Applications in Plant Sciences

Applications in Plant Sciences 2014 2(12): 1400044

Application Article

Gene prediction and annotation in *Penstemon* (Plantaginaceae): A workflow for marker development from extremely low-coverage genome sequencing[1]

Paul D. Blischak[2,3], Aaron J. Wenzel[2], and Andrea D. Wolfe[2]

[2]Department of Evolution, Ecology, and Organismal Biology, Ohio State University, 318 W. 12th Avenue, Columbus, Ohio 43210 USA

- *Premise of the study: Penstemon* (Plantaginaceae) is a large and diverse genus endemic to North America. However, determining the phylogenetic relationships among its 280 species has been difficult due to its recent evolutionary radiation. The development of a large, multilocus data set can help to resolve this challenge.
- *Methods:* Using both previously sequenced genomic libraries and our own low-coverage whole-genome shotgun sequencing libraries, we used the MAKER2 Annotation Pipeline to identify gene regions for the development of sequencing loci from six extremely low-coverage *Penstemon* genomes (~0.005×–0.007×). We also compared this approach to BLAST searches, and conducted analyses to characterize sequence divergence across the species sequenced.
- *Results:* Annotations and gene predictions were successfully added to more than 10,000 contigs for potential use in downstream primer design. Primers were then designed for chloroplast, mitochondrial, and nuclear loci from these annotated sequences. MAKER2 identified longer gene regions in all six *Penstemon* genomes when compared with BLASTN and BLASTX searches. The average level of sequence divergence among the six species was 7.14%.
- *Discussion:* Combining bioinformatics tools into a workflow that produces annotations can be useful for creating potential phylogenetic markers from thousands of sequences even when genome coverage is extremely low and reference data are only available from distant relatives. Furthermore, the output from MAKER2 contains information about important gene features, such as exon boundaries, and can be easily integrated with visualization tools to facilitate the process of marker development.

Key words: 454 pyrosequencing; bioinformatics; BLAST; MAKER2; *Penstemon.*

Penstemon Mitch. has been the subject of many ecological and systematic studies, due in large part to its diversity of forms, habitats, and pollination syndromes (Wolfe et al., 2006). Its 280 species are distributed across North America with states of the Intermountain Region (USA) boasting as many as 45–75 species each (Nold, 1999; Lindgren and Wilde, 2003). Works by Pennell (1920, 1935), Keck (1932, 1936), Straw (1956a, b), and Holmgren (1984) provide thorough detailing of the differing morphologies and difficult taxonomy of this genus. Wolfe et al. (2006) conducted the largest molecular phylogenetic study on *Penstemon* to date with 163 species included in their parsimony analyses of the nuclear ITS and chloroplast *trnC-D* and *trnT-L* regions. These analyses provided some insights into the phylogenetic relationships within the genus, such as confirming the basal position of the subgenus *Dasanthera*, and suggesting at least 10 independent origins of the hummingbird pollination syndrome. However, relationships for taxa within and among subgenera, sections, and subsections were not always consistent with current taxonomy, and relationships within clades having strong support were largely unresolved. Wolfe et al. (2006) hypothesized that the genus has undergone a rapid evolutionary radiation and that the markers used were not sufficiently variable to determine the relationships in the tips of the tree. The topologies of the nuclear ITS and chloroplast trees were also incongruent, most likely as a result of hybridization or incomplete lineage sorting. Many naturally occurring hybrids have been discovered and studied in *Penstemon* (Wolfe et al., 1998a, b; Wilson and Valenzuela, 2002; Datwyler and Wolfe, 2004), making the inference of species boundaries and sister relationships particularly troublesome. Furthermore, the recent radiation of the genus makes inference of the phylogeny all the more difficult due not only to incomplete lineage sorting, but also to factors such as gene flow after speciation (Leaché et al., 2014). Thus, the need for a large-scale, multilocus data set to resolve the relationships within the genus, in light of the potential sources of discordance, is paramount.

High-throughput sequencing technologies greatly facilitate the creation of large marker sets, and numerous approaches exist for their generation (Davey et al., 2011; Good, 2011; Cronn et al., 2012). Among these is low-coverage whole-genome shotgun sequencing (WGS), or genome skimming, which sequences a small fraction of the genome for characterization and

[1]Manuscript received 30 May 2014; revision accepted 7 November 2014.

The authors thank M. Zianni and A. McCoy for assistance with next-generation sequencing; P. Jourdan for growing *Penstemon centranthifolius*, *P. grinnellii*, and other *Penstemon* species and for supporting the low-coverage whole genome shotgun sequencing runs completed in this research; M. Stevens for helpful discussion regarding the genomic reduction using restriction-site conservation (GR-RSC) study on *Penstemon*; and three anonymous reviewers for helpful comments on the manuscript.

[3]Author for correspondence: blischak.4@osu.edu

doi:10.3732/apps.1400044

marker development via random shearing of the DNA followed by high-throughput sequencing (Straub et al., 2012). High-copy chloroplast genomes can often be fully recovered and can be used to create sequencing primers (Straub et al., 2011). Another application for low-coverage WGS includes the identification of microsatellite loci (Jennings et al., 2011; Castoe et al., 2012). The discovery of low-copy nuclear loci, on the other hand, can be more difficult as very-low-coverage genome skimming may produce only small fragments (e.g., ~400–500 bp on the Roche 454) when reads cannot be assembled into contigs. Sifting through these thousands of small fragments to find useful information is a major challenge of the postsequencing process. However, many bioinformatic tools exist for characterizing data from next-generation sequencing (NGS) runs, including identifying regions through BLAST searches, mapping the reads to a reference genome using BLAT, as well as de novo approaches such as gene prediction (Altschul et al., 1990; Kent, 2002; Stanke and Waack, 2003; Korf, 2004). A common problem is that many organisms of biological interest do not have reference genomes, requiring researchers to rely on alignments to sequences derived from phylogenetically distant species in GenBank or other sequence repositories. De novo sequence characterization does remove the dependence on the amount and quality of data in sequence databases, but training ab initio gene predictors can be difficult, and gene prediction algorithms can inflate the number of genes identified because of false positives (Yandell and Ence, 2012).

Software that includes multiple different types of gene-finding tools for NGS data offers a potential solution. The MAKER2 Annotation Pipeline is one such bioinformatic tool that combines many useful analyses such as repeat masking (RepeatMasker; Smit et al., 1996), ab initio gene prediction (Semi-HMM-based Nucleic Acid Parser [SNAP]; Korf, 2004), and expressed sequence tag (EST) and protein alignments (BLAST [Altschul et al., 1990], Exonerate [Slater and Birney, 2005]) to annotate genomic contigs (Cantarel et al., 2008; Holt and Yandell, 2011). It also has the capability to act as a wrapper program for the training of gene prediction algorithms, such as SNAP or Augustus, by iteratively updating the parameters in the hidden Markov models (HMM) that are used by these programs to identify gene regions from NGS data (Stanke and Waack, 2003; Korf, 2004; Cantarel et al., 2008). This functionality greatly facilitates the training of gene prediction algorithms that particularly pose an analytical challenge (Cantarel et al., 2008). The EST and protein libraries input to MAKER2 for sequence alignments are used in combination with gene predictions to produce annotations, and would ideally be from the organism being annotated (e.g., from a previous transcriptome sequencing project). However, for laboratories completing NGS projects that yield only small portions of the genome(s) of the targeted organism(s), these resources are typically not available, which only adds to the problem of identifying potential loci, especially if genome coverage is very low.

