocarpa* (*211*, *213*-*215*) previously confirmed monolignol biosynthesis genes and inspected for their gene model structures and evidence of transcription on the *Amborella* genome browser (http://www.amborella.org).

Analysis of glycosyltransferase genes involved in cell wall biosynthesis - Glycosyltransferases were initially identified via BLAST, manually verified, and phylogenetic trees constructed following the methods described in Harholt *et al.* (*216*). Approximate likelihood-ratio test (*217*) values below 0.7 are shown on branches in the tree of Fig. S30.

The carbohydrate-active enzymes (CAZyome) (*67*) of *Amborella* include a smaller number of cell wall biosynthesis genes, but all the gene families seen in monocots and eudicots are present in *Amborella*. The diversity in cell wall genes in more derived angiosperms may result from gene or genome duplications during the radiation of angiosperms. For example, the fucosyltransferase, *FUT1*, which fucosylates xyloglucan (XG) is a member of family GT37. *FUT1* alone catalyzes this reaction, while FUT4 and 6 are involved in AGP-fucosylation (*218*). *Arabidopsis* and *Oryza* have experienced independent gene family expansions not observed in *Amborella* (Fig. S30). Family GT37 features a clade of *Physcomitrella*, *Selaginella*, and *Amborella* genes while the third *Amborella* gene is placed in a small clade with one putative ortholog in rice and one, *AtFUT3*, in *Arabidopsis*, which is known not to fucosylate XG (*219*). Gene models for all 10 enzymes involved in lignin biosynthesis have been identified in the *Amborella* genome (Table S28), with similar numbers of putatively functional orthologs in *Populus* and *Amborella*.

11: Transposable Element Annotation

Structural Annotation of LTR Retrotransposons – Annotation of LTR retrotransposons combined computational annotation predictions from LTR_Struc (*220*), LTR_Finder (*134*), LTRharvest (*221*) and LTR DeNovo (*133*). These *ab-initio* annotation programs identify putative full-length LTR retrotransposons in genomic sequence based on diagnostic structural features including long terminal repeats (LTRs) and target site duplications (TSDs). The database of 631 tRNAs produced by Aragorn (*124*) was used to predict the location of primer binding sites using LTR_Finder (*134*) and LTRdigest (*222*). Protein coding regions within these structurally predicted LTR retrotransposons were identified by profile hidden Markov model (pHMM) searches using the *batch_hmmer.pl* program of RepMiner (http://repminer.sourceforge.net/) and the hmmer search capabilities of LTRdigest (*222*) to search against a database of retrotransposon pHMMs downloaded from PFAM v. 25 (*223*) and GyDB v. 2.0 (*224*).

To facilitate combination of the 34,270 LTR retrotransposon predictions generated by multiple annotation programs, the computational evidences were converted to a common sequence ontology compliant GFF3 format (*225*) using the DAWGPAWS plant genome annotation pipeline (*21*). The predicted LTR retrotransposon models were then loaded into a GMOD SeqFeature-Store database (http://gmod.org) and combined using the *db_cluster_features.pl* program within DAWGPAWS. This process identified all spatially overlapping LTR retrotransposon models, and selected the longest LTR from each spatially overlapping set to be representative of that set. The resulting 15,778 putative LTR retrotransposons were then clustered into similarity-based groups using an all-by-all BLASTn analysis of the representative LTR sequences with an E-value threshold of $1x10^{-5}$ and returning the top 1,000 significant matches. The BLAST similarity matrix was used to cluster the sequences into groups by single-linkage clustering with subclusters defined by Markov clustering with MCL (*226*) using the default parameters. False positive predictions that resulted from simple repeat sequences in close proximity were identified by BLAST2Seq (*227*) analysis of individual LTRs using the *batch_blast2seq_self.pl* program of DAWGPAWS removing all sequences that had a blast2self length index that was greater than five. Additionally, 558 false positives resulting from endogenous pararetroviruses (EPRVs) computationally identified by as LTR retrotransposons by LTR_Finder (*134*), LTRharvest (*221*), or LTR DeNovo (*133*) were removed. The resulting set of sequences was clustered into families using a combination of MCL identified clusters and a similarity threshold of 70 percent identity in LTR sequences. The 70 percent identity was chosen over the more traditional 80 percent identity score (*71*) since over 10% of individual LTR retrotransposons insertions had LTRs that were more than 80% divergent from each other. Family nomenclature followed the classification system of Wicker *et al.* (*71*) and each family was given a unique name by random assignment using an online resource for roleplay game character naming (http://wownamegenerator.com/) that combined random syllables together to generate a pronounceable name.

LTR retrotransposon models were considered full-length if long terminal repeats were flanked by computationally identified target site duplications and had internal coding domains sufficient to categorize the model to a superfamily. Individual elements were classified into *Gypsy* and *Copia* superfamilies based on the location of the *reverse transcriptase* and *integrase* domains when these domains could be ascertained (Fig. 5 of the main text). When these protein domains were

not available, superfamily identity was based on BLAST (*227*) searches against known *Gypsy* and *Copia* LTR retrotransposons downloaded from existing transposable element databases (*224*, *228*). This process resulted in the identification of 4,326 full-length LTR retrotransposon models that could be grouped into 188 families with multiple copies in the genome. Of these, 111 families were identified as *Gypsy* elements and 77 as *Copia* elements.

Nucleotide sequence divergence between the LTRs of full-length models was used to estimate the time since insertion following previously described methods (*229*, *230*). Individual pairs of LTRs were aligned with CLUSTALW v 1.83 (*231*) and sequence divergence was estimated using the baseml v 4.4b program of PAML (*105*). Assuming an intergenic nucleotide substitution rate that was roughly twice as slow in *Amborella* as observed in the grasses (*230*), a substitution rate of 6.5×10^{-9} per site per year was used to convert LTR sequence divergence values into the estimated years since insertion (Table S29; Fig. 5 of the main text). This rate was used to make the essentially qualitative point that the LTR retro insertions occurred long ago. If we had used the rate that had been estimated for grasses, the inferred insertion dates would be twice as old.This process was automated using the *date_ltr_divergence.pl* script, and a family network graph visualization of the distribution of ages was generated using the RepMiner program (http://repminer.sourceforge.net/).

Methylation patterns around LTR-retrotransposons were inferred from analysis of bisulfite sequence data (see Bisulfite sequencing section, above). On average, annotated LTR-retrotransposons showed high levels of cytosine methylation in CpG, CHG, and CHH contexts relative to surrounding regions. However, levels of background and LTR-retrotransposon body methylation varied among families as a function of median transposon age. For example, the Copia and Gypsy families with the youngest median estimated insertion ages - *Dekena* (8.3 million years; 670 elements) and *Mugnar* (4.4 my; 502 elements), respectively - show clear increases in all three forms of methylation, whereas the older Gypsy families *Ronus* (11.7 my; 184 elements) and *Minar* (11.9 my; 316 elements) are in regions with high background methylation and elevated methylation is not observed within the element (Fig. S31). Elevated methylation is observed for an older Copia family, *Lukyla* (10.9 my; 101 elements) residing in regions with background levels mode similar to regions with genes (Fig. S31) and younger LTR retroelements.

To generate a database of representative full-length LTR retrotransposons, 5’ LTR sequences were clustered using all-by-all BLAST with an E-value threshold of 1×10^{-5} followed by clustering and classification in RepMiner (http://repminer.sourceforge.net/). Exemplar sequences were identified by affinity propagation based clustering (*232*) such that each individual family was represented by at least a single full-length model. These exemplars were screened against other databases including *Amborella* TE identified using *ab-initio* approaches described below to remove contaminating and potentially misidentified sequences. The resulting curated database of 1,137 full-length exemplar LTR retrotransposons was comprised of 668 *Gypsy* elements and 469 *Copia* elements.

Structural Annotation of non-LTR Retrotransposons – The structural annotation of LINEs (Long INterspersed Elements) used the MGEScan-nonLTR (*233*) annotation program. This pHMM based annotation pipeline identified 324 putative full-length LINEs. These sequences were assigned to superclass by a nearest-neighbor based BLAST search against published TE

databases (*228*, *234*) allowing 309 LINEs to be assigned to the *L1* superfamily and 15 to the *RTE* superfamily (*71*). LINE sequences were clustered into families using an All-By-All BLASTn search using a clustering threshold of 70% nucleotide identity over at least 70% of the length of the model sequences. This classification process used the *batch_pairwise_global_align_by_list.pl* program of the RepMiner package (http://repminer.sourceforge.net). Exemplar sequences for each family were identified by affinity propagation based clustering (*232*) such that each individual family was represented by at least a single full-length model. The resulting exemplar database 218 LINEs is comprised of 10 *RTE* LINE sequences and 208 *L1* LINE sequences.

Structural Annotation of DNA Transposons – The consensus sequences of the conserved DDE/D conserved transposase domains of each DNA transposon superfamily were obtained from Dr. Yao-Wu Yuan (*235*). These consensus sequences were used as query to search the *Amborella* genome assembly by tBLASTn on TARGeT (*236*) with an E-value cutoff of 0.1. Flanking DNA sequences within 10 kb upstream and downstream of the matched region were retrieved, and the ends of putative elements were determine by aligning two closely related elements with their 20-kb flanking sequences using NCBI-BLAST 2 Sequences (*237*). For elements in which the TIRs and TSDS could not be determined, 1 kb of sequences flanking the TARGeT matched region were retrieved to serve as a representative element.

As was observed with the retrotransposons, the DNA transposons (class II elements) were highly degenerate. Further, there was no evidence of recent transposon activity. Unlike transposon populations characterized in other genomes, all DNA transposons in the *Amborella* genome exhibited substantial divergence. For example a comparison of phylograms for PIF /Harbinger elements in the rice and *Amborella* genomes illustrates the qualitative difference in the diversification of annotated elements in these two genomes (Fig. S32). Whereas the rice genome harbors identical and near-identical elements, all branching events in the tree for *Amborella* elements are ancient. Moreover, a greater breadth of sequence diversity among annotated elements was found for all DNA transposon families observed in the *Amborella* genome. Like the LTR retrotransposons, DNA elements have accumulated in the genome over millions of years.

For identification of putative Helitrons, consensus protein sequences of Helicase domains from a database of classified Angiosperm helicases obtained from RepBase (*228*) were used as TARGeT query sequences against the *Amborella* genome assembly using the tBLASTn algorithm with a E-value threshold of 0.1. Flanking DNA Sequences were manually searched for Helitron diagnostic terminal motifs including 5’ CT and 3’ CATG, and small foldback structures in the 3’ end. No intact Helitrons were discovered using this methodology. Sequences flanking 2 Kb of the matching domains were thus taken as putative representative Helitrons, and were further curated by trimming non-helitron sequences by comparing these exemplars to other representative transposable elements in the Amborella genome.

Consensus sequences of small non-autonomous elements were generated using MITE-Hunter (*238*) and were used to search the genome with an E-value cutoff of 0.1 and all other default parameters suggested by the authors. The hits matching terminal inverted repeat (TIRs) between 150 bp and 1.5 kb and were filtered from the set of putative MITEs. The number and sizes of hAT, PIF, Mutator, and undefined MITE annotations is listed in Table S30.

Repeat Based Annotation of Transposable Elements –

RepeatScout consensus sequences - The RepeatScout program (*239*) was used to identify high-copy-number repeats in the 43,234 contigs that comprise the *Amborella* genome assembly (AmTr v1.0). A RepeatScout frequency index with an L-mer length of 16 nucleotide bases was first generated using the contigs as input. A RepeatScout analysis with default options used this frequency index and the genomic contig sequences to build 15,969 high-copy-number consensus sequences. The predicted consensus sequences that were less than 50 bp long and more than 50% low-complexity as identified by nseg (*240*) were removed to yield a size and complexity filtered database of 14,570 consensus sequences.

Genome Annotation of RepeatScout Results - The *batch_mask.pl* script of the DAWGPAWS (*21*) program was used to annotate the 5,745 *Amborella* scaffolds for the occurrence of the RepeatScout identified consensus sequences. This search used the WU-BLAST engine (http://blast.wustl.edu) with default RepeatMasker (http://www.repeatmasker.org) parameters. The GFF3 files resulting from this analysis were loaded to a custom database to count the occurrence of the RepeatScout matches. This RepeatMasker search identified 425,389,334 bases (~64%) in the *Amborella* genome scaffolds that match the RepeatScout database or were otherwise identified as low-complexity sequences. The count analysis revealed that 9,896 of the RepeatScout sequences occur more than 10 times in the genome scaffolds, and 3,570 RepeatScout sequences occur more than 100 times.

Classification of RepeatScout Consensus Sequences - For similarity based classification of RepeatScout consensus sequences, NCBI BLASTx and tBLASTx (*227*) were used to query the RepeatScout repeat sequences against multiple databases of known transposable elements that included RepBase amino acid sequences v. 16.03 (*228*), TIGR Plant Repeat Database v. 20060315 (*241*), and mips-REdat v 4.3 (*234*). All searches used an E-value threshold of $1x10^{-5}$ and the top 250 alignments were returned. These BLAST results were loaded into a custom designed MS Access database, and were hand curated to assign RepeatScout sequences to a superfamily (*71*) when possible matches were identified. The curation of the BLAST results generally used the collection of top hits from each database to assign sequences to a superfamily while also considering the spatial distribution of all matches across the RepeatScout consensus model. This process allowed for the classification of 1,745 sequences to a TE superfamily. Within the retrotransposon (Class I) elements, these similarity classified sequences included both *Gypsy* and *Copia* LTR retrotransposons as well as *L1* and *RTE* LINES. Multiple copies of Endogenous Pararetrovirus sequences were also identified using this approach. Within the DNA Transposon (Class II) elements, the similarity-identified superfamilies include *hAT*, *Mutator*, and *PIF-Harbinger* TIR elements as well as *Helitrons*. No SINE retrotransposons were identified using this approach, and TIR DNA transposons belonging to the *CACTA*, *Tc1-Mariner*, and *P element* superfamilies were not found despite their occurrence other plant lineages (*71*, *235*).

A tRNAScan-SE Cove analysis (*125*) of the filtered RepeatScout sequences was conducted in order to structurally identify putative tRNA derived SINES (*68*, *242*). This search identified tRNA signatures in 15 consensus sequences, with three identified as pseudo-tRNAs. Two of these pseudo-tRNA matches were very short (63 bp and 64 bp) while the third is 624 bp long and occurred only 26 times in the genome. In total these three putative tRNA derived SINEs have

108 occurrences in the *Amborella* genome assembly, and at most comprise approximately 21 kB of the *Amborella* genome.

Annotation of Genomic Transposable Element Composition - The structurally identified TE exemplars and the classified RepeatScout annotated transposable elements were combined to generate a comprehensive database of *Amborella* TEs. Combining these repeat-based and structurally identified TE models ensured that all high-copy number TEs are included in the database regardless of the structural completeness of the individual insertions while at the same time representing any low copy-number TEs that have at least a single insertion that could be structurally identified. An all-by-all BLAST analysis of the combined database was manually curated for cross-contamination among predicted transposable elements. The resulting database of 2,845 *Amborella* TE sequences classified to superfamily was used to annotate the genomic composition of TEs in the genome (Table S31).

The RepeatMasker program (http://www.repeatmasker.org) with the WU-BLAST (http://blast.wustl.edu) engine was used to search the *Amborella* genome scaffolds with default parameters. The resulting RepeatMasker output was converted to standard GFF3 format using the *cnv_repmask2gff.pl* program of DAWGPAWS (*21*) and the resulting file was loaded to a GenCart MySQL database (http://sourceforge.net/projects/gencart/) using the *load_gff2gencartdb.pl* script. Spatially overlapping alignments were identified and trimmed from the RepeatMasker result database using the *gencart_analysis_remove_overlaps.pl* program. The lengths of the trimmed alignments were then summed for each superfamily, ignoring all matches that were less than 100 bases in length.

12: Small RNA Annotation and Analysis

Small RNA-seq data processing – Small RNA-seq from three separate libraries (two from leaves, one from female floral tissues) was performed. For the 'ATR1' leaf library, processed data with adapter sequences already removed were obtained from GEO accession GSM712477. For the other two libraries, the 3' adapter sequences were trimmed using 'trim_illumina_sRNA_fastq.pl' version 0.3 (http://axtell-lab-psu.weebly.com/tools.html) with settings -e 1 -q Y -a TGGAATTC. The resulting trimmed reads ranged between 15-28 nts in length. The small RNAs were then aligned to a reference genome containing the version 1.0 nuclear genome assembly as well as the mitochondrial and plastid genome assemblies using Bowtie 0.12.7 (*243*) with parameters -f -v1 --best --strata -k 50 -S. The results were processed using 'Prep_bam.pl' version 0.1.1 from the ShortStack package (*244*) to produce a sorted and indexed BAM formatted small RNA alignment file containing NH:i: tags, indicating the number of alignments for each read. Table S32 lists the numbers of aligned reads, total alignments, and NCBI GEO accession numbers for these data sets.

***de novo* annotation of small RNA genes –** The three separate small RNA-seq BAM alignments were merged using the SAMtools (*245*) 'merge' command. The merged data were then analyzed with ShortStack version 0.2.1 (*244*), with default settings except --mindepth 60. A file of inverted repeats found by einverted (*246*) analysis of the genome assembly with a maximum repeat size of 10,000 nts was input into the ShortStack run. This analysis resulted in the annotation of 71,722 loci corresponding to areas of significant small RNA accumulation (Table S33). Table S34 shows the classification of these loci with respect to the predominant small RNA size ("DicerCall") and hairpin association. Note that Table S34 incorporates the *MIRNA* annotation modifications detailed below in the *MIRNA* annotation methods section.

***MIRNA* annotation methods** – The initial ShortStack annotations were filtered to ensure high-quality annotation of *Amborella MIRNA* loci, as well as to ensure that homologs of miRNA families known in other species were not unduly discarded. The following filtering steps were performed:

1. Loci initially called as *MIRNA*s were filtered based on the presence of the mature miRNA sequence in at least two of the three libraries. Those that failed this step were demoted to 'HP' status (e.g. a non-*MIRNA* hairpin).
2. Remaining *MIRNA* loci whose DicerCall was "N" (indicating that < 85% of all mapped small RNAs were within 20-24 nt in size) were filtered to retain only those whose annotated mature miRNA sequence was between 20-24 nt in size. Those that failed were demoted to 'HP' status.
3. The subset of loci demoted to 'HP' in step 1 that had a DicerCall other than "N" were examined with respect to existing miRBase (*247*) 19 annotations. Loci with zero or one mismatch to any known mature miRNA sequences from another species were rescued and re-promoted to *MIRNA* status.
4. Loci initially annotated as 'HP' (e.g., a non-*MIRNA* hairpin) that had a DicerCall other than "N", and that possessed all *MIRNA*-like features except for the sequencing of the predicted miRNA* were examined for similarity to known mature miRNAs from miRBase 19. Loci with zero or one mismatch to any known mature miRNA sequences from another species were promoted to *MIRNA* status.

Finally, paralogs amoung novel *MIRNA* loci (80 loci) were identified based on similarity of mature miRNAs (≤3nt mismatches) and/or stem-loop sequences. Several novel *MIRNA* loci had ≥4 mismatches in mature miRNA region but showed high sequence similarity along the entire stem-loops. Hence these were also annotated as paralogus loci (miR8552, 8553 and miR8565 families). Thus 121 *MIRNA* loci, from 88 distinct families, were annotated as a result of this process (Table S35).

An unusually high number of 24 nt miRNAs in *Amborella* – Of the 121 *MIRNA* loci, 62 produced 23- or 24-nt mature miRNAs (Table S35). An additional four loci produced 20-22-nt mature miRNAs but had an overall small RNA production profile dominated by 23-24-nt RNAs (because of miRNA* sequences and variable processing) (Table S35). Thus, 66 *Amborella MIRNA* loci have substantial production of 23-24-nt RNAs. All but two of these loci were members of *Amborella*-specific families, so that most (64 out of 75) of the *Amborella*-specific *MIRNA* loci were dominated by 23-24-nt RNAs, which in most cases were also the mature miRNAs themselves. 23-24-nt RNAs in plants are typically associated with heterochromatic siRNAs, not miRNAs. Some examples of such 'long' miRNAs have been reported in other plants (*248*, *249*). However, the high prevalence of 23-24-nt miRNAs observed in *Amborella* has not been reported in any other plant species to date. For example, only 11 of the 338 (3.3%) mature miRNA sequences of *Arabidopsis* in miRBase (version 19) are 23-24 nts in length, compared to the 51.2% we observed in *Amborella*. Like other flowering plants, *Amborella* contains a very high number of 23-24-nt siRNA loci (Table S34), raising the question of whether our *MIRNA* annotation methodology resulted in a large number of mis-annotations. To test this, we extracted three *Arabidopsis thaliana* small RNA libraries (two from leaves, one from inflorescences; data were randomly sampled from NCBI GEO accessions GSM738727 and GSM738731) of identical sizes as our input *Amborella* data sets. The data were then merged and analyzed with ShortStack using the same settings as used in the *Amborella* analysis. This resulted in the annotation of 118 *MIRNA* loci, only four of which (3.4%) produced mature miRNAs 23 or 24 nts in length. This result demonstrates that the high number of 23-24-nt miRNAs observed in *Amborella* is not the result of a pervasive methodological problem.

***Amborella* miRNA targets** – A non-redundant set of 104 mature *Amborella* miRNAs was used for identifying targets in the EVM-annotated *Amborella* coding sequences (version 27 third release). TargetFinder 1.6 (*250*) (http://carringtonlab.org/resources/targetfinder) was used to predict targets using an alignment score cutoff of 3. This resulted in the prediction of 60 unique targets for 19 of the 88 *Amborella MIRNA* families (Table S36). Best *Arabidopsis* BLASTp hits for the predicted *Amborella* targets were determined (using a maximum E-value of 1e-5). The TAIR10 annotations of the *Arabidopsis* best hits were utilized for identification of putative target gene function. Many of these initial predictions were later experimentally validated (see below). Details of the target predictions that were not validated are provided in Table S37.

Degradome (also known as PARE: Parallel analysis of RNA ends) sequencing identifies the 5' ends of uncapped mRNAs and can be used to experimentally determine miRNA targets in plants (*251*, *252*). To assist in identification of *Amborella* miRNA targets, two degradome libraries (one from leaves and one from female flowers) were produced, using the method described by German *et al.* (*253*). After removal of adapter sequences, the reads were mapped to the *Amborella* coding sequences (EVM annotation, version 27 third release) using Bowtie 0.12.7 (parameters: -f --norc -v 2 -a --best –strata). The GEO accession, processed read counts, and

mapping percentages are summarized in Table S38. The mappings were filtered to ensure that the first six nucleotides of the reads aligned perfectly to the coding sequences. Filtered mappings were then analyzed via the Cleaveland version 3 pipeline (http://axtell-lab-psu.weebly.com/cleaveland.html) to identify sliced miRNA targets at a p-value cutoff of 0.05. A total of 7,102 TargetFinder1.6 alignments (alignment score cutoff 7) was provided as an exhaustive source of possible miRNA-target alignments for Cleaveland analysis. A total of 61 targets for 29 *MIRNA* families was identified based on degradome evidence (Table S36), and 28 of these targets overlapped with predictions made in TargetFinder with an alignment score cutoff of 3 (Fig. S33). Details of the miRNA targets verified by degradome data are provided in Table S39.

Both target prediction based on alignments and validation based on degradome data were biased towards conserved *MIRNA* families. Conserved *MIRNA* families generally had targets that are clear homologs of their known targets from other species, underscoring the point that miRNA-target relationships for conserved *MIRNA*s are invariant in plants over long timescales. In contrast, most of the *Amborella*-specific miRNA families had few or no easily predicted or validated targets. This suggests that, as a class, the *Amborella*-specific miRNAs may have functions other than the canonical *trans* targeting of highly complementary mRNA targets. This dichotomy in target identification between conserved and non-conserved miRNA families is typical of what has been observed in other plant species (*250*, *254*).

Genomic features of *Amborella* small RNA genes – Both in terms of gene number and small RNA abundance, *MIRNA*s are a minor component of the *Amborella* small RNA inventory (Table S34). Small RNA clusters formed by non-Dicer processes were one of the major classes, accounting for 21.5% of the loci and 79.2% of the total small RNA abundance. We defined non-Dicer small RNA genes as those where less than 85% of the mapped small RNAs were within the 20-24-nt size range. The overall distribution of the fraction of reads between 20-24 nt in length was bimodal, and a cutoff of 85% corresponded to a point that ensured a robust and relatively conservative separation between the two peaks (Fig. S34). We suspect that most of the non-Dicer small RNA loci are breakdown products from abundant cellular RNAs (e.g. rRNA, tRNA, etc.), although we do not discount the possibility that some could represent other classes of functional molecules.

We next compared the locations of small RNA loci to other genome-wide features. We considered four categories of annotations. Exons, introns, and intergenic regions are mutually exclusive features. The fourth category, transposable element-related sequences (TE-related), can overlap with any of the other three. Both the absolute numbers of overlapping loci (Table S40) and the total of non-redundant overlapping nucleotides (Table S41) were calculated for each pairwise comparison of feature types. Enrichment or depletion was then calculated based on the ratio of the observed to the expected number of overlapping nucleotides (Fig. S35). The expected number of overlapping nucleotides for any pairwise comparison is given by the formula

$$E = (x/g) * (y/g) * g$$

where E is the expected number of overlapping nucleotides under the null hypothesis of random location, x is the total number of non-redundant nucleotides of feature type 1, y is the total number of non-redundant nucleotides of feature type 2, and g is the total genome size. Most

types of small RNA genes were depleted for overlaps with exons, especially the loci dominated by 23-24-nt RNAs. The 23-24-nt loci, especially the siRNAs, were enriched for overlaps with TE-related sequences (Fig. S35), and most overlapped with one or more of the TE-related loci (Table S40). The canonically sized 20-22-nt *MIRNA*s had a pattern typical of *MIRNA*s from other plants, with a strong depletion of overlaps with TE-related loci and introns, and a tendency to overlap intergenic regions (Fig. S35). In stark contrast, the 23-24-nt *MIRNA*s had a slight depletion in intergenic regions, and a very strong enrichment within introns. Indeed, nearly half (29 of 66) of the 23-24-nt *MIRNA* loci overlapped an intron (Table S40). The enrichment for intron-association also extended to the 23-24-nt hpRNA class, indicating that many non-*MIRNA* hairpins that generate predominantly 23-24-nt RNAs also originate from introns in *Amborella*. This strong bias toward intronic locations for 23-24-nt *MIRNA* and hpRNA loci has not been reported in any other plant species.

Identification of tasiRNA-like phasiRNA loci (PHAS) in *Amborella* – Phasing analysis was performed as described previously (*255*). All loci with a phasing score ≥12.5 in any one library are considered as PHAS loci and were subjected to manual inspection. In total, 19 PHAS loci were identified, comprising 10 annotated, protein-coding genes (Fig. S36) and nine intergenic regions (IGRs) (Fig. S37). Combining small RNA and PARE data with target predictions, five miRNAs were identified as triggers for these PHAS loci (all validated in the PARE data), but no triggers were found for 11 of the loci. These trigger miRNAs included three miRNAs that were missed during the initial automated miRNA-discovery analysis (a paralog of the miR8558 locus called miR8558b, miR2950, and a novel family miR8615; Table S42). miR8558b was found to trigger phased RNA synthesis from several NBS-LRR disease-resistance genes (Table S43), reminiscent of the activity of the sequence-diverse miR482/2118 superfamily of miRNAs (*72-74*). Indeed, miR8558a and miR8558b are likely part of the miR428/2118 superfamily as they can be aligned with 3-4 mismatches to some previous miR428/2118 annotations (Fig. S38). Interestingly, the MIR8558a and MIR8558b loci are clustered together on scaffold 17. Two other small RNA loci (Cluster_15097 and Cluster_15098; Table S33) are found between these two *MIRNA* loci, and both are dominated by 22-nt small RNAs with high sequence identity to the mature miR8558a and miR8558b miRNAs (Fig. S38). Although both of these clusters lack discernable hairpin structures, all four produce similar 22-nt RNAs with sequence similarity to the miR482/2118 superfamily. We therefore conclude that the miR482/2118 superfamily and its ability to spawn phased siRNAs from NBS-LRR targets was present in the MRCA of angiosperms.

Other PHAS transcripts encode an LRR-kinase (TAS3), two RNA-dependent RNA polymerase 6 proteins (RDR6, located nearby on the same scaffold), an auxin signaling F-box protein and one hypothetical protein (Table S43). The triggers for most of the nine IGR PHAS loci were not identified, making it difficult to define their relationship to known, *Arabidopsis* TAS loci. But because miR390 is the conserved trigger for TAS3 in *Arabidopsis* and other plants, the PHAS locus encoding an LRR-kinase, triggered by miR390, could be a TAS3 homolog in *Amborella*.

miRNA families of the ancestral angiosperm – We used small RNA expression evidence from a wide panel of plants (NCBI GEO accession: GSE28755), in combination with miRBase version 19 annotations, to track the conservation patterns of the 90 *Amborella* miRNA families(88 families from automated ShortStack discovery and 2 families from PHAS loci analysis). All 20-24-nt small RNA reads from the plant small RNA-seq data set that were

represented by at least two reads in a library were aligned to the mature miRNA sequences from the 94 *Amborella* miRNA families, allowing three mismatches and up to 2 nt of overhanging nucleotides at either the 5’ or 3’ end, or both. Alignments were performed using SeqMap (*256*) followed by output filtering and reformatting by custom-written Perl scripts. For each miRNA family investigated, the sum of abundances of all the identified variants in each library was tallied.

Nineteen of these families were robustly expressed in nearly all flowering plants as well as in most gymnosperms (Fig. S39, top section); all of these families had previously been annotated in other plant species, per miRBase version 19 (*247*). An additional eight families showed expression in some plant lineages other than *Amborella*, but with conspicuous absences from many species (Fig. S39, middle section). These include three miRNA families previously annotated in other plants (miR477, miR828, and miR2950) and five additional families annotated here for the first time. The simplest explanation for the patterns of expression of these families is that they were all present in the ancestral angiosperm (and gymnosperm lineages, with the possible exception of miR477), and then lost later in angiosperm diversification. However, because these data are expression-based, it is also possible that some of these apparent losses are instead due to lack of expression in the sampled tissues/conditions. Altogether, we conclude that at least 27 miRNA families were present in the ancestral angiosperm. The other 63 families have scant evidence of expression in other plants (Fig. S39, bottom section) and do not have putative homologs in miRBase, based on sequence similarity; we thus conclude these are *Amborella*-specific. The high number of apparently species-specific miRNAs in *Amborella* is typical for most plant species (*257*).

13: Population Genomic Analyses

Sampling – Preliminary population genetic variation and geographic structure were inferred from an analysis of 10 microsatellite loci (*258*) and were used to select one *Amborella trichopoda* individual from each of 12 natural populations representing the geographic and genetic diversity of *Amborella* on the mainland of the New Caledonia archipelago (see *78*).

These sites were sampled in Poncet *et al*.'s microsatellite analysis (*78*), and each individual [sample numbers AmTr_822_NCAM_013/Amieu (15); AmTr_823_NCAOC_010/Aoupinié C (13); AmTr_824_NCBA_02/Ba Houailou (11); AmTr_825_NCBO_010/Boregaou (13); AmTr_826_NCDOB_010/Dogny B (16); AmTr_827_NCMF_01/Mé Foméchawa (15); AmTr_828_NCMO_08/Mé Ori (12); AmTr_829_NCNAA_011/Nakada A (18); AmTr_830_NCPO_012/Ponandou (14); AmTr_831_NCPWB_011/Pwicaté B (18); AmTr_832_NCTC_012/Tchamba (9); AmTr_833_NCTOB_012/Tonine B (19)] was also included in their study (see *78* for location and voucher information). In addition, sequence data for samples from the University of California Santa Cruz (Santa Cruz) and the Bonn Botanical Garden (Bonn) were included in the analyses. The Santa Cruz sample is the individual from which the v. 1 genome is derived; however, the collection data for this individual do not precisely place it within a known population from New Caledonia. The Santa Cruz and Bonn samples were included to determine (1) if they derive from the same natural population, (2) from which population they were sampled, and (3) which populations are most genetically similar.

DNA extraction and sequencing – DNA was extracted from silica-dried leaves using a standard plant CTAB DNA extraction protocol, and quality was evaluated on an Agilent Bioanalyzer DNA1000 chip. Standard short-insert Illumina TruSeq libraries were prepared according to the manufacturer's recommended protocol with an average insert size of 225 bp. Library quality and insert size were validated with a preliminary multiplexed sequencing run using the Illumina MiSeq prior to full sequencing with 2 x 100 bp paired-end protocol on four lanes of an Illumina HiSeq2000.

Read Filtering – Illumina reads for the12 populations from New Caledonia, the Santa Cruz genotype, and the Bonn genotype were extensively filtered for artificial duplicates and short reads, and trimmed for Illumina sequencing adapter and poor-quality ends.

A custom python script was used to identify and filter artificial duplicate sequences resulting from PCR amplification from Illumina library preparation and sequencing reactions. The sequence prefixes (in this case the first 25 bases) of each paired read were concatenated. Read prefix sequences were used because they represent the high-quality portion of Illumina reads. If two or more paired-end sequences share identical concatenated prefixes, they are classified as artificial duplicate clusters. Only one of the paired-end sequences was retained in each cluster, and the rest were discarded as artificial duplicates. This process should remove nearly all duplicates; results are summarized in Table S44.

Reads remaining after removal of artificial duplicates were processed through an adapter removal program called Cutadapt (version 1.1) (*259*), which trimmed the remains of Illumina adapter sequences using the following parameters: --error-rate (maximum allowed error rate) 0.1; --times (try to remove adapters at most) 1; --overlap (minimum overlap length between the read

and the adapter) 5; --minimum-length (discard trimmed reads that are shorter than minimum-length) 0. After adapter removal, the Trimmomatic program (*118*) was used for trimming low-quality regions and removing short reads using the following parameters: HEADCROP (number of bases to remove from the start of the read) 0; LEADING (minimum quality required to keep a base at the start of a read) 3; TRAILING (minimum quality required to keep a base at the end of a read) 3; SLIDINGWINDOW: window size of 4 with a minimum average quality of 15; MINLEN (specifies the minimum length of reads to be kept) 75 for sequences with 101 bp as raw read length and 60 for sequences with 82 bp as raw read length. Only sequences with paired-end information were used in further analysis; the quantity of paired reads left after all the above described filtering and trimming steps is summarized for each population in Table S41.

Read Alignment – We aligned the set of paired-end reads against the *Amborella* reference genome v. 1.0 using BWA (v0.6.2) (*260*). Median depths of aligned reads for each population are summarized in Table S41. BWA makes use of the Burrows-Wheeler Transform (BWT) (*261*) to create a permanent index of the reference sequence, which may be re-used across alignment runs, and is able to achieve very good speed and memory efficiency. It uses the technique of *backward search* (*262*, *263*) with BWT and is able to effectively mimic the top-down traversal on the prefix trie (hereafter, trie) of the genome with a relatively small memory footprint (*264*) and to count the number of exact hits of a string in linear time regardless of the size of the genome. For inexact search, BWA samples from the implicit trie the distinct substrings that are less than k edit distance from the query read. Because exact repeats are collapsed on one path on the trie, it does not need to align the reads against each copy of the repeat.

