METHODS AND TECHNIQUES

A multi-step comparison of short-read full plastome sequence assembly methods in grasses

William P. Wysocki,[1] Lynn G. Clark,[2] Scot A. Kelchner,[3] Sean V. Burke,[1] J. Chris Pires,[4] Patrick P. Edger,[5] Dustin R. Mayfield,[4] Jimmy K. Triplett,[6] J. Travis Columbus,[7] Amanda L. Ingram[8] & Melvin R. Duvall[1]

1 *Biological Sciences, Northern Illinois University, 1425 W. Lincoln Hwy, DeKalb, Illinois 60115-2861, U.S.A.*
2 *Ecology, Evolution and Organismal Biology, Iowa State University, 353 Bessey Hall, Ames, Iowa 50011-1020, U.S.A.*
3 *Biological Sciences, Idaho State University, 921 S. 8th Ave, Pocatello, Idaho 83209-8007, U.S.A.*
4 *Division of Biological Sciences, University of Missouri, 1201 Rollins St, Columbia, Missouri 65211, U.S.A.*
5 *Department of Plant and Microbial Biology, University of California – Berkeley, Berkeley, California 94720, U.S.A.*
6 *Department of Biology, Jacksonville State University, 144B Martin Hall, Jacksonville, Alabama 36265, U.S.A.*
7 *Rancho Santa Ana Botanic Garden & Claremont Graduate University, 1500 North College Avenue, Claremont, California 91711, U.S.A.*
8 *Department of Biology, Wabash College, P.O. Box 352, Crawfordsville, Indiana 47933, U.S.A.*
Author for correspondence: *William P. Wysocki, wwysoc2@gmail.com*
ORCID: J.C.P., http://www.orcid.org/ J. Chris Pires: 0000-0001-9682-2639

DOI http://dx.doi.org/10.12705/634.5

Abstract Technological advances have allowed phylogenomic studies of plants, such as full chloroplast genome (plastome) analysis, to become increasingly popular and economically feasible. Although next-generation short-read sequencing allows for full plastomes to be sequenced relatively rapidly, it requires additional attention using software to assemble these reads into comprehensive sequences. Here we compare the use of three de novo assemblers combined with three contig assembly methods. Seven plastome sequences were analyzed. Three of these were Sanger-sequenced. The other four were assembled from short, single-end read files generated from next-generation libraries. These plastomes represented a total of six grass species (Poaceae), one of which was sequenced in duplicate by the two methods to allow direct comparisons for accuracy. Enumeration of missing sequence and ambiguities allowed for assessments of completeness and efficiency. All methods that used de Bruijn-based de novo assemblers were shown to produce assemblies comparable to the Sanger-sequenced plastomes but were not equally efficient. Contig assembly methods that utilized automatable and repeatable processes were generally more efficient and advantageous when applied to larger scale projects with many full plastomes. However, contig assembly methods that were less automatable and required more manual attention did show utility in determining plastomes with lower read depth that were not able to be assembled when automatable procedures were implemented. Although the methods here were used exclusively to generate grass plastomes, these could be applied to other taxonomic groups if previously sequenced plastomes were available. In addition to comparing sequencing methods, a supplemental guide for short-read plastome assembly and applicable scripts were generated for this study.

Keywords ACRE; de novo assembly; next-generation sequencing; plastome; Poaceae

Supplementary Material The Electronic Supplement (Tables S1–S2; Appendix S1: Guide to short-read plastome assembly) is available in the Supplementary Data section of the online version of this article at http://www.ingentaconnect.com/content/iapt/tax

■ INTRODUCTION

Systematic studies of land plants have advanced in the context of molecular phylogenetic research. Access to next-generation sequencing (NGS) methods has given plant systematists the ability to analyze genome-scale data for large numbers of terminal taxa to investigate evolutionary relationships. Taxa within the grass family (Poaceae) are of particular interest due to the economic importance of cereal grains, their use in functional genomic studies (Botiri & al., 2008) and their complex evolutionary history. These phylogenomic studies—which are genome-scale phylogenetic studies—have been shown to offer improved resolution, stronger support, and allow for more confident estimates of divergence dates. In this context, complete chloroplast genomes (plastomes) are tools of demonstrated phylogenetic utility for studies at different taxonomic levels. Intrageneric phylogenomic studies show fine-scale relationships, document microstructural events, and offer explanations for historical biogeographic patterns (Cronn & al., 2008; Parks & al., 2009, 2012; Burke & al., 2012, 2014).

Received: 22 Jan 2014 | returned for first revision: 3 May 2014 | last revision received: 27 May 2014 | accepted: 27 May 2014 | published online ahead of inclusion in print and online issues: 5 Aug 2014 ||

Intergeneric studies show broader evolutionary patterns even in taxa with slowly evolving plastomes, indicate patterns of genome evolution, and show lineage-specific rate variations (Zhang, Y.J. & al., 2011; Hand & al., 2013). Studies within families have resolved formerly intractable phylogenetic issues and revealed readily interpretable patterns of molecular evolution (Leseberg & Duvall, 2009; Duvall & al., 2010; Wu & Ge, 2012).

Growing use of complete plastomes by plant systematists does not reflect uniform methodological choices for determining and assembling these data. In part, this is due to the rapid changes of the NGS technologies, with periodic advances often accompanied by increases in read length, which directly impact the efficiency and accuracy of assembly. The development of new software tools for assembly is another factor. However, these do not entirely account for the diversity of methods of plastome assembly employed in published reports. Although there are numerous comparative studies of different methods of assembling large genomes (Lin & al., 2011; Zhang, W. & al., 2011; Liu & al., 2012; and many others), we find a single such study for assembling complete plastomes (Steele & al., 2012). Plastome assembly presents challenges such as a large inverted repeat region, mitogenomes with similar sequences that are likely intermingled in the same pool of reads, and AT-richness, which introduces periodic regions of low sequence complexity. Plastomes also present unique opportunities for phylogenomic analysis, foremost of which are their highly conserved sequences and structures. Major structural changes have been occasionally well documented at intergeneric or interfamilial levels (e.g., Doyle & al., 1992; Cosner & al., 2004). However, plastomes from monocot congeners such as *Acorus americanus* (Raf.) Raf. (NC010093) and *A. calamus* L. (NC007407) and those from *Oryza sativa* L. "Japonica Group" (NC001320) and "Indica Group" (NC008155), *O. meridionalis* N.Q.Ng (NC016927) and *O. nivara* S.D.Shartma & Shastry (NC005973) show 99.54% and 99.49% nucleotide identities in alignments of these plastomes respectively (Wysocki, unpub. comparisons).

In existing studies of complete plastomes, different combinations of assembly methods have been used. Prior to assembly, NGS reads may be trimmed based on sequence quality (although see Paszkiewicz & Studholme, 2010 for a contrasting view). Reads may also be filtered by comparison against published plastomes, discarding those that fail to meet a threshold nucleotide identity (e.g., Hand & al., 2013). The risk here is that low frequency events caused by intergenomic recombinations within a species may be missed. During assembly, de novo methods implemented in a variety of software packages, which follow different algorithms, may be used. Assemblies may be accompanied by a reference-guided step, where contigs are aligned to the plastome of a closely related species (Zhang, Y.J. & al., 2011), or even where an intermediate pseudoreference is created that is chimeric between the de novo sequences and the reference (Whittall & al., 2010). Some Sanger-sequencing of plastome fragments may also be performed to close gaps in the assembly (Hand & al., 2013), verify assembled sequences in mutation hotspots (Whittall & al., 2010) or to identify the boundaries of the major inverted repeats (Zhang, Y.J. & al., 2011).

