EXTENDED PDF FORMAT
SPONSORED BY
Move Forward with Confidence
R&D Systems Kits for Stem Cell Verification
>Learn More
www.rndsystems.com

The *Amborella* Genome and the Evolution of Flowering Plants

Amborella Genome Project

Science **342**, (2013);
DOI: 10.1126/science.1241089

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by clicking here.

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines here.

The following resources related to this article are available online at www.sciencemag.org (this information is current as of January 8, 2014):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:
http://www.sciencemag.org/content/342/6165/1241089.full.html

Supporting Online Material can be found at:
http://www.sciencemag.org/content/suppl/2013/12/18/342.6165.1241089.DC1.html

A list of selected additional articles on the Science Web sites **related to this article** can be found at:
http://www.sciencemag.org/content/342/6165/1241089.full.html#related

This article **cites 82 articles**, 32 of which can be accessed free:
http://www.sciencemag.org/content/342/6165/1241089.full.html#ref-list-1

This article has been **cited by** 1 articles hosted by HighWire Press; see:
http://www.sciencemag.org/content/342/6165/1241089.full.html#related-urls

Downloaded from www.sciencemag.org on January 8, 2014

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. Copyright 2013 by the American Association for the Advancement of Science; all rights reserved. The title *Science* is a registered trademark of AAAS.

The *Amborella* Genome and the Evolution of Flowering Plants

Amborella Genome Project*†

READ THE FULL ARTICLE ONLINE
http://dx.doi.org/10.1126/science.1241089

Cite this article as *Amborella* Genome Project, *Science* **342**, 1241089 (2013). DOI: 10.1126/science.1241089

Introduction: Darwin famously characterized the rapid rise and early diversification of flowering plants (angiosperms) in the fossil record as an "abominable mystery." Identifying genomic changes that accompanied the origin of angiosperms is key to unraveling the molecular basis of biological innovations that contributed to their geologically near-instantaneous rise to ecological dominance.

Methods: We provide a draft genome for *Amborella trichopoda*, the single living representative of the sister lineage to all other extant flowering plants and use phylogenomic and comparative genomic analyses to elucidate ancestral gene content and genome structure in the most recent common ancestor of all living angiosperms.

Results: We reveal that an ancient genome duplication predated angiosperm diversification. However, unlike all other sequenced angiosperm genomes, the *Amborella* genome shows no evidence of more recent, lineage-specific genome duplications, making *Amborella* particularly well suited to help interpret genomic changes after polyploidy in other angiosperms. The remarkable conservation of gene order (synteny) among the genomes of *Amborella* and other angiosperms has enabled reconstruction of the ancestral gene arrangement in eudicots (~75% of all angiosperms). An ancestral angiosperm gene set was inferred to contain at least 14,000 protein-coding genes; subsequent changes in gene content and genome structure across disparate flowering plant lineages are associated with the evolution of important crops and model species. Relative to nonangiosperm seed plants, 1179 gene lineages first appeared in association with the origin of the angiosperms. These include genes important in flowering, wood formation, and responses to environmental stress. Unlike other angiosperms, the *Amborella* genome lacks evidence for recent transposon insertions while retaining ancient and divergent transposons. The genome harbors an abundance of atypical lineage-specific 24-nucleotide microRNAs, with at least 27 regulatory microRNA families inferred to have been present in the ancestral angiosperm. Population genomic analysis of 12 individuals from across the small native range of *Amborella* in New Caledonia reveals geographic structure with conservation implications, as well as both a recent genetic bottleneck and high levels of genome diversity.

Discussion: The *Amborella* genome is a pivotal reference for understanding genome and gene family evolution throughout angiosperm history. Genome structure and phylogenomic analyses indicate that the ancestral angiosperm was a polyploid with a large constellation of both novel and ancient genes that survived to play key roles in angiosperm biology.

FIGURES IN THE FULL ARTICLE

SUPPLEMENTARY MATERIALS

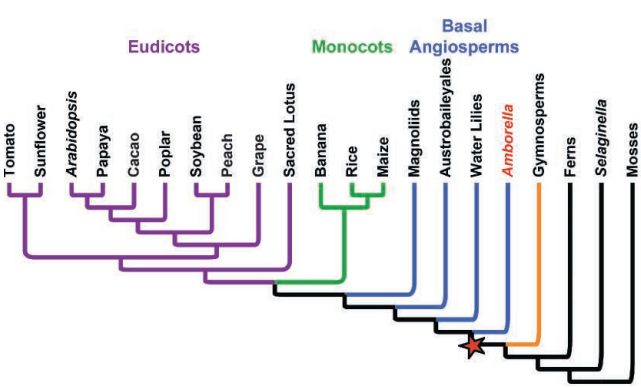
RELATED ITEMS IN *SCIENCE*

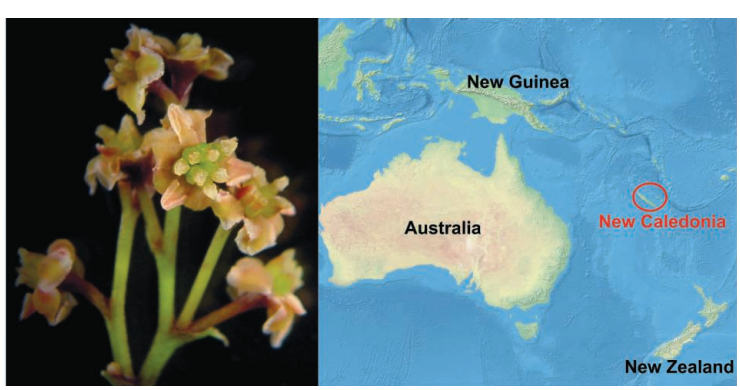
K. Adams, Genomic clues to the ancestral flowering Plant. *Science* **342**, 1456–1457 (2013). DOI: 10.1126/science.1248709

D. W. Rice, Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm *Amborella*. *Science* **342**, 1468–1473 (2013). DOI: 10.1126/science.1246275

S. Chamala, Assembly and validation of the genome of the nonmodel basal angiosperm *Amborella*. *Science* **342**, 1516–1517 (2013). DOI: 10.1126/science.1241130

***Amborella trichopoda*, an understory shrub endemic to New Caledonia, is the sole surviving sister species of all other living flowering plants (angiosperms).** The *Amborella* genome provides an exceptional reference for inferring features of the first flowering plants and identifies an ancient angiosperm-wide whole-genome duplication (red star). *Amborella* flowers have spirally arranged tepals, unfused carpels (female; shown), and laminar stamens.





*All authors and their affiliations and contributions are listed at the end of the paper.
†Corresponding author. E-mail: cwd3@psu.edu. Contact information for working groups is provided in the authorship details.

The *Amborella* Genome and the Evolution of Flowering Plants

***Amborella* Genome Project*†**

Amborella trichopoda is strongly supported as the single living species of the sister lineage to all other extant flowering plants, providing a unique reference for inferring the genome content and structure of the most recent common ancestor (MRCA) of living angiosperms. Sequencing the *Amborella* genome, we identified an ancient genome duplication predating angiosperm diversification, without evidence of subsequent, lineage-specific genome duplications. Comparisons between *Amborella* and other angiosperms facilitated reconstruction of the ancestral angiosperm gene content and gene order in the MRCA of core eudicots. We identify new gene families, gene duplications, and floral protein-protein interactions that first appeared in the ancestral angiosperm. Transposable elements in *Amborella* are ancient and highly divergent, with no recent transposon radiations. Population genomic analysis across *Amborella*'s native range in New Caledonia reveals a recent genetic bottleneck and geographic structure with conservation implications.

The origin of angiosperms (flowering plants) prompted one of Earth's greatest terrestrial radiations, famously characterized by Charles Darwin as "an abominable mystery" (*1*). The oldest angiosperm fossils date from 130 to 136 million years ago (Ma), but the crown age for the angiosperms has been estimated to be at least 160 Ma (*2*–*7*). The origin of the flowering plants was followed by a rapid rise to ecological dominance before the end of the Cretaceous. Angiosperms have since diversified to at least 350,000 species, occupying nearly all terrestrial and many aquatic environments. Angiosperms provide the vast majority of human food and contribute massively to global photosynthesis and carbon sequestration. Understanding angiosperm evolution and diversification is therefore essential to elucidating key processes that underlie the assembly of biotic associations and entire ecosystems.

Paleobotany, phylogenetics, and developmental biology have dramatically reshaped views of the origin and early diversification of angiosperms (*8*). Most phylogenetic analyses examining chloroplast (*9*–*12*), large multigene nuclear (*6*, *13*, *14*), and chloroplast, mitochondrial, and nuclear genes combined (*15*) strongly support *Amborella trichopoda*, an understory shrub endemic to New Caledonia, as the single sister species to all other extant angiosperms (Fig. 1) (*16*). Sister lineages such as *Amborella*, when compared with other key lineages, can provide unique insights into ancestral characteristics, including genome structure and gene content. Specifically, comparisons of the *Amborella* genome reported here to other sequenced angiosperm genomes distinguish the genomic features of the most recent common ancestor (MRCA) of all extant flowering plants from those acquired later within individual angiosperm lineages.

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

†All authors with their affiliations and contributions appear at the end of this paper.

Genome Assembly and Annotation

The genome of *Amborella* was sequenced and assembled using a whole-genome shotgun approach that combined more than 23 Gb of single- and paired-end sequence data (~30×) obtained from multiple sequencing platforms (table S1) (*17*, *18*). Our assembly comprises 5745 scaffolds totaling 706 Mb, 81% of an earlier genome size estimate of 870 Mb (*19*) and 94% of our sequence-based estimate of 748 Mb (*17*, *18*), with a mean scaffold length of 123 kb, an N50 length of 4.9 Mb, and a maximum scaffold length of 16 Mb (table S2). Ninety percent of the assembled genome is contained in 155 scaffolds larger than 1.1 Mb.

We evaluated the quality of the assembly using an integrated strategy of comparison with available finished bacterial artificial chromosome (BAC) contig sequences (*20*), a BAC-based physical map (*20*), fluorescence in situ hybridization (FISH), and whole-genome (optical) mapping (*18*). Accurate and nearly complete coverage of the regions previously characterized through BAC sequencing (*20*) and congruence (99%) with the available physical map verify that the local contig assemblies are of high quality. FISH-based mapping of scaffold ends to chromosomes has thus far confirmed 306 Mb (44%) of the genome assembly (*18*).

