A Protocol for Targeted Enrichment of Intron-Containing Sequence Markers for Recent Radiations: A Phylogenomic Example from Heuchera (Saxifragaceae)

Authors: Folk, Ryan A., Mandel, Jennifer R., and Freudenstein, John V.

Source: Applications in Plant Sciences, 3(8)

Published By: Botanical Society of America

URL: https://doi.org/10.3732/apps.1500039
A protocol for targeted enrichment of intron-containing sequence markers for recent radiations: A phylogenomic example from Heuchera (Saxifragaceae)

RYAN A. FOLK2,4, JENNIFER R. MANDEL3, AND JOHN V. FREUDENSTEIN2

2Herbarium, The Ohio State University, Columbus, Ohio 43212 USA; and 3Department of Biology, University of Memphis, Memphis, Tennessee 38152 USA

Key words: COS marker; Heuchera; intron enrichment; Saxifragaceae; targeted enrichment.

The necessity of multilocus data sets for phylogenetic analysis is now largely taken for granted. Low-copy nuclear sequences have long been advocated for the future of plant phylogenetic reconstruction, especially at low taxonomic levels where their high rate of evolution (particularly for introns) results in more-resolved phylogenetic hypotheses (Sang, 2002). These types of markers are also critical for accounting for phylogenetic discord across the genome, caused by such widespread phenomena as lineage sorting and hybridization. From the coalescent point of view, the canonical nonrecombining organellar genome (such as the typical angiosperm plastome, which has been heavily relied upon for developing Sanger markers) is essentially a single long phylogenetic marker with a single underlying history. This means that recent progress in the massive sequencing of entire organellar genomes (e.g., Ruhfel et al., 2014) does not necessarily equate in itself to adequate genomic sampling of phylogenetic signal for some questions, particularly at lower taxonomic levels, although such data sets may achieve high levels of phylogenetic resolution. The cost of whole genome shotgun sequencing has fallen dramatically in recent years, yet it remains costly to produce complete eukaryotic genomes at the scale typical for population and phylogenetic study (but see Cao et al., 2011), necessitating economical methods for broad subsampling of genomic loci.

Genome skimming (Straub et al., 2012) is simple and can result in large amounts of rDNA, chloroplast, and mitochondrial data, yet this alone may not offer enough independent loci if either coalescent methods or the gene trees themselves are of interest.

Reduced-representation methods for next-generation sequencing (NGS) have emerged as a new standard for the collection of phylogenetic and population-genetic data. Restriction-based methodologies such as RADSeq are most commonly used for population studies, and targeted sequencing methods, usually based on target enrichment via synthesized RNA probes, have gained traction at higher levels for phylogenetics (McCormack et al., 2013, table 1; Cronn et al., 2012, table 5). For the taxonomic middle ground, at the level of species complex, appropriate methodologies remain unclear. While the best methodology may be lineage-specific to some extent, there remains a paucity of literature examining NGS methodology performance for
Aspects of target enrichment relating to wet-laboratory methods, off-target genome assembly, and phylogenetic reconstruction are now well covered in exon-based enrichment methods (Mandel et al., 2014; Weitemier et al., 2014). Some recent studies have also included intronic probes (e.g., Govindarajulu et al., 2015); our work contributes to this publication series by focusing on the effectiveness of intron enrichment and long-target enrichment for closely related species. While we were collecting these data, a study also including intron probes was published for *Medicago* L. (de Sousa et al., 2014); these authors also point out the use of introns for lower-level taxonomic questions. However, this study only reports experimental statistics on entire loci—discussion and data on the effectiveness of intron enrichment are absent. Additionally, their locus design methodology cannot be used on the majority of organisms that lack a high-quality annotated reference genome, underlining the need for broader methodologies and data exploration for use by the phylogenomic community.

### METHODS

**Materials**—The DNAs we used were primarily of high quality and prepared from cultivated material or quickly dried field material. However, we also successfully sequenced a herbarium specimen of *H. acutifolia* Rose, collected in 1984; we have had further success with specimens collected as long ago as the 1940s (unpublished). The same *H. acutifolia* sample that was successful for this study previously proved impossible to sequence for low-copy markers using a Singer protocol (Folk and Freudenstein, 2014; similar observations in Cronn et al., 2012). As an initial test of our method, we present results from 15 individuals selected as taxonomically representative of the major sections of the genus (Appendix 1).