A robust test for accuracy of plastome assembly would be to compare duplicated sequences from the same plant using two different methods, such as Sanger and Illumina. To this end a Sanger-sequenced plastome of *Neyraudia reynaudiana* (Kunth) Keng ex Hitchc. along with NGS sequenced plastome of the same individual was used to test their accuracy. Additionally we used Sanger-sequenced plastomes of two other species (*Arundinaria gigantea* (Walt.) Muhl., *Pharus latifolius* L.) along with NGS sequenced plastomes from closely related congeners of these species to perform a somewhat less stringent test, similar to the strategy employed by Steele & al. (2012). This is justified because of the high identities of plastomes between grass and other monocot congeners (see above).

The study species represent three major lineages of grasses. From the PACMAD clade (acronym abbreviates the subfamilial membership for Panicoideae, Arundinoideae, Chloridoideae, Micrairoideae, Aristidoideae, and Danthonioideae) the chloridoid *Neyraudia reynaudiana* was selected and compared against itself for the two sequencing methods. From the BEP clade (acronym for: Bambusoideae, Ehrhartoideae, and Pooideae) the New World bambusoid species (following Clark & Triplett, 2007) *Arundinaria gigantea* and both of its congeners, *A. tecta* (Walt.) Muhl. and *A. appalachiana* Triplett & al., were selected. Finally, from one lineage of the deeply diverging grade of grasses *Pharus latifolius* and the congeneric species *P. lappulaceus* Aubl. were selected. By assessing the accuracy, completeness, and efficiency of these assembly methods we compared different approaches to assembly and established guidelines for full plastome determination of grasses using short reads of approximately 100 base pairs (bp) produced from single-read libraries.

■ MATERIALS AND METHODS

DNA extraction and Sanger sequencing. — Silica-dried leaf samples were obtained from five species: *Arundinaria appalachiana*, U.S.A., *J. Triplett JT099* (ISC); *A. tecta*, U.S.A., *J. Triplett JT173* (ISC); *Neyraudia reynaudiana*, China, *J.T. Columbus 5302* (RSA); *Pharus latifolius*, U.S.A., *J. Triplett 421* (MO); and *P. lappulaceus*, U.S.A., *J. Triplett 422* (MO). Leaf tissue was homogenized manually in liquid nitrogen before extraction. The DNA extraction protocol using the Qiagen DNeasy Plant Mini Kit (Qiagen, Valencia, California, U.S.A.) was followed. The extraction from fresh leaf tissue of a sixth species *Arundinaria gigantea* was similarly performed as described in Burke & al. (2012).

Methods for obtaining complete plastomes using Sanger sequencing generally followed those described by Dhingra & Folta (2005). Overlapping segments of these plastomes were amplified using universal plastome primers for the inverted repeat (IR) regions (Dhingra & Folta, 2005) and primers specific for other regions of the grass plastome (Leseberg & Duvall, 2009). Each primer pair flanks a region of approximately 1200 bp. There are 125 such regions, 28 of which lie within the major inverted repeat and do not require duplicate sequencing except to locate the IR boundaries. Touchdown

PCR was performed with all primer pairs using the "round I" conditions described by Dhingra & Folta (2005). Failure of PCR when using the main set of primers was addressed with the alternative methods of Morris & Duvall (2010) including the design of species-specific primers (Electr. Suppl.: Table S1; Burke & al., 2012). Amplicons were purified using the Wizard SV PCR Clean-up System (Promega, Madison, Wisconsin, U.S.A.) and sent for sequencing at Macrogen (Seoul, South Korea). Quality of sequence information was verified and sequence identities were confirmed in duplicated bidirectional and overlapping sequences. Sequence assembly was performed using Geneious Pro v.6.1.6 (Biomatters, Auckland, New Zealand). All manual sequence manipulations in this study were performed using the Geneious Pro software package. By these methods a draft plastome of *Neyraudia reynaudiana* (88% complete) was determined and complete plastomes were determined for *Arundinaria gigantea* (Burke & al., 2012) and *Pharus latifolius* (Jones & al., 2014).

Next-generation sequencing (NGS). — Starting quantities of total genomic DNA from *Neyraudia reynaudiana*, *Arundinaria appalachiana*, *A. tecta*, and *Pharus lappulaceus* were determined by measurement at A260 with a Nanodrop 1000 (ThermoFisher Scientific, Wilmington, Delaware, U.S.A.) to be approximately 1.5 µg each. DNAs were diluted to approximately 2 ng/µl and sheared into ~300 bp fragments using a Bioruptor sonicator (Diagenode, Denville, New Jersey, U.S.A.) in two 12 min periods, inverting the tubes between periods. Sonicated DNA preparations were purified and concentrated with the MinElute Extraction Kit (Qiagen). Single-read libraries were prepared using the TruSeq sample preparation low-throughput protocol (gel method) following manufacturer instructions (Illumina, San Diego, California, U.S.A.). Sequencing was performed on a HiSeq 2000 instrument (Illumina) at Iowa State University (Ames, U.S.A.). Reads produced by this method were 99 bp in length. The single-reads were first quality filtered using DynamicTrim v.2.1 from the SolexaQA software package (Cox & al., 2010) with default settings, and then sequences shorter than 25 bp in length (default) were removed with LengthSort v.2.1 from the same package. The quality of the reads was then assessed using FastQC v.0.10.1 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

The next portion of the methods is an overview of how the reads presented here were assembled and then scaffolded into full plastomes using different methods at both of these steps. A guide to assembling a full plastome using these methods is included in the Electronic Supplement to this article as Appendix S1.

De novo assembly of NGS reads into contigs

(1) Greedy assembly. — The de novo assembly program in the Geneious Pro v.6.1.6 package provides a greedy assembly algorithm, similar to that of a multiple sequence alignment, which aligns all possible combinations of reads and assesses identity to produce contigs (http://www.geneious.com). Because of the computational burden associated with greedy assembly on complete sets of reads, further filtration was performed on each set of reads to include only those with high identity to published plastome sequences. This filtration step was performed using the BLASTn software package (Altschul & al., 1997), which used a full plastome as a query against the read file database and an e-value threshold of 10^{-3}. The full plastome for *Panicum virgatum* L. (NC015990) was used as a query for the *N. reynaudiana* reads, *Bambusa oldhamii* Munro (NC012927) was used as a query for the *A. tecta* and *A. appalachiana* reads, and *Anomochloa marantoidea* Brongn. (NC014062) was used as a query for the *P. lappulaceus* reads. These taxa were used as a reference instead of the available conspecific or congeneric plastomes to simulate a more likely situation in which a plastome from the same species or congener would not typically be available. No filtration step was used prior to the other de novo assembly methods. Python scripts were used to extract matched sequences from the read files. The greedy assembly software was used to assemble these extracted sequences into contigs.

(2 and 3) Single-pass de novo assembly using de Bruijn-based assemblers. — "Single-pass" de novo assemblies were performed. There are a large number of de Bruijn-based assembly packages and we screened four of these in preliminary assemblies including Edena v.3 (Hernandez & al., 2008), SOAPdenovo v.1.02 (Li & al., 2010), SPAdes v.2.4.0 (Bankevich & al., 2012; http://bioinf.spbau.ru/spades), and Velvet v.1.2.08 (Zerbino & Birney, 2008; http://www.ebi.ac.uk/~zerbino/velvet/). Two packages, SPAdes (2) and Velvet (3), which were specifically designed for the assembly of small genomes, were ultimately selected based on larger contig sets that had high plastome homology and speed of assembly, although some other assembly packages may also be appropriate for use in plastome determination.