Annotation of protein-coding genes and repetitive elements was performed with DAWGPAWS (*17*, *21*). Despite the different histories of ancient whole-genome duplication (WGD; paleopolyploidy), the number of predicted protein-coding genes in the *Amborella* genome is similar to the number given in the most recent *Arabidopsis thaliana* reference genome annotation (TAIR10, http://www.arabidopsis.org). Evidence Modeler (*22*) was used to integrate gene annotations, producing 26,846 automated high-confidence gene predictions, 20,301 (76%) of which are supported by transcript evidence. Additionally, 17,089 gene models contain one or more introns, with 86.9% of the splice sites supported by transcript evidence. Refined gene models were further curated through manual comparisons with *Amborella* complementary DNA transcript assemblies, gene family analyses, and homologous full-length genes from other species (*17*). Many of the resulting gene models included very long introns relative to other annotated genomes [for example, mean intron length is 1528 bp in *Amborella*, compared to 165, 966, and 1017 bp in *Arabidopsis thaliana*, grape (*Vitis vinifera*), and Norway spruce (*Picea abies*), respectively] (*17*). Annotated high-confidence protein-coding gene models occupied 152 Mb (~21.5% of the genome assembly), including 25.4 Mb of exon sequence. A conservative estimate of 17,095 alternatively spliced protein isoforms was predicted for 6407 intron-containing genes, and multiple splice site variants were inferred for 37.5% of the genes with two or more exons (*17*).

Gene body methylation is generally conserved in monocots and eudicots (*23*) and has been hypothesized to play an important regulatory role in eukaryotic genomes, distinct from the silencing of transposons (*24*). Whereas gene body methylation is not seen in mosses or lycophytes (*25*), bisulfite sequence mapping indicates that gene body methylation is prevalent in *Amborella* (fig. S5), suggesting that it is an ancestral feature found in the the MRCA of flowering plants.

Angiosperm-Wide Genome Duplication

Intragenomic syntenic analysis of *Amborella* provides clear structural evidence of an ancient WGD event. An *Amborella* versus *Amborella* structural comparison shows numerous, duplicate colinear genes (syntenic homeologs) (Fig. 2A and fig. S9). Forty-seven intra-*Amborella* syntenic blocks were identified containing 466 *Amborella* gene pairs inferred to be descendants of this WGD event (Fig. 2A and table S10). Syntenic blocks contain an average of 10 homeologous gene pairs, and the longest block contains 23 gene pairs. Collectively, these 47 blocks include 6565 gene models (out of 26,846), indicating that about one-quarter of the annotated *Amborella* gene space maps to assembly scaffolds exhibiting synteny-based signal for an ancient WGD event.

Previous examinations of plant genomes have shown that polyploidy has been a prominent feature in the evolutionary history of angiosperms and that WGD events have had major impacts on genome structure and gene family evolution (*7*, *26*–*30*). Although most paleopolyploid events detected to date are associated with specific angiosperm families or smaller clades, an older paleohexaploidization (genome triplication), referred to as *gamma*, has been confirmed in the common ancestor of most eudicots (*26*–*28*, *30*). If the *Amborella* WGD revealed in this study was an internal, lineage-specific event, a 2:3 syntenic depth ratio would be expected between *Amborella* and *Vitis*. Instead, structural analysis shows a clear

1:3 relationship of *Amborella* and *Vitis* syntenic blocks that map to the *gamma* paleopolyploidy (Fig. 2B and figs. S6 to S8), indicating that the WGD detected in *Amborella* is not lineage-specific and likely occurred in an ancient common ancestor of the two species, thereby confirming that the divergence of *Amborella* predates *gamma* (*20*, *26*, *27*).

Phylogenomic analyses of 11,519 gene families confirm that dispersed, duplicated genes specific to *Amborella* are uncommon (282 nontandem gene pairs), especially when compared to older gene family expansions shared across angiosperm or seed plant lineages (473 orthogroups with at least 50% bootstrap support) (*17*). The age distribution of the pre-angiosperm gene duplications is bimodal (fig. S17), with the two peaks corresponding to the same ancestral angiosperm (*epsilon*) and ancestral seed plant (*zeta*) genome duplications inferred in previous analyses based on transcriptome data (*7*). *Zeta* has escaped syntenic detection in this and other studies of angiosperm synteny, presumably because of extreme gene loss and rearrangements that have accumulated since this hypothesized ancient event more than 300 Ma. To confirm further that the syntenic, duplicated blocks correspond to the same angiosperm-wide duplications discovered through phylogenomics, we manually curated six large duplicated blocks (Fig. 2B and fig. S10). Phylogenetic analysis of 155 syntenic gene pairs from these large blocks supports the placement of the *epsilon* genome duplication on the branch leading to the MRCA of extant angiosperms (77 of 155 gene trees resolved *epsilon* with bootstrap values of 80% or greater; see table S11).

In summary, *Amborella* genome structure demonstrates no evidence of WGD since this lineage diverged from the rest of the angiosperms at least 160 Ma. However, analyses indicate that paralagous gene copies associated with the *epsilon* WGD resulted from duplication shortly before the diversification of all living angiosperms (*7*). This event represents the most ancient WGD known in plants for which structural evidence persists. The *Amborella* genome therefore provides a unique evolutionary reference for elucidating genome content and structure in the MRCA of extant angiosperms and for resolving the timing of WGDs and single-gene losses and gains that have contributed to the diversification of the angiosperms (*8*).

Ancestral Gene Order in Core Eudicots

We combined scaffold-level information from *Amborella* with chromosome-level data from the eudicot rosid lineages of grape (*V. vinifera*), peach (*Prunus persica*), and cacao (*Theobroma cacao*) to reconstruct the hypothetical structure of seven inferred pre-hexaploidization chromosomes in the ancestor of the core eudicots. These three species were chosen because they have retained structurally similar genomes and clear patterns of paralogy among syntenic gene copies (fig. S11), enabling us to assign most genes to one of seven groups of three homeologous chromosomes or segments (*26*, *27*, *30*, *31*). A comprehensive analysis of *Amborella* and the three subgenomes from the representative rosids (combining a number of computational techniques) (*29*, *31*, *32*) enabled a completely automated reconstruction of ancestral gene order beyond the level of "contiguous ancestral regions" [compare (*33*)]. Figure 2C shows the orthologous gene alignments between one of the ancestral chromosomes, an *Amborella* genome

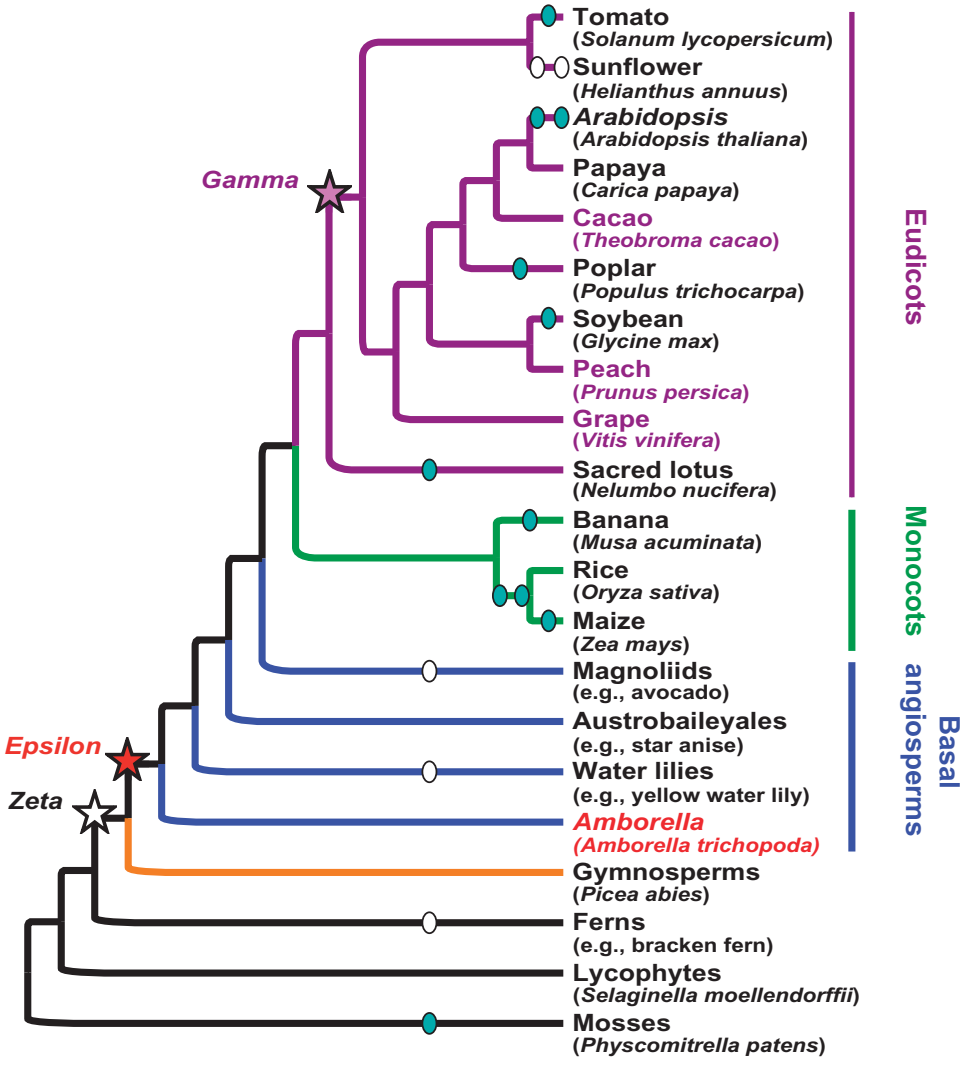


Fig. 1. *Amborella* is sister to all other extant angiosperms. An overview of land plant phylogeny is shown, including the relationships among major lineages of angiosperms. Representatives with sequenced genomes are shown for most lineages (scientific names in parentheses); however, basal angiosperms (all of which lack genome sequences except for *Amborella*) and nonflowering plant lineages are indicated by their larger group names. Hypothesized polyploidy events in land plant evolution are overlaid on the phylogeny with symbols. The red star indicates the common ancestor of angiosperms and the evolutionary timing of the *epsilon* WGD (*7*). The evolutionary timing of *zeta* (*7*) and *gamma* (*26*, *27*, *82*) polyploidy events are shown with empty and purple stars, respectively. The peach, cacao, and grape genomes (purple text) were used with the *Amborella* genome to reconstruct the gene order in the pre-*gamma* core eudicot (Fig. 2C). Additional polyploidy events are indicated with ellipses. Events supported by genome-scale synteny analyses are filled, whereas those supported only with frequency distributions of paralogous gene pairs (Ks) or phylogenomic analyses are empty (*34*–*37*, *83*, *84*).