**Genomic reference development**—As a draft genomic reference, we used 3.7 Gbp of short-read data (20,288,896 reads) from *H. parviflora var. saurensis* R. A. Folk. Paired-end sequencing was performed by The Ohio State University Molecular and Cellular Imaging Center in six multiplexed runs with samples from unrelated projects (five Illumina MiSeq runs with 300-bp reads; one Illumina HiSeq lane with 100-bp reads; all data uploaded to the Short Read Archive, accession SRP058301). This level of sequencing corresponds to approximately 7.5× expected average nuclear genomic coverage using 2× values from Godsoe et al. (2013). While over-sampling of high-copy loci will reduce the actual coverage of the single-copy portion of the genome, most nuclear contigs consistently showed 5–8× coverage. These data were assembled in Velvet (Zerbino and Birney, 2008) using a kmer length of 64, expected insert length of 700 bp, standard deviation of insert length 100 bp, and minimum contig length of 1000, to develop reference contigs for the low-copy portion of the genome. This resulted in 18,344 contigs between 1000 and 7360 bp; a longer kmer length results in longer contigs, but these were not necessary for marker development. A BLAST search against organelar references confirmed that organelar and rDNA sequences were not present in this batch of contigs.

Because we wished to explore the possibility of off-target assembly of high-copy loci, we used the skimming data from *H. parviflora var. saurensis* to assemble complete chloroplast and mitochondrial genomes (GenBank accessions KR478645 and KR559021, respectively). The chloroplast genome was assembled by mapping all reads (trimmed with Geneious default settings) using the native Geneious assembler (version 7); all contigs were distinguished by BLAST searches against the *Vitis vinifera* L. mitochondrial genome (GenBank accession NC_007913), and the complete lack of mis-mapping regions was confirmed. The resultant contig was manually screened for regions of mis-mapped reads (always due to large indels in intergenic spacers); at these points the consensus sequence was broken into subsequences. These subsequences were subjected to iterative assembly in Geneious (25 iterations), repeated until all consensus sequences of these assemblies overlapped (by de novo assembly in Geneious with default settings) and could be combined into a single circle. The read mapping was repeated on this resultant genome draft, and the complete lack of mis-mapping regions was confirmed.

The mitochondrial assembly method was largely similar to that of the chloroplast; however, a reasonably close reference mitochondrial genome was not available for the initial read-mapping. Geneious de novo assembly was performed with default settings on the entire data set, divided into three approximately equal parts to reduce RAM requirements to feasible levels. Mitochondrial contigs were distinguished by BLAST searches against the *Vitis vinifera* L. mitochondrial genome (GenBank accession NC_007913). This procedure was found to produce very long mitochondrial contigs compared to preliminary Velvet and SOAP-Liger assemblies as long as 122,611 bp. Comparing the assemblies revealed that the Geneious contigs contained all of the mitochondrial sequence assembled by the other programs, but in fewer pieces. Consensus sequences of these contigs (*n* = 39) were used in iterative assembly, circular reference construction, and assembly validation as above.
Organellar genome annotations were performed using Geneious, based on *H. sanguinea* (chloroplast) or *V. vinifera* (mitochondrion); synteny comparisons were performed using the Mauve plug-in (Darling, 2004) in Geneious.

**Locus design**—To obtain putatively single-copy loci in the absence of comparative genomic data within Saxifragaceae, Velvet contigs (n = 18,334) were searched against a panel of conserved single-copy sequences (hereafter “COS loci”) from the Arabidopsis genome (Kozik et al., unpublished; but see http://www.cgd.bud.c_VARS_070_SX/ Arabidopsis; also used in Mandel et al., 2014; n = 3714) using BLAST, yielding 3251 hits that were retained for RNA transcript queries. It is important to note that for these retained loci, we had two sources of assurance of their low-copy (likely single-copy) status: (a) they are single-copy in Arabidopsis, and (b) they had coverage consistent with single-copy status in the *H. parviflora* Bartl., genome scaffolding work. This means that a fully or mostly assembled genome is not needed to identify low-copy targets, although such resources are useful when they are available (Weitemier et al., 2014). We used GeneSeqer (Brendel et al., 2004) to identify intron-containing loci among these hits. GeneSeqer aligns transcriptome sequences to genomic sequences using several available splice-site models, evaluates putative alignments based both on sequence similarity and splice-site structure, and outputs gapped sequence alignments. In our experience, these alignments are more accurate in the context of large indels (= introns) than those based on simple sequence alignment such as BLAST. The splice-site model we used was based on Arabidopsis thaliana (L.) Heynh. Our transcriptome sequences came from a SOX-Seq (Liu et al., 2013) transcriptome (courtesy of the Weitz project, D. Doltis and G. Wong; assembly parameters available at https://pods.pi. plantcollaborative.org/ /wiki/display/iptol/Access+to+on+1+K+data+set), and the genomic data were the set of *H. parviflora* Velvet contigs that matched the COS marker data set. All GeneSeqer alignments were evaluated by eye to verify acceptable exon sequence identity (i.e., no obvious paralogy or spurious alignment) between *H. sanguinea* and *H. parviflora* var. *saurensis*. We further favored loci with at least two introns >100 bp long for locus development. A minority of chosen loci had a single longer intron, and two loci were included that were entirely exonic but appeared unusually variable based on alignment. We designed a 278-locus panel, synthesized by MYcroarray (Ann Arbor, Michigan) as 8634 120-bp oligomers tiled at 3× coverage of the target loci. The mean locus length was 1362 bp (range 555–3672 bp), for a total targeted length of 378,553 bp.