"Single-pass" de novo assemblies were performed using Velvet and SPAdes. For purposes of comparison, the parameters for both programs were set to twelve identical *k*-mer lengths (units of read overlap). A minimum *k*-mer length of 19 was chosen and increased by steps of 6 bp 11 times until the maximum *k*-mer length, 85, was reached. The shorter *k*-mer length is used to achieve assembly in areas of lower coverage while the longer *k*-mer length is used to keep this method consistent with the next section in which longer contigs as used as input. Although this procedure includes 12 assemblies, we refer to it as a "single pass" assembly, to distinguish it from our alternative approach ("iterative assembly", see below). One Velvet assembly was performed for each *k*-mer length and the contigs generated from each of these were pooled into single files. Note that SPAdes performed this task automatically. The N50 of each contig set was calculated using a Python script.

Iterative de novo assembly. — The contigs generated by performing single-pass de novo assemblies using Velvet were assembled into larger contigs using the same procedure. This method could not be performed with the SPAdes software package because of a software limitation on user-provided sequence length, in which initially assembled contig files could not be reassembled. Assemblies were repeated until substantially larger contigs were no longer being generated, which occurred after the third assembly. The output from each step was pooled into one file and assembled de novo one additional time. These results were then pooled with the previous de novo assembly results because sequences that are not incorporated into Velvet assemblies are excluded from the output files. To summarize, in this iterative approach, assemblies were performed three times and were finally applied to all intermediate de novo results in a fourth and final Velvet assembly (Fig. 1). Python scripts were used to automate this repetitive task.

Contig scaffolding

(4) Map to reference. — This type of scaffolding used the previously assembled Sanger-sequenced plastomes of closely related species to assess the position of each DNA fragment. One IR region was omitted from the reference plastome to reduce the number of erroneous hits to the inverted repeat boundaries. The Geneious Pro software package was used to map the contigs generated onto their positions within

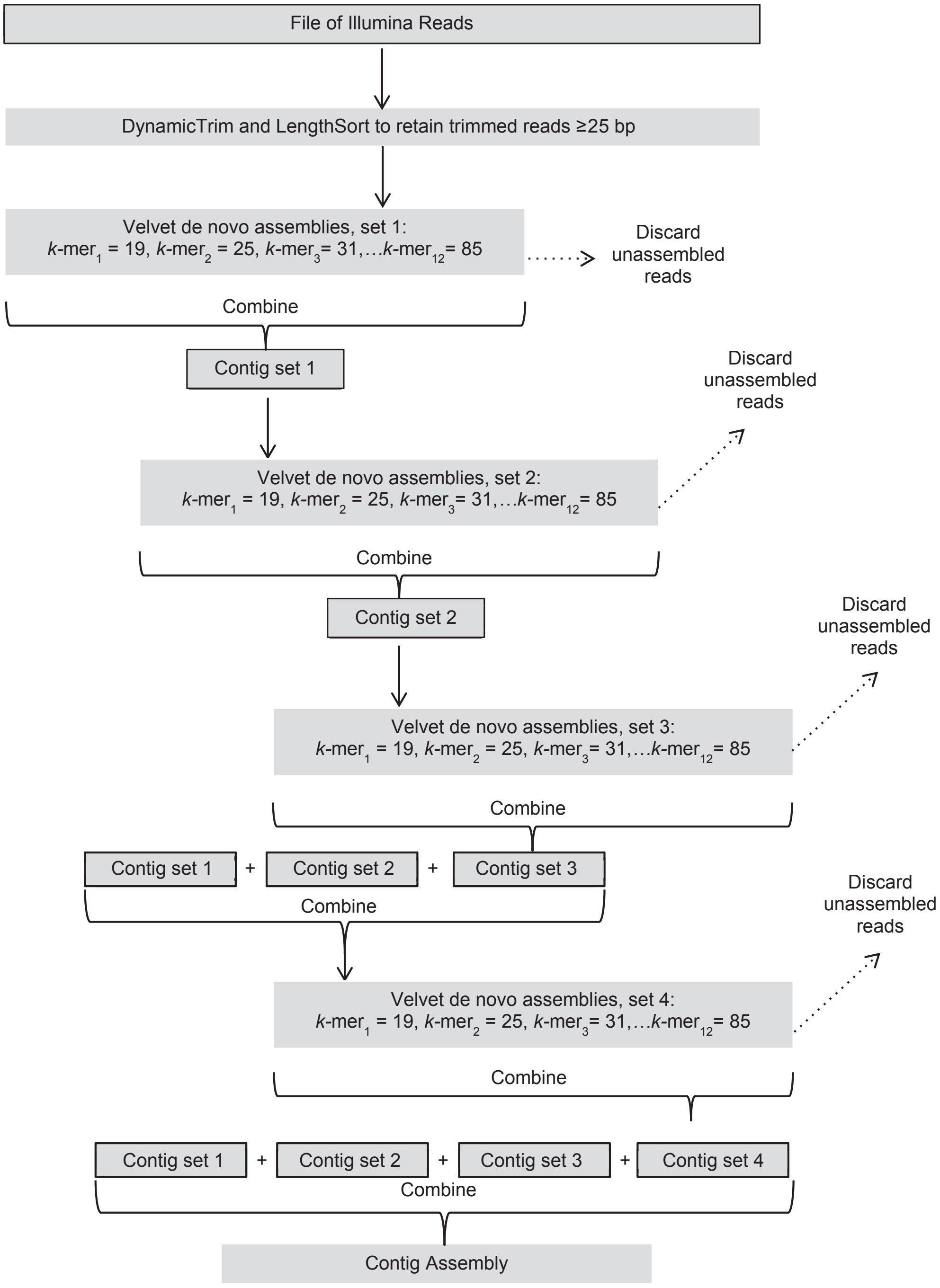
File of Illumina Reads
DynamicTrim and LengthSort to retain trimmed reads ≥25 bp
Velvet de novo assemblies, set 1:
k-mer$_1$ = 19, k-mer$_2$ = 25, k-mer$_3$= 31,…k-mer$_{12}$= 85
Discard unassembled reads
Combine
Contig set 1
Discard unassembled reads
Velvet de novo assemblies, set 2:
k-mer$_1$ = 19, k-mer$_2$ = 25, k-mer$_3$= 31,…k-mer$_{12}$= 85
Combine
Contig set 2
Discard unassembled reads
Velvet de novo assemblies, set 3:
k-mer$_1$ = 19, k-mer$_2$ = 25, k-mer$_3$= 31,…k-mer$_{12}$= 85
Combine
Contig set 1 + Contig set 2 + Contig set 3
Combine
Discard unassembled reads
Velvet de novo assemblies, set 4:
k-mer$_1$ = 19, k-mer$_2$ = 25, k-mer$_3$= 31,…k-mer$_{12}$= 85
Combine
Contig set 1 + Contig set 2 + Contig set 3 + Contig set 4
Combine
Contig Assembly

Fig. 1. Summary of iterative de novo assembly using Velvet. Boxes with borders indicate either read or contig files. Boxes without borders indicate processes. Points at which contig files are combined are indicated.

plastomes. Each contig set produced by de novo assembly was mapped to its respective full plastome reference that was used to query the reads file using BLASTn as previously. While the mapping methods of Steele & al. (2012), which used the published plastome from each taxon as a reference, would allow for more accurate assemblies, we used closely related reference plastomes instead for reasons stated above.

