BioOne COMPLETE

A Targeted Enrichment Strategy for Massively Parallel Sequencing of Angiosperm Plastid Genomes

Authors: Stull, Gregory W., Moore, Michael J., Mandala, Venkata S., Douglas, Norman A., Kates, Heather-Rose, et al.

Source: Applications in Plant Sciences, 1(2)

Published By: Botanical Society of America

URL: https://doi.org/10.3732/apps.1200497

BioOne Complete (complete.BioOne.org) is a full-text database of 200 subscribed and open-access titles in the biological, ecological, and environmental sciences published by nonprofit societies, associations, museums, institutions, and presses.

Your use of this PDF, the BioOne Complete website, and all posted and associated content indicates your acceptance of BioOne's Terms of Use, available at www.bioone.org/terms-of-use.

Usage of BioOne Complete content is strictly limited to personal, educational, and non - commercial use. Commercial inquiries or rights and permissions requests should be directed to the individual publisher as copyright holder.

BioOne sees sustainable scholarly publishing as an inherently collaborative enterprise connecting authors, nonprofit publishers, academic institutions, research libraries, and research funders in the common goal of maximizing access to critical research.

Downloaded From: https://bioone.org/journals/Applications-in-Plant-Sciences on 28 Mar 2024
Terms of Use: https://bioone.org/terms-of-use

Applications in Plant Sciences

Applications in Plant Sciences 2013 1(2): 1200497

Protocol Note

A targeted enrichment strategy for massively parallel sequencing of angiosperm plastid genomes[1]

Gregory W. Stull[2,3,8], Michael J. Moore[4], Venkata S. Mandala[4], Norman A. Douglas[4], Heather-Rose Kates[3,5], Xinshuai Qi[6], Samuel F. Brockington[7], Pamela S. Soltis[3,5], Douglas E. Soltis[2,3,5], and Matthew A. Gitzendanner[2,3,5]

[2]Department of Biology, University of Florida, Gainesville, Florida 32611-8525 USA; [3]Florida Museum of Natural History, University of Florida, Gainesville, Florida 32611-7800 USA; [4]Department of Biology, Oberlin College, Oberlin, Ohio 44074-1097 USA; [5]Genetics Institute, University of Florida, Gainesville, Florida 32611 USA; [6]Key Laboratory of Conservation Biology for Endangered Wildlife of the Ministry of Education, and Laboratory of Systematic and Evolutionary Botany and Biodiversity, College of Life Sciences, Zhejiang University, Hangzhou 310058 People's Republic of China; and [7]Department of Plant Sciences, University of Cambridge, Cambridge CB13EA United Kingdom

- *Premise of the study:* We explored a targeted enrichment strategy to facilitate rapid and low-cost next-generation sequencing (NGS) of numerous complete plastid genomes from across the phylogenetic breadth of angiosperms.
- *Methods and Results:* A custom RNA probe set including the complete sequences of 22 previously sequenced eudicot plastomes was designed to facilitate hybridization-based targeted enrichment of eudicot plastid genomes. Using this probe set and an Agilent SureSelect targeted enrichment kit, we conducted an enrichment experiment including 24 angiosperms (22 eudicots, two monocots), which were subsequently sequenced on a single lane of the Illumina GAIIx with single-end, 100-bp reads. This approach yielded nearly complete to complete plastid genomes with exceptionally high coverage (mean coverage: 717×), even for the two monocots.
- *Conclusions:* Our enrichment experiment was highly successful even though many aspects of the capture process employed were suboptimal. Hence, significant improvements to this methodology are feasible. With this general approach and probe set, it should be possible to sequence more than 300 essentially complete plastid genomes in a single Illumina GAIIx lane (achieving ~50× mean coverage). However, given the complications of pooling numerous samples for multiplex sequencing and the limited number of barcodes (e.g., 96) available in commercial kits, we recommend 96 samples as a current practical maximum for multiplex plastome sequencing. This high-throughput approach should facilitate large-scale plastid genome sequencing at any level of phylogenetic diversity in angiosperms.

Key words: next-generation sequencing; phylogenomics; plastid genomes.

Over the past few years, complete plastid genome sequencing has emerged as a powerful and increasingly accessible tool for plant phylogenetics, facilitated by rapid advances in next-generation sequencing (NGS) technologies (e.g., Moore et al., 2006, 2007, 2010; Jansen et al., 2007; Cronn et al., 2008, 2012). Many aspects of the plastid genome, including its structural simplicity, relatively small size, and highly conserved gene content, make it ideally suited for next-generation sequencing and assembly. Additionally, its wealth of characters, useful across many taxonomic levels, makes it an excellent resource for phylogenetic studies across the plant branch of the tree of life. Plastome-scale phylogenetic studies have, for example, clarified relationships among major angiosperm lineages (Moore et al., 2007, 2010; Jansen et al., 2007) and resolved recent, rapid radiations in *Pinus* (Parks et al., 2009). Plastid genomes also have great potential for population genetic and phylogeographic studies (e.g., Whittall et al., 2010), particularly as a complement to multiple unlinked nuclear loci, although this application of large-scale plastid data sets has been underexplored compared to deeper-level phylogenetic studies.

The ever-increasing capacities of next-generation sequencers, particularly the Illumina platforms, coupled with the high-copy nature of the plastid genome, have made it possible to multiplex numerous samples of whole-genomic DNA (gDNA) on a single lane and still recover sufficient coverage to assemble complete or nearly complete plastid genomes (e.g., Cronn et al., 2008, 2012; Steele et al., 2012; Straub et al., 2012). However, given that plastid DNA typically constitutes only ~0.5–13% of gDNA samples (Steele et al., 2012; Straub et al., 2012), this approach expends much of the sequencing capacity on nuclear reads, significantly reducing the number of plastomes that can be sequenced in parallel. Consequently, this limits the scalability

[1]Manuscript received 17 September 2012; revision accepted 9 December 2012.

The authors thank K. Kupsch for sharing material of *Monococcus echinophorus*, M. Heaney for providing DNA of *Nolina brittoniana*, M. Croley and R. Mostow for assistance with library construction, Andy Crowl and Nico Cellinese for proving the *Campanula* sample, and R. Cronn for helpful advice. We also thank K. E. Holsinger and two anonymous reviewers for their helpful comments on the manuscript. This research was funded by National Science Foundation grants DBI-0735191 and EF-0431266 and the Oberlin College Office of Sponsored Programs.