**Library preparation**—Genomic DNAs were extracted as described previ- ously (Folk and Freudenstein, 2014). DNA was quantified with a Qubit BR- assay (Life Technologies, Carlsbad, California), and quality was evaluated using a Nanodrop 2000 (Thermo Scientific, Waltham, Massachusetts, USA). Problematic DNAs were cleaned with a QIAquick kit (QIAGEN, Valencia, California, USA); any DNAs too dilute for library preparation were concentrated by vacuum centrifugation. One microgram of clean genomic DNA in 60 μL was sonicated using a Covaris machine (model S220; Covaris, Woburn, Massachusetts, USA), aiming for fragment size of 500–700 bp. Libraries were prepared with 55.5 μL of sonicated DNA, using a NEBNext Ultra kit (New England Biolabs, Ipswich, Massachusetts, USA), following the manufacturer’s protocols with the following modifications: size selection aim for a 500–700 bp range (i.e., 30 μL AMP beads [Beckman Coulter, Brea, California, USA]) for the first size-selection step, and 15 μL for the second step), and PCR amplification mostly used six cycles (eight cycles for the herbarium specimen). Libraries were barcoded using NEBNext Multiplex oligos (New England Biolabs). Lib- rary quality was verified using an Agilent Bioanalyzer (Agilent Technologies, Santa Clara, California, USA), and libraries were again quantified by Qubit. Any libraries that were too dilute for sequence hybridization were concentrated in a vacuum centrifuge.

**Probe-based target enrichment**—Verified libraries were used in a MYcroarray MyBaits in-solution hybridization protocol using our custom phylogenetic loco- cus panel, following manufacturer protocols (version 2) with the following modifications: eight libraries were pooled per MyBaits reaction (60 ng each for 480 ng total input DNA in 6 μL) to reduce reagent cost and required input DNA, and hybridization was conducted for 36 h. PCR conditions used an annealing temperature of 60 °C, an elongation time of 45 s, 15 total cycles, and 10 μL reaction template. Target-enriched library pools were verified for size distribution on an Agilent Bioanalyzer, and quantified with an Illumina library quantification kit (KAPA, Wilmington, Massachusetts, USA). Because adapter dimers were problematic for one of the pools, all pools were cleaned following the paramagnetic bead-based cleaning procedure in the NEBNext protocol but with 0.8 volumes of AMPure beads to 1 volume of enriched library. For all target enrichment pools, efficiency of enrichment was evaluated with two qPCR assays (qPCR primers, thermocycler profile, and other information in Appendix S1). Enriched samples that amplified at least 10 cycles earlier than unenriched samples (>1000-fold enrichment assuming 100% primer efficiency) were considered successful.

Target-enriched library pools were again pooled and diluted for sequencing on an Illumina MiSeq (W. Harry Feinstone Center for Genomic Research, University of Memphis, Tennessee; 500-cycle kit, Illumina V2 chemistry). The MiSeq controller software was set to trim adapters. We downloaded all raw barcodes; the raw output has been uploaded to the National Center for Biotechnology Information (NCBI) Sequence Read Archive (accession SRP057104). Read end trimming was performed in Trimmomatic 0.33 (Bolger et al., 2014) with a sliding window of 20 bp and a quality score of Q20 or greater.

**Low copy marker sequence assembly**—Assembly of data for each species reused as a reference the initial 278 contigs of *H. parviflora* var. *saurensis* that were used for probe design. We reasoned that contigs that were sufficient for designing efficient enrichment probes should also be able to inform se- quence assembly, avoiding the need for computationally demanding and often unsatisfactory de novo methods. We mapped reads to the 278 references using the relatively indel-tolerant BWA program (v. 0.7.12; Li and Durbin, 2009; Li, 2013), with the default settings for paired-end data; finally, the resultant contigs were imported into Geneious. Consensus sequences were extracted from these contigs in Geneious using default settings, including trimming the consensus sequence to the length of the reference sequence. Each locus was aligned individually using MAFFT 7.017 (Katoh et al., 2009) with default settings. One of the loci (labeled Locus 4 in the probe panel, wherein loci are numbered in order of descending length; data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.4cn66 [Folk et al., 2015]) was found to yield only partial sequences from many taxa, resulting in large amounts of missing data (see Results); this has been excluded from the analyses, leaving 277 effectively enriched loci.