To find the suffix array coordinates of good hits of each individual read, we used the following input arguments: -n 4 to set the maximum edit distance of the alignment equal to 4; -o 1 to set the maximum number of gap opens equal to 1; -l 32 to use a seed of length 32; and -k 2 to set the maximum edit distance in the seed part of the read equal to 2. When generating the alignments of paired-end reads, we additionally used the input argument -a 800 to set the maximum insert size for a read pair to be considered being mapped properly equal to 800 bp.

SNP calling and SNP Selection – For each of the individuals from New Caledonia, Santa Cruz (only RUN02 was used in population studies), and Bonn, SNPs were called relative to the reference using FreeBayes (version 0.9.6) (*265*), a Bayesian genetic variant detector designed to find small polymorphisms, with the following parameter settings: --pvar (reports sites if the probability that there is a polymorphism at the site is greater than specified) 0.75; --theta (expected mutation rate or pairwise nucleotide diversity among the population under analysis) 0.01; --ploidy (ploidy for the analysis) 2; --no-indels (ignore insertion and deletion alleles); --min-mapping-quality (exclude alignments from analysis if they have a mapping quality less than specified) 30; --min-base-quality (exclude alleles from analysis if their supporting base quality is less than specified) 20; --min-supporting-quality (in order to consider an alternative allele, at least one supporting alignment must have mapping quality 30, and one supporting allele must have base quality 20) 30,20; --min-alternate-fraction (require at least 0.5 fraction of observations supporting an alternative allele within a single individual in order to evaluate the position) 0.5; --min-alternate-count (require at least this count of observations supporting an alternative allele within a single individual in order to evaluate the position) run at 5, 10, and 15. The numbers of SNPs called at the three thresholds of alternative counts are 5,131,595 for min-alternate-count = 5; 3,415,227 for min-alternate-count = 10; and 1,903,437 for min-alternate-count = 15. We

selected min-alternate-count = 5 and used the set of over 5 million SNPs for subsequent analyses. For each individual, a modified 'reference' sequence was generated by comparing the reads of the sample with the *Amborella* v. 1.0 reference genome and 'correcting' the v. 1.0 sequence to the sample variants for each contig. In this way, a set of 'modified reference genomes' was created and saved as a set of fasta files on which all subsequent analyses were conducted.

Assessing Genetic Diversity – Individual alignments were split by scaffold, and multifasta scaffold files were generated using Perl scripts. Population genetic parameters, including the number of segregating sites (*S*), Watterson estimator (θ_w), nucleotide diversity (π), and Tajima's D statistics, were estimated using DnaSAM (*75*) (Table S45). These statistics were estimated for 100-kb windows and plotted across the lengths of the scaffolds (Fig. S40) using CIRCOS (*266*).

Inferring Changes in Effective Population Size through Time – We applied the PSMC model, which was originally applied to human and other mammalian genomes, to study the history of *Amborella* effective population size (N_e). PSMC infers the local time to the most recent common ancestor of the present-day genome on the basis of the local density of heterozygotes by use of a hidden Markov model in which the observation is a single diploid sequence (*267*). PSMC utilizes sequence reads as mapped to a reference genome to estimate historical fluctuations in N_e. For scaling N_e, PSMC requires input of an estimated per-year mutation rate (μ). Per-generation mutation rates have been shown to be generally related to genome size across a variety of organisms (*268*). Taking this into account, we interpolated the mutation rates for several plant species, including *Amborella*, using Lynch's (*268*) published rate/genome-size relationship (Table S46). To scale PSMC results to real time, we assumed 4 years per *Amborella* generation (*g*) and a per-generation mutation rate (μ) of 14.1×10^{-9}. PSMC was otherwise conducted using default parameters.

Mapping Geographic Structure – Population genetic structure was identified using STRUCTURE (*269*). The SNP vcf file was modified for use in STRUCTURE using an R script (*270*). For the analysis, a randomly selected subset of 12,000 SNPs was used, and different population numbers (K) were evaluated, with K ranging from 1 to 12 (Fig. S41A). A burn-in of 20,000 iterations and 100,000 iterations of the Markov Chain Monte Carlo (MCMC) were used. Modelling decisions included the use of an admixture model and correlated allele frequencies. In addition, for each set of parameters, the analysis was repeated 12 times with a new random subset of SNPs to evaluate the consistency on the probability of the data. Bar plots representing the results of repeated analysis are described (Fig. S41). Values of K were tested for significance following Evanno *et al*. (*271*). To verify that the results were consistent, regardless of the SNPs selected in the first set of iterations, the analysis was repeated with an independent random selection of 12,000 SNPs. Three significant clusters were identified when only the 12 population samples were included in the analysis (Fig. S42), however, when the cultivated individuals from Santa Cruz and Bonn were included to infer the genetic cluster(s) from which these two individuals originated, four significant clusters were identified (Fig. 6 of the main text).

General Summary of Microsatellite Results – The resequencing results largely agree with those obtained in the analysis of 10 microsatellite loci with more extensive population level sampling (*78*). Nearly the same four clusters were detected in both analyses; however individuals

sampled from Aoupinié were associated with the central cluster in microsatellite analyses rather than the northern cluster as shown in Fig. 6 (main text) and Fig. S42. This population is located on the boundary between the northern and central clusters, and the locality was inferred to be a refugium during the last glacial maximum during the Pleistocene. The difference may be explained by the sampling of only one individual for genome resequencing.

Supplementary Figures:

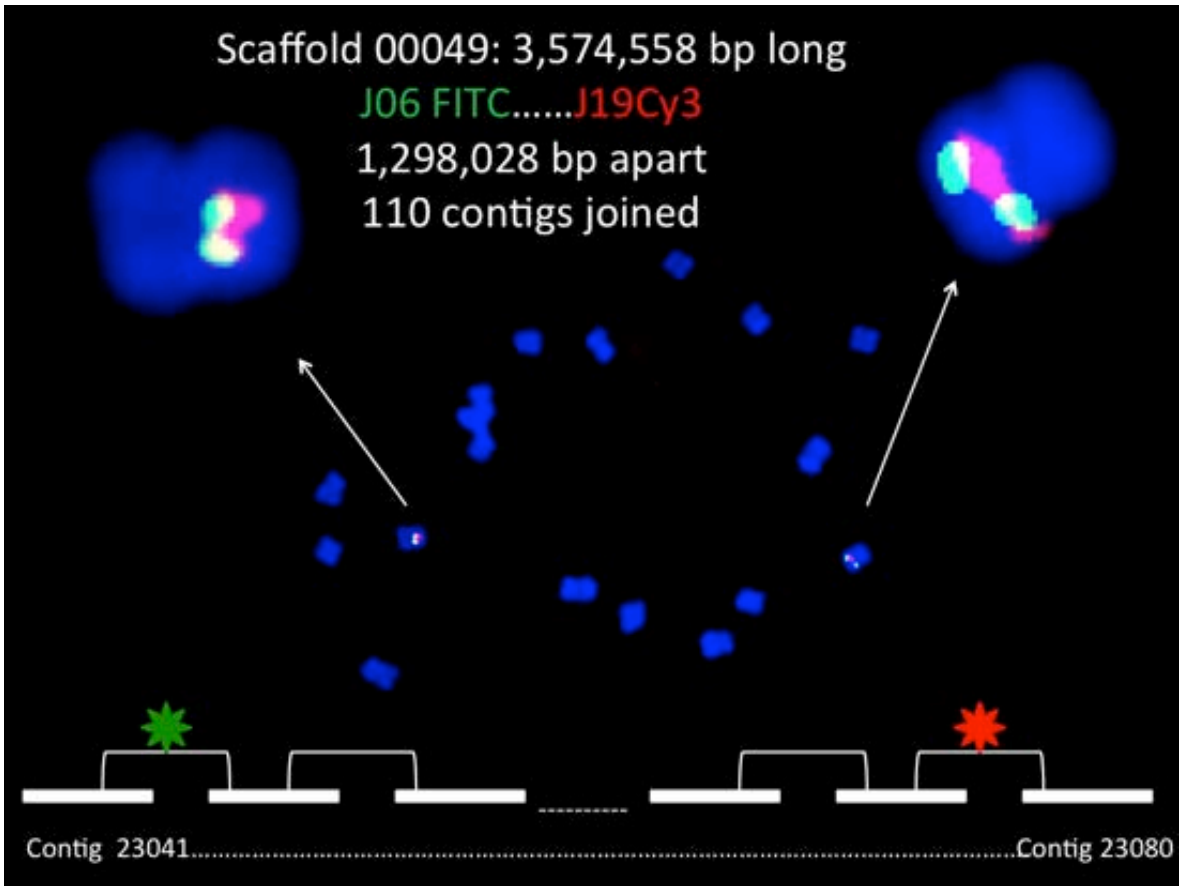
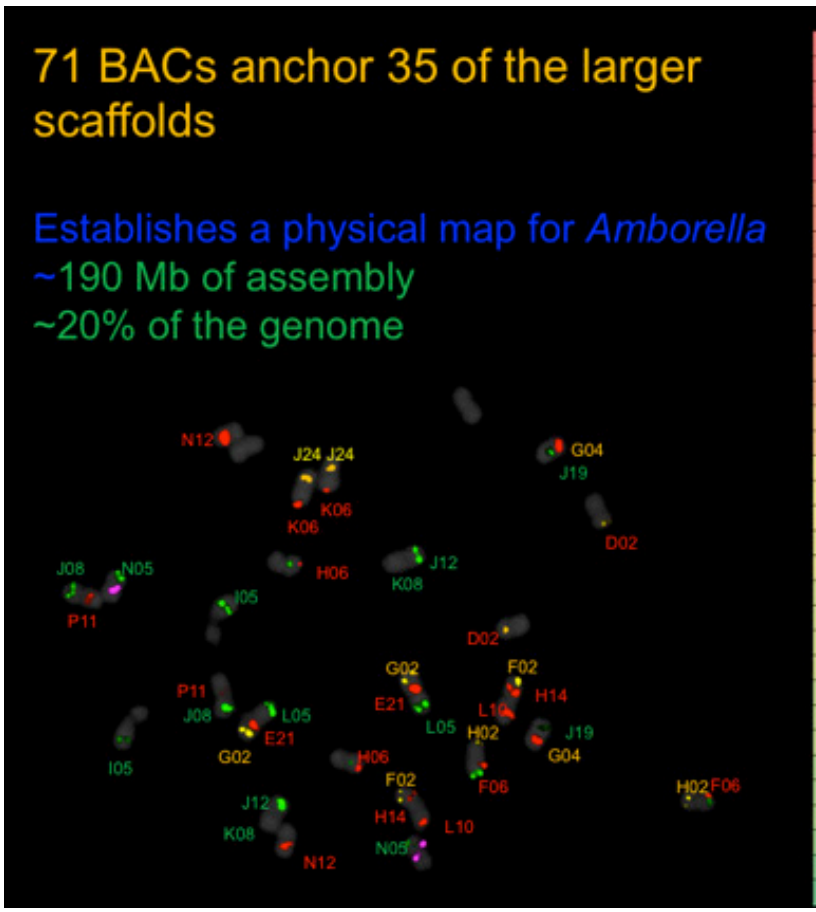


Figure S1. Assembly validation by FISH to verify co-localization of BAC-based probes that map to large scaffolds.



Chromosome	Scaffold	Length
1	15	7405751
1	20	6874971
1	21	6860876
1	25	6368207
1	93	2935351
1	130	1605649
2	2	11522362
2	4	10537363
2	32	5900637
2	45	5143656
2	50	4903570
2	57	4629110
2	81	3248084
3	48	5155698
3	68	3967838
4	10	9150321
4	38	5499001
5	56	4878105
5	109	2285430
5	136	1544621
6	13	8123900
6	66	3926670
7	9	9389330
7	39	5379737
7	72	3666727
7	148	1337138
8	12	8757819
8	23	6666511
8	80	3254339
9	29	6285685
9	44	5106096
9	55	4791638
10	74	3745557
11	40	5311347
12	131	1574043

Figure S2. FISH karyotype and the chromosomal assignment of 35 scaffolds in the genome assembly.

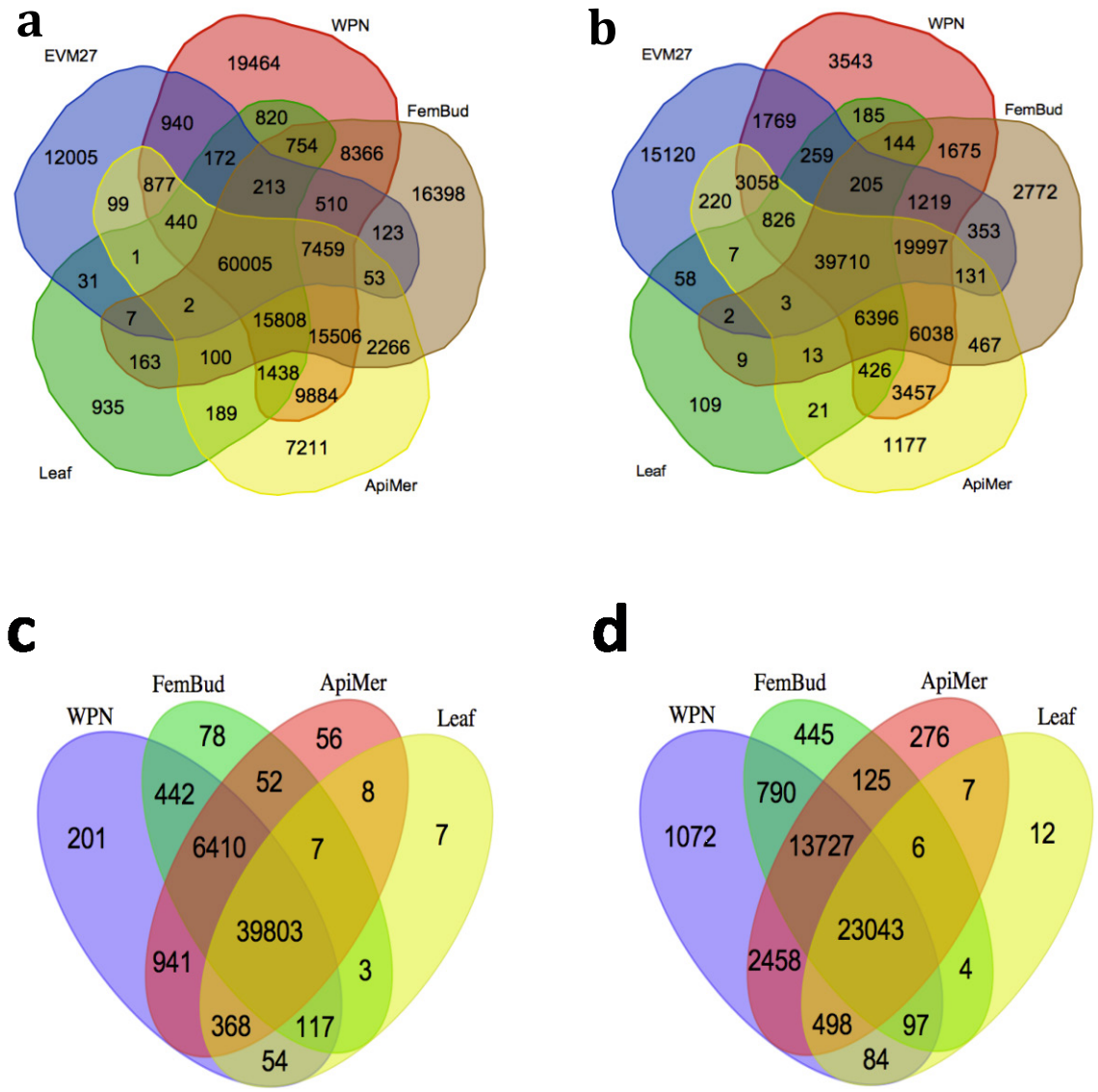


Figure S3. Comparison of splice junction usage among tissues. Splice junctions compared originated from the entire EVM 27 gene annotation set (a & b) or from isoforms of genes predicted to undergo alternative splicing (c & d). The Venn diagrams illustrate counts of junctions specific to one tissue, or shared between multiple tissues. (WPN – whole plant normalized; FemBud – female bud; ApiMer – apical meristem). (a) Splice junction with at least two reads support in each tissue; (b) splice junction with at least twenty reads support in each tissue; (c) splice site overlap with AS gene only and at least two reads support junction; (d) splice site overlap with AS gene only and at least 10 reads support junction.

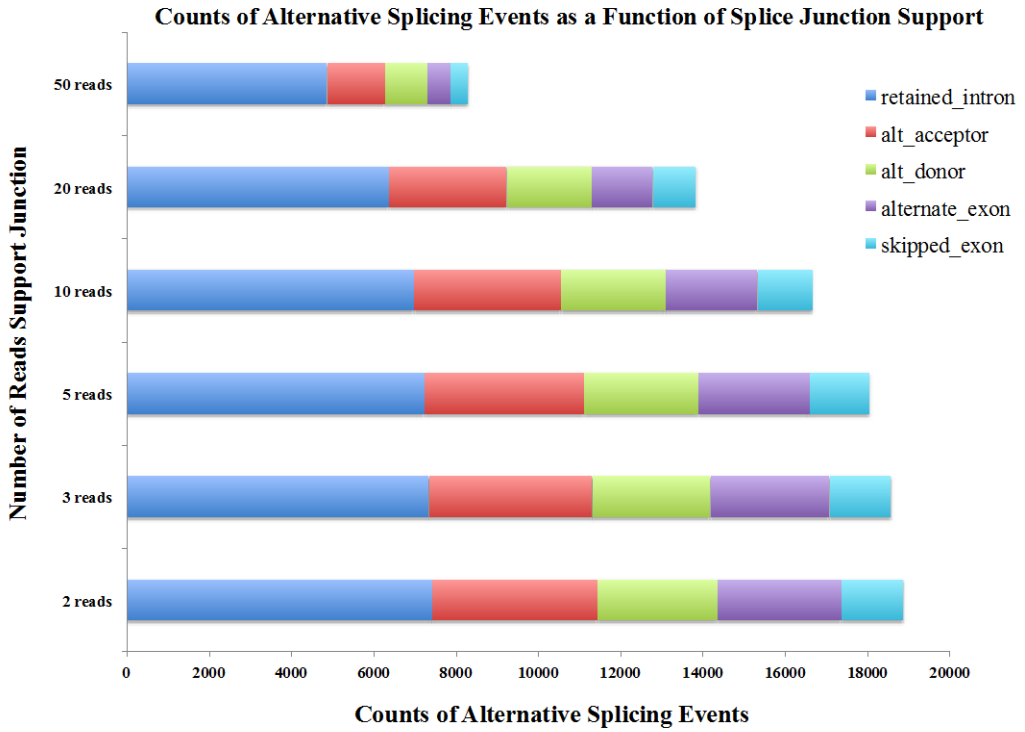


Figure S4. Counts of alternative splicing events as a function of splice junction support (2 reads, 3 reads, 5 reads, 10 reads and 20 reads).

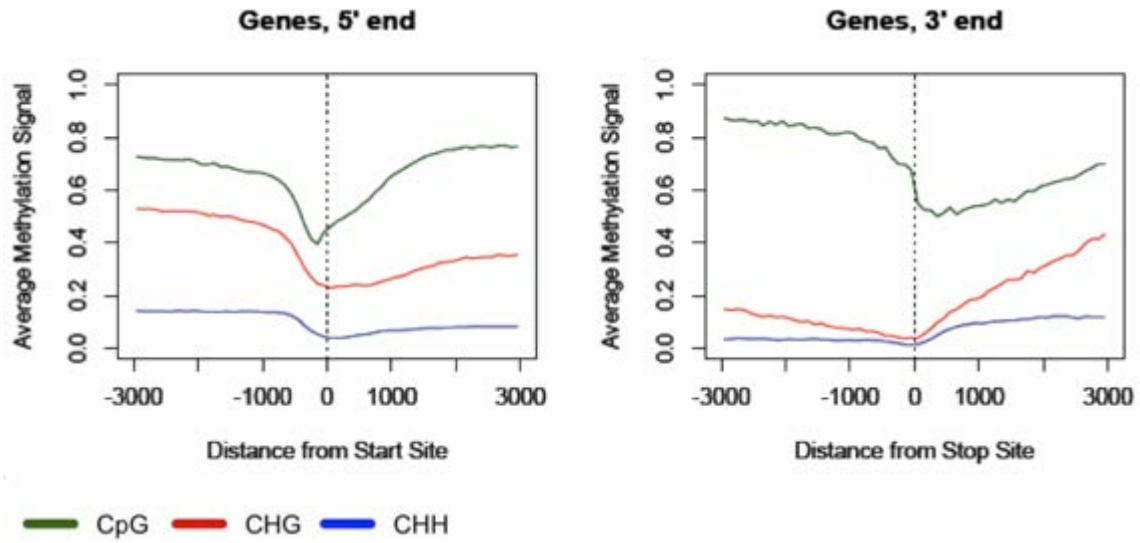


Figure S5. Gene body methylation signals relative to translation start (left) and stop sites (right) of predicted protein coding genes. *Amborella* genes exhibit reduced methylation just upstream from the translation start site and gene body methylation similar to the pattern seen in other characterized flowering plant methylomes.

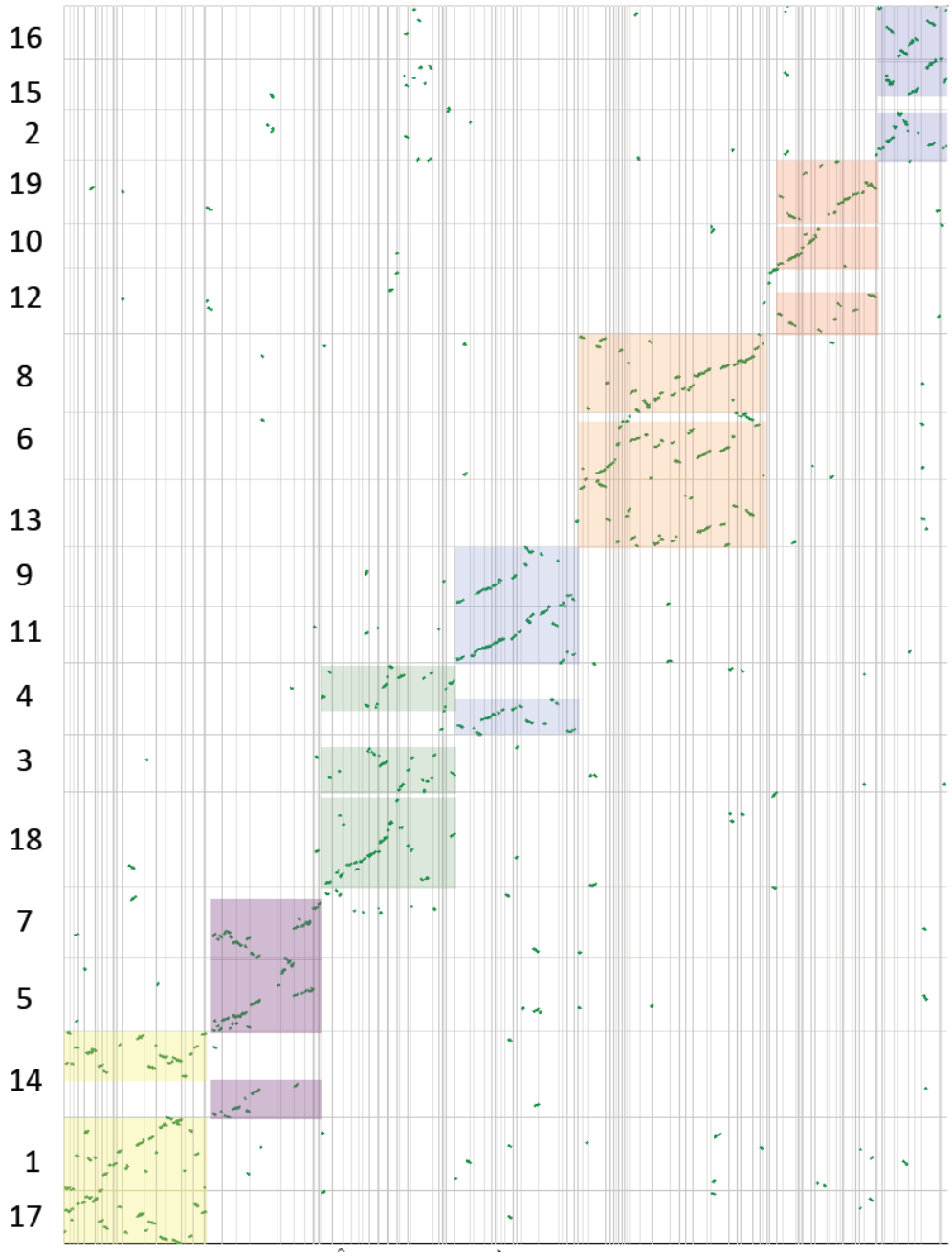


Figure S6. Syntenic path dotplot (*130*) of *Amborella* (x-axis) versus grape (y-axis). Note the 1:3 syntenic depth relationship between grape and *Amborella*.

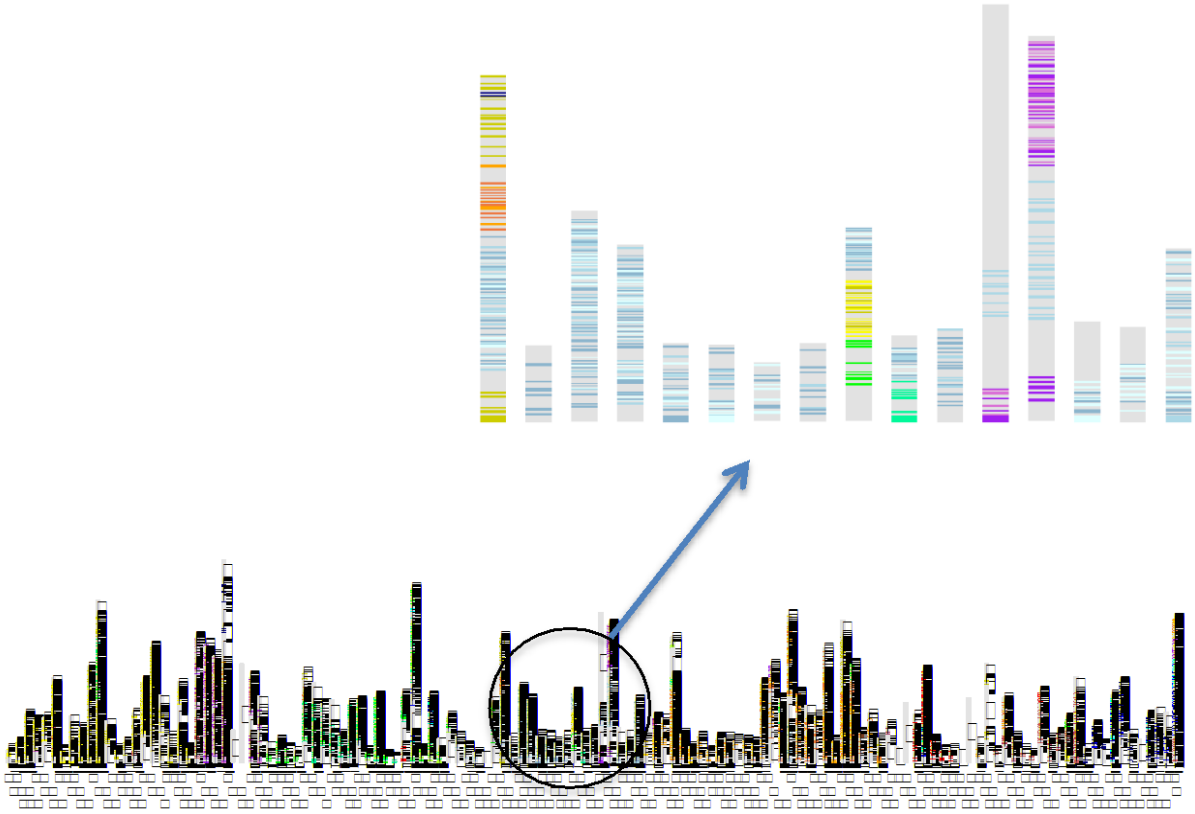


Figure S7. *Amborella* contigs showing positions of orthologs to grape triplicated genes. Three shades (dark, medium and light) of each of seven colors uniquely refer to three homeologous grape chromosomes, as in Fig. S11. Contigs ordered by SynMap as in Fig. S6. Height of each *Amborella* contig proportional to the number of genes on it, with a horizontal stripe at the position of each ortholog to a grape gene. Gray areas represent genes that have no ortholog in grape triplicated regions. Expanded detail (above) shows how genes colored three shades of the same color are interleaved on *Amborella* contigs, reflecting the fractionation pattern of the grape homeologs. There is no concentration of a single shade (reflecting a single grape chromosomes) on any *Amborella* scaffold, confirming that the gamma triplication occurred after the divergence of *Amborella* from the other angiosperms.

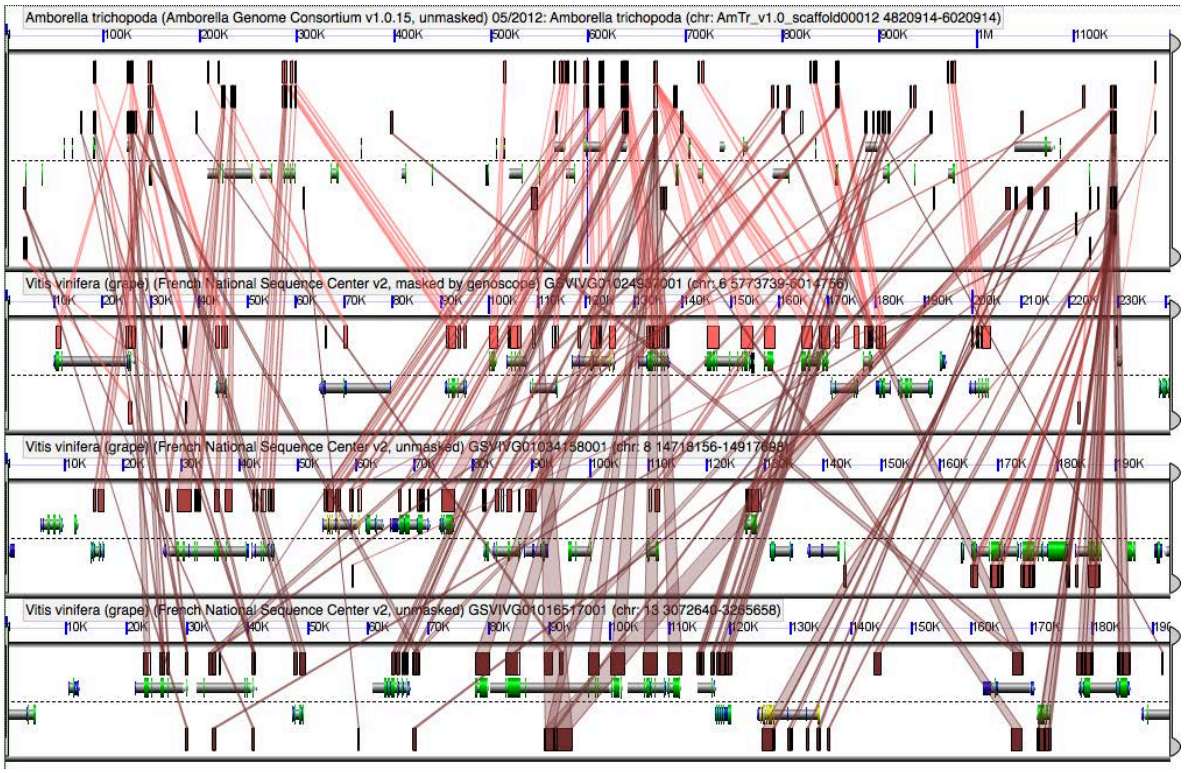


Figure S8. High-resolution analysis of three syntenic grape regions and a single *Amborella* region. Note that the gene content of the three grape regions is fractionated, and the *Amborella* region contains the entire gene content of the grape regions combined. Results may be regenerated: http://genomevolution.org/r/4e0q.

