BioOne COMPLETE

Geneious! Simplified Genome Skimming Methods for Phylogenetic Systematic Studies: A Case Study in Oreocarya (Boraginaceae)

Authors: Ripma, Lee A., Simpson, Michael G., and Hasenstab-Lehman, Kristen

Source: Applications in Plant Sciences, 2(12)

Published By: Botanical Society of America

URL: https://doi.org/10.3732/apps.1400062

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 29 Mar 2024
Terms of Use: https://bioone.org/terms-of-use

Applications in Plant Sciences 2014 2(12): 1400062

Application Article

Geneious! Simplified genome skimming methods for phylogenetic systematic studies: A case study in *Oreocarya* (Boraginaceae)[1]

Lee A. Ripma[2,4], Michael G. Simpson[2], and Kristen Hasenstab-Lehman[3]

[2]Department of Biology, San Diego State University, San Diego, California 92182-4614 USA; and [3]Rancho Santa Ana Botanic Garden, 1500 N. College Avenue, Claremont, California 92117 USA

- *Premise of the study:* As systematists grapple with how to best harness the power of next-generation sequencing (NGS), a deluge of review papers, methods, and analytical tools make choosing the right method difficult. *Oreocarya* (Boraginaceae), a genus of 63 species, is a good example of a group lacking both species-level resolution and genomic resources. The use of Geneious removes bioinformatic barriers and makes NGS genome skimming accessible to even the least tech-savvy systematists.
- *Methods:* A combination of de novo and reference-guided assemblies was used to process 100-bp single-end Illumina HiSeq 2000 reads. A subset of 25 taxa was used to test the suitability of genome skimming for future systematic studies in recalcitrant lineages like *Oreocarya*.
- *Results:* The nuclear ribosomal cistron, the plastome, and 12 mitochondrial genes were recovered from all 25 taxa. All data processing and phylogenomic analyses were performed in Geneious. We report possible future multiplexing levels and published low-copy nuclear genes represented within de novo contigs.
- *Discussion:* Genome skimming represents a much-improved primary data collection over PCR+Sanger sequencing when chloroplast DNA (cpDNA), nuclear ribosomal DNA (nrDNA), and mitochondrial DNA (mtDNA) are the target sequences. This study details methods that plant systematists can employ to study their own taxa of interest.

Key words: Amsinckiinae; Geneious; genome skimming; next-generation sequencing (NGS); *Oreocarya*; phylogenomics.

The power of next-generation sequencing (NGS) is transforming the study of nonmodel plant taxa (Soltis et al., 2013). Sweeping statements about the utility of NGS to answer previously intractable questions are commonplace in systematics journals. The initial bioinformatic hurdle and the fact that NGS technology can be used in different ways (see review by Godden et al., 2012; Soltis et al., 2013) inhibit many systematists from beginning studies. Briefly, many NGS library preparation methods rely on genome reduction, including targeting the transcriptome (e.g., Wen et al., 2013), nuclear loci (e.g., Weitemier et al., 2014), or the plastome (e.g., Stull et al., 2013). Reduction techniques capturing large numbers of nuclear loci require baseline genomic knowledge (see review by Cronn et al., 2012). In contrast, systematists can use the NGS genome skimming method (Straub et al., 2012) to assemble the high-copy fraction of total genomic DNA (gDNA) into the nuclear ribosomal cistron (nrDNA), plastome (cpDNA), and individual mitochondrial genes (mtDNA) without genome reduction during library preparation. With shallow sequencing of the nuclear DNA (nDNA), deeper sequencing for the high-copy fraction of gDNA is achieved (hence "skimming"). Additionally, these data generate baseline information from the nDNA to identify known single-copy and low-copy nuclear genes (LCNG) that are potentially fruitful for future targeted sequencing studies (Straub et al., 2012). The genome skimming method has been used to produce family-level phylogenies (Malé et al., 2014), species-level phylogenies (Parks et al., 2009; Straub et al., 2012), and infra-species phylogenies (Whittall et al., 2010; Kane et al., 2012).

Reads from a genome skim can be assembled with many bioinformatically complex methods, for example: the alignreads pipeline (Straub et al., 2011), the command line Velvet assembler (Zerbino and Birney, 2008), Python scripts from the OBITools package (Malé et al., 2014), Trinity (Grabherr et al., 2011; e.g., Bock et al., 2014), or various custom scripts (e.g., Kane et al., 2012). Even basic programming skills required to assemble sequences present a hurdle for many systematists (Godden et al., 2012; Soltis et al., 2013; L. A. Ripma, personal observation). Comparable methods can be implemented in Geneious Pro (Geneious version 7.1.5; Biomatters Ltd., Auckland, New Zealand [http://www.geneious.com/]), a program with a user-friendly graphical user interface (GUI). The use of Geneious for processing genome skim reads was first presented to the authors at the Botany 2012 workshop entitled "Introduction

[1]Manuscript received 22 July 2014; revision accepted 7 November 2014.

The authors thank S. Straub and the Liston laboratory for providing the exposure to genome skimming that inspired this study; R. B. Kelley for sharing extensive knowledge of *Oreocarya* natural history; and C. M. Guilliams, S. Derkarabetian, and three anonymous reviewers for helpful comments on the manuscript. Funding is acknowledged from the Riverside County Community Foundation's Desert Legacy Grant, the California Native Plant Society, the American Society of Plant Taxonomists, Southern California Botanists, and Northern California Botanists.

[4]Author for correspondence: leeripma@gmail.com

doi:10.3732/apps.1400062

 © 2014 Ripma et al. Published by the Botanical Society of America. This work is licensed under a Creative Commons Attribution License (CC-BY-NC-SA).

Downloaded From: https://bioone.org/journals/Applications-in-Plant-Sciences on 29 Mar 2024
Terms of Use: https://bioone.org/terms-of-use

to Next Generation Sequencing" (Liston, 2012; Straub, 2012). Our study demonstrates that in the genus *Oreocarya* Greene (Boraginaceae), reads from a genome skim can be assembled into nrDNA, cpDNA, and mtDNA sequences at levels suitable for phylogenetic inference, solely using GUI programs.

Oreocarya, a genus of slow-growing perennials distributed in mostly xeric habitats (Bresowar and McGlaughlin, 2014), of approximately 63 species and 72 taxa (Kelley and Ripma, in preparation for *Flora of North America*, vol. 15), presents an ideal system for demonstrating the utility of new NGS methods. To date, species-level resolution in the genus has consistently proven elusive due to a lack of parsimony informative characters (PICs) (Marushak, 2003; Bresowar and McGlaughlin, 2011; Hasenstab-Lehman and Simpson, 2012). Most studies have supported the monophyly of *Oreocarya* (Hasenstab-Lehman and Simpson, 2012; Nazaire and Hufford, 2012; Weigend et al., 2013), placing it in a clade referred to as the subtribe Cryptanthinae (Hasenstab-Lehman and Simpson, 2012) or "*Cryptantha* clade" (Weigend et al., 2013), referred to here as subtribe Amsinckiinae (Brand, 1931). As other studies have shown, variation in DNA sequences is often insufficient to resolve lower-level taxonomic relationships using traditional markers (Parks et al., 2009; Whittall et al., 2010; Godden et al., 2012); therefore, further systematic studies of *Oreocarya* required a new approach.

Several authors have reviewed the possibilities of NGS in plant systematics (Straub et al., 2012; Godden et al., 2012; Lemmon and Lemmon, 2013; Soltis et al., 2013). In our study, the genome skimming method was selected due to a lack of baseline genomic resources to design nuclear exon probes (Cronn et al., 2012), the knowledge that specimens of many taxa in future studies of the Amsinckiinae would be silica dried or from herbarium sheets, and the relative low-cost of gDNA library preparation. The goals of this study are to (1) develop and present methods for processing genome skimming data in the user-friendly program Geneious, and (2) test the feasibility of genome skimming for systematic studies in *Oreocarya* and Amsinckiinae and inform these future studies.

MATERIALS AND METHODS

***Taxon sampling*—**DNA was extracted from silica-dried leaf samples ($n = 17$) collected concurrently with vouchered specimens, or taken directly from recently collected herbarium specimens ($n = 8$). Collections are housed at the San Diego State University herbarium (SDSU) or Jepson Herbarium at University of California, Berkeley (UC) (Appendix 1). Sampling included 19 *Oreocarya* taxa and six outgroups from Amsinckiinae (Table 1). Because the genome skimming method is evaluated as a method to continue systematic studies of *Oreocarya*, samples represent the taxonomic breadth of Higgins's (1971) groups within the genus.

