Correspondence

The *Amborella* genome: an evolutionary reference for plant biology

Douglas E Soltis[1], Victor A Albert[2,3], Jim Leebens-Mack[4], Jeffrey D Palmer[5], Rod A Wing[6], Claude W dePamphilis[7], Hong Ma[7], John E Carlson[8], Naomi Altman[9], Sangtae Kim[10], P Kerr Wall[7], Andrea Zuccolo[6] and Pamela S Soltis[11]

Addresses: [1]Department of Botany and the Genetics Institute, University of Florida, Gainesville, FL 32611, USA. [2]Joint Centre for Bioinformatics in Oslo, University of Oslo and Rikshospitalet HF, Blindern, NO-0316 Oslo, Norway. [3]Department of Biological Sciences, University at Buffalo (SUNY), Buffalo, NY 14260-1300, USA. [4]Department of Plant Biology, University of Georgia, Athens, GA 30602, USA. [5]Department of Biology, Indiana University, Bloomington, IN 47405, USA. [6]Department of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA. [7]Department of Biology, the Huck Institutes of the Life Sciences, and the Institute of Molecular Evolutionary Genetics, Pennsylvania State University, University Park, PA 16802, USA. [8]School of Forest Resources, Pennsylvania State University, University Park, PA 16802, USA. [9]Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA. [10]National Institute of Biological Resources, Incheon 404-170, Korea. [11]Florida Museum of Natural History and the Genetics Institute, University of Florida, Gainesville, FL 32611, USA.

Correspondence: Pamela S Soltis. Email: psoltis@flmnh.ufl.edu

Published: 10 March 2008

Genome ***Biology*** 2008, **9:**402 (doi:10.1186/gb-2008-9-3-402)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2008/9/3/402

Abstract

The nuclear genome sequence of *Amborella trichopoda*, the sister species to all other extant angiosperms, will be an exceptional resource for plant genomics.

The origin and evolution of the angiosperms is one of the great terrestrial radiations and has had manifold effects on the global biota. Today, flowering plants generate the vast majority of human food, either directly or indirectly as animal feed, and account for a huge proportion of land-based photosynthesis and carbon sequestration. With a fossil record that extends back to just over 130 million years ago, flowering plants have diversified to include 250,000 to possibly 400,000 species occupying nearly every habitable terrestrial environment, and many aquatic ones. Understanding how angiosperms have accomplished this feat over a relatively short span of evolutionary time will elucidate many of the key processes underlying the assembly of Earth's plant/animal associations and entire ecosystems.

Many scientists have understood the importance of broad, comparative genome sequencing since the beginning of the *Arabidopsis thaliana* and rice (*Oryyza sativa*) genome sequencing projects [1-4]. *Arabidopsis*, a relative of cabbage, had already become the premier model for plant genetics, and half the world's dependence on rice for food makes that crop plant an important model for the genetic architecture of traits important to humanity. More recently, poplar (*Populus trichocarpa*), grapevine (*Vitis vinifera*) and papaya (*Carica papaya*) have been sequenced as genomic models for woody crop plants [5-12]. These advances have been motivated by the realization that understanding the structure and evolution of plant genomes would contribute to society through enhancements to agriculture and forestry [13].

However, the few angiosperm nuclear genomes that have been sequenced so far reside on just two limbs within the angiosperm branch of the Tree of Life [14,15] and, therefore, aid us little in understanding the characteristics of the last common ancestor of all angiosperms (Figure 1). Many key angiosperm innovations, such as the origin of the flower and fruit, diverse pollination systems and double fertilization, large water-conducting vessel elements, diverse biochemical pathways, and many of the specific genes that regulate