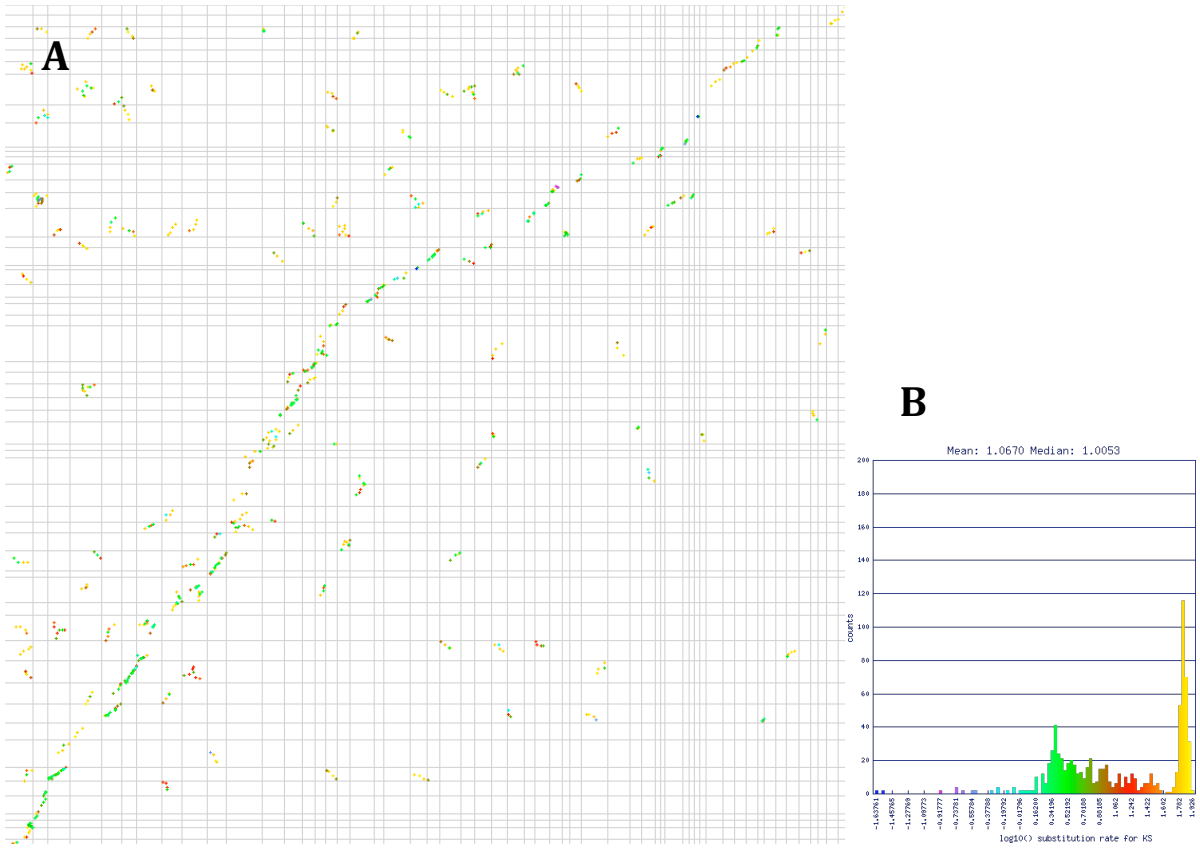
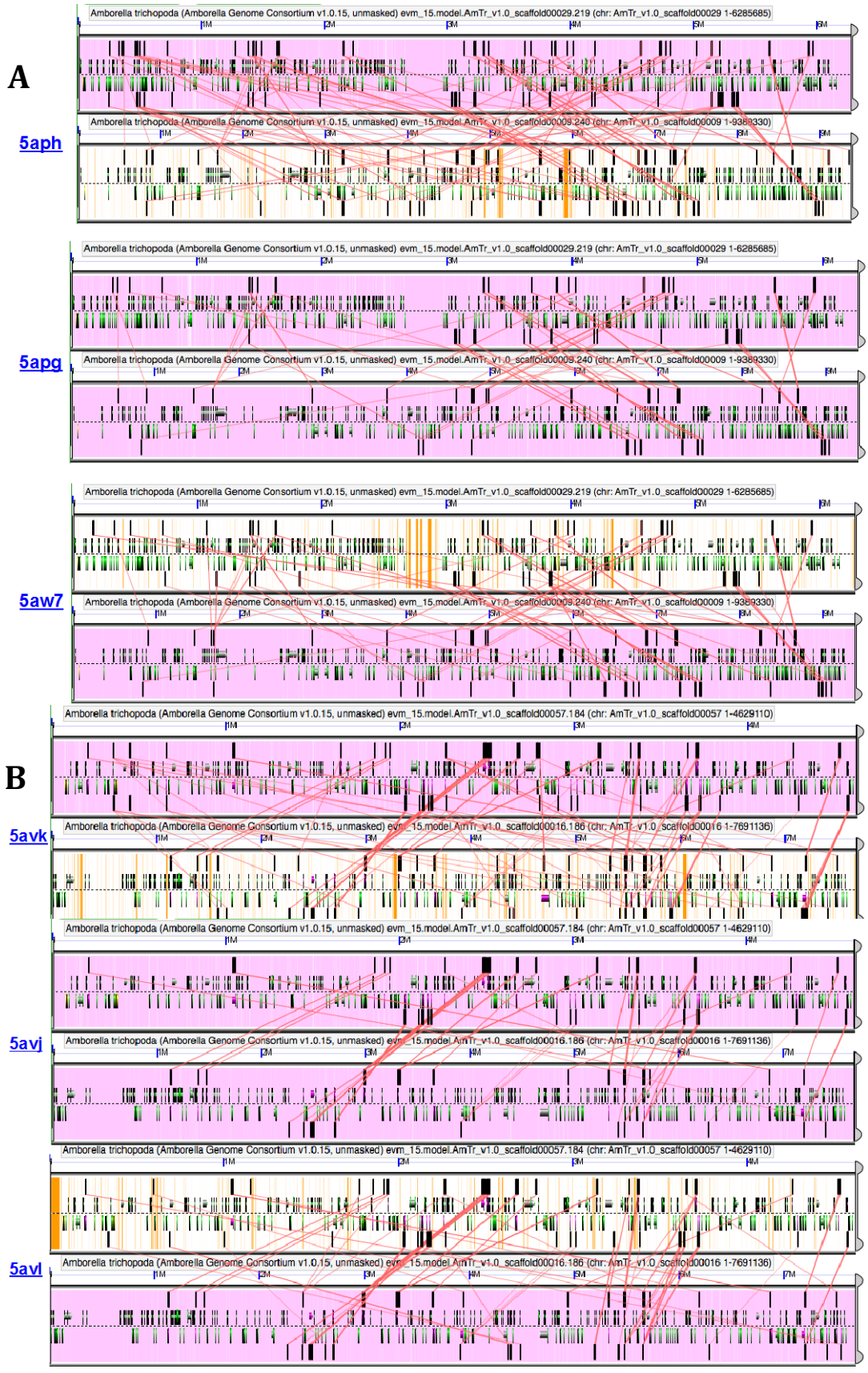
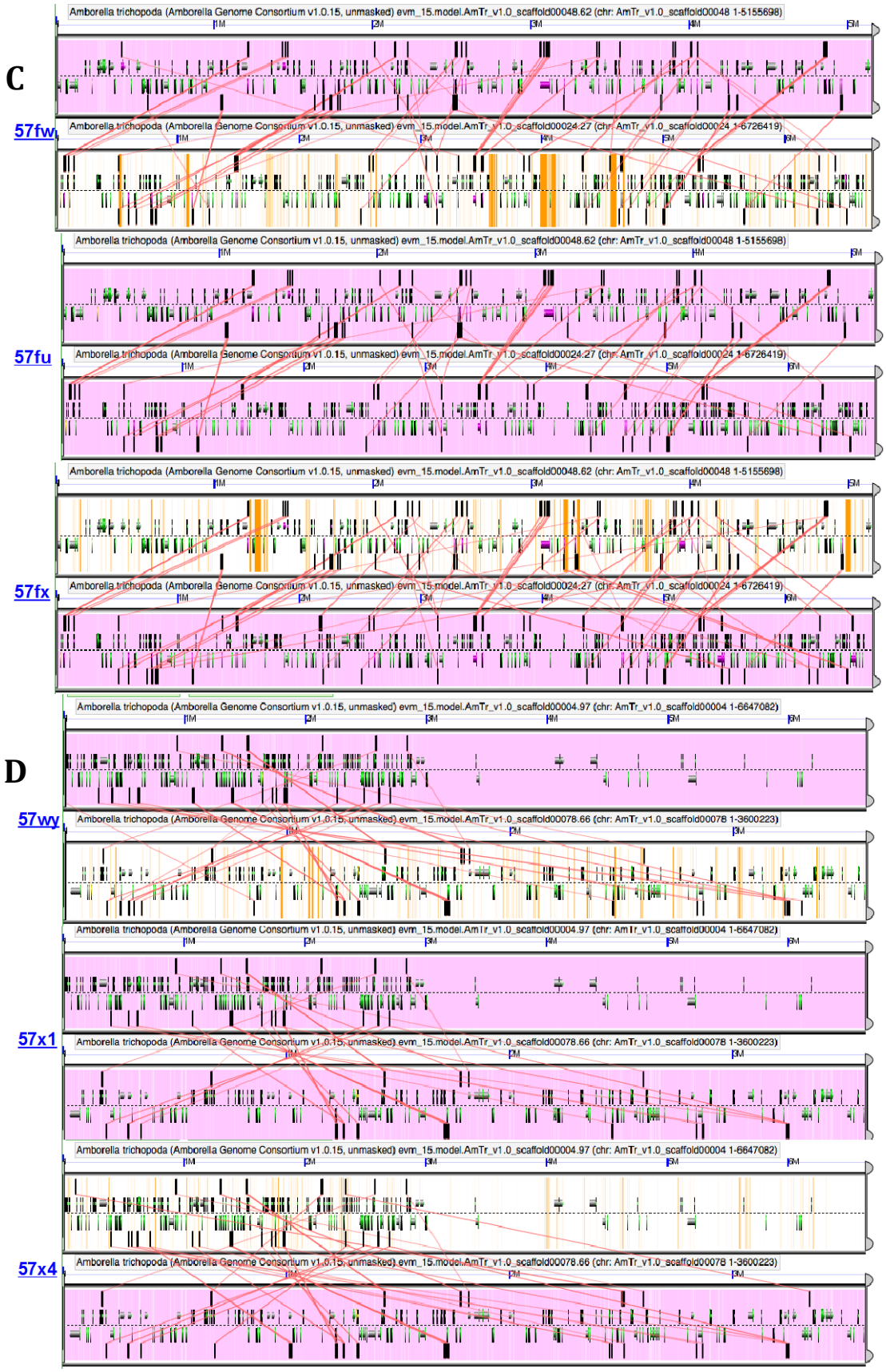


Figure S9. Synteny within the *Amborella* genome (A) Self-self syntenic dotplot of *Amborella* with contigs ordered and oriented by their syntenic path. Dots are colored by their synonymous mutation values. Green dots are all from the same age distributions. Red/orange/yellow dots are noise. (B) Histogram of K_s values.


A
5aph
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00029.219 (chr: AmTr_v1.0_scaffold00029 1-6285685)
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00009.240 (chr: AmTr_v1.0_scaffold00009 1-9389330)
5apg
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00029.219 (chr: AmTr_v1.0_scaffold00029 1-6285685)
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00009.240 (chr: AmTr_v1.0_scaffold00009 1-9389330)
5aw7
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00029.219 (chr: AmTr_v1.0_scaffold00029 1-6285685)
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00009.240 (chr: AmTr_v1.0_scaffold00009 1-9389330)
B
5avk
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00057.184 (chr: AmTr_v1.0_scaffold00057 1-4629110)
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00016.186 (chr: AmTr_v1.0_scaffold00016 1-7691136)
5avj
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00057.184 (chr: AmTr_v1.0_scaffold00057 1-4629110)
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00016.186 (chr: AmTr_v1.0_scaffold00016 1-7691136)
5avl
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00057.184 (chr: AmTr_v1.0_scaffold00057 1-4629110)
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00016.186 (chr: AmTr_v1.0_scaffold00016 1-7691136)


C
57fw
57fu
57fx
D
57wy
57x1
57x4
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00048.62 (chr: AmTr_v1.0_scaffold00048 1-5155698)
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00024.27 (chr: AmTr_v1.0_scaffold00024 1-6726419)
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00004.97 (chr: AmTr_v1.0_scaffold00004 1-6647082)
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00078.66 (chr: AmTr_v1.0_scaffold00078 1-3600223)

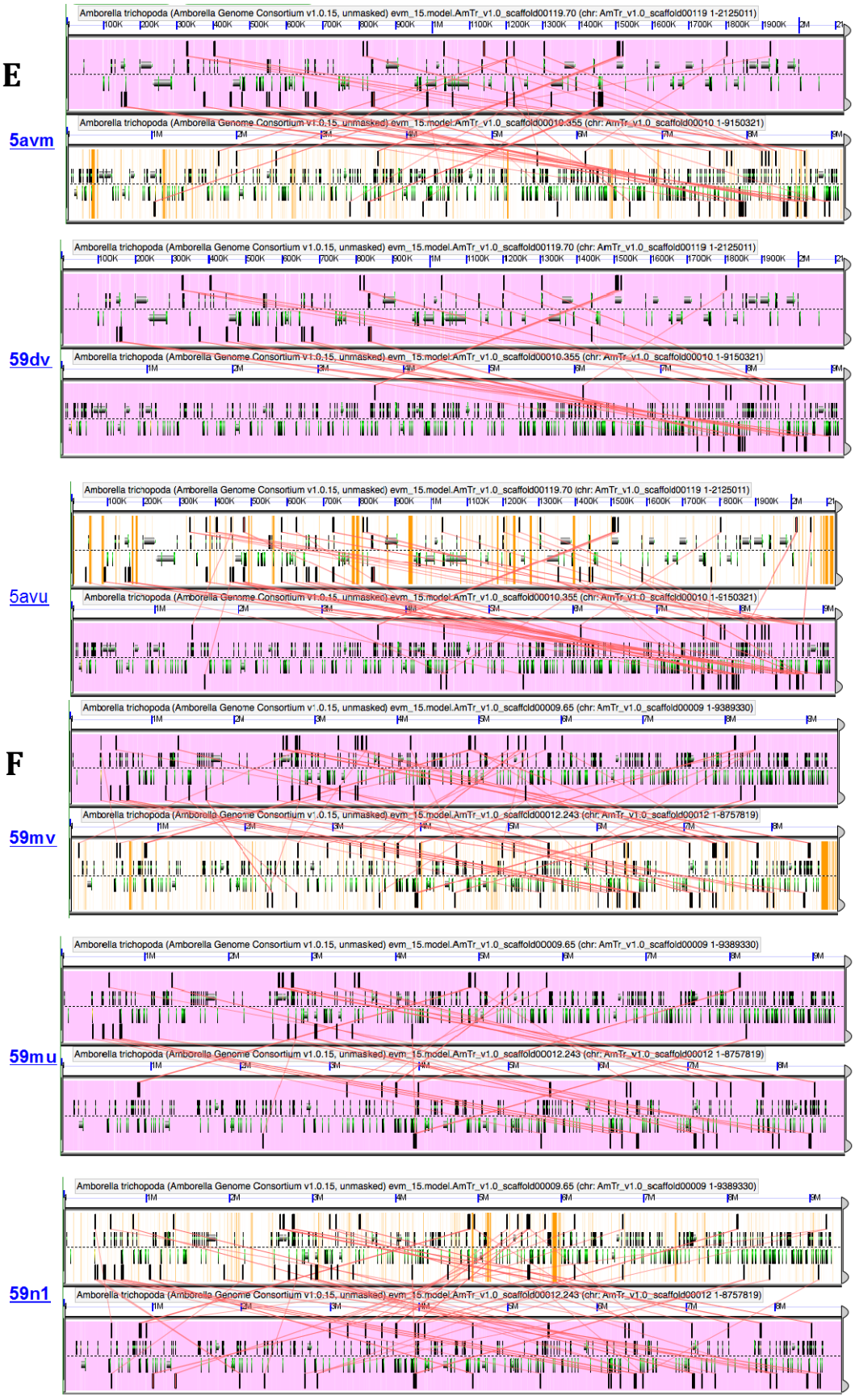

E
5avm
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00119.70 (chr: AmTr_v1.0_scaffold00119 1-2125011)
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00010.355 (chr: AmTr_v1.0_scaffold00010 1-9150321)
59dv
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00119.70 (chr: AmTr_v1.0_scaffold00119 1-2125011)
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00010.355 (chr: AmTr_v1.0_scaffold00010 1-9150321)
5avu
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00119.70 (chr: AmTr_v1.0_scaffold00119 1-2125011)
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00010.355 (chr: AmTr_v1.0_scaffold00010 1-9150321)
F
59mv
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00009.65 (chr: AmTr_v1.0_scaffold00009 1-9389330)
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00012.243 (chr: AmTr_v1.0_scaffold00012 1-8757819)
59mu
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00009.65 (chr: AmTr_v1.0_scaffold00009 1-9389330)
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00012.243 (chr: AmTr_v1.0_scaffold00012 1-8757819)
59n1
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00009.65 (chr: AmTr_v1.0_scaffold00009 1-9389330)
Amborella trichopoda (Amborella Genome Consortium v1.0.15, unmasked) evm_15.model.AmTr_v1.0_scaffold00012.243 (chr: AmTr_v1.0_scaffold00012 1-8757819)

Figure S10. Synteny analysis of six pairs of *Amborella* scaffolds showing evidence of an old whole genome duplication event. HSPs (High-Scoring Segment Pairs) are indicated by wedge shaped red colored connectors. Each pair of scaffolds is shown with three different analyses where non-coding sequence is masked in order to minimize obfuscation of the syntenic signal due to repeat sequences. Each panel represents a genomic region where the dashed line separates the upper and lower strand of DNA; gene models are represented as composite arrows with CDS sequence colored green; the background is colored based on whether sequence is present (white), unsequenced (N; orange), or masked (X; purple). Each pair of *Amborella* scaffolds represents syntenic genomic blocks based on the occurrence of a collinear homologous gene pairs. Each analysis can be interactively viewed in GEvo by replacing '####' with the four character code indicated at the left side of each analysis in the following CoGE stable URL: http://genomevolution.org/r/#### (e.g. http://genomevolution.org/r/5apg for the fully masked analysis of scaffold 9 and 29)
The pairs of scaffolds represented are:
(A) AmTr_v1.0_scaffold00009 vs AmTr_v1.0_scaffold00029;
(B) AmTr_v1.0_scaffold00057 vs AmTr_v1.0_scaffold00016;
(C) AmTr_v1.0_scaffold00048 vs AmTr_v1.0_scaffold00024;
(D) AmTr_v1.0_scaffold00078 vs AmTr_v1.0_scaffold00004;
(E) AmTr_v1.0_scaffold00119 vs AmTr_v1.0_scaffold00010;
(F) AmTr_v1.0_scaffold00009 vs AmTr_v1.0_scaffold00012.

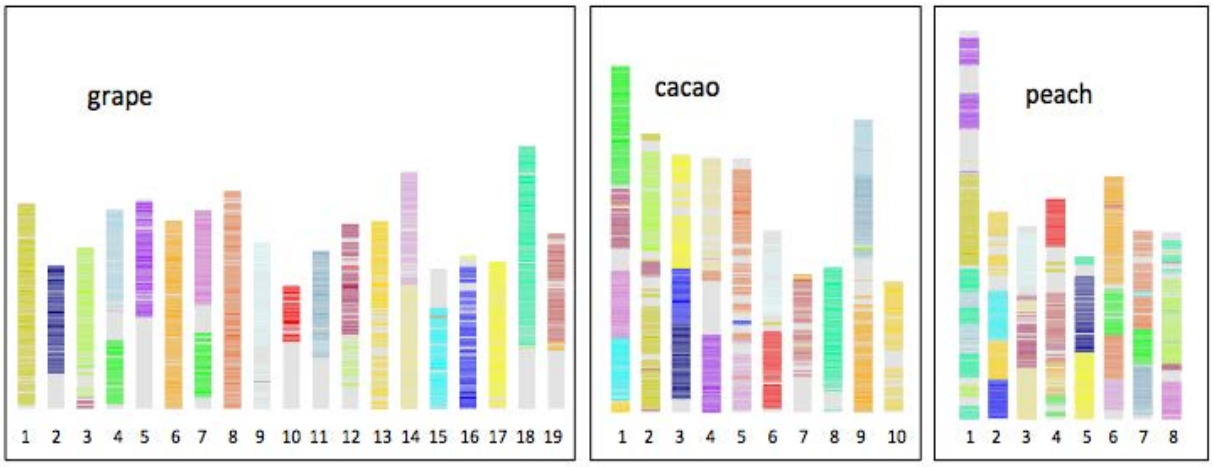


Figure S11. Chromosomal distribution of triplicated regions in the grape, cacao and peach genomes. Three shades (dark, medium and light) of each of seven colors represent homeologous chromosomes or regions deriving from the same ancestor chromosome, as in Fig. S7. Vertical axis within each genome proportional to the number of genes on the chromosomes, with a horizontal bar at the position of each gene. Gray areas represent genes that are not in colored regions (*136*, *162-166*).

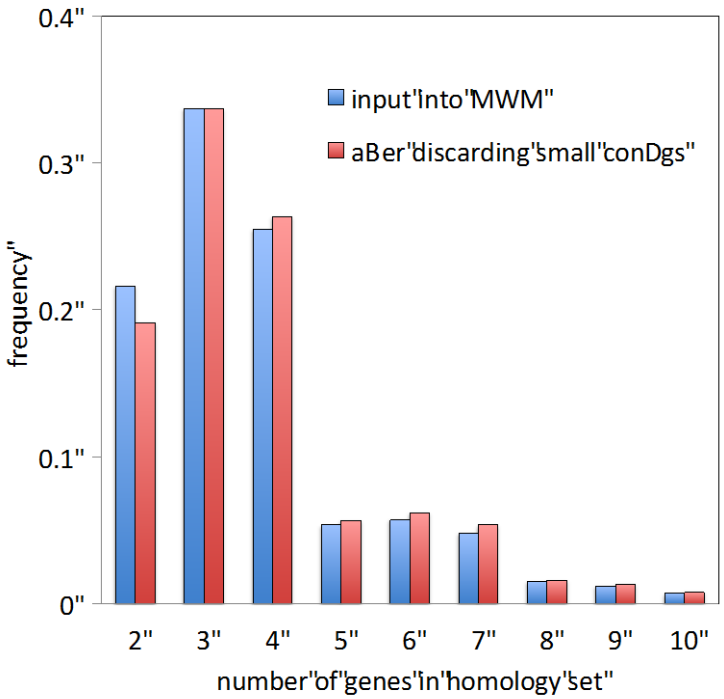


Figure S12. Improvement in homology set size after discarding small contigs.

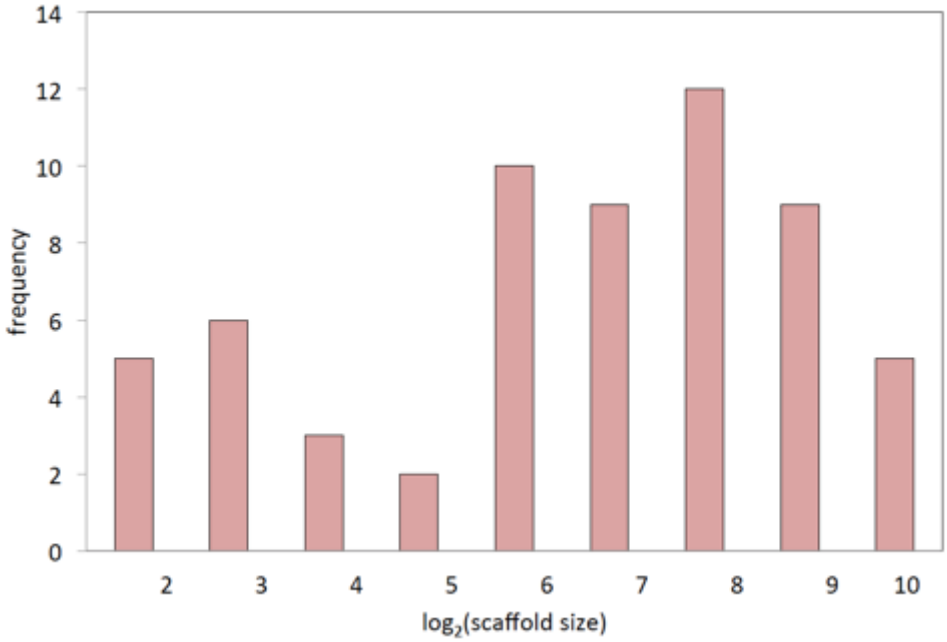


Figure S13. Distribution of scaffold sizes. 45 of 61 scaffolds contain more than 32 genes.

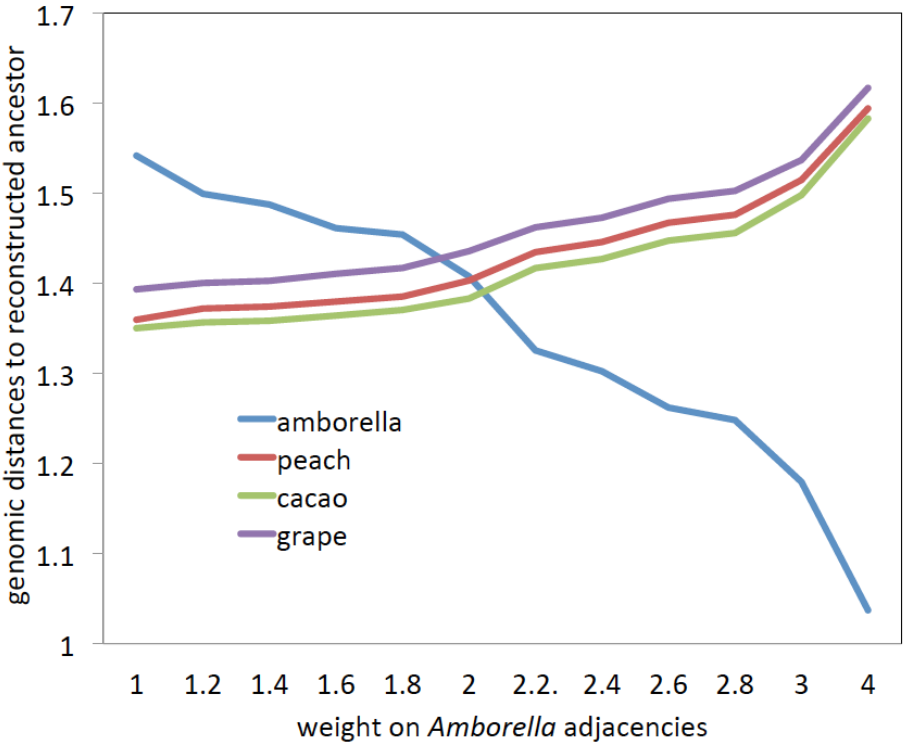


Figure S14. Genomic costs as a function of relative weight on *Amborella* contig adjacencies compared to those in the rosids.

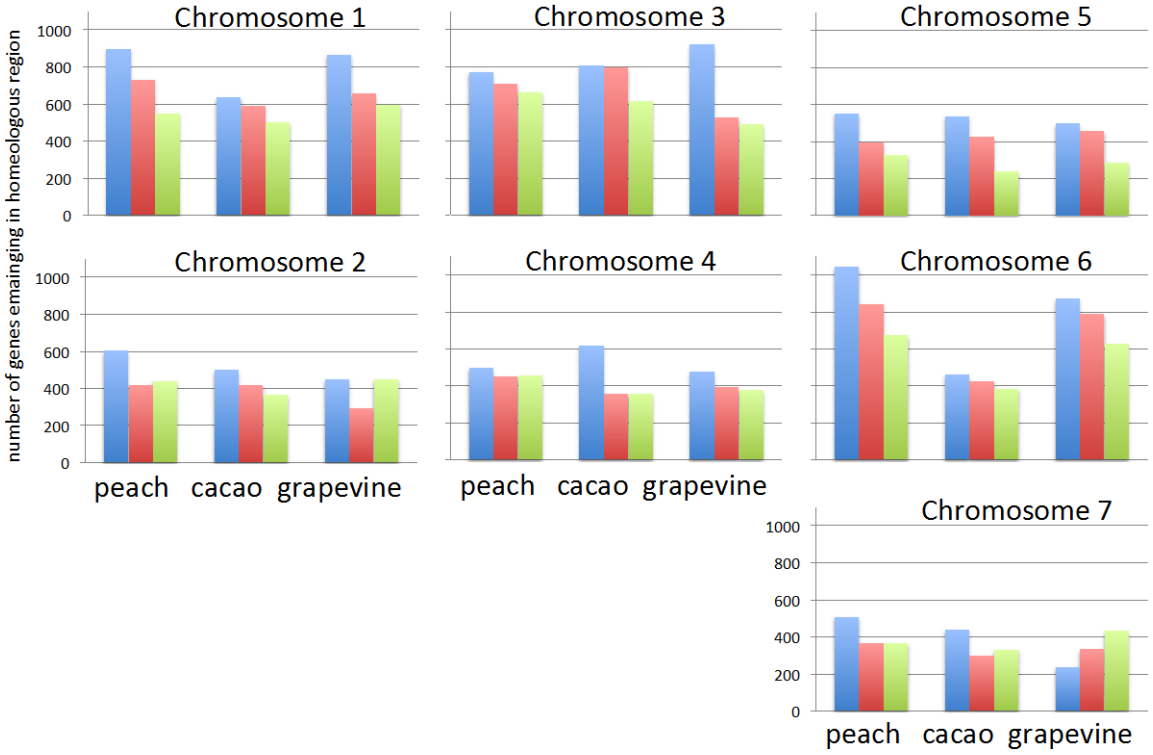


Figure S15. Number of genes retained from seven ancestral chromosomes in three homeologous daughter regions in peach, cacao and grape. Daughter region with the largest number of genes in peach, and homeologous regions in cacao and grape, are colored blue. Under a random null hypothesis, the probability that 13 out of 14 times dominant peach homeologs regions are orthologous to the dominant homeologs in another species is significant at the 0.00001 level (binomial model with p=1/3), without even having to take into account the numerical differences in the retention rates. This dominance of one daughter must have arisen before the grape-eurosid split, and is consistent with the more recent participation of the "blue" genome in the sequence of polyploidization events, as was recently suggested in the analysis of the *Nelumbo nucifera* genome (*82*, *163*, *166*).

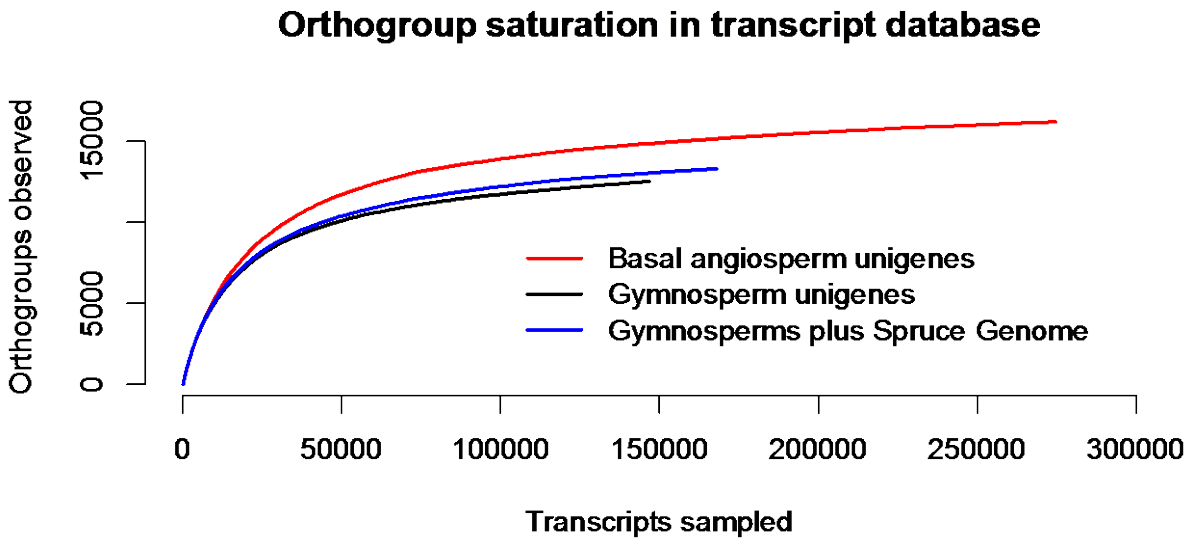


Figure S16. Orthogroup detection in basal angiosperm and gymnosperm transcript assemblies and the Norway spruce genome. Transcripts from 16 gymnosperms and four basal angiosperms were sampled in random order and the cumulative number of orthogroups observed was plotted, showing saturation.

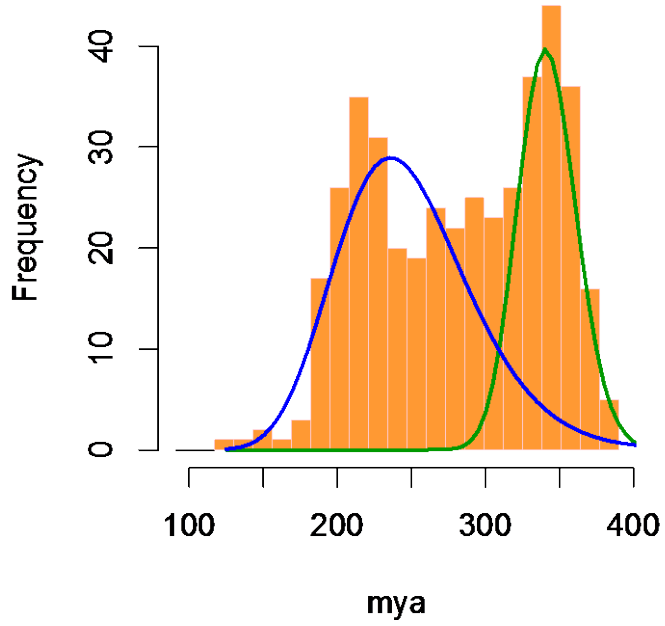


Figure S17. Age distribution of ancient duplications that are shared by *Amborella* and the rest of the angiosperms. The inferred divergence times for 414 ancestral duplication nodes identified through phylogenomic analysis (bootstrap support over 50%) were analyzed by EMMIX to determine whether these duplications occurred randomly over time or are clustered in time. Two statistically significant components were identified with mean ages estimated at 244 mya (blue line; 62% of the nodes) and 341 mya (green line; 38% of the nodes).

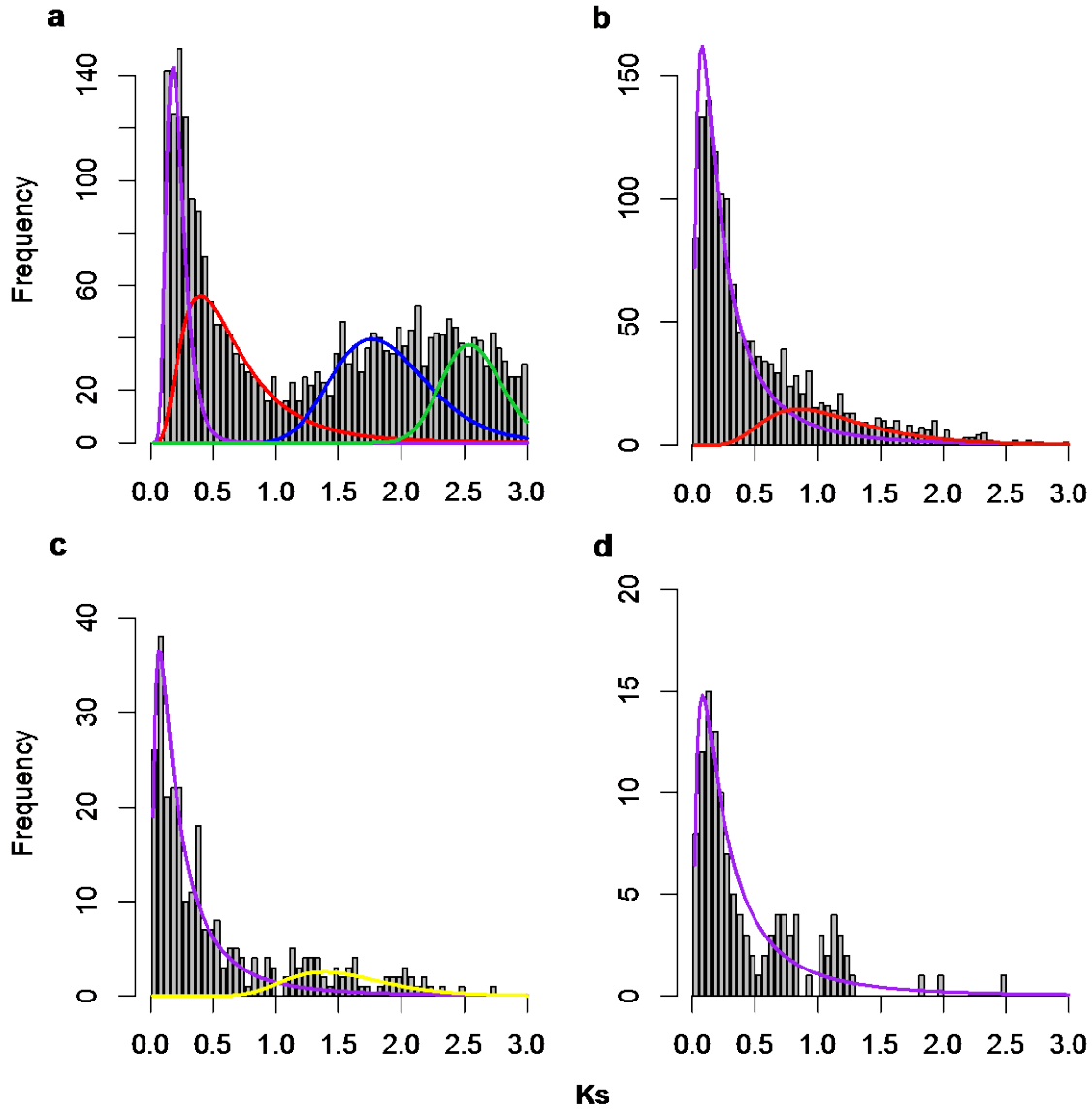


Figure S18. K_s distributions of paralogous gene pairs in *Amborella*. A) Pairwise K_s divergences for 4402 reciprocal 'best hit' gene pairs in the *Amborella* genome. Paralogous pairs of genes were identified from all-by-all BLASTP searches, and duplicated pairs (we define tandem duplications as those closely related genes falling within five genes of one another) were removed before parsing the best reciprocal hits. Colored lines superimposed on the K_s distribution represent significant duplication components identified by a likelihood mixture model (*37, 163, 166-168*). 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. Four statistically significant components: purple-0.20-0.21, red-0.56-0.31, blue-1.84-0.30, green-2.56-0.18. B) Pairwise K_s divergences for 1536 tandem duplication gene pairs across whole genome. Two statistically significant components were identified: purple-0.22-0.78, red-1.04-0.22. C) Pairwise K_s divergences for 282 non-tandem terminal duplication gene pairs. Two statistically significant components: purple-0.20-0.82, yellow-1.49-0.18. D) Pairwise K_s divergences for 121 tandem terminal duplication gene pairs. One component: purple-0.27-1.