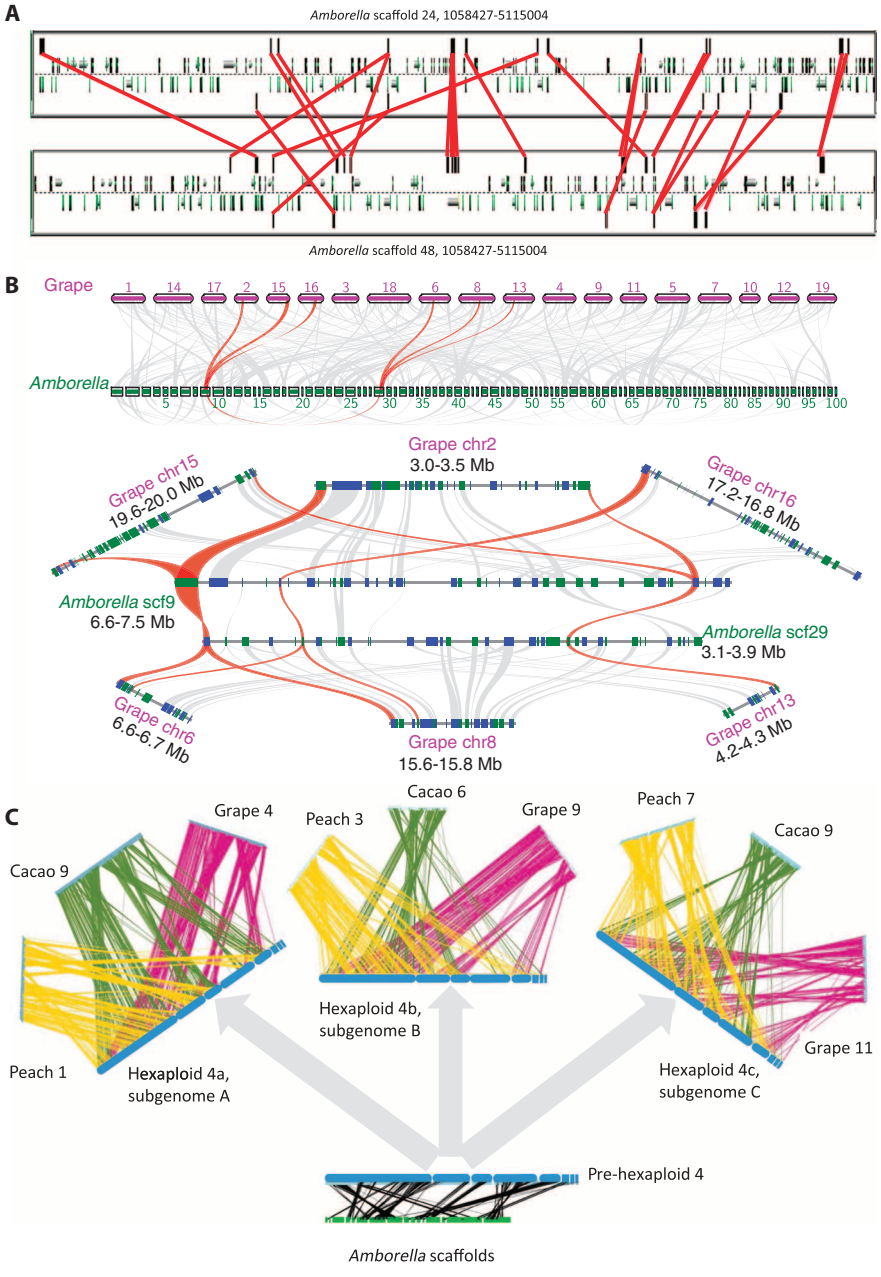


Fig. 2. Synteny analysis of *Amborella*. (**A**) High-resolution analysis of *Amborella-Amborella* intragenomic syntenic regions putatively derived from the ancestral angiosperm (*epsilon*) WGD. Note the series of colinear genes between the two regions. Intragenomic syntenic regions from *Amborella* are shown when scaffolds are compared and appear as a series of colinear genes between the two regions. (**B**) Macrosynteny and microsynteny between genomic regions in *Amborella* and grape. Top: Macrosynteny patterns between grape and *Amborella* and within *Amborella* scaffolds (only scaffolds 1 to 100 are shown). Each *Amborella* region aligns with up to three regions in grape that resulted from the *gamma* hexaploidization event in early core eudicots (*27*). Syntenic regions within the *Amborella* genome were derived from the *epsilon* WGD before the origin of all extant angiosperms (*7*). An exemplar set of blocks, showing two homeologous *Amborella* regions derived from this early WGD, aligns to three distinct grape regions (derived from *gamma*), with eight parallel regions in total. Bottom: Microsynteny is shown among the eight regions (noted above). Blocks represent genes with orientation on the same strand (blue) or reverse strand (green); shades represent matching gene pairs. (**C**) Gene order alignments between one of the seven hypothesized ancestral core eudicot chromosomes (blue bar), the three post-hexaploidization copies of this chromosome for peach, cacao, and grape chromosomes descending from it (top of figure), and a subset of the *Amborella* scaffolds (green, bottom of figure). Similar configurations were obtained for the other six ancestral chromosomes.

scaffold, and triplicated blocks of genes in the rosid genomes. This analysis, which uses *Amborella* as an outgroup to the three eudicot genomes, would not have been possible without *Amborella* or another (as yet undiscovered) non-eudicot genome that retains a large amount of syntenic signal. The prevalence of WGDs in monocots and other basal angiosperms (*34*–*37*) limits the possibility of identifying such genomes. Similar patterns for all seven ancestral core eudicot chromosomes (*17*) illustrate the utility of the *Amborella* genome for reconstructing ancestral genomes within the angiosperms, thus clarifying the divergence of subgenomes after WGD events. In the case of the reconstructed core eudicot ancestor, tracking the syntenically retained descendant blocks in the three rosids reveals a consistent pattern of subgenome dominance (fig. S15). This pattern, which governs the fractionation likelihood of gene triplets generated by the *gamma* event, is not evident from the direct comparison of the extant genomes alone and highlights the value of ancestral genome reconstructions enabled by *Amborella*.

Ancestral Angiosperm Gene Content

To assess the origin and history of angiosperm genes using *Amborella* genes as an anchor, we clustered protein-coding genes from 22 sequenced land plant genomes selected for their phylogenetic representation into 53,136 orthogroup clusters (narrowly defined gene lineages; table S12), with annotations provided by the associated pfam domain and full Gene Ontology (GO) terms for genes contained in these clusters (table S16). We further merged the orthogroups into 6054 super-orthogroup clusters representing more inclusively circumscribed gene families (*17*). The broader circumscription of super-orthogroups allows for the clustering of more divergent homologs, thus increasing the likelihood that they represent truly distinct gene families. Phylogenetic analyses of super-orthogroups can help to root orthogroup phylogenies and resolve the relationships among related orthogroups.

We estimated the ancestral gene content at key nodes in land plant phylogeny and modeled the changes of orthogroups occurring along each branch (Fig. 3 and tables S13 to S15). The largest changes in gene family content appear to have occurred evolutionarily recently along terminal branches, or are shared among closely related taxa, such as within the tomato (Solanaceae) or crucifer (Brassicaceae) families. Large numbers of orthogroup gains were also inferred along the deeper branches leading to all angiosperms (3285 new orthogroups using parsimony reconstruction) and to grasses (4281 new orthogroups) (Fig. 3 and tables S13 to S15). However, because this analysis does not include genome sequences from ferns and gymnosperms, it cannot distinguish between orthogroups originating with euphyllophytes (ferns plus seed plants), seed plants, or angiosperms; consequently, all of these orthogroups are reconstructed along the stem branch leading to angiosperms. We sorted the inferred gene set of the recently published Norway spruce genome (*38*), plus gymnosperm and basal

angiosperm transcript assemblies, into this gene classification, and manually reevaluated the origin of orthogroups around the MRCAs of seed plants and angiosperms, thereby resolving or refining the origin of 5210 orthogroups, 1179 (23%) of which are specific to angiosperms or have diverged sufficiently such that none of the gymnosperm homologs were detected, with 4031 (77%) present in the MRCA of seed plants (table S13).

The large number of orthogroups first appearing in angiosperms suggests that a diverse collection of novel gene functions was likely associated with the origin of flowering plants. Analyses of GO annotations for genes in angiosperm-derived orthogroups revealed the origin of orthogroups with functions associated with key innovations defining the flowering plant clade (table S16) (*17*). GO annotations related to reproduction (flower development, reproductive developmental process, pollination, and similar terms), including MADS-box gene lineages (see below), were overrepresented in this set of orthogroups. Genes with roles in *Arabidopsis* floral development (table S17) are included in 201 orthogroups, 18 of which were evolutionarily derived in the MRCA of angiosperms. Significant enrichments were also observed for several classes of regulatory genes (transcription, regulation of gene expression and of cellular, biochemical, and metabolic processes) as well as genes involved in various developmental processes. These include genes involved in carpel development (*CRABS CLAW*), endosperm development (*AGL62*), stem cell maintenance in meristems (*WUSCHEL*), and flowering time (*FRIGIDA*), suggesting that they might be key components underlying the origin of the flower.

Once a functional flower evolved, genetic innovations related to reproductive biology continued. Indeed, many gene lineages with genes inferred to have specific stamen (*39*), carpel (*39*), and ovule (*40*) functions apparently arose after the origin of angiosperms, within evolutionarily derived angiosperm lineages (table S18).

Whereas the origin of the flower may be partly explained by novel gene lineages that first appeared with the origin of the angiosperms, other floral genes, including putative B-class (that is, petal- and stamen-specific) gene targets (*41*), predate the origin of angiosperms. More than 70% of the gene lineages with known roles in flowering, including genes involved in floral timing and initiation (*CO*, *SOC1*, *VIN3*, *VEL1*), meristem identity (*ULT1*, *TFL2*), and floral structure (*AFO*, *AP2*, *ETT*, *HUA2*, *HEN4*, *KAN*, *RPL*, *JAG*), were present in the MRCA of all extant seed plants (table S16) (*17*). Orthogroups for other major components of the floral regulatory pathway are older still, with core components of the pathway present in the ancestral vascular plant (for example, *LFY*, phytochromes, *CLV*, *SKP1*, *GA1*, *SEU*, *HEN1*, and *FVE*).

Together, these observations suggest that orthologs of most floral genes existed long before their specific roles were established in flowering, and that they were later co-opted to serve floral functions. After the origin of angiosperms, new genes originated or were recruited to refine or more narrowly parse functions associated with flower development. This pattern is consistent with the observation that the floral organ transcriptional program is canalized (entrained) in eudicots relative to the less organ-constrained transcriptomes of earlier-diverging, less species-diverse angiosperm lineages (*42*).