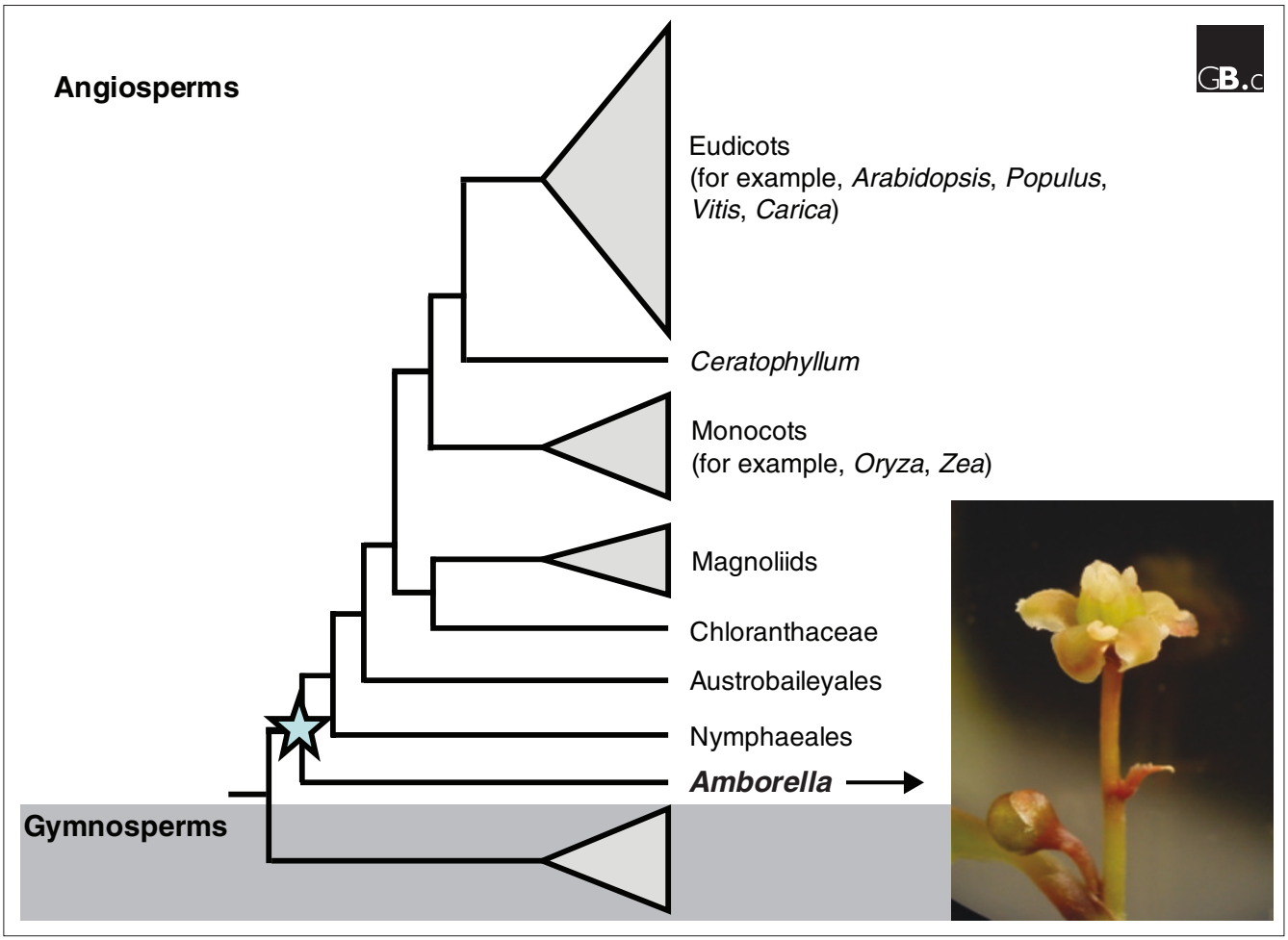

Figure 1
The position of *Amborella* in the angiosperm phylogenetic tree. Taxa for which whole-genome sequences have been published are indicated in parentheses. The node highlighted by a star on the tree identifies the 'ancestral angiosperm', or most recent common ancestor of all living angiosperms. An *Amborella* genome sequence will allow the ancestral genes and genomic features of living angiosperms to be identified and will provide the essential root for angiosperm comparative genomics. Based on [14,15].

key growth and developmental processes, appeared first among the basal angiosperm lineages [16-20]. A thorough understanding of processes that shape genes and genomic features, and of the many similarities and differences between model monocots (for example, *Oryza*) and eudicots (for example, *Arabidopsis*), requires a perspective based on evolutionary lineages. Such perspectives can be obtained only through analysis of an appropriately broad sampling of genomes, including lineages branching from the most basal node on the angiosperm tree [21]. But which basal angiosperm(s) should be given the highest priority for sequencing in the near future?