***DNA isolation and sequencing*—**Genomic DNA was isolated using a modified cetyltrimethylammonium bromide (CTAB) protocol (Doyle and Doyle, 1987; Friar, 2005). DNA samples were prepared for sequencing by Global Biologics (Columbia, Missouri, USA) using the following protocol: DNA samples were quantitated using the Qubit dsDNA HS Assay Kit and Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, California, USA) and integrity was checked using the Advanced Analytical Technologies Fragment Analyzer and Genomic DNA kit (Ames, Iowa, USA); high-molecular-weight DNA (>15 kb) samples showing no degradation were considered suitable for libraries. A 500–1000-ng sample of DNA was normalized to 40 µL in a low-bind 96-well microplate and sheared to ~300 bp using the Q700 Sonicator (QSonica, Newtown, Connecticut, USA). The fragmented DNA was blunt-end repaired, 3′ adenylated, and ligated with multiplex compatible adapters using the NEXTflex DNA Sequencing Kit for Illumina (catalog no. 514104; Bioo Scientific, Austin, Texas, USA) prior to being size selected to retain ~200–400-bp fragments with Agencourt AMPure XP SPRI beads (Beckman Coulter, Brea, California, USA). PCR enrichment selectively amplified fragments containing DNA with adapters on both ends. Library validation used the Fragment Analyzer followed by quantitation with the Qubit dsDNA HS Assay Kit and the qPCR kit for Illumina (Kapa Biosystems, Wilmington, Massachusetts, USA). Equimolar amounts of each library were pooled at 10 nM for sequencing. High-throughput sequencing used the Illumina HiSeq 2000 genetic analysis system (San Diego, California, USA) at the University of Delaware Sequencing and Genotyping Center for Run 1 (single-end 100-bp reads) and the University of California at Riverside Genomics Core for Run 2 (single-end 101-bp reads).

***DNA quality control filtering*—**Raw read quality control and filtering used PRINSEQ (Schmieder and Edwards, 2011) with the following parameters: all exact sequence duplicates, reads with a mean quality Phred score below 30, and reads with more than one N were removed. Both the 3′ and 5′ end were trimmed to a Phred quality score of 30 using a window size of 1 (Straub et al., 2013). Any read less than 50 bp in length was removed. The barcode to multiplex a sample was removed from the corresponding read pool. Post–quality control reads were imported into Geneious in FASTQ format, and are hereafter referred to as read pools.

***De novo assembly*—**All assemblies in this study were performed on a MacBook Pro (Apple Inc., Cupertino, California, USA) with a 2.7-GHz Intel Core i7 and 16 GB of memory. A de novo assembly was performed for each read pool using the Geneious de novo assembler with default settings. A consensus sequence of each contig greater than 100 bp in length was saved with a 75% threshold for sequence matching (80% is the threshold used by Whittall et al., 2010; Parks et al., 2012; Straub et al., 2012). Positions with under 5× coverage were converted to sequence base calling ambiguity (Ns), and International Union of Pure and Applied Chemistry (IUPAC) ambiguity codes were retained. The de novo assembly contigs were used to recover nearly complete plastid sequences for use in downstream reference-guided assemblies. Plastid contigs were identified by MegaBLAST searching all contigs against the *Solanum lycopersicum* L. (AM087200) plastome using an *E*-value of 1e-10 (Wu et al., 2006), a k-mer length of 24, a scoring match-mismatch of 1-2, and a linear open extend gap cost. Note that recovered plastome sequences were only refined through iterative assemblies; no primers were designed to PCR verify gene boundaries, confirm sequences, or verify assembled gene order. Therefore, these de novo–assembled partial plastomes are not suitable for studying molecular evolution; rather, these sequences are assembled to a level where homologous plastid sequences can be recovered from all samples for use in downstream phylogenetic analyses. It should be noted that a closely related (or not so closely related, see Straub et al. [2012] for a discussion) published reference sequence could be used instead of this de novo method. The de novo method is presented here as an efficient way to generate a reference sequence for nonmodel organisms and was employed in this study due to lack of close references.

A de novo assembly of each read pool was also performed using the Geneious Velvet plugin (version 1.2.10; Zerbino and Birney, 2008) with a k-mer length of 37 (the result of Velvet Optimizer), a minimum contig length of 74, and default settings.

***Identification of LCNG*—**To identify the presence of LCNG gene sets in genome skimming data, which may have utility in future studies of *Oreocarya* and Amsinckiinae, the following published gene sets were obtained: (1) conserved orthologous set (COS) (Fulton et al., 2002), (2) single-copy conserved orthologous genes (COSII) (Wu et al., 2006), and (3) shared single-copy genes (SSC) (Duarte et al., 2010). The set of 1006 COS and 2592 COSII genes were downloaded from the Sol Genomics Network (SGN; http://solgenomics.net/), and the set of 959 SSC shared among *Arabidopsis* (DC.) Heynh., *Populus* L., *Vitis* L., and *Oryza* L. were downloaded from GenBank (Benson et al., 2013) and made into a custom database in Geneious. The Velvet de novo–assembled contigs were MegaBLAST searched against these LCNG sets using the settings above.

***Ribosomal cistron assembly*—**To make a reference sequence for the ribosomal cistron, a 483-bp sequence from *Oreocarya humilis* Greene (JQ513418) with a complete 5.8S gene and a partial sequence of both internal transcribed spacer regions (ITS1 and ITS2) was obtained from GenBank; this was the only taxon with a sequence from the ribosomal cistron that was both in this study and on GenBank. A reference-guided assembly of the *O. humilis* read pool was

Downloaded From: https://bioone.org/journals/Applications-in-Plant-Sciences on 29 Mar 2024
Terms of Use: https://bioone.org/terms-of-use

Downloaded From: https://bioone.org/journals/Applications-in-Plant-Sciences on 29 Mar 2024
Terms of Use: https://bioone.org/terms-of-use

TABLE 1. Taxa sampled in this study, preservation type, raw and post–quality control read numbers, sequencing depth, and library content for each genome region.

Taxon (population identifier)	Accession no.	Preservation type	Raw read no.	Read no. post-QC	Reads retained (%)	nrDNA depth	% nrDNA	cpDNA depth	% cpDNA	mtDNA depth	% mtDNA
Cryptantha maritima[a]	SDSU 20050	H	3,515,019	2,366,721	67.33	396.6	1.10	246.2	12.99	28.5	4.62
Cryptantha muricata var. *muricata*[a]	SDSU 19537	H	2,553,590	2,219,773	86.93	233.6	0.68	109.2	6.08	12.9	2.12
Cryptantha torreyana var. *torreyana*	SDSU 20124	S	3,634,668	2,922,047	80.39	530.3	1.31	227.7	10.27	31.7	4.46
Dasynotus daubenmirei	SDSU 20343	H	2,271,537	2,037,136	89.68	535.5	1.69	67.4	4.05	10.8	1.90
Eremocarya micrantha	SDSU 18956	H	3,281,361	2,627,734	80.10	334.3	0.92	85.3	4.21	18.1	2.75
Oreocarya celosioides	SDSU 20113	S	5,549,018	4,520,946	81.50	639.9	1.01	205.2	5.96	40.8	3.72
Oreocarya crymophila	SDSU 20116	S	3,662,963	2,902,007	79.20	434.5	1.06	199.8	9.05	56.1	8.04
Oreocarya flavoculata	SDSU 20030	S	9,697,938	7,593,640	78.30	1129.1	1.08	329.6	5.74	50	2.74
Oreocarya hoffmannii	SDSU 20036	S	5,339,723	4,250,549	79.60	661.3	1.12	245.1	7.60	50.4	4.92
Oreocarya humilis subsp. *humilis*	SDSU 20029	S	4,420,140	3,463,693	78.40	609.4	1.26	173.7	6.57	28.3	3.33
Oreocarya hypsophila	SDSU 20086	S	4,256,094	3,304,770	77.60	548.1	1.18	239.9	9.57	36.2	4.52
Oreocarya nubigena ("Horseshoe Meadows")	SDSU 20004	S	9,197,464	6,866,170	74.70	1135.8	1.20	496	9.64	102.8	6.36
Oreocarya nubigena ("Mammoth Mtn.")	SDSU 20055	S	1,760,164	1,423,937	80.90	338.8	1.69	142	13.02	17.8	4.99
Oreocarya nubigena ("Mono Craters")	SDSU 20094	S	5,054,148	4,050,809	80.10	446.9	0.79	131.6	4.24	35.9	3.64
Oreocarya nubigena ("Sierran granite")	SDSU 20079	S	5,452,126	4,490,708	82.40	450.7	0.72	156.5	4.56	41.9	3.85
Oreocarya nubigena ("Sonora Pass")	SDSU 20098	S	3,020,887	2,427,679	80.40	336.1	0.98	187.1	10.11	32.8	5.56
Oreocarya schoolcraftii	SDSU 20123	S	6,139,787	5,000,894	81.50	586.3	0.84	219.3	5.76	52.7	4.38
Oreocarya setosissima	SDSU 20242	H	5,280,735	4,093,537	77.50	558.3	0.95	95.9	3.05	65.1	6.66
Oreocarya sobolifera	SDSU 20210	H	4,256,120	3,495,577	82.10	546	1.11	149.6	5.60	29.3	3.45
Oreocarya subretusa ("Mt. Eddy")	SDSU 20232	H	5,112,245	3,974,139	77.70	529.6	0.99	180.5	5.95	63.6	6.72
Oreocarya subretusa ("type location")	SDSU 20107	S	2,774,358	2,224,590	80.20	346.8	1.11	94.9	5.55	17.3	3.10
Oreocarya subretusa ("Warner Mts.")	SDSU 20110	S	3,096,293	2,463,930	79.60	291.9	0.84	202.5	10.80	42.5	7.13
Oreocarya suffruticosa var. *abortiva*	SDSU 20024	S	9,309,461	7,259,178	78.00	1563.3	1.56	402.8	7.38	75.8	4.40
Oreocarya virgata[a]	SDSU 20117	S	3,001,782	2,342,067	78.02	406.1	1.14	169.1	8.97	37.8	6.21
Pectocarya penicillata	UC 1965571	H	2,151,733	1,741,444	80.90	326.2	1.35	145.6	10.92	17.6	4.02
Maximum			9,697,938	7,593,640	89.68	1,563.30	1.69	496	13.02	102.8	8.04
Minimum			1,760,164	1,423,937	67.33	233.6	0.68	67.4	3.05	10.8	1.90
Mean (±SE)[b]			4,551,574 (±435,943)	3,602,547 (±333,732)	79.72 (±0.79)	556.6 (±60.6)	1.11 (±0.05)	196.1 (±19.6)	7.51 (±0.57)	39.9 (±4.3)	4.54 (±0.32)
IQR			2,318,836	1,883,828	2.88	239.5	0.25	85.7	4.04	22.1	2.10

