Assembly and Validation of the Genome of the Nonmodel Basal Angiosperm *Amborella*

Srikar Chamala *et al.*

Science **342**, 1516 (2013);
DOI: 10.1126/science.1241130

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

References and Notes

1. D. Martinez *et al.*, *Nat. Biotechnol.* **26**, 553–560 (2008).
2. E. A. Bayer, J. P. Belaich, Y. Shoham, R. Lamed, *Annu. Rev. Microbiol.* **58**, 521–554 (2004).
3. R. H. Doi, A. Kosugi, *Nat. Rev. Microbiol.* **2**, 541–551 (2004).
4. I. A. Kataeva *et al.*, *J. Bacteriol.* **191**, 3760–3761 (2009).
5. A. Lochner *et al.*, *Appl. Environ. Microbiol.* **77**, 4042–4054 (2011).
6. S. J. Yang *et al.*, *Appl. Environ. Microbiol.* **75**, 4762–4769 (2009).
7. Y. Vazana, S. Moraïs, Y. Barak, R. Lamed, E. A. Bayer, *Appl. Environ. Microbiol.* **76**, 3236–3243 (2010).
8. E. Berger, D. Zhang, V. V. Zverlov, W. H. Schwarz, *FEMS Microbiol. Lett.* **268**, 194–201 (2007).
9. V. Zverlov, S. Mahr, K. Riedel, K. Bronnenmeier, *Microbiology* **144**, 457–465 (1998).
10. S. Shoemaker *et al.*, *Biotechnology* **1**, 691–696 (1983).
11. J. O. Baker *et al.*, *Appl. Biochem. Biotechnol.* **45-46**, 245–256 (1994).
12. J. O. Baker, C. I. Ehrman, W. S. Adney, S. R. Thomas, M. E. Himmel, *Appl. Biochem. Biotechnol.* **70-72**, 395–403 (1998).
13. L. C. Textor *et al.*, *FEBS J.* **280**, 56–69 (2013).
14. L. P. Walker, C. D. Belair, D. B. Wilson, D. C. Irwin, *Biotechnol. Bioeng.* **42**, 1019–1028 (1993).
15. M. Gruno, P. Väljamäe, G. Pettersson, G. Johansson, *Biotechnol. Bioeng.* **86**, 503–511 (2004).
16. W. Liebl, J. Gabelsberger, K. H. Schleifer, *Mol. Gen. Genet.* **242**, 111–115 (1994).
17. M. M. Patel, R. M. Bhatt, *J. Chem. Technol. Biotechnol.* **53**, 253–263 (1992).
18. M. J. Playne, *Biotechnol. Bioeng.* **26**, 426–433 (1984).
19. C. E. Wyman *et al.*, *Bioresour. Technol.* **96**, 2026–2032 (2005).
20. H. Alizadeh, F. Teymouri, T. I. Gilbert, B. E. Dale, *Appl. Biochem. Biotechnol.* **124**, 1133–1141 (2005).
21. S. P. S. Chundawat *et al.*, *Bioresour. Technol.* **101**, 8429–8438 (2010).
22. R. Brunecky *et al.*, *Biotechnol. Bioeng.* **102**, 1537–1543 (2009).
23. S. P. S. Chundawat *et al.*, *Energ. Environ. Sci.* **4**, 973–984 (2011).
24. M. G. Resch *et al.*, *Energy Environ. Sci.* **6**, 1858 (2013).

Acknowledgments: This work was supported by the BioEnergy Science Center (BESC). BESC is a U.S. Department of Energy (DOE) Bioenergy Research Center supported by the Office of Biological and Environmental Research in the U.S. DOE Office of Science. We acknowledge colleagues at the Biomass Conversion Research Laboratory at Michigan State University for providing the AFEX-pretreated materials. Structures have been deposited in the Protein Data Bank with PDB codes 4DOD (GH9), 4DOE (GH9-CB), and 4EL8 (GH48).

Supplementary Materials

www.sciencemag.org/content/342/6165/1513/suppl/DC1
Materials and Methods
Figs. S1 to S17
Supplementary Text S1 to S6
Tables S1 to S8
References (*25*–*35*)

5 August 2013; accepted 8 November 2013
10.1126/science.1244273

Assembly and Validation of the Genome of the Nonmodel Basal Angiosperm *Amborella*

Srikar Chamala,[1]* Andre S. Chanderbali,[1,2]* Joshua P. Der,[3] Tianying Lan,[4] Brandon Walts,[1] Victor A. Albert,[4] Claude W. dePamphilis,[3] Jim Leebens-Mack,[5] Steve Rounsley,[6] Stephan C. Schuster,[7,8,9] Rod A. Wing,[10,11] Nianqing Xiao,[12] Richard Moore,[12] Pamela S. Soltis,[2,13] Douglas E. Soltis,[1,2,13] W. Brad Barbazuk[1,13]†