Here we present the application of free bioinformatic tools to annotate NGS data from six extremely low-coverage (~0.005×–0.007×) genomic libraries for different *Penstemon* species representing a large phylogenetic range of the genus. The low-coverage nature of our data presents a unique challenge in that most low-coverage WGS projects have a greater representation of the genome than just a fraction of a percent. For example, Straub et al. (2011) employed "low-coverage" WGS of *Asclepias syriaca* L. and had 0.5× coverage of the genome, which is two orders of magnitude greater than the depth of coverage of our genomes. An obvious solution to this challenge would be to sequence the genome at a higher depth of coverage, or to use multiple approaches such as transcriptome sequencing or targeted enrichment in combination with genome skimming (Good, 2011; Weitemier et al., 2014). However, despite the decreasing costs of NGS technologies (Davey et al., 2011; Straub et al., 2012), financial constraints can still prevent the acquisition of higher-coverage WGS data or the use of multiple sequencing strategies. Thus, tools that can extract information from extremely low-coverage WGS data are helpful. To this end, we employ a workflow (Fig. 1) centered on the MAKER2 Annotation Pipeline, which provides a practical framework for identifying contigs containing gene regions, even when the majority of the sequences are short (~400–500 bp) and genomic resources are only available from distant relatives of the target organism. It also allows for the direct characterization of genomic features, such as exon boundaries, which can be used to design primers for future PCR-based sequencing efforts. This fits well with our larger goal to resolve the phylogeny of *Penstemon*, as we plan to use the markers developed here for the targeted enrichment of PCR amplicons using parallel tagged sequencing (PTS; Meyer et al., 2008; e.g., O'Neill et al., 2013). Finally, we compare this workflow to more standard approaches that use searches for sequence similarity (BLAST) and contrast the amount of potentially useful data resulting from each approach.

MATERIALS AND METHODS

***Sequence data*—**A previous NGS study on *Penstemon* employed genomic reduction using restriction-site conservation (GR-RSC) in combination with 454 pyrosequencing to study the genome content of four species, and to develop single-nucleotide polymorphism (SNP) and simple sequence repeat (SSR) markers (Dockter et al., 2013). As the name indicates, GR-RSC differs from low-coverage WGS through its use of restriction enzymes (rather than random shearing of the DNA), followed by size selection and high-throughput sequencing (Maughan et al., 2009). Contigs from Dockter et al. (2013) were downloaded from GenBank with the following accession numbers: *Penstemon fruticosus* (Pursh) Greene (AKKJ01), *P. davidsonii* Greene (AKKI01), *P. dissectus* Elliott (AKKH01), and *P. cyananthus* Hook. (AKKG01).

Whole-genome shotgun sequencing of P. centranthifolius *and* P. grinnellii—DNAs from one accession of *P. centranthifolius* (Benth.) Benth. and *P. grinnellii* Eastw., included in previous studies (Wolfe and Elisens, 1994, 1995), were used for low-coverage WGS. The selected accessions showed no evidence of introgressive hybridization (Wolfe and Elisens, 1994, 1995). These samples were normalized to a concentration of approximately 50 ng/μL and sent to the Plant Microbe Genomics Facility (PMGF, Columbus, Ohio, USA) for next-generation sequencing. Library preparation, sequencing, and assembly were completed by the PMGF. Briefly, separate libraries were created for *P. centranthifolius* and *P. grinnellii* by sonically shearing the DNA samples and attaching individual barcodes. Samples were then pooled and sequenced on four of eight partitions of a picotitre plate on the Roche 454 platform (454 Life Sciences, a Roche Company, Branford, Connecticut, USA). Contigs were assembled using Newbler version 2.8 (454 Life Sciences, a Roche Company) with a minimum contig length of 100 bp.

***The MAKER2 Annotation Pipeline*—**MAKER2 runs on three control files that are generated at the command line to direct the program to all of the needed executables and sequence libraries, as well as to define parameter values for sequence alignments and gene predictions (Appendix S1). Input for MAKER2 includes the genomic contigs to be annotated, a library of ESTs for sequence alignments and gene predictions, and a library of protein sequences for alignment to the contigs translated in all reading frames. The software programs and databases that are dependencies of the MAKER2 Annotation Pipeline were installed locally and are listed in Table 1. All analyses were carried out on a Dell desktop computer (Dell, Round Rock, Texas, USA) with 2 GB of RAM running CentOS Linux version 6.3.