Many of the novel gene lineages that first arose in angiosperms play no specific role in reproductive processes. Orthogroups containing genes with specific functions in vessel formation (*VND7* and *NAC083*) also first appeared at this time, even though *Amborella* does not produce vessels, but only tracheids (see below). Perhaps surprisingly, the most highly enriched GO terms in orthogroups derived in angiosperms were associated with homeostatic processes (GO:0042592; 18.9-fold enrichment). Relevant to the importance of plant-herbivore coevolution in the diversification of angiosperms and insects (*43*, *44*), the next most highly enriched GO classification was for genes involved in response to external stimuli (GO:0009605; 10.9-fold enrichment), including those with expression elicited by herbivory.

Enrichment patterns for functional categories were similar in the ancestral seed plant and ancestral angiosperm (table S16), including novel lineages of genes involved in reproductive, regulatory, and developmental processes. GO classifications associated with pollen-pistil interaction and epigenetic modification were enriched in orthogroups arising on the branch leading to seed plants, but not in the lineage leading to the ancestral angiosperm (table S16), the former perhaps indicating that some angiosperm-specific reproductive features predated angiospermous (enclosed ovule) reproduction.

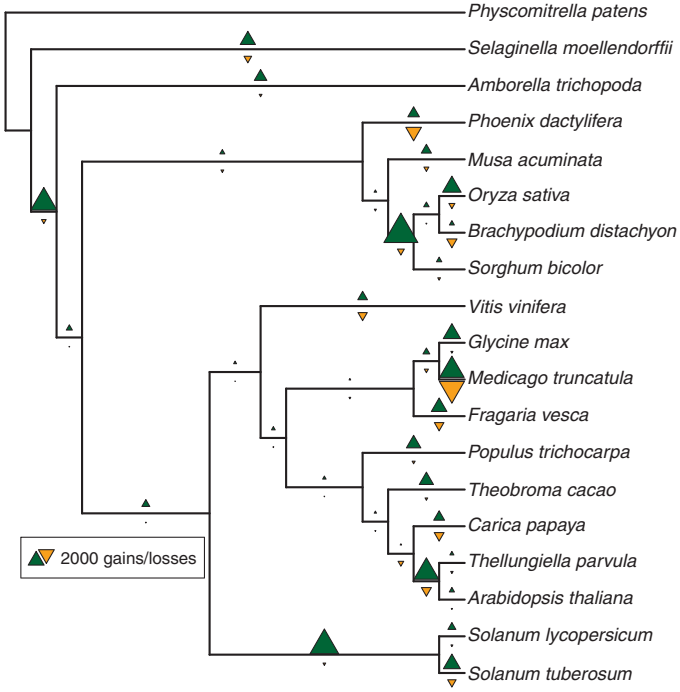


Fig. 3. Ancestral reconstruction of gene family content in land plants. Orthogroup gains and losses are inferred from the global gene family classification of proteins from sequenced plant genomes using a Wagner parsimony framework (*17*). Triangles are proportional to the number of orthogroup gains (green) and losses (orange). Actual values for the gains and losses in this analysis are provided in table S14; an analogous likelihood-based analysis is provided in table S15.

Gene Family Expansions in Angiosperms

Expansions of many gene families are evident in *Amborella*, and phylogenetic analyses indicate that such expansions occurred in the ancestral angiosperm, accompanying innovations associated with angiosperm origin. Using *Amborella* as a reference, we examined patterns of gene family diversification in angiosperm evolution, often in association with phenotypic divergence among angiosperm lineages.

MADS-Box Genes

MADS-box transcription factors are among the most important regulators of flower development. The *Amborella* genome encodes 36 MADS-box genes (table S19) (*17*), fewer than in other angiosperms (for example, *Arabidopsis* and rice), but consistent with the lack of a lineage-specific WGD. These genes belong to 21 clades, each of which includes genes from at least one other major lineage of angiosperms, implying that a minimum set of 21 MADS-box genes existed in the MRCA of extant angiosperms (figs. S19 and S20). The *Amborella* genome reveals that floral organ identity genes from eight major lineages (that is, *AP1*/*SQUA*, *AP3*/*DEF*, *PI*/*GLO*, *AG*, *STK*, *AGL2*/*SEP1*, *AGL9*/*SEP3*, and *AGL6*; Fig. 4A) existed in the MRCA of extant angiosperms and were likely derived from three ancestral lineages in the MRCA of extant seed plants. These data support the hypothesis that duplication and diversification of floral MADS-box genes likely occurred before the origin of extant angiosperms, despite being tightly associated with the origin of the flower. Furthermore, the previously presumed monocot-specific *OsMADS32* and eudicot-specific *TM8* gene lineages (fig. S20) (*45*–*47*) have orthologs in *Amborella*, suggesting that they were likely present in the earliest angiosperms and were subsequently lost in eudicots or monocots, respectively.

MADS-box transcription factors in floral development form dimers or higher-level complexes that bind to their targets with more complex patterns than those in gymnosperms (*48*). We conducted a comprehensive series of yeast two-hybrid assays among the *Amborella* floral MADS-box transcription factors. The protein-protein interaction (PPI) patterns in *Amborella* (fig. S21) are generally consistent with those in other angiosperms, and show clear differences from those in gymnosperms. For example, the B-function *AP3*/*DEF* and *PI*/*GLO* genes represent duplicate lineages in early angiosperms, arising after the divergence from the gymnosperms. The *Amborella* AP3 and PI homologs form heterodimers, as in other angiosperms, whereas the single AP3/PI homologs in gymnosperms form only homodimers, with heterodimers only occurring between recent duplicates of the AP3/PI homologs (Fig. 4B). B function is essential for the development of petals and other petal-like organs, which represent one of the most prominent novel floral features and exhibit extraordinary diversity in form; therefore, evolutionary shifts in PPI patterns after gene duplications, along with changes in gene sequence and expression patterns, likely have been crucial for functional innovations in the regulatory network for reproductive organ development and the origin of the flower (*49*), as well as for functional diversification of the many floral forms among lineages of angiosperms (*50*).

Glycogen Synthase Kinase 3 (GSK3) genes

GSK3 genes encode signal transduction proteins with roles in a variety of biological processes in eukaryotes. In contrast to their low copy numbers in animals, *GSK3* genes are numerous in land plants and have diverse functions, including floral development in angiosperms (*51*). Five *GSK3* loci that were present in the ancestral angiosperm have subsequently diversified among major angiosperm lineages, but a sixth ancestral locus has been detected only in *Amborella* (fig. S22). Thus, among flowering plants, *Amborella* alone may contain all the *GSK3* gene lineages that arose before the origin of extant angiosperms, underscoring the importance of *Amborella* for reconstructing the ancestral angiosperm genome (*52*).

Seed Storage Globulins

Seed storage proteins, including globulins, are critical for embryo and early seedling development in seed plants. These proteins are embedded in the very diverse cupin superfamily, which is distributed across the tree of life (*53*). The 11S legumin-type globulins are widespread across the seed plant phylogeny [for example, (*54*–*56*)]. Three distinct 11S legumin-type globulins have been identified in proteomic analyses of the globulin fraction in *Amborella* seeds (table S21) (*17*). Comparisons of the *Amborella* globulin-coding gene sequences to other seed plants revealed that key cysteine residues contributing to disulfide bonding between subunits and the absence of Intron IV, found in gymnosperms, are conserved characteristics of angiosperm legumins (fig. S25). In contrast, a conserved 52-residue region present in soybean, and thought to be important for mature hexamer formation (*57*), was apparently derived after the divergence of *Amborella* from other angiosperms (fig. S26). Globally, both structural (fig. S26) and phylogenetic (fig. S27 and table S22) analyses support the view that *Amborella* 11S globulins can both be reminiscent of those in monocots and eudicots and exhibit specific features of corresponding seed storage proteins in basal angiosperms and gymnosperms.

Terpene Synthase Genes

Terpenoids constitute the largest class of plant secondary metabolites and play important roles in plant ecological interactions (*58*). Biosynthesis of plant terpenoids is driven by terpene synthases (TPS). The *Amborella* TPS family contains more than 30 members, comparable in size to those of other angiosperms. However, the sesquiterpene synthase subfamily a (TPS-a), which is present in dicots, monocots, and Magnoliaceae but absent in gymnosperms and non-seed plants (*59*), is also absent in *Amborella* (fig. S28) (*17*). This indicates that the occurrence and diversification of this subfamily likely happened after the divergence of *Amborella* from other angiosperms, although its presence or absence in other basal angiosperms still needs to be established. Sesquiterpene synthases are involved in the production of C15 terpenoids, which are involved in diverse biological processes including the production of floral scents used to attract pollinators. *Amborella* lacks any detectable floral volatiles (*60*), and the expansion of the TPS-a subfamily may therefore have played an important role in the subsequent radiation of flowering plants.

Cell Wall and Lignin Genes

Secondary cell walls of woody plants contain lignin (*61*), facilitating water transport and mechanical support in xylem (*62*). Most gymnosperms (cycads,

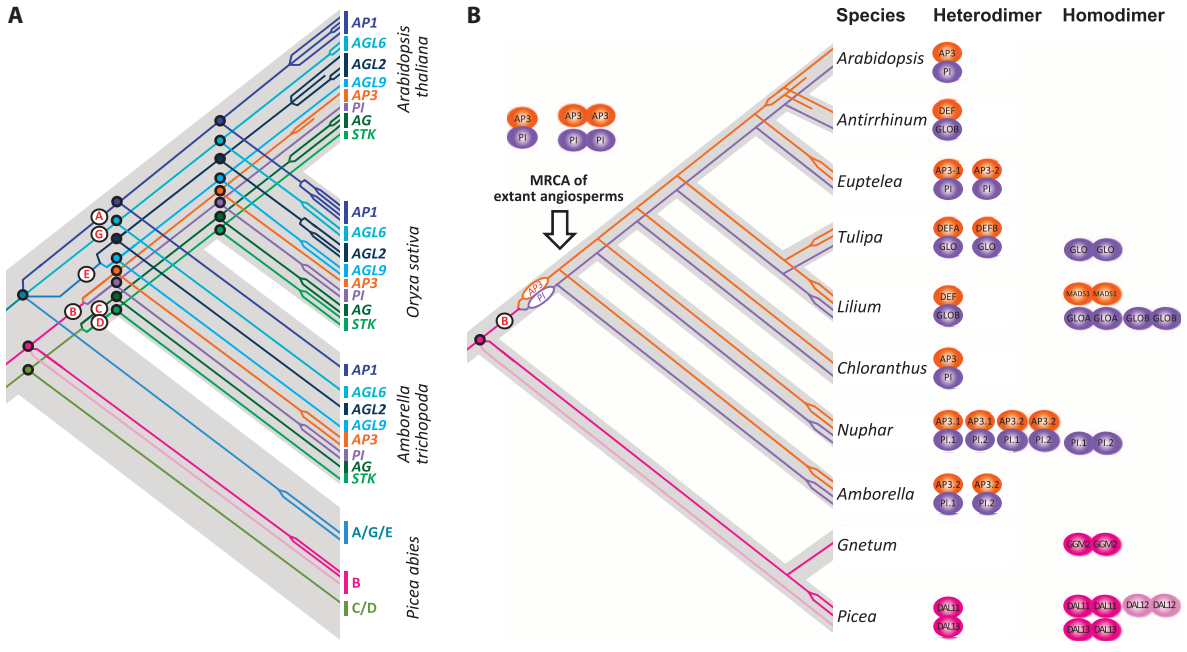


Fig. 4. *Amborella* as the reference for understanding the molecular developmental genetics of flower evolution. (**A**) A schematic diagram showing the evolutionary history of floral MADS-box genes. Note that all of the eight major gene lineages existed in the MRCA of extant angiosperms. (**B**) Evolutionary changes in the ability of B-class MADS-box proteins to form homodimers and heterodimers. In gymnosperms, the proteins of B-class genes can only form homodimers or semi-homodimers (that is, heterodimers formed by products of recently duplicated genes), whereas in the MRCA of extant angiosperms, they gained the ability to form heterodimers between members of different lineages. The interrupted lines represent previously described gene loss events (*85*, *86*).

