Genome Biology

Genome Biology

A physical map for the Amborella trichopoda genome sheds light on the evolution of angiosperm genome structure

Genome Biology 2011, **12**:R48 doi:10.1186/gb-2011-12-5-r48

Andrea Zuccolo (azuccolo@ag.arizona.edu)
John E Bowers (jebowers@uga.edu)
James C Estill (jamesestill@gmail.com)
Zhiyong Xiong (xiongz@missouri.edu)
Meizhong Luo (mzluo@mail.hzau.edu.cn)
Asthway Sebastian (aswathyseb@gmail.com)
Jose' LUIS Goicoechea (jlgoicoe@ag.arizona.edu)
Kristi Collura (kcollur@Ag.arizona.edu)
Yeisoo Yu (yeisooyu@ag.arizona.edu)
Yuannian Jiao (yxj129@psu.edu)
Jill Duarte (jmr464@psu.edu)
Haibao Tang (tanghaibao@gmail.com)
Saravanaraj Ayyampalayam (raj@plantbio.uga.edu)
Steve Rounsley (rounsley@email.arizona.edu)
Dave Kudma (dkudrna@ag.arizona.edu)
Andrew H Paterson (paterson@plantbio.uga.edu)
J CHRIS Pires (piresjc@missouri.edu)
Andre Chanderbali (achander@ufl.edu)
Douglas E Soltis (psoltis@flmnh.ufl.edu)
Srikar Chamala (vaalbert@buffalo.edu)
Brad Barbazuk (bbarbazuk@ufl.edu)
Pamela S Soltis (psoltis@flmnh.ufl.edu)
Victor A Albert (vaalbert@buffalo.edu)
Hong Ma (hxm16@psu.edu)
Dina Mandoli (dina.mandoli@mac.com)
Jody Banks (banksj@purdue.edu)
John E Carlson (jec16@psu.edu)
Jeffrey Tomkins (jtmkns@clemson.edu)
Claude W dePamphilis (cwd3@psu.edu)
Rod A Wing (rwing@ag.arizona.edu)
Jim Leebens-Mack (jleebensmack@plantbio.uga.edu)

ISSN 1465-6906

Article type Research

Genome Biology

Genome Biology

Submission date 26 November 2010

Acceptance date 27 May 2011

Publication date 27 May 2011

Article URL http://genomebiology.com/2011/12/5/R48

A physical map for the *Amborella trichopoda* genome sheds light on the evolution of angiosperm genome structure

Andrea Zuccolo[1], John E Bowers[2], James C Estill[2], Zhiyong Xiong[3], Meizhong Luo[1,4], Aswathy Sebastian[1], Jose' Luis Goicoechea[1], Kristi Collura[1], Yeisoo Yu[1], Yuannian Jiao[5], Jill Duarte[5], Haibao Tang[2,6,9], Saravanaraj Ayyampalayam[2], Steve Rounsley[7,8], Dave Kudrna[1], Andrew H Paterson[2,9], J Chris Pires[3], Andre Chanderbali[10], Douglas E Soltis[10], Srikar Chamala[10], Brad Barbazuk[10], Pamela S Soltis[11], Victor A Albert[12], Hong Ma[5.13], Dina Mandoli[14], Jody Banks[15], John E Carlson[16], Jeffrey Tomkins[17], Claude W dePamphilis[5], Rod A Wing[1] and Jim Leebens-Mack[2*]

[1]Arizona Genomics Institute, School of Plant Sciences and BIO5 Institute for Collaborative Research, University of Arizona, 1657 E. Helen Street, Tucson, AZ 85721, USA

[2]Department of Plant Biology, University of Georgia, 4504 Miller Plant Sciences, Athens, GA 30602, USA

[3]Department of Biological Sciences, University of Missouri, 371B Life Sciences Center Columbia, MO 65211, USA

[4]College of Life Sciences and Technology, Huazhong Agricultural University, Wuhan, Hubei 430070, China

[5]Intercollege Graduate Degree Program in Plant Biology and Institute of Molecular Evolutionary Genetics, Huck Institutes of the Life Sciences, The Pennsylvania State University, 405 Life Sciences Building, University Park, Pennsylvania 16802, USA

[6]Department of Plant and Microbiology, College of Natural Resources, University of California, 311 Koshland Hall, Berkeley, CA, USA

[7]School of Plant Sciences and BIO5, University of Arizona, 1657 East Helen Street, Tucson, AZ 85721, USA

[8]Dow Agrosciences LLC, 9330 Zionsville Road, Indianapolis, IN 46268, USA

[9]Plant Genome Mapping Laboratory, University of Georgia, 111 Riverbend Road, Athens, GA, 30605, USA

[10]Department of Biology, University of Florida, 220 Bartram Hall, Gainesville, FL 32611, USA

[11]Florida Museum of Natural History, Museum Road & Newell Drive, University of Florida, Gainesville, FL 32611, USA

[12]Department of Biological Sciences, University at Buffalo (SUNY), 637 Hochstetter Hall, Buffalo, NY 14260, USA

[13]State Key Laboratory of Genetic Engineering, School of Life Sciences, Institute of Plant Biology, Center for Evolutionary Biology, and Institutes of Biomedical Sciences, Fudan University, 220 Handan Road, Shanghai 200433, China

[14]Northern Lights, 4500 NE 40th Street, Seattle WA 98105, U.S.A.

[15]Department of Botany and Plant Pathology, Purdue University, B028 Whistler Hall, West Lafayette, IN 47906, USA

[16]School of Forest Resources, The Pennsylvania State University, 323 Forest Resources Building, University Park, PA 16802, USA

[17]Clemson University Genomics Institute, Clemson University, 51 Cherry St, Clemson, NC 29634, USA

*Corresponding Author: Jim Leebens-Mack jleebensmack@plantbio.uga.edu

Abstract

Background

Recent phylogenetic analyses have identified *Amborella trichopoda*, an understory tree species endemic to the forests of New Caledonia, as sister to a clade including all other known flowering plant species. The *Amborella* genome is a unique reference for understanding the evolution of angiosperm genomes because it can serve as an outgroup to root comparative analyses. A physical map, BAC end sequences and sample shotgun sequences provide a first view of the 870 Mbp *Amborella* genome.

Results

Analysis of *Amborella* BAC ends sequenced from each contig suggests that the density of LTR retrotransposons is negatively correlated with that of protein coding genes. Syntenic, presumably ancestral, gene blocks were identified in comparisons of the *Amborella* BAC contigs and the sequenced *Arabidopsis thaliana, Populus trichocarpa, Vitis vinifera* and *Oryza sativa* genomes. Parsimony mapping of the loss of synteny corroborates previous analyses suggesting that the rate of structural change has been more rapid on lineages leading to *Arabidopsis* and *Oryza* compared with lineages leading to *Populus* and *Vitis*. The gamma paleohexiploidy event identified in the *Arabidopsis, Populus* and *Vitis* genomes is shown to have occurred after the divergence of all other known angiosperms from the lineage leading to *Amborella*.

Conclusions

When placed in the context of a physical map, BAC end sequences representing just 5.4% of the *Amborella* genome have facilitated reconstruction of gene blocks that existed in the last common ancestor of all flowering plants. The *Amborella* genome is an invaluable reference for inferences concerning the ancestral angiosperm and subsequent genome evolution.

Background

The origin and rapid diversification of the angiosperms (flowering plants) were pivotal events in the evolutionary history of Earth's biota. Over the last 130-150 million years angiosperms have diversified to include ~350,000 species occupying nearly all habitable terrestrial and many aquatic environments. Angiosperms generate the vast majority of human food either directly or indirectly as animal feed, and they account for a huge proportion of land-based photosynthesis and carbon sequestration. Comparative analyses of genome sequences and gene function for a growing number of species are shedding light on how gene and genome duplications have contributed to the diversification within major flowering plant lineages (e.g. *Rosidae*, *Asteridae*, *Monocotyledoneae* [1]), but elucidation of the genetic and genomic processes underlying the key innovations associated with the origin of flowering plants (e.g., typically bisexual flowers, endosperm formation, double fertilization, ovules with two integuments, seed development within the carpel) requires comparisons between lineages that diverged from the last common ancestor of all extant angiosperms [2, 3].

Recent phylogenetic analyses have identified *Amborella trichopoda*, an understory tree or shrub species endemic to the forests of New Caledonia, as the sister species to all other extant angiosperms [4-8]. *Amborella* is no more "ancient" or "primitive" than any other extant flowering plant species, but comparisons between *Amborella* and other angiosperms are allowing researchers to triangulate on characteristics of their last common ancestor. Using a similar approach, researchers have used the complete genome sequence of platypus, *Ornithorhynchus anatinus*, representing the sister group of all other extant mammals, to elucidate mammalian genome evolution [9].

Previous comparisons of transcriptome content [10], gene expression patterns [11-13], and gene function [14, 15] between *Amborella* and other flowering plant species have suggested that much of the floral development program that has been characterized in *Arabidopsis*, snapdragon and maize existed in the last common ancestor of extant angiosperms. While gene duplications in the MADS-box transcription factor family likely contributed to the earliest floral development regulatory networks [11, 12, 16-19],

it is not clear whether these were single gene duplications or the product of polyploidization. Genome duplications have occurred repeatedly throughout angiosperm history [20-23] but there is uncertainty in the timing of polyploidy events relative to the origin of the angiosperms and important innovations in flowering plant history [24].

Here we describe a BAC-based draft physical map for *Amborella trichopoda* and use BAC end sequences (BES) to compare the structure of the *Amborella* genome to representative eudicot (*Vitis*, *Populus* and *Arabidopsis*) and grass (*Oryza*) genomes. Comparative analyses of sequences for two large contiguous regions (487.3 and 629.7 kb in the *Amborella* genome) were also performed. In addition we use a large transcriptome assembly to identify BAC ends matching protein-coding sequences [25]. Our aim here is to begin to investigate whether regions of these genomes have remained syntenic throughout angiosperm history, and determine whether ancient genome duplications discovered in eudicot and grass genomes [26-29] occurred before or after the divergence of these lineages from the *Amborella* lineage. In addition, the physical map and sequence analyses establish a framework for future studies of all flowering plant genomes including the *Amborella* genome itself.

Results and Discussion

BAC library and physical map

The structure and composition of the 870 Mbp/C [30] *Amborella trichopoda* genome was investigated through physical mapping of clones from a 5.2X coverage BAC library. The library was constructed after partial digest of high-molecular-weight DNA with *Hind* III. The library, which comprises 36,684 BAC clones with an estimated average insert size of 123 Kb, is available through the Arizona Genomics Institute [31]. The BAC library was double spotted in high density onto Hybond N+ filters. All 36,684 clones were end-sequenced, and a physical map was constructed after high information content fingerprinting (HICF; [32, 33]). A total of 32,719 fingerprinted BACs was assembled into 3106 contigs and 1356 singletons using the program FPC version 7.2 [34].

The quality of the physical map was assessed by screening the arrayed library with probes developed for *Amborella* homologs for eight genes that have been found to be single-copy in sequenced plant genomes [35, 36]. Probes derived from *Amborella* cDNA clones or PCR amplicons were putative homologs of following single-copy *Arabidopsis* genes: *ASD* (At1g14810), *DWARF1* (At3g19820), *GIGANTEA* (At1g22770), *LEAFY* (At5g61850), a dienelactone hydrolase gene (At2g32520), a cytochrome-C-oxidase related gene (At4g37830), *EIF3K* (At4g33250) and a hypothetical protein-coding gene with strong similarity to rice gene Os02g0593400 (At5g63135). All verified positive clones mapped to the same FPC contig for 6 of the 8 probes (Figure S1 Additional file 1). Positive clones for the *EIF3K* and the hypothetical protein-coding gene probes were each distributed between two FPC contigs and inspection of the HICF bands for these contigs suggests that the genes have been duplicated in the *Amborella* lineage. In accord with the expected library coverage, the single copy nuclear gene probes hybridized to 3-13 clones (mean 6.9).