Recent phylogenetic analyses [14,15,17, 22] have identified *Amborella trichopoda*, a large shrub known only from the island of New Caledonia, as the single 'sister species' to all other living flowering plants. *Amborella* therefore offers the unparalleled potential to 'root' analyses of all angiosperm features, from gene families to genome structure, and from physiology to morphology. Furthermore, as the branching-point for *Amborella* is situated 'between' gymnosperms and all other angiosperms, a genome sequence for *Amborella* would help characterize processes that distinguish these two lineages of extant seed plants. The nuclear genome sequence of *Amborella* would contribute uniquely to efforts to reconstruct characteristics of the 'ancestral angiosperm'. The importance of *Amborella* in this regard is already widely appreciated [19,23]. Two recent papers, in fact, point specifically to basal angiosperms, including *Amborella*, as obvious choices for future nuclear genome sequencing efforts [24,25].

The genome structure of the ancestral angiosperm is currently much debated: did a whole-genome duplication predate or coincide with the origin of angiosperms (perhaps catalyzing innovation) or did the whole-genome duplication reported for several lineages of basal angiosperms [26] occur after the

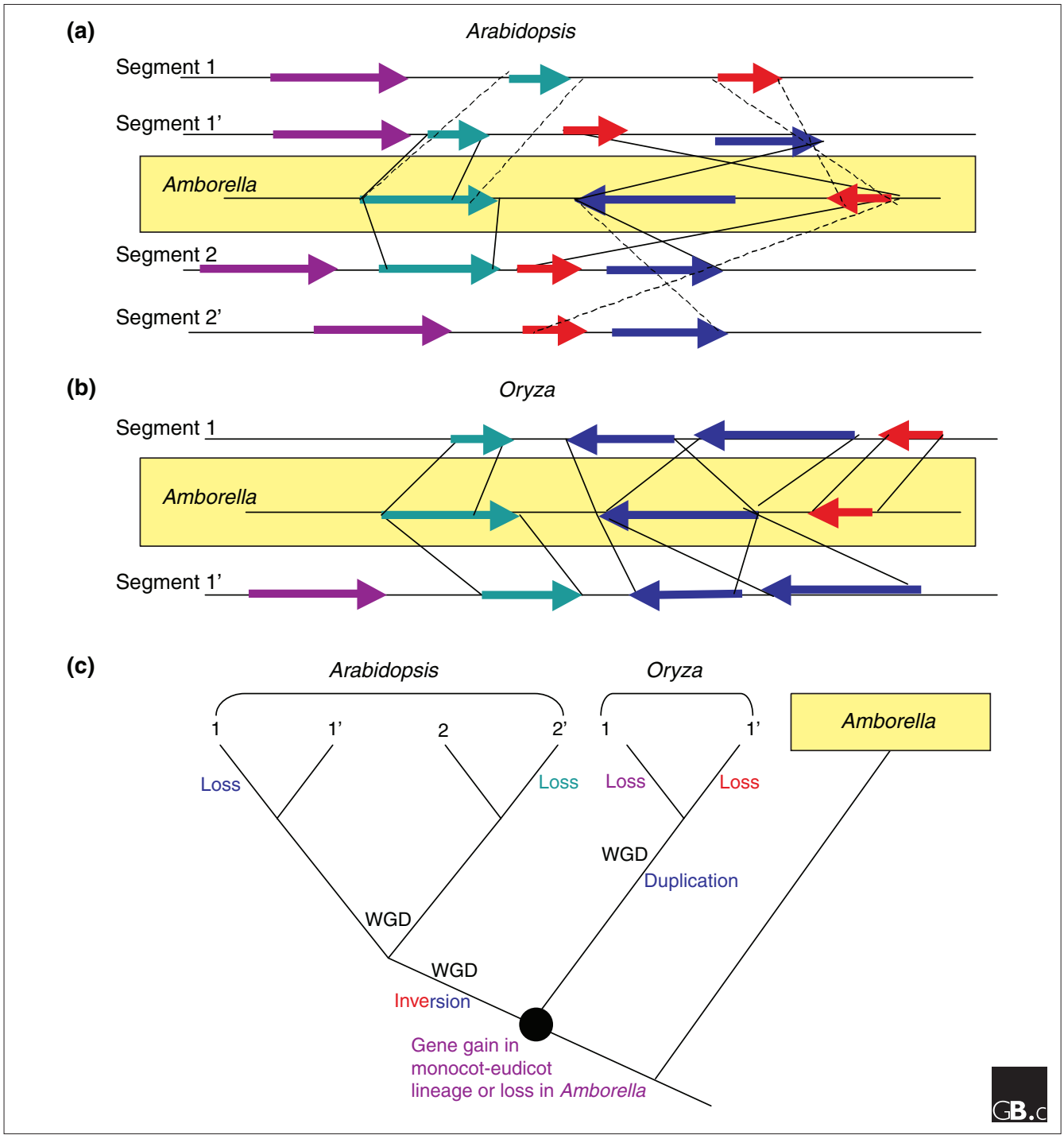

Figure 2

Sequencing the nuclear genome for *Amborella* will root comparisons of monocot and eudicot genome sequences. **(a,b)** Sequence-based comparisons of the *Amborella* sequence (highlighted in yellow) with (a) *Arabidopsis* and (b) rice (*Oryza*) sequences for homologous genome segments (1, 1', 2 and 2') identify homologous genomic regions and genes (shown by colored arrows) that have undergone duplications and presumed gene loss in different segments. **(c)** From such comparisons investigators can identify the timings of segmental duplications and inversions, gene gains and losses, and whole-genome duplications (WGDs) in these three lineages. The large black circle indicates the monocot-eudicot split. The *Amborella* sequence resolves the timing of an inversion and a tandem duplication (versus loss of a duplicate) that distinguish homologous *Arabidopsis* and rice segments. Taken together, the map comparisons imply that the orientation of the green, blue and red genes in the *Amborella* sequence matches that in the common ancestor of monocots and eudicots. We can also infer that the purple gene was present in the common ancestor of monocots and eudicots. However, the homologous region would have to be sequenced in a gymnosperm to determine whether this gene was gained on the lineage leading to monocots and eudicots, or was present in the common ancestor of eudicots, monocots and *Amborella* and lost in the lineage leading to *Amborella*.

