Using partial genomic fosmid libraries for sequencing complete organellar genomes

Joel R. McNeal[1,2], James H. Leebens-Mack[1], Kathiravetpillai Arumuganathan[3], Jennifer V. Kuehl[4], Jeffrey L. Boore[4,5], and Claude W. dePamphilis[1]

[1]The Pennsylvania State University, University Park, PA, [2]Harvard University, Cambridge, MA, [3]Benaroya Research Institute at Virginia Mason, Seattle, WA, [4]DOE Joint Genome Institute and Lawrence Berkeley National Laboratory, Walnut Creek, CA, and [5]University of California, Berkeley, CA, USA

Organellar genome sequences provide numerous phylogenetic markers and yield insight into organellar function and molecular evolution. These genomes are much smaller in size than their nuclear counterparts; thus, their complete sequencing is much less expensive than total nuclear genome sequencing, making broader phylogenetic sampling feasible. However, for some organisms, it is challenging to isolate plastid DNA for sequencing using standard methods. To overcome these difficulties, we constructed partial genomic libraries from total DNA preparations of two heterotrophic and two autotrophic angiosperm species using fosmid vectors. We then used macroarray screening to isolate clones containing large fragments of plastid DNA. A minimum tiling path of clones comprising the entire genome sequence of each plastid was selected, and these clones were shotgun-sequenced and assembled into complete genomes. Although this method worked well for both heterotrophic and autotrophic plants, nuclear genome size had a dramatic effect on the proportion of screened clones containing plastid DNA and, consequently, the overall number of clones that must be screened to ensure full plastid genome coverage. This technique makes it possible to determine complete plastid genome sequences for organisms that defy other available organellar genome sequencing methods, especially those for which limited amounts of tissue are available.

INTRODUCTION

Unlike eukaryotic nuclear genomes, organellar genomes occur in high copy number per cell and are of a size more amenable for complete sequencing. Gene orthology is typically clear even across a wide taxonomic range; thus, organellar genes provide a disproportionately large fraction of genes currently used for phylogeny (1). Furthermore, comparisons of organellar genomes can provide insights into the evolutionary transformations from cyanobacteria and proteobacteria into plastids and mitochondria, the functions of these organelles, and the patterns of co-evolution that have occurred with the many nuclear genes whose products function inside of these organelles.

The earliest organelle genome sequences were generated by digesting, cloning, and mapping purified organellar DNA, followed by sequencing small fragments individually from the clone bank (2). With the advent of cost-effective, high-throughput sequencing, genome sequences are being generated more efficiently by shotgun cloning directly from organellar DNA isolations, performing a single sequencing read from each end of a large number of randomly selected clones, then assembling these into a complete genome sequence computationally. There are several possibilities for preparing a template that is acceptable for this process and, for some taxa, these have become simple and reliable protocols (megasun.bch.umontreal.ca/People/lang/FMGP/methods/mtDNA.html). Intact organelles can be isolated, most often by sucrose or Percoll gradient centrifugations (3), and in some cases, the differences in base composition and topology (i.e., circular versus linear DNA) between organellar and nuclear DNAs can be exploited using bisbenzimide or cesium chloride gradients to isolate organellar DNA for sequencing

(4). Large quantities of fresh tissue are typically necessary to produce small amounts of organellar DNA (although it is often possible to amplify these small amounts using rolling circle amplification or RCA). Even after enrichment, the low proportion of organellar to nuclear DNA can lead to significant nuclear contamination (>50% of the total DNA) in many species, including those with large nuclear genomes or interfering polyphenolics. Another method is to amplify large sections of organellar DNA by long-PCR between regions for which primers exist, which has been used effectively for many animal mitochondrial DNAs (mtDNAs) and occasionally for plastid DNAs (ptDNAs) as well (5). Jansen et al. (3) review current land plant ptDNA isolation and sequencing methods.

Although these procedures have succeeded for a variety of plastid genomes, many organisms exist for which they are not feasible. It is difficult or impossible to produce significantly enriched organellar DNA from many plants, even with large quantities of fresh tissue. The PCR method (5) eliminates the need for enriched ptDNA, but is only practical if the genome is not highly rearranged or if gene order is known via prior mapping. A set of PCR primers spaced around the entire plastid genome is necessary, and amplification-induced artifacts may occur. Heterotrophic plants often exhibit both rapid sequence divergence and unusual plastid ultrastructure that make these procedures infeasible; accordingly, the complete sequence of only one heterotrophic angiosperm has been published (6). The method we present enables plastid genome sequencing from both parasitic and nonparasitic plants using small amounts of fresh, frozen, or desiccated tissue and should be equally applicable for sequencing mitochondrial genomes.