The correlation between HICF bands and the number of BACs included in each FPC contig was 0.655 for all contigs and 0.917 after removing two contigs derived from the chloroplast and mitochondrial genomes and one contig composed largely of repetitive elements (Figure S2 in Additional file 1). We used a calibration of average insert size (123 kb) over the average number of HICF bands per BAC clone (128) to obtain a rough estimate of FPC contig lengths. Of 77 FPC contigs with 39 or more BACs (not including the contigs with the plastome and repetitive elements), estimated lengths ranged from 308 to 1,429 kb.

BAC end sequencing (BES) was performed on all fingerprinted BACs producing 69,466 Sanger reads with an average length of 695 bp after quality and vector trimming. This corresponds to 48.25 Mbp, or roughly 5.4% of the *Amborella* genome. BAC end sequences were related to the physical map and used to identify regions of synteny between regions of the *Amborella* genome and the sequenced *Arabidopsis*, *Populus, Vitis* (grape), and *Oryza* (rice) genomes (see below). In addition, end sequences were used to verify the identity of the three excluded FPC contigs described above. All BES mapping at least 100 bp apart on the plastid genome [37] were found in the same FPC contig. This contig included just 532 BACs, indicating very low (1.6%) plastid DNA contamination.

Characterization of repeats in BAC end and shotgun sequences

Repeat composition and frequency in the *Amborella* genome were characterized through analysis of the BAC end and whole genome survey sequences. Reads were first compared with sequences in Repbase (v. 15.08; [38]) using BLASTN [39]. In order to minimize the effect of divergence between *Amborella* genes and homologous repeats from other species, we used relaxed BLASTN settings (-q -4 –r 5) to accommodate an estimated 160 million years of sequence divergence since the last common ancestor of extant flowering plants [8, 40-42] while maintaining rigorous support for significant hits (E-value threshold was set at 1e-10). All BAC end sequences without significant hits were then compared with the non-redundant protein database in GenBank using BLASTX and an E-value threshold of e-5. Finally, the remaining sequences without matches in Repbase or the GenBank nr database were compared with sequences that did have matches in either database using BLASTN with an E-value threshold of 1.0e-10. We report results both

excluding these "internal" BLAST searches and including them (I). Together these results provide estimates of TE content based on conservative and more comprehensive (and possibly more permissive; I) search strategies.

With the more comprehensive strategy (I), slightly more than half of all the *Amborella* BESs matched known transposable element sequences (TEs). Not surprisingly, the most highly represented TE class was LTR retrotransposons, accounting for 7.65% (I: 30.01%) of all BESs and 57.5% (I: 56.58%) of all those with hits to Repbase. Hits to Ty1-*copia* type sequences were slightly more common (3.11%; I:13.79%) than matches to Ty3-*gypsy*-like LTRs (3.50%; I:12.09%); the remaining LTR retrotransposon matches (1.04%; I: 4.13%) were not classified. LINEs also represented a significant fraction of *Amborella* BAC ends: 2.70% (I:11.60%) of the total, 19.98% of all the repeats (I: 22.22%). This is noteworthy because LINEs are usually significantly less numerous than LTR retrotransposons in plant genomes [43-47] with some notable exceptions such as the element *del2* in *Lilium speciosum* [48]. The complete set of DNA-TE-related BESs account for just 1.63% (I: 4.51%) of the total, and the most represented classes are those of hAT and MuDR elements: 0.92% (I:2.41%) and 0.49% (I:1.04%) of the total BESs, respectively. Results from the same analyses replicated on the set of 2695 random sheared Sanger sequences **(Table 1)** and 648,519 454 reads (Table S1; Additional file 1) are generally in very good agreement with those obtained using BES data.

A *de novo* search for novel MITE elements overlooked by the similarity search approach was carried out using the pipeline MUST [49]. The most abundant candidates identified by the pipeline were manually inspected to confirm features typical of MITE

elements such as small size, terminal inverted repeats (TIR), high A+T nucleotide content and target site duplications (TSD). Three putative high-copy MITEs were identified. All of these were small elements (174-500 bp) with TIRs, TSDs, and A+T contents greater than 65% (Figure S3; Additional file 1). Repeat copy numbers estimated from the BESs and random sheared sequences were extrapolated to obtain genome-wide estimates using the procedure developed by Hawkins et al. [50]. Copy number ranges from 3300 copies for MITE_2 to 17000 copies for MITE_1. The estimates inferred from BESs were generally consistent with those calculated for random sheared reads (with the possible exception of MITE_3) **(Table 2).**

The conserved Reverse Transcriptase domains of LTR retrotransposons and LINEs were collected and used to estimate maximum likelihood trees (**Figure 1).** In the case of LTR retroelements, the trees indicate substitution rate heterogeneity (i.e. variation root-to-tip distances) and no evidence for recent retrotranspositional bursts of single families (i.e. short terminal branches). In the case of LINEs, the phylogenetic tree displays very long branches suggestive of an ancient diversification or very rapid substitution rates. As has been described for other plants [51], *Amborella* LINEs exhibit high sequence divergence and extreme heterogeneity.

The *Amborella* BESs were also searched for microsatellites (i.e. simple sequence repeats, SSRs); for comparison, the search was also conducted on the *Amborella* random sheared reads and on BES (from other *Hind*III BAC libraries) from *Glycine* (soybean) and *Oryza rufipogon*. In comparison to the other two species, *Amborella* shows a higher frequency of SSRs, particularly mono- and dinucleotide repeats, with a particularly high frequency of “AG” dinucleotide microsatellites. The results of SSR analysis in BESs

were confirmed by those obtained from the randomly sheared *Amborella* sequences (**Table 3**).

Repeat profiles in the shotgun sequences were also assessed using Tallymer to characterize K-mer frequencies [52]. The *Amborella* K-mer frequency profiles were compared with those of *Arabidopsis thaliana*, *Oryza sativa* (rice), *Sorghum bicolor* and *Zea mays* (maize). While the *Amborella* genome size is closest to Sorghum's (870 and 740 Mbp/C, respectively), its K-mer frequency profiles were more similar to those of *Arabidopsis* and rice, with much smaller genome sizes (157 and 490 Mbp/1C, respectively; [53])(**Figure 2**).

Distribution of BES with matches to protein-coding regions of reference genomes

All BAC end and shotgun sequences were compared to the GenBank nr database using BLASTX [39] with an e-value threshold of 1e-5. After the removal of sequences similar to TEs, the overall frequencies of sequences finding matches in the protein database were 11.9% and 8.05% for the BES and Sanger shotgun sequences, respectively. For BESs from FPC contigs with 10 or more BACs, we found a negative correlation between the frequencies of BESs matching protein-coding genes and LTR retrotransposons ($r = -0.423$, $p<0.0001$). As has been described for other genomes [54-56], gene density seems to be negatively correlated with retrotranposon density in the *Amborella* genome.

Identification of syntenic blocks between *Amborella, Arabidopsis*, rice, poplar and grape

Taking advantage of the availability of a Phase I physical map assembly, we mapped the *Amborella* contigs onto the genomes of *Arabidopsis thaliana*, *Populus trichocarpa, Vitis vinifera,* and *Oryza sativa.* We focused on the 77 largest contigs with at least 39 clones.

BLAST analyses of BES were done within the context of their linkages within FPC contigs. All of the contig BESs classified as repeats (see above) were discarded. Those remaining were compared against the four reference genomes. Because of the large evolutionary time that separates *Amborella* from the other four sequenced genomes [41, 42, 57], the comparisons were carried out at the protein level using tBLASTX; only the best hits were taken into account. *Amborella* FPC contigs were considered for further analyses if at least two BESs had matches with bit scores greater than 80 (typically a maximum e-value of 1.0E-20 over 100 amino acidic residues) to loci separated by less than 500 kb within one of the four genomes being compared. Positive matches were used as anchors to circumscribe 4 Mbp tracts within the reference genomes and a second, more focused tBLASTX search was performed comparing the BESs with these regions. An e-value threshold of 1.0E-4 was used for the second set of tBLASTX searches and all significant hits were used to identify syntenic regions. We considered a contig as anchored if the contig had at least four positive hits (e-value lower than 1.0e-4) to at least 3 distinct genes.

Non-repetitive BESs were also compared to a database of 246,196 *Amborella* cDNA unigenes assemblies with lengths greater than 100 bp. These cDNAs were derived from comprehensive sequencing of nine cDNA libraries (**Table 4**; [25]). Sixty- six percent of the non-repetitive BESs matched cDNA sequences in BLASTN searches with an e-value cutoff of 1.0e-10.

Using the search strategy described above, 29 large *Amborella* BAC contigs (>39 BAC clones) showed synteny with at least one of the four sequenced genomes, and nine of these showed synteny with at least one region in all four genomes. All BES mapping

to these syntenic regions also exhibited significant matches to the sequences in the Amborella cDNA assembly (**Table 4**, **Table S2** Additional file 1). Whereas 25 of these *Amborella* BAC contigs mapped to at least one tract in the *Vitis* genome, 15, 16, and 24 contigs were found to be syntenic with one or more tracts in the *Oryza*, *Arabidopsis*, and *Populus* genomes, respectively **(Table S2** Additional file 1**).** These results provide a novel, albeit coarse, first view of the ancestral genome for all flowering plants and the timing of rearrangements and other structural changes (e.g. genome duplications, fractionation, chromosomal fissions and fusions) that have reduced synteny between the monocot and eudicot genomes analyzed here (**Figure 3**). Parsimony mapping of synteny loss onto a phylogeny consisting of *Amborella* and the other four species indicates variation in rates of change in genome structure. In agreement with previous studies [29, 45], *Vitis* seems to have been the most stable of the sequenced genomes, and the rate of change slowed in the lineage leading to *Populus* following divergence from the lineage leading to *Arabidopsis* (**Figure 3**).

Paleopolyploidy in Angiosperm genomes

Paleopolyploidy events have been well characterized in all four sequenced genomes analyzed here [29, 45, 58-60], and the syntenic *Amborella* FPC contigs described above often match multiple regions in these genomes. The most ancient of these paleopolyploidy events is the so-called γ triplication that has been inferred to have occurred before the divergence of the *Asteridae* (represented by tomato, *Solanum lycopersicon*) and the *Rosidae*, including *Vitis*, *Populus* and *Arabidopsis* [29]. Given the very incomplete view of the *Amborella* genome that is available in the BES data, we are not able to assess synteny between *Amborella* FPC contigs. Nevertheless, comparisons

between the *Amborella* contigs and sets of syntenic blocks in the *Vitis* genome indicate that the γ triplication most likely occurred sometime after the divergence of all other angiosperms from the lineage leading to *Amborella*.

All BES were compared to all annotated protein-coding genes in the *Vitis* genome placed within the context of the pre-triplication ancestral gene blocks and post-triplication syntenic segments identified by Tang et al. [29]. A total of 328 *Amborella* FPC contigs had between two and eight genes with significant best BLASTX matches (e-values $\leq$ 1.0E-6) to *Vitis* genes corresponding to pre-triplication gene blocks in the ancestral genome. In most of these cases (199/328; Additional file 2), best hits were distributed between two or three homeologous (i.e. post-triplication) syntenic *Vitis* genome segments. Of the remaining 129 *Amborella* FPC contigs with BES showing significant BLASTX hits to a single *Vitis* subgenome (i.e. single copy of a triplicated ancestral block), most (113) included just 2 genes mapping to the ancestral *Vitis* gene blocks (14 including 3 genes, and 2 including 4 genes) (Additional file 2). All 21 FPC contigs with best BLASTX matches to 5 or more genes within the ancestral *Vitis* blocks were distributed among two or three post-triplication subgenomes. Complete sequences for the *Amborella* BAC contigs may reveal more even distribution of segments among *Vitis* subgenomes, but the results described here suggest that triplication, fractionation and divergence of homeologous segments in the *Vitis* genome postdate the divergence between lineages leading to *Vitis* and *Amborella* (i.e. the last common ancestor of all extant angiosperms).