Note: H = herbarium sheet; IQR = interquartile range; QC = quality control; S = silica gel dried; SE = standard error.

[a] Run 2,1/53 samples, read length 101 bp; all remaining taxa from Run 1, 1/38 samples, read length 100 bp.

[b] Data are not normally distributed but SE is presented to give a general sense of variation around the mean, see also IQR.

implemented in Geneious with medium-low sensitivity, default settings, and 100 iterations. A consensus contig was saved using a 75% masking threshold, and a gap masked areas with coverage under 20× (although Straub et al. [2012] used 25× for a single nucleotide polymorphism [SNP], 5× for a base shared with the reference sequence, and masked with Ns). The resulting sequence was annotated using the "find annotations" feature in Geneious, transferring annotations with a 50% or greater similarity from relatives with annotated sequences on GenBank: *Amsinckia lycopsoides* Lehm. (JQ388495) for 5.8S and the two ITS regions, *Vahlia capensis* (L. f.) Thunb. (AF479182) for the 26S gene, and *Ehretia acuminata* R. Br. (HQ384690) for the 18S gene. A Sanger sequence of external transcribed spacer (ETS) from *Oreocarya confertiflora* Greene (Guilliams and Baldwin, unpublished data) was used to annotate the approximate boundary of ETS. The resulting annotated *O. humilis* cistron was trimmed to exclude the nontranscribed spacer (NTS), a portion of the intergenic spacer (IGS). This was used for the reference-guided assembly of the remaining read pools in Geneious using 25 iterations, medium-low sensitivity, and default settings. A consensus contig was generated for each sample using a 75% threshold. Areas with less than 20× sequence coverage were masked with gaps, and IUPAC ambiguity codes were retained. Sequences were aligned using the Geneious MAFFT plugin (version 7.017; Katoh et al., 2002) with default settings. Alignments were examined for misaligned areas; these were aligned by eye or excluded. Sequence portions that were not represented among all samples, contained gaps, and/or contained ambiguity codes were removed using the "strip alignments" feature in Geneious (Fig. 1).

***Plastome assembly*—**The de novo assembly of *Pectocarya penicillata* (Hook. & Arn.) A. DC. (UC1965571) generated a 124,868-bp partial plastome sequence. This was annotated from the complete plastome of *S. lycopersicum* using the "find annotations" feature in Geneious to transfer annotations at 50% or greater similarity. Annotations were translated in Geneious and examined by eye; problematic annotations were removed. The goal in this study is to generate homologous plastid sequences from each sample, not to generate fully annotated plastomes. Reference-guided assembly to the annotated *P. penicillata* plastome was implemented in Geneious, with default settings and 25 iterations of the read pool from each sample. The methods for generating a cpDNA consensus sequence, sequencing editing, and alignment follow those employed for the nrDNA cistron (Fig. 2).

***Mitochondrial gene assembly*—**Straub et al. (2012) used the longest mtDNA contigs from the de novo assembly for reference-guided assembly of each read pool and subsequent phylogenetic inference. However, plant mitochondria undergo frequent structural rearrangements (Knoop, 2004; Woloszynska, 2010), meaning that genes rather than partial or complete genomes are suitable for phylogenetic inference (Godden et al., 2012; Malé et al., 2014). Preliminary assemblies revealed the mitochondrial content from each sample did not appear to be uniform; introns and intergenic regions were represented among some but not all samples. However, coding regions were consistently recovered among all samples. Plant mtDNA markers have been less used in plant phylogenetic studies (Godden et al., 2012) and are usually assumed to have conservative rates of evolution, although this is not true for all plant genera (Cho et al., 2004; Knoop, 2004). This study presents a Geneious-based method, conceptually similar to Malé et al. (2014) to recover mtDNA exons from genome skimming data. The *Nicotiana tabacum* L. mitochondrion (BA000042) was obtained from GenBank and was modified to include only one copy of each annotated repeat

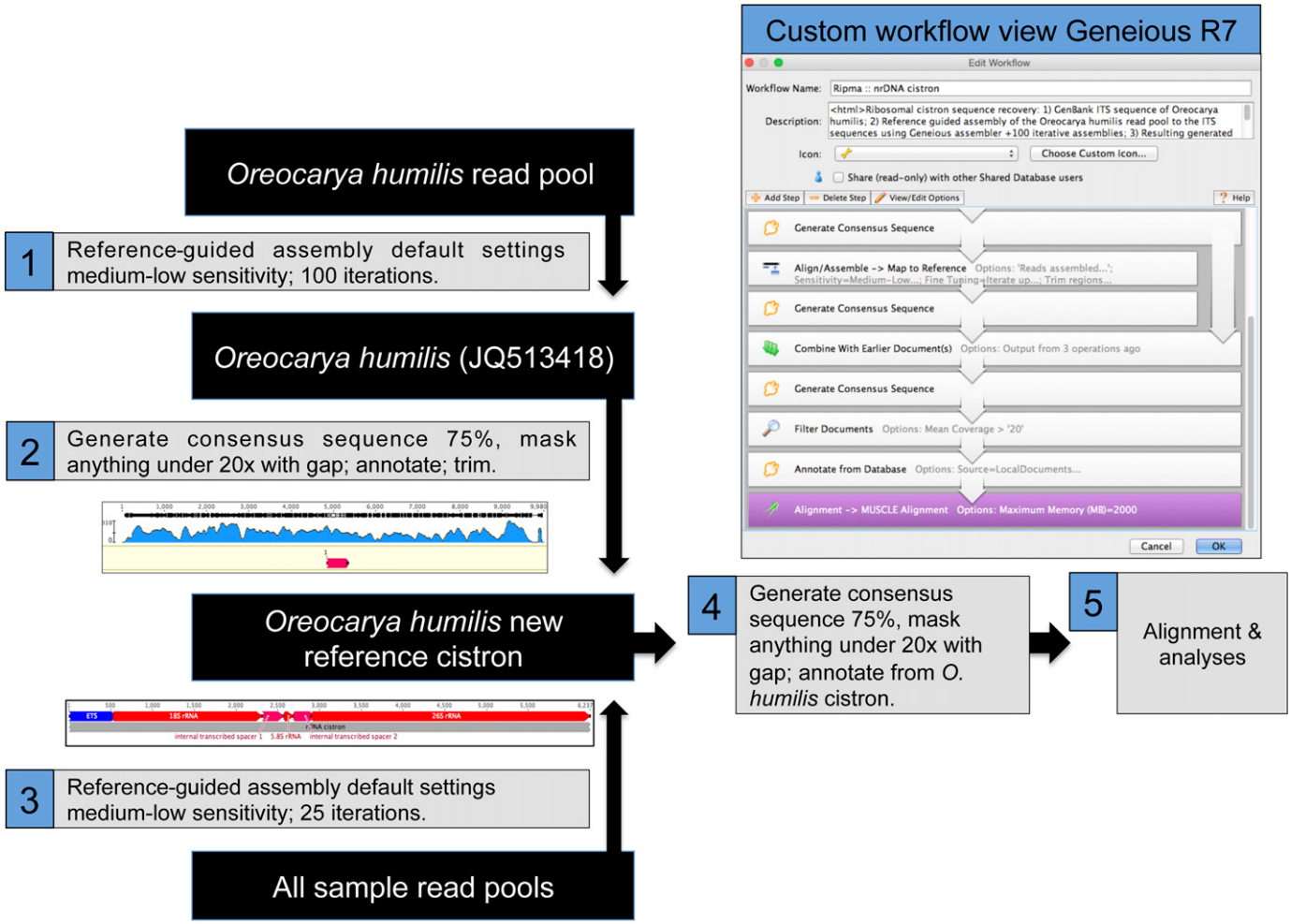


Fig. 1. Illustrated workflow for generating the nuclear ribosomal cistron using Geneious.