[8]Author for correspondence: gwstull@ufl.edu

doi:10.3732/apps.1200497

Downloaded From: https://bioone.org/journals/Applications-in-Plant-Sciences on 28 Mar 2024
Terms of Use: https://bioone.org/terms-of-use

of plastid genome sequencing for large-scale phylogenetic and phylogeographic studies when funding is limited.

By increasing the abundance of plastid DNA relative to the nuclear and mitochondrial genomes, targeted enrichment strategies for the plastid genome offer a promising means of vastly increasing the number of plastomes that can be multiplexed on a single lane. Some researchers have used long-range PCR to amplify segments of the plastid genome as one enrichment strategy (e.g., Cronn et al., 2008; Njuguna et al., 2013). However, such methods are more time-intensive and require appropriate primer design as well as high-quality DNA to ensure amplification of the long segments. Another method of enriching for plastids is through sucrose gradient centrifugation during DNA extraction (e.g., Moore et al., 2006), but this requires large amounts (frequently >5 g) of fresh tissue. In contrast, hybridization-based methods of plastid enrichment, which use oligonucleotide probes (or "baits") to capture plastid targets, show considerable potential for broad applicability given their ability to enrich degraded samples (e.g., DNA from herbarium material) and their utility across large phylogenetic distances (when the probe design incorporates sequences from phylogenetically diverse samples) (e.g., Cronn et al., 2012). However, these plastid capture methods, while promising, have until now only been developed for *Pinus* (Cronn et al., 2012; Parks et al., 2012). Designing a plastid probe set of broad phylogenetic applicability has not been attempted.

Several commercial kits have been developed for hybridization-based targeted enrichment using custom probe sets (e.g., Agilent SureSelect, Roche Nimblegen, MYcroarray), and the offerings are rapidly changing. Here we present a hybridization-based method for targeted enrichment of angiosperm plastid genomes, using a custom set of RNA probes designed from 22 previously sequenced eudicot plastomes (see Table 1) and an early version of the Agilent SureSelect technology. We demonstrate the utility of this probe-based approach with results from an enrichment experiment that involved 24 angiosperms (22 species of eudicots and two species of monocots) multiplexed on a single lane of the Illumina GAIIx (Illumina Inc., San Diego, California, USA) subsequent to enrichment. The success of this experiment illustrates the utility of the capture method in general and the broad applicability of the probe set in particular. This capture method, or improvements thereto, will enable a significant increase in the number of angiosperm plastid genomes that can be multiplexed on the Illumina platform. This, in turn, will dramatically decrease per-genome sequencing costs, making large-scale sequencing of plastid genomes a feasible option for any phylogenetic or phylogeographic study. Furthermore, the broad phylogenetic utility of the probe set employed here makes this method applicable for plastome-based evolutionary studies across not only eudicots, but also monocots and potentially all angiosperms.

METHODS AND RESULTS

***Probe design*—**RNA probes ("baits") were designed by Genotypic Technology Ltd. (Bangalore, India) from the complete plastid genomes of 22 eudicot species, selected to represent much of the phylogenetic breadth of eudicots (Table 1). We chose to limit bait design to eudicots to maximize the utility of the bait array for plastid phylogenomics throughout this clade, which includes approximately 75% of angiosperm diversity (Drinnan et al., 1994; Soltis et al., 2005) and has been the subject of ongoing research in our laboratories (e.g., Jian et al., 2008; Wang et al., 2009; Brockington et al., 2009; Moore et al., 2010; Arakaki et al., 2011). For each input genome, 120-bp baits were designed, with 50-bp overlap (~2× tiling). To minimize representational bias of highly conserved regions of the plastid genome (e.g., rRNA genes) during hybridization capture, bait sequences for all genomes were compared using BLAST, and only baits with <90% identity to all other baits were retained in the final bait design. In all, ~55 000 baits were included in the final design. The bait sequences and coordinates are available in Appendix S1.

***Sampling*—**To test the efficacy of the bait array for plastome capture, we constructed Illumina libraries for 24 species (Table 2), representing 22 eudicots and two monocots. The 22 eudicots span the phylogenetic diversity of the clade, including species from *Rosidae*, *Asteridae*, and *Caryophyllales* (sensu Cantino et al., 2007). These species were also selected to test the effects on plastome capture of increasing phylogenetic distance from the sequences included in the bait design. For example, we constructed libraries for one species that was part of the bait design (*Cucumis sativus*), one species (*Oenothera hartwegii*) that is

Table 1. Eudicot plastomes used for probe design.

Taxon	Family (Order)	GenBank accession no.
Antirrhinum majus L.	Plantaginaceae (Lamiales)	Unpublished data (M. J. Moore)
Arabidopsis thaliana (L.) Heynh.	Brassicaceae (Brassicales)	NC_000932
Citrus sinensis (L.) Osbeck	Rutaceae (Sapindales)	NC_008334
Cornus florida L.	Cornaceae (Cornales)	Unpublished data (M. J. Moore)
Cucumis sativus L.	Cucurbitaceae (Cucurbitales)	NC_007144
Dillenia indica L.	Dilleniaceae (Dilleniales)	Unpublished data (M. J. Moore)
Ficus sp.	Moraceae (Rosales)	Unpublished data (M. J. Moore)
Gossypium hirsutum L.	Malvaceae (Malvales)	NC_007944
Helianthus annuus L.	Asteraceae (Asterales)	NC_007977
Ilex cornuta Lindl. & Paxton	Aquifoliaceae (Aquifoliales)	Unpublished data (M. J. Moore)
Liquidambar styraciflua L.	Altingiaceae (Saxifragales)	Unpublished data (M. J. Moore)
Lonicera japonica Thunb.	Caprifoliaceae (Dipsacales)	Unpublished data (M. J. Moore)
Nandina domestica Thunb.	Berberidaceae (Ranunculales)	NC_008336
Nerium oleander L.	Apocynaceae (Gentianales)	Unpublished data (M. J. Moore)
Oenothera biennis L.	Onagraceae (Myrtales)	NC_010361
Oxalis latifolia Kunth	Oxalidaceae (Oxalidales)	Unpublished data (M. J. Moore)
Platanus occidentalis L.	Platanaceae (Proteales)	NC_008335
Plumbago auriculata Lam.	Plumbaginaceae (Caryophyllales)	Unpublished data (M. J. Moore)
Populus trichocarpa Torr. & A. Gray	Salicaceae (Malpighiales)	NC_009143
Spinacia oleracea L.	Amaranthaceae (Caryophyllales)	NC_002202
Staphylea colchica Steven	Staphyleaceae (Crossosomatales)	Unpublished data (M. J. Moore)
Ximenia americana L.	Olacaceae (Santalales)	Unpublished data (M. J. Moore)