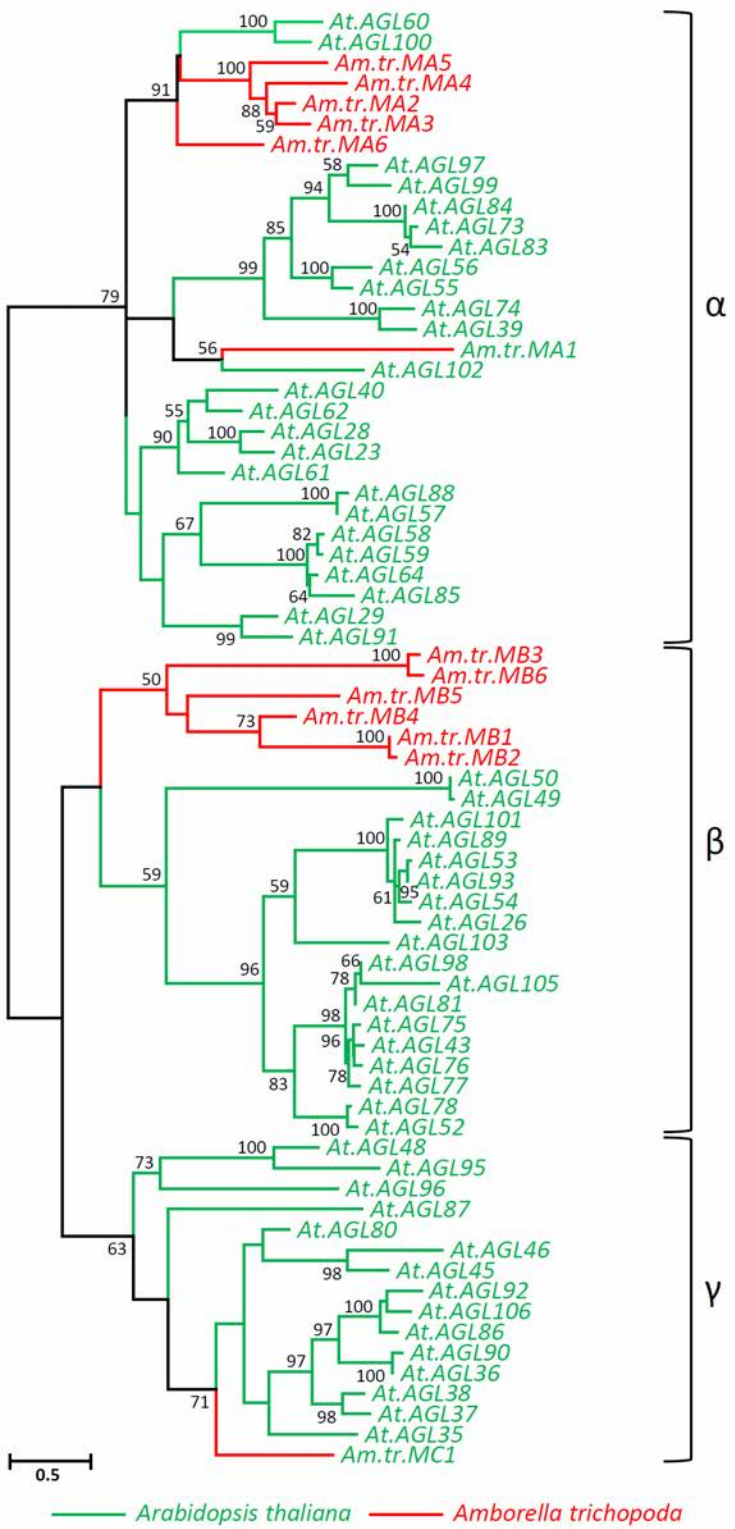


Figure S19. Phylogenetic relationships of type I MADS-box genes in *Amborella trichopoda* and *Arabidopsis thaliana*. The tree was constructed using maximum likelihood in PhyML. Bootstrap values > 50% are shown.

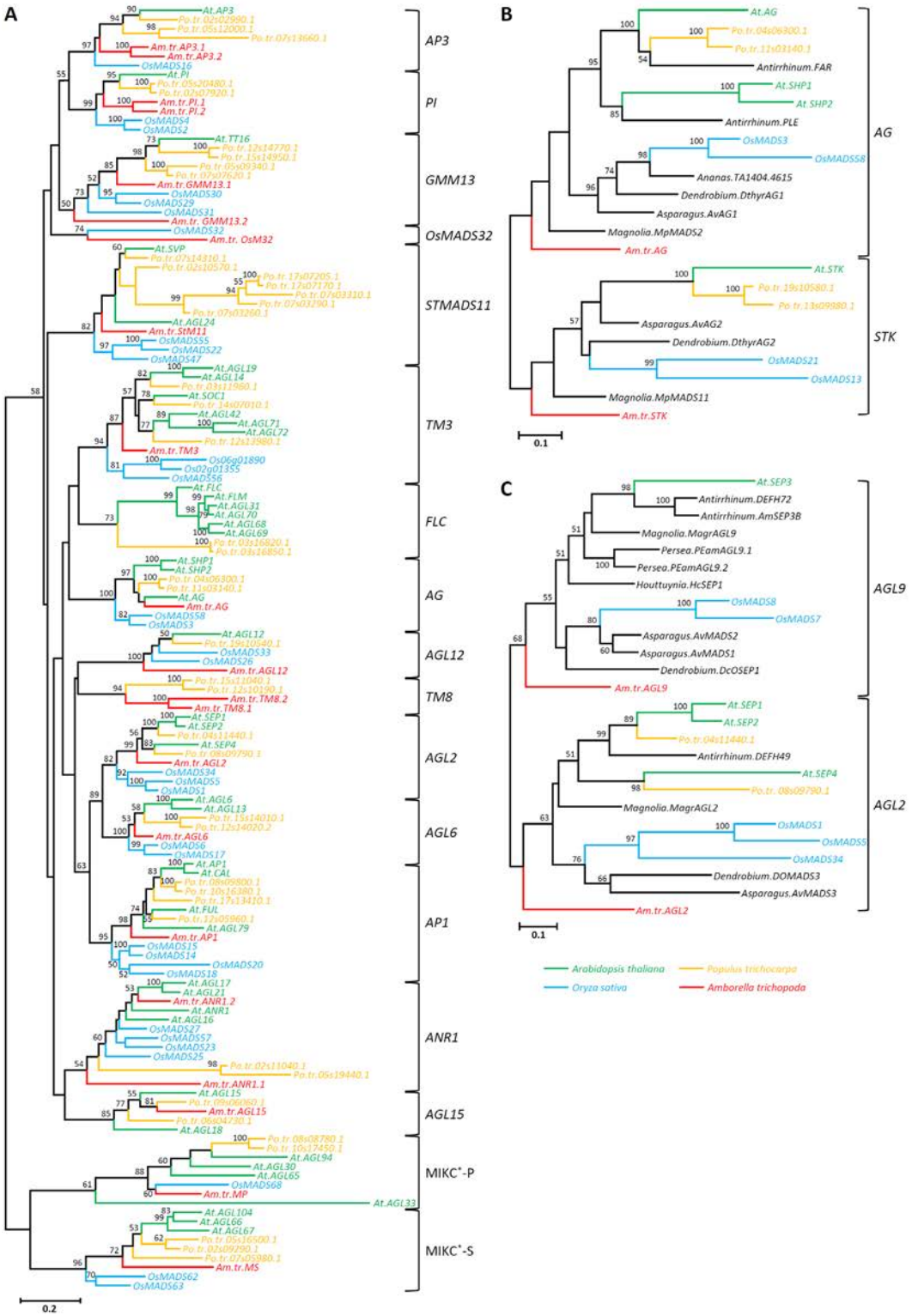


Figure S20. Phylogenetic relationships of type II MADS-box genes in *Amborella trichopoda*, *Arabidopsis thaliana*, *Populus trichocarpa* and *Oryza sativa*. (A) All of type II MADS-box genes except *STK* and *AGL9* lineages. (B) *AG* and *STK* lineages. (C) *AGL2* and *AGL9* lineages. The tree was constructed by using maximum likelihood in PhyML. Bootstrap values > 50% are shown.

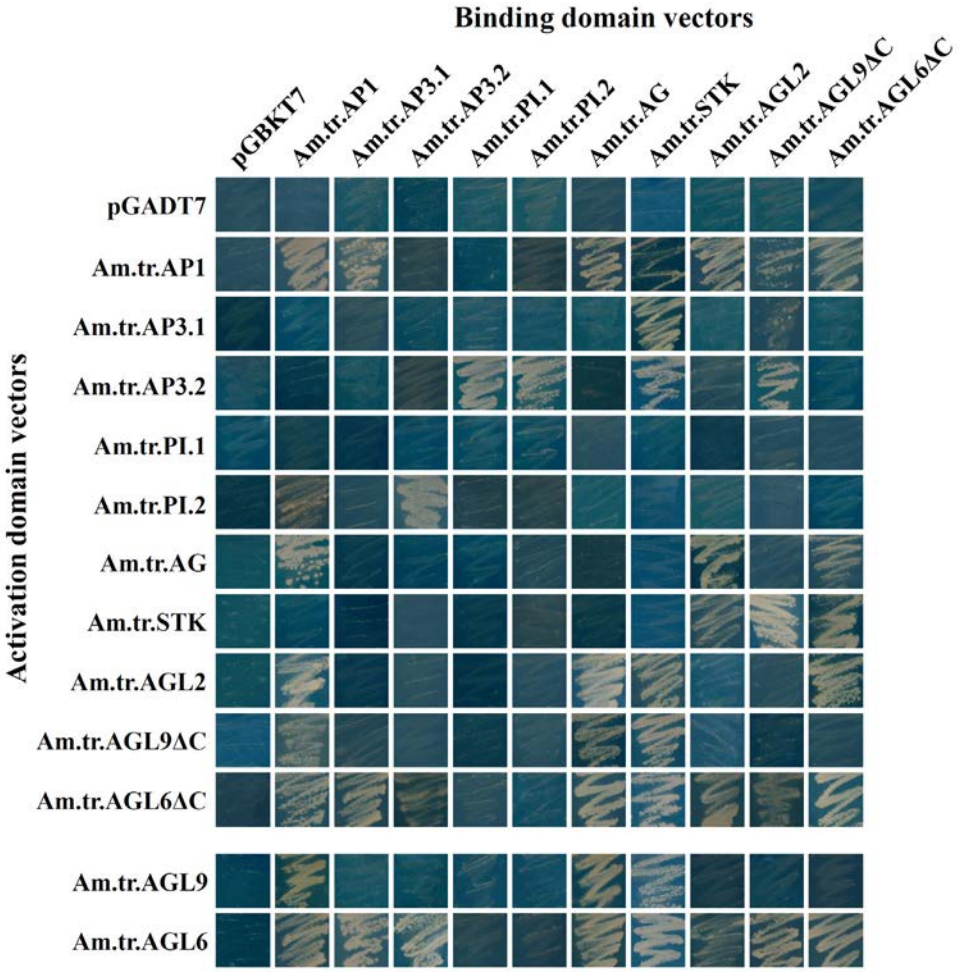


Figure S21. Interaction patterns between proteins of the *Amborella trichopoda* floral MADS-box genes as revealed by yeast two-hybrid assays. Δ C indicates the truncation of the C-terminal regions of *Am.tr.AGL9* and *Am.tr.AGL6* because the binding domain vector constructs of the full-length candidate genes auto-activated the expression of reporter genes.

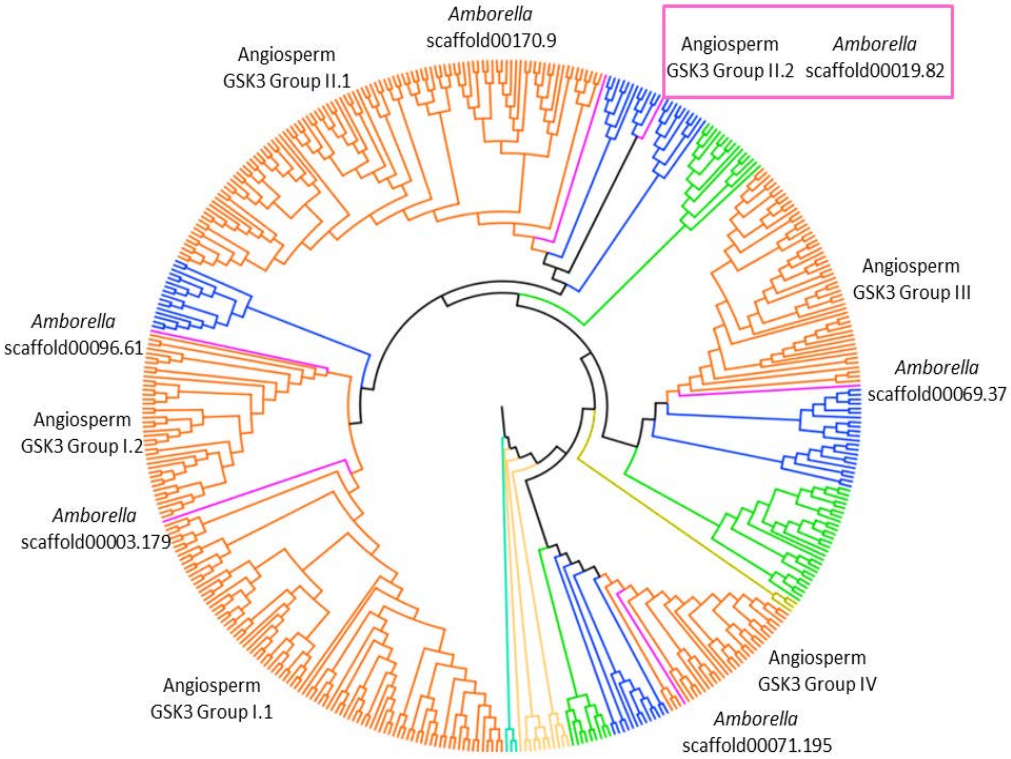


Figure S22. Phylogeny of plant *GSK3* genes. Line colors indicate genes from algae (turquoise), liverworts and mosses (light brown), lycophytes (brown-green), monilophytes (green), gymnosperms (blue), and angiosperms (orange) excluding *Amborella*, which are indicated in magenta. Five *Amborella* genes are sister to all or most angiosperm genes in five pan-angiosperm gene lineages. A sixth *Amborella* gene is placed within a grade of gymnosperm genes.

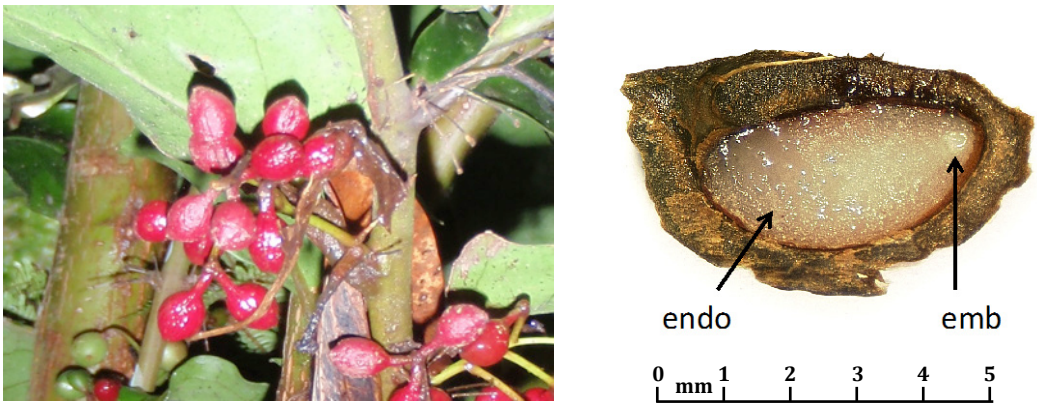


Figure S23. Left, *Amborella trichopoda* fruits on a tree. Right, transverse cut of the fruit without the fleshy exocarp, showing the localization of the endosperm (endo) and embryo (emb) in the seed within the sclerotized endocarp/mesocarp.

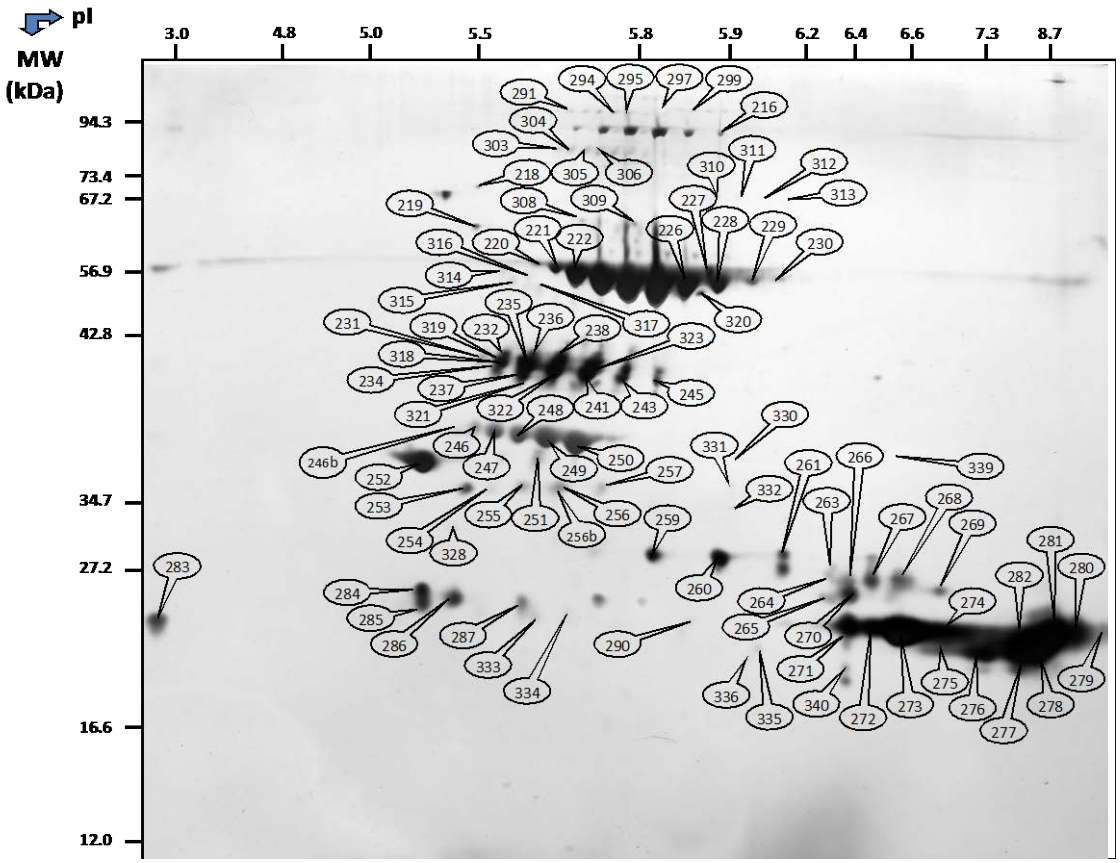


Figure S24. Reference map for *Amborella* 11S globulins. Localization of the 99 protein spots identified by MS (labeled from 216 to 340) on a reference 2DE gel of 11S globulins from mature *Amborella* seeds. Peptide sequences are listed in Table S21.

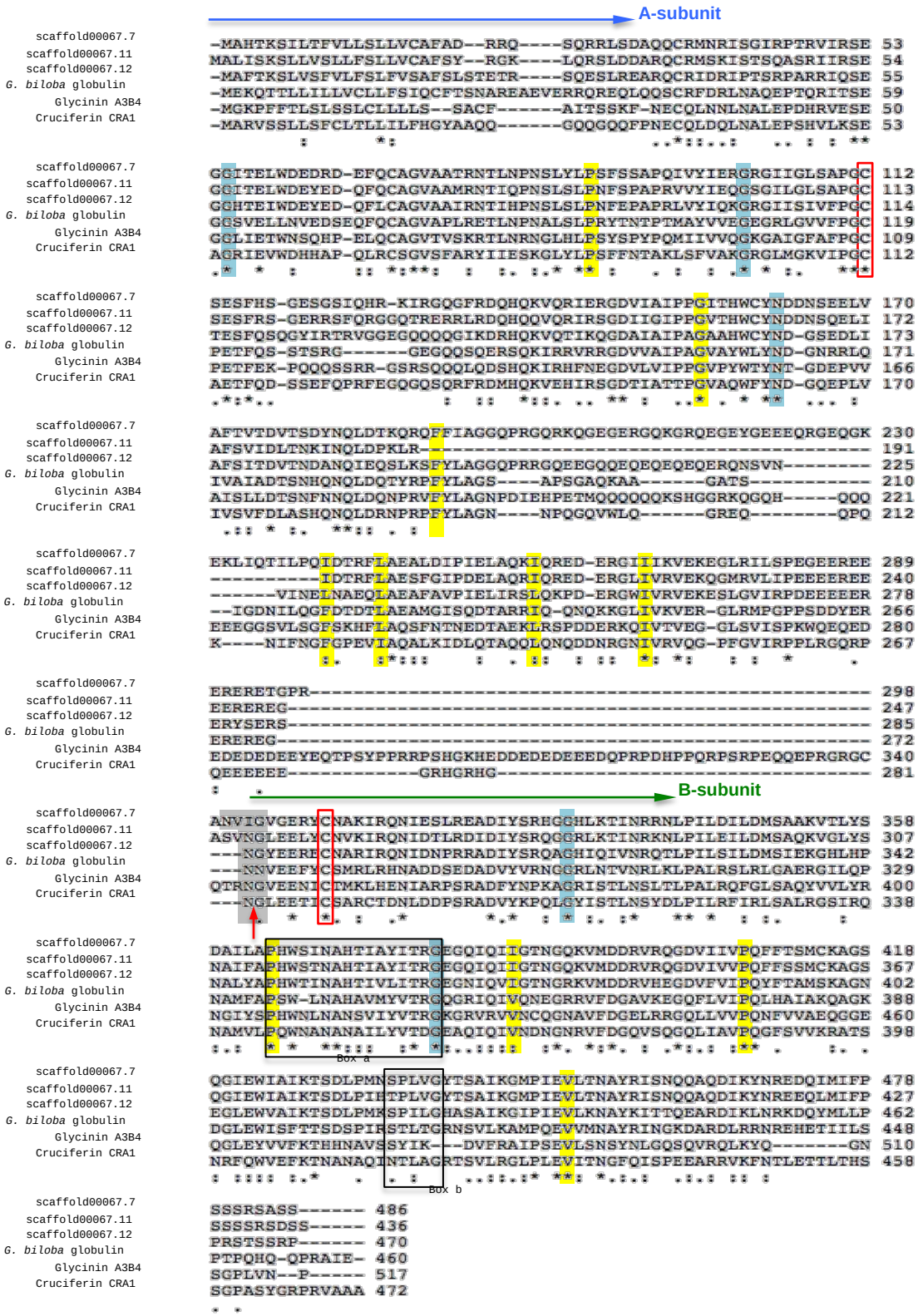
evm_27.model.AmTr_v1.0_scaffold00067.7
MAHTKSILTFVLLSLLVCAFADRRQSQRRLSDAQQCRMNRISGIRPTRVIRSEGGITELWDEDRDEFQCAGV
AATRNTLNPNSLYLPSFSSAPQIVYIERGRGIIGLSAPGCSESFHSGESGSIQHRKIRGQGFRDQHQKVQRIER
GDVIAIPPGITHWCYNDDNSEELVAFTVTDVTSDYNQLDTKQRQFFIAGGQPRGQRKQGEGERGQKGRQE
GEYGEEEQRGEQGKEKLIQTILPQIDTRFLAEALDIPIELAQKIQREDERGIIIKVEKEGLRILSPEGEEREEERE
RETGPRANVIGVGERYCNAKIRQNIESLREADIYSRHGGHLKTINRRNLPILDILDMSAAKVTLYSDAILAPH
WSINAHTIAYITRGEGQIQIIGTNGQKVMDDRVRQGDVIIVPQFFTSMCKAGSQGIEWIAIKTSDLPMNSPLV
GYTSAIKGMPIEVLTNAYRISNQQAQDIKYNREDQIMIFPSSSRSASS

evm_27.model.AmTr_v1.0_scaffold00067.12
MAFTKSLVSFVLFSLFVSAFSLSTETRSQESLREARQCRIDRIPTSRPARRIQSEGGHTEIWDEYEDQFLCAGV
AAIRNTIHPNSLSLPNFEPAPRLVYIQKGRGIISIVFPGCTESFQSQGYIRTRVGGEGQQQQGIKDRHQKVQTI
KQGDAIAIPAGAAHWCYNDGSEDLIAFSITDVTNDANQIEQSLKSFYLAGGQPRRGQEEGQQEQEQEQEQE
RQNSVNVINELNAEQLAEAFAVPIELIRSLQKPDERGWIVRVEKESLGVIRPDEEEEERERYSERSNGYEERE
CNARIRQNIDNPRRADIYSRQAGHIQIVNRQTLPILSILDMSIEKGHLHPNALYAPHWTINAHTIVLITRGEGN
IQVIGTNGRKVMDDRVHEGDVFVIPQYFTAMSKAGNEGLEWVAIKTSDLPMKSPILGHASAIKGIPIEVLKN
AYKITTQEARDIKLNRKDQYMLLPPRSTSSRP

evm_27.model.AmTr_v1.0_scaffold00067.11
MALISKSLLVSLLFSLLVCAFSYRGKLQRSLDDARQCRMSKISTSQASRIIRSEGGITELWDEYEDQFQCAGV
AAMRNTIQPNSLSLPNFSPAPRVVYIEQGSGILGLSAPGCSESFRSGERRSFQRGGQTRERRLRDQHQQVQRI
RSGDIIGIPPGVTHWCYNDDNSQELIAFSVIDLTNKINQLDPKLRIDTRFLAESFGIPDELAQRIQREDERGLIV
RVEKQGMRVLIPEEEEREEEEREREGASVNGLEELYCNVKIRQNIDTLRDIDIYSRQGGRLKTINRKNLPILEI
LDMSAQKVGLYSNAIFAPHWSTNAHTIAYITRGEGQIQIIGTNGQKVMDDRVRQGDVIVVPQFFSSMCKAG
SQGIEWIAIKTSDLPIHTPLVGYTSAIKGMPIEVLTNAYRISNQQAQDIKYNREEQLMIFPSSSSRSDSS

Figure S25. Inferred protein sequences for the three *Amborella* gene models with 11S globulin peptides mapping them. Highlighted regions show peptide matches.

A


A-subunit
B-subunit
Box a
Box b
scaffold00067.7
scaffold00067.11
scaffold00067.12
G. biloba globulin
Glycinin A3B4
Cruciferin CRA1

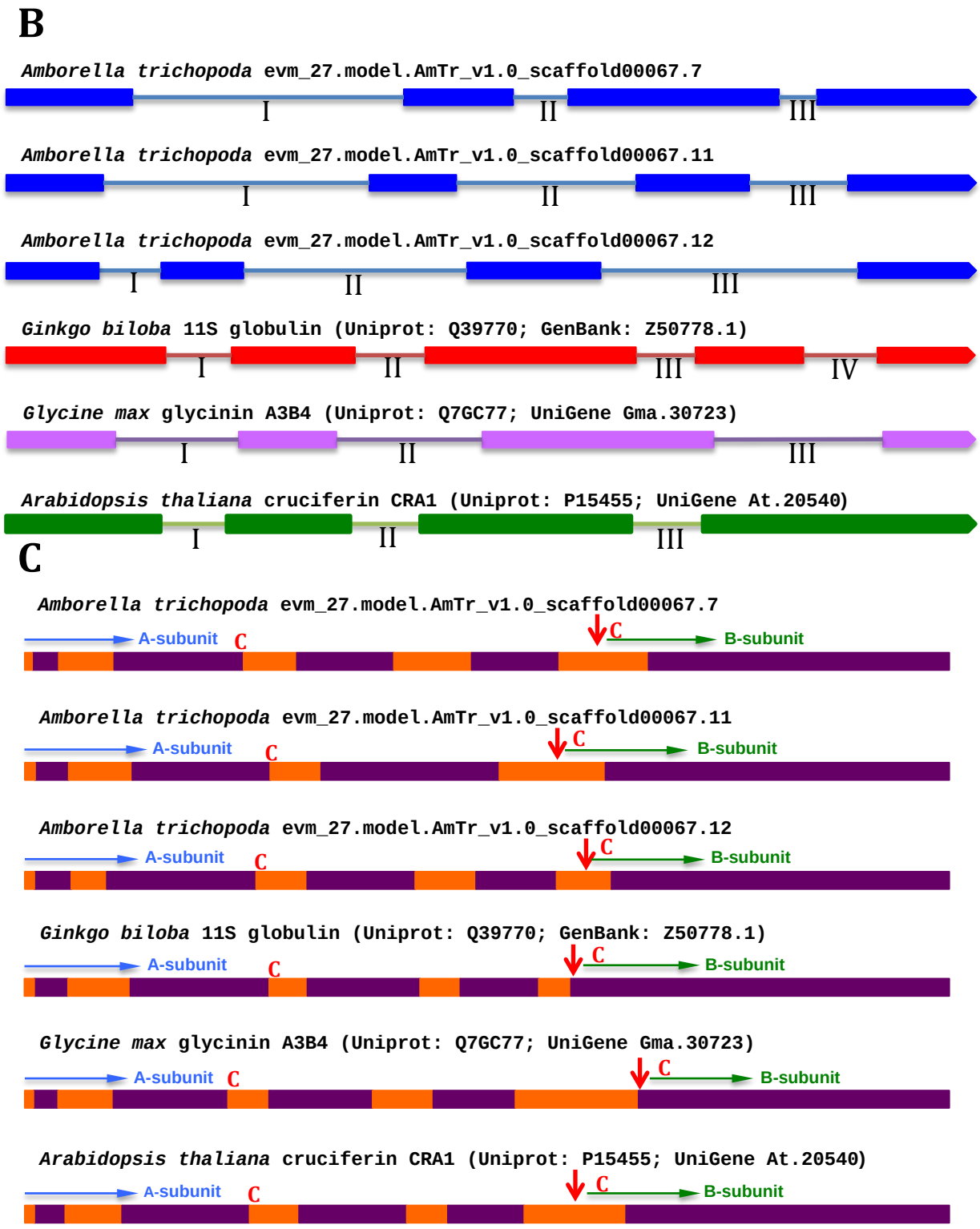


Figure S26. 11S globulin family of *Amborella trichopoda.*
(A) Protein sequence alignment of 11S globulin family from *Amborella*, *Glycine max* (Uniprot: Q7GC77), *Ginkgo biloba* (Uniprot: Q39770) and *Arabidopsis thaliana* (Uniprot: P15455). Multiple sequence alignment was performed with ClustalW2 program. The BLOSUM matrix

was used to score the alignment. The Multiple Sequence alignment parameters were: Gap Open = 5; Gap Extension = 0.05; Gap Distances = 1; No End Gaps = Yes. The red boxes show the two Cys involved in the conserved disulfide bond linking acidic (A)- and basic (B)- subunits of 11S globulins. The red arrow indicates the conserved cutting site by a vacuolar processing enzyme/asparaginyl endopeptidase (VPE/Asn-endopeptidase). The blue and green arrows point the A- and B-subunits respectively. The amino acids highlighted with gray shading indicate the conserved motif presumably involved in the recognition of the cutting site just behind the first Asn residue (N). The amino acids highlighted with yellow shading indicate the residues involved in trimer formation of the precursor forms and stabilization of 11S globulins. The amino acids highlighted with blue shading indicate the residues involved in the correct globular folding of 11S globulins. Black boxes a and b indicate regions that are homologous to PX14G (Box a) and SX3G (Box b) motifs adjacent to the position of introns III and IV respectively.
(B) Schematic structural description of 11S globulin gene models is displayed here indicating exons (boxes) and introns (lines) positions for *Amborella*, *Glycine max*, *Ginkgo biloba* and *Arabidopsis*. Introns are labeled with Roman numbers.
(C) 11S-Globulins predicted structural organization from *Amborella*, *Glycine max*, *Ginkgo biloba* and *Arabidopsis*. Protein structure was predicted with Phyre2 (http://www.sbg.bio.ic.ac.uk/phyre2). Orange boxes represent disordered regions. Purple boxes represent organized regions (including alpha helix and beta strand). The red arrow shows the conserved cleavage site by a vacuolar processing enzyme/asparaginyl endopeptidase (VPE/Asn-endopeptidase). The red letter C shows the two Cys involved in the conserved disulfide bond linking the A- and B- subunits of 11S globulins. The blue and green arrows point the A- and B-subunits respectively.

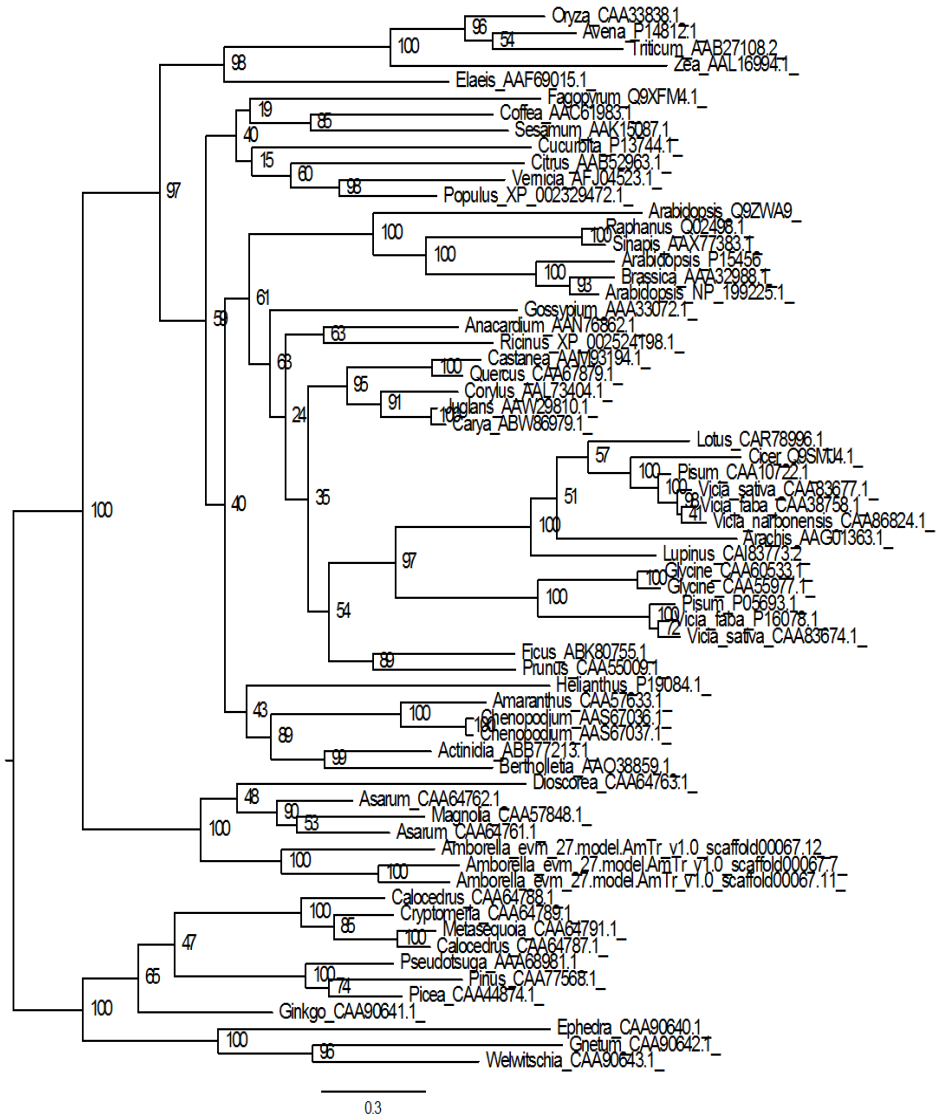


Figure S27. Phylogenetic tree of 11S globulin proteins in seed plants shows that angiosperm homologs form a well-supported clade. Estimated relationships among eudicots, monocots, magnoliids, and *Amborella* suggest clade-specific duplications within the *Amborella* and eudicot lineages. Seed plant 11S globulin genes included in the phylogenetic analysis are listed in Table S22.