(5) *In silico* genome walking. — Contigs generated using de novo assembly were scaffolded by locating overlapping regions. First the longest contig with plastome homology within Poaceae was selected. A region of 20–45 bp in length at the end of each contig was used as a query in the pool of generated contigs. When an exact match was located, the contig that contained this match was concatenated to the end of the initial contig (minus the overlapping region). Genome walking was performed both manually and using Python scripts to automate the process.

(6) Anchored conserved region extension (ACRE). — A rough-draft plastome alignment of 75 taxa within Poaceae was generated (Duvall & al., unpub.) using MAFFT v.1.2 (Katoh & al., 2005). MAFFT was also used for all subsequent alignments in this study. Regions that were identically conserved among all 75 taxa and were greater than 19 bp in length were put into an input file (Electr. Suppl.: Table S2). This length was chosen because it is long enough to reasonably assess homologous regions and short enough to locate complete family-specific conserved regions. For each of these 85 regions, the largest contig that included the region was located and scaffolded in the order in which they appeared in the plastome by combining overlapping regions. This original ACRE method was then automated by the first author, written as a Python script. These scripts as well as the ones used for *in silico* genome walking can be found at http://sourceforge.net/projects/grassplastome/.

Final assessment

Final plastome assembly. — The quality of all preliminary scaffoldings was assessed by aligning the scaffoldings to the respective reference plastome used previously for read filtration using MAFFT. This allowed for the percentage of each plastome assembled to be calculated for each contig assembly method. Each plastome assembly that covered more than 80% of its reference was used in subsequent analysis. Missing regions in each assembly were inserted manually by locating flanking overlapping regions from each respective read file or the combined contigs file for larger gaps. Large changes such as indel mutations greater than 20 bp in length or large regions of substitutions were verified by locating overlapping sequences in the original read file. Final plastome assemblies were arranged and trimmed to a proper orientation for sequence alignment. The 5′ end of the large single-copy region was positioned on the 5′ end of the total assembly while the 5′ end of the inverted repeat region A (IRa) was positioned at the 3′ end of the total assembly. The IRa region was omitted from alignment because the inclusion of two inverted repeats would double the representation of that region during subsequent analyses. These boundaries were located using the methods in Burke & al. (2012) by identifying the region where the end of the sequence matches the 3′ IRb boundary and where the beginning of the sequence meets the 5′ IRb boundary using BLAST.

Assessment of contig assembly methods. — Assemblies generated using each method were assessed using one criterion for accuracy, the percent identity that the assembly shared with its partnered Sanger-sequenced plastome, and two criteria for completeness, the quantity of missing sequence in the plastome prior to final plastome assembly and the number of ambiguous nucleotide sites that could not be resolved by the contig assembly. Percent identity was assessed by aligning the completed plastome generated using NGS data to either the Sanger-sequenced plastome of the same species (*N. reynaudiana*) or the Sanger-sequenced plastome of its congener (*A. appalachiana*, *A. tecta*, *P. lappulaceus*). Nucleotide sites in the alignment of the *N. reynaudiana* plastomes, where gaps were present in the Sanger sequence, were omitted so that only the Sanger-sequenced 88% of the plastome was represented while calculating the percent similarity.

Assessment of read depth. — As a final assessment, each complete plastome was subjected to a reference mapping using only each respective trimmed and filtered read set. Because of the unusually low number of reads associated with *A. tecta*, the mapping for this species used a more sensitive setting along with a gap optimization step. This allowed for the mean, minimum and maximum read depth to be calculated using the Geneious Pro software. By mapping the reads onto the final assembly we ensured that read depth measurements were accurate. A consensus sequence was also generated from the read mapping which was aligned to the assembled plastome as further verification.

■ RESULTS

Sequences of complete plastomes were determined and deposited in GenBank for five species—*Neyraudia reynaudiana*, *Arundinaria appalachiana*, *A. tecta*, *Pharus latifolius*, and *P. lappulaceus.* Note that the sixth plastome from *Arundinaria gigantea* was previously sequenced and assembled (Burke & al., 2012). GenBank accession numbers and comparative lengths of plastome regions and subregions are given in Table 1.

After trimming for quality and filtering out short reads, the single-read files for *N. reynaudiana*, *A. appalachiana*, *A. tecta*, and *P. lappulaceus* contained 7.96, 5.42, 0.92 and 1.75 million reads, respectively. Fragments in each set were 25–99 bp in length, had a median length of 99 bp, and a mean length of 93–94 bp. After de novo assembly (Methods 1–3; Fig. 2) the longest contig in each set with grass chloroplast homology ranged from 2699 to 48,146 bp in length. The N50 of each contig set ranged from 86 to 8114. The results for both of these measurements are reported in Table 2.

Out of the 48 combinations of de novo assembly and contig scaffolding methods tested here, 12 of these met the 80%

Table 1. GenBank accession numbers and lengths of regions and subregions for plastomes used in this study.

		Lengths [bp]			
		Total	LSC[a]	SSC[b]	IR[c]
Neyraudia reynaudiana	KF356392	135,367	80,616	12,695	21,028
Arundinaria gigantea	JX235347	138,935	82,641	12,700	21,797
A. appalachiana	KC817462	139,547	83,223	12,717	21,804
A. tecta	KC817463	139,499	83,162	12,730	21,804
Pharus latifolius	JN032131	142,077	83,341	12,530	23,103
P. lappulaceus	KC311467	141,928	83,188	12,536	23,102

[a] Large single-copy region [b] Short single-copy region [c] Inverted-repeat region