Downloaded From: https://bioone.org/journals/Applications-in-Plant-Sciences on 28 Mar 2024
Terms of Use: https://bioone.org/terms-of-use

TABLE 2. Eudicot and monocot species included in this study, with voucher information and assembly statistics.

Taxon	Family (Order)	Voucher (Herbarium)	No. of plastid reads/ total reads	% Plastid reads	% Plastid reads (unenriched)*	% Plastome recovered	Mean coverage
Acleisanthes lanceolata (Wooton) R. A. Levin	Nyctaginaceae (Caryophyllales)	R. Merkel 8 (OC)	1 478 311/2 001 153	73.9	17.7	99.83	1091
Campanula erinus L.	Campanulaceae (Asterales)	A. Crowl 42 (FLAS)	292 895/833 412	35	4.1	95.8	176
Cucumis sativus L.	Cucurbitaceae (Cucurbitales)	cv. 'Calypso' (Seminis Vegetable Seeds)	2 131 764/2 495 346	85.4	N/A	100	1408
Dicranocarpus parviflorus A. Gray	Asteraceae (Asterales)	M. Moore 655 (OC)	1 660 972/2 408 461	69	14.5	99.99	1250
Frankenia L. sp.	Frankeniaceae (Caryophyllales)	S. F. Brockington (s.n.)	2 658 678/3 839 653	69	N/A	84.5	2088
Glinus dahomensis (Fenzl) A. Chev.	Molluginaceae (Caryophyllales)	S. F. Brockington (cultivated from seed, s.n.)	152 181/413 956	36.8	N/A	93.3	87
Limeum L. sp.	Limeaceae (Caryophyllales)	S. F. Brockington (cultivated from seed, s.n.)	2 402 594/3 316 313	72.4	N/A	98.7	1515
Limonium limbatum Small	Plumbaginaceae (Caryophyllales)	M. Moore 694 (OC)	47 113/81 348	58	N/A	97.7	32.4
Mentzelia perennis Wooton	Loasaceae (Cornales)	M. Moore 917 (OC)	654 939/767 318	85.4	14.3	100	467
Microtea debilis Sw.	Phytolaccaceae (Caryophyllales)	M. Rimachi 11128 (TEX/LL)	580 514/1 146 486	50.6	N/A	95	375
Monococcus echinophorus F. Muell.	Phytolaccaceae (Caryophyllales)	S. F. Brockington (s.n. Burringbar Botanic Gardens Nursery)	652 299/1 014 800	64.3	5.2	97.41	477
Nama carnosum (Wooton) C. L. Hitchc.	Boraginaceae (unplaced lamiid)	M. Moore 678 (OC)	1 069 755/1 606 440	66.6	N/A	99.96	693
Nepenthes alata Blanco	Nepenthaceae (Caryophyllales)	M. Moore 1145 (OC)	528 035/1 106 057	47.7	N/A	82.6	378
Nerisyrenia linearifolia (S. Watson) Greene	Brassicaceae (Brassicales)	M. Moore 671 (OC)	2 462 357/3 247 079	75.8	N/A	99.99	1573
Nolina brittoniana Nash	Asparagaceae (Asparagales)	J. M. Heaney (FLAS)	106 808/333 995	32	N/A	96	64
Oenothera hartwegii Benth.	Onagraceae (Myrtales)	M. Moore 628 (OC)	985 316/1 816 515	54.2	N/A	99.6	566
Petiveria alliacea L.	Phytolaccaceae (Caryophyllales)	L. Majure 4132 (FLAS)	12 249/27 892	43.9	2.5	84.9	9
Phaulothamnus spinescens A. Gray	Achatocarpaceae (Caryophyllales)	M. Moore 976 (OC)	3 445 475/6 452 382	53.4	N/A	99.99	2321
Physena madagascariensis Thouars ex Tul.	Physenaceae (Caryophyllales)	2007-895 (Kew Living Collection)	442 864/601 799	73.6	N/A	82	332
Sarcobatus vermiculatus (Hook.) Torr.	Sarcobataceae (Caryophyllales)	M. Moore 813 (OC)	314 733/652 540	48.2	N/A	94.01	236
Simmondsia chinensis (Link) C. K. Schneid.	Simmondsiaceae (Caryophyllales)	1972-3169 (Kew Living Collection)	861 126/1 432 458	60.1	N/A	93.8	577
Sporobolus nealleyi Vasey	Poaceae (Poales)	M. Moore 659 (OC)	378 858/824 691	45.9	4.5	99.44	312
Stegnosperma Benth. sp.	Stegnospermataceae (Caryophyllales)	S. F. Brockington (s.n.)	1 115 117/1 827 385	61	N/A	96.7	894.1
Tamarix L. sp.	Tamaricaceae (Caryophyllales)	M. Moore 320 (FLAS)	485 054/1 063 674	45.6	N/A	88.3	292.2

Note: N/A = not applicable.

*The data under "% Plastid reads (unenriched)" were taken from separate GAIIx or HiSeq runs without enrichment for the plastome (A. C. Crowl, unpublished *Campanula erinus* data; M. J. Moore, unpublished data for the rest).

congeneric with another species in the bait array (*Oenothera biennis*), species that are in different genera but the same family as species in the bait array (e.g., *Dicranocarpus parviflorus* vs. *Helianthus annuus*; both are Asteraceae), and species that are phylogenetically distant from all other taxa in the bait design (e.g., *Mentzelia perennis* [Loasaceae], *Acleisanthes lanceolata* [Nyctaginaceae]). We also included two monocots—*Nolina brittoniana* (Asparagaceae) and *Sporobolus nealleyi* (Poaceae)—to test whether the probes were effective beyond eudicots.