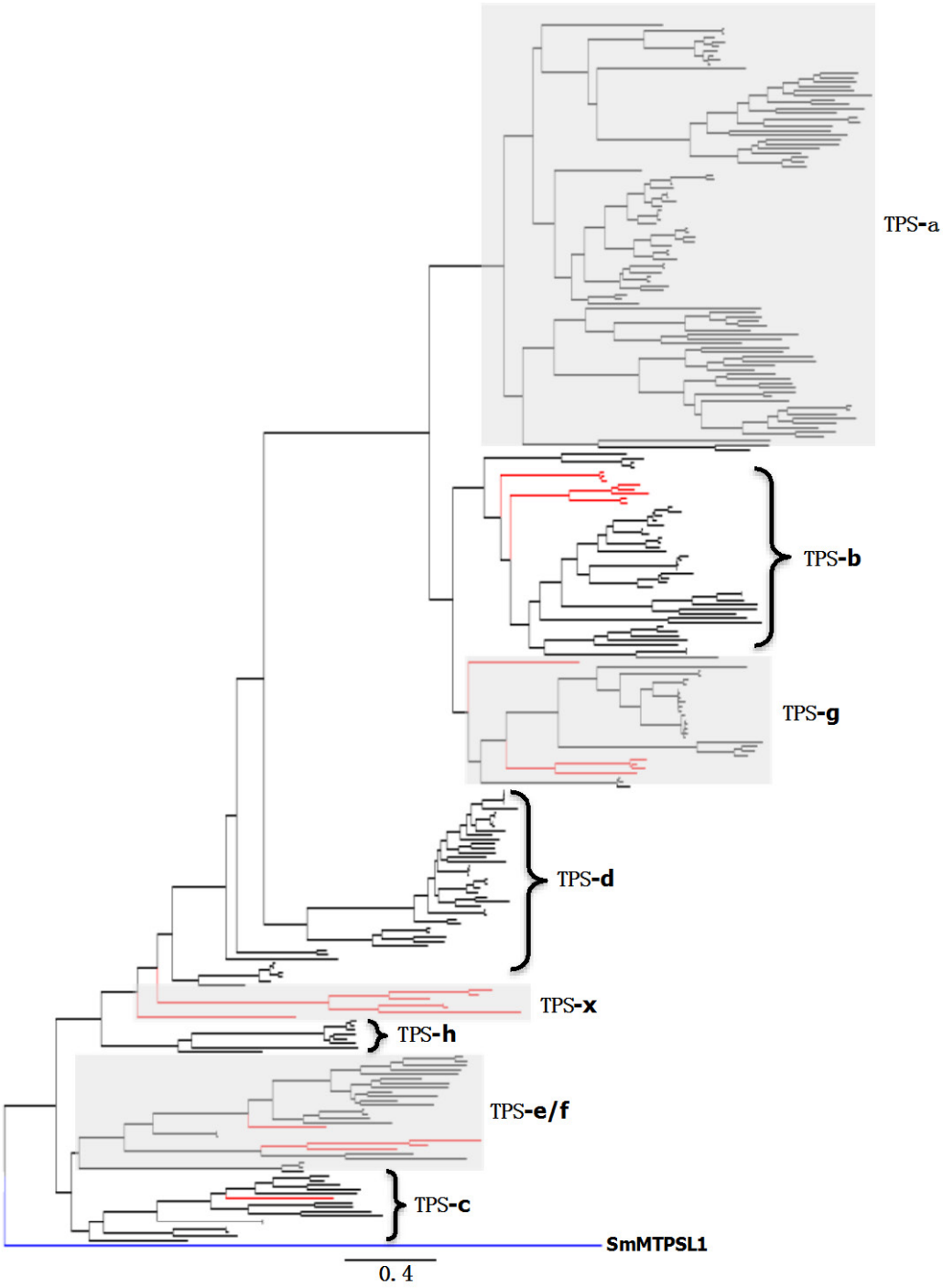


Figure S28. Phylogenetic tree of the 25 putative full-length *Amborella* TPSs (in red) and the TPSs from other plants (in black), as described in Chen *et al*. (*59*). A microbial TPS-like protein, SmMTPSL1, from *Selaginella moellendorffii* (in blue) (*204*) was used as an outgroup. *TPS-a*, *b*, *c*, *d*, *e/f*, *g*, and *h* represent known TPS subfamilies. A group of TPSs from *Amborella* that did not belong to any of the existing subfamilies was tentatively grouped into the *TPS-x* subfamily.

(A)

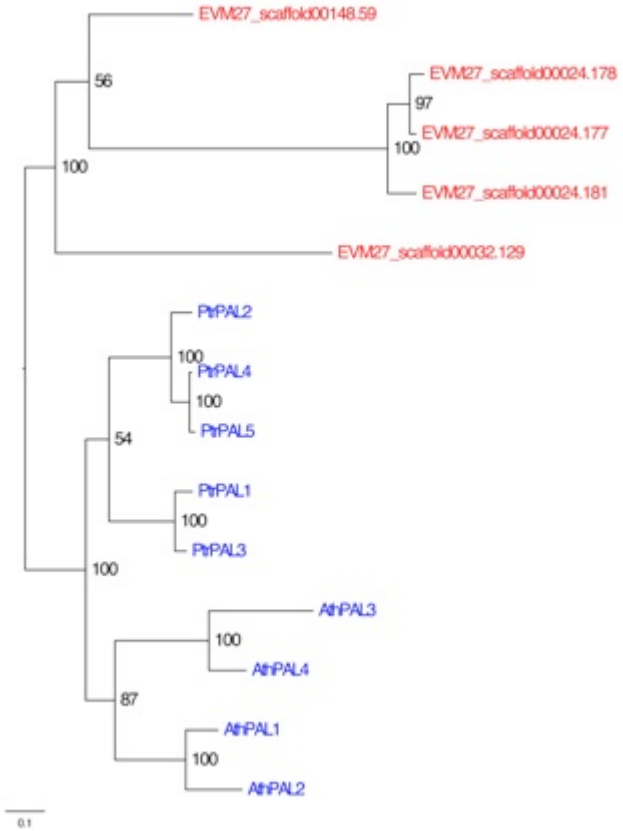


Figure S29 continued

(B)

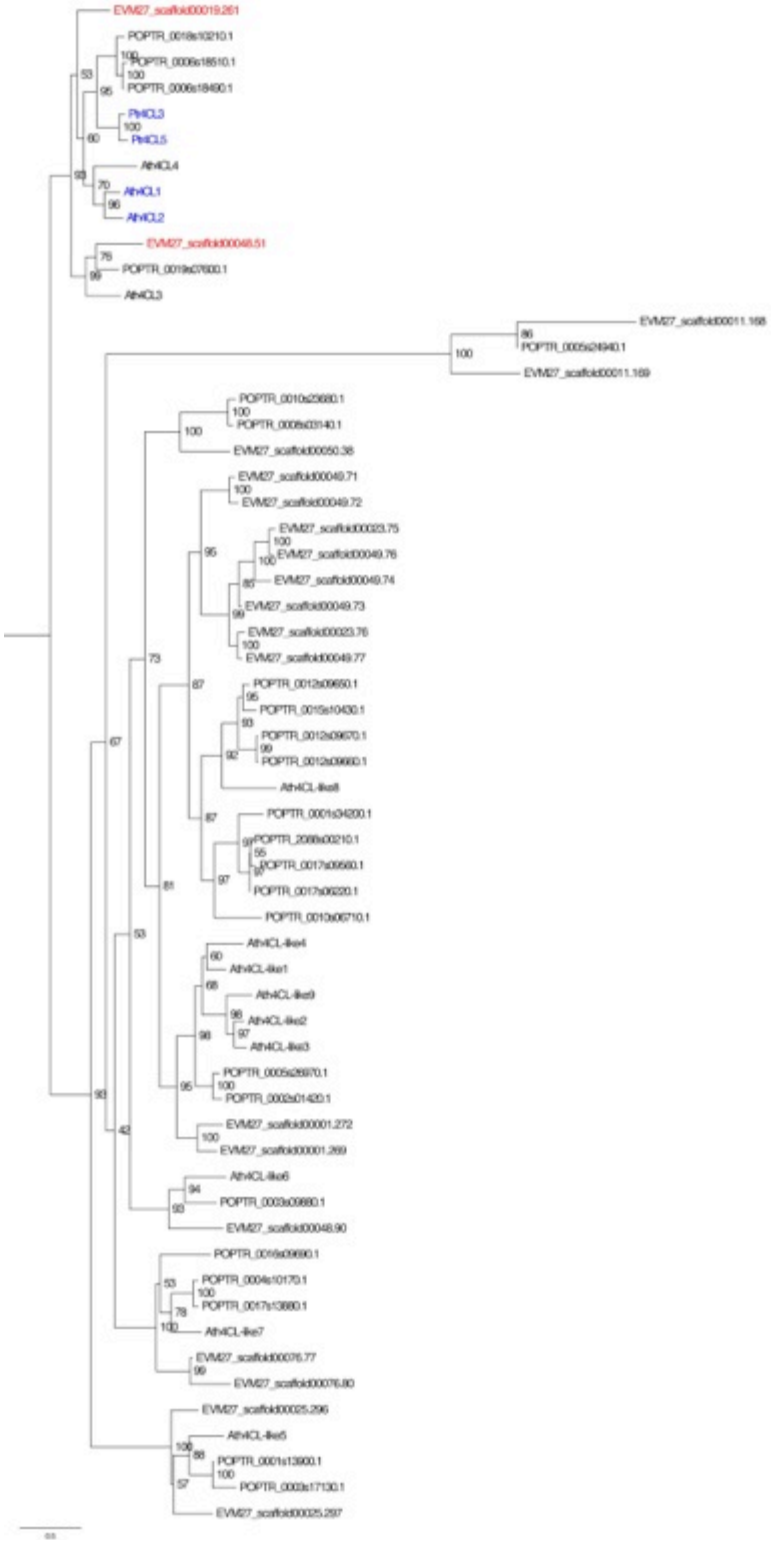


Figure S29 continued

(C)

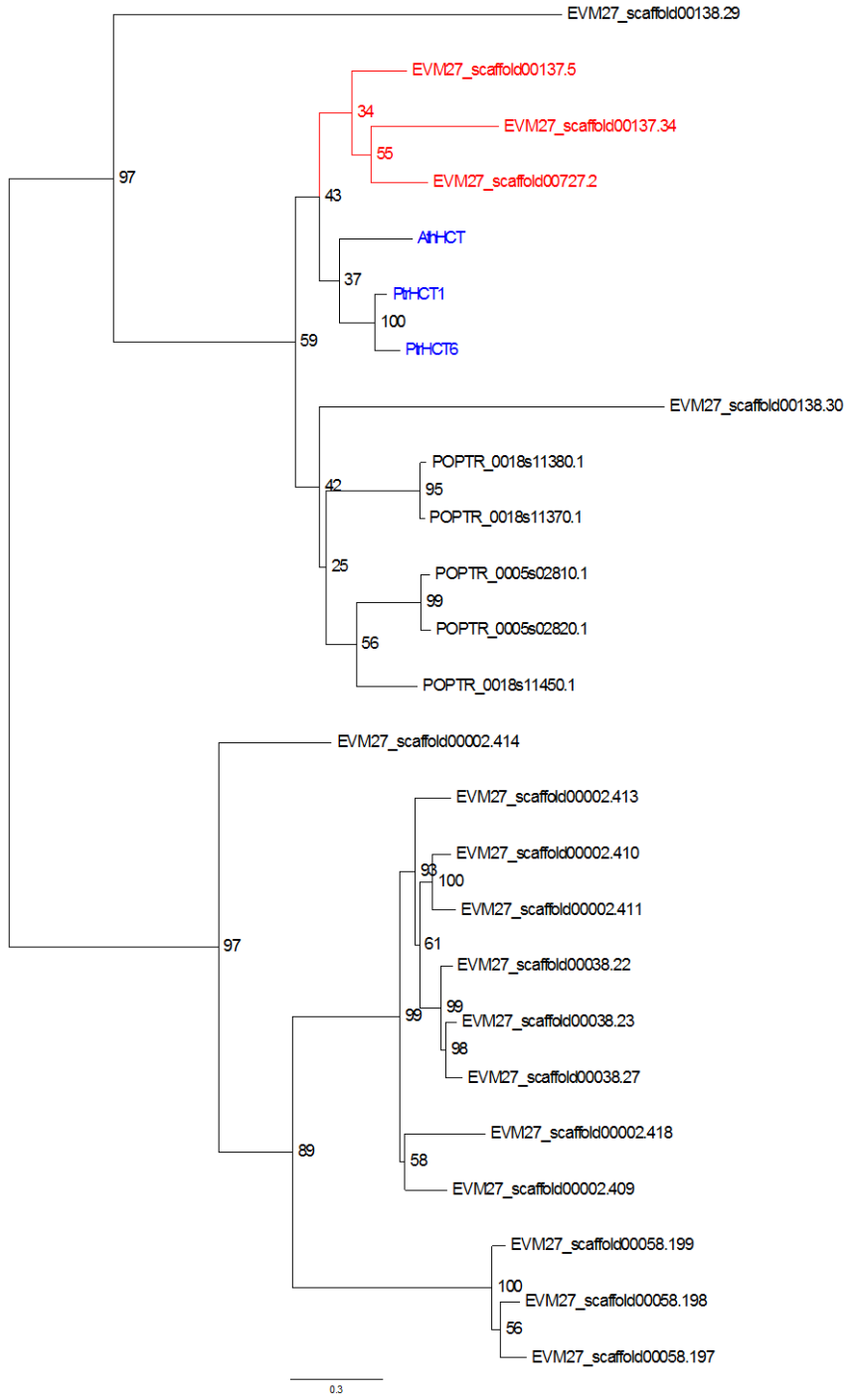


Figure S29 continued

(D)

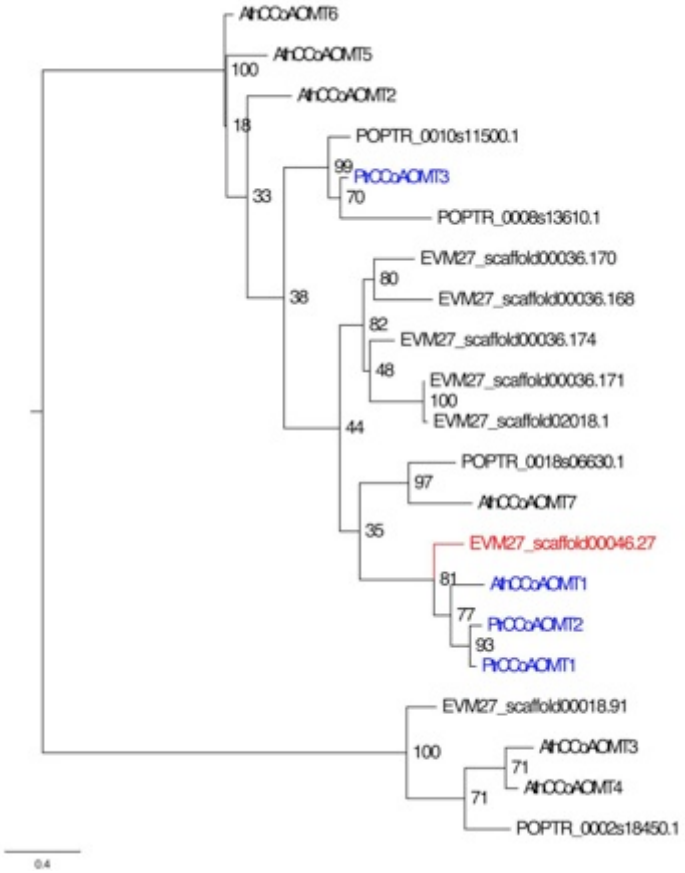


Figure S29 continued

(E)

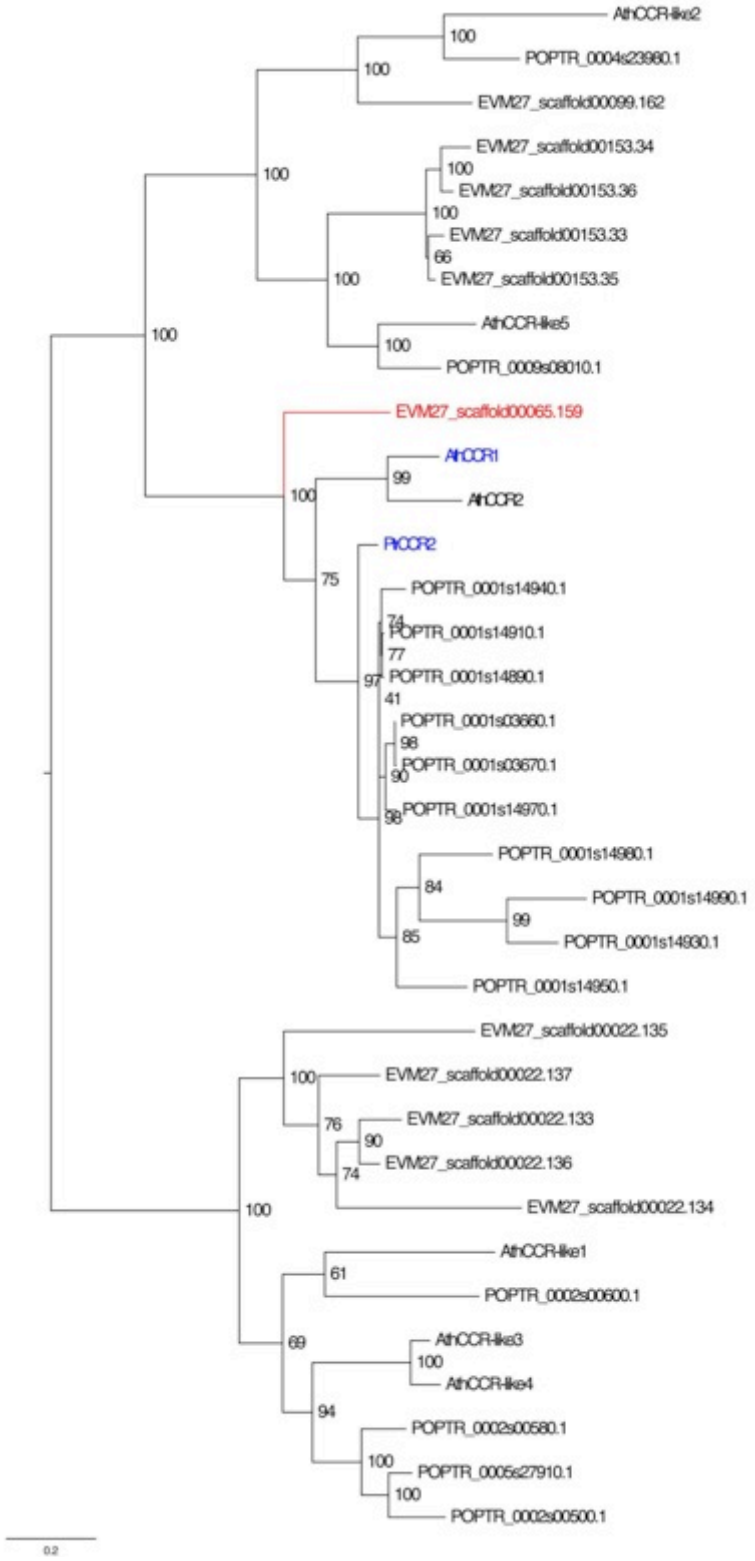


Figure S29 continued

(F)

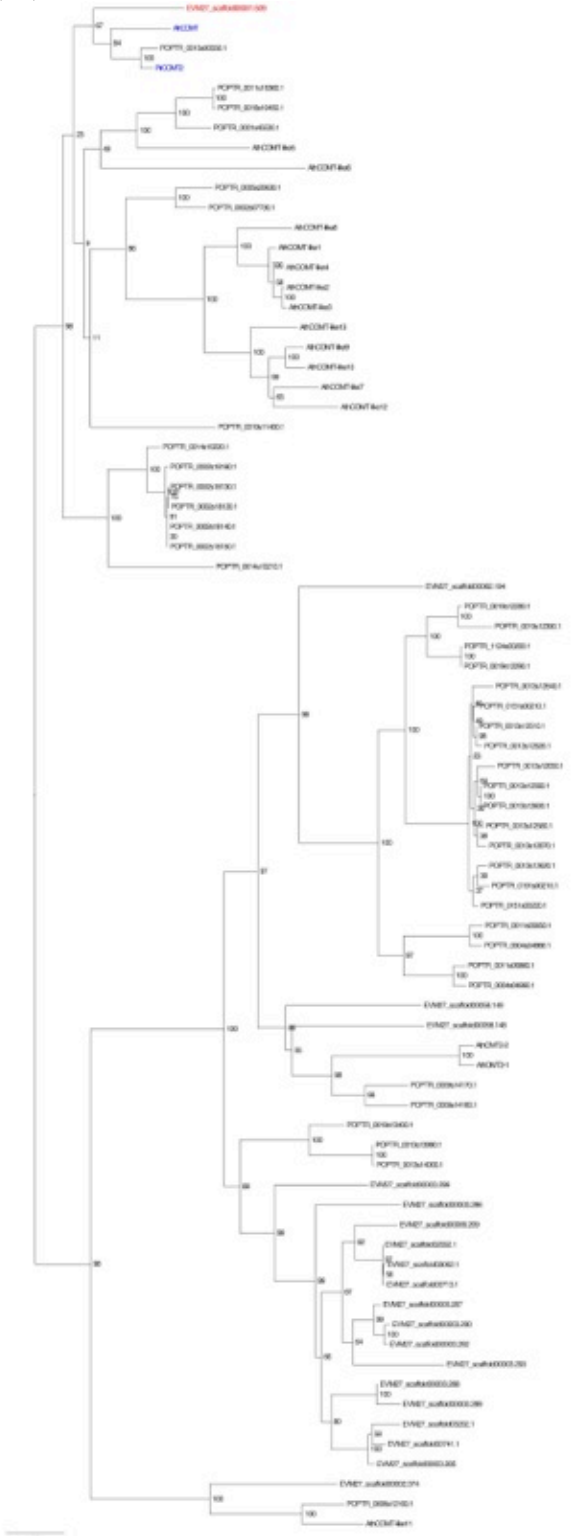


Figure S29 continued

(G)

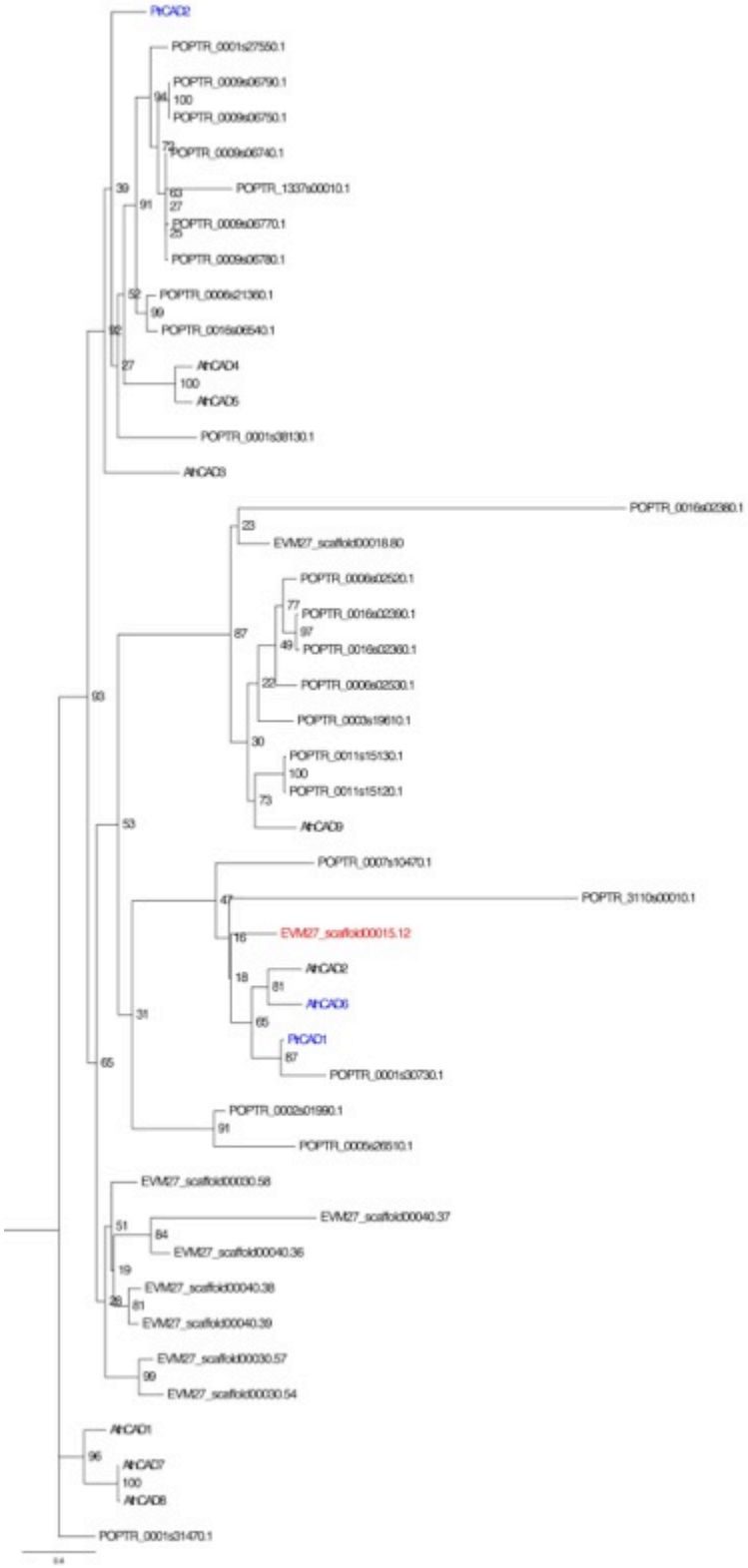


Figure S29 continued

(H)

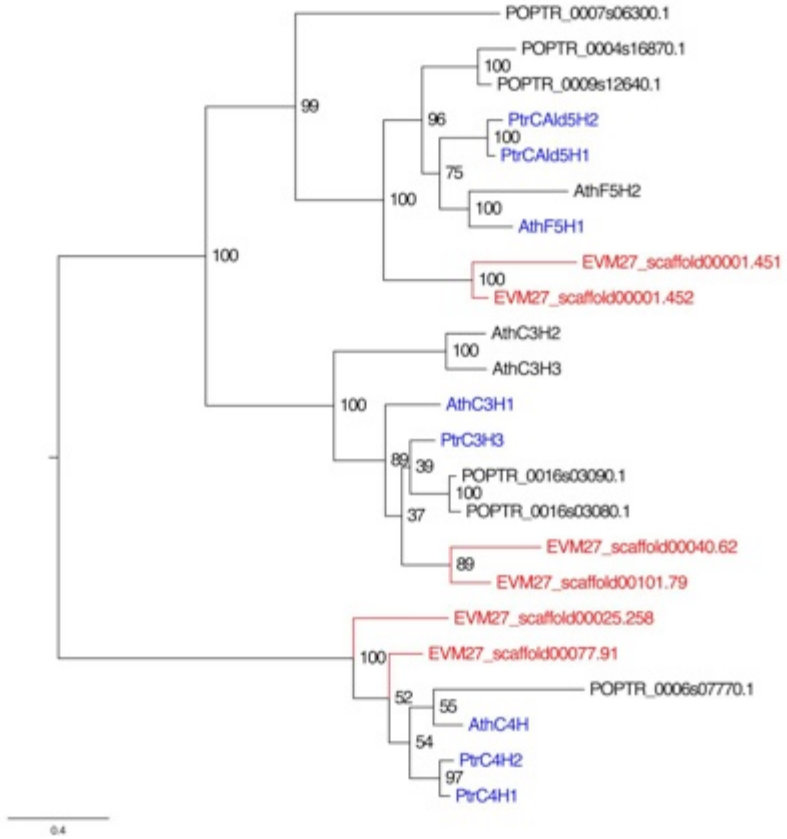


Figure S29. Maximum likelihood phylogenetic trees of 10 phenylpropanoid biosynthesis gene families in *Arabidopsis*, *Populus trichocarpa and Amborella trichopoda.* (A) PAL, (B) 4CL, (C) HCT, (D) CCoAOMT, (E) CCR, (F) COMT, (G) CAD, and (H) C3H, C4H and Cald5H. Taxa labeled in blue represent genes known monolignol biosynthesis in *Arabidopsis* and *Populus trichocarpa*. Genes labeled in red represent putative *Amborella trichopoda* monolignol biosynthesis genes.

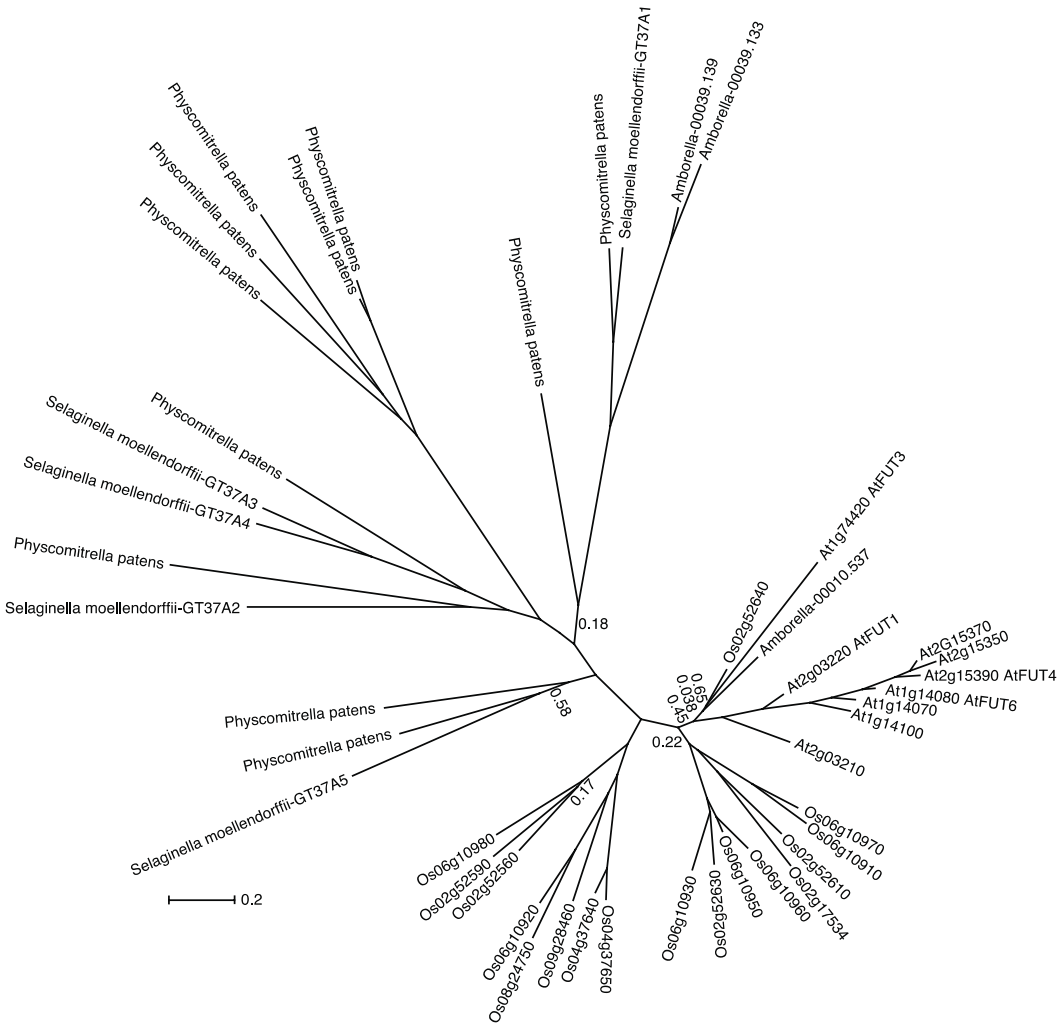


Figure S30. Phylogenetic tree of plant members of CAZy *GT37*. Orthologs in *Amborella* are present in the clade of *GT37* that also contains *Physcomitrella* and *Selaginella GT37*. A second clade contains an *Amborella* gene and a single ortholog from *Arabidopsis* and *Oryza GT37*. *Amborella* does not have orthologs in any of the clades of *GT37* that contain multiple *Arabidopsis* and *Oryza* members.