Ginkgo, and conifers) have a predominant guaiacyl (G) subunit type of lignin, whereas the gnetophytes (<100 species) and the woody angiosperms have a lignin characterized by a copolymer of syringyl S and G subunits (S/G lignin). S/G lignin has also been found in the lycophyte *Selaginella moellendorffii*, suggesting that S/G lignin may have evolved more than once in plant evolution (*63*). S/G lignin is associated with cells involved in mechanical support, whereas G-type lignin has been associated with water transport (*64*). *Amborella* produces an S/G type of lignin, but the relative proportion of S subunits is much lower (13%) than values typical of woody angiosperms (50 to 70%) (tables S24 to S26). The low S/G ratio of *Amborella* might represent an ancestral condition that was transitional between gymnosperms and other angiosperms. However, the underlying genes of lignin precursor biosynthesis in *Amborella* are typical of woody angiosperms (table S27 and fig. S29, A to H). Although *Amborella* lacks vessels, in contrast to nearly all other angiosperms (*65*), the wood cell walls of *Amborella* are xylan-rich (table S24), typical of angiosperms (*66*). *Amborella* contains all of the carbohydrate-active enzyme families found in angiosperms (table S28) (*67*), but lacks many of the more derived clades of genes seen in other angiosperms. Indeed, much of the diversity of cell wall genes in angiosperms (for example, glycosyltransferase family 37; fig. S25) appears to result from gene duplication after the divergence of *Amborella* from other angiosperms.

Transposable Element Content in *Amborella*

As in the Norway spruce genome (*38*), the average age of identifiable transposable elements (TEs) in *Amborella* is considerably older than that of other angiosperm genomes [for example, (*68*–*70*)]. Likewise, ancient, full-length long terminal repeat (LTR) retrotransposons were identifiable in *Amborella* more than an estimated 40 million years after insertion (*17*). Wicker *et al.* (*71*) established the convention of separating LTR retrotransposons exhibiting more than 80% divergence in their terminal repeats into distinct families, but nearly 10% of individual *Amborella* LTR elements show a greater degree of divergence between their terminal repeats. Therefore, we used a clustering approach to circumscribe TE families. Median estimated insertion times for LTR subfamilies with two or more detectable TEs ranged from 4.0 to 17.6 Ma. A large class of *Gypsy* LTR retrotransposons with 502 annotated TEs experienced the most recent burst of activity 0.5 Ma (Fig. 5) (*17*). Endogenous pararetroviruses (EPRVs) were a relatively large component of the repeat landscape, comprising 2.4% of the assembled *Amborella* genome. Similar to *Sorghum bicolor*, which has a comparable genome size, TEs and EPRVs account for 57.2% of the nonambiguous nucleotides in the *Amborella* genome (668 Mb, table S30), but TE insertion times estimated for the *Amborella* genome are much older than inferred for *Sorghum* (*64*). Only four of the common superclasses of DNA TEs were observed (table S30); CACTA and TC1/Mariner-type elements were not detected. Most DNA TEs were highly degraded, with highly divergent sequences and missing terminal inverted repeats, again suggesting the persistence of identifiable elements over millions of years. The lack of recent transposon activity in the *Amborella* genome may be due to very effective silencing or the loss of active transposases.

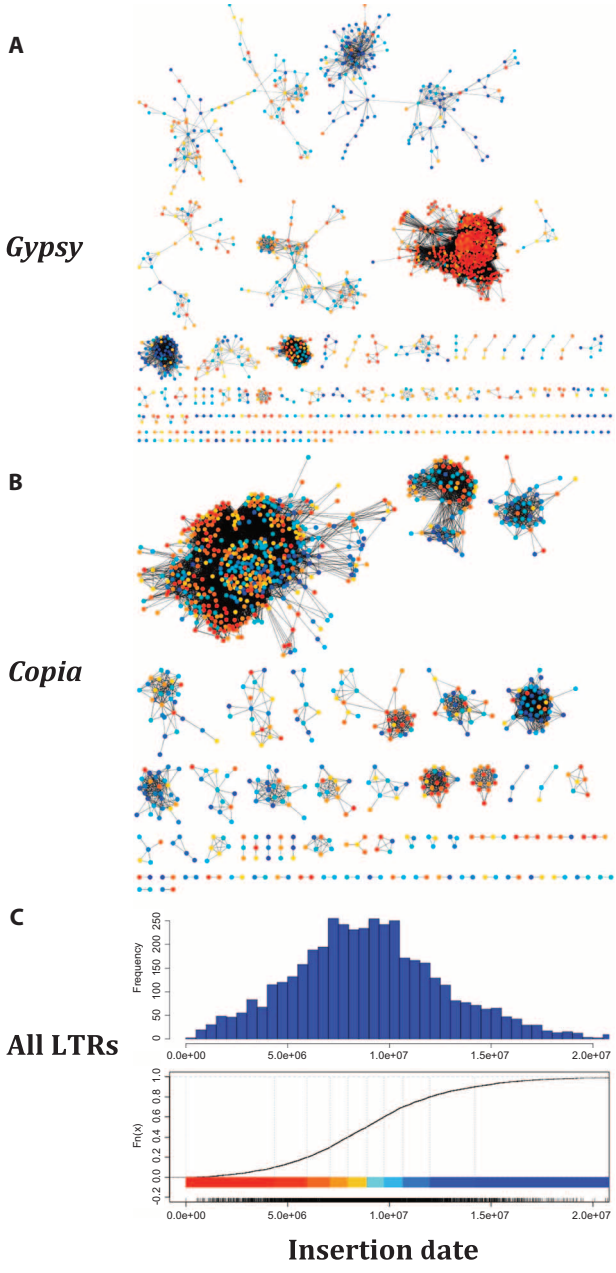


Fig. 5. Classification and insertion dates of LTR transposons in the *Amborella* genome. *Gypsy* (**A**) and *Copia* (**B**) LTR transposons are clustered into putative families, and individual elements are colored by their estimated insertion dates. Cool colors (for example, blue) represent older insertions, whereas warm colors (for example, red) represent more recent insertion dates. (**C**) Although some LTR transposon families have been active over the last 5 million years (for example, the large *Gypsy* cluster), the estimated insertion dates for the majority of elements are more than 10 Ma. See (*17*) for median insertion dates for each cluster (table S29s).

Evolution of Small RNAs

More than 56,000 discrete loci generating apparent regulatory small RNAs 20 to 24 nucleotides (nt) in size were identified by analysis of small RNA-seq data (*17*). Most small RNA loci had features consistent with those of heterochromatic small interfering RNAs (siRNAs) (*24*), indicating that heterochromatic siRNAs were present in the MRCA of all angiosperms. We also identified 124 *MIRNA* loci corresponding to 90 distinct families; 27 of these microRNA (miRNA) families, including 5 newly discovered ones, were likely present in the ancestral angiosperm. Most of these families (19 of them) are broadly conserved in other angiosperms, whereas 8 have evidence suggestive of later losses during angiosperm diversification. Inferred targets of the ancestral miRNA families were generally homologous to known miRNA-target relationships in other angiosperms, demonstrating that these relationships have been conserved since the earliest angiosperms despite the one-to-several rounds of polyploidy that separate *Amborella* from most other flowering plants. The other 63 miRNA families appear to be lineage-specific, and we could verify targets for just 14 of them. Surprisingly, most (78%) of

these lineage-specific miRNAs were 23 to 24 nt in size, rather than the 20- to 22-nt size typical of plant miRNAs. In contrast, none of the conserved miRNAs were 23 to 24 nt in size. The frequency of 23- to 24-nt miRNAs in *Amborella* is higher (>2×) than for any other land plant reported. Similar to the results for *Medicago* (*72*) and members of Solanaceae (*73*, *74*), several phased siRNA loci were nucleotide binding site–leucine-rich repeat (NB-LRR) disease resistance genes targeted by miRNAs in the miR482/2118 superfamily. Therefore, phased siRNA production from NB-LRR genes was likely present in the MRCA of angiosperms.

Population Genomics and Conservation Implications

Amborella is restricted to wet tropical forests on isolated slopes of New Caledonia. The genomes of 12 individuals of *Amborella*, sampled from nearly all known populations, were resequenced to assess the levels and patterns of genetic variation within this endemic species. These 12 individuals harbor levels of genetic diversity (θw = 0.0017, π = 0.0021) similar to those reported for species of *Populus*, which are also outcrossing perennials (table S45). The average Tajima's *D* across the genome (*75*) is positive (*D* = 0.8137), perhaps indicating balancing selection, although demographic processes such as population subdivision, a recent bottleneck, or migration can also produce a positive value. However, the genome exhibits significant among-locus and among-scaffold variance in allelic variation (fig. S40). Some regions, such as scaffold 1, are highly polymorphic and heterogeneous across their length, whereas other regions are nearly invariant with negative Tajima's *D* (for example, scaffold 31; fig. S40), consistent with multiple alternative explanations, such as recent selective sweeps and/or a mixed mating system.