Downloaded From: https://bioone.org/journals/Applications-in-Plant-Sciences on 29 Mar 2024
Terms of Use: https://bioone.org/terms-of-use

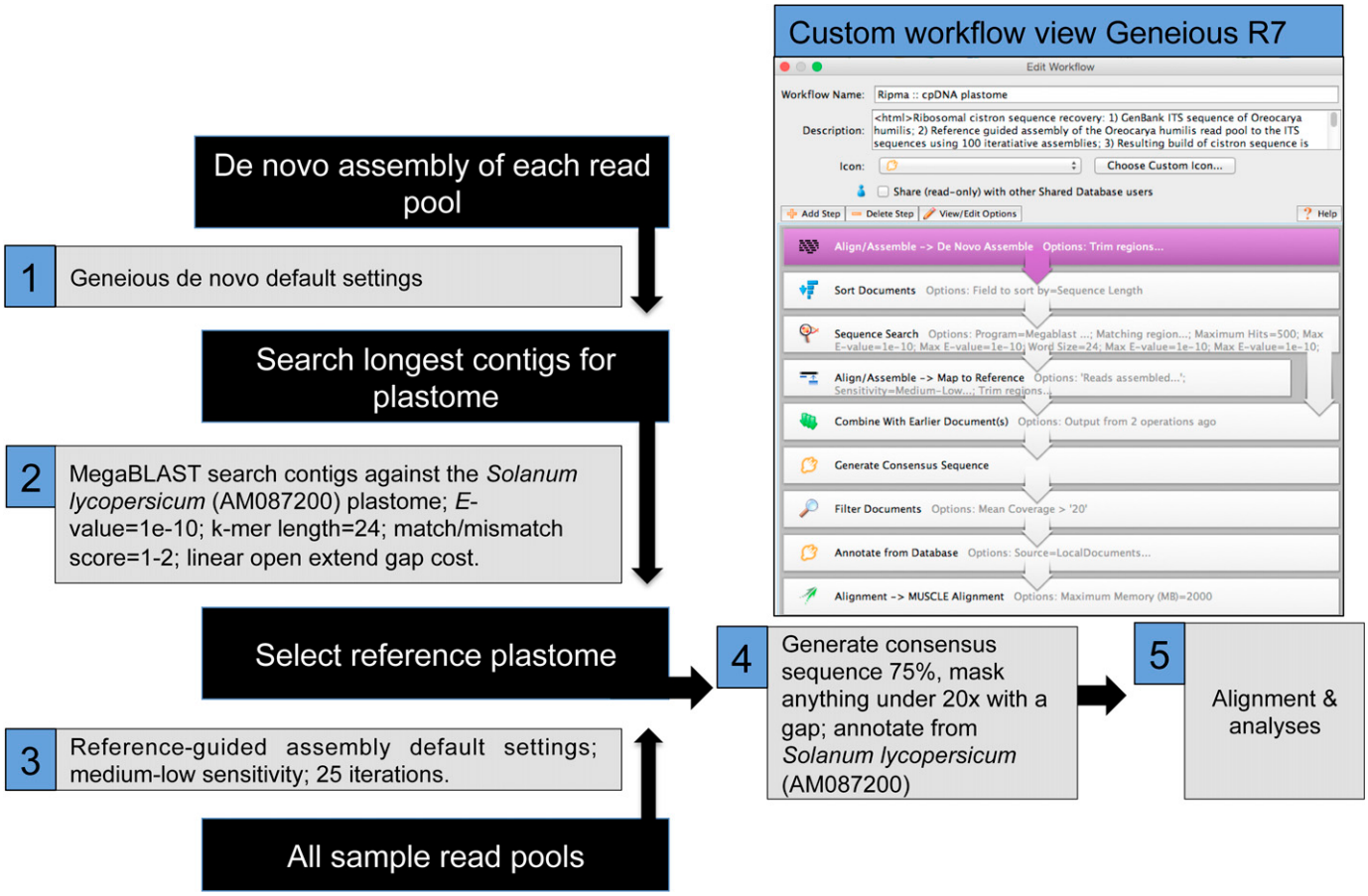


Fig. 2. Illustrated workflow for generating chloroplast DNA using Geneious.

region (final length 396,065 bp). A reference-guided assembly to this modified sequence was performed for each sample read pool. A consensus contig was saved using the same methods presented for the nrDNA cistron. A single contig from each sample was made into a custom database in Geneious, henceforth referred to as the mtDNA contig bin. Each *N. tabacum* exon was separated using the extract annotations feature in Geneious, and these exons were BLASTN searched against *S. lycopersicum* using an *E*-value of 1e-10, a k-mer length of 15, a scoring match-mismatch of 2-3, and a 5-2 open extend gap cost (this BLASTN search is more likely to find matches than the MegaBLAST search used elsewhere). Only *N. tabacum* exons with no match to chloroplast sequences were retained. Each retained exon was MegaBLAST searched against the mtDNA contig bin using a query centric alignment output and the settings for LCNG gene searches. The result was an alignment of each *N. tabacum* exon and the corresponding sequence(s) from each sampled taxon. Only alignments with a single copy from each sample were retained, and several of these exons were partial. Exons were aligned using the MAFFT plugin with default settings. Sequences were edited with the same methods as those used for the nrDNA cistron (Fig. 3).

***Depth of coverage, genomic library content, and PICs*—**To calculate mean depth of coverage for each genomic target, the number of nucleotides that mapped to the reference sequence was divided by the total length of the reference sequence. Library genomic content was calculated by dividing the number of reads that mapped to the reference nrDNA, cpDNA, and modified mtDNA sequence by the total number of reads in each sample pool (Straub et al., 2012). PICs were calculated by analyzing each final alignment in the Geneious GARLI plugin (version 2.0; Zwickl, 2006); the "info tab" displays variable characters and PICs.

***Multiplexing level*—**Straub et al. (2012) presents the following formula to calculate multiplexing level: ML = (LC*CF*PTG)/(CD*TaG), where ML = multiplex level possible, LC = lane capacity of the sequencing instrument in base pairs, CF = correction for reads lost to quality control and adapters, PTG = proportion of reads mapping to the target genome (i.e., the library content for the target genome), CD = coverage depth desired (e.g., 30×), and TaG = length in base pairs of the target genome. This formula was used to calculate multiplexing levels if future samples contained both the mean and minimum cpDNA library content values as Run 1 and Run 2. Values are based on the cpDNA as the genomic target, because a sufficient sequencing depth for the cpDNA will recover both the nrDNA cistron and many mtDNA exons. This calculation is of paramount importance in making genome skimming affordable and is widely applicable to other study systems due to the conserved length of plant plastomes; it can easily be adjusted to the particulars of any NGS run.

***Phylogenetic analyses*—**Sequences were analyzed using maximum likelihood (ML) in the RAxML (Stamatakis, 2006) Geneious plugin with a GTR+GAMMA model of nucleotide evolution. Each genome was analyzed separately, with the nrDNA partitioned by gene (ETS, 18S, ITS1, 5.8S, ITS2, and 26S), the mtDNA partitioned into 12 exons, and the cpDNA unpartitioned. A concatenated analysis was performed of all data with the partitions above. All analyses were run with *P. penicillata* set as the outgroup based on the results of Hasenstab-Lehman and Simpson (2012). To assess support, 10,000 rapid bootstrap (BS) replicates were done for every analysis, with clades having a BS value of 70 or greater considered highly supported (Stamatakis et al., 2008). The topology with the highest ML from each genome was analyzed using the species tree program STAR (Liu et al., 2009) on the STRAW server (Shaw et al., 2013); STAR uses the topology of individual gene trees to generate a species tree. The 10,000 BS trees from each genome were used to assess support for the STAR species tree. Resulting trees were viewed in Geneious and formatted in Adobe Illustrator CS (Adobe Systems, San Jose, California, USA).