MATERIALS AND METHODS

### DNA Isolation and Partial Genomic Library Construction

Fresh material from *Cuscuta exaltata*, *Cuscuta obtusiflora* (parasitic), and *Ipomoea purpurea*

(autotrophic) was grown from seed. Tissue from *Yucca schidigera* (autotrophic) was collected and snap-frozen in liquid nitrogen. Nuclear genome sizes of all species were determined by flow cytometry following the protocol described previously (7). One gram tissue from each plant was pulverized to powder by mortar and pestle after being frozen in liquid nitrogen for 20 s. DNA was extracted in 10 mL buffer using a 2× cetyltrimethylammonium bromide (CTAB) procedure (8) with 1% polyethylene glycol (PEG) 8000 in the buffer. After isopropanol precipitation, DNA was spooled out, rinsed with 70% ethanol, and resuspended in 500 µL water. To clean and concentrate the DNA, it was reprecipitated by adding 125 µL 4 M NaCl plus 625 µL 13% PEG 8000 and incubated on ice for 20 min before centrifugation at 13,000× *g* at 4°C for 15 min. DNA pellets were resuspended in 75 µL water. DNA fragments ranging from 40 to 45 kb were excised from a 0.8% agarose gel using field inversion gel electrophoresis (FIGE).

The CopyControl™ Fosmid Library Production kit (Epicentre Biotechnologies, Madison, WI, USA) was used to construct partial genomic DNA libraries. Concentration of size-selected, end-repaired DNA was determined using Molecular Probes™ PicoGreen® (Invitrogen, Carlsbad, CA, USA) dye and flourimetry. Appropriate quantities of DNA were ligated and packaged according to the manufacturer's protocol.

Identifying Plastid Clones

Fosmid clones were plated as infected *Escherichia coli* on LB-agar plus 12.5 µg/mL chloramphenicol. A QPix2™ robot (Genetix, Boston, MA, USA) was used to organize clones into 384-well plates and to grid colonies onto nylon membranes (Q-Performa™; Genetix) soaked in LB broth plus 12.5 µg/mL chloramphenicol. Gridding patterns that allowed rapid identification of specific clones after hybridization were used (Figure 1), and each clone was replicated at least six times per filter. Colonies were grown on the filters for 16 h. Afterwards, filters were allowed to soak up denaturing
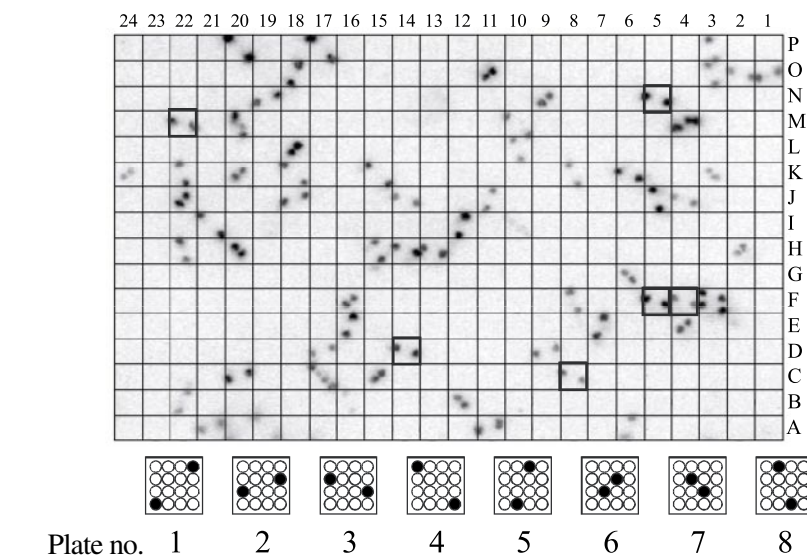
Figure 1. Macroarray screen of fosmid clones using pooled plastid probes. Eight plates, each containing 384 clone cultures from a partial genomic fosmid library of *Cuscuta obtusiflora*, were spotted onto the filter in a known pattern. Squares on the grid are labeled along the outer edge corresponding to the 384 wells of the plates. Each grid square contains clones corresponding to that well from all eight plates, and each clone is replicated twice within the square in a particular pattern unique to each of the eight plates (shown below the grid). In total, 6144 spots representing 3072 unique clones were screened in this particular image, of which approximately 66 positively hybridized to the plastid probes. Six clones from plate 3 (wells C8, D14, F4, F5, and N5, shown with bold borders) were randomly chosen for end-sequencing and internal PCR testing to determine what portion of the plastid genome they contained.