Genome sequencing with next-generation sequence (NGS) technologies can now be applied to organisms pivotal to addressing fundamental biological questions, but with genomes previously considered intractable or too expensive to undertake. However, for species with large and complex genomes, extensive genetic and physical map resources have, until now, been required to direct the sequencing effort and sequence assembly. As these resources are unavailable for most species, assembling high-quality genome sequences from NGS data remains challenging. We describe a strategy that uses NGS, fluorescence in situ hybridization, and whole-genome mapping to assemble a high-quality genome sequence for *Amborella trichopoda*, a nonmodel species crucial to understanding flowering plant evolution. These methods are applicable to many other organisms with limited genomic resources.

Amborella (*1*, *2*) has been identified as the single sister species to all other living angiosperms and is a pivotal reference for comparison to other angiosperms (*3*). However, *Amborella* is not a genetic model and has no existing genetic map, genetic resources, or genome sequence. Although next-generation sequencing (NGS) provides deep genomic sequence coverage at low cost, short-read assembly remains difficult, and assessing assembly accuracy is problematic without independently derived genomic maps. We produced a whole-genome assembly for *Amborella* from a mixed data set of 454, Illumina, and Sanger bacterial artificial chromosome (BAC)–end sequences, evaluated the assembly using fluorescence in situ hybridization (FISH), and improved contiguity using whole-genome mapping. FISH has broad utility (*4*), but has not been used in de novo genome assembly. Likewise, whole-genome mapping has been used to assemble bacterial genomes (*5*, *6*), but has only recently been applied to complex genomes of model organisms (*7*, *8*) to assist with scaffolding and correction of well-advanced genome assemblies.

More than 23 Gb of quality-filtered (*9*) DNA sequence comprising single-end (SE) 454-FLX, SE 454-FLX+ reads, 11-kb paired-end (PE) 454-FLX, 3-kb PE Illumina HiSeq, and Sanger-sequenced BAC-end reads (*10*) were combined and assembled (table S1). Assembly (*9*) resulted in 5745 scaffolds totaling 706 Mb (table S5) with a mean scaffold size of 123 kb and an N50 size of 4.9 Mb, and N90 scaffold metrics that indicate that 90% of our assembled sequence resides within 155 scaffolds greater than 1.1 Mb in length (table S5).

Flow cytometry was used to estimate the size of the *Amborella* genome at ~870 Mb (*11*), while our sequence-based size assessments (*9*, *10*, *12*, *13*) suggest that the *Amborella* genome size is closer to 748 Mb. Our high-quality sequence represents an average depth of coverage of ~31×, and the assembly covers >94% of the genome.

Long contig and scaffold assemblies are required to understand genome structure, enable gene identification, and support subsequent comparative, structural, and population genomics studies. We sought long continuous stretches of assembled sequence that represent all, or a major fraction of, the *Amborella* genome. Coverage of two finished BAC contigs (*10*) by assembled sequence contigs suggests that these two regions were faithfully represented in the assembly (figs. S9 and S10) (*9*), and all 155 of our N90 scaffolds incorporate physically mapped BAC-end sequences.

The accuracy of the genome assembly was further assessed by FISH analysis (*9*). BACs assembled in 104 scaffolds containing 430 Mb (68%) of the genome assembly were cytogenetically localized by FISH to assess scaffold integrity (Fig. 1, fig. S11, and table S8). This analysis confirmed contiguity across major regions (56%) of 66 scaffolds containing 306 Mb (44%) of the genome assembly. Notably, co-assembled BACs that were

[1]Department of Biology, University of Florida, Gainesville, FL 32611, USA. [2]Florida Museum of Natural History, University of Florida, Gainesville, FL 32611, USA. [3]Department of Biology and Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA. [4]Department of Biological Sciences, University at Buffalo (State University of New York), Buffalo, NY 14260, USA. [5]Department of Plant Biology, University of Georgia, Athens, GA 30602, USA. [6]Dow AgroSciences, 9330 Zionsville Road, Indianapolis, IN 46268, USA. [7]Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA 16802, USA. [8]Center for Comparative Genomics and Bioinformatics, The Pennsylvania State University, University Park, PA 16802, USA. [9]Singapore Centre on Environmental Life Sciences Engineering, Nanyang Technological University, 637551 Singapore. [10]Arizona Genomics Institute, University of Arizona, Tucson, AZ 5721, USA. [11]School of Plant Sciences and BIO5 Institute for Collaborative Research, University of Arizona, Tucson, AZ 85721, USA. [12]OpGen, Inc., 708 Quince Orchard Road, Gaithersburg, MD 20878, USA. [13]Genetics Institute, University of Florida, Gainesville, FL 32610, USA.