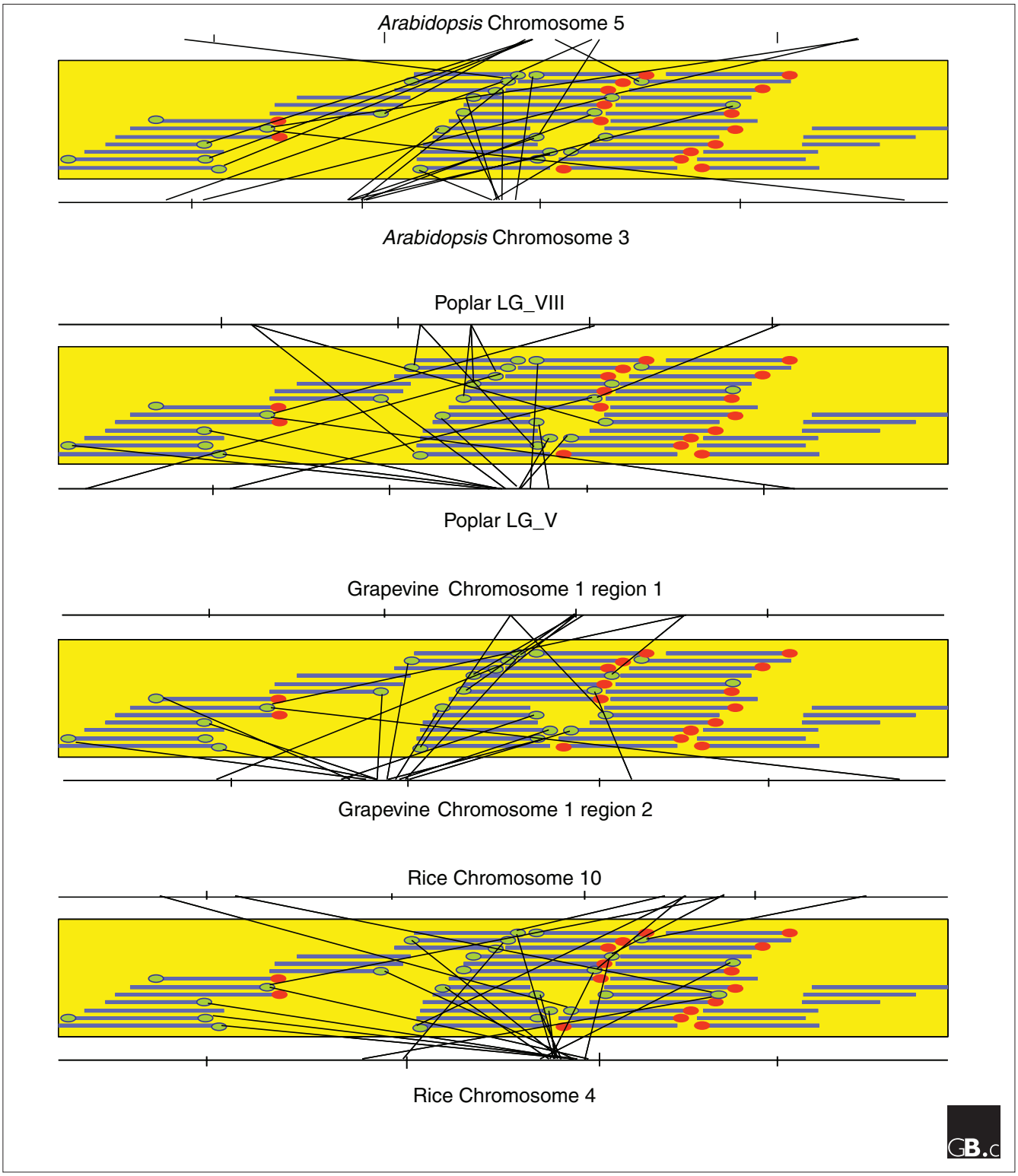

Figure 3
Synteny of the *Amborella* genome with other plant genomes. Illustrated here is a physical map of a 0.65 Mb region of the *Amborella* nuclear genome (highlighted in yellow) showing synteny with segments in each of the *Arabidopsis*, poplar, grapevine, and rice genomes. Two homologous segments are shown in each case: one above and one below the *Amborella* map. The physical map is based on high information content fingerprinting of an *Amborella* BAC library. Synteny was inferred over 5 Mb tracts of sequenced genomes on the basis of BAC-end sequences matching the reference genomes with TBLASTX bit scores of greater than 80. Red and green ovals depict BAC-end *Amborella* sequences with significant hits to known transposable elements and protein-coding genes, respectively.

divergence of *Amborella*? Was the common ancestor of *Vitis*, *Populus*, and *Arabidopsis* an ancient hexaploid that arose after the monocot-eudicot split? Did a separate genome-wide duplication occur early in monocot evolutionary history [8,11]? The answers to these questions are crucial for understanding angiosperm genome evolution and the diversification of flowering plants themselves. The *Amborella* Genome Project will address fundamental questions relating to the early evolution of gene content and genome structure in angiosperms (Figure 2), while providing comprehensive genomic resources for researchers studying all aspects of angiosperm biology [27].