solution (0.5 M NaOH, 1.5 M NaCl) from saturated blotter paper for 4 min. This process was repeated with fresh denaturing solution using bottom-heat from a glass plate placed over a boiling water bath. The filters were then placed on blotter paper soaked in 1.5 M NaCl, 1 M Tris solution for 4 min at room temperature and dried for 10 min. Colonies were immersed in a proteinase K solution (0.1 M NaCl, 50 mM Tris, 50 mM EDTA, 1× Sarkosyl®, 100 mg/L proteinase K) for 50 min at 37°C, dried, baked for 2 h at 80°C, and cross-linked under UV light for 2 min.

PCR products ranging from 200 to 700 nucleotides were generated from the plastid genes *rps2*, *rps4*, *rpl16*, *rps7*, *rbcL*, and *psaC* for all species; *psbA* and a PCR product from *psbE* to *psbJ* were also amplified for *Yucca*. These products were pooled at equal molar concentration, diluted to approximately 5 ng/µL, and radioactively labeled with [∝-^{32}P]dATP according to the Strip-EZ® DNA protocol (Ambion, Austin, TX, USA). Excess radionucleotide was removed with Centri-Spin™ columns (Princeton Separations, Adelphia, NJ, USA).

Filters were prehybridized in 5× NaCl/NaH$_2$PO$_4$/EDTA (SSPE), 5× Denhardt's solution (9), 0.5% sodium dodecyl sulfate (SDS), and 0.1 mg/mL fragmented salmon sperm DNA for 1 h at 68°C. Radioactive probes were diluted to 250 µL in 10 mM EDTA, denatured at 90°C for 10 min, and hybridized to the filters at 68°C overnight. Filters were first washed in 2× SSPE and 0.5% SDS at room temperature, followed by a wash in 2× SSPE/0.5% SDS, a wash of 0.3× SSPE/0.5% SDS, a wash in 2× SSPE/0.5% SDS at 55°C, and a wash of 0.3× SSPE at room temperature. Wash durations were 15 min. The filters were enclosed in plastic wrap and exposed on phosphorimaging screens overnight. Screen images were captured, and plastid clones were identified by positive hybridizations.

Selecting Clones for Sequencing

Randomly selected positive clones were grown for 15 h in 5 mL of Terrific Broth plus 12.5 µg/mL chloramphenicol. This culture (0.5 mL) was added to 4.5 mL LB broth plus 12.5 µg/mL chloram-
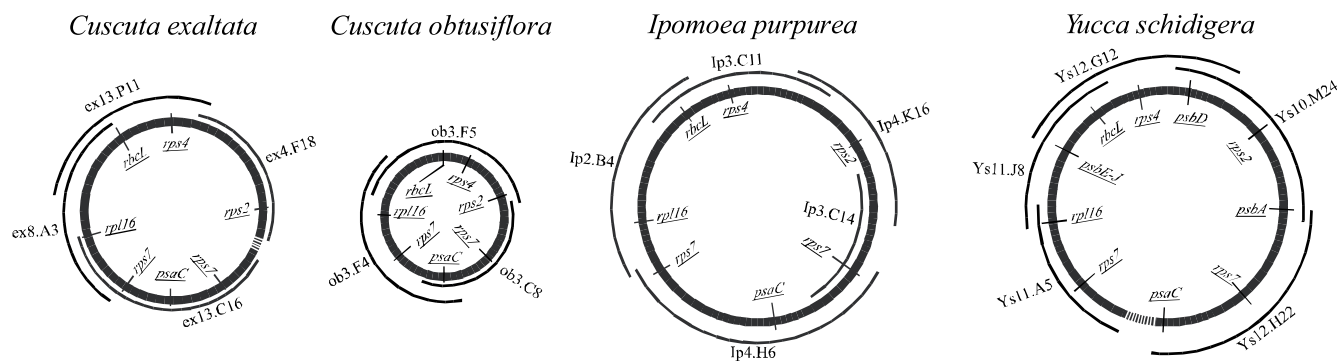
Figure 2. End-sequenced clone coverage on plastid genomes. Both ends of selected clones were sequenced to determine relative coverage of the plastid genome. Sequence strand directionality and internal PCR assays for a variety of plastid genes were also used to identify any genome rearrangements that may have occurred and could possibly confuse mapping. Minimal subsets of clones necessary for optimum coverage were used for shotgun sequencing and are shown as solid arcs. Relative locations of the gene probes used for hybridization are marked on the circular genome map, with underlined gene labels for each probe inside the circles. Genome maps are drawn to scale relative to one another.