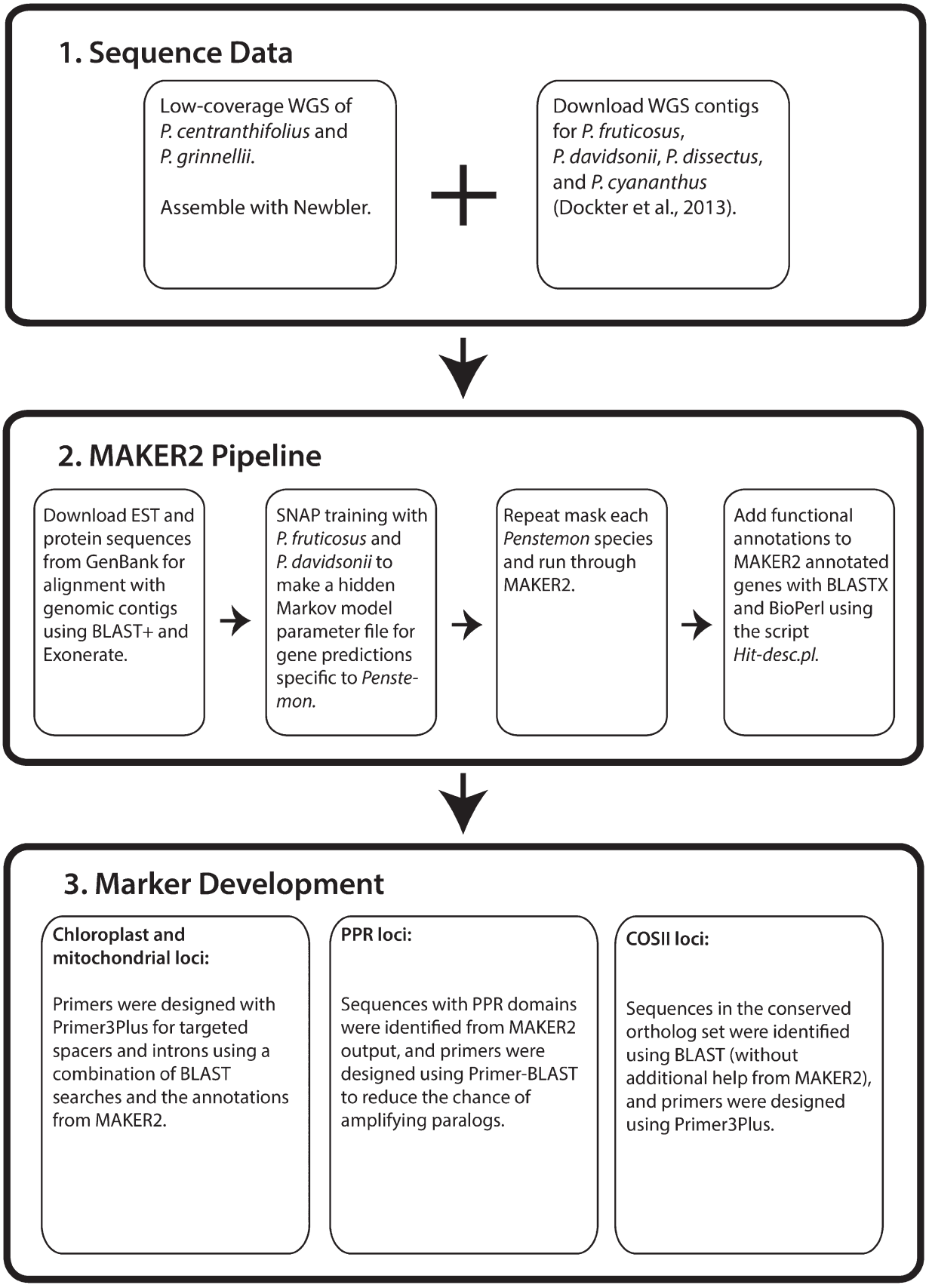

Fig. 1. Workflow used for marker development from six low-coverage *Penstemon* genomes using the MAKER2 Annotation Pipeline.

Table 1. MAKER2 dependencies, version numbers, and websites.

Program	Version	Website[a]
MAKER2 (Holt and Yandell, 2011)	update 07-22-2012	http://www.yandell-lab.org/software/maker.html
RepeatMasker (Smit et al., 1996)	open-3.3.0	http://www.repeatmasker.org
RMBLAST (Smit et al., 1996)	2.2.27	http://www.repeatmasker.org/RMBlast.html
RepBase	update 20120418	http://www.girinst.org/
RM database	update 20120418	http://www.girinst.org/
SNAP (Korf, 2004)	N/A; downloaded 15 Dec. 2012	http://korflab.ucdavis.edu/software.html
Legacy BLAST (Altschul et al., 1990)	2.2.26	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/LATEST/
BLAST+ (Camacho et al., 2009)	2.2.27	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST
Exonerate (Slater and Birney, 2005)	2.2.0	http://www.ebi.ac.uk/~guy/exonerate
Perl	5.10	http://www.perl.org
BioPerl (Stajich et al., 2002)	1.6.1	http://www.bioperl.org/wiki/Getting_BioPerl
Augustus[b] (Stanke and Waack, 2003)	2.5	http://augustus.gobics.de/binaries/old

[a] URLs for websites as of 14 November 2014.
[b] Required for installation but not used in our analyses.

EST and protein libraries—ESTs for all available species of Plantaginaceae and Orobanchaceae were downloaded from GenBank. Orobanchaceae was chosen because it is a closely related family in the order Lamiales (Olmstead et al., 2001) and also because it is a group that has been widely studied (e.g., Parasitic Plant Genome Project [Westwood et al., 2012]). Protein sequences for *Arabidopsis thaliana* (L.) Heynh. and *Solanum lycopersicum* L. were downloaded from the UniProt database, and protein sequences for all available species of Plantaginaceae were downloaded from GenBank. All sequences are available from the Dryad Digital Repository (http://doi.org/10.5061/dryad.f6s22; Blischak et al., 2014).

Training ab initio gene predictor—To create an HMM parameter file for *Penstemon*, the ab initio gene predictor SNAP was iteratively trained in two rounds in a manner similar to the methods of Cantarel et al. (2008). First, sequences of *P. fruticosus* were repeat masked and run through MAKER2 using EST alignments and SNAP, with an HMM parameter file pretrained for *A. thaliana*, to look for gene regions. The resulting gene predictions were then used to create a new parameter file to be used by SNAP for future gene predictions. Using the new parameters, we ran MAKER2 a second time using only SNAP to generate gene predictions, and no EST alignments. This second step was repeated once more for a total of three training runs on the *P. fruticosus* contigs, producing a preliminary HMM parameter file on which we based the second round of training.

In the second round of training, contigs for *P. fruticosus* and *P. davidsonii* were repeat masked, pooled, and reassembled using CAP3, a genome assembly program that takes input sequences, as well as any available base quality information, to build contigs using sequence overlap and alignment scores to produce consensus contigs from the input data (Huang and Madan, 1999). Reassembled contigs were combined with all contigs not combined with any others (singletons) to form a single training library of sequences (Dryad-files D2 [available from the Dryad Digital Repository: http://doi.org/10.5061/dryad.f6s22]; Blischak et al., 2014). The new set of sequences was run through MAKER2 with EST alignments and SNAP using the HMM file from the first round of training. As before, we used the resulting gene predictions to update the HMM file and reran MAKER2 on the combined set of contigs. This second step was repeated four more times using the same iterative process of updating the SNAP HMM parameter file with each new run of MAKER2, resulting in six training runs from round 2 and a total of nine training runs from both rounds. Repeat masking was left on during the training process, in addition to the repeat masking done before running MAKER2.

Running MAKER2 and BLAST—Raw contigs for the six species of *Penstemon* were repeat masked using species-specific repeat settings (Smit et al., 1996). This repeat masking was done separately from any other repeat masking conducted during the training of SNAP, and the resulting masked contigs were used for all subsequent analyses. Masked contigs were then run through MAKER2 using the trained SNAP HMM parameter file for *Penstemon* and EST-based predictions turned on. Additional repeat masking was turned off and all other settings in the control files were left at their default values. As a comparison, we conducted BLAST searches against the EST and protein libraries that we used in MAKER2. Alignments against these libraries were done with the same settings used by MAKER2 for BLASTN (minimum *E*-value of 1×10^{-10}, percent identity of 85%) and BLASTX (minimum *E*-value of 1×10^{-6}), with each search being conducted twice to compare the amount of output between recording all hits (unrestricted), or restricting the output to only the single best hit (max_target_seqs = 1).

We used two measures for comparing the output of MAKER2 and BLAST: (1) the length of the annotations/alignments, and (2) the amount of output generated by the two methods (i.e., the number of sequences identified as containing gene regions). The lengths of the MAKER2 annotations were determined by first combining all the annotations of every contig for each species into individual files using the *gff3_merge* utility script. MAKER2 annotations were then pulled from the GFF3 files with the *Filter and Sort::Extract features* tool from the Galaxy online bioinformatics portal (Giardine et al., 2005). Output files containing MAKER2 annotations and BLASTN and BLASTX alignments were then imported into R to compare the distribution of annotation/alignment lengths (R Core Team, 2014). The amount of output generated by each program was simply the number of lines printed in the output file and is an important factor to consider because it ultimately determines the quantity of data that will have to be sorted through when developing a set of markers.

***Functional annotation using BLAST and BioPerl*—**Contigs receiving annotations from MAKER2 were extracted and put into separate FASTA files for each species using the *gff3_merge* and *maker2zff* utility scripts (Cantarel et al., 2008). Statistics for the annotated contigs were calculated using the *fathom* command in SNAP designated with the *-gene-stats* flag at the command line (Korf, 2004). Functional annotations were added to these contigs with BLASTX and a custom Perl script using the BioPerl module *Bio::DB::GenBank* (Stajich et al., 2002). Three local databases were created for the annotations: (1) all RefSeq plant proteins, (2) all *A. thaliana* RefSeq proteins, and (3) all *S. lycopersicum* RefSeq proteins (Pruitt et al., 2012). Settings for BLASTX restricted the output to 10 hits per query alignment, a best hit overhand of 0.1, and a minimum *E*-value of 1e-10 (Altschul et al., 1990, 1997). All BLASTX alignments were completed using BLAST+ version 2.2.27 (Camacho et al., 2009). Functions were added to the BLASTX hits via the Perl script by pulling out the accession number for each hit sequence from the tab-delimited BLAST output file (-outfmt 6) and then querying GenBank for its functional description.