Downloaded From: https://bioone.org/journals/Applications-in-Plant-Sciences on 29 Mar 2024
Terms of Use: https://bioone.org/terms-of-use

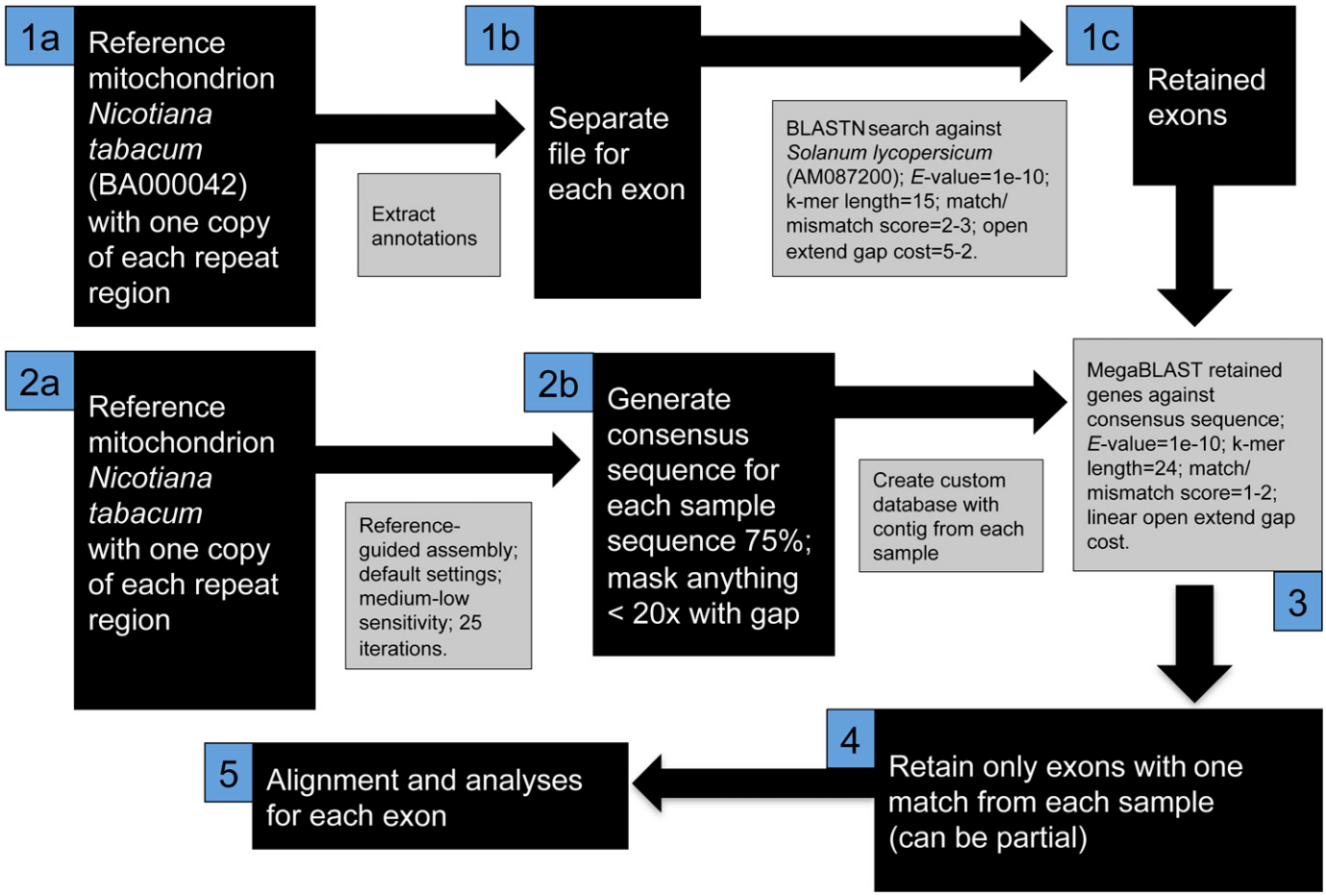


Fig. 3. Illustrated workflow for generating mitochondrial DNA exons using Geneious.

RESULTS

***DNA sequencing and quality control filtering*—**Run 1 on the Illumina HiSeq 2000 lane resulted in 154,126,153 reads from 38 samples (102,447,426 from the 21 samples in this study). Run 2 resulted in 170,464,908 reads from 53 samples (11,341,928 from the four samples in this study). Samples returned between 1,760,164 and 9,697,938 reads, for a mean read number of 4,441,574 (±SE 435,943). Reads retained per sample following quality control were between 1,423,937 and 7,593,640 with a mean post–quality control read pool number of 3,602,547 (±SE 333,732). Reads retained ranged from 67.33% to 89.68%, and detailed results are presented in Table 1.

***De novo assembly and identification of LCNG*—**The Geneious de novo assembly resulted in many partial plastome contigs. The longest was a 124,868-bp sequence from *P. penicillata*; this was used for reference-guided assembly of all the read pools. The Velvet de novo contigs contained MegaBLAST matches to a total of 552 LCNG (38 COS, 461 COSII, and 53 SCC). The LCNG names, number of hits, and length of the longest hit are presented in Appendix S1. Note that the majority (91%) of hits to LCNG matched only a single Velvet de novo contig. Therefore, LCNG alignments cannot be extracted from genome skimming data for use in phylogenetic analysis (as they are for mtDNA exons); rather this is a tool for identifying LCNG present in the sampled organisms.

***Ribosomal cistron assembly, depth of coverage, and library content*—**Total nrDNA cistron sequencing depths were between 233.6× and 1563.3×, with a mean depth of 556.6× (±SE 60.6×). The total amount of nrDNA present in the samples was between 0.68% and 1.69%, with a mean of 1.11% (±SE 0.05%) (Table 1). A cistron sequence from each sample was recovered, with a total aligned length of 6418, reduced to 5866 without gaps and ambiguities. The whole data set contained 2.56% PICs, while the ingroup (*Oreocarya* only) contained 0.32% PICs (Table 2).

***Plastome assembly, depth of coverage, and library content*—**Total cpDNA sequencing depths were between 67.4× and 496.0×, with a mean depth of 196.1× (±SE 19.6×). The total amount of cpDNA present in the samples was between 3.05% and 13.02%, with a mean of 7.51% (±SE 0.57%) (Table 1). A plastome sequence from each sample was recovered, with a total aligned length of 130,148, reduced to 115,745 without gaps and ambiguities. The whole data set contained 0.48% PICs while the ingroup contained 0.09% PICs (Table 2).

***Mitochondrial exon assembly, depth of coverage, and library content*—**Total mtDNA sequencing depths were between 10.8× and 102.8×, with a mean depth of 39.9× (±SE 4.3×). The total amount of mtDNA present in the samples was between 1.90% and 8.04%, with a mean of 4.54% (±SE 0.32%) (Table 1). A total of 12 mtDNA exons were recovered, with a total aligned

Downloaded From: https://bioone.org/journals/Applications-in-Plant-Sciences on 29 Mar 2024
Terms of Use: https://bioone.org/terms-of-use

TABLE 2. Final aligned sequence length excluding gaps and ambiguities, variable characters, and parsimony informative characters for the nuclear ribosomal DNA, chloroplast DNA, and mitochondrial DNA.

Genome region	Aligned sequence length	All taxa: variable characters	All taxa: PICs	All taxa: % PICs	*Oreocarya*: variable characters	*Oreocarya*: PICs	*Oreocarya*: % PICs
Nuclear ribosomal DNA	5866	320	150	2.56	68	19	0.32
Chloroplast DNA	115,745	4101	556	0.48	586	104	0.09
Mitochondrial exons	2661	1049	505	18.98	919	407	15.30
Total	124,272	5470	1211	0.97	1573	530	0.43

Note: PICs = Parsimony informative characters.

length of 4978, reduced to 2661 without gaps and ambiguities. The mtDNA gene set contained 18.98% PICs while the ingroup contained 15.3% PICs (Table 2).

Multiplexing level—The *P. penicillata* plastome from the de novo assembly, with one copy of the inverted repeat region (IRR), was used as the genomic target to calculate future multiplexing capacity with a target depth of 30× (Straub et al., 2012). If future samples return the same mean as Run 1, 245 samples could be multiplexed in a lane; if future samples return the same minimum value, 94 could be multiplexed in a lane. For Run 2 the mean multiplexing level was 293 and the minimum was 124 (Table 3).

Phylogenetic analyses—Cladograms for each genome region, the concatenated data set, and a coalescent-based analysis are presented in Fig. 4A–E; phylograms (inset) were transformed into cladograms so that relationships among taxa are visible, as *Oreocarya* has very short branch lengths. In all cladograms the monophyly of *Oreocarya* is strongly supported; relationships within *Oreocarya* with no resolution using Sanger sequencing are resolved with strong support, discussed below. Multiple samples of the same taxon (*O. nubigena* Greene and *O. subretusa* (I. M. Johnst.) Abrams) were not monophyletic. The STAR species tree (Fig. 4E) shows topological incongruence among the three gene trees, and incongruence is also present between the species tree and the concatenated tree.

DISCUSSION

This study achieves Goal 1, to develop and present user-friendly methods for processing genome skimming data without the use of complex bioinformatics programs. The study demonstrates that reads from a genome skim can be assembled into nrDNA, cpDNA, and mtDNA sequences to a level suitable for phylogenetic inference solely using Geneious. It should be noted that Geneious is a proprietary program that currently (October 2014) costs US$395 for a student license and US$795 for a noncommercial license. Free programs can be used piecemeal to achieve the same results Geneious offers in a complete software package. We feel that Geneious greatly simplifies file formatting, phylogenetic analyses, sequence queries, and GenBank submission (to name a few). The custom database feature is a powerful and easy-to-use search tool, which was instrumental in this study.