phenicol and induced to high plasmid copy number following the CopyControl protocol. Minipreps were performed using mini alkaline-lysis (9) followed by precipitation with one-fourth volume 4 M NaCl and equal volume PEG 8000 at 4°C for 20 min. Pellets were resuspended in 20 µL water, and DNA concentrations were determined on an Eppendorf® Biophotometer™ (Eppendorf, Westbury, NY, USA).

T7 forward primer and pCC1/pEpiFOS reverse primer (sequence in CopyControl protocol) were used to sequence the ends of each fosmid insert on a CEQ™ 8000 Genetic Analysis System (Beckman Coulter, Fullerton, CA, USA). DNA template (2.5 µg) and 5 µmol primer were used, with other parameters following those provided by Beckman Coulter for bacterial artificial chromosome (BAC) end sequencing. Sequences were used in BLASTN (10) searches to verify the position of fosmid inserts within plastid genomes. Directionality of the end sequences was checked relative to the plastid genome of *Nicotiana tabacum* (2) to identify major genomic inversions. PCR tests were conducted with the genes used as probes to confirm that the clones spanned the regions indicated by end sequences. A minimally overlapping set of clones covering the plastid genome was chosen for each species. Those fosmid clone preparations were sheared by repeatedly driving the DNA through a narrow aperture using a HydroShear™ device (GeneMachines, San Carlos, CA, USA). After enzymatic end repair, gel purification of fragments

approximately 3 kb, and cloning into pUC18, 384 clones were picked from each fosmid preparation. These clones were robotically processed through RCA and sequenced from each end (3). Vector sequences were screened out and reads were assembled into complete circular maps. Detailed protocols are available at the Joint Genome Institute (JGI) web site www.jgi.doe.gov. Two gaps in coverage of <4 and 6 kb for *C. exaltata* and *Y. schidigera*, respectively, were PCR-amplified and sequenced on the CEQ 8000 Genetic Analysis System following standard manufacturer's procedures rather than sequencing additional clones.

RESULTS AND DISCUSSION

This method successfully produced plastid genome sequences for all species. Five fosmid clones were necessary for coverage of *Ipomoea* and *Yucca*, four for *C. exaltata*, and three for *C. obtusiflora*. Average fosmid insert size was 38 kb (ranging from 32 to 47 kb), and clone locations are shown in Figure 2. The full plastid genome inverted repeat (IR) was only sequenced once in *C. exaltata*, and no polymorphisms between the two IRs were detected in the other species.

Drastic differences in percentage of positively hybridizing clones were observed across species (Table 1). This percentage is expected to be proportional to the amount of ptDNA relative to other DNA (nuclear plus mitochondrial) in the tissue, assuming DNA

from all compartments shears equally during the isolation process. Base composition was similar for all species examined (37.4%–38.1% GC) and did not significantly impair fosmid cloning, but could affect cloning efficiency in other extreme cases. A number of other factors, including nuclear genome size, amount of mtDNA, number of plastids per cell, and number of ptDNAs per plastid, could affect this ratio. Tissue age may also influence relative abundance of ptDNA (11). Estimates of nuclear genome size for *Ipomoea* and *C. obtusiflora* were similar, yet the percentage of plastid clones in *Ipomoea* was over three times higher than in *C. obtusiflora*. However, because the plastid genome size of *C. obtusiflora* is only about half of that in *Ipomoea*, the observed results deviate only slightly from the number of plastid clones expected if ptDNAs of both species were in equal copy number per cell. Although *C. exaltata* is more chlorophyllous than *C. obtusiflora*, over 10 times as many clones positively hybridized for *C. obtusiflora* (Table 1). Nuclear DNA content of *C. exaltata* was estimated to be over 25 times that of *C. obtusiflora*, indicating that nuclear genome size is more crucial in determining percentage of plastid clones than tissue type or photosynthetic ability.

