Title
The Amborella genome and the evolution of flowering plants.

Permalink
https://escholarship.org/uc/item/06v4x3nz

Journal
Science (New York, N.Y.), 342(6165)

ISSN
0036-8075

Author
Amborella Genome Project

Publication Date
2013-12-01

DOI
10.1126/science.1241089

Peer reviewed
The *Amborella* Genome and the Evolution of Flowering Plants

*Amborella Genome Project*†

**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 eu dicots (~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 ancient 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**

- **Fig. 1.** *Amborella* is sister to all other extant angiosperms.
- **Fig. 2.** Synteny analysis of *Amborella*.
- **Fig. 3.** Ancestral reconstruction of gene family content in land plants.
- **Fig. 4.** *Amborella* as the reference for understanding the molecular developmental genetics of flower evolution.
- **Fig. 5.** Classification and insertion dates of LTR transposons in the *Amborella* genome.
- **Fig. 6.** Population genomic diversity in *Amborella*.

**Supplementary Materials**

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

**Related Items in Science**


*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 duplications. 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.

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 Arabidopsis 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 polyploid 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

*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.
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 non-tandem gene pairs), especially when compared to older gene family expansions shared across angiosperm or seed plant lineages (473 orthologous 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 syntenies, 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 syntenic-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 paralogous 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).

**Ancient 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. 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.
angiosperm transcript assemblies, into this gene classification, and manually reevaluated the origin of orthogroups around the MRCA 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 gene lineages 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 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.

**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 OeMADS32 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.

---

**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.
MADS-box transcription factors play a role in floral development, forming dimers or higher-level complexes that bind to their targets with complex patterns. 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 PIGL 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 were identified in the ancestral angiosperm lineage and 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 can maintain 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 Introns IV, found in gymnosperms, are conserved characteristics of angiosperm legumin (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 nonseed 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,
**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.

**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. Inferrred 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

---

**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 S29a).
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 (Ne), 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), Ne may have been much lower at 5000 individuals or less (Fig. 6A). The reduction in Ne 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.

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× objective 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 (genomeevolution.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 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 subgenera 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 in 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 (Ne) of Amborella over time. Genetic divergence among populations was assessed using STRUCTURE.

References and Notes

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
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 0921789-C.W.D., H.R., W.B.R., P.S.S., D.E.S., V.A.K., J.L., S.M., D.P., U.D.P., and additional funding from NSF's Plant Collaborative to P.S.S. and D.E.S. Author contributions are included in the Supplementary Materials.

Authorship information

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. Andre S. Chanderbali, B. J. Haas, Claude Bernard Lyon, and the author for correspondence (*) are indicated with- out order.

Research leadership: Victor A. Albert1,2, W. Bradley Babick3,2, Claude W. dePamphilis4,5,6,† (wdp@psu.edu), Joshua P. Der,6,4,† (jderr@buffalo.edu), Ronald Determann,18 James Leebens-Mack,7,4,8 (jleebensmack@uga.edu), Hong Ma,4,8,17 (Hong.Ma@buffalo.edu), and Blake C. Meyers,21 (blakecm@psu.edu). The Amborella Genome Project.


1. Materials and methods

1.1. Plant materials

Amborella trichopoda fruit material was collected from the field site at the University of Florida. DNA was extracted from fresh leaf material using the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany). Total DNA was quantified using a NanoDrop 2000c spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA).

1.2. Genome assembly

A draft genome assembly was generated using the CLC Genomics Workbench (CLC bio, Aarhus, Denmark) and the SPAdes Genome Assembler (Bankevich et al., 2012). The assembly was circularized using the circular genome tool in the GenomeTrio application (CCGC, France) and was then polished using the Pilon software (Hickey et al., 2012) and the SNPfix tool in the GenomeTrio application (CCGC, France).

1.3. Gene prediction

Gene prediction was performed using the AUGUSTUS gene prediction program (Zdobnov and Seluanov, 2001) with the default parameters. The protein sequences were then predicted using the TransDecoder tool (Kretzschmar et al., 2015).

1.4. Genome annotation

The genome was annotated using a combination of BLAST (Altschul et al., 1990), InterProScan (Zdobnov and Apweiler, 2001), and Gene Ontology (Ashburner et al., 2000) searches. The annotations were further refined using the Cytoscape software (Shannon et al., 2003) and the UCSC Genome Browser (Kent et al., 2002).

1.5. Genome statistics

The genome statistics were calculated using the RepeatMasker software (Smit et al., 2000) and the RepeatFinder tool (Zhang et al., 2014). The number of genes was estimated using the BWA-MEM tool (Li and Durbin, 2009) and the Tophat software (Kim et al., 2009). The gene expression levels were calculated using the RNA-Seq tool (Mortazavi et al., 2008) and the DESeq software (Anders and Huber, 2010).