Some of the species sampled here—and in some cases, the same genomic libraries—were also sequenced in separate Illumina GAIIx or HiSeq (Illumina Inc.) runs (100-bp, single-end or paired-end reads) without enrichment for the plastid genome. Specifically, the following species were sequenced using both enriched and unenriched libraries: *Acleisanthes lanceolata*, *Campanula erinus* (same library), *Dicranocarpus parviflorus*, *Mentzelia perennis*, *Monococcus echinophorus* (same library), *Petiveria alliacea* (same library), *Sarcobatus vermiculatus* (same library), and *Sporobolus nealleyi.* This overlap presents an excellent opportunity to compare both depth and evenness of plastome coverage obtained using enriched vs. unenriched samples.

***Library construction*—**Genomic DNA (1–15 μg) was fragmented using a Covaris E220 sonicator (Covaris, Woburn, Massachusetts, USA) with the following parameters to produce fragmented DNA with a target peak of 500 bp: duty cycle = 5%; intensity = 3; cycles per burst = 200; time = 80 s. The NEBNext DNA Library Prep Master Mix Set for Illumina kit (Cat no.: E6040L, New England BioLabs, Ipswich, Massachusetts, USA) was then used to construct Illumina libraries with the sonicated DNAs and 24 different 5-bp barcodes from Craig et al. (2008). We followed the manufacturer's protocol for library construction, except that half reactions were used for most libraries to reduce per-sample

Downloaded From: https://bioone.org/journals/Applications-in-Plant-Sciences on 28 Mar 2024
Terms of Use: https://bioone.org/terms-of-use

preparation costs. Following adapter ligation, 300–400-bp fragments (insert size ~200–300 bp) were excised and purified from agarose gels using the Freeze 'N Squeeze kit (Bio-Rad, Hercules, California, USA). The size-selected libraries were then enriched using the Phusion High-Fidelity PCR Master Mix (New England BioLabs) with the following PCR program: one cycle of 98°C for 30 s; 14–18 cycles of 98°C for 10 s, 65°C for 30 s, and 72°C for 30 s; and one cycle of 72°C for 5 min, followed by a hold at 4°C. Adapter dimers were removed from enriched libraries using 0.85 volume per sample of Agencourt AMPure XP beads (Beckman Coulter, Brea, California, USA). After AMPure purification, samples were quantified using a 2100 Bioanalyzer (Agilent, Santa Clara, California, USA) and pooled into a single, equimolar mix in preparation for plastid genome capture using a SureSelect Target Enrichment Kit (Agilent) with the custom RNA baits described above.

***Plastid genome enrichment and sequencing*—**We stress that the methods described here deviate substantially from the manufacturer's protocols (see http://www.genomics.agilent.com/GenericB.aspx?PageType=Custom&SubPageType=Custom&PageID=3120). Additionally, the kit we used has been updated as Agilent has continued to refine its enrichment products (see http://www.genomics.agilent.com/CollectionSubpage.aspx?PageType=Product&SubPageType=ProductDetail&PageID=3033). We provide the information not only as a record of our methods, but also to illustrate the robustness of the kit and to encourage further experimentation among other users.

Specifically, three significant deviations were made from the manufacturer's recommendations. First, for many reasons beyond our control, the kit was nine months past the manufacturer's expiration date when it was used—clearly we would not recommend using an expired kit, but our success should reassure others who may find themselves with similarly outdated kits. Second, the kit contains blockers for the adapters that prevent nonspecific capture via adapter-adapter annealing. We used an older kit with blockers for single-end adapters, while our libraries had barcoded paired-end adapters—thus, we did not have the correct blockers in the mix. Lastly, all 24 barcoded libraries were pooled for a single capture, although the SureSelect protocol recommends selecting individual barcoded libraries followed by pooling of samples. Agilent now offers preselection pooling of barcoded libraries, although this is currently limited to 10 libraries, and the cost, while somewhat lower than 10 individual samples, is still significantly higher than one sample. Hence, performing a single selection on pooled barcoded samples is a significant and previously unsupported deviation from the manufacturer's protocol. However, again we think that our results indicate that this method will work in many situations, and this approach is the only cost-effective option for enrichment of a small region such as the plastid genome.

Other than the three significant changes discussed above, we followed the protocol outlined for the SureSelect kit (version 1.2, April 2009), using the custom RNA baits described above. After plastid genome enrichment, the 24-library pool was amplified using the Phusion High-Fidelity Master Mix (New England BioLabs) and the following program: one cycle of 98°C for 30 s; 18 cycles of 98°C for 10 s, 57°C for 30 s, and 72°C for 30 s; and one cycle of 72°C for 7 min, followed by a hold at 4°C. The amplified product was then cleaned using AMPure XP beads and sequenced on a single lane of an Illumina GAIIx at the Interdisciplinary Center for Biotechnology Research (University of Florida) with 100 cycles and single-end reads. The sequencing run generated 47 491 666 reads.

***Plastome assembly*—**Prior to plastome assembly, the reads were barcode-sorted using Novocraft (http://www.novocraft.com/main/index.php) and quality-filtered using Sickle (https://github.com/najoshi/sickle) or the FASTQ Quality Filter (FASTX-Toolkit; http://hannonlab.cshl.edu/fastx_toolkit/). The number of reads obtained for each library is shown in Table 2. De novo assemblies were conducted with the quality-filtered reads using the VelvetOptimizer script provided with Velvet (Zerbino and Birney, 2008; k-mer range: 43–81; http://bioinformatics.net.au/software.velvetoptimiser.shtml) or Geneious (using default settings and medium to high sensitivity; http://www.geneious.com/). The resulting de novo contigs were then assembled against the most closely related available reference plastome (Table 2). Prior to reference-based assembly, we removed one of the inverted repeat regions from each reference. After assembly of the contigs to the reference, we filled in as many gaps as possible by assembling the quality-filtered reads to the reference using Geneious. Any remaining gaps were filled with Ns. Regions with very low coverage in the read-to-reference assembly (below 5× coverage) were also masked with Ns. Following assembly, we used DOGMA (Wyman et al., 2004) to annotate the plastid genomes, allowing an examination of sequence depth distribution in relation to coding vs. noncoding regions of the plastome.