In addition, two features of *Amborella*'s truly extraordinary mitochondrial genome raise compelling questions that warrant the sequencing of the *Amborella* nuclear genome. First, the *Amborella* mitochondrial genome is extraordinarily rich in 'foreign' genes acquired by horizontal gene transfer, far richer than any other plant mitochondrial genome [28]. These foreign genes were acquired from a wide range of donors. These findings raise important questions that can best be addressed with a complete nuclear genome sequence. For instance, is the *Amborella* nuclear genome also exceptionally rich in foreign sequences, and were these sequences acquired from the same donors as the foreign mitochondrial sequences? The *Amborella* nuclear genome sequence will enable subsequent experiments to determine what roles, if any, foreign nuclear genes play in *Amborella*. Second, the *Amborella* mitochondrial genome is exceptionally large, and much of the extra DNA is of unknown origin (Rice DW, Richardson AO, Young GJ, Sanchez-Puerta MV, Zhang Y, CWD, Knox EB, Munzinger J, Boore J, JDP, unpublished observations). We suspect that much of this unknown DNA was probably acquired from *Amborella*'s nuclear genome, a hypothesis that can only be tested once a complete nuclear sequence is available.

Ongoing deep transcriptome sequencing and physical mapping [26,29,30] form the foundation for this important project. *Amborella* cDNA sequences have already rooted gene trees and illuminated the timing of gene diversification relative to the origin of the angiosperms for many gene families ([31-34] and Duarte JR, Wall PK, Barakat A, Zhang J, Cui L, Landherr LL, Leebens-Mack J, Ma H, CWD, Kim S, *et al.*, unpublished observations), and the potential for further evolutionary orientation of other gene families is great. The generation and analysis of a bacterial artificial chromosomes (BAC) fingerprint/end sequence physical map of the relatively small, 870 Mb *Amborella* genome [26] is already yielding new and exciting information about the genome structure of the earliest angiosperms and the retention of some syntenic blocks throughout angiosperm history (Figure 3). The physical map will also serve as a framework for assembling the sequence of the *Amborella* genome.

Given the available genomic infrastructure, the importance of *Amborella* as the sister to all other extant angiosperms, the large community of plant biologists who require a universal evolutionary reference for their studies, and the availability of cost-effective, ultra-high-throughput DNA sequencing technologies, it is our opinion that the *Amborella* genome is in an extremely strong position to warrant complete sequencing in the near future. Thus, the stage is set for a large-scale international *Amborella* genome sequencing initiative in support of fundamental and applied plant sciences, and we enthusiastically advocate such an endeavor.

Acknowledgements

This work was supported in part by NSF grant PGR-0638595, DBI-207202 and NIH grant RO1-GM-70612.

References

1. Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.** *Nature* 2000, **408**:796-815.
2. **The *Arabidopsis* Information Resource** [http://www.arabidopsis.org]
3. International Rice Genome Sequencing Project: **The map-based sequence of the rice genome.** *Nature* 2001, **441**:337-340.
4. **Rice Annotation Database** [http://rad.dna.affrc.go.jp]
5. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, *et al.*: **The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray).** *Science* 2006, **313**:1596-1604.
6. **The International Populus Genome Consortium** [http://www.ornl.gov/sci/ipgc]
7. **JGI Populus trichocarpa v1.1** [http://genome.jgi-psf.org/Poptr1_1]
8. 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, Mica E, Jublot D, Poulain J, Bruyère 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-U465.
9. **International Grape Genome Program - IGGP** [http://www.vitaceae.org]
10. **Grape Genome Browser** [http://www.genoscope.cns.fr/externe/English/Projets/Projet_ML]
11. Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, Fitzgerald LM, Vezzulli S, Reid J, Malacarne G, Iliev D, Coppola G, Wardell B, Micheletti D, Macalma T, Facci M, Mitchell JT, Perazzolli M, Eldredge G, Gatto P, Oyzerski R, Moretto M, Gutin N, Stefanini M, Chen Y, Segala C, Davenport C, Demattè L, Mraz A, *et al.*: **A high quality draft consensus sequence of the genome of a heterozygous grapevine variety.** *PLoS ONE* 2007, **2**:e1326.
12. Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KLT, Salzberg SL, Feng L, Jones MR, Skelton RL, Murray JE, Chen C, Qian W, Shen J, Du P, Eustice M, Tong E, Tang H, Lyons E, Paull RE, Michael TP, Wall K, Rice DW, Albert H, Wang M-L, Zhu YJ, *et al.*: **The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus).** *Nature*, in press.
13. Committee on Objectives for the National Plant Genome Initiative: *2003-2008, National Research Council: The National Plant Genome Initiative: Objectives for 2003-2008.* Washington, DC, USA: National Academies Press; 2002.
14. Jansen RK, Cai Z, Raubeson LA, Daniell H, dePamphilis CW, Leebens-Mack J, Müller 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 USA* 2007, **104**:19369-19374.