The overall positive value of Tajima's *D* is consistent with a decrease in population size through time, as also demonstrated by an analysis of population genomic history using the pairwise sequentially Markovian coalescent (PSMC) (*76*) model, which has recently been applied to plant genomes (*77*). PSMC analysis of all 14 *Amborella* individuals, including the reference genome, the cultivated Bonn specimen, and the 12 locality-specific exemplars (Fig. 6A), reveals that the variation present in these modern genomes coalesces between 0.9 and 2 Ma. Confidence intervals for PSMC analyses of each individual are consistent with the hypothesis that at least two distinct *Amborella* sublineages with different levels of genetic diversity converged by 800,000 years ago, followed by admixture and a subsequent bottleneck event between 300,000 and 400,000 years ago, and by some recovery of genetic diversity thereafter. *Amborella* may therefore have undergone a series of population bottlenecks over the past 900,000 years, including one as recent as 100,000 years ago, represented by individual NCNAA (Fig. 6A). At the time of putative sublineage admixture (vertical line), effective population size (N_e), as averaged among all sequenced accessions, approximated 37,500 individuals, whereas in the recent event in NCNAA's past (where the PSMC plot reaches the ordinate axis), N_e may have been much lower at 5000 individuals or less (Fig. 6A). The reduction in N_e associated with any of these bottlenecks could have contributed to increased genetic structure among populations and linkage disequilibrium (LD). Increased LD may contribute to the size and persistence of genomic regions affected by selective sweeps, if they have occurred. Further analyses, with greater population sampling, are needed to distinguish the relative roles of selection, inbreeding, and other processes in shaping genome variability in *Amborella*.

Genetic variation among *Amborella* populations is significantly structured into four geographic clusters of populations on New Caledonia (Fig. 6B), corresponding roughly to populations in (i) the northern part of the range (blue cluster), (ii) the central part of the range (red cluster), (iii) a small region west of cluster 2, and (iv) a single disjunct location at the southern end of the distribution. These results are consistent with an independent analysis and extensive sampling of the 12 populations using microsatellite loci (*78*). Population genomic analyses tell a tale of dynamic genome evolution in this narrowly distributed plant species, the sole extant member of a lineage that shared a common ancestor with all other extant angiosperms about 160 Ma. Despite its restricted distribution, *Amborella* maintains substantial genetic diversity, with substructure among four population clusters. As ongoing effects of an expanding human population (for example, mining operations, fires, urbanization, and invasive species introduction) threaten the unique flora of this biodiversity hotspot, conservation efforts in New Caledonia should focus on preserving and managing the genetic diversity of New Caledonia's endemic species, including *A. trichopoda*.

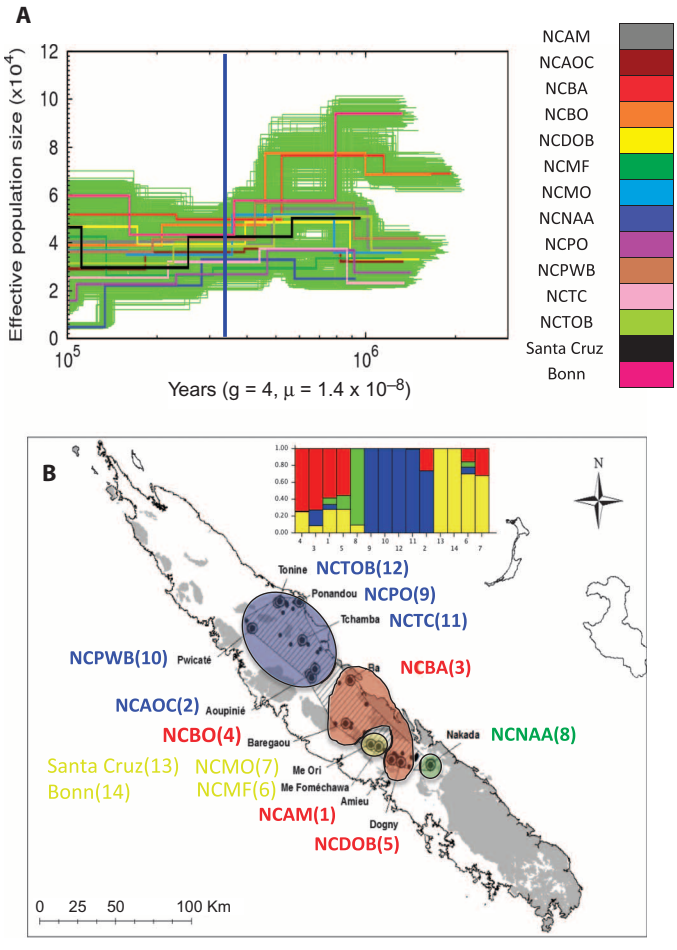


Fig. 6. Population genomic diversity in *Amborella*. (**A**) Plots of PSMC results for 14 individuals: 12 from separate populations on Grande-Terre, New Caledonia, the reference genome (Santa Cruz), and an additional cultivated individual (Bonn), indicated in the color panel (right) and with bootstrap clouds (for each genome analyzed) co-plotted in green. Times more recent than 10^5 years, where PSMC can be less reliable, are excluded. A vertical bar is drawn over the plot at about 325,000 years before present to indicate the timing of species-wide decline of effective population size, interpreted as a genetic bottleneck. (**B**) Results of STRUCTURE analysis, showing four significant genetic clusters of 12 individuals from natural populations. Additionally, the cultivated individual from the Bonn Botanical Garden and the reference genome (Santa Cruz) are clustered with individuals from Mé Foméchawa and Mé Ori [see (*17*)].

Conclusions

The phylogenetic position, conservation of genome structure, and absence of a lineage-specific polyploidy event have made the *Amborella* genome a unique and valuable reference that facilitates interpretation of major genomic events in flowering plant evolution, including the polyploid origin of angiosperms and a genomic hexaploidization event in eudicots. *Amborella* has enabled the identification of an ancestral gene set for angiosperms of at least 10,088 genes, including many that resulted from the ancestral angiosperm genome duplication, thereby helping to elucidate the origin of genes critical in flowering and other processes. The ancestral angiosperm-wide genome duplication apparent in the *Amborella* genome not only serves as a genetic marker for the origin of extant angiosperms, but it may also have set in motion a

series of events as numerous genes evolved novel functions, eventually leading to modern flowering plants. As the only extant member of an ancient lineage, *Amborella* provides a unique window into the earliest events in angiosperm evolution.

Materials and Methods

Sequencing and Assembly

Plant material for the reference genome sequence was obtained from a plant in cultivation since 1975 at the University of California at Santa Cruz Botanical Garden and additional clones located at the Atlanta Botanic Garden and the University of Florida. Single end genomic 454-FLX and SE 454-FLX+, DNA sequences, 11-kb paired-end 454-FLX reads, 3-kb PE Illumina HiSeq reads, and Sanger sequenced BAC end sequence reads were filtered to remove organellar contaminants, reads of short length or poor quality, artificial duplicates, and chimeras. After filtering, the read collection was pooled and assembled with the Roche Newbler assembler V2.6 [see (*18*) for details].

Genome Annotation and Database Development

Protein-coding genes, transposons, and endogenous viral sequences within the assembled genome were annotated iteratively using a variety of homology-based and de novo prediction algorithms integrated within the DAWGPAWS package (*21*). Initial gene model and transposon annotations were curated, and refined models were used to train ab initio prediction programs. The PASA annotation pipeline (*79*) was used to identify and classify alternative splicing events by aligning Newbler assembled 454 and Sanger expressed sequence tags (ESTs) and Trinity RNA-Seq assemblies. Three small RNA libraries and two degradome libraries were sequenced and used for annotation of small RNA–producing loci (including miRNAs, phased siRNAs, and heterochromatic siRNAs) and their targets. All resulting gene and transposon predictions, as well as alternative splicing annotations, have been placed in appropriate databases accessible through the *Amborella* Genome Database (http://www.amborella.org/) and National Center for Biotechnology Information (NCBI) (BioProject PRJNA212863).

Cytogenetics

Fluorescently labeled BACs were applied to mitotic chromosome spreads from root tips following Kato *et al.* (*80*). A Zeiss Axio Imager.M2 fluorescence microscope with an X-Cite Series 120 Q Lamp (EXFO Life Sciences) was used for visualization, and images were captured with a 100× objec7tive lens and a microscope-mounted AxioCam MRm digital camera (Zeiss) in conjunction with Axiovision version 4.8 software (Zeiss).

Synteny Analyses

For uncovering within-genome WGDs, we used the SynMap tool in the online CoGe portal (http://genomevolution.org/CoGe/), specifying a minimum number of colinear genes per window size to define putative syntenic regions. These regions were subsequently compared and confirmed using the microsynteny tool GEvo, also in CoGe. Blocks determined to represent the pan-angiosperm duplication event were further studied using phylogenomic methods to ascertain whether duplication patterns on trees concurred with a region-wide duplication model.

Scaffolds containing up to 10 orthologous and paralogous genes in common syntenic context from *Amborella* and three *gamma* subgenomes of three rosids were ordered using maximum weight matching to produce a hypothetical ancestral core eudicot genome with seven chromosomes. Each of the subgenomes mapped to virtually the whole length of the appropriate reconstructed chromosome. The reconstructed genes show a much clearer pattern of pan-rosid fractionation bias in extant genomes than is apparent without evidence derived from the *Amborella* genome scaffolds.

Global Gene Family Circumscription and Analysis

A global plant gene family classification was created using OrthoMCL (*81*) for the annotated protein set of *Amborella* and 21 other land plant genomes. The gene families (orthogroups) were populated with the gene models from the Norway spruce genome and a large collection of EST assemblies from basal angiosperms and other gymnosperms. We analyzed the evolutionary history of gain and loss of orthogroups and estimated the gene families present in the MRCA of living angiosperms using both parsimony and likelihood methods. Genome-wide analyses were performed, as well as more focused studies of genes with roles in flower development.

To study the history of ancient gene duplications in angiosperms and seed plants, we performed maximum likelihood phylogenetic analysis of 11,519 orthogroups that contained *Amborella* genes. Gene duplications were scored on the basis of taxa present in the daughter lineages to identify angiosperm-wide, seed plant–wide, and *Amborella*-specific gene duplications (*7*). Possible genome duplications were identified from statistically significant peaks in the distributions of synonymous divergences and estimated ages of gene duplication events. Six of the largest syntenic blocks in the *Amborella* genome were also used for manual curation of syntenic duplicates and phylogenetic analysis of gene families containing duplicated genes present on paralogous genomic blocks.

Targeted Gene Family Analyses

To illustrate the value of the *Amborella* genome as a reference for understanding the evolutionary history of gene families associated with angiosperm innovations or divergence among angiosperm lineages, we examined the phylogenetic history of MADS-box, GSK3, TPS, and cell wall and lignin genes. Yeast two-hybrid analysis of MADS-box proteins in *Amborella* was used to identify heterodimeric PPIs found only in angiosperms. Proteomic and phylogenetic analysis of seed storage globulin proteins validated protein-coding gene models as well as examined protein features that separate angiosperms from earlier land plant lineages.