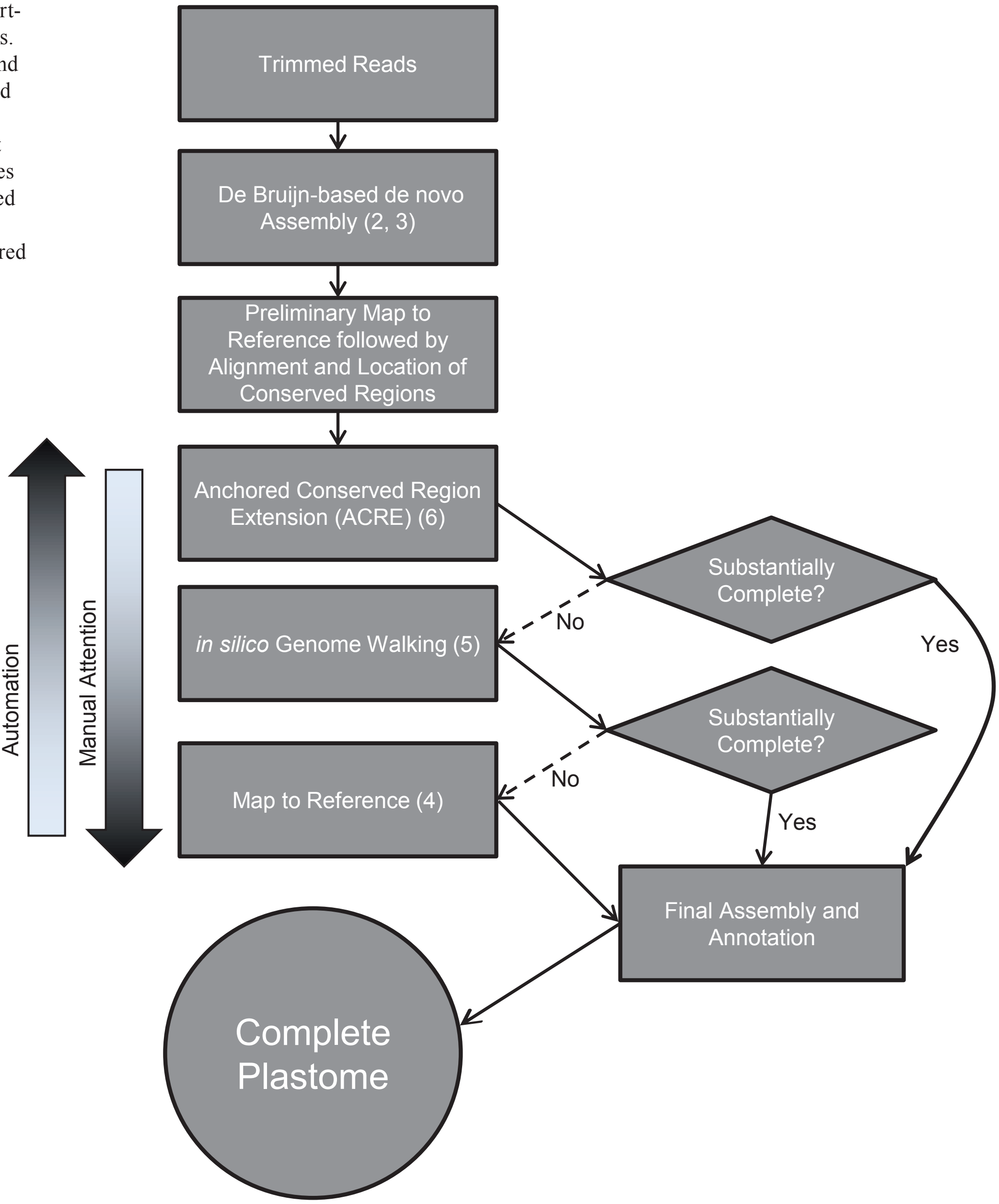
Trimmed Reads
De Bruijn-based de novo Assembly (2, 3)
Preliminary Map to Reference followed by Alignment and Location of Conserved Regions
Anchored Conserved Region Extension (ACRE) (6)
in silico Genome Walking (5)
Map to Reference (4)
Substantially Complete?
No
Yes
Substantially Complete?
No
Yes
Final Assembly and Annotation
Complete Plastome
Automation
Manual Attention

Fig. 2. Overall summary of short-read plastome assembly process. Numbers in each box correspond to each assembly tool or method as outlined in the Materials and Methods section. Note that method 1, greedy assembly, does not appear in this figure. Shaded arrow directionality indicates increasing automation or required manual attention.

plastome coverage threshold. Note that the contigs generated by the de novo greedy assembly method were the shortest, had the most missing and ambiguous sites and produced the lowest percent plastome coverage after scaffolding. Because the percent plastome coverage did not exceed the target 80%, full plastomes were not generated using these contig sets.

(4) Map to reference. — This method generated entire plastome sequences that possessed similarity with each corresponding Sanger-sequenced plastome ranging from 39.98% to 98.87% before manual attention was given to the sequence and a final assembly was produced. These sequences were missing or ambiguous at 297–56,413 nucleotide sites with the greatest number of these sites being generated when SPAdes was used to assemble the reads de novo even though this method often produced the longest contigs. The results for each map to reference assembly are reported in Table 3. The mappings that were manually resolved into a final assembly after using Velvet iteratively for *N. reynaudiana*, *A. appalachiana*, *A. tecta*, and *P. lappulaceus* showed 99.38%, 99.06%, 98.96%, and 98.94% similarity (Table 4), respectively, to their partnered Sanger-sequenced plastome. These were the only set of reference mappings that were assembled into full plastomes because Velvet produced the highest amount of coverage for each species (Table 3) and the iterative strategy produced larger contigs than the single-pass strategy (Table 2).

(5) *In silico* genome walking. — This method was able to initially generate greater than 80% of a plastome for only one

Table 2. The N50 values and largest contigs generated for each set of reads using the Greedy assembly function included in the Geneious Pro package, SPAdes, and Velvet. Velvet assemblies were performed using both the single-pass and iterative methods.

Species	De novo assembly method	N50	Largest contig [bp]
Neyraudia reynaudiana	Geneious	388	3,431
	SPAdes	401	48,146
	Single-Pass Velvet	125	24,265
	Iterative Velvet	153	36,174
Arundinaria appalachiana	Geneious	8,114	28,388
	SPAdes	506	28,504
	Single-Pass Velvet	86	20,022
	Iterative Velvet	105	23,075
A. tecta	Geneious	281	2,823
	SPAdes	506	4,817
	Single-Pass Velvet	93	2,699
	Iterative Velvet	99	4,991
Pharus lappulaceus	Geneious	1,327	8,554
	SPAdes	395	18,863
	Single-Pass Velvet	97	5,429
	Iterative Velvet	106	13,154

Table 3. Summary of the results of the contig scaffoldings performed using the reference mapping function in Geneious Pro v.6.1.6 after de novo assembly using Velvet, SPAdes, or the assembly function included in the Geneious Pro package. Velvet assemblies were performed using both the single-pass and iterative methods.

Species	De novo assembly software	Missing or ambiguous nucleotide sites	% identity w/ Sanger-sequence
Neyraudia reynaudiana[a]	Iterative Velvet	1,898	97.01
	Single-Pass Velvet	577	98.87
	SPAdes	47,486	53.74
	Geneious	44,455	65.72
Arundinaria appalachiana	Iterative Velvet	355	97.61
	Single-Pass Velvet	297	97.40
	SPAdes	18,985	81.34
	Geneious	16,371	85.33
A. tecta	Iterative Velvet	22,228	70.04
	Single-Pass Velvet	34,689	69.29
	SPAdes	56,413	39.98
	Geneious	31,524	72.21
Pharus lappulaceus	Iterative Velvet	1,469	92.11
	Single-Pass Velvet	3,114	91.07
	SPAdes	11,566	84.84
	Geneious	15,647	82.95

[a] Percent similarity with the Sanger plastome for *Neyraudia* was calculated over the 88% completed.

set of contigs (*N. reynaudiana* assembled de novo using Velvet iteratively) and substantially less for the rest of the contig sets (Table 5). The scaffolded contigs from *Neyraudia* assembled using Velvet also required regions to be manually rearranged to conserve gene order and complete both inverted repeat sequences. After rearrangement of regions and final assembly was performed, this method produced an assembly that was 99.46% similar to its partnered partially Sanger-sequenced plastome.

(6) ACRE. — This assembly method was attempted with the contigs generated using both Velvet and SPAdes for each of the four taxa. Before any manual attention was given to produce a final assembly, these methods generated from 99.63% (*N. reynaudiana* contigs assembled de novo iteratively using Velvet iteratively) to 4.11% (*A. tecta* contigs assembled de novo using SPAdes) of the full plastome. Table 5 reports these percentages. When the assemblies from *A. tecta* were omitted from the set, using the ACRE method yielded 81.08%–99.63% of the plastome automatically. Because this analysis concatenated overlapping contigs with matching motifs, it also generated no ambiguous nucleotide sites. After final assembly was completed, ACRE methods produced assemblies that were 98.48% to 99.66% similar to their partnered Sanger-sequenced plastomes (Table 4).