***Primer development for chloroplast, mitochondrial, and nuclear loci*—**Characterization and mapping of the chloroplast sequences obtained from the GR-RSC and low-coverage WGS runs was done using the MUMmer (version 3.23; Kurtz et al., 2004) suite of alignment tools. Repeat masked contigs for each *Penstemon* species were aligned to the chloroplast genome of *S. lycopersicum* (AC_000188.1), after removing one copy of the inverted repeat region, using NUCmer. Following the initial alignment, mapped contigs were filtered (using the *delta-filter* utility) to achieve a one-to-one mapping of the contig query sequences to the reference chloroplast genome. Genomic coordinates and summary statistics (percent coverage, sequence length) of the alignments were extracted with the *show-coords* tool (Kurtz et al., 2004).

Due to the difference in sequencing techniques used (see Results), *P. centranthifolius* and *P. grinnellii* had many large contigs (>10 kb) that were from the chloroplast and mitochondrial genomes. We narrowed our search for chloroplast

regions to only include intron sequences pulled from the chloroplast genome of *S. lycopersicum* and the intergenic spacers used in Shaw et al. (2005; species *Gratiola brevifolia* Raf.) using BLAST searches. BLAST searches were also used for targeting mitochondrial regions using introns from the mitochondrial genome of *Mimulus guttatus* DC. (NC_018041.1). Our approach was to combine the annotations from MAKER2 with these preliminary BLAST alignments to identify the contigs containing the targeted regions. The identified contigs were imported into the Apollo Genome Browser, along with all of their annotations, to build primers that were anchored in the predicted/annotated exons to sequence across the target introns and intergenic spacers. This was done by matching up the coordinates of the BLAST alignments to the annotated exons by hand and recording the coordinates of the intron/spacer start and end positions for primer design (Lewis et al., 2002). Primers for all targeted chloroplast and mitochondrial regions were built using Primer3Plus online with default settings, targeting the desired regions using the recorded coordinates for the locations of the introns and intergenic spacers (Rozen and Skaletsky, 2000). All BLAST searches were done using TBLASTX in the BLAST+ toolkit and default settings (Camacho et al., 2009).

A number of our annotated contigs from the MAKER2 runs also contained pentatricopeptide repeat (PPR) domains. PPR loci belong to a multigene family (ca. 450 PPR genes in *A. thaliana*) and have been shown to be highly variable and useful for phylogenetic inference in plants. A large proportion of them are also intronless (Yuan et al., 2009). PPR genes contain repeat motifs of 35 amino acids that can vary in the number of repeat units, and are used for posttranscriptional processing in the chloroplast and mitochondria (Yuan et al., 2009). We identified 14 PPR loci among our contigs and designed primers for them using Primer-BLAST with *A. thaliana* as the reference database against which the sequences were aligned (Ye et al., 2012). Primer-BLAST was chosen over Primer3Plus to reduce the chance of designing primers that could potentially amplify paralogous loci. Primers were tested for successful amplification using the PCR conditions of Yuan et al. (2009) and verified on a 1% agarose gel.

In an analysis separate from our MAKER2 annotations, we also searched for single-copy nuclear genes from the conserved ortholog set (COSII) for euasterids using TBLASTX with default settings. COSII sequences were downloaded from The *Arabidopsis* Information Resource (TAIR; http://arabidopsis.org) using the *Bulk Data Retrieval* tool and accession numbers provided in Wu et al. (2006). We selected only COSII loci that were identified in two or more of our species, and aligned those sequences in MEGA5 using the MUSCLE alignment tool to assess the proportion of variable sites (Edgar, 2004; Tamura et al., 2011). Primers for sequences containing hits to COSII loci were then designed to target the entire contig with Primer3Plus online and default settings, using the longest contig from the set of two or more identified by a given COSII locus (Rozen and Skaletsky, 2000). In addition, we screened 11 of the most variable COSII loci using the PCR conditions described in Wu et al. (2006), with successful amplification verified by the presence of a band in the correct size range on a 1% agarose gel.

***Pairwise sequence variation*—**To assess the amount of nucleotide differentiation among the six *Penstemon* species sequenced, we conducted pairwise BLAST searches of all the 454 contigs assembled for each species. We used each of the six species as both the set of query sequences and as the database against which alignments were made, resulting in a total of 30 pairwise comparisons. Searches were done using BLASTN with a minimum *E*-value of 1×10^{-10} and recording only the best hit. Pairwise sequence differentiation (proportion of variable sites) was calculated as the number of mismatched sites divided by the total alignment length for each alignment in a given BLASTN search. The mean and standard error for the proportion of variable sites were calculated in R, along with the overall mean and standard error for all BLASTN alignments (R Core Team, 2014).

RESULTS

***Sequence data (Next-gen sequencing of* P. centranthifolius *and* P. grinnellii*)*—**454 pyrosequencing of *P. centranthifolius* and *P. grinnellii* yielded 301,622 and 218,457 reads, with an average read length of 428.6 bp and 425.5 bp, respectively. Assembled sequences resulted in 8436 contigs (4,719,701 bp) for *P. centranthifolius* and 6927 contigs (3,874,098 bp) for *P. grinnellii* (Table 2). For both of the assemblies, only about one-third of the 454 reads were successfully assembled into contigs. This is likely due to the nature of the low-coverage technique used for sequencing, with reads from the chloroplast and mitochondrial genomes assembling into larger contigs and reads from the nuclear genome remaining unassembled. The estimated genome size of *P. grinnellii* is 686 Mb, putting the genome coverage (including the organellar genomes) at 0.006× (Broderick et al., 2011). *Penstemon centranthifolius* is a close relative of *P. grinnellii*, and therefore would be expected to have a genome that is similar in size, putting the genome coverage for *P. centranthifolius* at 0.007×. The assembled contigs are available through the WGS Database in GenBank (*P. centranthifolius*: JPFH01, *P. grinnellii*: JPFI01). Downloaded sequence data from Dockter et al. (2013) are as follows: *P. fruticosus* = 4770 contigs (2,319,038 bp; 0.005×); *P. davidsonii* = 4880 contigs (2,375,230 bp; 0.005×); *P. dissectus* = 5361 contigs (2,628,091 bp; 0.005×); *P. cyananthus* = 9712 contigs (4,622,258 bp; 0.006×). In total, we collected 40,086 genomic contigs from six species of *Penstemon* comprising 20,538,516 bp of sequence data to be used for annotation and marker development.

TABLE 2. Low-coverage WGS sequencing and assembly statistics for *Penstemon centranthifolius* and *P. grinnellii*.