Although this method worked well for these plants, there are some caveats. Ability to detect small organellar genomes is limited by the minimum insert size of the library. Small plastid genomes probably occur in concat-

Short Technical Reports

Table 1. Number of Clones Screened and Identified for Each Species

Species	Clones Screened (no.)	Positives (no.)	Positives (%)	Estimated pg/2 C Nuclear Genome Size (SD)
Ipomoea purpurea	1536	120	7.81	1.51 (0.020)
Cuscuta exaltata	6144	10	0.16	41.86 (0.559)
Cuscuta obtusiflora	6144	140	2.28	1.58 (0.022)
Yucca schidigera	4608	56	1.21	6.45 (0.050)

enated forms that would be clonable by this method (12), but any organellar genomes existing as fragments <40 kb would not be included in a fosmid library and would require building libraries with smaller insert sizes. This method also requires plastid probes <80 kb apart that can be hybridized against the library. Genomes for which insufficient PCR primers exist could be heterologously probed with sequences from related taxa using less stringent hybridization conditions. Once one plastid clone is identified, its end sequences can be used to reprobe the library and reveal adjacent clones in both directions. Highly rearranged genomes could confound identifying a proper set of plastid clones. Although interpretation is complicated by the presence of a fosmid vector ligated to insert DNA, restriction mapping of clones could be used to confirm complete genome coverage. However, end sequencing and an increased number of internal PCR tests on each clone should nearly always suffice.

A final caveat is the possibility of false positive hybridizations from laterally transferred ptDNA to either the mitochondrial or nuclear genome. Although lateral transfer of ptDNA to the nucleus occurs at high frequency (13,14), such transfers are typically much smaller in size than a 40-kb fosmid insert (15), and any transfer to the nuclear genome that exists in low copy is less likely to be detected than true plastid clones. Transfer of ptDNA to the mitochondrial genome is much more detectable because, like the plastid genome, it exists in high copy number per cell (16). We detected two clones with inserts suspected to be of mitochondrial origin. End sequences of a strongly hybridizing clone for *Ipomoea* gave BLASTN results similar

to regions of the *Beta vulgaris* mitochondrial genome (GenBank® accession no. NC 002511). One *C. exaltata* clone possessed plastid sequences as best Basic Local Alignment Search Tool (BLAST) hits on both ends, and PCR tests showed it contained all expected plastid probes. However, most genes in this clone were obvious pseudogenes with early stop codons or large truncations. Some pseudogenes were present in multiple copies, and many internal rearrangements existed, although pseudogene sequences were not extremely diverged from true plastid sequences. Rapid structural change but slow mutation rates are characteristic of plant mitochondrial genomes (16), indicating this clone was probably a large fragment of ptDNA transferred to the mitochondrial genome, where it has become nonfunctional. Transfers of genetic material from the plastid to the mitochondrion this large are not found in currently sequenced angiosperm mitochondrial genomes, but large intergenomic transfers are not unexpected given that in one ecotype of *Arabidopsis thaliana*, a nearly full copy of the mitochondrial genome is present on a nuclear chromosome (17).

Despite these caveats, this method is an effective way of obtaining complete plastid genomes from as little as 1 g of tissue, even from plants for which extracting purified ptDNA is impossible or that have extensive genome rearrangements. Small quantities of frozen or silica gel dried plant material generally produce sufficient DNA quantity with high molecular weight fragments falling within the size range necessary for fosmid cloning. Even though the fosmid vector is proportionally 15%–20% of the DNA that is shotgun-sequenced, practically no finishing sequencing was necessary

for the plastid genomes generated with this method, whereas other land plant ptDNA shotgunsequencing methods rarely approach 80% efficiency (3).

Although we used plant plastid genomes as an example, this method could easily be extended to large mitochondrial genomes. For both mitochondrial and plastid genomes, BAC libraries could be used instead of fosmid libraries, assuming the insert sizes were less than the overall size of the in vivo organellar genome fragments. It would take fewer BAC clones than fosmid clones to cover an organellar genome, but BAC libraries are more difficult to generate and require sizable amounts of fresh material for DNA extraction (18). Finally, this method could be used to separate organellar DNA of organisms in close association, such as endophytes and endosymbiotic organisms and their hosts. As long as species-specific probes could be generated, organellar genomes could be readily attainable without contamination.