Read depth. — Read depth varied between taxa but stayed consistent within taxa for each of the different assembly methods. Mean read depth values fell within a broad range from 16.4 to 132.7. Mean, minimum, and maximum read depth for each assembled plastome are reported in Table 4.

■ DISCUSSION

Key to the successful use of plastomes for grass systematics are methods to assemble short NGS reads economically, accurately, rapidly, and in a largely automated manner with little subsequent need for manual adjustment. The use of single-read libraries can economize the production of next-generation sequence files. Over the past eight years, many de Bruijn-based de novo assembly software packages have been released (Zhang, W. & al., 2011) and have been shown to assemble short reads effectively and conservatively. However, as the reads are assembled conservatively into contigs complete plastomes are not always produced, especially when single-read technology, which is more cost effective under certain conditions, is used (as opposed to paired-end reads). This creates a need to assemble them using somewhat less stringent methods. The contigs assembled by this type of software can be used to assemble a smaller genome by aligning them to a closely related taxon where gene order is largely conserved, as in most Poaceae, and by filling in any gaps with reads or smaller contigs using flanking overlap. While this method is fairly simple and accurate it can become very time consuming and laborious when applied to a larger-scale study and may, in some cases, introduce bias.

The assemblies produced using each method were assessed using three criteria: their identity with each partnered Sanger plastome, the quantity of missing sequence, and the number of ambiguous sites. Because of the longer reads and targeted nature of Sanger sequencing, a plastome sequenced using this method can be a reliable aid in testing whether shorter and

Table 4. The similarity between plastome assemblies generated using NGS data and their partnered Sanger-sequenced plastome. Each assembly method includes a de novo assembly followed by contig scaffolding. Plastome assemblies that were less than 80% complete after contig scaffolding are not included here.

				Read depth (per bp)		
Species	Assembly	% Identity with Sanger	Polymorphic nucleotide sites	Mean	Min	Max
Neyraudia reynaudiana[a]	Velvet (iter)[b]-ACRE	99.66	343	132.7	33	439
	Velvet (SP)[c]-ACRE	99.66	345	132.5	1	439
	SPAdes-ACRE	99.66	344	132.7	33	439
	Velvet (iter)-Walking	99.46	549	132.7	33	439
	Velvet (iter)-MTR[d]	99.38	618	129.8	2	445
Arundinaria appalachiana	Velvet (iter)-ACRE	99.24	889	16.8	1	1,725
	SPAdes-ACRE	99.35	766	16.4	1	171
	Velvet (iter)-MTR	99.06	1,102	16.4	1	170
A. tecta	Velvet (iter)-MTR	98.96	1,228	31.8	1	709
Pharus lappulaceus	Velvet (iter)-ACRE	98.48	1,542	18.8	1	86
	SPAdes-ACRE	98.54	1,739	18.8	1	86
	Velvet (iter)-MTR	98.94	1,259	18.9	1	86

[a] *N. reynaudiana* % identity with its partnered Sanger-sequenced plastome was calculated using the 88% of the existing, Sanger-sequenced plastome.
[b] Velvet(iter): Velvet run iteratively according to methods outlined in Fig. 1.
[c] Velvet(SP): Velvet run once.
[d] MTR: Map to reference method.

randomly targeted NGS reads are assembled accurately. Since an accurate assembly is crucial in subsequent phylogenomic analyses, percent similarity to its partnered Sanger sequence is the most important criterion. The efficiency of the assembly, which can be quantified using the number of gaps and ambiguous sites, is of less concern because complications can be easily identified and resolved. However, this does present a concern because of the amount of time and labor required to accurately perform these adjustments. At first glance, repairing a small number of assemblies seems manageable, but as the number of assemblies grows, this task becomes an increasing hindrance to achieving final downstream analyses.

One criterion that was not emphasized as an assessment of assembly quality here is the N50 of each contig file. The N50, used in many NGS assembly studies, is a weighted median statistic for assessing the distribution of contig lengths, where a higher value reflects a greater proportion of longer contigs. We report N50 scores (Table 2). Studies with the objective of generating the largest possible contigs for purposes of assembling large sequences such as eukaryotic chromosomes (e.g., Li & al., 2010; Nowrousian & al., 2010) require de novo assembly of a contig set with a higher N50. In the study presented here large contigs are useful in establishing preliminary assemblies while the smaller contigs are also useful for gap bridging and resolution of ambiguities. In addition, the N50 statistic is not comparable between different assemblies if their combined lengths are not equal (Miller & al., 2010). Other factors such as the presence of nuclear, mitochondrial, and microbial sequences within each contig set also make the comparison of N50 values between assembled contig sets less useful for the methods of grass plastome assembly considered here since a large contig could potentially represent one of these other sources of DNA.

After completion, *N. reynaudiana* and *A. appalachiana* exhibited greater than 99% similarity with the Sanger-sequenced plastome. Plastome assemblies for *A. tecta*, and *P. lappulaceus* exhibited greater than 98% similarity with each of their Sanger-sequenced congeneric plastomes (Table 4). While two plastomes from the same individual would be expected to be indistinguishable regardless of which sequencing method was used, sequencing artifacts such as erroneous base calls by capillary sequencing software and incorrectly incorporated bases during early stages of PCR can result in low frequency nucleotide polymorphisms between the two assemblies. The partially Sanger-sequenced plastome for *N. reynaudiana* and its next-gen assemblies contained polymorphisms at less than 0.5% of the nucleotide sites.

Prior to contig scaffolding, methods for de novo assembly of reads exhibited no clear patterns in effectiveness. The plastome for *N. reynaudiana* was most accurately assembled using Velvet iteratively, but was only marginally more accurate than the single-pass strategy using either of the de Bruijn assemblers used here. Although using single-pass Velvet generally allowed for fewer ambiguities after subsequent reference mapping (Method 4), this strategy did not produce more than 56% of a complete plastome when combined with genome walking (Method 5) or ACRE (Method 6) in the other three taxa. The SPAdes software package did perform somewhat more accurately and with comparable efficiency than

Table 5. Percent of each Sanger-sequenced plastome covered by first de novo assembling using Velvet, SPAdes, or the greedy assembly function included in the Geneious Pro package on each set of reads and then applying the ACRE or *in silico* genome walking scaffolding method to each set of contigs. Velvet assemblies were performed using both the single-pass and iterative methods.

Species	De novo assembly method	ACRE (%)	Walking (%)
Neyraudia reynaudiana	Iterative Velvet	99.63	89.62
	Single-Pass Velvet	95.85	24.52
	SPAdes	98.75	49.72
	Geneious	25.12	2.92
Arundinaria appalachiana	Iterative Velvet	92.94	23.00
	Single-Pass Velvet	55.19	20.29
	SPAdes	81.08	23.93
	Geneious	79.29	24.18
A. tecta	Iterative Velvet	7.79	4.26
	Single-Pass Velvet	14.87	2.64
	SPAdes	17.36	4.11
	Geneious	18.54	2.40
Pharus lappulaceus	Iterative Velvet	84.12	13.77
	Single-Pass Velvet	31.83	8.43
	SPAdes	93.60	18.19
	Geneious	62.99	7.16

the iterative Velvet approach when combined with the ACRE method in *A.appalachiana* and *P. lappulaceus*. However, the iterative Velvet approach executed more rapidly than SPAdes assemblies.