Methods presented here are largely congruent with Straub et al. (2011, 2012), Bock et al. (2014), and Malé et al. (2014), albeit in a more user-friendly interface. A key difference is that Straub et al. (2012) and Bock et al. (2014) used large fragments of mtDNA for phylogenetic inference that included introns and intergenic regions, while Malé et al. (2014) and this study inferred relationships using only coding mtDNA sequences. The mtDNA exons presented in this study contain higher levels of PICs than the other genomes, but before concluding that there are elevated levels of mitochondrial evolution in *Oreocarya* (demonstrated for *Plantago* L. and *Pelargonium* L'Hér. ex Aiton in Cho et al. [2004]) primers should be designed to ensure that PCR sequences match the in silico results (e.g., Straub et al., 2011).

Goal 2 in this study was to examine the feasibility of genome skimming for future studies of *Oreocarya* and Amsinckiinae. The phylogenetic relationships presented here show more resolution in *Oreocarya* than in any study to date. The nearly complete nrDNA and cpDNA sequences reveal very low levels of PICs in *Oreocarya* (0.32% and 0.09%, respectively). These results explain the polytomies in other studies using traditional methods (Marushak, 2003; Bresowar and McGlaughlin, 2011; Hasenstab-Lehman and Simpson, 2012). Although the sequence variation is low, 7/17 within-ingroup nodes are resolved with BS support >70 in the cpDNA, while only 3/17 nodes are resolved in the nrDNA. The mtDNA sequences contain more

TABLE 3. Multiplexing calculations for the mean and minimum CF and PTG values from Run 1 and Run 2 when the genomic target is the plastome with one copy of the inverted repeat region and a sequencing depth of 30×.

Parameters	Equation abbreviation	Run 1 mean	Run 1 minimum	Run 2 mean	Run 2 minimum
Read length (bp)		100	100	101	101
Total reads generated in a lane		156,000,000	156,000,000	170,464,908	170,464,908
Lane capacity (nucleotides)	LC	15,600,000,000	15,600,000,000	17,046,490,800	17,046,490,800
Reads passing quality filters	CF	0.7960	0.7470	0.8049	0.6733
Proportion mapping to genomic target	PTG	0.0741	0.0305	0.0802	0.0405
Coverage depth desired	CD	30	30	30	30
Target genome size	TaG	124,868	124,868	124,868	124,868
Multiplexing possible	ML	245.5	94.7	293.8	124.2

Downloaded From: https://bioone.org/journals/Applications-in-Plant-Sciences on 29 Mar 2024
Terms of Use: https://bioone.org/terms-of-use

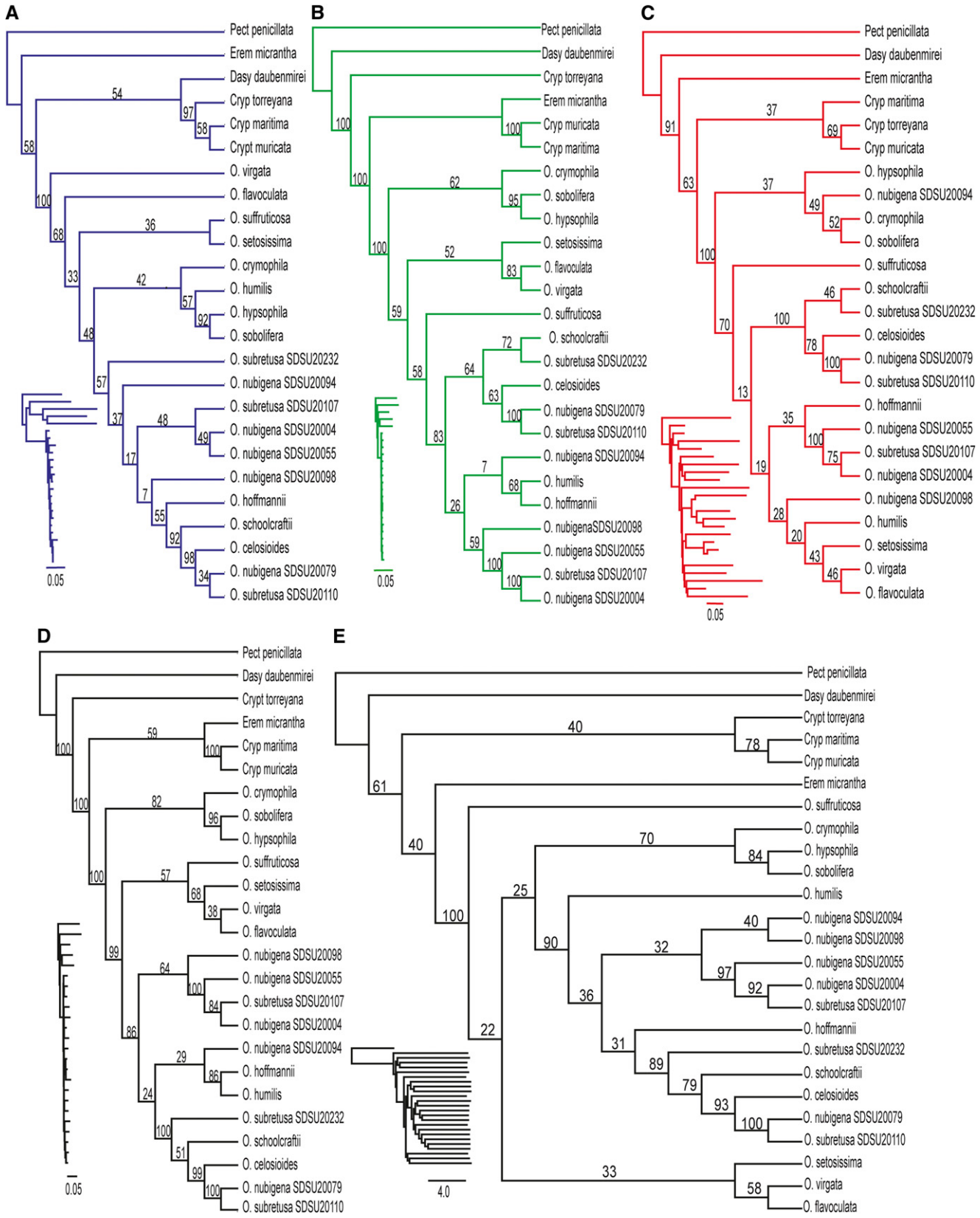


Fig. 4. The results of maximum likelihood RAxML phylogenetic analyses for 19 *Oreocarya* and six outgroups obtained from the nuclear ribosomal DNA (panel A), the chloroplast DNA (panel B), all mitochondrial genes (panel C), all data concatenated (panel D), and a STAR species tree (panel E). The support values from 10,000 bootstrap replicates are displayed. Cladograms are shown so relationships are visible, with inset phylograms to show branch lengths.

Downloaded From: https://bioone.org/journals/Applications-in-Plant-Sciences on 29 Mar 2024
Terms of Use: https://bioone.org/terms-of-use

PICs but only 6/17 nodes are resolved. Both the concatenated tree and the species tree have more resolved nodes than the individual analyses, 10/17 and 9/17 resolved nodes, respectively. The genome skimming method recovers loci from three separate genomes, two of these are uniparentally inherited, and one that could obscure phylogenetic signal because it is known to occur in arrays across nonhomologous chromosomes (Baldwin et al., 1995). Issues with the ribosomal cistron (especially ITS) are thoroughly reviewed elsewhere (Alvarez and Wendel, 2003); however, for many plant groups it remains an accessible phylogenetic tool. Genome skimming generates the entire ribosomal cistron, which can aid phylogenetic inference in closely related groups of plants due to the enhanced rate of nucleotide substitution in the ITS1, ITS2, and ETS regions (Baldwin et al., 1995; Baldwin and Markos, 1998) in a more cost-effective manner than Sanger sequencing of these regions separately. Our phylogenetic inferences recover conflicting topologies among all gene trees and a STAR analysis is presented. Although STAR may not be an appropriate method to combine only three gene trees, this analysis demonstrates that a coalescent-based approach is possible with genome skimming products.

Genome skimming cannot produce a large data set of orthologous nuclear genes, which are necessary as the trend in phylogenetics moves toward coalescent-based analyses. Attempts were made to find previously unpublished orthologous nuclear sequences within contigs generated from a genome skim, but nuclear genes recovered in the fragment data were represented among too few of the samples to be of use in phylogenetic reconstruction, and paralogy could not be determined at such a low sequencing depth of the nuclear genome. However, Hyb-Seq probes can be designed using nuclear genes identified in the fragments generated from a genome skim (Straub et al., 2011; Godden et al., 2012; Cronn et al., 2012; Grover et al., 2012; Lemmon and Lemmon, 2013).

One of the more valuable aspects of genome skimming is that libraries can be prepared with dried samples, and although no formal comparison was made between preservation types (herbarium sheet vs. silica dried), samples with both preservation types generated nearly complete nrDNA, cpDNA, and 12 mtDNA exons. This result demonstrates that herbarium sheets are a viable way to extract gDNA for genome skimming library preparation. This is important for the future study of Amsinckiinae, as preservation of fresh material is difficult when taxa are spread throughout remote areas of North and South America. The sequence variability within the limited samples of Amsinckiinae was much higher than that of *Oreocarya* alone (0.97% vs. 0.43%; Table 2). Genome skimming is now being used to collect sequence data for a larger study of the Amsinckiinae.