Statistics	*P. centranthifolius*	*P. grinnellii*
Sequencing statistics		
Total number of reads	301,622	218,457
Total base pairs sequenced	129,272,235	92,989,517
Assembly statistics		
% reads assembled	33.04%	36.02%
Total assembled base pairs	4,719,701	3,874,098
No. of contigs	8436	6927
No. of contigs (>500 bp)	3200	2677
Contig N50 (>500 bp)	984	1012

***MAKER2 annotations and gene predictions*—**Iterative training of the ab initio gene predictor SNAP resulted in an HMM parameter file that was able to produce consistent predictions for *Penstemon*. Repeat masking was done prior to training, as well as during the running of MAKER2, to ensure that repetitive regions did not affect the generation of the gene prediction parameter file. This was a more conservative approach than the single round of masking done prior to the actual annotation stage. Our method of training also deviated from that of Cantarel et al. (2008) in that we did not reduce the number of contigs used during training after the first round. This was done because we did not have long contigs containing multiple genes and the total number of base pairs in the entire training data set was on the order of 10 Mb, which was a much smaller training set than that used by Cantarel et al. (2008). The main difference between MAKER2 annotations and SNAP gene predictions is that MAKER2 combines the results of gene predictions (SNAP) and sequence alignments (BLASTN, BLASTX, Exonerate) to annotate contigs using quality measures to ensure better accuracy. Gene prediction algorithms, on the other hand, work alone when not incorporated into a pipeline and can be less accurate than a combined approach such as the one used by MAKER2 (Cantarel et al., 2008; Holt and Yandell, 2011; Yandell and Ence, 2012). In total, 1895 genes were fully annotated by MAKER2 and 8469 were predicted by SNAP, although we expect that the annotations/predictions from the different GR-RSC genomes will overlap (likewise for the low-coverage WGS genomes), making the number of unique genes identified smaller than the total number (Table 3). MAKER2 annotations were also longer on average when compared to predictions made by SNAP (Appendix S2). Functional annotations added to contigs with MAKER2 annotations with BLASTX and the custom Perl script

TABLE 3. Comparison of the amount of output for sequences identified to contain gene regions using MAKER2, SNAP, BLASTN, and BLASTX. Results for unrestricted BLASTN/BLASTX searches are given along with the searches that reported only the best hit.

Penstemon species	MAKER2 annotations	SNAP gene predictions	BLASTN unrestricted/best hit	BLASTX unrestricted/best hit
P. centranthifolius	486	2437	823/77	150,172/1839
P. grinnellii	365	1898	805/79	88,619/1378
P. fruticosus	238	784	659/164	107,867/1511
P. davidsonii	238	827	695/169	101,621/1561
P. dissectus	230	801	779/197	103,041/1563
P. cyananthus	338	1689	858/228	173,980/2740

(Appendix S3) are available from the Dryad Digital Repository (http://doi.org/10.5061/dryad.f6s22; Blischak et al., 2014).

***BLAST searches*—**Comparisons between MAKER2 and BLAST based on the two measures of annotation/alignment length and quantity of output showed that MAKER2 identified longer gene regions on average and had less output than similar BLAST searches. For BLASTN searches to the EST library, the amount of output generated was much less than that for MAKER2 and BLASTX searches. BLASTX searches to the protein library generated much more output, especially when the number of hits was unrestricted. This is likely due to the fact that many sequences in the databases used were being targeted multiple times by our contigs. However, restricted BLASTX searches still produced more output than MAKER2 (Table 3). The distributions of annotation/alignment length are given in Fig. 2, with means and standard errors reported in the top right corner for each species.

***Chloroplast, mitochondrial, and nuclear markers for* Penstemon—**Despite the longer contigs in the assemblies for *P. centranthifolius* and *P. grinnellii*, we were unable to recover a full chloroplast genome after mapping our contigs to the plastome of *S. lycopersicum* (*P. centranthifolius* = 81.5%, *P. grinnellii* = 78.9%). Mapping of the reads from the GR-RSC genomes were highly fragmented and did not cover a majority of the chloroplast genome of *S. lycopersicum*. Primers designed for specifically targeted genes with Primer3Plus resulted in 28 COSII, 11 chloroplast, and 10 mitochondrial loci. Chloroplast and mitochondrial markers were designed from *P. centranthifolius* and *P. grinnellii*. The COSII loci were designed from the GR-RSC genomes. This discrepancy between the nuclear and organellar content of the low-coverage WGS vs. GR-RSC libraries is a result of the sequencing techniques used to gather the data. Low-coverage WGS of *P. centranthifolius* and *P. grinnellii* produced high amounts of chloroplast