Performing the ACRE method after running Velvet iteratively or SPAdes single-pass also successfully revealed a 596 bp insertion in the *psbE*-*petL* intergenic spacer of *A. appalachiana*. This sequence was absent in the Sanger-sequenced *A. gigantea*, but found in all other Arundinarieae (Burke & al., 2012). Note that a reference-guided assembly of the plastome of *A. appalachiana* using *A. gigantea* as a reference failed to detect this large insertion because of the bias imposed by the reference (Wysocki, unpub. obs.). This shortcoming of reference mapping assembly demonstrates that improvements can be made to assemblies by using methods that rely less on reference sequence data.

The reference mapping scaffolding method did generate substantial amounts of the plastome sequence, however the sheer number of ambiguous nucleotide sites presents a problem for efficient assembly. To generate a full and accurate plastome, each of these ambiguous sites would require a verification using sequences within the original read file. This would require the arduous, time-consuming and computationally taxing endeavor of querying millions of reads for motifs flanking each of the potentially thousands of ambiguous sites. Note that reference assembly cannot be accomplished if no closely related reference sequences are available.

Although the reference mapping method (Method 4) can be labor intensive, it does occasionally possess an indispensable role in plastome assembly. The contigs generated from the *A. tecta* reads did not contain enough coverage or overlap to complete a substantial amount of the plastome sequence using genome walking or ACRE, likely because the original read file contained the fewest reads. Reference mapping after using the iterative Velvet strategy did produce a substantial amount of the plastome and by manually repairing the ambiguities a full plastome was able to be generated.

In silico genome walking (Method 5) can function well as a method for de novo assembly if contigs are well represented around the entire plastome. However, genome walking will fail in areas that contain little overlap, even if larger areas are well represented. The overlap does require an exact sequence match, which presents a problem for regions that contain ambiguities in poorly covered areas of the plastome. This method also exhibits problems when approaching IR region boundaries. Genome walking software cannot distinguish between assembling a plastome downstream in one IR region and upstream in the other copy. Even when the repeat regions are assembled correctly, a product of this type of assembly performed on a circular sequence may also produce a fragment with regions in an order that may need to be manually rearranged for alignment purposes. While these complications can be remedied with ease by locating the boundaries of each region of the plastome and disassembling the sequence and placing each region where it belongs, this also consumes time and labor and cannot be conventionally streamlined. One benefit of genome walking is that preliminary data, such as draft or reference plastomes, are not necessary. Plastome assembly using genome walking only requires that a plastome-homologous sequence that was assembled de novo from reads is identified. This method could potentially allow for de novo assembly of plastomes within taxa that lack a Sanger-sequenced reference plastome and in which gene order may not be preserved, such as between the plastomes of Poaceae and non-grass monocots. Another application for genome walking is its ability to span highly variable regions where a reference sequence could not be utilized. This could allow for nuclear genes with large introns to be extracted from contig assemblies and the reads themselves.

While most complete assemblies performed here showed similarity greater than 98%, the ACRE method (6) produced full plastome assemblies that required the least amount of manual attention. The ACRE method can be effective in rapidly generating accurate datasets for large-scale intrafamilial analyses when preliminary data from previously sequenced plastomes that allow identification of conserved regions are available, as from Poaceae, Asteraceae, and Fabaceae. The abundance of preliminary data for these families can be attributed in part to their large size and widespread scientific interest. However, this type of assembly will become applicable to other families as more plastomes are sequenced and conserved regions can be identified. The utility of this type of assembly can also be attributed to the map to reference method. While a quick map to reference on many taxa without manual resolution of ambiguous sites can produce erroneous results in phylogenomic analyses and will not clarify some of the unique features of a sequence, it can aid in finding ultra-conserved regions within a small genome. These regions are fundamental to performing the ACRE assembly method, which can effectively compensate for the previously mentioned weaknesses of a map to reference.

■ CONCLUSIONS

When performing a phylogenomic study that includes numerous taxa, automation of assembly becomes essential to eliminate tedious and time-consuming steps. Because of this, small genomes should be assembled first using the most automatable methods (ACRE, *in silico* genome walking), which can largely complete the assembly using read files with high coverage and overlap. This can be followed by less automatable processes (map to reference, manual assembly), which may be necessary for taxa that possess a smaller number of reads and do not produce successful results when subjected to automatable processes (Fig. 2).

For purposes of measurement, methods performed in this study were kept consistent and comparable. However, a variety of adjustments could be applied to the protocols outlined here to ensure that more successful, efficient, and complete assemblies can be performed. One obvious adjustment is the insertion of all of the reads into each respective contig file prior to assemblies such as mapping to a reference or *in silico* genome walking. These reads may ensure that overlapping regions with insufficient read depth, which may have been lost during initial de novo assembly, are represented and can potentially eliminate ambiguities and holes. Another potential adjustment is to combine two or more contig sets that were generated from the same reads file, but used more than one de novo assembly software package or algorithm.

Phylogenomic studies are quickly becoming more large-scale and widespread. This makes efficient and high-throughput pipelines for assembling short-read sequences increasingly crucial. In Poaceae, applications range from agrostology, biodiversity and bioconservation studies, bioengineering and functional ecology. While the scope of the aforementioned methods may be limited to highly conserved plastomes, they can be utilized to perform other tasks by altering parameters. As technological capabilities increase and allow for longer sequencing reads and more advanced computing power, modified versions of the methods used in this paper can be put to use for plastome determination with more efficient assemblies that produce more accurate results.

■ ACKNOWLEDGMENTS

We thank L. Attigala, J. Cotton, T. Hajek, A. Moore, and Y. Yin for technical assistance. This work was supported in part by the Plant Molecular Biology Center, the Department of Biological Sciences at Northern Illinois University and the National Science Foundation under Grant Numbers DEB-1120761 to LGC, SAK, and MRD, DEB-1146603 to JCP, PPE and DRM, NPGI-1202793 to PPE, and DEB-0920147 to JTC. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

■ LITERATURE CITED

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. 1997. Gapped BLAST and PSIBLAST: A new generation of protein database search programs. *Nucl. Acids Res.* 25: 3389–3402. http://dx.doi.org/10.1093/nar/25.17.3389

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A. & Pevzner, P.A. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Computat. Biol.* 19: 455–477. http://dx.doi.org/ 10.1089/cmb.2012.0021

Botiri, E., Coleman-Derr, D., Lazo, G.R., Anderson, O.D. & Gu, Y.Q. 2008. The complete chloroplast genome of *Brachypodium distachyon*: Sequence comparison and phylogenetic analysis of eight grass plastomes. *B. M. C. Res. Notes*. 1: 61. http://dx.doi.org/10.1186/1756-0500-1-61

Burke, S.V., Grennan, C.P. & Duvall, M.R. 2012. Plastome sequences of two New World bamboos, *Arundinaria gigantea* and *Cryptochloa strictiflora* (Poaceae), extend phylogenomic understanding of Bambusoideae. *Amer. J. Bot.* 99: 1951–1961. http://dx.doi.org/ 10.3732/ajb.1200365

Burke, S.V., Clark, L.G., Triplett, J.K., Grennan, C.P. & Duvall, M.R. 2014. Biogeography and phylogenomics of New World Bambusoideae (Poaceae), revisited. *Amer. J.Bot.* 101: 886–891. http://dx.doi.org/10.3732/ajb.1400063

Clark, L.G. & Triplett, J.K.. 2007. *Arundinaria*. Pp. 17–18 in: Barkworth, M.E., Capels, K.M., Long, S., Anderton, L.K. & Piep, M.B. (eds.), *Flora of North America north of Mexico*, vol. 24. New York: Oxford University Press.