Analysis of complete sequences for two *Amborella* BAC contigs

Two of the larger (~ 500 kb) BAC contigs (IDs 431 and 1003) mapping to multiple segments in all four sequenced reference genomes were identified for further investigation. A minimum tiling path was constructed for each contig, and florescence *in situ* hybridizations were performed to verify that the BACs mapped to a single contiguous region in the *Amborella* genome (**Figure 4**). Each BAC in the tiling paths was subcloned and sequenced to 8X coverage on an ABI 3730xl sequencer. Gaps were closed for each scaffold, and contiguous 487,318 and 629,678 bp phase II sequences were assembled for contigs 431 and 1003, respectively.

The DAWGPAWS suite of scripts was used to organize *ab initio* gene predictions, BLAST results and the output of repeat identification tools [61, 62]. *Ab initio* gene predictions were generated using FGENESH [63], AUGUSTUS [64], SNAP [65], GeneID [66] and GenScan [67]. In addition, *Amborella* EST sequences produced by the 454 Titanium platform (2,943,273 reads; total read size= ~776 Mbp; average read length = 263.60 bp) and Sanger sequencing (38147 reads; total read size= ~21.3 Mbp; average read length = 559.57 bp) were splice-aligned to the contigs using GMAP (Genomic Mapping and Alignment Program) [68] with the PASA (Program to Assemble Spliced Alignments) genome annotation tool [69]. All predictions were manually compared with BLASTX results against gene annotations from *Arabidopsis* [70], *Vitis* [45], *Zea mays* [56], *Medicago* [71], *Oryza* [72, 73], and *Sorghum* [55] as well as tBLASTx results against the *Amborella* transcript assemblies. GBrowse views of gene annotations and BLAST results for each contig are available at the Ancestral Angiosperm Genome Project website [25].

Rigorous assessments of synteny between these *Amborella* contigs and the aforementioned four angiosperm genomes were performed using LASTZ [74, 75]. Dotplots comparing the *Amborella* contigs and the *Vitis* genome show that contigs are syntenic with previously triplicated blocks [29]. Regions of contig 1003 match genes on syntenic segments of chromosomes 1, 14 and 17 in the *Vitis* genome (**Figure 5)** and contig 431 mapped to syntenic portions of *Vitis* chromosomes 6, 8 and 13 (**Figure 6**). These findings support the conclusion from the BES analyses suggesting that the γ triplication occurred after the first branching event in the phylogeny of extant angiosperms.

At least two genome duplications (ρ and σ) have been inferred to have occurred within the monocot lineage leading to rice since divergence of monocots and eudicots [28]. These duplications were evident in comparisons with both *Amborella* contigs. Regions of contig 1003 were found to be syntenic with portions of rice chromosomes 2 and 4 derived from the ρ duplication and a portion of chromosome 10 (**Figure 5**) which is related to these two regions through the earlier σ duplication [28]. The LASTZ analysis of contig 431 revealed synteny with seven regions in the rice genome (**Figure 6**) and one of the "putative ancestral regions" (PAR 17) characterized by Tang et al. [28]. These PARs were defined as regions of synteny between the rice and *Vitis* genomes. Phylogenetic analyses of genes in *Amborella* contig 431 and syntenic regions of the rice and *Vitis* genomes may elucidate the timing of the γ triplication and genome duplications evident in synteny analyses of the rice genome relative to the divergence of monocots and eudicots.

Phylogenetic analyses of gene families represented in sequenced *Amborella* contigs

While the fractionation process has resulted in the loss of most duplicated genes following the ancient polyploidy events evident in the syntenic *Vitis* and rice segments shown in **Figures 5** and **6**, duplicate *Vitis* genes have been retained for homologs of three *Amborella* genes located on contig 431 (**Figures 6a)**. These genes were used to search the PlantTribes gene family database [35]. The three gene sets identified in the synteny analysis correspond to three gene families (auxin-independent growth promoter, ceramidase and plant uncoupling mitochondrial protein) circumscribed through OrthoMCL clustering [76] of gene annotations from the available *Arabidopsis*, *Carica* (papaya), *Populus*, *Medicago* (alfalfa), *Glycine*, *Cucumis* (cucumber), *Vitis*, *Mimulus*, *Oryza*, *Sorghum*, *Selaginella* (spike moss) and *Physcomitrella* genomes. Homologous genes sampled from exemplar asterid, ranunculid, non-grass monocot and gymnosperm species were obtained from EST assembly databases [25, 77, 78] and were added to each gene family set. Sequences in each gene family set were aligned using MUSCLE [79], and RAxML [80] run with the GTRGAMMA substitution model was used to obtain maximum likelihood estimates of gene trees.

Inspection of the resulting gene trees shows support for the inference drawn from the BAC end sequence analysis. The γ triplication (hexaploidy event) clearly occurred after *Amborella* diverged from other extant angiosperm lineages (**Figure 7)**. The placement of the γ triplication with respect to the divergence of monocots and eudocots or core eudicots and the Ranunculales varies among the three gene trees. This incongruence among gene trees is likely due to artifacts associated with substitution rate variation and insufficient taxon sampling. Analyses of additional gene families with broader taxon sampling will be necessary to obtain better resolution for the timing of the

γ triplication with respect to the divergence of monocot, eudicots, Ranunculales (i.e. "basal" eudicots) and core eudocots.

Conclusions

Amborella trichocarpa is the sister species to the large clade encompassing all other extant flowering plants. As such, comparative analyses of *Amborella* and other flowering plants offer a uniquely informative perspective on the most recent common ancestor of all extant angiosperms. The physical map and BAC end sequences described in this study provide a low-resolution view of the *Amborella* genome. Nonetheless, these data shed light on genomic features of the last common ancestor of flowering plants. Moreover, the *Amborella* genome provides a unique reference for understanding genome evolution throughout angiosperm history. When placed in the context of the physical map, BAC end sequences representing just 5.4% of the *Amborella* genome allowed reconstruction of ancestral gene blocks in regions represented by 29 BAC contigs and inference of the timing of structural mutations that disrupted these blocks (**Figure 3**).

Analyses of BAC end sequences and BAC contigs also indicate that the ancient γ polyploidy event inferred from the *Arabidopsis* [58], *Carica* [81], *Populus* [60], and *Vitis* [45] genomes occurred after the *Amborella* lineage diverged from the rest of the angiosperms. Therefore, if the origin of angiosperms was associated with a genome duplication as has been hypothesized elsewhere [16, 20, 23], that polyploidy event predated the γ event.

Materials and methods

BAC library construction: Protocols for DNA megabase preparation, library construction, picking and arraying proposed in Luo and Wing [82] were followed.

Fingerprinting: SNaPshot fingerprinting technique was adopted [32] with the modifications described by Kim et al. [83]. Snapshot reactions were loaded into ABI 3730xl DNA sequencers. Analysis of data for each contig was carried out using the ABI Data Collection Program.

Physical map construction: Fingerprints were assembled into contigs using the program FPC version 7.2 [34]. The initial assembly was carried out using a Sulston score threshold of e-50 followed by three rounds of dequeuing at the same stringency and auto-merging of contigs at e-21.

BAC end extraction and sequencing: BAC DNA was extracted and end sequenced from 36684 clones using the methods described by Ammiraju et al. [83, 84]. Sequence quality assessment and trimming were carried out using the programs Phred [85] and Lucy [86].

Random Sheared Library: Random sheared library was constructed as previously described [87].

cDNA Sequencing and Assembly: Additional Sanger ESTs were generated from available male and female flower bud cDNA libraries [10] (**Table 4**). Libraries for 454 sequencing were constructed from the tissues listed in Table 4 using the Mint cDNA synthesis kit (Evrogen). Total RNAs fore cDNA synthesis were isolated using a combination of CTAB extraction and the RNeasy Plant Mini kit (Qiagen) as previously described for basal angiosperms [11]. Two rounds of messenger RNA isolation were performed with the Poly(A)Purist™ mRNA Purification Kit (Ambion Inc.) according to

the manufacturer's recommendation. Contaminant DNA was removed with DNA-free™ (Ambion Inc.) and mRNA quality was verified using a Bioanalyzer (Agilent Inc.). Vector and adaptor sequences were trimmed from 454 Titanium (2,943,273 reads; total read size= ~776 Mbp; average read length = 263.60 bp) and Sanger sequences (38147 reads; total read size= ~21.3 Mbp; average read length = 559.57 bp) using seqclean [88] and assembled using MIRA [89].

Similarity searches, repeat classification and contig anchoring: Similarity searches were carried out using the programs BLASTN and BLASTX [39]. BLASTN was run under relaxed settings (-q -4 –r 5) in order to accommodate the evolutionary distance between *Amborella* and the species included in the repeats databases used; the significance threshold was set at 1e-10. In the case of BLASTX searches the threshold was set at 1e-5 or 1e-4 for the BES synteny analysis. tBLASTX was used to anchor the contigs to the reference genomes (see results for details).

Databases: The databases used in similarity searches were RepBase version 15.08 [38], the GenBank non-redundant (nr) database, and the *Oryza*, *Arabidopsis*, *Vitis* and *Populus* genome sequences.

Validation of repeats search and MITEs identification: The program MUST [49] was used for *de novo* characterization of highly repeated sequences; results were then inspected for the presence of MITEs features. Inverted repeats were identified manually parsing the results of dot-plot comparisons made using the program “Dotter” [90].

SSR searches: Microsatellites were identified using the program Sputnik [91]. SSR composition, length and distribution were parsed and analyzed using the tools and the strategy used by Morgante et al. [92].

Fluorescence *in situ* hybridization (FISH): FPC contigs were validated by hybridizing BAC DNAs to *Amborella* chromosome squashes. DNAs were prepared for BACs mapping to the middle and both ends of BAC contigs 431 and 1003 and used to prepare fluorescently labelled BAC-FISH probes. Chromosome squashes were prepared from root tips and labeled BAC-FISH probes were prepared as described by Xiong et al. [93].

Contig sequencing and annotation: Minimum tiling paths (MTPs) of 7 and 6 BACs were identified for contigs 1003 and 431, respectively, by the visual inspection of the FPC assemblies. Adjacent clones were chosen based on their reciprocal position and probability value associated to their overlapping fingerprinted bands as shown by FPC. Sequencing of selected MTP BACs was done to phase II quality as previously described [73]. Phase II BAC sequences were then assembled into 1003 and 431contig sequences based on dot plot comparisons and overlap similarity between adjacent clones.

Perl scripts available from the DAWGPAWS package [61, 62] were used to convert computational annotation results from multiple sources into a single GFF3 file for combined evidence annotation in Apollo [94] and publication in Gbrowse [95]. *Ab initio* gene annotation programs used in this process included FGENESH [63] AUGUSTUS [64], SNAP [65], GeneID [66] and GenScan [67]. Because *Amborella*-specific gene model parameterizations were not available for these programs, multiple plant models were used for each *ab initio* program. The sequence of the entire contig was BLASTx ($e < 1x10^{-5}$) searched against gene annotations from *Arabidopsis* [70], *Vitis* [45], *Zea mays* [56], *Medicago* [71], *Oryza* [72], and *Sorghum* [55] as well as tBLASTx ($e < 1x10^{-5}$) searched against a database of comprehensive *Amborella* transcript assemblies [25]. In addition, *Amborella* EST sequences (reads and assemblies; **Table 4**)

were splice-aligned to the contigs using GMAP (Genomic Mapping and Alignment Program) [68] with the PASA (Program to Assemble Spliced Alignments) genome annotation tool [69]. The gene models and BLAST search results were manually combined into gene models using the Apollo genome annotation curation tool [94].

Synteny analysis of sequenced BAC contigs with *Vitis* and *Oryza* genomes: Sequenced *Amborella* BAC contigs 431 (487,318 bp) and 1003 (629,678 bp) were compared to the IRGSP rice genome assembly (version 5) and the Genoscope 12X *Vitis* genome assembly using LASTZ and default parameters. Prior to LASTZ comparisons, all genomic sequences were masked using NCBI's WindowMasker to remove simple repeats. Significant matches after repeat masking were visualized as dot plots. Gene annotations for the rice and *Vitis* genomes were obtained from the Rice Annotation Project [96] and Genoscope [97] respectively, and plotted on the vertical axes of the dot plots (**Figures 5,6**). FGENESH [63] annotations for the *Amborella* contigs were included on the horizontal axes of the dot plots. LASTZ scores were summed for all aligned *Amborella*-rice or *Amborella*-*Vitis* blocks within 100 Kb of each other in sequenced genomes. All regions with summed scores > 100000 were considered as syntenic and included in **Figures 5** and **6**.