15. Moore MJ, Bell CD, Soltis PS, Soltis DE: **Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms.** *Proc Natl Acad Sci USA* 2007, **104:**19363-19368.
16. Soltis DE, Soltis PS, Albert VA, Oppenheimer DG, dePamphilis CW, Ma H, Frohlich MW, Theissen G, Floral Genome Project Research Group: **Missing links: the genetic architecture of flower and floral diversification.** *Trends Plant Sci* 2002, **7:**22-31.
17. Soltis DE, Soltis PS, Endress PK, Chase MW: *Phylogeny and Evolution of Angiosperms.* Sunderland, MA, USA: Sinauer; 2005.
18. Williams JH, Friedman WE: **Identification of diploid endosperm in an early angiosperm lineage.** *Nature* 2002, **415:**522-526.
19. Friedman WE: **Embryological evidence for developmental lability during early angiosperm evolution.** *Nature* 2006, **441:**337-340.
20. Duarte JM, Wall PK, Zahn LM, Soltis PS, Soltis DE, Leebens-Mack J, Ma H, Carlson JE, dePamphilis CW: **Utility of *Amborella trichopoda* and *Nuphar advena* ESTs for phylogeny and comparative sequence analysis.** *Taxon*, in press.
21. Committee on the National Plant Genome Initiative: Achievements and Future Directions, National Research Council: *Achievements of the National Plant Genome Initiative and New Horizons in Plant Biology.* Washington, DC, USA: National Academies Press; 2008. [http://www.nap.edu/catalog.php?record_id=12054]
22. Soltis PS, Soltis DE, Chase MW: **Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology.** *Nature* 1999, **402:**402-404.
23. 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 USA* 2005, **102:**4649-4654.
24. Pryer KM, Schneider H, Zimmer EA, Banks JA: **Deciding among green plants for whole genome studies.** *Trends Plant Sci* 2002, **7:**550-554.
25. Jackson S, Rounsley S, Purugganan M: **Comparative sequencing of plant genomes: choices to make.** *Plant Cell* 2006, **18:**1100-1104.
26. 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.
27. **AMBORELLA** [http://www.amborella.org]
28. Bergthorsson U, Richardson AO, Young GJ, Goertzen LR, Palmer JD: **Massive horizontal transfer of mitochondrial genes from diverse land plant donors to the basal angiosperm *Amborella*.** *Proc Natl Acad Sci USA* 2004, **101:**17747-17752.
29. 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.
30. Soltis DE, Ma H, Frohlich MW, Soltis PS, Albert VA, Oppenheimer DG, Altman NS, dePamphilis C, Leebens-Mack J: **The floral genome: an evolutionary history of gene duplication and shifting patterns of gene expression.** *Trends Plant Sci* 2007, **12:**358-367.
31. Kim S, Yoo MJ, Albert VA, Farris JS, Soltis PS, Soltis DE: **Phylogeny and diversification of B-function MADS-box genes in angiosperms: evolutionary and functional implications of a 260-million-year-old duplication.** *Am J Bot* 2004, **91:**2102-2118.
32. Kim S, Soltis PS, Wall K, Soltis DE: **Phylogeny and domain evolution in the APETALA2-like gene family.** *Mol Biol Evol* 2006, **23:**107-120.
33. Zahn LM, King HZ, 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.
34. Yoo MJ, Albert VA, Soltis PS, Soltis DE: **Phylogenetic diversification of glycogen synthase kinase 3/SHAGGY-like kinase genes in plants.** *BMC Plant Biol* 2006, **6:**3.