***Assembly statistics*—**The percent completeness of the newly assembled plastomes (vs. the reference genomes used) is presented in Table 2, which also shows the mean coverage of each assembly and the percentage of reads that assembled to the plastome reference. The enrichment efficiency across the 24 samples (i.e., the percentage of reads that assembled to the plastid genome) was on average 59%. The mean plastome coverage, averaged across the 24 species sequenced, was 717×. Examination of the coverage graphs superimposed on the annotated assemblies revealed that the sequence depth is generally nonuniform across the genome, with large spikes in depth clearly present at the coding regions (Figs. 1 and 2). This general pattern, evident across all 24 assemblies, is particularly pronounced in the species more distantly related to those included in the probe design (Figs. 1 and 2). These coverage spikes are also generally accompanied by tails of decreasing depth on either side, usually around 150–400 bp in length, roughly corresponding to the insert sizes of the libraries sequenced.

DISCUSSION

Constructing large data sets of complete (or nearly so) plastid genomes is becoming increasingly feasible due to the ever-increasing sequencing capacities of NGS instruments, particularly the Illumina GAIIx and HiSeq 2000/2500, which currently allow for parallel sequencing of 12–16 (GAIIx) or 36–48 (HiSeq 2000/2500) plastid genomes from pooled, unenriched gDNA samples. Targeted enrichment strategies for the plastid genome offer a promising means of vastly increasing the number of plastid genomes that can be sequenced in parallel, which in turn would dramatically decrease per-sample sequencing costs and increase the accessibility of plastid genome sequencing for routine phylogenetic as well as population and phylogeographic studies. The enrichment approach described in this paper shows considerable promise as a relatively simple and universal means of plastid genome enrichment (across eudicots and monocots, and potentially all angiosperms), making large-scale sequencing of angiosperm plastid genomes a more cost-effective (and therefore broadly accessible) practice.

***Increasing the limits of parallel plastome sequencing*—**A sequencing depth of ~30–50× is recognized as the minimum threshold needed for high-quality assembly of plastid genomes (Straub et al., 2012). Based on the mean coverage obtained across the 24 samples included in this study (717×), it should be theoretically possible to multiplex as many as 344 samples on a single lane of the Illumina GAIIx to obtain ~50× coverage following plastid enrichment using the probe set described here. By coupling this enrichment strategy with the even higher sequencing capacity of the HiSeq 2000 or 2500—which can yield ~187 500 000 reads per lane in a single run (Glenn, 2011)—we estimate that it should be possible, theoretically, to multiplex up to ~1300 samples and still obtain ~50× coverage of the plastid genome (given that the capacity of the HiSeq is roughly four times that of the GAIIx). This method of plastid enrichment therefore substantially increases the number of plastomes that can be sequenced in parallel. However, given the difficulties of pooling numerous samples proportionally, attempts to multiplex ~300 or more samples would probably lead to considerable variation in read numbers obtained per library. Additionally, the number of barcodes available in current adapter sets is limited (e.g., up to 96 in the NEXTflex DNA Barcode kit, Bioo Scientific, Austin, Texas, USA), and designing/purchasing adapter sets with more than 300 barcodes might be prohibitively expensive. Therefore, we suggest 96 samples as a current practical maximum for plastome multiplexing using this targeted enrichment method, but we encourage approaches to expand

Downloaded From: https://bioone.org/journals/Applications-in-Plant-Sciences on 28 Mar 2024
Terms of Use: https://bioone.org/terms-of-use