Population Genomics

To assess the levels and patterns of genetic variation within *A. trichopoda*, we sequenced the genomes of 12 individuals representing nearly all of the known natural populations of the species, the reference plant, and an additional accession from the Bonn Botanical Garden. Sequences were mapped to the reference genome using BWA. We used basic population genetic measures to infer levels of diversity and applied the PSMC model, originally applied to human and other mammalian genomes, to study the effective population size (N_e) of *Amborella* over time. Genetic divergence among populations was assessed using STRUCTURE.

References and Notes

1. C. Darwin, in *Letter to Hooker*, F. Darwin, A. C. Seward, Eds. (John Murray, London, UK, 1903).
2. C. D. Bell, D. E. Soltis, P. S. Soltis, The age and diversification of the angiosperms re-revisited. *Am. J. Bot.* **97**, 1296–1303 (2010). doi: 10.3732/ajb.0900346; pmid: 21616882
3. E. M. Friis, K. R. Pedersen, P. R. Crane, Cretaceous angiosperm flowers: Innovation and evolution in plant reproduction. *Palaeogeogr. Palaeocl.* **232**, 251–293 (2006). doi: 10.1016/j.palaeo.2005.07.006
4. S. Magallón, K. W. Hilu, D. Quandt, Land plant evolutionary timeline: Gene effects are secondary to fossil constraints in relaxed clock estimation of age and substitution rates. *Am. J. Bot.* **100**, 556–573 (2013). doi: 10.3732/ajb.1200416; pmid: 23445823
5. J. A. Doyle, Molecular and fossil evidence on the origin of angiosperms. *Annu. Rev. Earth Planet Sci.* **40**, 301–326 (2012). doi: 10.1146/annurev-earth-042711-105313
6. N. Zhang, L. Zeng, H. Shan, H. Ma, Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. *New Phytol.* **195**, 923–937 (2012). doi: 10.1111/j.1469-8137.2012.04212.x; pmid: 22783877
7. Y. Jiao *et al.*, Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97–100 (2011). doi: 10.1038/nature09916; pmid: 21478875
8. D. E. Soltis, C. D. Bell, S. Kim, P. S. Soltis, Origin and early evolution of angiosperms. *Ann. N. Y. Acad. Sci.* **1133**, 3–25 (2008). doi: 10.1196/annals.1438.005; pmid: 18559813
9. R. K. Jansen *et al.*, Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 19369–19374 (2007). doi: 10.1073/pnas.0709121104; pmid: 18048330
10. M. J. Moore, C. D. Bell, P. S. Soltis, D. E. Soltis, Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 19363–19368 (2007). doi: 10.1073/pnas.0708072104; pmid: 18048334
11. M. J. Moore, P. S. Soltis, C. D. Bell, J. G. Burleigh, D. E. Soltis, Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 4623–4628 (2010). doi: 10.1073/pnas.0907801107; pmid: 20176954
12. M. J. Moore *et al.*, Phylogenetic analysis of the plastid inverted repeat for 244 species: Insights into deeper-level angiosperm relationships from a long, slowly evolving

sequence region. *Int. J. Plant Sci.* **172**, 541–558 (2011). doi: 10.1086/658923

13. J. G. Burleigh *et al.*, Genome-scale phylogenetics: Inferring the plant tree of life from 18,896 gene trees. *Syst. Biol.* **60**, 117–125 (2011). doi: 10.1093/sysbio/syq072; pmid: 21186249
14. E. K. Lee *et al.*, A functional phylogenomic view of the seed plants. *PLoS Genet.* **7**, e1002411 (2011). doi: 10.1371/journal.pgen.1002411; pmid: 22194700
15. D. E. Soltis *et al.*, Angiosperm phylogeny: 17 genes, 640 taxa. *Am. J. Bot.* **98**, 704–730 (2011). doi: 10.3732/ajb.1000404; pmid: 21613169
16. While *Amborella* and water lilies (Nymphaeales) together have been reported to form the sister group to other angiosperms in some analyses (*87*–*89*), this topology does not impact any of the inferences made in this paper.
17. Supplementary materials for this article are available on *Science* Online.
18. S. Chamala *et al.*, Assembly and validation of the genome of the nonmodel basal angiosperm *Amborella*. *Science* 10.1126/science.1241130 (2013).
19. I. J. Leitch, L. Hanson, DNA C-values in seven families fill phylogenetic gaps in the basal angiosperms. *Bot. J. Linn. Soc.* **140**, 175–179 (2002). doi: 10.1046/j.1095-8339.2002.00096.x
20. A. Zuccolo *et al.*, A physical map for the *Amborella trichopoda* genome sheds light on the evolution of angiosperm genome structure. *Genome Biol.* **12**, R48 (2011). doi: 10.1186/gb-2011-12-5-r48; pmid: 21619600
21. J. C. Estill, J. L. Bennetzen, The DAWGPAWS pipeline for the annotation of genes and transposable elements in plant genomes. *Plant Methods* **5**, 8 (2009). doi: 10.1186/1746-4811-5-8; pmid: 19545381
22. B. J. Haas *et al.*, Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008). doi: 10.1186/gb-2008-9-1-r7; pmid: 18190707
23. S. Takuno, B. S. Gaut, Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 1797–1802 (2013). doi: 10.1073/pnas.1215380110; pmid: 23319627
24. J. A. Law, S. E. Jacobsen, Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11**, 204–220 (2010). doi: 10.1038/nrg2719; pmid: 20142834
25. A. Zemach, I. E. McDaniel, P. Silva, D. Zilberman, Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* **328**, 916–919 (2010). doi: 10.1126/science.1186366; pmid: 20395474
26. Y. Jiao *et al.*, A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* **13**, R3 (2012). doi: 10.1186/gb-2012-13-1-r3; pmid: 22280555
27. O. Jaillon *et al.*, The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007). doi: 10.1038/nature06148; pmid: 17721507
28. X. Argout *et al.*, The genome of *Theobroma cacao*. *Nat. Genet.* **43**, 101–108 (2011). doi: 10.1038/ng.736; pmid: 21186351
29. E. Lyons *et al.*, Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol.* **148**, 1772–1781 (2008). doi: 10.1104/pp.108.124867; pmid: 18952863
30. International Peach Genome Initiative *et al.*, The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat. Genet.* **45**, 487–494 (2013). pmid: 23525075
31. C. Zheng, K. Swenson, E. Lyons, D. Sankoff, in *Algorithms in Bioinformatics* (Springer, Berlin, 2011), pp. 364–375.
32. Z. Galil, Efficient algorithms for finding maximum matching in graphs. *ACM Comput. Surv.* **18**, 23–38 (1986). doi: 10.1145/6462.6502
33. S. Jung *et al.*, Whole genome comparisons of *Fragaria*, *Prunus* and *Malus* reveal different modes of evolution between Rosaceous subfamilies. *BMC Genomics* **13**, 129 (2012). doi: 10.1186/1471-2164-13-129; pmid: 22475018
34. D. E. Soltis *et al.*, Polyploidy and angiosperm diversification. *Am. J. Bot.* **96**, 336–348 (2009). doi: 10.3732/ajb.0800079; pmid: 21628192
35. Y. Van de Peer, A mystery unveiled. *Genome Biol.* **12**, 113 (2011). doi: 10.1186/gb-2011-12-5-113; pmid: 21635712
36. S. Proost, P. Pattyn, T. Gerats, Y. Van de Peer, Journey through the past: 150 million years of plant genome evolution. *Plant J.* **66**, 58–65 (2011). doi: 10.1111/j.1365-313X.2011.04521.x; pmid: 21443623
37. L. Cui *et al.*, Widespread genome duplications throughout the history of flowering plants. *Genome Res.* **16**, 738–749 (2006). doi: 10.1101/gr.4825606; pmid: 16702410
38. B. Nystedt *et al.*, The Norway spruce genome sequence and conifer genome evolution. *Nature* **497**, 579–584 (2013). doi: 10.1038/nature12211; pmid: 23698360
39. F. Wellmer, J. L. Riechmann, M. Alves-Ferreira, E. M. Meyerowitz, Genome-wide analysis of spatial gene expression in Arabidopsis flowers. *Plant Cell* **16**, 1314–1326 (2004).doi: 10.1105/tpc.021741; pmid: 15100403
40. D. J. Skinner, C. S. Gasser, Expression-based discovery of candidate ovule development regulators through transcriptional profiling of ovule mutants. *BMC Plant Biol.* **9**, 29 (2009). doi: 10.1186/1471-2229-9-29; pmid: 19291320
41. S. E. Wuest *et al.*, Molecular basis for the specification of floral organs by APETALA3 and PISTILLATA. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 13452–13457 (2012). doi: 10.1073/pnas.1207075109; pmid: 22847437
42. A. S. Chanderbali *et al.*, Conservation and canalization of gene expression during angiosperm diversification accompany the origin and evolution of the flower. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 22570–22575 (2010). doi: 10.1073/pnas.1013395108; pmid: 21149731
43. P. R. Ehrlich, P. H. Raven, Butterflies and plants—A study in coevolution. *Evolution* **18**, 586–608 (1964).
44. D. J. Futuyma, A. A. Agrawal, Macroevolution and the biological diversity of plants and herbivores. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 18054–18061 (2009). doi: 10.1073/pnas.0904106106; pmid: 19815508
45. R. Arora *et al.*, MADS-box gene family in rice: Genome-wide identification, organization and expression profiling during reproductive development and stress. *BMC Genomics* **8**, 242 (2007). doi: 10.1186/1471-2164-8-242; pmid: 17640358
46. J. Díaz-Riquelme, D. Lijavetzky, J. M. Martínez-Zapater, M. J. Carmona, Genome-wide analysis of MIKC^{C}-type MADS box genes in grapevine. *Plant Physiol.* **149**, 354–369 (2009). doi: 10.1104/pp.108.131052; pmid: 18997115
47. C. H. Leseberg, A. Li, H. Kang, M. Duvall, L. Mao, Genome-wide analysis of the MADS-box gene family in *Populus trichocarpa*. *Gene* **378**, 84–94 (2006). doi: 10.1016/j.gene.2006.05.022; pmid: 16831523
48. S. De Bodt, J. Raes, Y. Van de Peer, G. Theissen, And then there were many: MADS goes genomic. *Trends Plant Sci.* **8**, 475–483 (2003). doi: 10.1016/j.tplants.2003.09.006; pmid: 14557044
49. T. Hernández-Hernández, L. P. Martínez-Castilla, E. R. Alvarez-Buylla, Functional diversification of B MADS-box homeotic regulators of flower development: Adaptive evolution in protein–protein interaction domains after major gene duplication events. *Mol. Biol. Evol.* **24**, 465–481 (2007). doi: 10.1093/molbev/msl182; pmid: 17135333
50. C. Liu *et al.*, Interactions among proteins of floral MADS-box genes in basal eudicots: Implications for evolution of the regulatory network for flower development. *Mol. Biol. Evol.* **27**, 1598–1611 (2010). doi: 10.1093/molbev/msq044; pmid: 20147438
51. Y. Saidi, T. J. Hearn, J. C. Coates, Function and evolution of "green" GSK3/Shaggy-like kinases. *Trends Plant Sci.* **17**, 39–46 (2012). doi: 10.1016/j.tplants.2011.10.002; pmid: 22051150
52. X. Qi, A. S. Chanderbali, G. K. S. Wong, D. E. Soltis, P. S. Soltis, Phylogeny and evolutionary history of glycogen synthase kinase 3/SHAGGY-like kinase genes in land plants. *BMC Evol. Biol.* **13**, 143 (2013). doi: 10.1186/1471-2148-13-143; pmid: 23834366
53. J. M. Dunwell, A. Purvis, S. Khuri, Cupins: The most functionally diverse protein superfamily? *Phytochemistry* **65**, 7–17 (2004). doi: 10.1016/j.phytochem.2003.08.016; pmid: 14697267
54. K. P. Häger, B. Müller, C. Wind, S. Erbach, H. Fischer, Evolution of legumin genes: Loss of an ancestral intron at the beginning of angiosperm diversification. *FEBS Lett.* **387**, 94–98 (1996). doi: 10.1016/0014-5793(96)00477-2; pmid: 8654576
55. M. Adachi *et al.*, Crystal structure of soybean 11S globulin: Glycinin A3B4 homohexamer. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 7395–7400 (2003). doi: 10.1073/pnas.0832158100; pmid: 12771376
56. C. Li, M. Li, J. M. Dunwell, Y. M. Zhang, Gene duplication and an accelerated evolutionary rate in 11S globulin genes are associated with higher protein synthesis in dicots as compared to monocots. *BMC Evol. Biol.* **12**, 15 (2012). doi: 10.1186/1471-2148-12-15; pmid: 22292855
57. M. R. Tandang-Silvas *et al.*, Conservation and divergence on plant seed 11S globulins based on crystal structures. *Biochim. Biophys. Acta* **1804**, 1432–1442 (2010). doi: 10.1016/j.bbapap.2010.02.016; pmid: 20215054
58. J. Gershenzon, N. Dudareva, The function of terpene natural products in the natural world. *Nat. Chem. Biol.* **3**, 408–414 (2007). doi: 10.1038/nchembio.2007.5; pmid: 17576428
59. F. Chen, D. Tholl, J. Bohlmann, E. Pichersky, The family of terpene synthases in plants: A mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *Plant J.* **66**, 212–229 (2011). doi: 10.1111/j.1365-313X.2011.04520.x; pmid: 21443633
60. L. B. Thien *et al.*, The population structure and floral biology of *Amborella trichopoda* (Amborellaceae). *Ann. Missouri Bot. Gard.* **90**, 466–490 (2003). doi: 10.2307/3298537
61. P. Albersheim, A. Darvill, K. Roberts, R. Sederoff, A. Staehelin, *Plant Cell Walls* (Garland Science, New York, 2011).
62. K. V. Sarkanen, C. H. Ludwig, *Lignins—Occurrence, Formation, Structure and Reactions* (John Wiley & Sons, New York, 1971).
63. J. K. Weng, X. Li, J. Stout, C. Chapple, Independent origins of syringyl lignin in vascular plants. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 7887–7892 (2008). doi: 10.1073/pnas.0801696105; pmid: 18505841
64. L. A. Donaldson, Lignification and lignin topochemistry—An ultrastructural view. *Phytochemistry* **57**, 859–873 (2001). doi: 10.1016/S0031-9422(01)00049-8; pmid: 11423137
65. T. S. Feild *et al.*, Structure and function of tracheary elements in *Amborella trichopoda*. *Int. J. Plant Sci.* **161**, 705–712 (2000). doi: 10.1086/314293
66. T. E. Timell, Wood hemicelluloses: Part II. *Adv. Carbohydr. Chem.* **20**, 409–483 (1965). doi: 10.1016/S0096-5332(08)60304-5
67. B. L. Cantarel *et al.*, The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenomics. *Nucleic Acids Res.* **37**, D233–D238 (2009). doi: 10.1093/nar/gkn663; pmid: 18838391
68. R. S. Baucom *et al.*, Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genet.* **5**, e1000732 (2009). doi: 10.1371/journal.pgen.1000732; pmid: 19936065
69. T. Wicker, B. Keller, Genome-wide comparative analysis of *copia* retrotransposons in Triticeae, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual *copia* families. *Genome Res.* **17**, 1072–1081 (2007). doi: 10.1101/gr.6214107; pmid: 17556529
70. Tomato Genome Consortium, The tomato genome sequence provides insights into fleshy fruit evolution, *Nature* **485**, 635–641 (2012). doi: 10.1038/nature11119; pmid: 22660326
71. T. Wicker *et al.*, A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.*