*These authors contributed equally to this work.
†Corresponding author. E-mail: bbarbazuk@ufl.edu

cytogenetically mapped to different chromosomes indicated potential misassemblies in only two scaffolds (table S8). A karyotyping cocktail differentially labeled all 13 *Amborella* chromosome pairs and anchored major sections of 35 FISH-validated scaffolds to the karyotype (Fig. 2). In total, the cytogenetic cocktail directly placed 101 Mb (58%) of scaffolds with a total length of 176 Mb (~25%) of the assembly onto chromosomes (table S8). However, multiple BACs from 37 scaffolds containing 154 Mb produced inconclusive genome-wide centromeric signals. Sequence alignments associated with the promiscuous probes indicate extensive sequence similarity and the presence of tandem repeats associated with the centromeric regions of the *Amborella* chromosomes.

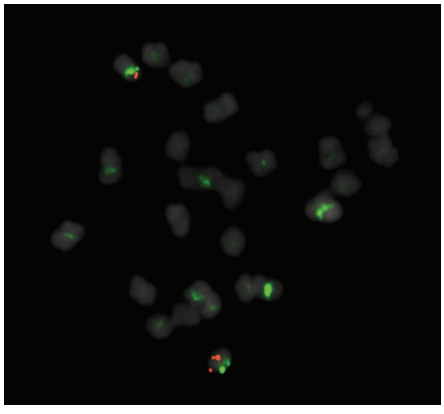
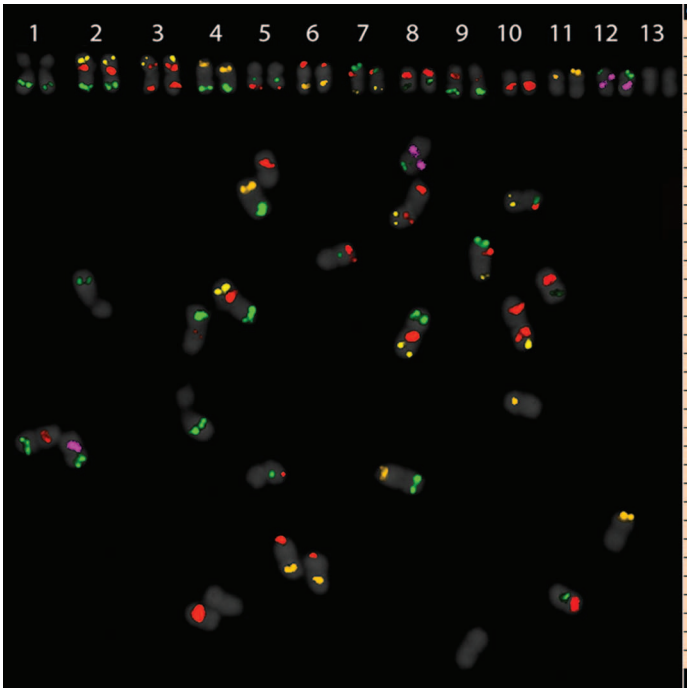


Fig. 1. FISH support of scaffold 7. Two BACs, AT_SBa0003A05 (green) and AT_SBa0003H23 (red), localize 8.2 Mb apart within the assembly scaffold 7 (9.5 Mb). Their colocalized FISH signals unambiguously support the assembly contained between their positional coordinates. Secondary green signals represent repetitive elements in AT_SBa0003A05.

Despite the extensive contiguity of the current draft assembly, gaps remain. Rather than constructing additional PE libraries to improve contiguity, a gap closure method based on whole-genome (formerly optical) mapping technology was undertaken in collaboration with OpGen, Inc. (Gaithersburg, MD, USA). Whole-genome mapping (*14*, *15*) permits assembly of whole-genome restriction endonuclease maps by digesting immobilized DNA molecules and determining the size and order of fragments.

We compared assembled scaffold sequences to single-molecule restriction maps generated with *Amborella* genomic fragments to identify potential joins and produce superscaffolds (*9*) (table S10). This improved our original assembly by a 2× increase in both N50 (4.9 to 9.3 Mb) and N90 (1.2 to 2.9 Mb) (table S5). Thirty joins were confirmed through a new assembly constructed after adding an additional 454 PE sequences and improving data filtering, and 20 joins were confirmed by FISH (*9*) (table S10).