Phylogenetic analysis: All alignments were carried out using the program "MUSCLE" [79] run under default settings. Maximum likelihood analyses were run on aligned DNA and amino acid sequences using RAxML [80] and the GTRGAMMA nucleotide substitution model.

Submission of data to GenBank databases: BAC end sequences (HR616970 - HR686434), full-length BAC sequences (AC243594.1 - AC243606.1), Sanger shotgun

sequences (HR614237 - HR616931), 454 shotgun sequences (SRP006044), Sanger EST (FD425831.1 - FD443502.1) and 454 cDNA sequences (SRX018174, SRX018165, SRX018164, SRX018163, SRX018157, SRX018156) were deposited in the appropriate NCBI GenBank sequence databases. All sequences are also available at the Ancestral Angiosperm Genome Project website [25].

Authors' contributions

JLM, AZ., RAW and CWD. designed and coordinated the study. The *Amborella* BAC library was constructed and characterized in the Arizona Genomics Institute (**AGI**) by DK, YY, KC , JLG. AS and ML. cDNA library production and sequencing was performed by AC at the University of Florida and assemblies were performed by SA. Funding for BAC library construction was obtained by DM, JB, JEC, JT, CWD and RAW. Comparative analyses were preformed by AZ, JEB, JCE, JD, HT, SR, AHP, DES, PSS, VAA, HM, CWD and JL-M. Florescence in situ hybridizations were performed by ZX and JCP BAC contig annotations were performed by JEB, JCE, SC, BB and JLM. AZ and JLM wrote the first draft of the manuscript and all authors contributed to refinement.

Acknowledgements

This work was supported with funding from National Science Foundation grants 0208502, 0638595 and 0922742. We also acknowledge helpful comments and suggestions provided by anonymous reviewers.

References

1. Cantino P, J. Doyle, S. Graham, W. Judd, R. Olmstead, D. Soltis, P. Soltis, and M. Donoghue: **Towards a Phylogenetic Nomenclature of Tracheophyta.** *Taxon* 2007, **56:**822-846.
2. Leebens-Mack JH, Wall PK, Duarte J, Zheng Z, Oppenheimer D, dePamphilis CW: **A genomics approach to the study of floral developmental genetics: strengths and limitations.** *Advances in Botanical Research* 2006, **44:**527-549.
3. Soltis DE, Albert VA, Leebens-Mack J, Palmer JD, Wing RA, dePamphilis CW, Ma H, Carlson JE, Altman N, Kim S, Wall PK, Zuccolo A, Soltis PS: **The Amborella genome: an evolutionary reference for plant biology.** *Genome Biol* 2008, **9:**402.
4. Mathews S, Donoghue MJ: **The root of angiosperm phylogeny inferred from duplicate phytochrome genes.** *Science* 1999, **286:**947-950.
5. Qiu YL, Lee J, Bernasconi-Quadroni F, Soltis DE, Soltis PS, Zanis M, Zimmer EA, Chen Z, Savolainen V, Chase MW: **The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes.** *Nature* 1999, **402:**404-407.
6. Soltis PS, Soltis DE, Chase MW: **Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology.** *Nature* 1999, **402:**402-404.
7. Jansen RK, Cai Z, Raubeson LA, Daniell H, Depamphilis CW, Leebens-Mack J, Muller KF, Guisinger-Bellian M, Haberle RC, Hansen AK, Chumley TW, Lee SB, Peery R, McNeal JR, Kuehl JV,Boore JL: **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* 2007, **104:**19369-19374.
8. Moore MJ, Bell CD, Soltis PS, Soltis DE: **Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms.** *Proc Natl Acad Sci U S A* 2007, **104:**19363-19368.
9. Warren WC, Hillier LW, Marshall Graves JA, Birney E, Ponting CP, Grutzner F, Belov K, Miller W, Clarke L, Chinwalla AT, Yang SP, Heger A, Locke DP, Miethke P, Waters PD, Veyrunes F, Fulton L, Fulton B, Graves T, Wallis J, Puente X S, Lopez-Otin C, Ordonez G R, Eichler E E, Chen L, Cheng Z, Deakin J E, Alsop A, Thompson K, Kirby P, et al.: **Genome analysis of the platypus reveals unique signatures of evolution.** *Nature* 2008, **453:**175-183.
10. Albert VA, Soltis DE, Carlson JE, Farmerie WG, Wall PK, Ilut DC, Solow TM, Mueller LA, Landherr LL, Hu Y, Buzgo M, Kim S, Yoo MJ, Frohlich MW, Perl-Treves R, Schlarbaum SE, Bliss BJ, Zhang X, Tanksley SD, Oppenheimer DG, Soltis PS, Ma H, dePamphilis CW, Leebens-Mack JH: **Floral gene resources from basal angiosperms for comparative genomics research.** *BMC Plant Biol* 2005, **5:**5.
11. Kim S, Koh J, Yoo MJ, Kong H, Hu Y, Ma H, Soltis PS, Soltis DE: **Expression of floral MADS-box genes in basal angiosperms: implications for the evolution of floral regulators.** *Plant J* 2005, **43:**724-744.
12. Soltis DE, Chanderbali AS, Kim S, Buzgo M, Soltis PS: **The ABC model and its applicability to basal angiosperms.** *Ann Bot* 2007, **100:**155-163.

13. Vialette-Guiraud AC, Adam H, Finet C, Jasinski S, Jouannic S, Scutt CP: **Insights from ANA-grade angiosperms into the early evolution of CUP-SHAPED COTYLEDON genes.** *Ann Bot* 2011.
14. Fourquin C, Vinauger-Douard M, Chambrier P, Berne-Dedieu A, Scutt CP: **Functional conservation between CRABS CLAW orthologues from widely diverged angiosperms.** *Ann Bot* 2007, **100:**651-657.
15. Fourquin C, Vinauger-Douard M, Fogliani B, Dumas C, Scutt CP: **Evidence that CRABS CLAW and TOUSLED have conserved their roles in carpel development since the ancestor of the extant angiosperms.** *Proc Natl Acad Sci U S A* 2005, **102:**4649-4654.
16. Zahn LM, Kong H, Leebens-Mack JH, Kim S, Soltis PS, Landherr LL, Soltis DE, Depamphilis CW, Ma H: **The evolution of the SEPALLATA subfamily of MADS-box genes: a preangiosperm origin with multiple duplications throughout angiosperm history.** *Genetics* 2005, **169:**2209-2223.
17. Zahn LM, Leebens-Mack J, DePamphilis CW, Ma H, Theissen G: **To B or Not to B a flower: the role of DEFICIENS and GLOBOSA orthologs in the evolution of the angiosperms.** *J Hered* 2005, **96:**225-240.
18. Zahn LM, Leebens-Mack JH, Arrington JM, Hu Y, Landherr LL, dePamphilis CW, Becker A, Theissen G, Ma H: **Conservation and divergence in the AGAMOUS subfamily of MADS-box genes: evidence of independent sub- and neofunctionalization events.** *Evol Dev* 2006, **8:**30-45.
19. Shan H, Zahn L, Guindon S, Wall PK, Kong H, Ma H, dePamphilis CW, Leebens-Mack J: **Evolution of plant MADS box transcription factors: evidence for shifts in selection associated with early angiosperm diversification and concerted gene duplications.** *Mol Biol Evol* 2009, **26:**2229-2244.
20. Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, Albert, VA, Ma H, dePamphilis CW: **Widespread genome duplications throughout the history of flowering plants.** *Genome Res* 2006, **16:**738-749.
21. Van de Peer Y, Fawcett JA, Proost S, Sterck L, Vandepoele K: **The flowering world: a tale of duplications.** *Trends Plant Sci* 2009, **14:**680-688.
22. Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH: **The frequency of polyploid speciation in vascular plants.** *Proc Natl Acad Sci U S A* 2009, **106:**13875-13879.
23. De Bodt S, Maere S, Van de Peer Y: **Genome duplication and the origin of angiosperms.** *Trends Ecol Evol* 2005, **20:**591-597.
24. Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, dePamphilis CW, Wall PK, Soltis PS: **Polyploidy and angiosperm diversification.** *Am J Bot* 2009, **96:**336-348.
25. Ancestral Angiosperm Genome Project **[http://ancangio.uga.edu/]**
26. Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang H, Wang X, Bowers J, Paterson A, Lisch D, Freeling M: **Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids.** *Plant Physiol* 2008, **148:**1772-1781.

27. Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH: **Synteny and collinearity in plant genomes.** *Science* 2008, **320:**486-488.
28. Tang H, Bowers JE, Wang X, Paterson AH: **Angiosperm genome comparisons reveal early polyploidy in the monocot lineage.** *Proc Natl Acad Sci U S A* 2010, **107:**472-477.
29. Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH: **Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps.** *Genome Res* 2008, **18:**1944-1954.
30. Leitch I, Hanson L: **DNA C-values in seven families fill phylogenetic gapsin the basal angiosperms.** *Botanical Journal of the Linnean Society* 2002, **140:**175-179.
31. Arizona Genome Institute **[http://www.genome.arizona.edu/orders/direct.html?library=AT_SBa].**
32. Luo MC, Thomas C, You FM, Hsiao J, Ouyang S, Buell CR, Malandro M, McGuire PE, Anderson OD, Dvorak J: **High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis.** *Genomics* 2003, **82:**378-389.
33. Nelson WM, Bharti AK, Butler E, Wei F, Fuks G, Kim H, Wing RA, Messing J, Soderlund C: **Whole-genome validation of high-information-content fingerprinting.** *Plant Physiol* 2005, **139:**27-38.
34. Soderlund C, Humphray S, Dunham A, French L: **Contigs built with fingerprints, markers, and FPC V4.7.** *Genome Res* 2000, **10:**1772-1787.
35. Wall PK, Leebens-Mack J, Muller KF, Field D, Altman NS, dePamphilis CW: **PlantTribes: a gene and gene family resource for comparative genomics in plants.** *Nucleic Acids Res* 2008, **36:**D970-976.
36. Duarte JM, Wall PK, Edger PP, Landherr LL, Ma H, Pires JC, Leebens-Mack J, dePamphilis CW: **Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryza and their phylogenetic utility across various taxonomic levels.** *BMC Evol Biol* 2010, **10:**61.
37. Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH: **Analysis of the Amborella trichopoda chloroplast genome sequence suggests that amborella is not a basal angiosperm.** *Mol Biol Evol* 2003, **20:**1499-1505.
38. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update, a database of eukaryotic repetitive elements.** *Cytogenet Genome Res* 2005, **110:**462-467.
39. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
40. Leebens-Mack J, Raubeson LA, Cui L, Kuehl JV, Fourcade MH, Chumley TW, Boore JL, Jansen RK, depamphilis CW: **Identifying the basal angiosperm node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone.** *Mol Biol Evol* 2005, **22:**1948-1963.
41. Magallon S: **Using fossils to break long branches in molecular dating: a comparison of relaxed clocks applied to the origin of angiosperms.** *Syst Biol* 2010, **59:**384-399.