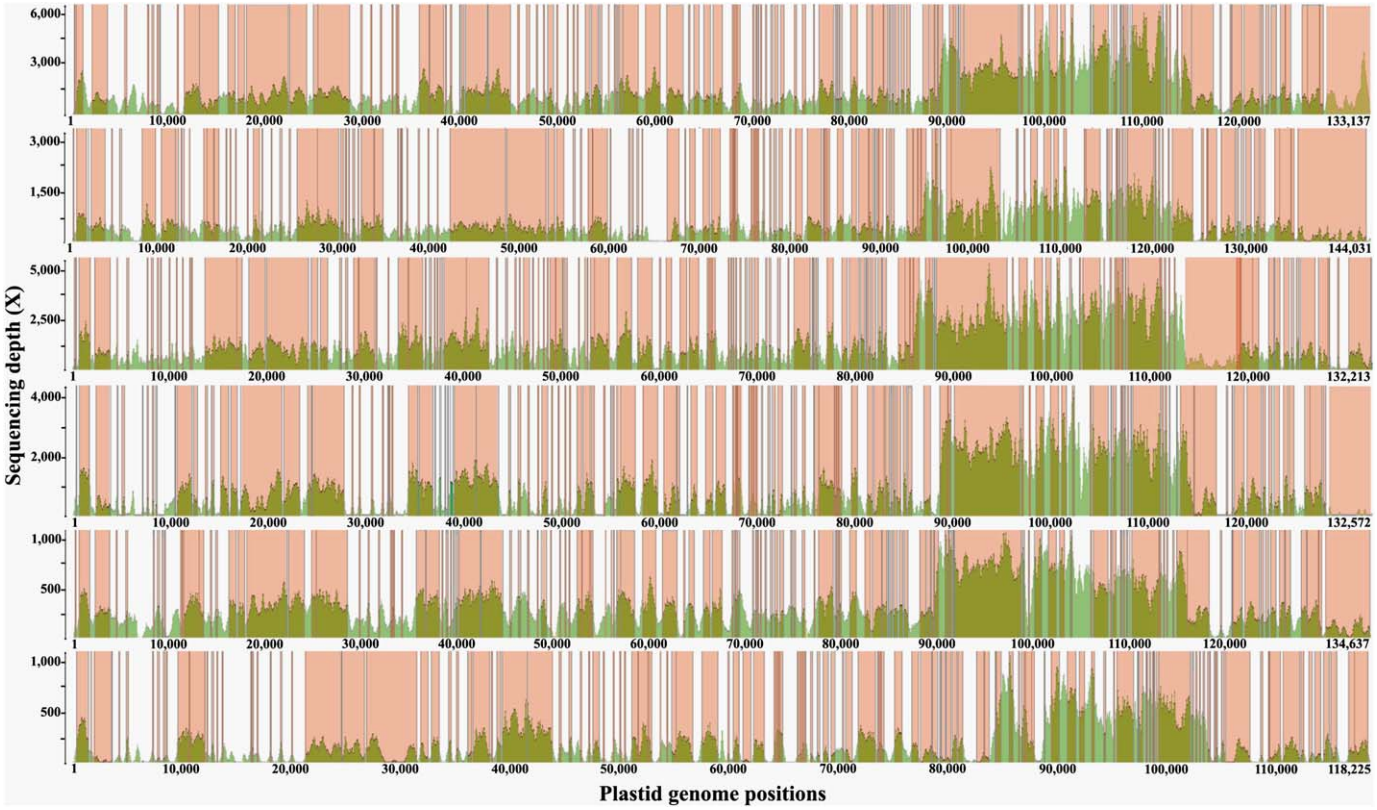


Fig. 1. Coverage graphs for six species included in this study, representing, from top to bottom, increasing phylogenetic distance from the taxa included in the probe design. From top to bottom, the species (and their closest relation to taxa included in the probe design) are: *Cucumis sativus* (same species), *Oenothera hartwegii* (same genus), *Dicranocarpus parviflorus* (same family), *Acleisanthes lanceolata* (same order), *Mentzelia perennis* (same order), *Sporobolus nealleyi* (monocot; outside the probe set's target clade). The coding regions are highlighted in red to show sequence depth obtained for coding vs. noncoding regions.

the number of multiplexed samples beyond 96, particularly to take advantage of the capacity of the HiSeq and other newer instruments that will continue to grow sequencing capacity.

Utility of the probe set—The enrichment strategy described here represents the first attempt to design a plastid probe set across a phylogenetically diverse set of samples (22 eudicot plastomes), making it broadly applicable for angiosperm plastid genome sequencing. This approach proved highly successful in recovering complete to essentially complete plastomes with impressively high coverage across most taxa tested, including monocots (Table 2). However, the depth of coverage was consistently uneven across the genome, with considerable spikes in sequence depth evident at the coding regions (Figs. 1 and 2). Because in many cases we had conspecific references available for plastome assembly, we believe this pattern reflects actual differences in depth of coverage across the genome, rather than an artifact of poor assembly due to a divergent reference. Several studies (Gnirke et al., 2009; Mamanova et al., 2010; Cronn et al., 2012; Lemmon et al., 2012) have demonstrated the importance of relatively long insert lengths for recovering more rapidly evolving (and hence divergent) spacer regions, which are usually flanked by more conserved genes that are more likely to hybridize with baits (Lemmon et al., 2012). Studies requiring more variable portions of the plastome (e.g., shallow phylogenetic and phylogeographic investigations) should therefore consider targeting relatively large insert sizes to increase the sequence coverage of these variable regions when using this plastid enrichment approach. In our study, we targeted 200–300-bp inserts, resulting in tails of decreasing sequence depth ~200–300 bp long on either side of the coverage spikes at the coding regions. Larger inserts would proportionally increase the span of the depth tails flanking the coding regions, thus capturing spacer/intronic regions with greater coverage.

Although the probe set outlined here shows immediate promise for essentially complete plastome sequencing in eudicots and monocots (which collectively represent >95% of angiosperm diversity), we anticipate that its applicability should extend to Magnoliidae, Chloranthaceae, and basal angiosperm lineages (Amborellaceae, Nymphaeales, Austrobaileyales), given that many of the probes target highly conserved coding regions of the plastid genome. However, at increasing phylogenetic distances from eudicots, the probe set will likely recover only the more conserved plastid regions, leaving behind the spacers and rapidly evolving regions useful for species- or population-level investigations (unless relatively large inserts are targeted for enrichment and sequencing). For example, Cronn et al. (2012) showed that probes designed from a single species of *Pinus* (*P. thunbergii*) could be used to enrich conserved plastid regions (i.e., those with >80% pairwise sequence identity) in a very distantly related angiosperm species (*Gossypium raimondii*). These results demonstrate that plastid probes can be successfully used for targeted enrichment (of at least highly conserved regions) across extensive phylogenetic distances.

Downloaded From: https://bioone.org/journals/Applications-in-Plant-Sciences on 28 Mar 2024
Terms of Use: https://bioone.org/terms-of-use