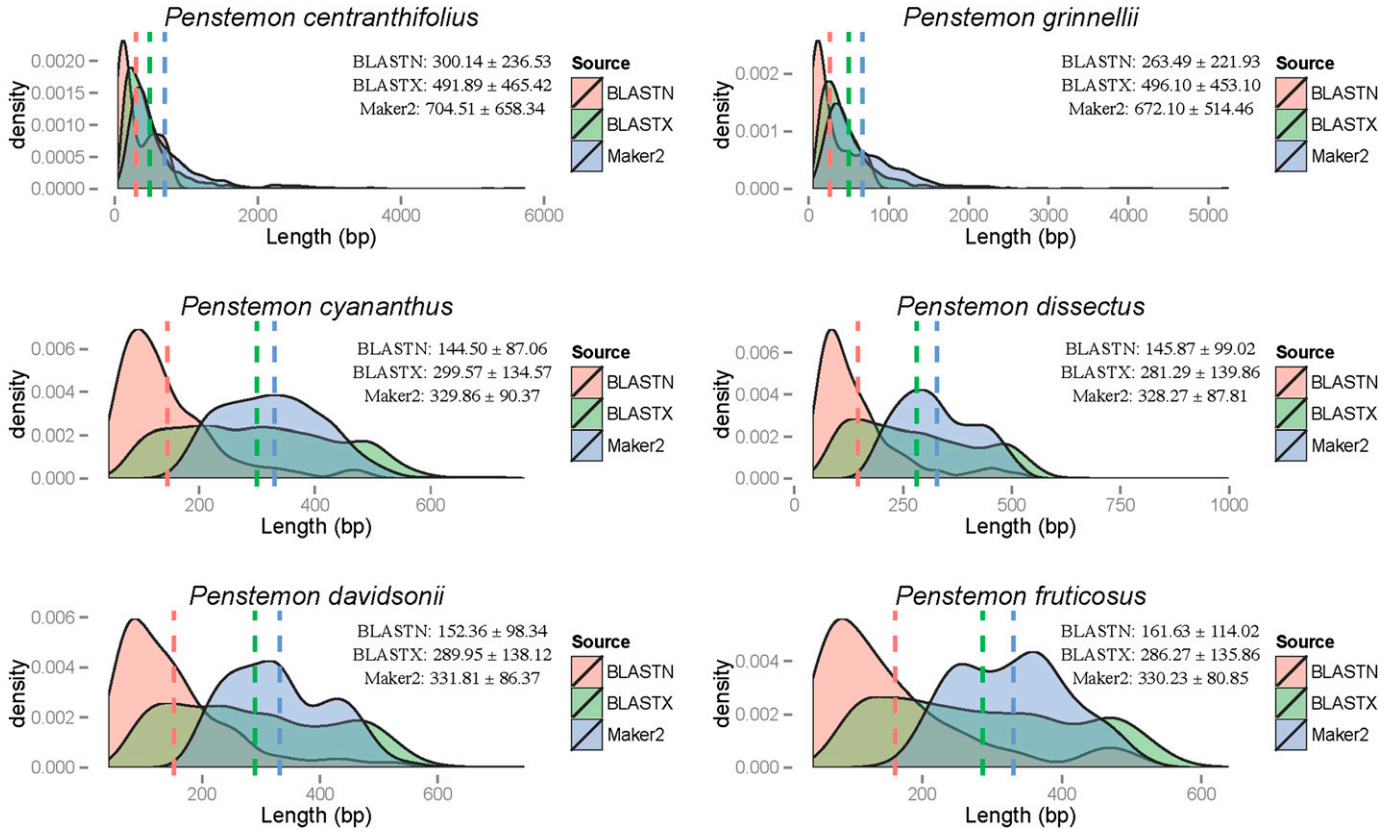

Fig. 2. Comparing MAKER2 annotations to best-hit BLAST searches against ESTs (BLASTN) and protein sequences (BLASTX) for the six species of *Penstemon* sequenced. Mean sequence lengths are plotted as dashed, vertical lines. Means ± SEs are also given in the upper right corner of each graph.

and mitochondrial data, but comparatively little nuclear data. Sequencing of *P. cyananthus*, *P. davidsonii*, *P. dissectus*, and *P. fruticosus* using GR-RSC, on the other hand, produced only small fragments of the organellar genomes but had more representation of the nuclear genome.

Of the 11 COSII markers tested, 10 were successfully amplified on the first attempt. Of the 14 PPR markers tested, only seven were successfully amplified on the first try. Further attempts to amplify the remaining PPR loci were not conducted. The direct cause of the 50% failure rate in the PPR loci was unknown. However, given that the PPR gene family has many copies (potentially hundreds), we believe that the failure may be due to issues with paralogy. Primers for all loci developed here are given in Appendix S4.

***Sequence variation in* Penstemon—**Among the six species of *Penstemon* that were sequenced, the amount of sequence variation ranged, on average, from 3.62% for *P. fruticosus* vs. *P. davidsonii* to 8.88% for *P. centranthifolius* vs. *P. fruticosus* (Fig. 3). This result is congruent with the current understanding of the relationships in the genus, as alignments between the other four species with either *P. fruticosus* or *P. davidsonii* (both members of subgenus *Dasanthera*, the earliest branching lineage of *Penstemon*) typically had the largest amounts of sequence variation, and alignments between these two species contained comparatively little sequence variation. Of the non-*Dasanthera* species surveyed, *P. grinnellii* and *P. centranthifolius* are the closest relatives (Wolfe et al., 2006), with a sequence variation level of ~5.7%. The average amount of sequence variation among all six species was 7.14%.

***Supplemental material*—**All supplemental figures and files, including an example protocol for using MAKER2, are available in the supplementary material accompanying this article (Appendices S1–S6). Sequence libraries and functional annotation files, along with the Perl script for adding annotations and the HMM file for *Penstemon* are available from the Dryad Digital Repository (http://doi.org/10.5061/dryad.f6s22; Blischak et al., 2014).

DISCUSSION

Despite the potential roadblocks to annotating our sequence data (extremely low-coverage, short contigs, genomic resources from distant relatives only), we were able to successfully identify hundreds of gene regions using MAKER2. By combining the functionality of many programs into one single pipeline, MAKER2 offers a way to simultaneously run many analyses to fully characterize genomic sequence data (Cantarel et al., 2008). Comparable approaches such as BLAST searches are also helpful, especially when targeting specific sequences such

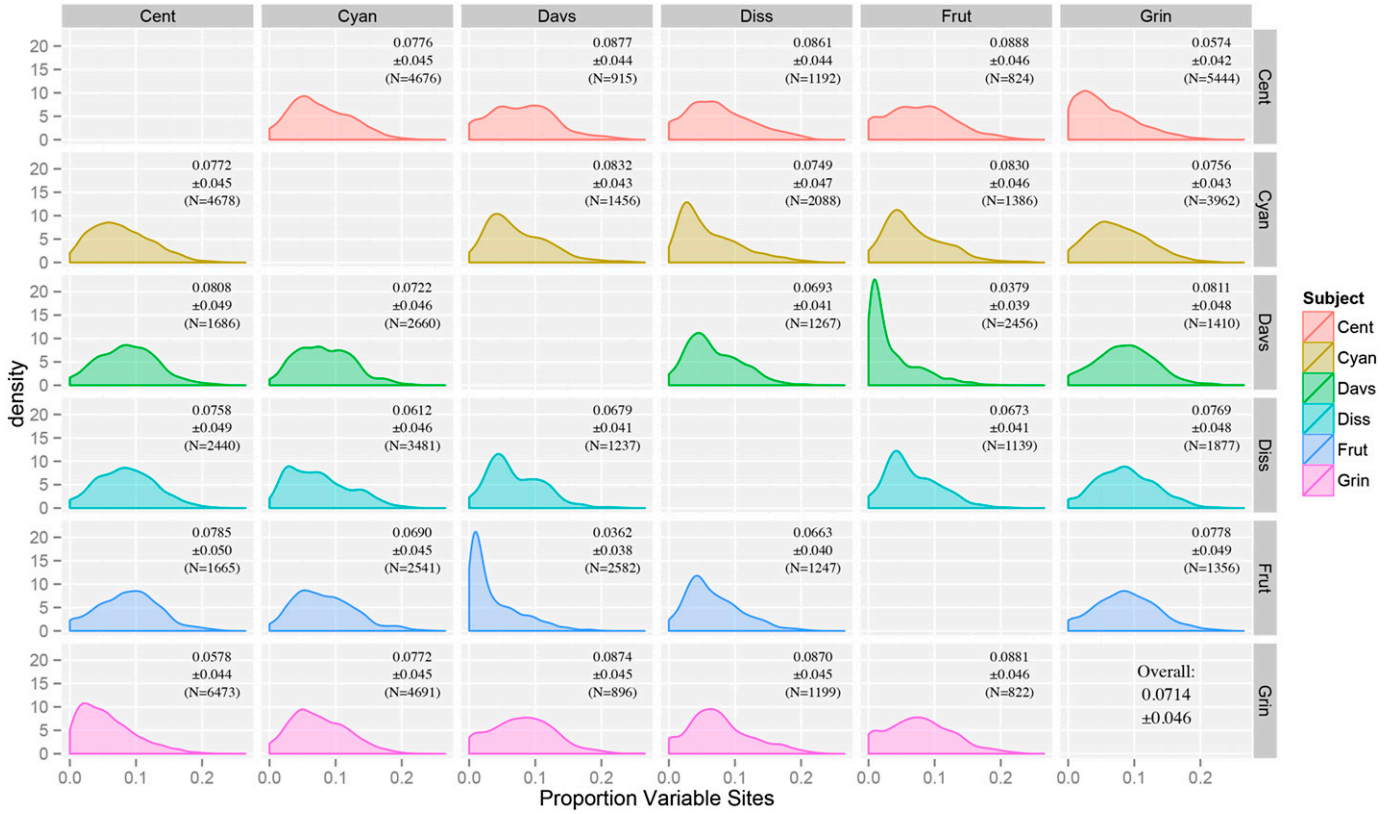

Fig. 3. Plot of pairwise comparisons of sequence variation among the six low-coverage genomes using BLASTN (Cent = *P. centranthifolius*, Cyan = *P. cyananthus*, Davs = *P. davidsonii*, Diss = *P. dissectus*, Frut = *P. fruticosus*, Grin = *P. grinnellii*). Rows represent the species used as the database, and columns represent the species used as the query (e.g., row Cent, column Grin represents a BLASTN search with *P. grinnellii* as the query and *P. centranthifolius* as the database). Mean sequence variation ± SEs and sample size are shown in the upper right corner of each graph. Note that the matrix is not symmetric due to differences between using the same set of sequences as both a query and as a database for a BLAST search (e.g., Frut vs. Cyan ≠ Cyan vs. Frut).