42. Smith SA, Beaulieu JM, Donoghue MJ: **An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants.** *Proc Natl Acad Sci U S A* 2010, **107:**5897-5902.
43. Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, Deragon JM, Westerman RP, Sanmiguel PJ, Bennetzen JL: **Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome.** *PLoS Genet* 2009, **5:**e1000732.
44. IRGSP: **The map-based sequence of the rice genome.** *Nature* 2005, **436:**793-800.
45. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Micac E, Jublot D, Poulain J, Bruyere C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, et al: **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla.** *Nature* 2007, **449:**463-467.
46. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, D. Xu, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, et al: **Genome sequence of the palaeopolyploid soybean.** *Nature* 2010, **463:**178-183.
47. Vershinin AV, Druka A, Alkhimova AG, Kleinhofs A, Heslop-Harrison JS: **LINEs and gypsy-like retrotransposons in Hordeum species.** *Plant Mol Biol* 2002, **49:**1-14.
48. Leeton PR, Smyth DR: **An abundant LINE-like element amplified in the genome of Lilium speciosum.** *Mol Gen Genet* 1993, **237:**97-104.
49. Chen Y, Zhou F, Li G, Xu Y: **MUST: a system for identification of miniature inverted-repeat transposable elements and applications to Anabaena variabilis and Haloquadratum walsbyi.** *Gene* 2009, **436:**1-7.
50. Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF: **Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium.** *Genome Res* 2006, **16:**1252-1261.
51. Schmidt T: **LINEs, SINEs and repetitive DNA: non-LTR retrotransposons in plant genomes.** *Plant Mol Biol* 1999, **40:**903-910.
52. Kurtz S, Narechania A, Stein JC, Ware D: **A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes.** *BMC Genomics* 2008, **9:**517.
53. KEW C-Value Database **[http://data.kew.org/cvalues/]**
54. Bowers JE, Arias MA, Asher R, Avise JA, Ball RT, Brewer GA, Buss RW, Chen AH, Edwards TM, Estill JC, Exum H E, Goff V H, Herrick K L, Steele C L, Karunakaran S, Lafayette G K, Lemke C, Marler B S, Masters S L, McMillan J M, Nelson L K, Newsome G A, Nwakanma C C, Odeh R N, Phelps C A, Rarick E A, Rogers C J, Ryan S P, Slaughter K A, Soderlund C A, et al: **Comparative physical mapping links conservation of microsynteny to chromosome**

structure and recombination in grasses. *Proc Natl Acad Sci U S A* 2005, **102:**13206-13211.
55. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, et al: **The Sorghum bicolor genome and the diversification of grasses.** *Nature* 2009, **457:**551-556.
56. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, et al: **The B73 maize genome: complexity, diversity, and dynamics.** *Science* 2009, **326:**1112-1115.
57. Bell CD, D. E. Soltis, P. Soltis: **The age and diversification of angiosperms re-revisited.** *American Journal of Botany* 2010, **97:**1296-1303.
58. Bowers JE, Chapman BA, Rong J, Paterson AH: **Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events.** *Nature* 2003, **422:**433-438.
59. Paterson AH, Bowers JE, Chapman BA: **Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics.** *Proc Natl Acad Sci U S A* 2004, **101:**9903-9908.
60. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov Schein A, J, Sterck L, Aerts A, Bhalerao R R, Bhalerao R P, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen G L, Cooper D, Coutinho P M, Couturier J, Covert S, Cronk Q, Cunningham R,et al: **The genome of black cottonwood, Populus trichocarpa (Torr. & Gray).** *Science* 2006, **313:**1596-1604.
61. Estill JC, Bennetzen JL: **The DAWGPAWS pipeline for the annotation of genes and transposable elements in plant genomes.** *Plant Methods* 2009, **5:**8.
62. DAWGPAWS **[http://dawgpaws.sourceforge.net]**
63. FGENESH **[http://softberry.com]**
64. Stanke M, Schoffmann O, Morgenstern B, Waack S: **Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources.** *BMC Bioinformatics* 2006, **7:**62.
65. Korf I: **Gene finding in novel genomes.** *BMC Bioinformatics* 2004, **5:**59.
66. Blanco E, Abril JF: **Computational gene annotation in new genome assemblies using GeneID.** *Methods Mol Biol* 2009, **537:**243-261.
67. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, **268:**78-94.
68. Wu TD, Watanabe CK: **GMAP: a genomic mapping and alignment program for mRNA and EST sequences.** *Bioinformatics* 2005, **21:**1859-1875.
69. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Jr., Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL, White O: **Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies.** *Nucleic Acids Res* 2003, **31:**5654-5666.

70. Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, Radenbaugh A, Singh S, Swing V, Tissier C, Zhang P, Huala E: **The Arabidopsis Information Resource (TAIR): gene structure and function annotation.** *Nucleic Acids Res* 2008, **36:**D1009-1014.
71. Cannon SB, Sterck L, Rombauts S, Sato S, Cheung F, Gouzy J, Wang X, Mudge J, Vasdewani J, Schiex T, Spannagl M, Monaghan E, Nicholson C, Humphray SJ, Schoof H, Mayer KF, Rogers J, Quetier F, Oldroyd GE, Debelle F, Cook DR, Retzel EF, Roe BA, Town CD, Tabata S, Van de Peer Y, Young ND: **Legume genome evolution viewed through the Medicago truncatula and Lotus japonicus genomes.** *Proc Natl Acad Sci U S A* 2006, **103:**14959-14964.
72. Itoh T, Tanaka T, Barrero RA, Yamasaki C, Fujii Y, Hilton PB, Antonio BA, Aono H, Apweiler R, Bruskiewich R, Bureau T, Burr F, Costa de Oliveira A, Fuks G, Habara T, Haberer G, Han B, Harada E, Hiraki AT, Hirochika H, Hoen D, Hokari H, Hosokawa S, Hsing YI, Ikawa H, Ikeo K, Imanishi T, Ito Y, Jaiswal P, Kanno M, et al: **Curated genome annotation of Oryza sativa ssp. japonica and comparative genome analysis with Arabidopsis thaliana.** *Genome Res* 2007, **17:**175-183.
73. Project IRGS: **The map-based sequence of the rice genome.** *Nature* 2005, **436:**793-800.
74. Harris RS: **Improved pairwise alignment of genomic DNA.** *PhD.* Pennsylvania State University, Biology; 2007.
75. Miller Lab Software **[http://www.bx.psu.edu/miller_lab/].**
76. Li L, Stoeckert CJ, Jr., Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13:**2178-2189.
77. Duvick J, Fu A, Muppirala U, Sabharwal M, Wilkerson MD, Lawrence CJ, Lushbough C, Brendel V: **PlantGDB: a resource for comparative plant genomics.** *Nucleic Acids Res* 2008, **36:**D959-965.
78. PlantGDB **[http://www.plantgdb.org/]**
79. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC Bioinformatics* 2004, **5:**113.
80. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22:**2688-2690.
81. Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, Salzberg S L, Feng L, Jones M R, Skelton R L, Murray J E, Chen C, Qian W, Shen J, Du P, Eustice M, Tong E, Tang H, Lyons E, Paull R E, Michael T P, Wall K, Rice D W, Albert H, Wang M L, Zhu Y J, et al: **The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus).** *Nature* 2008, **452:**991-996.
82. Luo M, Wing RA: **An improved method for plant BAC library construction.** *Methods Mol Biol* 2003, **236:**3-20.
83. Kim H, San Miguel P, Nelson W, Collura K, Wissotski M, Walling JG, Kim JP, Jackson SA, Soderlund C, Wing RA: **Comparative physical mapping between Oryza sativa (AA genome type) and O. punctata (BB genome type).** *Genetics* 2007, **176:**379-390.

84. Ammiraju JS, Luo M, Goicoechea JL, Wang W, Kudrna D, Mueller C, Talag J, Kim H, Sisneros NB, Blackmon B, Fang E, Tomkins JB, Brar D, MacKill D, McCouch, N. Kurata, G. Lambert, D. W. Galbraith, K. Arumuganathan, K. Rao, Walling SJ, Gill N, Yu Y, SanMiguel P, Soderlund C, Jackson S, Wing RA: **The Oryza bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus Oryza.** *Genome Res* 2006, **16:**140-147.
85. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome Res* 1998, **8:**175-185.
86. Chou HH, Holmes MH: **DNA sequence quality trimming and vector removal.** *Bioinformatics* 2001, **17:**1093-1104.
87. Zuccolo A, Sebastian A, Talag J, Yu Y, Kim H, Collura K, Kudrna D, Wing RA: **Transposable element distribution, abundance and role in genome size variation in the genus Oryza.** *BMC Evol Biol* 2007, **7:**152.
88. SeqClean **[http://sourceforge.net/projects/seqclean/]**
89. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WE, Wetter T, Suhai S: **Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs.** *Genome Res* 2004, **14:**1147-1159.
90. Sonnhammer EL, Durbin R: **A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis.** *Gene* 1995, **167:**GC1-10.
91. Sputnik **[http://espressosoftware.com/sputnik/index.html]**
92. Morgante M, Hanafey M, Powell W: **Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes.** *Nat Genet* 2002, **30:**194-200.
93. Xiong Z, Kim JS, Pires JC: **Integration of genetic, physical, and cytogenetic maps for Brassica rapa chromosome A7.** *Cytogenet Genome Res* 2010, **129:**190-198.
94. Lee E, Harris N, Gibson M, Chetty R, Lewis S: **Apollo: a community resource for genome annotation editing.** *Bioinformatics* 2009, **25:**1836-1837.
95. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S: **The generic genome browser: a building block for a model organism system database.** *Genome Res* 2002, **12:**1599-1610.
96. Rice Annotation Project **[http://rapdb.dna.affrc.go.jp/]**
97. Grape Genome Browser **[http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis/]**

Table 1. Frequencies of BAC end sequences (BESs) and Sanger shot gun sequences (SGSs) matching sequences in Repbase (v. 15.08; [38])

		Absolute number in BESs	% BESs	% Repeats in BESs	Absolute number in SGSs	% SGSs	% Repeats in SGSs
	Type						
DNA-TEs	hAT	642 (1671)	0.92 (2.41)	6.84 (4.61)	20 (41)	0.74 (1.52)	5.73 (2.94)
	MuDR	343 (724)	0.49 (1.04)	3.65 (2.00)	7 (30)	0.26 (1.11)	2.00 (2.15)
	CACTA	27 (75)	0.04 (0.11)	0.29 (0.21)	0 (4)	0 (0.15)	0 (0.29)
	Helitrons	12 (69)	0.02 (0.10)	0.13 (0.19)	0 (3)	0 (0.11)	0 (0.22)
	Other	108 (595)	0.15 (0.86)	1.15 (1.64)	1 (24)	0.04 (0.89)	0.29 (1.72)
	Total	1132 (3134)	1.63 (4.51)	12.06 (8.64)	28 (102)	1.04 (3.78)	8.02 (7.31)
Retrotransposons	LTR Ty1-*copia*	2162 (9578)	3.11 (13.79)	23.02 (26.42)	64 (314)	2.37 (11.65)	18.34 (22.51)
	LTR Ty3-*gypsy*	2431 (8395)	3.50 (12.09)	25.89 (23.15)	129 (377)	4.78 (13.98)	36.96 (27.03)
	LTR not classified	720 (2868)	1.04 (4.13)	7.67 (7.91)	51 (139)	1.89 (5.16)	14.61 (0.96)
	LINEs	1876 (8055)	2.70 (11.60)	19.98 (22.22)	55 (294)	2.04 (10.91)	15.76 (21.08)
	SINEs	11 (183)	0.02 (0.26)	0.12 (0.50)	0 (4)	0 (0.15)	0 (0.29)
	Retro not classified	1058 (4046)	1.52 (5.82)	11.27 (11.16)	23 (165)	0.85 (6.12)	6.59 (11.83)
	Total	8258 (33125)	11.89 (47.69)	87.94 (91.36)	321 (1293)	11.91 (47.96)	91.98 (92.69)
Total		9390 (36259)	13.52 (52.20)	100 (100)	349 (1395)	12.95 (51.74)	100 (100)

Results in parentheses include Internal BlastN searches

Table 2. Putatively high-copy MITEs identified in the BESs and SGSs using MUST pipeline (see Figure S3 Additional file 1)

	Length	Inv. Repeats length	BES hits	Copy number estimate	SGS hits	Copy number estimate	AT%
MITE_1	358	26	542	~ 17000	18	~ 17200	68.80
MITE_2	190	19	140	~3300	8	~3100	68.70
MITE_3	516	47	394	~ 17900	8	~ 11300	75.20