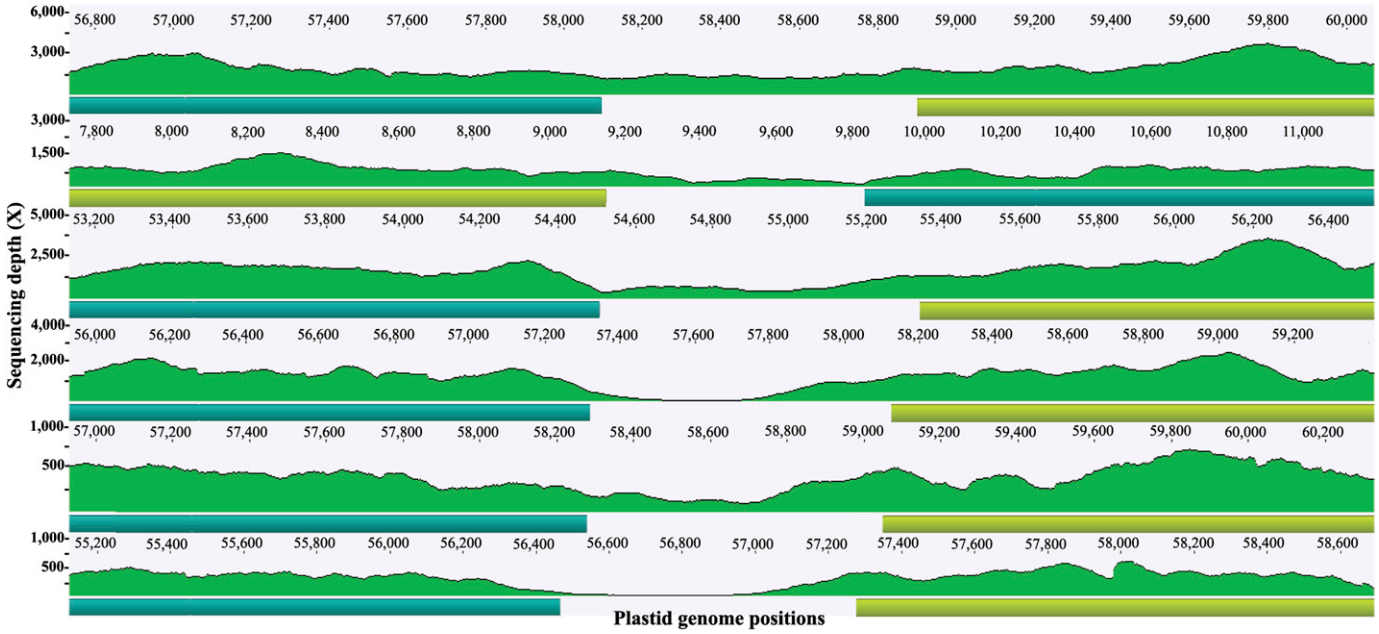


Fig. 2. Close-up of the *atpB-rbcL* spacer, from the same six species shown in Fig. 1, highlighting differences in sequence depth obtained for coding vs. noncoding regions. As in Fig. 1, the phylogenetic distance from taxa included in the probe set increases from top to bottom. The coding regions *atpB* and *rbcL* are indicated by the blue and yellow bars, respectively.

Considerations for multiplex sequencing—The low overall coverage obtained for some of the 24 libraries sequenced for this experiment is probably due to uneven pooling of libraries prior to hybridization enrichment. No phylogenetic pattern is evident in those taxa that had low coverage, and fairly close relatives of these low-coverage samples had much higher coverage. For example, *Sarcobatus* and *Acleisanthes* had extremely high coverage using SureSelect, whereas *Petiveria* had low coverage; all three taxa belong to the clade of Phytolaccaceae + Nyctaginaceae, and all have similar genome structures. When multiplexing large numbers of libraries, even small errors in DNA quantification can lead to significant differences in read numbers that can be compounded by the additional enrichment step after hybridization. Hence, it is crucial to quantify DNA concentration accurately in each library prior to pooling. Multiple methods are possible, including Bioanalyzer (Agilent), the Qubit 2.0 fluorometer (Life Technologies, Grand Island, New York, USA), and quantitative real-time PCR (qPCR). Because qPCR simultaneously amplifies and quantifies DNA samples, it more accurately quantifies the "sequenceable" portion of the library (i.e., the amount of DNA with successfully ligated adapters) and is thus the most accurate method overall; the Bioanalyzer and the Qubit, on the other hand, determine the total quantity of DNA in the sample regardless of adapter ligation. Likewise, fewer cycles should be used to amplify the plastid-enriched library pool. In the experiment outlined here, we used 18 cycles to amplify the 24-plex capture; this might have exacerbated the unequal enrichment of the library pool and consequently led to disparities in the number of reads obtained from each sample in the sequencing run.

Alternative sequencing strategies—Although the method outlined here represents an excellent means of large-scale plastid genome sequencing with great potential for plant phylogenetics and phylogeography, it by no means displaces the importance of alternative NGS strategies in the plant systematics community. For example, genome skimming (also known as genome survey sequencing), which involves low-coverage sequencing of whole-genomic samples, is an effective approach for recovering complete to essentially complete plastid genomes (up to ~48 on a single HiSeq 2000/2500 lane), as well as partially complete mitochondrial genomes and a wealth of nuclear data (Straub et al., 2012; Steele et al., 2012). This method is attractive in that it yields data from all three plant genomes for phylogeny reconstruction without the extra effort/cost associated with targeted enrichment, but it is important to note that, at present, considerably fewer samples can be sequenced in parallel with genome skimming compared to enrichment-based approaches, especially when using the GAIIx instrument. Moreover, only the high-copy nuclear elements (e.g., the rDNA cistron) are usually sequenced with >5× coverage in multiplex genome skimming. The shallow coverage obtained for low-copy nuclear regions may be sufficient for PCR primer design or probe development (for nuclear targeted enrichment) but generally precludes both the determination of orthology/paralogy and the immediate use of these regions in phylogenetic analysis. Targeted nuclear enrichment—employing baits designed to capture hundreds of single/low-copy nuclear loci—represents another promising yet underexplored NGS method for plant systematics. Lemmon et al. (2012) demonstrated how genomic resources could be used to develop a nuclear probe set with utility across vertebrates—a vast phylogenetic distance including ~500 million years of evolutionary history. Using available genomic or transcriptomic resources (e.g., the 1KP dataset: http://www.onekp.com/), similar probe sets could be developed for major plant clades, allowing for the recovery of hundreds of unlinked nuclear loci across hundreds of multiplexed samples.

These three alternative strategies—plastid enrichment/sequencing, genome skimming, and nuclear enrichment/sequencing—all have advantages and disadvantages related to

Downloaded From: https://bioone.org/journals/Applications-in-Plant-Sciences on 28 Mar 2024
Terms of Use: https://bioone.org/terms-of-use

their cost, time investment, and data output. Although the extra time and effort required for the hybridization-enrichment step is relatively minor compared to the effort required for gDNA library preparation, targeted enrichment kits (e.g., Agilent SureSelect, Roche Nimblegen, MYcroarray) are a somewhat costly investment. Therefore, plastid genome hybridization enrichment will be most efficient in terms of time and money for projects that involve sequencing of hundreds of plastid genomes. For smaller-scale phylogenetic projects, genome skimming remains an excellent and relatively cost-effective means of multiplexing plastid genomes. The increasing availability of nuclear genomic resources makes the development of probe sets for nuclear enrichment a viable and promising NGS strategy, with potential for large-scale sequencing of hundreds of independent nuclear loci. This study and others (Cronn et al., 2012; Lemmon et al., 2012) highlight the general effectiveness of hybridization-based enrichment across relatively large phylogenetic distances, offering promise for the development of nuclear probe sets for major plant clades. Researchers should carefully consider these points and others (Cronn et al., 2012; Steele et al., 2012; Straub et al., 2012; Lemmon et al., 2012) when deciding which sequencing strategy best suits the budget and data requirements of their phylogenetic and phylogeographic studies.