8, 973–982 (2007). doi: 10.1038/nrg2165; pmid: 17984973
72. J. Zhai *et al.*, MicroRNAs as master regulators of the plant *NB-LRR* defense gene family via the production of phased, *trans*-acting siRNAs. *Genes Dev.* **25**, 2540–2553 (2011). doi: 10.1101/gad.177527.111; pmid: 22156213
73. F. Li *et al.*, MicroRNA regulation of plant innate immune receptors. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 1790–1795 (2012). doi: 10.1073/pnas.1118282109; pmid: 22307647
74. P. V. Shivaprasad *et al.*, A microRNA superfamily regulates nucleotide binding site–leucine-rich repeats and other mRNAs. *Plant Cell* **24**, 859–874 (2012). doi: 10.1105/tpc.111.095380; pmid: 22408077
75. A. J. Eckert, J. D. Liechty, B. R. Tearse, B. Pande, D. B. Neale, DnaSAM: Software to perform neutrality testing for large datasets with complex null models. *Mol. Ecol. Resour.* **10**, 542–545 (2010). doi: 10.1111/j.1755-0998.2009.02768.x; pmid: 21565054
76. H. Li, R. Durbin, Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011). doi: 10.1038/nature10231; pmid: 21753753
77. E. Ibarra-Laclette *et al.*, Architecture and evolution of a minute plant genome. *Nature* **498**, 94–98 (2013). doi: 10.1038/nature12132; pmid: 23665961
78. V. Poncet *et al.*, Phylogeography and niche modelling of the relict plant *Amborella trichopoda* (Amborellaceae) reveal multiple Pleistocene refugia in New Caledonia. *Mol. Ecol.* (2013). doi: 10.1111/mec.12554; pmid: 24118476
79. B. J. Haas *et al.*, Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003). doi: 10.1093/nar/gkg770; pmid: 14500829
80. A. Kato *et al.*, Chromosome painting for plant biotechnology. *Methods Mol. Biol.* **701**, 67–96 (2011). doi: 10.1007/978-1-61737-957-4_4; pmid: 21181525
81. L. Li, C. J. Stoeckert Jr., D. S. Roos, OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003). doi: 10.1101/gr.1224503; pmid: 12952885
82. R. Ming *et al.*, Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.). *Genome Biol.* **14**, R41 (2013). doi: 10.1186/gb-2013-14-5-r41; pmid: 23663246
83. M. S. Barker, in *Plant Genome Diversity*, J. Greilhuber, J. Dolezel, J. F. Wendel, Eds. (Springer, Vienna, 2013), vol. 2, pp. 245–253.
84. M. S. Barker *et al.*, Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol. Biol. Evol.* **25**, 2445–2455 (2008). doi: 10.1093/molbev/msn187; pmid: 18728074
85. T. Gerats, M. Vandenbussche, A model system for comparative research: *Petunia*. *Trends Plant Sci.* **10**, 251–256 (2005). doi: 10.1016/j.tplants.2005.03.005; pmid: 15882658
86. R. S. Lamb, V. F. Irish, Functional divergence within the APETALA3/PISTILLATA floral homeotic gene lineages. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 6558–6563 (2003). doi: 10.1073/pnas.0631708100; pmid: 12746493
87. Y. L. Qiu *et al.*, Angiosperm phylogeny inferred from sequences of four mitochondrial genes. *J. Syst. Evol.* **48**, 391–425 (2010). doi: 10.1111/j.1759-6831.2010.00097.x
88. V. V. Goremykin *et al.*, The evolutionary root of flowering plants. *Syst. Biol.* **62**, 50–61 (2013). doi: 10.1093/sysbio/sys070; pmid: 22851550
89. C. Finet, R. E. Timme, C. F. Delwiche, F. Marlétaz, Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. *Curr. Biol.* **20**, 2217–2222 (2010). doi: 10.1016/j.cub.2010.11.035; pmid: 21145743

Acknowledgments: Sequencing reads, reference genome assembly, and gene annotations of *Amborella trichopoda* are available from NCBI (BioProject PRJNA212863). The *Amborella* genome is also available in CoGe (http://genomevolution.org/CoGe/) and at the *Amborella* Genome Database (http://www.amborella.org), where additional tools for comparative genomic analysis are available. This work was funded by the NSF Plant Genome Research Program (grant 0922742) to C.W.D., H.M., W.B.B., P.S.S., D.E.S., V.A.A., J.L.-M., S.R.W., J.D.P., and S.R., with additional funding from NSF's iPlant Collaborative to P.S.S. and D.E.S. Author contributions are included in the Supplementary Materials.

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,[19] 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,[1,20] 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,[4,23] 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]

[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 Sciences and 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, Ontario 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]Technische Universität Dresden, Institut für Botanik, 01062 Dresden, Germany. [20]Department of Biology, Chongqing University of Science and Technology, Chongqing 4000042, China. [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, 1871 Frederiksberg C, 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 and 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), UMR–Diversité, Adaptation et Développement des Plantes (DIADE), BP 64501, F-34394 Montpellier cedex 5, France. [43]Laboratorio Nacional de Genomica para la Biodiversidad, 36821 Irapuato, Mexico. [44]Scientific Computing Group, Heidelberg Institute for Theoretical Studies, Heidelberg 69118, Germany.

Supplementary Materials

www.sciencemag.org/content/342/6165/1241089/suppl/DC1
Text
Figs. S1 to S42
Tables S1 to S46
Additional Acknowledgment
References

10.1126/science.1241089