Copy number estimates based on procedure of Hawkins et al. [50]

Table 3. Simple Sequence Repeats (SSRs) identified in the BESs and SGSs

Repeat	Amborella (BES) Ppmb*	Amborella (RS) Ppmb*	SoyBean Ppmb*	Oryza rufipogon Ppmb*
mono	149.66	152.89	72.74	50.79
di	225.03	211.00	77.89	63.94
tri	72.49	78.96	110.01	144.06
tetra	89.88	90.70	100.67	102.25
penta	74.85	89.73	64.54	56.00
total	611.92	623.28	425.85	417.04

* presence per million base pairs

Table 4. Statistics for cDNA sequences included in multi-library transcriptome assembly of 246,196 unigenes with lengths greater than 100 bp

Tissue – Library Name	Sequencing Method	Number of Reads	Unscreened Reads	Total passing bases
Apical meristem - Atr12	454 FLX Titanium	794,746	688,305	201.90 MB
Male flowers - Atr15	454 FLX Titanium	277,023	255,213	73.49 MB
Old leaves - Atr14	454 FLX Titanium	280,097	260,563	73.49 MB
Old Stem - Atr13	454 FLX Titanium	259,431	238,156	68.70 MB
Pre meiotic female flower buds - Atr10	454 FLX GS	895,000	812,325	176.97 MB
Premeiotic female flower bud - Atr02	Sanger	13,263	13,141	7.17 MB
Premeiotic male flower bud - Atr01	Sanger	25,343	25,006	14.17 MB
Root - Atr11	454 FLX GS	324,070	300,275	64.88 MB
Stem - Atr16	454 FLX Titanium	410,098	388,436	120.03 MB

Assemblies and raw data can be downloaded from the Ancestral Angiosperm Genome Project website [25]. A BLAST portal for the assembly is also available at the project website.

Figure legends

Figure 1. ML trees for reverse transcriptase genes classified as (a) Copia-type and (b) Gypsy-type LTR and (c) LINE elements show rate heterogeneity and no recent expansive radiations (i.e. short terminal branches). Reverse transcriptase sequences were mined from BAC end sequence set.

Figure 2. K-mer analyses of Sanger shotguns sequences reveal low frequencies of short repeats in the *Amborella* genome relative to the sorghum and maize genomes.

Figure 3. Variation in rates of structural evolution evident in parsimony mapping of losses of synteny with 29 gene blocks inferred for the last common ancestor of all extant flowering plant lineages.

Figure 4. Hybridization of 3 BAC clones in the minimum tiling paths for contigs 1003 and 431 to mitotic squashes ($2n$=26) verifies the FPC assemblies. A-E results for contig 1003; F-J results for contig 431. A and F show all three BAC-FISH probes merged. E and J = DAPI staining. B, C, and D show each of three BACs (red, green, white) for contig 1003. G, H, and I show each of three BACs (red, green, white) for contig 431.

Figure 5. LASTZ dot plots comparing BAC contig 1003 syntenic regions in the (a) grape and (b) rice genomes.

Figure 6. LASTZ dot plots comparing BAC contig 431 syntenic regions in the (a) grape and (b) rice genomes.

Figure 7. Gene trees for (a) auxin-independent growth promoter (*AXI1*), (b) ceramidase and (c) plant uncoupling mitochondrial protein 1 [*PUMP1*] gene families show divergence of genes on *Amborella* contig 431 diverging from lineages leading to *Vitis* γ homeologs mapping to syntenic blocks on chromosomes 6, 8 and 13 (shown in red). Genes sampled from major angiosperm lineages are highlighted.

Additional files

Additional file 1
Title: Amb_Additional_file1.doc
Description: Supplemental tables and figures cited with additional details for the physical map and shotgun sequences.

Additional file 2
Title: Amb_Additional_file2.xls
Description: Synteny analysis of *Amborella* BAC ends and *Vitis* genes.

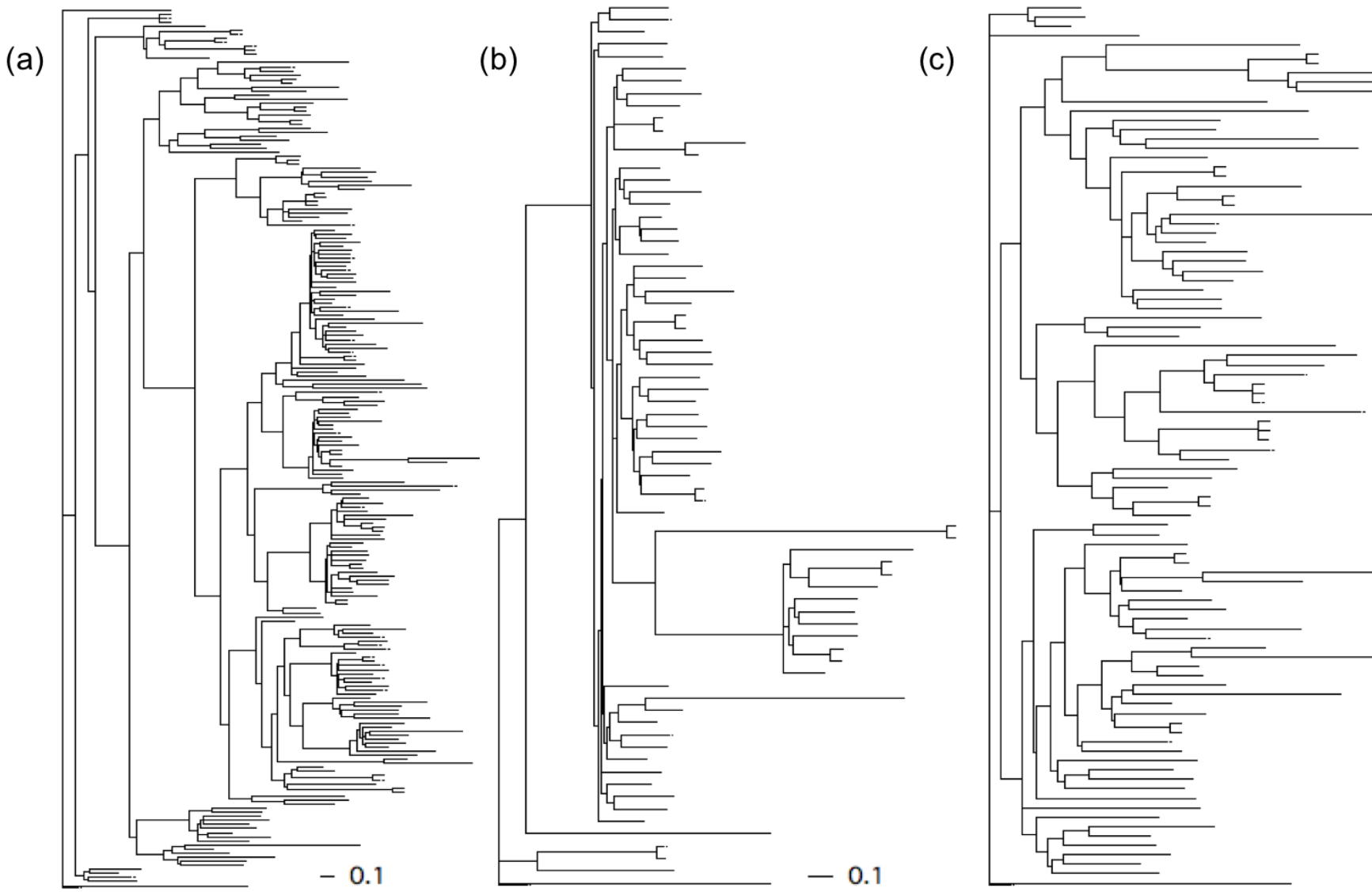


Figure 1

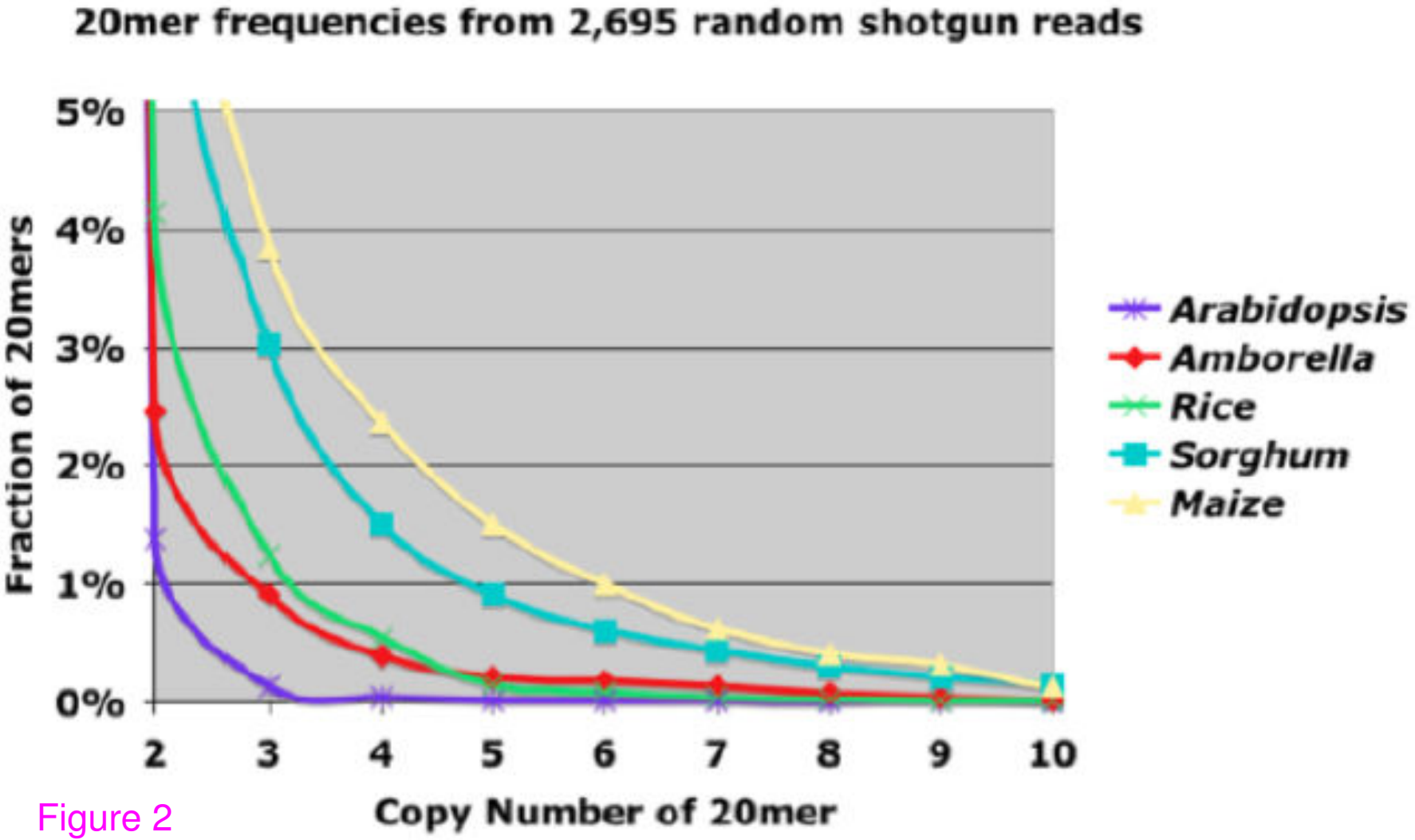

20mer frequencies from 2,695 random shotgun reads
Fraction of 20mers
5%
4%
3%
2%
1%
0%
2
3
4
5
6
7
8
9
10
Copy Number of 20mer
Arabidopsis
Amborella
Rice
Sorghum
Maize