as the COSII or *Arabidopsis-Populus-Vitis-Oryza* shared single-copy (APVO SSC) sets of low-copy nuclear genes (Wu et al., 2006; Duarte et al., 2010). However, our results showed that BLAST identified shorter gene regions on average and produced more variable results than MAKER2. The amount of output from BLAST (especially BLASTX; Table 3) was also orders of magnitude greater than that of MAKER2 when the number of hits was not restricted. Conducting BLAST searches can be sufficient for developing a set of sequencing loci, but a major advantage of using a pipeline such as MAKER2 for identifying gene regions is its use of a single output file that gathers all of the evidence provided by each source used by the software to annotate a contig. This becomes a particularly powerful resource when combining the output from MAKER2 with a visualization tool such as the Apollo Genome Browser, which can present the annotations for a given contig together with all of the evidence for the annotations (see Appendix S5). Furthermore, by identifying exon boundaries, finding variable regions to sequence becomes much easier as primers can be anchored in the exons bordering introns by direct visualization. For our data, the majority of the identified introns were in contigs from the chloroplast and mitochondrial genomes, which greatly facilitated primer design for those regions. Nuclear introns were much more difficult to characterize, but this problem could likely be circumvented by conducting deeper genomic sequencing. An additional benefit to using MAKER2 for our data is that we now have a gene prediction model that has been designed specifically for *Penstemon*. Such a model will be useful for any future WGS or other NGS projects involving the genus, and has the capability of being continually updated as we gather more data from transcriptome sequencing and higher-coverage WGS efforts.

It should be noted that the research here does not take full advantage of the entire suite of tools offered by MAKER2. The original intention of the program is to annotate full eukaryotic genomes, with contigs on the order of thousands to millions of base pairs long and libraries of high-coverage RNA-Seq data (Cantarel et al., 2008). More recently, a new version of MAKER, MAKER-P, has been released that is designed specifically to annotate plant genomes by taking into account the large amount of repetitive sequences that are often present (Campbell et al., 2014). Our utilization of the pipeline for NGS data that are characteristic of very low-coverage WGS may not be the typical application of such a program. However, it demonstrates that MAKER2 is capable of handling data from a wide range of NGS studies, not just the annotation of whole genomes. Thus, regardless of how developed the resources for an organism may be, MAKER2 can be successfully applied to help laboratories that are working with low-coverage WGS data to develop sets of markers.

Although our results show that MAKER2 is a useful program, there are a few things that should be considered before its use. The main drawback of using MAKER2 is that its many dependencies make the installation of the software nontrivial. Each individual program that is required by MAKER2 must be installed separately, and the installation of those programs may depend on others as well (Appendix S6). This multilevel dependency tree of software can create problems when trying to get MAKER2 to install properly when there is an issue with one of the underlying dependencies. The most common problems we experienced during the installation of MAKER2 were not having the proper compilers for programs written in C (GNU Compiler Collection [GCC]) and the occasional missing C library. Also, some of the Perl modules could not be installed through a direct connection to the Comprehensive Perl Archive Network (CPAN), requiring them to be downloaded and installed manually. MAKER2 is also only available on Unix-based operating systems such as Linux or Mac OS X and runs entirely from the command line. Thus, it bears the learning curve associated with running programs exclusively from a terminal window. Nevertheless, the documentation for MAKER2 on the Generic Model Organism Database (GMOD) website is quite helpful and has instructions for installing the dependencies as well as tutorials for running the program (http://gmod.org/wiki/MAKER). We have also provided example control files from our MAKER2 runs (Appendix S1) and a brief outline of our workflow (Appendix S6) to help other researchers learn to use this powerful program.

LITERATURE CITED

Altschul, S. F., W. Gish, W. Miller, E. W. Meyer, and D. J. Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215: 403–410.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* 25: 3389–3402.

Blischak, P. D., A. J. Wenzel, and A. D. Wolfe. 2014. Data from: Gene prediction and annotation in *Penstemon* (Plantaginaceae): A workflow for marker development from extremely low-coverage genome sequencing. Dryad Digital Repository. http://doi.org/10.5061/dryad.f6s22.

Broderick, S. R., M. R. Stevens, B. Geary, S. L. Love, E. N. Jellen, R. B. Dockter, S. L. Daley, and D. T. Lindgren. 2011. A survey of *Penstemon*'s genome size. *Genome* 54: 160–173.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. 2009. BLAST+: Architecture and applications. *BMC Bioinformatics* 10: 421.

Campbell, M. S., M. Yaw, C. Holt, J. C. Stein, G. D. Moghe, D. E. Hufnagel, J. Lei, et al. 2014. MAKER-P: A tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiology* 164: 513–524.

Cantarel, B. L., I. Korf, S. M. Robb, G. Parra, E. Ross, B. Moore, C. Holt, et al. 2008. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research* 18: 188–196.

Castoe, T. A., A. W. Poole, A. J. de Koning, K. L. Jones, D. F. Tomback, S. J. Oyler-McCance, J. A. Fike, et al. 2012. Rapid microsatellite identification from Illumina paired-end genomic sequencing in two birds and a snake. *PLoS ONE* 7: e30953.

Cronn, R. C., 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.

Datwyler, S. L., and A. D. Wolfe. 2004. Phylogenetic and biogeographic relationships of *Penstemon* subg. *Dasanthera*. *Systematic Botany* 29: 165–176.

Davey, J. W., P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen, and M. L. Blaxter. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews. Genetics* 12: 499–510.

Dockter, R. B., D. B. Elzinga, B. Geary, P. J. Maughan, L. A. Johnson, D. Tumbleson, J. Franke, et al. 2013. Developing molecular tools and insights into the *Penstemon* genome using genomic reduction and next-generation sequencing. *BMC Genetics* 14: 66.

Duarte, J. M., P. K. Wall, P. P. Edger, L. L. Landherr, H. Ma, J. C. Pires, J. Leebans-Mack, et al. 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.

Edgar, R. C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32: 1792–1797.

Giardine, B., C. Riemer, R. C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, et al. 2005. Galaxy: A platform for interactive large-scale genome analysis. *Genome Research* 15: 1451–1455.

Good, J. M. 2011. Reduced representation methods for subgenomic enrichment and next-generation sequencing. *In* V. Orgogozo and M. V. Rockman [eds.], Methods in molecular biology, vol. 772, 85–103. Humana Press, New York, New York, USA.

Holmgren, N. H. 1984. *Penstemon. In* A. Cronquist, A. H. Holmgren, N. H. Holmgren, J. L. Reveal, and P. K. Holmgren [eds.], Intermountain flora: Vascular plants of the intermountain west, vol. 4, 370–457. New York Botanical Garden, Bronx, New York, USA.