Our study revealed that even using the most conservative estimates, 94 samples can be multiplexed using single-end 100-bp reads (Table 3). The Straub et al. (2012) formula can be changed to reflect the particulars of an individual study and will reduce costs in the Amsinckiinae study. At the commencement of this study, barcoding kits were limited to 96 samples, but now barcodes for 384 samples (NuGEN Technologies, San Carlos, California, USA) and even 480 samples (Fluidigm, South San Francisco, California, USA) are available. Equipment startup costs for the preparation of gDNA libraries "in-house" can be expensive, and this study was made possible by outsourcing the library preparation to Global Biologics, who charged US$100 per sample for gDNA library preparation (100-bp reads) at the time of this study (February 2013), but costs have been reduced by nearly 30% over the past year. Outsourcing library preparation has a low startup cost and is available to systematics laboratories with limited resources. At the time of this study, an Illumina HiSeq 2000 lane (single-end reads) cost US$1500–$2000, with costs decreasing rapidly. Genome skimming generates the same product as the study by Stull et al. (2013) using library enrichment and massive multiplexing to generate high sequencing depth for target chloroplasts. However, gDNA extraction and library preparation for genome skimming are more straightforward and less expensive.

Few authors discuss standard methods to ensure genome-scale sequence editing and alignments are not misleading phylogenetic inference (although see Parks et al., 2012). Sequence editing in this study was conservative; any location with an ambiguity code or gap was excluded from analyses, simplified by the brilliant "strip alignments" feature in Geneious. Strict sequence editing resulted in the loss of PICs in sequences already plagued by low variability. Although concerted evolution is believed to homogenize nrDNA copies, multiple copies are evident when reads are mapped to the cistron (see Straub et al. [2012] for polymorphism levels), and our conservative methods excluded "polymorphic" sites in the nrDNA altogether.

As higher-level relationships among angiosperms are resolved, plant systematists will increasingly work at lower taxonomic levels (Soltis et al., 2011; Godden et al., 2012). Methods for elucidating finer relationships present challenges that are well illustrated in the genus *Oreocarya*. In addition to low PICs, multiple samples from the same taxon were recovered as nonmonophyletic, a result consistent with the findings of Straub et al. (2012) in multiple samples of *Asclepias* L. Coalescent theory predicts that the gene trees will fail to be reciprocally monophyletic in a rapid species radiation (Maddison, 1997; Kubatko and Degnan, 2007; Edwards, 2009), which could be the case in *Oreocarya*. The methods presented in this study will aid in the future systematic study of both *Oreocarya* and the Amsinckiinae and demonstrate the value of genome skimming in a group with few genomic resources. In addition to achieving the goals of our study and providing a valuable application of genome skimming, we conclude that if the objective is to infer a phylogeny using plastome and cistron data, then genome skimming is a less expensive and more efficient option than PCR+Sanger sequencing of several gene regions.

LITERATURE CITED

Alvarez, I., and J. F. Wendel. 2003. Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution* 29: 417–434.

Baldwin, B. G., M. J. Sanderson, J. M. Porter, M. F. Wojciechowski, C. S. Campbell, and M. J. Donoghue. 1995. The ITS region of nuclear ribosomal DNA: A valuable source of evidence on angiosperm phylogeny. *Annals of the Missouri Botanical Garden* 82: 247–277.

Baldwin, B. G., and S. Markos. 1998. Phylogenetic utility of the external transcribed spacer (ETS) of 18S–26S rDNA: Congruence of ETS and ITS trees of *Calycadenia* (Compositae). *Molecular Phylogenetics and Evolution* 10: 449–463.

Benson, D. A., M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. 2013. GenBank. *Nucleic Acids Research* 41: D36–D42.

Bock, D. G., N. C. Kane, D. P. Ebert, and L. H. Rieseberg. 2014. Genome skimming reveals the origin of the Jerusalem Artichoke tuber crop species: Neither from Jerusalem nor an artichoke. *New Phytologist* 201: 1021–1030.

Downloaded From: https://bioone.org/journals/Applications-in-Plant-Sciences on 29 Mar 2024
Terms of Use: https://bioone.org/terms-of-use

Brand, A. 1931. Borraginaceae-Borraginoideae-Amsinckieae. *In* A. Engler [ed.], Das Pflanzenreich, 204. Verlag von Wilhelm Engelmann, Leipzig, Germany.

Bresowar, G. E., and M. E. McGlaughlin. 2011. Phylogenetics of the genus *Cryptantha* subgenus *Oreocarya* (Boraginaceae): A Western North American endemic taxon. Botany 2011: A joint meeting of the American Fern Society, American Society of Plant Taxonomists, the Botanical Society of America and the Society for Economic Botany, St. Louis, Missouri, USA [online abstract]. Website http://2011.botanyconference.org/engine/search/index.php?func=detail&aid=296 [accessed 17 November 2014].

Bresowar, G. E., and M. E. McGlaughlin. 2014. Characterization of microsatellite markers isolated from members of *Oreocarya* (Boraginacae). *Conservation Genetics Resources* 6: 205–220.

Cho, Y., J. P. Mower, Y. L. Qiu, and J. D. Palmer. 2004. Mitochondrial substitution rates are extraordinarily elevated and variable in a genus of flowering plants. *Proceedings of the American Academy of Arts and Sciences* 101: 17741–17746.

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.

Doyle, J. J., and J. L. Doyle. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* 19: 11–15.

Duarte, J. M., P. K. Wall, P. P. Edger, L. L. Landherr, H. Ma, J. C. Pires, J. Leebens-Mack, and C. W. dePamphilis. 2010. Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evolutionary Biology* 10: 61.

Edwards, S. V. 2009. Is a new and general theory of molecular systematics emerging? *Evolution; International Journal of Organic Evolution* 63: 1–19.

Friar, E. A. 2005. Isolation of DNA from plants with large amounts of secondary metabolites. *In* E. A. Zimmer and E. H. Roalson [eds.], Methods in enzymology, vol. 395, Producing the biochemical data, Part B, 3–14. Academic Press, San Diego, California, USA.

Fulton, T., R. Van Der Hoeven, N. Eannetta, and S. Tanksley. 2002. Identification, analysis and utilization of a conserved ortholog set (COS) markers for comparative genomics in higher plants. *Plant Cell* 14: 1457–1467.

Godden, G. T., I. E. Jordon-Thaden, S. Chamala, A. A. Crowl, N. García, C. C. Germain-Aubrey, J. M. Heaney, et al. 2012. Making next-generation sequencing work for you: Approaches and practical considerations for marker development and phylogenetics. *Plant Ecology & Diversity* 5: 427–450.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 29: 644–652.

Grover, C. E., A. Salmon, and J. F. Wendel. 2012. Targeted sequence capture as a powerful tool for evolutionary analysis. *American Journal of Botany* 99: 312–319.

Hasenstab-Lehman, K. E., and M. G. Simpson. 2012. Cat's eyes and popcorn flowers: Phylogenetic systematics of the genus *Cryptantha* s. l. (Boraginaceae). *Systematic Botany* 37: 738–757.

Higgins, L. C. 1971. A revision of *Cryptantha* subgenus *Oreocarya*. *Brigham Young University Science Bulletin. Biological Series* 8: 1–62.

Kane, N., S. Sveinsson, H. Dempewolf, J. Y. Yang, D. Zhang, J. M. M. Engels, and Q. Cronk. 2012. Ultra-barcoding in cacao (*Theobroma* spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. *American Journal of Botany* 99: 320–329.

Katoh, K., K. Misawa, K. Kuma, and T. Miyata. 2002. MAFFT: A novel method for rapid multiple sequence alignment based on a fast Fourier transformation. *Nucleic Acids Research* 30: 3059–3066.

Knoop, V. 2004. The mitochondrial DNA of land plants: Peculiarities in a phylogenetic perspective. *Current Genetics* 46: 123–139.

Kubatko, L. S., and J. H. Degnan. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology* 56: 17–24.

Lemmon, M. E., and A. R. Lemmon. 2013. High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology Evolution and Systematics* 44: 99–121.

Liston, A. 2012. Introduction to next-generation sequencing. Botany 2012: Annual Meeting of the Botanical Society of America, Columbus, Ohio, USA [online abstract]. Website http://2012.botanyconference.org/engine/search/index.php?func=detail&aid=49 [accessed 17 November 2014].

Liu, L., L. Yu, D. K. Pearl, and S. V. Edwards. 2009. Estimating species phylogenies using coalescence times among sequences. *Systematic Biology* 58: 468–477.

Maddison, W. P. 1997. Gene trees in species trees. *Systematic Biology* 46: 523–536.

Malé, P. G., L. Bardon, G. Besnard, E. Coissac, F. Delsuc, J. Engel, E. Lhuillier, et al. 2014. Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. *Molecular Ecology Resources* 14: 966–975.