LITERATURE CITED

Arakaki, M., P.-A. Christin, R. Nyffeler, A. Lendel, U. Eggli, R. M. Ogburn, E. Spriggs, et al. 2011. Contemporaneous and recent radiations of the world's major succulent plant lineages. *Proceedings of the National Academy of Sciences, USA* 108: 8379–8384.

Brockington, S. F., R. Alexandre, J. Ramdial, M. J. Moore, S. Crawley, A. Dhingra, K. Hilu, et al. 2009. Phylogeny of the Caryophyllales sensu lato: Revisiting hypotheses on pollination biology and perianth differentation in the core Caryophyllales. *International Journal of Plant Sciences* 170: 627–643.

Cantino, P. D., J. A. Doyle, S. W. Graham, W. S. Judd, R. G. Olmstead, D. E. Soltis, P. S. Soltis, and M. J. Donoghue. 2007. Towards a phylogenetic nomenclature of *Tracheophyta. Taxon* 56: 822–846.

Craig, D. W., J. V. Pearson, S. Szelinger, A. Sekar, M. Redman, J. J. Corneveaux, T. L. Pawlowski, et al. 2008. Identification of genetic variants using bar-coded multiplexed sequencing. *Nature Methods* 5: 887–893.

Cronn, R., A. Liston, M. Parks, D. S. Gernandt, R. Shen, and T. Mockler. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Research* 36: e122.

Cronn, R., B. J. Knaus, A. Liston, P. J. Maughan, M. Parks, J. V. Syring, and J. Udall. 2012. Targeted enrichment strategies for next-generation plant biology. *American Journal of Botany* 99: 291–311.

Drinnan, A. N., P. R. Crane, and S. B. Hoot. 1994. Patterns of floral evolution in the early diversification of non-magnoliid dicotyledons (eudicots). *Plant Systematics and Evolution* 8(Supplement): 93–122.

Glenn, T. C. 2011. Field guide to next-generation DNA sequencers. *Molecular Ecology Resources* 11: 759–769.

Gnirke, A., A. Melnikov, J. Maguire, P. Rogov, E. M. LeProust, W. Broockman, T. Fennel, et al. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology* 27: 182–189.

Jansen, R. K., Z. Cai, L. A. Raubeson, H. Daniell, C. W. dePamphilis, J. Leebens-Mack, K. F. Müller, et al. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proceedings of the National Academy of Sciences, USA* 104: 19369–19374.

Jian, S., P. S. Soltis, M. Gitzendanner, M. Moore, R. Li, T. Hendry, Y. Qiu, et al. 2008. Resolving an ancient, rapid radiation in Saxifragales. *Systematic Biology* 57: 38–57.

Lemmon, A. R., S. A. Emme, and E. M. Lemmon. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology* 61: 727–744.

Mamanova, L., A. J. Coffey, C. E. Scott, I. Kozarewa, E. H. Turner, A. Kumar, E. Howard, et al. 2010. Target-enrichment strategies for next-generation sequencing. *Nature Methods* 7: 111–118.

Moore, M. J., A. Dhingra, P. S. Soltis, R. Shaw, W. G. Farmerie, K. M. Folta, and D. E. Soltis. 2006. Rapid and accurate pyrosequencing of angiosperm plastid genomes. *BMC Plant Biology* 6: 17–30.

Moore, M. J., C. D. Bell, P. S. Soltis, and D. E. Soltis. 2007. Using plastid genomic-scale data to resolve enigmatic relationships among basal angiosperms. *Proceedings of the National Academy of Sciences, USA* 104: 19363–19368.

Moore, M. J., P. S. Soltis, C. D. Bell, J. G. Burleigh, and D. E. Soltis. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proceedings of the National Academy of Sciences, USA* 107: 4623–4628.

Njuguna, W., A. Liston, R. Cronn, T.-L. Ashman, and N. Bassil. 2013. Insights into phylogeny, sex function and age of *Fragaria* based on whole chloroplast genome sequencing. *Molecular Phylogenetics and Evolution* 66: 17–29.

Parks, M., R. Cronn, and A. Liston. 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biology* 7: 84.

Parks, M., R. Cronn, and A. Liston. 2012. Separating the wheat from the chaff: Mitigating the effects of noise in a plastome phylogenomic data set from *Pinus* L. (Pinaceae). *BMC Evolutionary Biology* 12: 100.

Soltis, D. E., P. S. Soltis, P. K. Endress, and M. W. Chase. 2005. Phylogeny and evolution of the angiosperms. Sinauer, Sunderland, Massachusetts, USA.

Steele, P. R., K. L. Hertweck, D. Mayfield, M. R. McKain, J. Leebens-Mack, and J. C. Pires. 2012. Quality and quantity of data recovered from massively parallel sequencing: Examples in Asparagales and Poaceae. *American Journal of Botany* 99: 330–348.

Straub, S. C. K., M. Parks, K. Weitemier, M. Fishbein, R. C. Cronn, and A. Liston. 2012. Navigating the tip of the genomics iceberg: Next-generation sequencing for plant systematics. *American Journal of Botany* 99: 349–364.

Wang, H., M. J. Moore, P. S. Soltis, C. D. Bell, S. Brockington, R. Alexandre, C. C. Davis, et al. 2009. Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proceedings of the National Academy of Sciences, USA* 106: 3853–3858.

Whittall, J. B., J. Syring, M. Parks, J. Buenrostro, C. Dick, A. Liston, and R. Cronn. 2010. Finding a (pine) needle in a haystack: Chloroplast genome sequence divergence in rare and widespread pines. *Molecular Ecology* 19(Supplement): 100–114.

Wyman, S. K., R. K. Jansen, and J. L. Boore. 2004. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20: 3252–3255.

Zerbino, D. R., and E. Birney. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* 18: 821–829.

Downloaded From: https://bioone.org/journals/Applications-in-Plant-Sciences on 28 Mar 2024
Terms of Use: https://bioone.org/terms-of-use