Holt, C., and M. Yandell. 2011. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12: 491.

Huang, X., and A. Madan. 1999. CAP3: A DNA sequence assembly program. *Genome Research* 9: 868–877.

Jennings, T. N., B. J. Knaus, T. D. Mullins, S. M. Haig, and R. C. Cronn. 2011. Multiplexed microsatellite recovery using massively parallel sequencing. *Molecular Ecology Resources* 11: 1060–1067.

Keck, D. D. 1932. Studies in *Penstemon*: A systematic treatment of the section *Saccanthera. University of California Publications in Botany* 16: 367–426.

Keck, D. D. 1936. Studies in *Penstemon*. II. The section *Hesperothamnus. Madrono* 3: 200–219.

Kent, W. J. 2002. BLAT–The BLAST-like alignment tool. *Genome Research* 12: 656–664.

Korf, I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* 5: 59.

Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg. 2004. Versatile and open software for comparing large genomes. *Genome Biology* 5: R12.

Leaché, A. D., R. B. Harris, B. Rannala, and Z. Yang. 2014. The influence of gene flow on species tree estimation: A simulation study. *Systematic Biology* 63: 17–30.

Lewis, S. E., S. M. J. Searle, N. Harris, M. Gibson, V. Iyer, J. Richter, C. Wiel, et al. 2002. Apollo: A sequence annotation editor. *Genome Biology* 3: research0082- research0082.14.

Lindgren, D., and E. Wilde. 2003. Growing penstemons: Species, cultivars and hybrids. American Penstemon Society, Infinity Publishing, Haverford, Pennsylvania, USA.

Maughan, P. J., S. M. Yourstone, E. N. Jellen, and J. A. Udall. 2009. SNP discovery via genomic reduction, barcoding and 454-pyrosequencing in amaranth. *Plant Genome* 2: 260–270.

Meyer, M., U. Stenzel, and M. Hofreiter. 2008. Parallel tagged sequencing on the 454 platform. *Nature Protocols* 3: 267–278.

Nold, R. 1999. *Penstemons*. Timber Press, Portland, Oregon, USA.

Olmstead, R. G., C. W. DePamphilis, A. D. Wolfe, N. D. Young, W. J. Elisons, and P. A. Reeves. 2001. Disintegration of the Scrophulariaceae. *American Journal of Botany* 88: 348–361.

O'Neill, E. M., R. Schwartz, C. T. Bullock, J. S. Williams, H. B. Shaffer, X. Aguilar-Miguel, G. Parra-Olea, and D. W. Weisrock. 2013. Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North American tiger salamander (*Ambystoma tigrinum*) species complex. *Molecular Ecology* 22: 111–129.

Pennell, F. W. 1920. Scrophulariaceae of the central Rocky Mountain states. *Contributions from the United States National Herbarium* 20: 313–381.

Pennell, F. W. 1935. The Scrophulariaceae of eastern temperate North America. Academy of Natural Sciences of Philadelphia, Philadelphia, Pennsylvania, USA.

Pruitt, K. D., T. Tatusova, G. R. Brown, and D. R. Maglott. 2012. NCBI Reference Sequences (RefSeq): Current status, new features and genome annotation policy. *Nucleic Acids Research* 40: D130–D135.

R Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Website http://www.R-project.org/ [accessed 10 November 2014].

Rozen, S., and H. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. *In* S. Misener and S. A. Krawetz [eds.], Methods in molecular biology, vol. 132: Bioinformatics methods and protocols, 365–386. Humana Press, Totowa, New Jersey, USA.

Shaw, J., E. B. Lickey, J. T. Beck, S. B. Farmer, W. Liu, J. Miller, K. C. Siripun, et al. 2005. The tortoise and the hare II: Relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany* 92: 142–166.

Slater, G. S. C., and E. Birney. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6: 31.

Smit, A. F. A., R. Hubley, and P. Green. 1996. RepeatMasker Open-3.0, version 3.3.0. Website http://www.repeatmasker.org [accessed 20 November 2012].

Stajich, J. E., D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz, C. Dagdigian, G. Fuellen, et al. 2002. The BioPerl toolkit: Perl modules for the life sciences. *Genome Research* 12: 1611–1618.

Stanke, M., and S. Waack. 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics (Oxford, England)* 19(Supplement 2): ii215–ii225.

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

Straw, R. M. 1956a. Adaptive morphology of the *Penstemon* flower. *Phytomorphology* 6: 112–119.

Straw, R. M. 1956b. Floral isolation in *Penstemon. American Naturalist* 90: 47–53.

Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar. 2011. MEGA5: Molecular Evolutionary Genetic Analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* 28: 2731–2739.

Weitemier, K., S. C. K. Straub, R. C. Cronn, M. Fishbein, R. Schmickl, A. McDonnell, and A. Liston. 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences* 2: 1400042.

Westwood, J. H., C. W. dePamphilis, M. Das, M. Fernández-Aparicio, L. A. Honaas, M. P. Timko, E. K. Wafula, et al. 2012. The Parasitic Plant Genome Project: New tools for understanding the biology of *Orobanche* and *Striga. Weed Science* 60: 295–306.

Wilson, P., and M. Valenzuela. 2002. Three naturally occurring *Penstemon* hybrids. *Western North American Naturalist* 62: 25–31.

Wolfe, A. D., and W. J. Elisens. 1994. Nuclear ribosomal DNA restriction-site variation in *Penstemon* section *Peltanthera* (Scrophulariaceae): An evaluation of diploid hybrid speciation and evidence for introgression. *American Journal of Botany* 81: 1627–1635.

Wolfe, A. D., and W. J. Elisens. 1995. Evidence of chloroplast capture and pollen-mediated gene flow in *Penstemon* sect. *Peltanthera* (Scrophulariaceae). *Systematic Botany* 20: 395–412.

Wolfe, A. D., Q.-Y. Xiang, and S. R. Kephart. 1998a. Assessing hybridization in natural populations of *Penstemon* (Scrophulariaceae) using hypervariable intersimple sequence repeat (ISSR) bands. *Molecular Ecology* 7: 1107–1125.

Wolfe, A. D., Q.-Y. Xiang, and S. R. Kephart. 1998b. Diploid hybrid speciation in *Penstemon* (Scrophulariaceae). *Proceedings of the National Academy of Sciences, USA* 95: 5112–5115.

Wolfe, A. D., C. P. Randle, S. L. Datwyler, J. J. Morawetz, N. Arguedas, and J. Diaz. 2006. Phylogeny, taxonomic affinities, and biogeography of *Penstemon* (Plantaginaceae) based on ITS and cpDNA sequence data. *American Journal of Botany* 93: 1699–1713.

Wu, F., L. A. Mueller, D. Crouzillat, V. Pétiard, 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.

Yandell, M., and D. Ence. 2012. A beginner's guide to eukaryotic genome annotation. *Nature Reviews. Genetics* 13: 329–342.

Ye, J., G. Coulouris, I. Zaretskaya, I. Cutcutache, S. Rozen, and T. L. Madden. 2012. Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* 13: 134.

Yuan, Y. W., C. Liu, H. E. Marx, and R. G. Olmstead. 2009. The pentatricopeptide repeat (PPR) gene family, a tremendous resource for plant phylogenetic studies. *New Phytologist* 182: 272–283.