Cosner, M., Raubeson, L. & Jansen, R. 2004. Chloroplast DNA rearrangements in Campanulaceae: Phylogenetic utility of highly rearranged genomes. *B. M. C. Evol. Biol*. 4: 27. http://dx.doi.org/10.1186/1471-2148-4-27

Cox, M.P., Peterson, D.A. & Biggs, P.J. 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *B. M. C. Bioinf.* 11: 485. http://dx.doi.org/ 10.1186/1471-2105-11-485

Cronn, R., Liston, A., Parks, M., Gernandt, D.S., Shen, R. & Mockler, T. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucl. Acids Res.* 36: e122. http://dx.doi.org/ 10.1093/nar/gkn502

Dhingra, A. & Folta, K. 2005. ASAP: Amplification, sequencing & annotation of plastomes. *B. M. C. Genomics* 6: 176–189. http://dx.doi.org/ 10.1186/1471-2164-6-176

Doyle, J.J., Davis, J.I., Soreng, R.J., Garvin, D. & Anderson, M.J. 1992. Chloroplast DNA inversions and the origin of the grass family (Poaceae). *Proc. Natl. Acad. Sci. U.S.A*. 89: 7722–7726. http://dx.doi.org/ 10.1073/pnas.89.16.7722

Duvall, M., Leseberg, C.R., Grennan, C.P. & Morris, L.M. 2010. Molecular evolution and phylogenetics of complete chloroplast genomes in Poaceae. Pp. 437–450 in: Seberg, O., Petersen, G., Barfod, A.S., & Davis, J.I. (eds.), *Diversity, phylogeny, and evolution in the monocotyledons*. Aarhus: Aarhus University Press.

Hand, M.L., Spangenberg, G.C., Forster, J.W. & Cogan, N.O. 2013. Plastome sequence determination and comparative analysis for members of the *Lolium-Festuca* grass species complex. *G3 Genes Genomes Genetics* 3: 607–616. http://dx.doi.org/ 10.1534/g3.112.005264

Hernandez, D., Francois, P., Farinelli, L., Osteras, M. & Shrenzel, J. 2008. De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. *Genome Res.* 18: 802–809. http://dx.doi.org/ 10.1101/gr.072033.107

Jones, S.S., Burke, S.V. & Duvall, M.R. 2014. Phylogenomics, molecular evolution, and estimated ages of lineages from the deep phylogeny of Poaceae. *Pl. Syst. Evol.* 300: 1421–1436. http://dx.doi.org/10.1007/s00606-013-0971-y

Katoh, K., Kuma, K., Toh, H. & Miyata, T. 2005. MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucl. Acid. Res.* 33: 511–518. http://dx.doi.org/ 10.1093/nar/gki198

Leseberg, C.H. & Duvall, M.R. 2009. The complete chloroplast genome of *Coix lacryma-jobi* and a comparative molecular evolutionary analysis of plastomes in cereals. *J. Molec. Evol*. 69: 311–318. http://dx.doi.org/ 10.1007/s00239-009-9275-9

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Li, S., Yang, H., Wang, J. & Wang, J. 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* 20: 265–272. http://dx.doi.org/10.1101/gr.097261.109

Lin, Y., Li, J., Shen, H., Zhang, L. & Papasian, C.J. 2011. Comparative studies of de novo assembly tools for next-generation sequencing technologies. *Bioinformatics* 27: 2031–2037. http://dx.doi.org/ 10.1093/bioinformatics/btr319

Liu, X., Pande, P., Meyerhenke, H. & Bader, D. 2012. PASQUAL: Parallel techniques for next generation genome sequence assembly. *IEEE Trans. Parallel Distributed Systems* 24: 977–986. http://dx.doi.org/10.1109/TPDS.2012.190

Miller, J.R., Koren, S. & Sutton, G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics* 95: 315–327. http://dx.doi.org/ 10.1016/j.ygeno.2010.03.001

Morris, L.M. & Duvall, M.R. 2010. The chloroplast genome of *Anomochloa marantoidea* (Anomochlooideae; Poaceae) comprises a mixture of grass-like and unique features. *Amer. J. Bot.* 97: 620–627. http://dx.doi.org/ 10.3732/ajb.0900226

Nowrousian, M., Stajich, J.E., Chu, M., Engh, I., Espagne, E., Halliday, K., Kamerewerd, J., Kempken, F., Knab, B., Kuo, H.C., Osiewacz, H.D., Poggeler, S., Read, N.D., Seiler, S., Smith, K.M., Zickler, D., Kuck, U. & Freitag, M. 2010. De novo assembly of a 40Mb eukaryotic genome from short sequence reads: *Sordaria macrospora*, a model organism for fungal morphogenesis. *PLoS Genet.* 6: e1000891. http://dx.doi.org/10.1371/journal.pgen.1000891

Parks, M., Cronn, R. & Liston, A. 2009. Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *B. M. C. Biol.* 7: 84. http://dx.doi.org/ 10.1186/1741-7007-7-84

Parks, M., Cronn, R. & Liston, A. 2012. Separating the wheat from the chaff: Mitigating the effects of noise in a plastome phylogenomic data set from *Pinus* L. (Pinaceae). *B. M. C. Evol. Biol.* 12: 100. http://dx.doi.org/10.1186/1471-2148-12-100

Paszkiewicz, K. & Studholme, D.J. 2010. De novo assembly of short sequence reads. *Briefings Bioinf.* 11: 457–472. http://dx.doi.org/10.1093/bib/bbq020

Renzaglia, K.S., Schuette, S., Duff, R.J., Ligrone, R., Shaw, A.J., Mishler, B.D. & Duckett, J.G. 2007. Bryophyte phylogeny: Advancing the molecular and morphological frontiers. *Bryologist* 110: 179–213. http://dx.doi.org/ http://dx.doi.org/10.1639/0007-2745

Steele, P.R., Hertweck, K.L., Mayfield, D., McKain, M.R., Leebens-Mack, J. & Pires, J.C. 2012. Quality and quantity of data recovered from massively parallel sequencing: Examples in Asparagales and Poaceae. *Amer. J. Bot.* 99: 330–348. http://dx.doi.org/ 10.3732/ajb.1100491

Whittall, J.B., Syring, J., Parks, M., Buenrostro, J., Dick, C., Liston, A. & Cronn, R. 2010. Finding a (pine) needle in a haystack: Chloroplast genome sequence divergence in rare and widespread pines. *Molec. Ecol.* 19(s1): 100–114. http://dx.doi.org/10.1111/j.1365-294X.2009.04474.x

Wu, Z-Q. & Ge, S. 2012. The phylogeny of the BEP clade in grasses revisited: Evidence from the whole genome sequences of chloroplasts. *Molec. Phylogen. Evol.* 62: 573–578. http://dx.doi.org/ 10.1016/j.ympev.2011.10.019

Zerbino, D.R. & Birney, E. 2008. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18: 821–829. http://dx.doi.org/10.1101/gr.074492.107

Zhang, W., Chen, J., Yang, Y., Tang, Y., Shang, J. & Shen, B. 2011. A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. *PLoS ONE* 6: e17915. http://dx.doi.org/10.1371/journal.pone.0017915

Zhang, Y.J., Ma, P.F. & Li, D.Z. 2011. High-throughput sequencing of six bamboo chloroplast genomes: Phylogenetic implications for temperate woody bamboos (Poaceae: Bambusoideae). *PLoS ONE* 6: e20596. http://dx.doi.org/10.1371/journal.pone.0020596