Figure 2

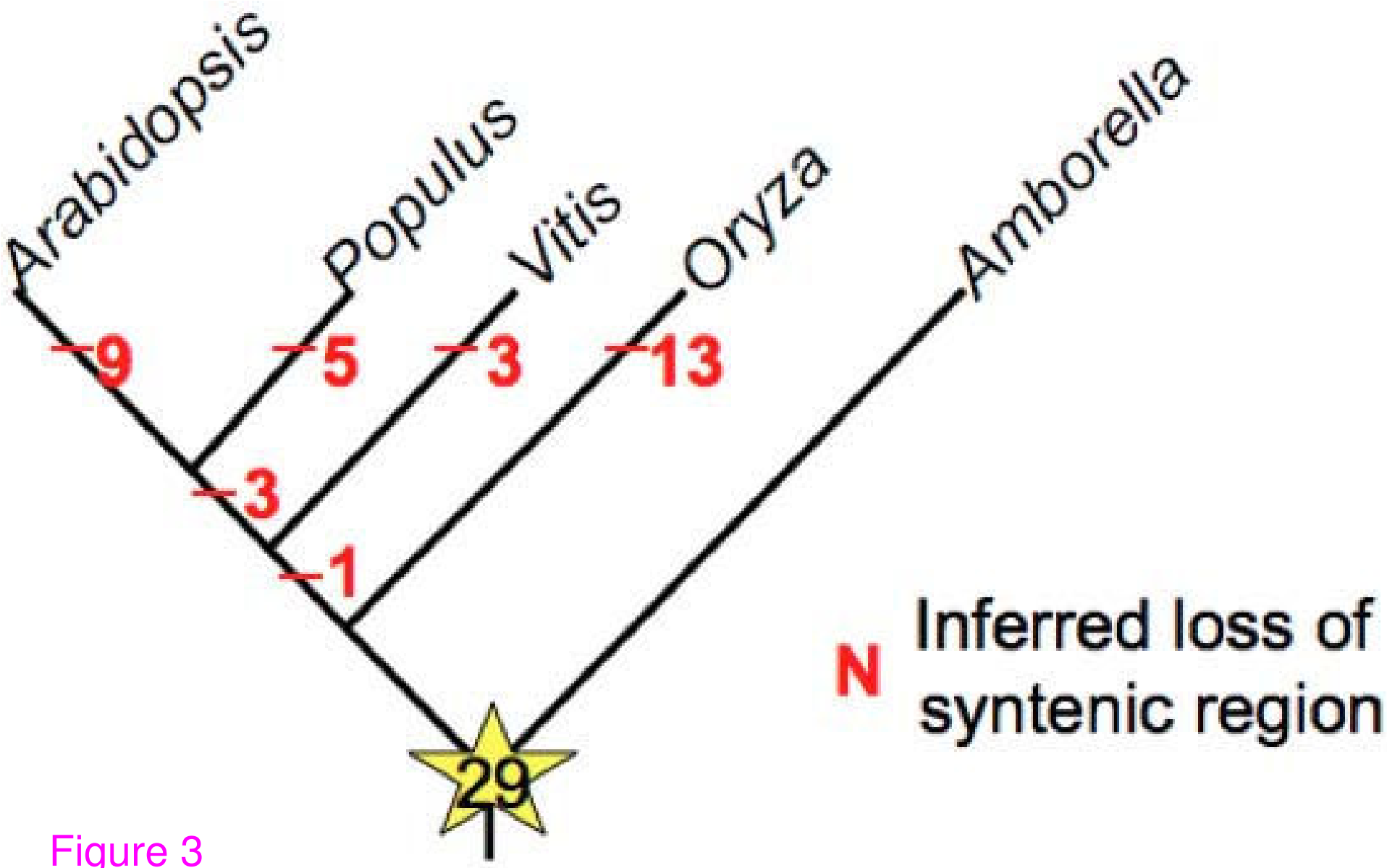

Arabidopsis
Populus
Vitis
Oryza
Amborella
9
5
3
13
3
1
29
N
Inferred loss of syntenic region

Figure 3

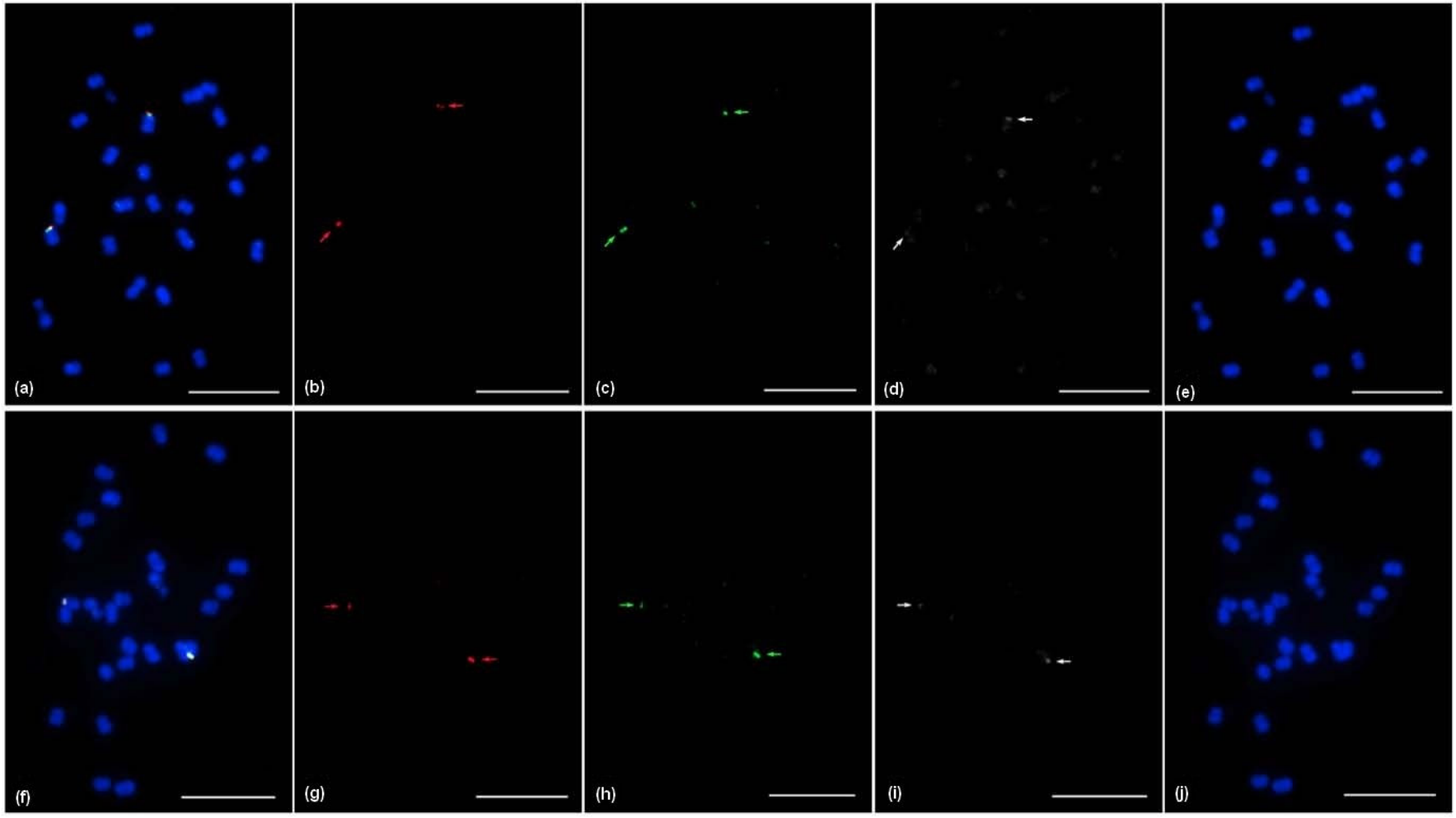


Figure 4

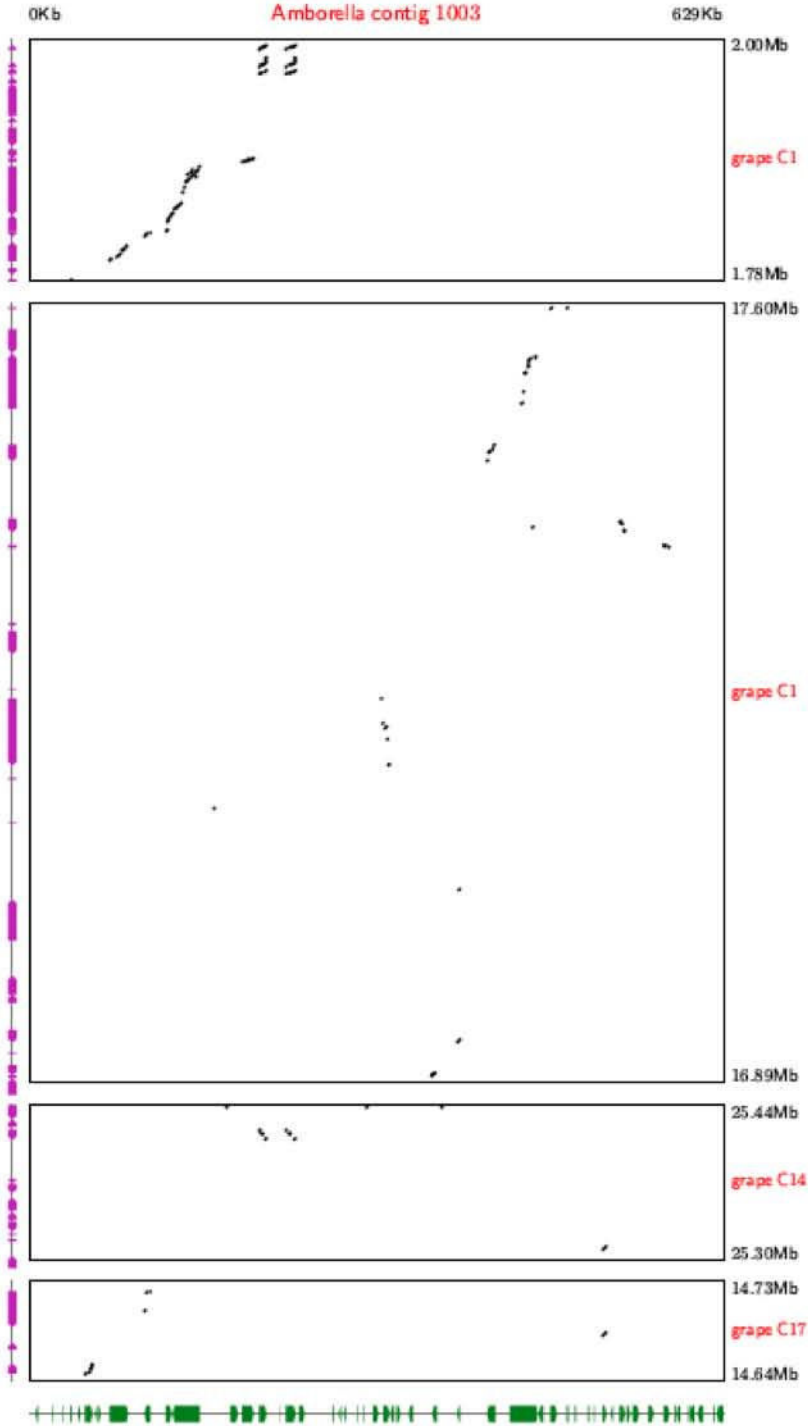
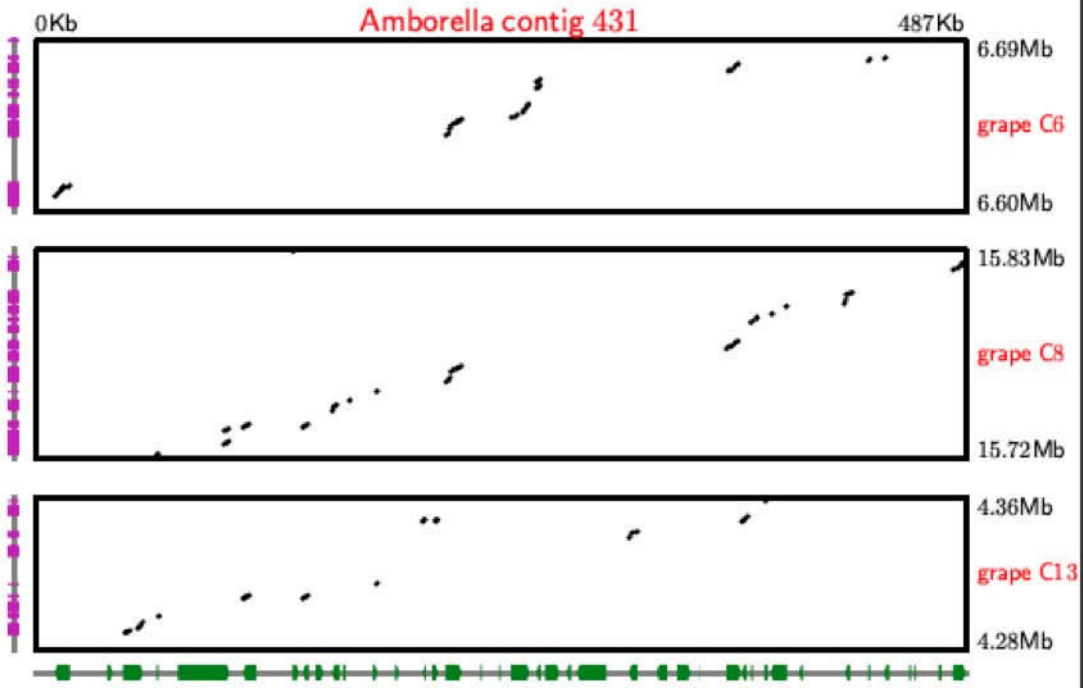