Marushak, T. M. 2003. Patterns of mating system evolution in *Cryptantha* section *Oreocarya* (Boraginaceae): A phylogenetic approach. Ph.D. dissertation, University of Maryland, College Park, Maryland, USA.

Nazaire, M., and L. Hufford. 2012. A broad phylogenetic analysis of Boraginaceae: Implications for the relationships of *Mertensia*. *Systematic Botany* 37: 758–783.

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.

Ripma, L. A., M. G. Simpson, and K. Hasenstab-Lehman. 2014. Data from: Geneious! Simplified genome skimming methods for phylogenetic systematic studies: A case study in *Oreocarya* (Boraginaceae). Dryad Digital Repository. http://doi.org/10.5061/dryad.50536.

Schmieder, R., and R. Edwards. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics (Oxford, England)* 27: 863–864.

Shaw, T., Z. Ruan, T. Glenn, and L. Liu. 2013. STRAW: A web server for species tree analysis. *Nucleic Acids Research* 41: W238–W241.

Soltis, D. E., S. A. Smith, N. Cellinese, K. J. Wurdack, D. C. Tank, S. F. Brockington, N. F. Refulio-Rodriguez, et al. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. *American Journal of Botany* 98: 704–730.

Soltis, D. E., M. A. Gitzendanner, G. Stull, M. Chester, A. Chanderbali, S. Chamala, I. Jordon-Thaden, et al. 2013. The potential of genomics in plant systematics. *Taxon* 62: 886–898.

Stamatakis, A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics (Oxford, England)* 22: 2688–2690.

Stamatakis, A., P. Hoover, and J. Rougemont. 2008. A fast bootstrapping algorithm for the RAxML web-servers. *Systematic Biology* 57: 758–771.

Straub, S. 2012. Botany 2012: Introduction to next generation sequencing workshop practical exercises. Website http://milkweedgenome.org/sites/default/files/workshopFiles/Botany_2012_NGS_workshop_exercises_0.pdf [accessed 1 August 2012].

Straub, S. C. K., M. Fishbein, T. Livshultz, Z. Foster, M. Parks, K. Weitemier, R. Cronn, and A. Liston. 2011. Building a model: Developing genomic resources for common milkweed (*Asclepias syriaca*) with low coverage genome sequencing. *BMC Genomics* 12: 211.

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

Straub, S. C. K., R. C. Cronn, C. Edwards, M. Fishbein, and A. Liston. 2013. Horizontal transfer of DNA from the mitochondrial to the plastid genome and its subsequent evolution in milkweeds (Apocynaceae). *Genome Biology and Evolution* 5: 1872–1885.

Stull, G. W., M. J. Moore, V. S. Mandala, N. A. Douglas, H.-R. Kates, X. Qi, S. F. Brockington, et al. 2013. A targeted enrichment strategy

Downloaded From: https://bioone.org/journals/Applications-in-Plant-Sciences on 29 Mar 2024
Terms of Use: https://bioone.org/terms-of-use

for massively parallel sequencing of angiosperm plastid genomes. *Applications in Plant Sciences* 1(2): 1200497.

Weigend, M., F. Leubert, F. Selvi, G. Brokamp, and H. H. Hilger. 2013. Multiple origins for Hound's tongues (*Cynoglossum* L.) and Navel seeds (*Omphalodes* Mill): The phylogeny of the borage family (Boraginaceae s.str.). *Molecular Phylogenetics and Evolution* 68: 604–618.

Weitemier, K., R. C. Cronn, M. Fishbein, R. Schmick, A. Mcdonnell, and A. Liston. 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences* 2(9): 1400042.

Wen, J., Z. Xiong, Z.-L. Nie, L. Mao, Y. Zhu, X.-Z. Kan, S. M. Ickert-Bond, et al. 2013. Transcriptome sequences resolve deep relationships of the grape family. *PLoS ONE* 8: e74394.

Whittall, J. B., S. 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: 100–114.

Woloszynska, M. 2010. Heteroplasmy and stoichiometric complexity of plant mitochondrial genomes-though this be madness, yet there's method in't. *Journal of Experimental Botany* 61: 657–671.

Wu, F., L. A. Mueller, D. Crouzillat, V. Petiard, and S. D. Tanksley. 2006. Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: A test case in the euasterid plant clade. *Genetics* 174: 1407–1420.

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

Zwickl, D. J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. dissertation, The University of Texas, Austin, Texas, USA.

Downloaded From: https://bioone.org/journals/Applications-in-Plant-Sciences on 29 Mar 2024
Terms of Use: https://bioone.org/terms-of-use

APPENDIX 1. Voucher information for collections used in this study.[a]

Taxon (population identifier)	Accession no.	Collection no.	Geographic coordinates	GenBank accession no. (nrDNA)	GenBank accession no. (cpDNA)
Cryptantha maritima (Greene) Greene	SDSU 20050	Simpson 3665	32.94626, −116.29714	KM213400	KP096536
Cryptantha muricata (Hook. & Arn.) A. Nelson & J. F. Macbr. var. *muricata*	SDSU 19537	Simpson 3142	34.50633, −119.87112	KM213423	KP096534
Cryptantha torreyana (A. Gray) Greene var. *torreyana*	SDSU 20124	Ripma 377	43.4617, −113.56226	KM213409	KP096524
Dasynotus daubenmirei I. M. Johnst.	SDSU 20343	Kelley 1951	46.23503, −115.65855	KM213413	KP096521
Eremocarya micrantha (Torr.) Greene	SDSU 18956	Guilliams 602	31.79778, −110.8091	KM213422	KP096527
Oreocarya celosioides Eastw.	SDSU 20113	Ripma 379	44.54885, −120.33336	KM213405	KP096525
Oreocarya crymophila (I. M. Johnst.) Jeps. & Hoover	SDSU 20116	Ripma 390	38.61707, −119.83705	KM213411	KP096532
Oreocarya flavoculata A. Nelson	SDSU 20030	Ripma 307	37.3857, −118.1805	KM213424	KP096526
Oreocarya hoffmannii (I. M. Johnst.) Abrams	SDSU 20036	Ripma 306	37.2635, −118.15706	KM213407	KP096537
Oreocarya humilis Greene subsp. *humilis*	SDSU 20029	Ripma 303	37.74431, −119.02917	KM213404	KP096530
Oreocarya hypsophila (I. M. Johnst.) Hasenstab & M. G. Simpson	SDSU 20086	Ripma 374	45.66983, −112.81633	KM213410	KP096523
Oreocarya nubigena Greene ("Horseshoe Meadows")	SDSU 20004	Ripma 312	36.4505, −118.163	KM213408	KP096528
Oreocarya nubigena Greene ("Mammoth Mtn.")	SDSU 20055	Ripma 301	37.62715, −119.03075	KM213402	KP096540
Oreocarya nubigena Greene ("Mono Craters")	SDSU 20094	Ripma 399	37.91395, −119.03845	KM213417	KP096542
Oreocarya nubigena Greene ("Sierran granite")	SDSU 20079	Ripma 363	37.41913, −118.7518	KM213419	KP096541
Oreocarya nubigena Greene ("Sonora Pass")	SDSU 20098	Ripma 395	38.2858, −119.64189	KM213406	KP096543
Oreocarya schoolcraftii (Tiehm) R. B. Kelley	SDSU 20123	Ripma 370	43.90865, −117.62695	KM213420	KP096522
Oreocarya setosissima (A. Gray) Greene	SDSU 20242	Kelley 1466	35.34968, −111.74575	KM213418	KP096531
Oreocarya sobolifera (Payson) R. B. Kelley	SDSU 20210	Kelley 1169	48.4816, −113.34305	KM213415	KP096535
Oreocarya subretusa (I. M. Johnst.) Abrams ("Mt. Eddy")	SDSU 20232	Kelley 928	41.31715, −122.48227	KM213416	KP096545
Oreocarya subretusa (I. M. Johnst.) Abrams ("type location")	SDSU 20107	Ripma 384	42.95651, −122.04713	KM213412	KP096539
Oreocarya subretusa (I. M. Johnst.) Abrams ("Warner Mts.")	SDSU 20110	Ripma 389	41.44734, −120.24316	KM213421	KP096544
Oreocarya suffruticosa (Torr.) Greene var. *abortiva* (Greene) J. F. Macbr.	SDSU 20024	Ripma 308	37.49136, −118.18608	KM213414	KP096529
Oreocarya virgata (Porter) Greene	SDSU 20117	Ripma 371	40.92065, −106.31195	KM213401	KP096538
Pectocarya penicillata (Hook. & Arn.) A. DC.	UC 1965571	Kelley 1967	34.3060, −117.46565	KM213403	KP096533

[a] mtDNA data available from the Dryad Digital Repository (http://doi.org/10.5061/dryad.50536; Ripma et al., 2014).

Downloaded From: https://bioone.org/journals/Applications-in-Plant-Sciences on 29 Mar 2024
Terms of Use: https://bioone.org/terms-of-use