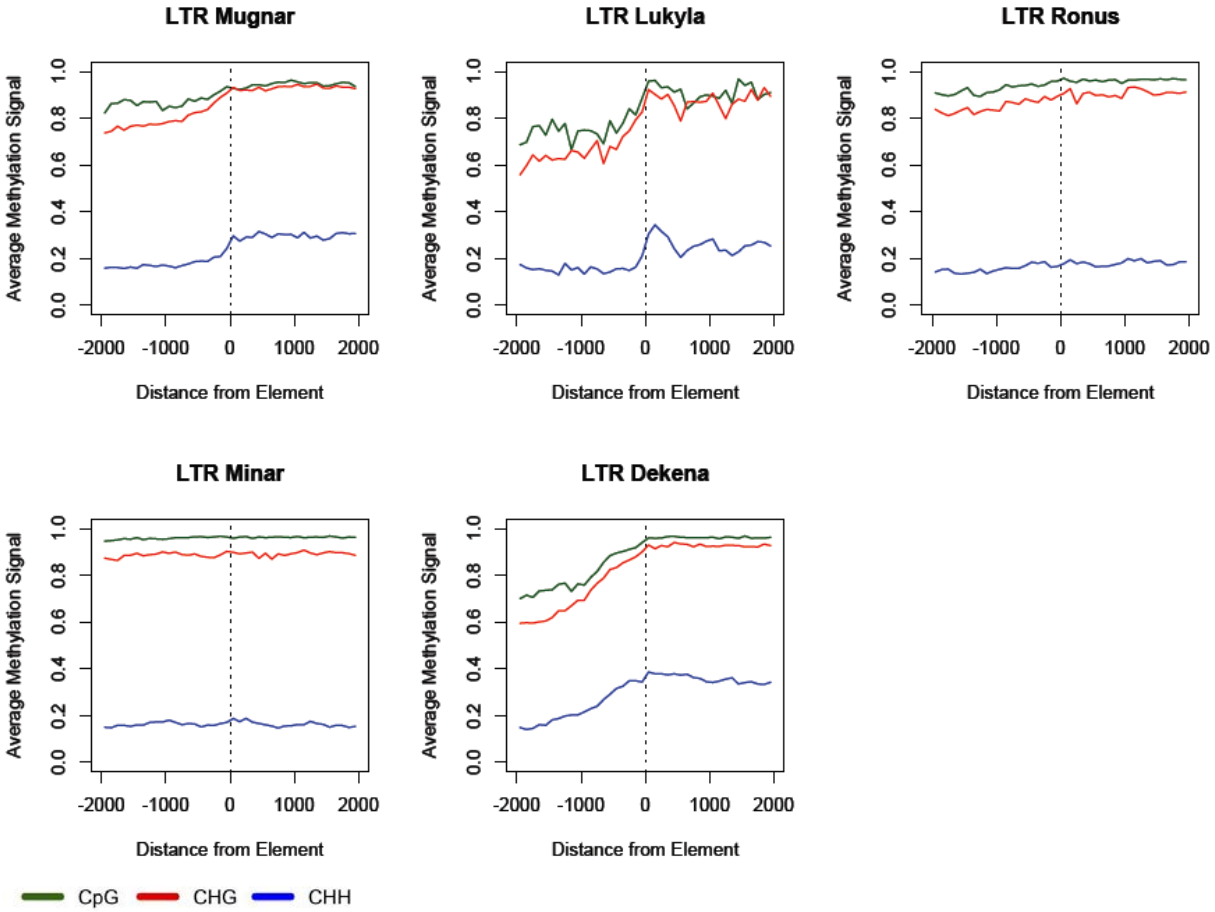


Figure S31. Frequency of cytosine methylation in CpG (green), CHG (red), and CHH (blue) contexts indicate that background and body methylation levels vary among the five most abundant LTR-retrotransposons families.

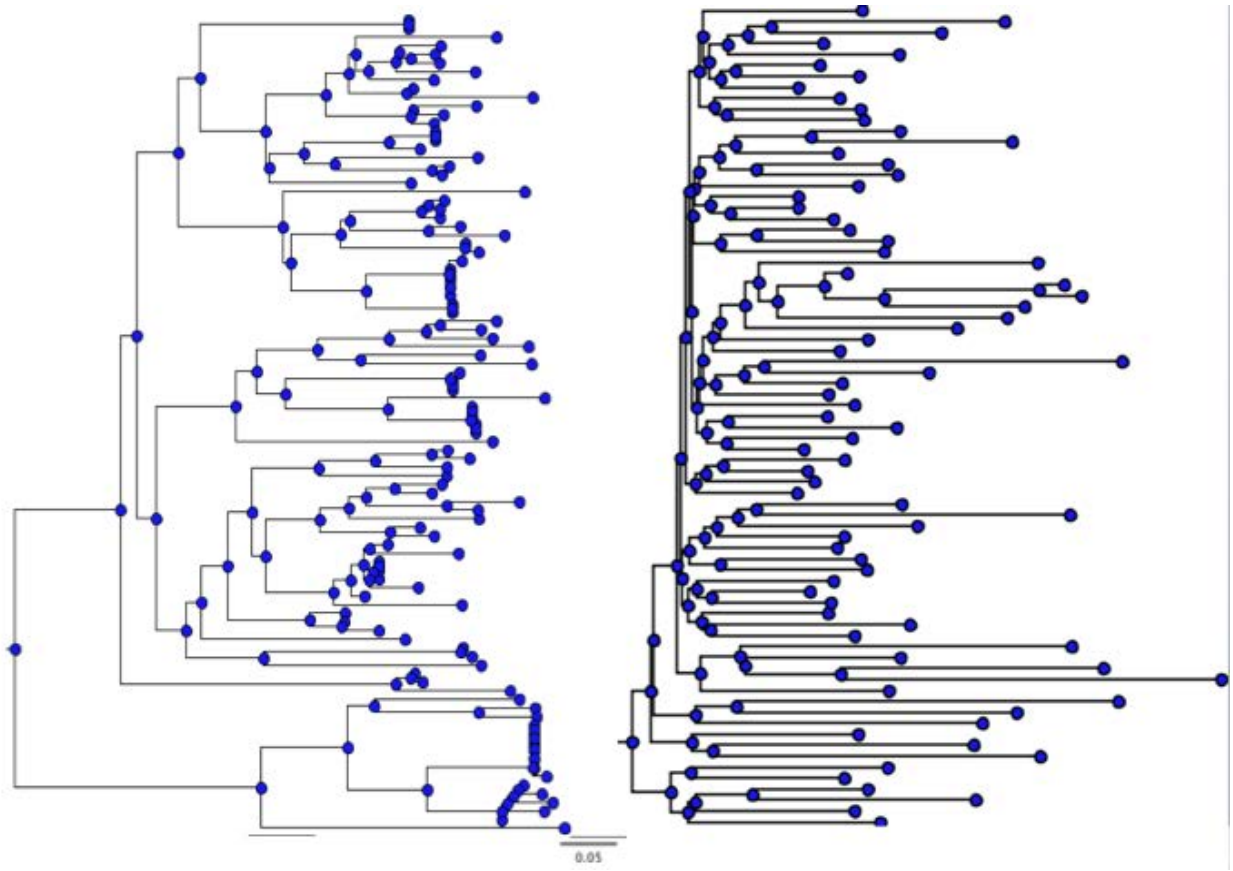


Figure S32. Comparison of phylograms for PIF/Harbinger elements from the rice (left) and *Amborella* (right) genomes reveal fundamentally different diversification rates. Blue circles at tips represent annotated elements and circles at nodes represent ancestral elements. Branch length scales are the same for both trees. All rice PIF /Harbinger elements and one clade in the *Amborella* PIF /Harbinger TE family are shown

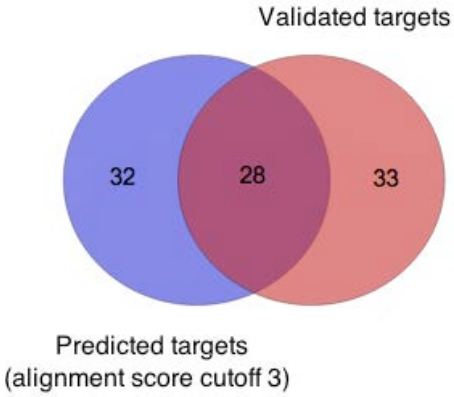


Figure S33. Comparison between predicted (high confidence) and degradome-validated *Amborella* miRNA targets.

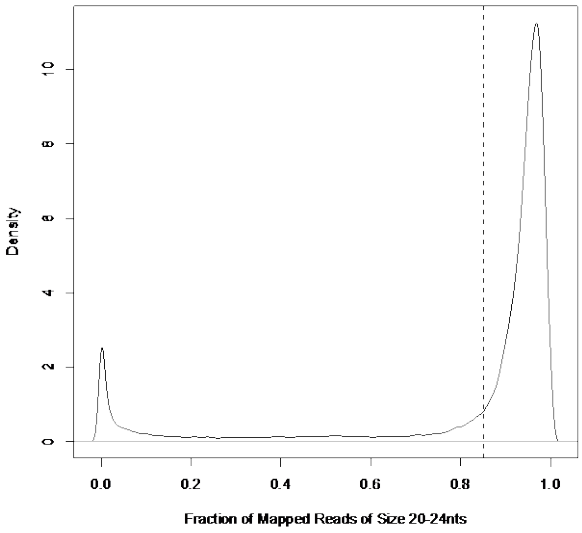


Figure S34. Separation of non-Dicer-derived from Dicer-derived small RNA genes. Kernel density plot of the fraction of all mapped reads within each cluster that were 20-24 nts in length. Dotted line indicates the cutoff value of 85%. n=71,722, Bandwidth = 0.006337.

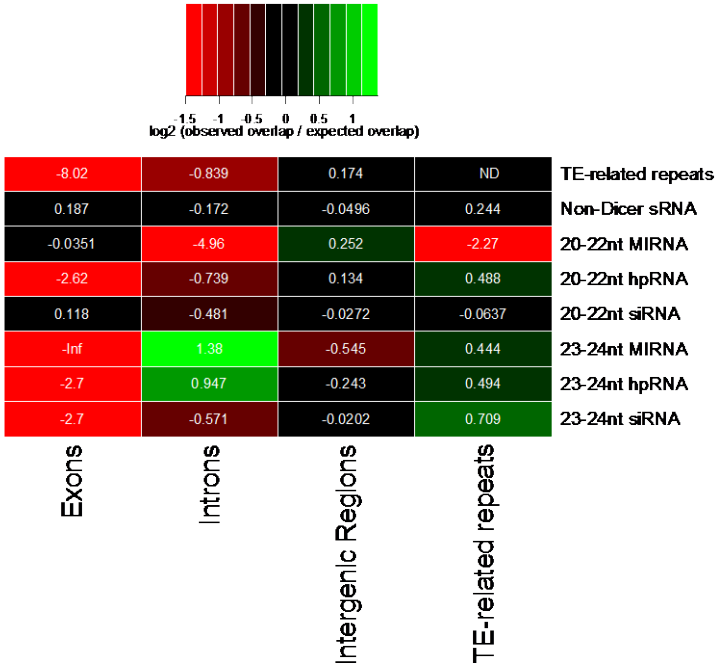


Figure S35. Analysis of overlaps between small RNA loci and other genomic features. Heatmap showing the $\log_2$ (observed overlapped bases / expected overlapped bases) for each of the pairwise comparisons shown. Cell values are shown in white text.

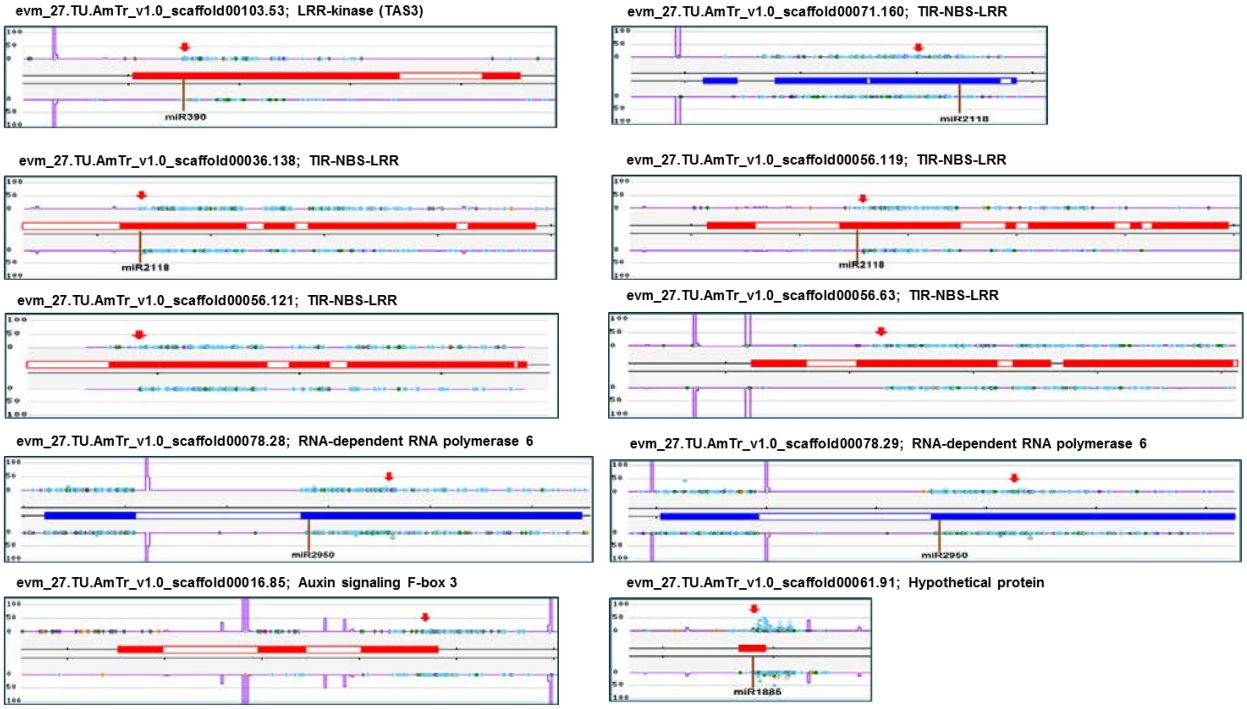


Figure S36. Annotated, protein-coding genes generate phased small RNAs in *Amborella*. Each panel shows a gene in the *Amborella* genome found to generate phased, secondary small RNAs ("PHAS" loci). Below each gene identifier (and predicted protein type) is a screenshot from the Meyers Lab website (http://mpss.udel.edu) showing the small RNA data aligned with the genomic region encoding the gene. Blue dots, the most significant class of phased secondary siRNAs, are 21-nt small RNAs, dark green dots are 22-nt small RNAs, and other colors indicate other small RNAs as indicated by the legend in the upper right corner of each image. Abundances are indicated in transcripts per million (TPM) and marked on the Y-axis. Blue or red boxes are annotated exons interrupted by introns (white boxes). Purple lines indicate the repetitiveness of the genome based on a simple k-mer frequency; yellow, pink, or orange shading indicates DNA transposons, retrotransposons, or inverted repeats (identified by RepeatMasker); the red arrow indicates the position of the highest phasing score. The locations corresponding to 22-nt miRNA triggers are also indicated by the brown line with a miRNA trigger name; PHAS loci lacking the trigger names had no miRNA triggers identified by the analysis pipeline.

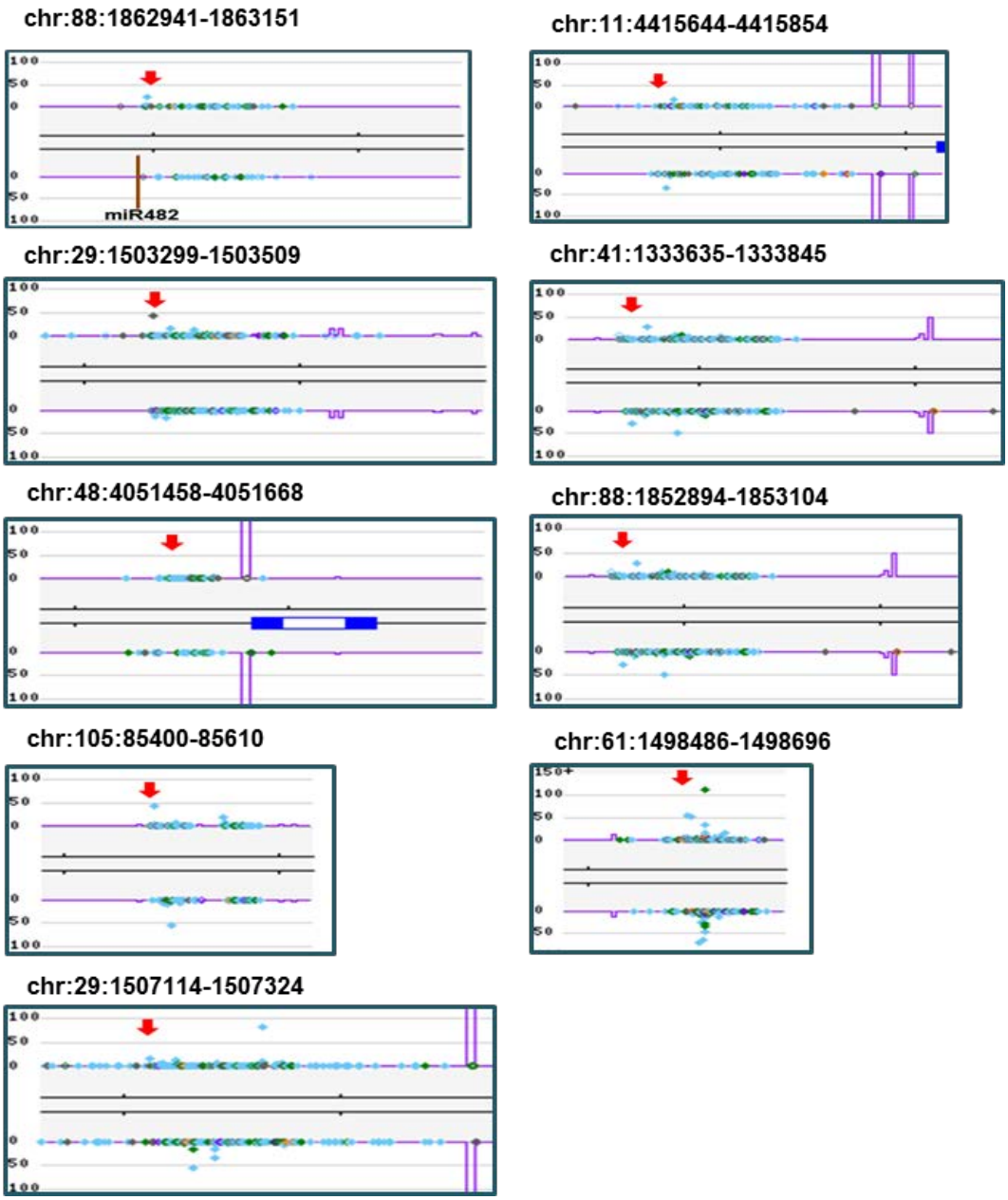


Figure S37. Non-coding PHAS loci in the *Amborella* genome. These panels show loci in the *Amborella* genome generating phased, secondary siRNAs at which no protein-coding genes were annotated. Thus, these are candidate non-coding precursors. In most cases, no miRNA triggers were identified. The data in each panel are as described in the legend for Figure S36.

```
UUUCCGAAUCCGCCUAUACCUG   -- atr-miR8558a
UUUCCGAUCCCGCCCAUGCCGU   -- atr-miR8558b
UUCCCGAUUCCGCCCAUGCCAU   -- atr-Cluster_15097 (dominant sequence)
UUUCCGAUUCCCCCCCAUGCCAU  -- atr-Cluster_15098 (dominant sequence)
UUCCCGAUGCCUCCUAUUCCUA   -- osa-miR2118c
UUUCCAAUUCCACCCAUUCCUA   -- sly-miR482
UUACCGAUUCCACCCAUUCCUA   -- mtr-miR2118
```

Figure S38. *Amborella* members of the sequence-diverse miR482/miR2118 superfamily. Alignments of the indicated mature miRNAs / small RNAs (5'-3'). Nucleotides in red differ from the consensus at that position.

Eudicotyledons
Asterids
Rosids
Poaceae
Ferns
Gymnosperms
Basalc Angiosperms
Solanaceae
Fabids
Malvids
Triticeae
Andropogoneae
Paniceae
miRNAc family
miR023
miR02Q
miR039
miR038
miR033
miR03-
miR03,
miR03Q
miR0-0
miR0-4
miR70Q
miR7Q9
miR7Q7
miR7Q8
miR7Q2
miR7Q3
miR7Q
miR7Q
miR272
miR8--
miR,4,
miR83
miR37
miR3Q
miR-9
miR-0
miR0
miR4
miR7
miR8
miR2
miR3
miR-
miR,
miRQ
miR09
miR00
miR04
miR07
miR08
miR02
miR03
miR0-
miR0,
miR0Q
miR49
miR40
miR44
miR47
miR48
miR42
miR43
miR4-
miR4,
miR4Q
miR79
miR70
miR74
miR77
miR78
miR72
miR73
miR7-
miR7,
miR7Q
miR89
miR80
miR84
miR87
miR88
miR82
miR8-
miR8,
miR8Q
miR29
miR20
miR24
miR27
miR28
miR22
miR23
miR2-
miR2,
miR2Q
miR39
miR30
miR34
miR38
miR32
miR33
miR3-
miR3,
miR-4
miR-7
0.9999RPM
0.999RPM
0.99RPM
0.01RPM
0.1RPM

Figure S39. Conservation patterns of 90 *Amborella* miRNA families across major representatives of land plants. Heatmap indicating miRNA abundance for the 90 *Amborella* miRNA families (in units of reads per million) across a diverse panel of plants. (NCBI GEO GSE28755). Families are presented in three groupings based on conservation pattern: Widespread (top group), some conservation but not universal (middle group), and *Amborella*-specific (bottom group). Plant name abbreviations: MQU *Marsilea quadrifolia*, CRU *Cycas rumphii*, GBI *Ginkgo biloba*, PAB *Picea abies*, AFI *Aristolochia fimbriata*, ATR *Amborella trichopoda*, NAD *Nuphar advena*, PAM *Persea americana*, CAN *Capsicum annum*, NTA *Nicotiana tabacum*, PHY *Petunia hybrida*, SLY *Solanum lycopersicum*, STU *Solanum tuberosum*, LSA *Lactuca sativa*, MGU *Mimulus guttatus*, CMA *Cucurbita maxima*, PTR *Populus trichocarpa*, CPA *Carica papaya*, CSI *Citrus sinensis*, GAR *Gossypium arboreum*, SLA *Silene latifolia*, VVI *Vitis vinifera*, HVU *Hordeum vulgare*, TAE *Triticum aestivum*, MGI *Miscanthus x gigantus*, SBI *Sorghum bicolor*, ZMA *Zea mays*, PVI *Panicum virgatum*, SIT *Setaria italica*, MAC *Musa acuminata*, ZMR *Zostera marina*.

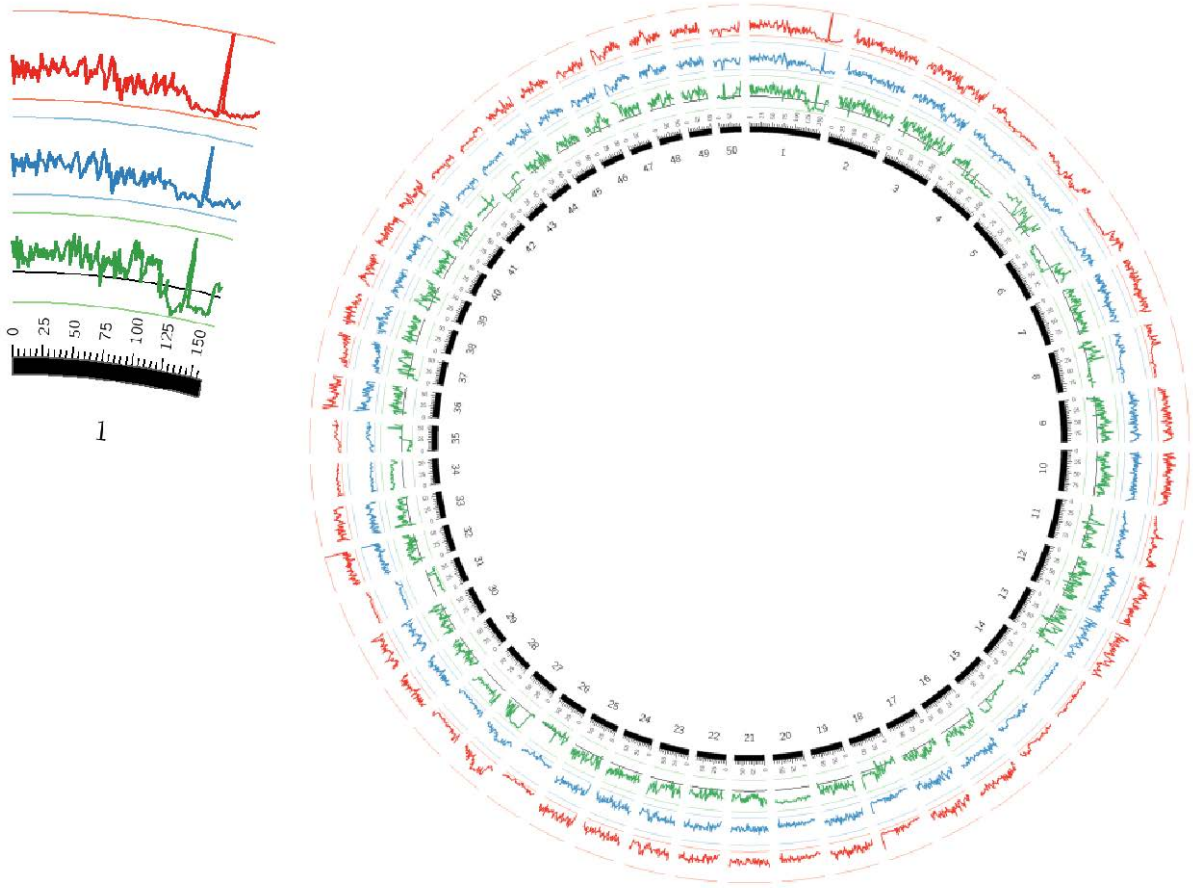


Figure S40. Distribution of genetic diversity across the genome. Density plot of nucleotide diversity (π), Watterson estimator (expected heterozygosity, θ_w), and Tajima's *D* statistics for the 50 largest scaffolds. Red lines indicate nucleotide diversity (π); the two red bounding lines represent maximum and minimum values, max=0.08, min=0. Blue lines represent θ_w; the two blue bounding lines represent maximum and minimum values, max=0.006, min=0. Green lines indicate Tajima's *D* statistics, the two green bounding lines represent maximum and minimum values, max=3.24, min=-1.67; the black line represents a value of Tajima's *D* of 0. Each scaffold unit represents 100,000 bp. Inset shows close-up of scaffold 1.

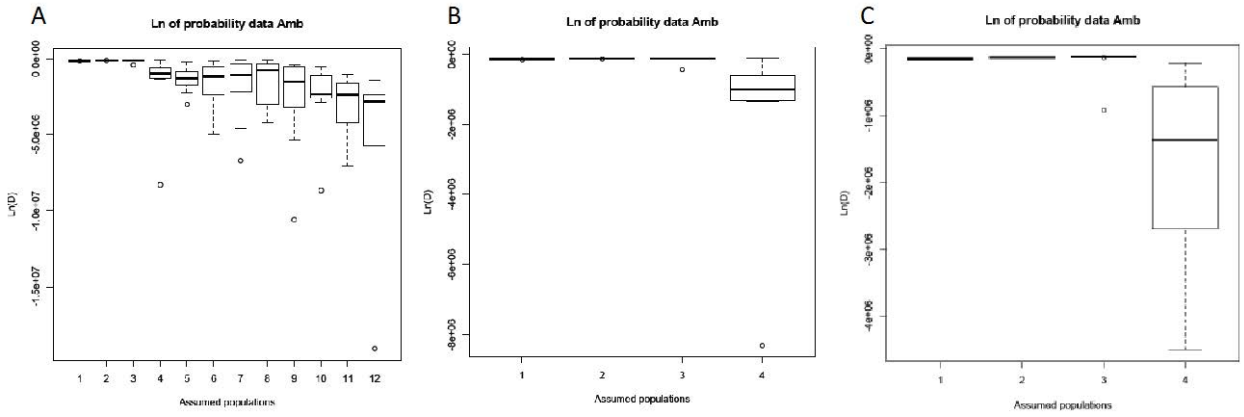


Figure S41. Selection of the number of clusters represented in the data for analyses containing only the 12 individuals sampled from wild populations. (A) Ln probability of the data for different numbers of populations (K) ranging from 1 to 12. (B) The Ln probability for K=1 to 4 in the first round of analyses. (C) The Ln probability of K=1 to 4 in the second round of analyses.

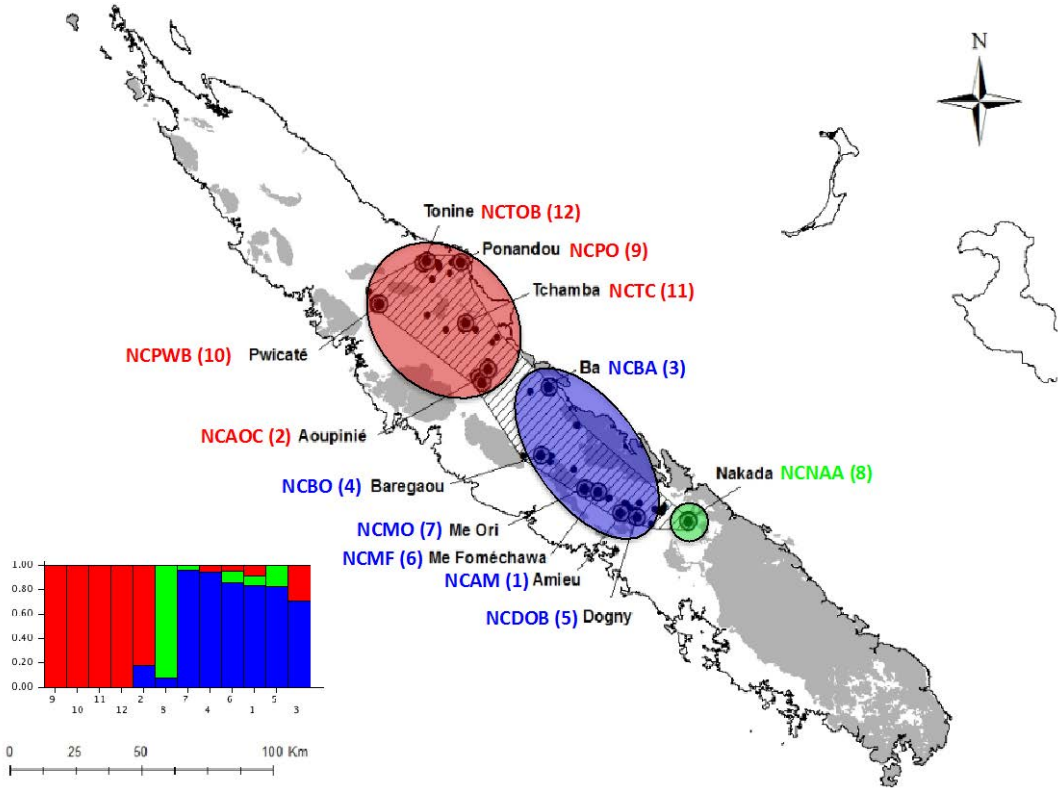


Figure S42. Population structure inferred using only 12 individuals sampled from wild populations of New Caledonia. Three significant clusters were recovered showing latitudinal geographic structure. When individuals in cultivation were sampled (Santa Cruz and Bonn), the

number of clusters increased to four, separating Mé Ori and Mé Foméchawa from the central cluster, in agreement with the microsatellite analyses (Fig. 6 of the main text).

Supplementary Tables:

Table S1. Sequencing and quality filtering statistics.

	(P)lates/ (L)anes	**Reads (millions)**		**Bases (billions)**		**Mean length**		**Insert size**
		Raw	Filtered	Raw	Filtered	Raw	Filtered	
454 FLX	28.5 P	38.8	27.6	14.2	10.5	365	381	N/A
454 FLX+	17 P	22.4	18.3	12.0	10.2	533	557	N/A
454 FLX mate-pairs	5 P	6.16	1.92	2.14	0.699	348	364	11 kb
Illumina mate-pairs	1 L	195	17.7	19.7	1.68	101	95	3 kb
Sanger BAC ends	N/A	0.070	0.0639	.048	.044	695	689	200 kb

Table S2. Whole genome shotgun *de novo* assembly statistics.

	Number	**Total length (Mb)**	**N50**	**L50 (kb)**	**N90**	**L90(kb)**	**Mean Length**
Contigs	43,234	668	6,448	29.4	7108	24.0	15,456
Scaffolds	5,745	706	50	4,900	155	1,200	122,947

Table S3. Illumina GAIIx RNA-Seq transcriptome data used in Cufflinks alignments and Trinity assembly.

RNA Source	**Reads**
Leaf Shoot Non-Normalized	26,293,810
Apical Meristem with Young Leaves	75,096,962
Premeiotic Female Bud	77,484,460
Whole Plant Normalized	149,743,562
Total	328,618,794

Table S4. Summary of *de novo* Trinity constructed contigs of *Amborella* RNA-Seq data.

Minimum contig length	201 bp
1st quartile	283 bp
Median contig length	470 bp
Mean contig length	1052 bp
3rd quartile	1278 bp
Maximum contig length	16,820 bp

Table S5. Summary statistics for the PASA aligned *Amborella* gene training set.

Average transcript length	5243 bp
Median transcript length	3551 bp
Average coding length	1419 bp
Median coding length	1,260 bp
Total length of transcript sequence	9,029,078 bp
Total length of coding sequence	2,444,529 bp
Total exons	10704 (6.22/gene)
Average exon length	229 bp
Median exon length	128 bp
Number of initial exons	1,368 (excludes single exon genes)
Average length of initial exons	313 bp
Number of internal exons	7533
Average Length of internal exon	155 bp
Average length of terminal exons	314 bp

Table S6. Four-fold cross validation results for *Amborella* trained TWINSCAN_EST compared to FGENESH (*272*) run with the Monocot parameter set.

		TWINSCAN_EST	**FGENESH**
Gene	Sensitivity	53.60%	28.51%
Gene	Specificity	38.58%	17.06%
Exon	Sensitivity	90.71%	75.93%
Exon	Specificity	77.16%	53.44%
Nucleotide	Sensitivity	98.70%	96.72%
Nucleotide	Specificity	85.00%	74.36%

Table S7. Number of alternatively spliced (AS) genes/isoforms identified at different junction support levels.

Junction Support	**2 Reads**	**3 Reads**	**5 Reads**	**10 Reads**	**20 Reads**	**50 Reads**
Genes (>=1 predicted AS)	6172	6106	5997	5656	5001	3617
AS Isoforms	14323	14146	13835	12899	10955	6813
Isoforms/Gene	2.32	2.32	2.31	2.28	2.19	2.88

Table S8. Transfer RNAs identified in the *Amborella* genome assembly.

Amino Acid	tRNA Copy Number
Ala	24
Arg	53
Asn	33
Asp	19
Cys	21
Gln	25
Glu	30
Gly	33
His	19
Ile	23
Leu	41
Lys	33
Met	40
Phe	25
Pro	35
SeC	4
Ser	72
Thr	54
Trp	21
Tyr	20
Val	30

Table S9. Transfer RNA triplet counts and codon usage in annotated protein coding genes in *Amborella*. Relative synonymous codon usage is abbreviated RSCU. (See large tables file.)

Table S10. Gene pairs identified on syntenic blocks within the *Amborella* genome, with orthogroup assignments and whether phylogenomic evidence supports a pre-angiosperm gene duplication. (See large tables file.)

Table S11. Duplications in curated syntenic blocks based on expanded orthogroup phylogenies. Data shown are syntenic block identity, orthogroup membership (or singleton status) of the two syntologous *Amborella* genes, whether trees were considered valid to score, and the detection of angiosperm-wide duplications supported with bootstrap values of at least 50% or 80% For additional information, see supplementary text section 5. (See large tables file.)

Table S12. Orthogroup classification summary. Orthogroup membership is summarized for 685,814 annotated protein coding genes in 19 sequenced plant genomes. Detailed annotation information is provided in Table S13. (See large tables file.)