0Kb
Amborella contig 1003
629Kb
2.00Mb
grape C1
1.78Mb
17.60Mb
grape C1
16.89Mb
25.44Mb
grape C14
25.30Mb
14.73Mb
grape C17
14.64Mb

Figure 5


0Kb
Amborella contig 431
487Kb
6.69Mb
grape C6
6.60Mb
15.83Mb
grape C8
15.72Mb
4.36Mb
grape C13
4.28Mb

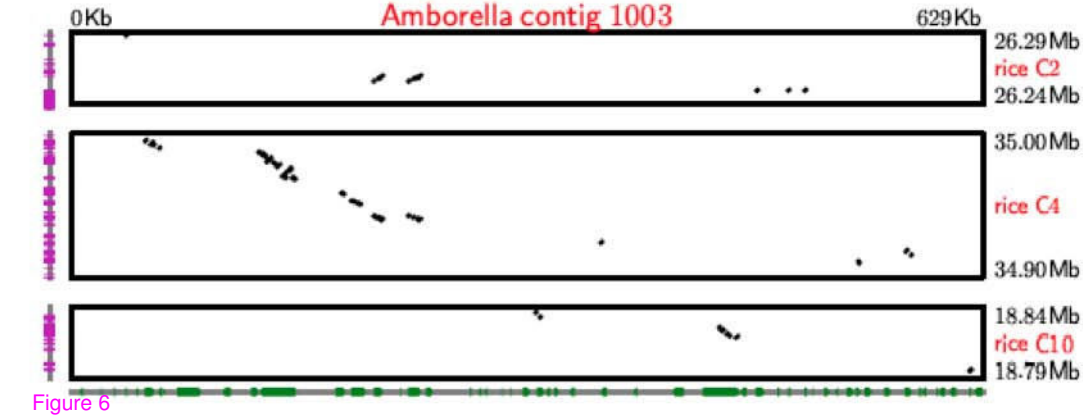

0Kb
Amborella contig 1003
629Kb
26.29Mb
rice C2
26.24Mb
35.00Mb
rice C4
34.90Mb
18.84Mb
rice C10
18.79Mb

Figure 6

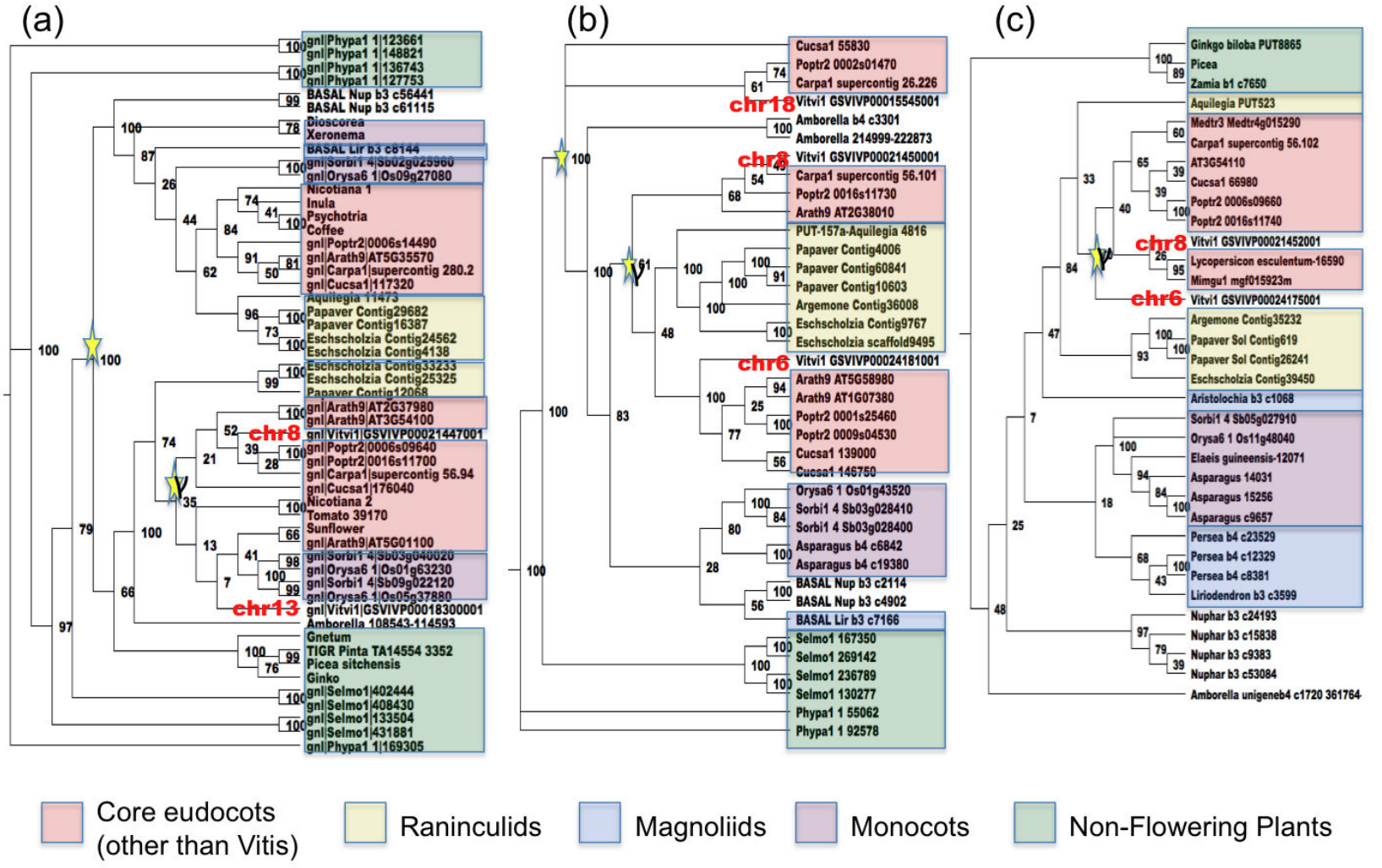

(a)
(b)
(c)
gnl|Phypa1 1|123661
gnl|Phypa1 1|148821
gnl|Phypa1 1|136743
gnl|Phypa1 1|127753
BASAL Nup b3 c56441
BASAL Nup b3 c61115
Dioscorea
Xeronema
BASAL Lir b3 c8144
gnl|Sorbi1 4|Sb02g025960
gnl|Orysa6 1|Os09g27080
Nicotiana 1
Inula
Psychotria
Coffee
gnl|Poptr2|0006s14490
gnl|Arath9|AT5G35570
gnl|Carpa1|supercontig 280.2
gnl|Cucsa1|117320
Aquilegia 11473
Papaver Contig29682
Papaver Contig16387
Eschscholzia Contig24562
Eschscholzia Contig4138
Eschscholzia Contig33233
Eschscholzia Contig25325
Papaver Contig12068
gnl|Arath9|AT2G37980
gnl|Arath9|AT3G54100
chr8
gnl|Vitvi1|GSVIVP00021447001
gnl|Poptr2|0006s09640
gnl|Poptr2|0016s11700
gnl|Carpa1|supercontig 56.94
gnl|Cucsa1|176040
Nicotiana 2
Tomato 39170
Sunflower
gnl|Arath9|AT5G01100
gnl|Sorbi1 4|Sb03g040020
gnl|Orysa6 1|Os01g63230
gnl|Sorbi1 4|Sb09g022120
gnl|Orysa6 1|Os05g37880
chr13
gnl|Vitvi1|GSVIVP00018300001
Amborella 108543-114593
Gnetum
TIGR Pinta TA14554 3352
Picea sitchensis
Ginko
gnl|Selmo1|402444
gnl|Selmo1|408430
gnl|Selmo1|133504
gnl|Selmo1|431881
gnl|Phypa1 1|169305
Cucsa1 55830
Poptr2 0002s01470
Carpa1 supercontig 26.226
chr18
Vitvi1 GSVIVP00015545001
Amborella b4 c3301
Amborella 214999-222873
chr8
Vitvi1 GSVIVP00021450001
Carpa1 supercontig 56.101
Poptr2 0016s11730
Arath9 AT2G38010
PUT-157a-Aquilegia 4816
Papaver Contig4006
Papaver Contig60841
Papaver Contig10603
Argemone Contig36008
Eschscholzia Contig9767
Eschscholzia scaffold9495
chr6
Vitvi1 GSVIVP00024181001
Arath9 AT5G58980
Arath9 AT1G07380
Poptr2 0001s25460
Poptr2 0009s04530
Cucsa1 139000
Cucsa1 146750
Orysa6 1 Os01g43520
Sorbi1 4 Sb03g028410
Sorbi1 4 Sb03g028400
Asparagus b4 c6842
Asparagus b4 c19380
BASAL Nup b3 c2114
BASAL Nup b3 c4902
BASAL Lir b3 c7166
Selmo1 167350
Selmo1 269142
Selmo1 236789
Selmo1 130277
Phypa1 1 55062
Phypa1 1 92578
Ginkgo biloba PUT8865
Picea
Zamia b1 c7650
Aquilegia PUT523
Medtr3 Medtr4g015290
Carpa1 supercontig 56.102
AT3G54110
Cucsa1 66980
Poptr2 0006s09660
Poptr2 0016s11740
chr8
Vitvi1 GSVIVP00021452001
Lycopersicon esculentum-16590
Mimgu1 mgf015923m
chr6
Vitvi1 GSVIVP00024175001
Argemone Contig35232
Papaver Sol Contig619
Papaver Sol Contig26241
Eschscholzia Contig39450
Aristolochia b3 c1068
Sorbi1 4 Sb05g027910
Orysa6 1 Os11g48040
Elaeis guineensis-12071
Asparagus 14031
Asparagus 15256
Asparagus c9657
Persea b4 c23529
Persea b4 c12329
Persea b4 c8381
Liriodendron b3 c3599
Nuphar b3 c24193
Nuphar b3 c15838
Nuphar b3 c9383
Nuphar b3 c53084
Amborella unigeneb4 c1720 361764
Core eudocots
(other than Vitis)
Raninculids
Magnoliids
Monocots
Non-Flowering Plants

Figure 7

Additional files provided with this submission:

Additional file 1: Amb_Additional_file1.doc, 971K
http://genomebiology.com/imedia/1981333735341649/supp1.doc
Additional file 2: Amb_Additional_file2.xls, 1577K
http://genomebiology.com/imedia/1069517148554677/supp2.xls