The *Amborella* assembly, as well as several recent plant whole-genome draft sequences (*13*, *16*, *17*), benefited from available collections of BAC-end sequences (*10*) that serve as very long (>150 kb) PE libraries. However, BAC clone resources are expensive and time-consuming to construct and evaluate, as is end-sequencing by low-throughput and high-cost Sanger sequencing. Therefore, as improvement in NGS technologies enables more nonmodel eukaryote whole-genome sequence projects, it is important to identify methods that permit long, accurate assemblies in the absence of large-insert clone resources. Superscaffolding facilitated by Genome-Builder can substitute for BAC-end sequences, as illustrated by our construction of an *Amborella* assembly (*9*) (tables S11 to S13). Although BACs were used as FISH probes in this study, they are not required for cytogenetic validation of an assembly; alternatively, probes could be developed using polymerase chain reaction amplification. Thus, sequencing is no longer a limiting factor, and the greatest challenge for many organisms will be accurate and highly contiguous genome assembly. A combination of FISH and whole-genome mapping, in concert with sequence filtering and assembly strategies described here, should prove successful even for genomes with a more complex repeat structure than that of *Amborella*.



Chromosome	Scaffold	Length (Mb)
1	15	7.41
1	21	6.86
1	25	6.37
1	130	1.61
2	14	7.56
2	22	6.71
2	32	5.9
2	45	5.14
2	57	4.63
2	83	3.07
3	11	8.97
3	12	8.76
3	48	5.16
3	68	3.97
4	10	9.15
5	36	5.61
5	56	4.88
5	109	2.29
5	136	1.54
6	13	8.12
6	66	3.93
7	9	9.39
7	39	5.38
7	67	3.94
7	72	3.67
7	148	1.34
8	23	6.67
8	80	3.25
8	165	0.91
9	29	6.29
9	44	5.11
9	171	0.76
9	183	0.54
12	7	9.5
12	131	1.57
Total	35	175.96

Fig. 2. FISH karyotype for *A. trichopoda*. BAC probes differentially label all chromosome pairs (one pair distinguished by the lack of fluorescent signal) and anchor 35 scaffolds (176 Mb) to the karyotype. Uniquely labeled chromosomes in the cytogenetic preparation (center) are arranged into homologous pairs (upper panel). Chromosomal assignments and sizes of cytogenetically localized scaffolds are tabulated.

References and Notes

1. R. K. Jansen *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 19369–19374 (2007).
2. M. J. Moore, C. D. Bell, P. S. Soltis, D. E. Soltis, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 19363–19368 (2007).
3. D. E. Soltis *et al.*, *Genome Biol.* **9**, 402 (2008).
4. M. Chester, A. R. Leitch, P. S. Soltis, D. E. Soltis, *Genes* **1**, 166–192 (2010).
5. P. Latreille *et al.*, *BMC Genomics* **8**, 321 (2007).
6. N. Nagarajan, T. D. Read, M. Pop, *Bioinformatics* **24**, 1229–1235 (2008).
7. N. D. Young *et al.*, *Nature* **480**, 520–524 (2011).
8. S. Zhou *et al.*, *PLOS Genet.* **5**, e1000711 (2009).
9. Materials and methods are available as supplementary materials on *Science* Online.
10. A. Zuccolo *et al.*, *Genome Biol.* **12**, R48 (2011).
11. I. J. Leitch, L. Hanson, *Bot. J. Linn. Soc.* **140**, 175–179 (2002).
12. B. Star *et al.*, *Nature* **477**, 207–210 (2011).
13. X. Xu *et al.*, *Nature* **475**, 189–195 (2011).
14. C. Aston, B. Mishra, D. C. Schwartz, *Trends Biotechnol.* **17**, 297–302 (1999).
15. D. C. Schwartz *et al.*, *Science* **262**, 110–114 (1993).
16. X. Argout *et al.*, *Nat. Genet.* **43**, 101–108 (2011).
17. A. D'Hont *et al.*, *Nature* **488**, 213–217 (2012).

Acknowledgments: Funded by grant 0922742 from the NSF-PGRP: National Science Foundation Plant Genome Research Program to V.A.A, W.B.B., C.W.D., J.L.M., S.R., D.E.S., and P.S.S. Sequence data are available from the National Center for Biotechnology Information (NCBI) Short Read Archive (SRA) under accession PRJNA212863 and NCBI BioProject ID 212863. Assemblies and additional data are available at http://www.amborella.org; FISH data and probe details are available at http://app.tolkin.org/projects/88. We acknowledge R. Winer (Roche) for technical assistance.

Supplementary Materials

www.sciencemag.org/content/342/6165/1516/suppl/DC1
Materials and Methods
Figs. S1 to S13
Tables S1 to S13
References (*18*–*37*)

28 May 2013; accepted 21 October 2013
10.1126/science.1241130