Table S13. Details of orthogroup membership and evolutionary history for genes from 19 sequenced plant genomes, with functional annotations and the inferred orthogroup membership for ancestral nodes. Data shown are: orthogroup ID (A); membership counts for sequenced genomes (B-T); inferred ancestral gene conent (U-AL); number of inferred gains, losses, expansions, and reductions for each orthogroup (AM-AP); number of EST contigs assigned to each orthogroup for the basal angiosperms *Nuphar*, *Aristolochia*, *Persea*, *Liriodendron* (AAGP EST, AQ); number of EST contigs assigned to each orthogroup for gymnosperms (Gymno EST, AR); inferred orthogroups originating with angiosperms (AS) and with seedplants (AT); ancestral tracheophyte orthogroups (AU) and orthogroups derived within angiosperms (AV); notes on orthogroup evolution (AW); phylogenomic evidence for duplication in the ancestral angiosperm with 50% (AX) or 80% (AY) bootstrap support; super orthogroup ID with inflation value 3.5 (AZ); annotation information from *Arabidopsis thalliana* (BA-BB); Pfam domans detected in the orthogroup with proportion of members having significant hits (BC); GO annotations associated with Pfam domains (BD-BF); known floral genes in *Arabidopsis thaliana* (BG-BJ). (See large tables file.)

Table S14. Wagner parsimony reconstruction (1.2 gain penalty) of orthogroup gains, losses, expansions, and contractions. N/A (Not applicable) is given for outroup lineages where ancestral genomes cannot be inferred. (See large tables file.)

Table S15. Maximum likelihood reconstruction of ancestral orthogroup gains, losses, expansions, and contractions. N/A (Not applicable) is given for outroup lineages where ancestral genomes cannot be inferred. (See large tables file.)

Table S16. Gene Ontology (GO) enrichment for orthogroups that are new in angiosperms and new in seedplants. GO classification of *Arabidopsis* genes in orthogroups first appearing with angiosperms or with seed plants were compared to GO classifications of the entire *Arabidopsis* genome background using AGRIGO analysis toolkit (*141*). Table provides: GO categories (Columns A-C); false discovery rate (FDR) corrected probability test that the GO category frequency is equal to the genome background for orthogroups gained with angiosperms (D) and seedplants (F), and their associated N-fold enrichments relative to background representation (E, G); statistics for orthogroups gained with angiosperms in comparison to the background (H-P); and seed plant node relative to background (Q-Y). (See large tables file.)

Table S17. Orthogroup origin and duplication history of 232 known floral genes from *Arabidopsis*. Data shown are Gene ID and symbol (Columns A and B), orthogroup membership (C), phylogenomic support for angiosperm wide (epsilon) duplication (D), seed plant wide duplication (E), inferred orthogroup gain in angiosperm (F), or seed plants (G), ancestral in vascular plants (G), or originating in a derived angiosprm lineage (I), and overall assessment of orthogroup origin (J). Also shown are superorthogroup assignment (K), and functional annotations for the orthogroup (L to Q). (See large tables file.)

Table S18. Diverse origins of floral genes. Genes with known roles in flower timing and development, downstreamB-class target genes, and genes with expression specific to various floral structures in *Arabidopsis* were classified to orthogroup and origin of the orthogroup was inferred as described. (See large tables file.)

Table S19. MADS-box genes in *Amborella trichopoda*.

Putative subfamily name		Gene name	Protein ID	Location	Cufflinks coverage	Source
type II	*AP1*	*Am.tr.AP1*	scaffold00047.105	scaffold00047_3064751_3090556	full coverage, 200-3000	this study
	AP3	*Am.tr.AP3.1*	scaffold00001.225	scaffold00001_4348140_4349431	full coverage, 1500-3800	(*174*); EST
	AP3	*Am.tr.AP3.2*	scaffold00066.97	scaffold00066_1372288_1376738	full coverage, 1000-3000	(*174*); EST
	PI	*Am.tr.PI.1*	scaffold00017.226	scaffold00017_5901058_5901961	full coverage, 2000-6000	(*174*);EST
	PI	*Am.tr.PI.2*	scaffold00089.36	scaffold00089_907931_909820	full coverage, 100-190	this study
	AG	*Am.tr.AG*	scaffold00021.296	scaffold00021_6744499_6778414	full coverage, 800-1700	(*175*); EST
	STK	*Am.tr.STK*	scaffold00071.203	scaffold00071_3440534_3462334	full coverage, 350-721	EST
	AGL2	*Am.tr.AGL2*	scaffold00047.121	scaffold00047_3344579_3375581	full coverage, 3000-6000	(*175*); EST
	AGL9	*Am.tr.AGL9*	scaffold00013.53	scaffold00013_1262781_1295413	full coverage, 1000-2400	(*176*); EST
	AGL6	*Am.tr.AGL6*	scaffold00001.413	scaffold00001_8828935_8888309	full coverage, 1900-4000	(*175*); EST
	AGL12	*Am.tr.AGL12*	scaffold00071.216	scaffold00071_3650392_3652966	partial coverage, 40-89	this study
	AGL15	*Am.tr.AGL15*	scaffold00053.185	scaffold00053_4857033_4870850	full coverage, 50-185	this study
	TM3	*Am.tr.TM3*	scaffold00001.409	scaffold00001_8708320_8773009	full coverage, 1000-2100	this study
	STMADS11	*Am.tr.StM11*	scaffold00127.17	scaffold00127_599439_654192	full coverage, 1500-4000	EST
	OsMADS32	*Am.tr.OsM32*	scaffold00001.226	scaffold00001_4397899_4400065	full coverage, 1000-2700	this study

	TM8	*Am.tr.TM8.1*	scaffold00013.57	scaffold00013_1472673_1472858	full coverage, 900-1900	this study
	TM8	*Am.tr.TM8.2*	scaffold00013.60	scaffold00013_1578058_1662108	full coverage, 50-320	EST
	ANR1	*Am.tr.ANR1.1*	scaffold00109.2	scaffold00109_29724_30613	partial coverage, 30-64	this study
	ANR1	*Am.tr.ANR1.2*	scaffold00046.134	scaffold00046_3995369_4044817	full coverage, 100-660	this study
	GMM13	*Am.tr.GMM13.1*	scaffold00001.461	scaffold00001_10158193_10163424	full coverage, 10-20	this study
	GMM13	*Am.tr.GMM13.2*	scaffold00002.466	scaffold00002_8048974_8049195	partial coverage, 1-2	this study
	MIKC*-P	*Am.tr.MP*	scaffold00010.504	scaffold00010_8389805_8396126	full coverage, 5-179	this study
	MIKC*-S	*Am.tr.MS*	scaffold00010.217	scaffold00010_4213059_4213797	partial coverage, 5-15	this study
type I	α	*Am.tr.MA1*	scaffold00140.17	scaffold00140_402687_403790	full coverage, 5-19	this study
	α	*Am.tr.MA2*	scaffold00159.17	scaffold00159_495293_495928	partial coverage, 1-4	this study
	α	*Am.tr.MA3*	scaffold00159.19	scaffold00159_504687_505322	full coverage, 1-17	this study
	α	*Am.tr.MA4*	scaffold00159.18	scaffold00159_499712_500182	partial coverage, 1-6	this study
	α	*Am.tr.MA5*	scaffold00050.92	scaffold00050_4694781_4695569	full coverage, 5-34	this study
	α	*Am.tr.MA6*	scaffold00025.385	scaffold00025_5897664_5898831	partial coverage, 1-3	this study
	β	*Am.tr.MB1*	scaffold00025.268	scaffold00025_3949692_3950678	full coverage, 2-36	this study
	β	*Am.tr.MB2*	scaffold00025.271	scaffold00025_3987163_3987890	unavailable	this study
	β	*Am.tr.MB3*	scaffold00176.19	scaffold00176_386806_3	full coverage, 5-72	this study

			87873		
β	*Am.tr.MB4*	scaffold00022.374	scaffold00022_6148497_6149633	unavailable	this study
β	*Am.tr.MB5*	scaffold00116.30	scaffold00116_1677969_1678568	full coverage, 2-50	this study
β	*Am.tr.MB6*	scaffold00017.X#	scaffold00017_3178608_3179039	full coverage, 2-16	this study
γ	*Am.tr.MC*	scaffold00095.150	scaffold00095_2797522_2798208	partial coverage, 1-5	this study

#: New predicted gene

Table S20. Copy numbers of MADS-box genes in four representive species.

Species	Type I	Type II	Total	Reference
Amborella trichopoda	13	23	36	this study
Arabidopsis thaliana	57	50	107	(*273*)
	60	45	105	this study
Oryza sativa	32	43	75	(*45*)
	27	37	64	this study
Populus trichocarpa	41	64	105	(*47*)
	39	45	84	this study

Table S21. 11S globulins identified from 2D-electrophoresis (2DE) individual spots from *Amborella trichopoda* seed protein extracts. The protein spots were analysed by LC-MS/MS on the PAPPSO plateform (Benoit Valot, Thierry Balliau, Michel Zivy, INRA Moulon, France; http://pappso.inra.fr). Based on the spectrum generated, proteins were identified using the X-Tandem software. "Spot ID" relates to the protein position on the 2DE reference map (see Fig. S24); "Description" was taken from the Amborella EVM 27 Predicted Proteins (http://www.amborella.org); The "log (E value)" is a statistical parameter that represents the number of peptides present at random in the database. It was calculated by the product of the Evalue of unique peptides identified in the protein spot (*194*). "Coverage" refers to the recovery rate of the protein by the identified peptides, expressed in %. "Spectra" corresponds to the number of spectra that allowed the protein identification from 2DE protein spots. "Uniques" refers to the number of unique peptides (different sequences) that led to the protein identification. "Peptide sequence" displays identified peptides for each protein spot. (See large tables file.)

Table S22. Seed plant 11S-globulin genes included in the phylogenetic analysis of Fig. S27.

Source Species	GenBank ID or *Amborella* Gene model ID
Actinidia chinensis	ABB77213.1
Amaranthus hypochondriacus	CAA57633.1
Amborella trichopoda	evm_27. model.AmTr_v1.0_scaffold00067.11
Amborella trichopoda	evm_27.model.AmTr_v1.0_scaffold00067.12
Amborella trichopoda	evm_27.model.AmTr_v1.0_scaffold00067.7
Anacardium occidentale	AAN76862.1
Arabidopsis thaliana	NP_199225.1
Arabidopsis thaliana	P15456
Arabidopsis thaliana	Q9ZWA9
Arachis hypogaea	AAG01363.1
Asarum europaeum	CAA64761.1
Asarum europaeum	CAA64762.1
Avena sativa	P14812.1
Bertholletia excelsa	AAO38859.1
Brassica napus	AAA32988.1
Calocedrus decurrens	CAA64787.1
Calocedrus decurrens	CAA64788.1
Carya illinoinensis	ABW86979.1
Castanea crenata	AAM93194.1
Chenopodium quinoa	AAS67036.1
Chenopodium quinoa	AAS67037.1
Cicer arietinum	Q9SMJ4.1
Citrus sinensis	AAB52963.1
Coffea arabica	AAC61983.1
Corylus avellana	AAL73404.1
Cryptomeria japonica	CAA64789.1
Cucurbita maxima	P13744.1
Dioscorea caucasia	CAA64763.1
Elaeis guineensis	AAF69015.1
Ephedra gerardiana	CAA90640.1
Fagopyrum esculentum	Q9XFM4.1
Ficus pumila var. awkeotsang	ABK80755.1
Ginkgo biloba	CAA90641.1
Glycine max	CAA55977.1
Glycine max	CAA60533.1
Gnetum gnemon	CAA90642.1
Gossypium hirsutum	AAA33072.1
Helianthus annuus	P19084.1
Juglans regia	AAW29810.1
Lotus japonicus	CAR78996.1
Lupinus albus	CAI83773.2
Magnolia salicifolia	CAA57848.1
Metasequoia glyptostroboides	CAA64791.1

Oryza sativa	CAA33838.1
Picea glauca	CAA44874.1
Pinus strobus	CAA77568.1
Pisum sativum	CAA10722.1
Pisum sativum	P05693.1
Populus trichocarpa	XP_002329472.1
Prunus dulcis	CAA55009.1
Pseudotsuga menziesii	AAA68981.1
Quercus robur	CAA67879.1
Raphanus sativus	Q02498.1
Ricinus communis	XP_002524198.1
Sesamum indicum	AAK15087.1
Sinapis alba	AAX77383.1
Triticum aestivum	AAB27108.2
Vernicia fordii	AFJ04523.1
Vicia faba	P16078.1
Vicia faba var. minor	CAA38758.1
Vicia narbonensis	CAA86824.1
Vicia sativa	CAA83674.1
Vicia sativa	CAA83677.1
Welwitschia mirabilis	CAA90643.1
Zea mays	AAL16994.1.

Table S23. The terpene synthase gene (*TPS*) family in *Amborella*.

Gene name	Gene ID	Scaffold location	Strand[a]	Protein size	Manual curation	*TPS* Sub-family	Ortho ID	Super ortho ID
TPS1	EVM_27.prediction.AmTr_v1.0_scaffold00002.146	2846574-2847309	-	125	no	ND[b]	Singleton	NA
TPS2	EVM_27.prediction.AmTr_v1.0_scaffold00003.311	6882485-6899747	-	550	no	G	1412	46
TPS3	EVM_27.prediction.AmTr_v1.0_scaffold00003.314	7042559-7051342	+	645	yes	e/f	Singleton	NA
TPS4	EVM_27.prediction.AmTr_v1.0_scaffold00007.400	9315947-9318970	-	515	no	B	607	46
TPS5	EVM_27.prediction.AmTr_v1.0_scaffold00007.401	9351307-9356611	-	410	no	B	607	46
TPS6	EVM_27.prediction.AmTr_v1.0_scaffold00011.117	5072736-5078333	-	681	yes	X	Singleton	NA
TPS7	EVM_27.prediction.AmTr_v1.0_scaffold00017.276	6908350-6909058	+	125	no	ND	Singleton	NA
TPS8	EVM_27.prediction.AmTr_v1.0_scaffold00019.25	434115-435399	-	385	no	X	Singleton	NA
TPS9	EVM_27.prediction.AmTr_v1.0_scaffold00032.208	4197672-4207107	+	908	no	e/f	1183	457
TPS10	EVM_27.prediction.AmTr_v1.0_scaffold00032.221	4410138-4413006	-	477	no	B	607	46
TPS11	EVM_27.prediction.AmTr_v1.0_scaffold00032.223	4441679-4444835	+	510	no	B	607	46
TPS12	EVM_27.prediction.AmTr_v1.0_scaffold00032.237	4658120-4661138	+	487	no	B	607	46
TPS13	EVM_27.prediction.AmTr_v1.0_scaffold00033.268	5812356-5812694	+	154	no	ND	Singleton	NA
TPS14	EVM_27.prediction.AmTr_v1.0_scaffold00033.270	5885131-5887171	+	279	no	B	24423	46

TPS15	EVM_27.prediction.AmTr_v1.0_scaffold00043.44	2894221-2894789	+	130	no	ND	Singleton	NA
TPS16	EVM_27.prediction.AmTr_v1.0_scaffold00043.45	2931510-2934181	-	483	no	G	1412	46
TPS17	EVM_27.prediction.AmTr_v1.0_scaffold00043.48	3109040-3109489	+	118	no	ND	Singleton	NA
TPS18	EVM_27.prediction.AmTr_v1.0_scaffold00043.49	3109772-3112167	+	268	no	G	Singleton	NA
TPS19	EVM_27.prediction.AmTr_v1.0_scaffold00043.50	3130881-3134338	-	588	no	G	1412	46
TPS20	EVM_27.prediction.AmTr_v1.0_scaffold00066.146	2102877-2198290	+	565	no	G	14740	46
TPS21	EVM_27.prediction.AmTr_v1.0_scaffold00066.195	2692114-2703023	+	727	no	e/f	8322	457
TPS22	EVM_27.prediction.AmTr_v1.0_scaffold00071.3	24725-28339	+	243	yes	X	Singleton	NA
TPS23	EVM_27.prediction.AmTr_v1.0_scaffold00103.20	419901-432053	-	795	no	C	716	457
TPS24	EVM_27.prediction.AmTr_v1.0_scaffold00169.1	570-2671	+	301	no	B	24423	46
TPS25	EVM_27.prediction.AmTr_v1.0_scaffold00169.3	38773-42048	+	403	no	B	24423	46
TPS26	EVM_27.prediction.AmTr_v1.0_scaffold00203.18	305786-306256	+	156	no	ND	20705	46
TPS27	EVM_27.prediction.AmTr_v1.0_scaffold00203.19	310270-312073	+	481	no	X	Singleton	NA
TPS28	EVM_27.prediction.AmTr_v1.0_scaffold00203.20	315355-315711	+	693	no	X	32600	457
TPS29	EVM_27.prediction.AmTr_v1.0_scaffold00260.1	25891-32995	-	638	yes	X	32548	46
TPS30	EVM_27.prediction.AmTr_v1.0_scaffold00352.1	8333-26137	-	371	yes	X	Singleton	NA

TPS31	EVM_27.prediction.AmTr_v1.0_scaffold00713.1	1321-2197	-	245	no	e/f	24374	457
TPS32	EVM_27.prediction.AmTr_v1.0_scaffold01463.1	1922-2490	-	160	no	ND	32612	46
TPS33	EVM_27.prediction.AmTr_v1.0_scaffold01573.1	2545-2876	-	144	no	ND	32600	457
TPS34	EVM_27.prediction.AmTr_v1.0_scaffold04075.1	1899-2467	-	160	no	ND	32612	46

[a] Orientation of open reading frame;
[b] "ND" stands for "not determined".

Table S24. Chemical composition of wood from *Amborella trichopoda.* Percentage is based on oven-dried wood meal weight (mean ± SE, n=2). Monosaccharides were determined by alditol acetate method via gas chromatography.

Chemical components	%
Lignin	
Acid-insoluble lignin	31.8 ± 0.5
Acid-soluble lignin	0.3 ± 0.0
Monosaccharides	
Rhamnose	0.1 ± 0.0
Arabinose	0.9 ± 0.1
Xylose	10.4 ± 0.7
Mannose	5.1 ± 0.3
Galactose	1.5 ± 0.2
Glucose	33.9 ± 0.3
Uronic acid	6.4 ± 0.2
Total	90.8 ± 1.4

Table S25. Lignin subunit composition of wood from *Amborella trichopoda.* Lignin subunit composition was determined by nitrobenzene oxidation (mean ± SE, n=2). H = (*p*-hydroxybenzaldehyde + *p*-hydroxybenzoic acid); V=(vanillin + vanillic acid); S = (syringaldehyde + syringic acid). All acids are in trace amounts. Mole percent is based on total lignin, assuming the average molecular mass of a lignin subunit is 180.

Lignin composition	Mole %
H	0.8 ± 0.0
V	35.9 ± 0.5
S	5.5 ± 0.0
Total yield	42.3 ± 0.5
S/V	0.15
H/V	0.02

Table S26. Taxonomic distribution of plants with lignin chemistry measurements. Plants are woody, except as mentioned. Values given are amounts of H, G (orV), and S lignin, normalized to G (or V) as 1.0. (See large tables file.)

Table S27. Putative monolignol biosynthesis genes in *Amborella trichopoda*. Genes were identified based on the orthologous relationships to the *Arabidopsis* (*212*) and *P. trichocarpa* (*274*) lignin toolbox genes at the protein sequence level.

Gene Family	***Amborella* gene model ID**
PAL (Phenylalanine ammonia-lyase)	scaffold00024.177
	scaffold00024.178
	scaffold00024.181
	scaffold00032.129
	scaffold00148.59
4CL (4-coumaric acid Co-A ligase)	scaffold00019.261
	scaffold00048.51
C4H (Cinnamic acid 4-hydroxylase)	scaffold00025.258
	scaffold00077.91
C3H (4-coumaric acid 3-hydroxylase)	scaffold00040.62
	scaffold00101.79
CAld5H (F5H) (Coniferaldehyde 5-hydroxylase)	scaffold00001.452
HCT (Hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyl transferase)	scaffold00137.5
	scaffold00137.34
	scaffold00727.2
HCT (Hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyl transferase)	scaffold00727.2
CCoAOMT (Caffeoyl CoA 3-*O*- methyltransferase)	scaffold00046.27
CCR (Cinnamoyl-CoA reductase)	scaffold00065.159
COMT (Caffeic acid *O*-methyl-transferase)	scaffold00001.509
CAD (Cinnamyl alcohol dehydrogenase)	scaffold00015.12
CES (Caffeoyl Shikimate Esterase)	scaffold00146.66

Table S28. Glycosyltransferases found in *Amborella*, listed according to CAZy family. Note that GT1 is not included in the table. (See large tables file.)

Table S29. Estimated ages for LTR retrotransposon subfamily clusters. (See large tables file.)

Table S30. MITE DNA transposons in the *Amborella* genome. Classification, count, and total size of hAT, PIF, Mutator-like, and undefined MITE annotations identified by MITE-Hunter after removing hits matching terminal inverted repeats in our *Amborella* repeat database.

Classification	Number	Size (kb)
hAT MITES		
hAT1	871	545
hAT2	56	82
hAT3	726	545
hAT4	212	172
hAT5	106	74
hAT6	281	360
hAT7	116	86
hAT8	324	217
hAT9	455	295
Total hAT MITES	**3147**	**2376**
PIF MITES		
PIF1	370	295
PIF2	335	340
Total PIF MITES	**705**	**635**
MULE MITES		
MULE1	285	266
MULE2	269	216
MULE3	31	33
MULE4	676	446
MULE5	606	274
MULE6	648	365
Total MULE MITES	**2515**	**1600**
Undefined MITES		
unknown1	120	57
unknown2	1650	922
unknown3	528	360
unknown4	27	33
unknown5	37	45
unknown6	62	33
unknown7	22	16
unknown8	276	201
unknown9	267	188
unknown10	907	467
unknown11	507	287
unknown12	31	20
unknown13	1610	987
unknown14	208	123
Total unknown MITES	**6252**	**3739**

Table S31. Overall transposon content in the *Amborella* genome.

Superfamily	Sum Length (Mb)	Feature Count	Average Length (bp)	Percentage of Assembled Genome
MITE (All from MITEHunter)	30.50	138,725	220	4.56
RLC (Copia LTR Retro)	91.18	296,197	307	13.65
RLG (Gypsy LTR Retro)	169.31	186,146	909	25.34
RIL (L1 LINE)	38.18	139,680	273	5.71
RIT (RTE LINE)	12.32	62,634	197	1.84
RLEPRV (Endogenous Pararetrovirus)	18.11	17,878	1,012	2.71
DTA (hAT DNA transposon)	11.03	22,007	501	1.65
DTH (PIF-Harbinger DNA transposon)	0.88	3,312	266	0.13
DTM (Mutator DNA transposon)	4.36	11,365	384	0.65
DHH (Helitron DNA transposon)	6.16	32,327	191	0.92
Protein Coding Gene Exons	25.40	109783	230	3.80

Table S32. Summary of *Amborella* small RNA mapping

Library	Tissue	GEO accession	Number of aligned reads	Total no. of alignments
ATR1	Leaves	GSM712477	3,764,154	18,589,854
Amleaf	Leaves	GSM1024602	15,994,122	69,823,071
AmOFF	Female flowers	GSM1024601	22,052,074	83,507,977

Table S33. Tab-delimited summary of all *Amborella trichopoda* small RNA loci. Table is the output of the ShortStack analysis run. (See large tables file).

Table S34. Summary of *Amborella* small RNA loci.

	Gene Frequencies			Gene Percentages			Total Abundance[a]			Abundance Percentages		
Dicer Call[b]	non-HP/MIR[c]	HP	MIR	non-HP/MIR	HP	MIR	non-HP/MIR	HP	MIR	non-HP/MIR	HP	MIR
N	14,020	1,370	8	19.55%	1.91%	0.01%	26,629,193	2,433,925	4,897	72.57%	6.63%	0.01%
20	48	7	8	0.07%	0.01%	0.01%	41,019	20,355	14,461	0.11%	0.06%	0.04%
21	3,426	520	42	4.78%	0.73%	0.06%	1,190,927	169,245	722,966	3.25%	0.46%	1.97%
22	1,251	183	4	1.74%	0.26%	0.01%	279,615	252,887	72,392	0.76%	0.69%	0.20%
23	371	126	1	0.52%	0.18%	0.00%	99,481	21,787	1,080	0.27%	0.06%	0.00%
24	45,106	5,173	58	62.89%	7.21%	0.08%	3,123,660	1,551,535	65,879	8.51%	4.23%	0.18%

[a] In units of repeat-normalized raw reads; from the merged data set

[b] DicerCall : Loci where < 85% of the small RNAs were not 20-24 nts in length have a DicerCall of "N". Otherwise, the DicerCall represents the most abundantly observed size of small RNA within the cluster

[c] HP: Hairpin-associated loci that are not microRNAs. MIR : microRNA loci

Non-Dicer Clusters

20-22 nt siRNAs

20-22 nt hpRNAs

20-22 nt *MIRNAs*

23-24 nt siRNAs

23-24 nt hpRNAs

23-24 nt *MIRNA*s

N *MIRNAs* .. seven of these were grouped with the 23-24-nt *MIRNA*s, while one was grouped with the 20-22-nt *MIRNAs*, based on the sizes of the annotated mature miRNA sequence.

Table S35. Text-based alignments of 121 *Amborella trichopoda MIRNA* loci. (Provided as a separate .txt file on the Science website).

Table S36. Summary of predicted (alignment score cutoff 3) and validated *Amborella* miRNA target genes

***Amborella MIRNA* Family**	**No. of predicted unique targets[a]**	***Arabidopsis* best BLASTp hit description[b]**	**No. of validated[c] unique targets**	***Arabidopsis* best BLASTp hit description[b]**
156	5	Squamosa promoter-binding protein-like (SBP domain)	5	Squamosa promoter-binding protein-like (SBP domain)
159	0		2	myb domain protein
160	2	Auxin response factor	3	Auxin response factor
164	2	NAC (No Apical Meristem) domain containing protein	2	NAC (No Apical Meristem) domain containing protein
166	5	Homeobox-leucine zipper family protein / lipid-binding START domain-containing protein SWITCH1	3	Homeobox-leucine zipper family protein / lipid-binding START domain-containing protein
167	0		1	Auxin response factor
168	0		1	Stabilizer of iron transporter SufD / Polynucleotidyl transferase
169	1	Stigma-specific Stig1 family protein	0	
171	2	GRAS family transcription factor	2	GRAS family transcription factor
172	4	Protein kinase superfamily protein with octicosapeptide/Phox/Bem1p domain Related to AP2.7	2	Integrase-type DNA-binding superfamily protein Related to AP2.7
319	1	myb domain protein	3	TCP family transcription factor Unknown protein myb domain protein
390	1	Leucine-rich repeat protein kinase family protein	0	
393	1	Auxin signaling F-box	3	auxin signaling F-box Glycosyltransferase family protein phenylalanyl-tRNA synthetase, putative / phenylalanine--tRNA ligase, putative

394	2	Galactose oxidase/kelch repeat superfamily protein	1	Cytochrome P450, family 94, subfamily B, polypeptide 3
396	10	Growth-regulating factor Galactose-binding protein	10	Growth-regulating factor Galactose-binding protein 31-kDa RNA binding protein chloroplast 30S ribosomal protein S20, putative
397	9	Laccase/Diphenol oxidase family protein ARM repeat superfamily protein tubby like protein 8	2	Laccase
398	0		5	6,7-dimethyl-8-ribityllumazine synthase / DMRL synthase / lumazine synthase / riboflavin synthase GroES-like zinc-binding alcohol dehydrogenase family protein copper chaperone for SOD1 proteasome inhibitor-related copper/zinc superoxide dismutase 2
2111	4	Galactose oxidase/kelch repeat superfamily protein Cytochrome P450 family ARM repeat superfamily protein	0	
8553	0		1	Tubulin beta 8
8558	0		2	Disease resistance protein (CC-NBS-LRR class) family Plant invertase/pectin methylesterase inhibitor superfamily protein
8560	0		1	TRICHOME BIREFRINGENCE-LIKE 13
8562	0		1	Heat shock factor binding protein
8565	0		2	-
8568	1	Disease resistance protein (TIR-NBS-LRR class) family	1	-

8573	0		1	Respiratory burst oxidase protein F
8578	3	Enoyl-CoA hydratase/isomerase D RNA processing factor 2 Pentatricopeptide repeat (PPR) superfamily protein	0	
8584	0		2	DNAJ-like 20 Monodehydroascorbate reductase 1
8586	2	POZ/BTB containin G-protein 1 Pentatricopeptide repeat (PPR) superfamily protein	0	
8587	0		1	Hercules receptor kinase 1
8595	0		1	Purine permease 5
8600	0		1	Pentatricopeptide (PPR) repeat-containing protein
8601	0		1	Glycine-rich protein family
8603	3	POZ/BTB containing G-protein 1 RNA-binding KH domain-containing protein	2	POZ/BTB containing G-protein 2
8610	2	ABC-2 type transporter family protein	0	
8613	0		1	RNA-binding KH domain-containing protein
Total unique targets	60		61	

[a] Targets predicted by TargetFinder 1.6 with alignment score cutoff of 3

[b] Simplified TAIR 10 annotation of the best *Arabidopsis* BLASTp hit (E-value threshold 1e-5) for corresponding *Amborella* miRNA target gene

[c] Targets validated by Cleaveland version 3 with p value cutoff of 0.05. Cleaveland pipeline was run with degradome tags mappings against EVM27 annotated *Amborella* coding sequences and TargetFInder 1.6 predicted miRNA-target alignments (score cutoff of 7) as input data sets.

Targets not identified

Table S37. Predicted Amborella miRNA targets (alignment score cutoff 3) that lack degradome evidence. (See large tables file).

Table S38. Summary of *Amborella* degradome libraries and alignments

Library	**Tissue**	**GEO accession**	**Number of aligned reads**	**Total no. of alignments after filtering**
AmLd2-2223	Leaves	GSM1024603	5,656,768	6,325,898
AmFF-2224	Female flowers	GSM1024604	4,754,704	5,169,645

Table S39. Details of cleaved Amborella miRNA targets identified using Cleaveland version 3 (p value cutoff 0.05) (See large tables file).

Table S40. Overlaps between small RNA loci and other genomic features by tallies*

		Exons	**Introns**	**Intergenic Regions**	**TE-related**
	Total loci in category	***109,783***	***82,271***	***27,438***	***823,265***
TE-related	***823,265***	793:773	108,598:22,593	689,376:24,672	
Non-Dicer	***14,839***	1,036:1,067	2,510:2,265	11,371:6,720	10,904:14,381
20-22-nt MIRNA	***55***	3:3	2:2	50:48	11:14
20-22-nt hpRNA	***707***	8:8	68:68	600:540	532:772
20-22-nt siRNA	***4,579***	198:201	630:570	3,590:2,577	3,218:4,681
23-24-nt MIRNA	***66***	0:0	29:29	34:34	56:117
23-24-nt hpRNA	***5,298***	46:52	905:745	4,150:3,159	4,871:7,412
23-24-nt siRNA	***45,464***	374:372	5,236:2,900	35,983:10,650	43,328:47,748
*The tallies of overlapping loci of each type are shown in the format [n overlapping of type row] : [n overlapping of type column]					

Table S41. Overlaps between small RNA loci and other genomic features by number of nucleotides

		Exons	**Introns**	**Intergenic Regions**	**TE-related**
	Total non-redundant nts category	***25,402,889***	***114,572,279***	***546,932,638***	***380,900,135***
TE-related	***380,900,135***	52,771	34,543,253	332,827,742	
Non-Dicer	***3,942,041***	161,423	567,436	2,949,295	2,517,966
20-22-nt MIRNA	***9,373***	329	49	8,644	1,048
20-22-nt hpRNA	***124,512***	730	12,104	105,801	94,161
20-22-nt siRNA	***1,411,432***	55,069	164,084	1,072,464	728,260
23-24-nt MIRNA	***25,456***	0	10,726	13,506	18,676
23-24-nt hpRNA	***1,736,132***	9,640	543,037	1,136,057	1,318,391
23-24-nt siRNA	***13,614,024***	75,589	1,486,879	10,395,410	11,998,485

Table S42. Three additional MIRNA loci discovered by manual inspection during analysis of PHAS loci (See large tables file).

Table S43. Summary of PHAS loci in *Amborella trichopoda*.

Trigger miRNA	Phased Gene	Gene Annotation	Maximum Phasing Score from All Libraries	Chr.	Strand	Start	End
Phased loci corresponding to annotated genes							
miR390	evm_27.TU.AmTr_v1.0_ scaffold00103.53	LRR-kinase (TAS3)	15.5	103	w	1,236,496	1,236,706
miR8558 b[1]	evm_27.TU.AmTr_v1.0_ scaffold00071.160	Disease resistance protein (TIR-NBS-LRR class), putative	19.1	71	c	2,544,994	2,545,204
miR8558 b[1]	evm_27.TU.AmTr_v1.0_ scaffold00036.138	Disease resistance protein (TIR-NBS-LRR class), putative	25.0	36	w	4,196,378	4,196,588
miR8558 b[1]	evm_27.TU.AmTr_v1.0_ scaffold00056.119	Disease resistance protein (TIR-NBS-LRR class), putative	41.1	56	w	2,127,552	2,127,762
n.d.	evm_27.TU.AmTr_v1.0_ scaffold00056.121	Disease resistance protein (TIR-NBS-LRR class), putative	18.7	56	w	2,150,865	2,151,075
n.d.	evm_27.TU.AmTr_v1.0_ scaffold00056.63	Disease resistance protein (TIR-NBS-LRR class), putative	16.3	56	w	1,040,215	1,042,607
miR2950	evm_27.TU.AmTr_v1.0_ scaffold00078.28	RNA-dependent RNA polymerase 6	18.9	78	c	536,566	536,776
miR2950	evm_27.TU.AmTr_v1.0_ scaffold00078.29	RNA-dependent RNA polymerase 6	20.1	78	c	555,251	555,461
n.d.	evm_27.TU.AmTr_v1.0_ scaffold00016.85	Auxin signaling F-box 3	19.6	16	w	1,703,641	1,703,851
miR8615	evm_27.TU.AmTr_v1.0_ scaffold00061.91	Hypothetical protein	22.5	61	w	1,489,812	1,490,022
Phased loci corresponding to intergenic loci							
miR8558 a[1]	chr:88:1862941-1863151	IGR	27.9	88		1,862,941	1,863,151
n.d.	chr:11:4415644-4415854	IGR	25.4	11		4,415,644	4,415,854
n.d.	chr:29:1503299-1503509	IGR	26.6	29		1,503,299	1,503,509
n.d.	chr:41:1333635-1333845	IGR	26.7	41		1,333,635	1,333,845

n.d.	chr:61:1498486-1498696	IGR	20.5	61	1,498,486	1,498,696
n.d.	chr:88:1852894-1853104	IGR	19.4	88	1,852,894	1,853,104
n.d.	chr:105:85400-85610	IGR	38.4	105	85,400	85,610
n.d.	chr:29:1507114-1507324	IGR	20.3	29	1,507,114	1,507,324
n.d.	chr:48:4051458-4051668	IGR	13.6	48	4,051,458	4,051,668

"n.d." means no trigger was detected.
"IGR" means intergenic region
[1] The miR8558 family is related to the miR482/2118 superfamily of miRNAs

Table S44. Population resequencing data summary

Population	Raw Reads (Size)	Average Read Size	Artificial Duplicates	Adapter Trimmed Reads	Clean Paired Reads	Aligned Median Read Depth	Heterozygosity (Avg. per 1 kb)
Amieu	128.1 M (12.9 Gb)	101	25.6 M	1.8 M	87.1 M	9X	1.39
Aoupinie C	88 M (8.9 Gb)	101	9.8 M	4 M	67.9 M	7X	1.20
Ba	328.9 M (33.2 Gb)	101	81 M	11.1 M	212.1 M	25X	1.59
Boregaou	375.7 M (37.9 Gb)	101	92.7 M	14.4 M	240.4 M	23X	1.30
Dogny B	118.8 M (12 Gb)	101	21.2 M	5 M	83.3 M	8X	1.14
Me Fomechawa	72.6 M (7.3 Gb)	101	10.9 M	1.9 M	52.6 M	6X	1.11
Me Ori	95.1 M (9.6 Gb)	101	7.3 M	1.6 M	75.8 M	8X	1.06
Nakada A	70.4 M (7.1 Gb)	101	5.7 M	0.7 M	55.4 M	6X	1.26
Ponandou	102.9 M (10.4 Gb)	101	8.8 M	1.9 M	81.1 M	8X	1.70
Pwicate	117.5 M (11.9 Gb)	101	11.5 M	4.8 M	91.4 M	10X	1.12
Tchamba	106.9 M (10.8 Gb)	101	8.2 M	2.5 M	85.2 M	8X	1.70
Tonine B	173.3 M (17.5 Gb)	101	37.1 M	2.2 M	115.3 M	12X	1.68
Santa Cruz (RUN 01)	69.6 M (5.7 Gb)	82	0.7 M	0.5 M	62.9 M	6X	1.26
Santa Cruz (RUN 02)	212.3 M (21.5 Gb)	101	3.3 M	2.7 M	172.2 M	23X	
Santa Cruz (RUN 03)	219.9 M (22.2 Gb)	101	3.5 M	2.7 M	178.8 M	23X	
Santa Cruz (RUN 04)	208.7 M (21.1 Gb)	101	3.2 M	2.7 M	170.3 M	22X	
Bonn	78.3 M (6.4 Gb)	82	0.9 M	0.5 M	72.3 M	7X	0.92

Table S45. Levels of genetic diversity in *Amborella* and species of *Populus*.