COMPETING INTERESTS STATEMENT

The authors declare no competing interests.

REFERENCES

1. **Savolainen, V., M.W. Chase, N. Salamin, D.E. Soltis, P.S. Soltis, A.J. Lopez, O. Fedrigo, and G.J.P. Naylor.** 2002. Phylogeny reconstruction and functional constraints in organellar genomes: plastid atpB and rbcL sequences versus animal mitochondrion. Syst. Biol. *51*:638-647.

2. **Shinozaki, K., M. Ohme, M. Tanaka, T. Wakasugi, N. Hayashida, T. Matsubayashi, N. Zaita, J. Chunwongse, et al.** 1986. The complete nucleotide sequence of the Tobacco chloroplast genome—its gene organization and expression. EMBO J. *5*:2043-2049.

3. **Jansen, R.K., L.A. Raubeson, J.L. Boore, C.W. dePamphilis, T.W. Chumley, R.C. Haberle, S.K. Wyman, A.J. Alverson, et al.** 2005. Methods for obtaining and analyzing whole chloroplast genome sequences. Methods Enzymol. *395*:348-384.

4. **Turmel, M., C. Lemieux, G. Burger, B.F. Lang, C. Otis, I. Plante, and M.W. Gray.** 1999. The complete mitochondrial DNA sequences of *Nephroselmis olivacea* and *Pedinomonas minor*: two radically different evolutionary patterns within green algae. Plant Cell *11*:1717-1729.

5. **Goremykin, V.V., K.I. Hirsch-Ernst, S. Wolfl, and F.H. Hellwig.** 2003. Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. Mol. Biol. Evol. *20*:1499-1505.

6. **Wolfe, K.H., C.W. Morden, and J.D. Palmer.** 1992. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. Proc. Natl. Acad. Sci. USA *89*:10648-10652.

7. **Arumuganathan, K. and E.D. Earle.** 1991. Estimation of nuclear DNA contents of plants by flow cytometry. Plant Mol. Biol. Rep. *9*:229-241.

8. **Doyle, J.J. and J.L. Doyle.** 1990. Isolation of plant DNA from fresh tissue. Focus *12*:13-15.

9. **Sambrook, J., E.F. Fritsch, and T. Maniatis.** 1989. Molecular Cloning: A Laboratory Manual. CSH Laboratory Press, Cold Springs Harbor, NY.

10. **Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. *215*:403-410.

11. **Rowan, B.A., D.J. Oldenburg, and A.J. Bendich.** 2004. The demise of chloroplast DNA in *Arabidopsis*. Curr. Genet. *46*:176-181.

12. **Bendich, A.J.** 2004. Circular chloroplast chromosomes: the grand illusion. Plant Cell *16*:1661-1666.

13. **Huang, C.Y., M.A. Ayliffe, and J.N. Timmis.** 2003. Direct measurement of the transfer rate of chloroplast DNA into the nucleus. Nature *422*:72-76.

14. **Stegemann, S., S. Hartmann, S. Ruf, and R. Bock.** 2003. High-frequency gene transfer from the chloroplast genome to the nucleus. Proc. Natl. Acad. Sci. USA *100*:8828-8833.

15. **Huang, C.Y., M.A. Ayliffe, and J.N. Timmis.** 2004. Simple and complex nuclear loci created by newly transferred chloroplast DNA in tobacco. Proc. Natl. Acad. Sci. USA *101*:9710-9715.

16. **Palmer, J.D. and L.A. Herbon.** 1988. Plant mitochondrial-DNA evolves rapidly in structure, but slowly in sequence. J. Mol. Evol. *28*:87-97.

17. **Lin, X., S.S. Kaul, S. Rounsley, T.P. Shea, M.I. Benito, C.D. Town, C.Y. Fujii, T. Mason, et al.** 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature *402*:761-768.

18. **Chalhoub, B., H. Belcram, and M. Caboche.** 2004. Efficient cloning of plant genomes into bacterial artificial chromosome (BAC) libraries with larger and more uniform insert size. Plant Biotechnol. J. *2*:181-188.

Address correspondence to Joel R. McNeal, Arnold Arboretum of Harvard University, 22 Divinity Avenue, Harvard University, Cambridge, MA 02138, USA. e-mail: jmcneal@oeb.harvard.edu