Species	π	θ_W	*D*	# of SNPs	Reference
Amborella trichopoda	0.0021	0.0017	0.8137	5,131,595	This study
Populus balsamifera	0.0026	0.0028	-0.233	3,506	(*275*)
P. balsamifera	0.0020	0.0033	---	412	(*276*)
P. tremula	0.0032	0.0030	0.239		(*277*)
*P. trichocarpa**	0.0026				(*278*)

*Note that the recent paper on *P. trichocarpa* by Slavov *et al*. (*279*) reports similar levels of diversity, although using different measures.

Table S46. Estimated mutation rate per generation (m) of *Amborella* and other plant species.

	Chromosomes (C)	**Genome size (Mb)**	**Mutation rate* (10^9base / generation)**	**Recombination rate (2C / Genome size)**
Solanum lycopersicum	12	781.67	15	1.60E-09
Mimulus guttatus	14	321.73	8.1	3.46E-09
Vitis vinifera	19	486.20	10.7	3.55E-09
Arabidopsis thaliana	5	119.67	4.1	2.44E-09
Carica papaya	9	331.27	8.2	2.20E-09
Populus trichocarpa	19	417.14	9.7	3.92E-09
Nicotiana tabacum	24	4434.00	50	9.60E-10
Amborella trichocarpa	13	731.00	14.1	1.84E-09

* Per-generation mutation rates was estimated using Lynch's {*Lynch:2010jh} published mutation rate-genome size relationship.

Additional Acknowledgments:

We thank the University of California, Santa Cruz Botanic Garden, the Atlanta Botanic Garden, and the Bonn Botanic Garden for providing *Amborella* material used in this study. We would also like to thank the South and North Provinces of New Caledonia for supplying sampling permit for Seed Proteomics and Population Genomic Analyses. Seed proteomic work was supported by grants from the Southern Province of New Caledonia and Fondation Paul Ars Cutolli.

Supplementary References:

90. V. Gomez-Alvarez, T. K. Teal, T. M. Schmidt, Systematic artifacts in metagenomes from complex microbial communities, *ISME J* **3**, 1314–1317 (2009).

91. B. Niu, L. Fu, S. Sun, W. Li, Artificial and natural duplicates in pyrosequencing reads of metagenomic data, *BMC Bioinformatics* **11**, 187 (2010).

92. M. Margulies *et al.*, Genome sequencing in microfabricated high-density picolitre reactors, *Nature* **437**, 376–380 (2005).

93. B. Star *et al.*, The genome sequence of Atlantic cod reveals a unique immune system, *Nature* **477**, 207–210 (2011).

94. Potato Genome Sequencing Consortium *et al.*, Genome sequence and analysis of the tuber crop potato, *Nature* **475**, 189–195 (2011).

95. S. Kurtz *et al.*, Versatile and open software for comparing large genomes, *Genome Biol* **5**, R12 (2004).

96. A. Berr, I. Schubert, Direct labelling of BAC-DNA by rolling-circle amplification, *Plant J* **45**, 857–862 (2006).

97. M. G. Grabherr *et al.*, Full-length transcriptome assembly from RNA-Seq data without a reference genome, *Nat Biotechnol* **29**, 644–652 (2011).

98. H.-H. Chou, M. Holmes, DNA sequence quality trimming and vector removal, *Bioinformatics* **17**, 1093 (2001).

99. Rice Annotation Project *et al.*, The Rice Annotation Project Database (RAP-DB): 2008 update, *Nucleic Acids Res* **36**, D1028–33 (2008).

100. T. Itoh *et al.*, Curated genome annotation of *Oryza sativa* ssp. *japonica* and comparative genome analysis with *Arabidopsis thaliana*, *Genome Res* **17**, 175–183 (2007).

101. D. Swarbreck *et al.*, The *Arabidopsis* Information Resource (TAIR): gene structure and function annotation, *Nucleic Acids Res* **36**, D1009–14 (2008).

102. A. Roberts, H. Pimentel, C. Trapnell, L. Pachter, Identification of novel transcripts in

annotated genomes using RNA-Seq, *Bioinformatics* **27**, 2325–2329 (2011).

103. M. Stanke *et al.*, AUGUSTUS: *ab initio* prediction of alternative transcripts, *Nucleic Acids Res* **34**, W435–9 (2006).

104. I. Korf, Gene finding in novel genomes, *BMC Bioinformatics* **5**, 59 (2004).

105. Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood, *Mol Biol Evol* **24**, 1586–1591 (2007).

106. C. Wei, M. R. Brent, Using ESTs to improve the accuracy of de novo gene prediction, *BMC Bioinformatics* **7**, 327 (2006).

107. I. Korf, P. Flicek, D. Duan, M. R. Brent, Integrating genomic homology into gene structure prediction, *Bioinformatics* **17 Suppl 1**, S140–8 (2001).

108. P. Flicek, E. Keibler, P. Hu, I. Korf, M. R. Brent, Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map, *Genome Res* **13**, 46–54 (2003).

109. A. E. Tenney *et al.*, Gene prediction and verification in a compact genome with numerous small introns, *Genome Res* **14**, 2330–2335 (2004).

110. R. H. Brown, S. S. Gross, M. R. Brent, Begin at the beginning: predicting genes with 5' UTRs, *Genome Res* **15**, 742–747 (2005).

111. W. J. Kent, BLAT—the BLAST-like alignment tool, *Genome Res* **12**, 656–664 (2002).

112. M. Garber, M. G. Grabherr, M. Guttman, C. Trapnell, Computational methods for transcriptome annotation and quantification using RNA-seq, *Nat Methods* **8**, 469–477 (2011).

113. C. Trapnell, L. Pachter, S. L. Salzberg, TopHat: discovering splice junctions with RNA-Seq, *Bioinformatics* **25**, 1105–1111 (2009).

114. G. S. C. Slater, E. Birney, Automated generation of heuristics for biological sequence comparison, *BMC Bioinformatics* **6**, 31 (2005).

115. T. Zhou *et al.*, Genome-wide identification of NBS genes in japonica rice reveals significant expansion of divergent non-TIR NBS-LRR genes, *Mol Genet Genomics* **271**, 402–415 (2004).

116. T. L. Bailey, N. Williams, C. Misleh, W. W. Li, MEME: discovering and analyzing DNA and protein sequence motifs, *Nucleic Acids Res* **34**, W369–73 (2006).

117. A. Lupas, M. Van Dyke, J. Stock, Predicting coiled coils from protein sequences, *Science* **252**, 1162–1164 (1991).

118. M. Lohse *et al.*, RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics, *Nucleic Acids Res* **40**, W622–7 (2012).

119. T. D. Wu, S. Nacu, Fast and SNP-tolerant detection of complex variants and splicing in short reads, *Bioinformatics* **26**, 873–881 (2010).

120. S. Kumar, M. L. Blaxter, Comparing de novo assemblers for 454 transcriptome data, *BMC Genomics* **11**, 571 (2010).

121. C. Trapnell *et al.*, Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation, *Nat Biotechnol* **28**, 511–515 (2010).

122. R. Lister *et al.*, Human DNA methylomes at base resolution show widespread epigenomic differences, *Nature* **462**, 315–322 (2009).

123. F. Krueger, S. R. Andrews, Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications, *Bioinformatics* **27**, 1571–1572 (2011).

124. D. Laslett, B. Canback, ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences, *Nucleic Acids Res* **32**, 11–16 (2004).

125. T. M. Lowe, S. R. Eddy, tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence, *Nucleic Acids Res* **25**, 955–964 (1997).

126. J. O. McInerney, GCUA: general codon usage analysis, *Bioinformatics* **14**, 372–373 (1998).

127. K. Lagesen *et al.*, RNAmmer: consistent and rapid annotation of ribosomal RNA genes, *Nucleic Acids Res* **35**, 3100–3108 (2007).

128. E. Lyons, B. Pedersen, J. Kane, M. Freeling, The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids, *Tropical Plant Biol.* **1**, 181–190 (2008).

129. B. J. Haas, A. L. Delcher, J. R. Wortman, S. L. Salzberg, DAGchainer: a tool for mining segmental genome duplications and synteny, *Bioinformatics* **20**, 3643–3646 (2004).

130. E. Lyons, M. Freeling, S. Kustu, W. Inwood, Using genomic sequencing for classical genetics in *E. coli* K12, *PLoS ONE* **6**, e16717 (2011).

131. H. Tang, E. Lyons, Unleashing the genome of *Brassica rapa*, *Front Plant Sci* **3**, 172 (2012).

132. R. J. Langham *et al.*, Genomic duplication, fractionation and the origin of regulatory novelty, *Genetics* **166**, 935–945 (2004).

133. M. Rho, J.-H. Choi, S. Kim, M. Lynch, H. Tang, *De novo* identification of LTR retrotransposons in eukaryotic genomes, *BMC Genomics* **8**, 90 (2007).

134. Z. Xu, H. Wang, LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons, *Nucleic Acids Res* **35**, W265–8 (2007).

135. D. Sankoff, C. Zheng, Fractionation, rearrangement and subgenome dominance, *Bioinformatics* **28**, i402–i408 (2012).

136. C. Zheng, E. Chen, V. A. Albert, E. Lyons, D. Sankoff, Ancient eudicot hexaploidy meets ancestral eurosid gene order, *BMC Genomics* **14**, S3 (2013).

137. J. C. Schnable, N. M. Springer, M. Freeling, Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss, *P Natl Acad Sci USA* **108**, 4069–4074 (2011).

138. S. R. Eddy, A new generation of homology search tools based on probabilistic inference, *Genome informatics International Conference on Genome Informatics* **23**, 205–211 (2009).

139. R. D. Finn, J. Clements, S. R. Eddy, HMMER web server: interactive sequence similarity searching, *Nucleic Acids Res* **39**, W29–37 (2011).

140. M. Csurös, Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood, *Bioinformatics* **26**, 1910–1912 (2010).

141. Z. Du, X. Zhou, Y. Ling, Z. Zhang, Z. Su, agriGO: a GO analysis toolkit for the agricultural community, *Nucleic Acids Res* **38**, W64–W70 (2010).

142. V. A. Albert *et al.*, Floral gene resources from basal angiosperms for comparative genomics research, *BMC Plant Biol* **5**, 5 (2005).

143. V. F. Irish, The flowering of *Arabidopsis* flower development, *Plant J* **61**, 1014–1028 (2010).

144. R. Sablowski, Genes and functions controlled by floral organ identity genes, *Semin. Cell Dev. Biol.* **21**, 94–99 (2010).

145. T. Ito, Coordination of flower development by homeotic master regulators, *Curr Opin Plant Biol* **14**, 53–59 (2011).

146. J. A. Higgins, P. C. Bailey, D. A. Laurie, Comparative genomics of flowering time pathways using *Brachypodium distachyon* as a model for the temperate grasses, *PLoS ONE* **5**, e10065 (2010).

147. A. Srikanth, M. Schmid, Regulation of flowering time: all roads lead to Rome, *Cell. Mol. Life Sci.* **68**, 2013–2037 (2011).

148. R. Melzer, G. Theissen, MADS and more: transcription factors that shape the plant, *Methods Mol Biol* **754**, 3–18 (2011).

149. K. Katoh, K.-I. Kuma, H. Toh, T. Miyata, MAFFT version 5: improvement in accuracy of multiple sequence alignment, *Nucleic Acids Res* **33**, 511–518 (2005).

150. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldon, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses, *Bioinformatics* **25**, 1972–1973 (2009).

151. A. Stamatakis, RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with

thousands of taxa and mixed models, *Bioinformatics* **22**, 2688–2690 (2006).

152. M. J. Sanderson, r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock, *Bioinformatics* **19**, 301–302 (2003).

153. S. A. Rensing *et al.*, The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants, *Science* **319**, 64–69 (2008).

154. P. Kenrick, P. R. Crane, The origin and early evolution of plants on land, *Nature* **389**, 33–39 (1997).

155. S.-M. Chaw, C.-C. Chang, H.-L. Chen, W.-H. Li, Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes, *J Mol Evol* **58**, 424–441 (2004).

156. J. A. Doyle, C. L. Hotten, in *Pollen and Spores: Patterns of Diversification*, S. Blackmore, S. H. Barnes, Eds. (Clarendon Press, 1991), pp. 169–195.

157. C. D. Bell, D. E. Soltis, P. S. Soltis, The age of the angiosperms: a molecular timescale without a clock, *Evolution* **59**, 1245–1258 (2005).

158. G. J. McLachlan, D. Peel, K. E. Basford, P. Adams, The EMMIX software for the fitting of mixtures of normal and t-components, *J Stat Softw* **4**, 1–14 (1999).

159. L. Cui *et al.*, Adaptive evolution of chloroplast genome structure inferred using a parametric bootstrap approach, *BMC Evol Biol* **6**, 13 (2006).

160. R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res* **32**, 1792–1797 (2004).

161. M. Suyama, D. Torrents, P. Bork, PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments, *Nucleic Acids Res* **34**, W609–12 (2006).

162. F. Garcia-Maroto *et al.*, New roles for MADS-box genes in higher plants, *Biol Plantarum* **46**, 321–330 (2003).

163. L. Gramzow, G. Theissen, A hitchhiker's guide to the MADS world of plants, *Genome Biol* **11**, 214 (2010).

164. M. Kwantes, D. Liebsch, W. Verelst, How MIKC* MADS-box genes originated and evidence for their conserved function throughout the evolution of vascular plant gametophytes, *Mol Biol Evol* **29**, 293–302 (2012).

165. M. D. Purugganan, S. D. Rounsley, R. J. Schmidt, M. F. Yanofsky, Molecular evolution of flower development: Diversification of the plant MADS-box regulatory gene family, *Genetics* **140**, 345–356 (1995).

166. G. Theissen *et al.*, A short history of MADS-box genes in plants, *Plant Mol Biol* **42**, 115–149 (2000).

167. D. E. Soltis *et al.*, The floral genome: an evolutionary history of gene duplication and shifting patterns of gene expression, *Trends Plant Sci* **12**, 358–367 (2007).

168. G. Theissen, H. Saedler, Plant biology. Floral quartets, *Nature* **409**, 469–471 (2001).

169. A. S. Rijpkema, T. Gerats, M. Vandenbussche, Evolutionary complexity of MADS complexes, *Curr Opin Plant Biol* **10**, 32–38 (2007).

170. A. S. Veron, K. Kaufmann, E. Bornberg-Bauer, Evidence of interaction network evolution by whole-genome duplications: A case study in MADS-box proteins, *Mol Biol Evol* **24**, 670–678 (2007).

171. J. D. Thompson, T. J. Gibson, F. Plewniak, F. Jeanmougin, D. G. Higgins, The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, *Nucleic Acids Res* **25**, 4876–4882 (1997).

172. K. B. Nicholas, H. B. Nicholas, Genedoc: A tool for editing and annotating multiple sequence alignments, distributed by the author, *EMBNEW NEWS* **4**, 1–4 (1997).

173. S. Guindon, O. Gascuel, A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood, *Syst Biol* **52**, 696–704 (2003).

174. S. Kim *et al.*, Phylogeny and diversification of B-function MADS-box genes in angiosperms: evolutionary and functional implications of a 260-million-year-old duplication, *Am J Bot* **91**, 2102–2118 (2004).

175. S. Kim *et al.*, Expression of floral MADS-box genes in basal angiosperms: implications for the evolution of floral regulators, *Plant J* **43**, 724–744 (2005).

176. L. M. Zahn *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).

177. C. P. Petersen, P. W. Reddien, Wnt signaling and the polarity of the primary body axis, *Cell* **139**, 1056–1068 (2009).

178. C. Schilde, T. Araki, H. Williams, A. Harwood, J. G. Williams, GSK3 is a multifunctional regulator of *Dictyostelium* development, *Development* **131**, 4555–4565 (2004).

179. B. W. Doble, J. R. Woodgett, GSK-3: tricks of the trade for a multi-tasking kinase, *J Cell Sci* **116**, 1175–1186 (2003).

180. M. C. Dornelas, A. A. Van Lammeren, M. Kreis, *Arabidopsis thaliana* SHAGGY-related protein kinases (AtSK11 and 12) function in perianth and gynoecium development, *Plant J* **21**, 419–429 (2000).

181. M. C. Dornelas, B. Lejeune, M. Dron, M. Kreis, The *Arabidopsis* SHAGGY-related protein kinase (ASK) gene family: structure, organization and evolution, *Gene* **212**, 249–257 (1998).

182. C. Jonak, H. Hirt, Glycogen synthase kinase 3/SHAGGY-like kinases in plants: an emerging family with novel functions, *Trends Plant Sci* **7**, 457–461 (2002).

183. O. Richard, N. Paquet, E. Haudecoeur, B. Charrier, Organization and expression of the GSK3/shaggy kinase gene family in the moss *Physcomitrella patens* suggest early gene multiplication in land plants and an ancestral response to osmotic stress, *J Mol Evol* **61**, 99–113 (2005).

184. B. Charrier, A. Champion, Y. Henry, M. Kreis, Expression profiling of the whole *Arabidopsis* shaggy-like kinase multigene family by real-time reverse transcriptase-polymerase chain reaction, *Plant Physiol* **130**, 577–590 (2002).

185. M.-J. Yoo, V. A. Albert, P. S. Soltis, D. E. Soltis, Phylogenetic diversification of *glycogen synthase kinase 3/SHAGGY-like kinase* genes in plants, *BMC Plant Biol* **6**, 3 (2006).

186. L. Rajjou *et al.*, Seed germination and vigor, *Annu Rev Plant Biol* **63**, 507–533 (2012).

187. P. R. Shewry, N. G. Halford, Cereal seed storage proteins: structures, properties and role in grain utilization, *J Exp Bot* **53**, 947–958 (2002).

188. K.-P. Häger, H. Braun, A. Czihal, B. Müller, H. Bäumlein, Evolution of seed storage protein genes: legumin genes of *Ginkgo biloba*, *J Mol Evol* **41**, 457–466 (1995).

189. T. B. Osborne, *The vegetable proteins* (Longmans, Green, and Co., New York, ed. 2, 1924).

190. J. Catusse, J.-M. Strub, C. Job, A. Van Dorsselaer, D. Job, Proteome-wide characterization of sugarbeet seed vigor and its tissue specific expression, *P Natl Acad Sci USA* **105**, 10262–10267 (2008).

191. M. M. Bradford, A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding, *Anal Biochem* **72**, 248–254 (1976).

192. H. Blum, H. Beier, H. J. Gross, Improved silver staining of plant proteins, RNA and DNA in polyacrylamide gels, *Electrophoresis* **8**, 93–99 (1987).

193. A. Shevchenko, M. Wilm, O. Vorm, M. Mann, Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels, *Anal Chem* **68**, 850–858 (1996).

194. B. Valot, O. Langella, E. Nano, M. Zivy, MassChroQ: a versatile tool for mass spectrometry quantification, *Proteomics* **11**, 3572–3577 (2011).

195. W. Albertin *et al.*, Linking post-translational modifications and variation of phenotypic traits, *Mol Cell Proteomics* **12**, 720–735 (2013).

196. D. F. Gruis, D. A. Selinger, J. M. Curran, R. Jung, Redundant proteolytic mechanisms process seed storage proteins in the absence of seed-type members of the vacuolar processing enzyme family of cysteine proteases, *Plant Cell* **14**, 2863–2882 (2002).

197. M. P. Scott, R. Jung, K. Müntz, N. C. Nielsen, A protease responsible for post-translational cleavage of a conserved Asn-Gly linkage in glycinin, the major seed storage protein of soybean, *P Natl Acad Sci USA* **89**, 658–662 (1992).

198. K.-P. Häger, C. Wind, Two ways of legumin‐precursor processing in conifers, *Eur J Biochem* **246**, 763–771 (1997).

199. K. Gallardo *et al.*, Proteomic analysis of arabidopsis seed germination and priming, *Plant Physiol* **126**, 835–848 (2001).

200. B. Bremer *et al.*, An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III, *Bot J Linn Soc* **161**, 105–121 (2009).

201. M. Wistrand, E. L. L. Sonnhammer, Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMER, *BMC Bioinformatics* **6**, 99 (2005).

202. D. J. McGarvey, R. Croteau, Terpenoid metabolism, *Plant Cell* **7**, 1015–1026 (1995).

203. R. Cao *et al.*, Diterpene cyclases and the nature of the isoprene fold, *Proteins* **78**, 2417–2432 (2010).

204. G. Li *et al.*, Nonseed plant *Selaginella moellendorffii* has both seed plant and microbial types of terpene synthases, *P Natl Acad Sci USA* **109**, 14711–14715 (2012).

205. D. M. Martin, J. Fäldt, J. Bohlmann, Functional characterization of nine Norway Spruce *TPS* genes and evolution of gymnosperm terpene synthases of the *TPS-d* subfamily, *Plant Physiol* **135**, 1908–1927 (2004).

206. T.-F. Yeh *et al.*, Rapid screening of wood chemical component variations using transmittance near-infrared spectroscopy, *J Agric Food Chem* **53**, 3328–3332 (2005).

207. C. W. Dence, in *Methods in Lignin Chemistry*, Springer Series in Wood Science. S. Y. Lin, C. W. Dence, Eds. (Springer-Verlag, Berlin, 1992), pp. 33–61.

208. A. B. Blakeney, P. J. Harris, R. J. Henry, B. A. Stone, A simple and rapid preparation of alditol acetates for monosaccharide analysis, *Carbohyd Res* **113**, 291–299 (1983).

209. T. Bitter, H. M. Muir, A modified uronic acid carbazole reaction, *Anal Biochem* **4**, 330–334 (1962).

210. R. Katahira, F. Nakatsubo, Determination of nitrobenzene oxidation products by GC and 1H-NMR spectroscopy using 5-iodovanillin as a new internal standard, *J Wood Sci* **47**, 378–382 (2001).

211. R. Shi *et al.*, Towards a systems approach for lignin biosynthesis in *Populus trichocarpa*: transcript abundance and specificity of the monolignol biosynthetic genes, *Plant Cell Physiol* **51**, 144–163 (2010).

212. J. Raes, A. Rohde, J. H. Christensen, Y. Van de Peer, W. Boerjan, Genome-wide characterization of the lignification toolbox in *Arabidopsis*, *Plant Physiol* **133**, 1051–1071 (2003).

213. H.-C. Chen *et al.*, Membrane protein complexes catalyze both 4- and 3-hydroxylation of cinnamic acid derivatives in monolignol biosynthesis, *P Natl Acad Sci USA* **108**, 21253–21258 (2011).

214. J. Liu, R. Shi, Q. Li, R. R. Sederoff, V. L. Chiang, A standard reaction condition and a single HPLC separation system are sufficient for estimation of monolignol biosynthetic pathway enzyme activities, *Planta* **236**, 879–885 (2012).

215. J. P. Wang *et al.*, Functional redundancy of the two 5-hydroxylases in monolignol biosynthesis of *Populus trichocarpa*: LC-MS/MS based protein quantification and metabolic flux analysis, *Planta* **236**, 795–808 (2012).

216. J. Harholt *et al.*, The glycosyltransferase repertoire of the spikemoss *Selaginella moellendorffii* and a comparative study of its cell wall, *PLoS ONE* **7**, e35846 (2012).

217. S. Guindon *et al.*, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0, *Syst Biol* **59**, 307–321 (2010).

218. Y. Wu *et al.*, Functional identification of two nonredundant Arabidopsis alpha(1,2)fucosyltransferases specific to arabinogalactan proteins, *J Biol Chem* **285**, 13638–13645 (2010).

219. R. Sarria *et al.*, Characterization of a family of *Arabidopsis* genes related to xyloglucan fucosyltransferase1, *Plant Physiol* **127**, 1595–1606 (2001).

220. E. M. McCarthy, J. F. McDonald, LTR_STRUC: a novel search and identification program for LTR retrotransposons, *Bioinformatics* **19**, 362–367 (2003).

221. D. Ellinghaus, S. Kurtz, U. Willhoeft, LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons, *BMC Bioinformatics* **9**, 18 (2008).

222. S. Steinbiss, U. Willhoeft, G. Gremme, S. Kurtz, Fine-grained annotation and classification of *de novo* predicted LTR retrotransposons, *Nucleic Acids Res* **37**, 7002–7013 (2009).

223. A. Bateman *et al.*, The Pfam protein families database, *Nucleic Acids Res* **32**, D138–41 (2004).

224. C. Llorens *et al.*, The Gypsy Database (GyDB) of mobile genetic elements: release 2.0, *Nucleic Acids Res* **39**, D70–4 (2011).

225. K. Eilbeck *et al.*, The Sequence Ontology: a tool for the unification of genome annotations, *Genome Biol* **6**, R44 (2005).

226. A. J. Enright, S. Van Dongen, C. A. Ouzounis, An efficient algorithm for large-scale

detection of protein families, *Nucleic Acids Res* **30**, 1575–1584 (2002).

227. S. F. Altschul *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res* **25**, 3389–3402 (1997).

228. J. Jurka *et al.*, Repbase Update, a database of eukaryotic repetitive elements, *Cytogenet Genome Res* **110**, 462–467 (2005).

229. P. SanMiguel, B. S. Gaut, A. Tikhonov, Y. Nakajima, J. L. Bennetzen, The paleontology of intergene retrotransposons of maize, *Nat Genet* **20**, 43–45 (1998).

230. J. Ma, J. L. Bennetzen, Rapid recent growth and divergence of rice nuclear genomes, *P Natl Acad Sci USA* **101**, 12404–12410 (2004).

231. M. A. Larkin *et al.*, ClustalW and ClustalX version 2 (2007), *Bioinformatics* **23**, 2947–2948 (2007).

232. B. J. Frey, D. Dueck, Clustering by passing messages between data points, *Science* **315**, 972–976 (2007).

233. M. Rho, H. Tang, MGEScan-non-LTR: computational identification and classification of autonomous non-LTR retrotransposons in eukaryotic genomes, *Nucleic Acids Res* **37**, e143 (2009).

234. M. Spannagl *et al.*, MIPSPlantsDB—plant database resource for integrative and comparative plant genome research, *Nucleic Acids Res* **35**, D834–D840 (2007).

235. Y.-W. Yuan, S. R. Wessler, The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies, *P Natl Acad Sci USA* **108**, 7884–7889 (2011).

236. Y. Han, J. M. Burnette, S. R. Wessler, TARGeT: a web-based pipeline for retrieving and characterizing gene and transposable element families from genomic sequences, *Nucleic Acids Res* **37**, e78 (2009).

237. T. A. Tatusova, T. L. Madden, BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences, *FEMS Microbiol Lett* **174**, 247–250 (1999).

238. Y. Han, S. R. Wessler, MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences, *Nucleic Acids Res* **38**, e199 (2010).

239. A. L. Price, N. C. Jones, P. A. Pevzner, *De novo* identification of repeat families in large genomes, *Bioinformatics* **21 Suppl 1**, i351–8 (2005).

240. J. C. Wootton, S. Federhen, Statistics of local complexity in amino acid sequences and sequence databases, *Computers & Chemistry* **17**, 149–163 (1993).

241. S. Ouyang, C. R. Buell, The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants, *Nucleic Acids Res* **32**, D360–D363 (2004).

242. C. Feschotte, U. Keswani, N. Ranganathan, M. L. Guibotsy, D. Levine, Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes, *Genome Biol Evol* **1**, 205–220 (2009).

243. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, *Genome Biol* **10**, R25 (2009).

244. M. J. Axtell, ShortStack: Comprehensive annotation and quantification of small RNA genes, *RNA* **19**, 740–751 (2013).

245. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools, *Bioinformatics* **25**, 2078–2079 (2009).

246. P. Rice, I. Longden, A. Bleasby, EMBOSS: the European Molecular Biology Open Software Suite, *Trends Genet* **16**, 276–277 (2000).

247. A. Kozomara, S. Griffiths-Jones, miRBase: integrating microRNA annotation and deep-sequencing data, *Nucleic Acids Res* **39**, D152–7 (2011).

248. P. Chellappan *et al.*, siRNAs from miRNA sites mediate DNA methylation of target genes, *Nucleic Acids Res* **38**, 6883–6894 (2010).

249. L. Wu *et al.*, DNA methylation mediated by a microRNA pathway, *Mol Cell* **38**, 465–475 (2010).

250. N. Fahlgren *et al.*, High-throughput sequencing of *Arabidopsis* microRNAs: evidence for frequent birth and death of *MIRNA* genes, *PLoS ONE* **2**, e219 (2007).

251. C. Addo-Quaye, T. W. Eshoo, D. P. Bartel, M. J. Axtell, Endogenous siRNA and miRNA targets identified by sequencing of the *Arabidopsis* degradome, *Curr Biol* **18**, 758–762 (2008).

252. M. A. German *et al.*, Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends, *Nat Biotechnol* **26**, 941–946 (2008).

253. M. A. German, S. Luo, G. Schroth, B. C. Meyers, P. J. Green, Construction of Parallel Analysis of RNA Ends (PARE) libraries for the study of cleaved miRNA targets and the RNA degradome, *Nat Prot* **4**, 356–362 (2009).

254. Z. Ma, C. Coruh, M. J. Axtell, *Arabidopsis lyrata* small RNAs: transient *MIRNA* and small interfering RNA loci within the *Arabidopsis* genus, *Plant Cell* **22**, 1090–1103 (2010).

255. E. De Paoli *et al.*, Distinct extremely abundant siRNAs associated with cosuppression in petunia, *RNA* **15**, 1965–1970 (2009).

256. H. Jiang, W. H. Wong, SeqMap: mapping massive amount of oligonucleotides to the genome, *Bioinformatics* **24**, 2395–2396 (2008).

257. M. J. Axtell, Evolution of microRNAs and their targets: are all microRNAs biologically

relevant? *Biochim Biophys Acta* **1779**, 725–734 (2008).

258. V. Poncet *et al.*, Microsatellite markers for *Amborella* (Amborellaceae), a monotypic genus endemic to New Caledonia, *Am J Bot* **99**, e411–4 (2012).

259. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads, *EMBnet J* **17**, pp. 10–12 (2011).

260. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics* **25**, 1754–1760 (2009).

261. M. Burrows, D. J. Wheeler, A block-sorting lossless data compression algorithm, *Systems Research Center, Technical Report*, 1–18 (1994).

262. P. Ferragina, G. Manzini, Opportunistic data structures with applications, *Foundations of Computer Science, Proceedings 41st Annual Symposium on*, 390–398 (2000).

263. R. A. Lippert, Space-efficient whole genome comparisons with Burrows-Wheeler transforms, *J Comput Biol* **12**, 407–415 (2005).

264. T. W. Lam, W. K. Sung, S. L. Tam, C. K. Wong, S. M. Yiu, Compressed indexing and local alignment of DNA, *Bioinformatics* **24**, 791–797 (2008).

265. E. Garrison, G. Marth, Haplotype-based variant detection from short-read sequencing, *arXiv preprint arXiv:1207.3907* (2012).

266. M. Krzywinski *et al.*, Circos: An information aesthetic for comparative genomics, *Genome Res* **19**, 1639–1645 (2009).

267. H. Li, R. Durbin, Inference of human population history from individual whole-genome sequences, *Nature* **475**, 493–496 (2011).

268. M. Lynch, Evolution of the mutation rate, *Trends Genet* **26**, 345–352 (2010).

269. J. K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data, *Genetics* **155**, 945–959 (2000).

270. R Development Core Team, *R: A language and environment for statistical computing, reference index version 2.15.1 (2012-06-22)* (The R Foundation for Statistical Computing, Vienna, Austria., 2012; http://www.R-project.org).

271. G. Evanno, S. Regnaut, J. Goudet, Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study, *Mol Ecol* **14**, 2611–2620 (2005).

272. A. A. Salamov, V. V. Solovyev, *Ab initio* gene finding in *Drosophila* genomic DNA, *Genome Res* **10**, 516–522 (2000).

273. L. Pařenicová *et al.*, Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in *Arabidopsis* new openings to the MADS world, *Plant Cell* **15**,

1538–1551 (2003).

274. H. Shi, B. Schmidt, W. Liu, W. Müller-Wittig, A parallel algorithm for error correction in high-throughput short-read data on CUDA-enabled graphics hardware, *J Comput Biol* **17**, 603–615 (2010).

275. M. S. Olson *et al.*, Nucleotide diversity and linkage disequilibrium in balsam poplar (*Populus balsamifera*), *New Phytol* **186**, 526–536 (2010).

276. S. R. Keller, M. S. Olson, S. Silim, W. Schroeder, P. Tiffin, Genomic diversity, population structure, and migration following rapid range expansion in the Balsam poplar, *Populus balsamifera*, *Mol Ecol* **19**, 1212–1226 (2010).

277. M. V. García, P. K. Ingvarsson, An excess of nonsynonymous polymorphism and extensive haplotype structure at the *PtABI1B* locus in European aspen (*Populus tremula*): a case of balancing selection in an obligately outcrossing plant? *Heredity* **99**, 381–388 (2007).

278. G. A. Tuskan *et al.*, The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray), *Science* **313**, 1596–1604 (2006).

279. G. T. Slavov *et al.*, Genome resequencing reveals multiscale geographic structure and extensive linkage disequilibrium in the forest tree *Populus trichocarpa*, *New Phytol* **196**, 713–725 (2012).