**Off-target sequence assembly**—We repeated the reference-mapping in BWA with mitochondrial and chloroplastic references to attempt to construct plastid genomes and mitochondrial genomes from off-target reads. For the chloroplast assembly, we used fully assembled plastomes from *H. parviflora* var. *saurensis*, *H. parishii* Rydb. (below), and *H. sanguinea* (M. Moore, unpub- lished data). The closest of these three references to each sample based on a preliminary Sanger chloroplast phylogeny (unpublished data) was used for reference-mapping to assist in assembling long, divergent intergenic regions. Development of a third plastome reference was necessary for the divergent in- tergenic regions of *H. parishii*, *H. elegans* Abrams, and *H. abramsii* Rydb.; this was constructed by mapping the *H. parishii* reads to the *H. parviflora* var. *saurensis* plastome in BWA and repeating the assembly refinement methods used for the *H. parviflora* plastome. Mitochondrial assembly used the fully assembled reference from *H. parviflora* var. *saurensis*. Because we desired purely mitochondrial signal for the mitochondrial phylogenetic analysis, this genome was blasted against the *H. parviflora* chloroplast genome and any hits were deleted from the mitochondrial assembly reference. Whole chloroplast genomes were aligned in MAFFT as with the single-copy markers; however, this was computationally infeasible for the much longer mitochondrial se- quences. For these we used Mauve, setting an assumption of collinear genomes (which was enforced for our contigs by the read-mapping method) but other- wise using default settings.

**Phylogenetic analysis**—Because chloroplast capture is a well-known pheno- menon in the genus *Heuchera* (Soltis et al., 1991; Soltis and Kuzoff, 1995; Folk, Mandel, and Freudenstein, unpublished data), and organellar genomes in general are thought to be more prone to displaying phylogenetic discordance caused by hybridization (Rieseberg and Soltis, 1991), we avoided including organellar genomes in all multilocus analyses and instead present these sepa- rately. We performed a concatenated analysis of the 277 low-copy nuclear loci in RAxML 8.1.3 (Stamatakis, 2006), using an unpartitioned GTR-GAMMA model and with support quantified by 1000 bootstrap replicates (matrices and tree files for the analyses of this paper available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.4cn66 [Folk et al., 2015]).

To examine the relative phylogenetic signal from exonic and intronic por- tions of the sequences, we performed an exon-only analysis and an intron-only analysis by alternatively excluding intron and exon sites, respectively. Approx- imately two thirds of our target sequence was intronic; hence analyzing this in its entirety would be an unfair comparison with the exonic matrix. Therefore, we
arbitrarily excluded the last 49.3% of the intron matrix (introns of approximately loci 88–277) to yield a matrix of equivalent length to the exon matrix. Each of these two matrices was analyzed in RaxML with an unpartitioned GTR-GAMMA model, with 1000 bootstrap replicates. The 277 gene alignments were also analyzed individually in RaxML to infer gene trees, with the GTR-GAMMA model and 1000 bootstrap replicates. The optimal trees and bootstrap samples from each gene tree were analyzed using a gene tree–based coalescent approach in MP-EST (Liu et al., 2010) and STAR (Liu et al., 2009) using the STRAW server (Shaw et al., 2013).

Miella pentandra Hook. was used as the outgroup for the nuclear analysis based on the results in Folk and Freudenstein (2014). Rooting is not as straightforward for the chloroplast data set because *Heuchera* is known to be polyphyletic for chloroplast markers due to frequent hybridization with closely related genera such as *Tiarella* L. (Soltis et al., 1991; Folk, Mandel, and Freudenstein, unpublished data). If a supposed outgroup had captured a *Heuchera* chloroplast genome, this would result in incorrectly rooting the tree within *Heuchera*. While a mitochondrial analysis has never been undertaken for the *Heuchera* group of genera, these concerns could also apply to the mitochondrial data.

To address these concerns, we repeated the chloroplast and mitochondrial read-mapping approach for publicly available short read data from *Saxifraga granulata* L. (one individual from Meer et al., 2014) to develop an appropriate outgroup sequence well outside the *Heuchera* group of genera.

The low-coverage of mitochondrial DNA and the greater divergence from the reference resulted in a greater amount of missing data than in other data sets, especially in intergenic regions; this was addressed by deleting all columns with 275% missing data (a trial was also made of deleting columns with 50% missing data, but this resulted in lowered support values on an identical topology; results not shown). Otherwise, the same phylogenetic analysis parameters from the low-copy nuclear markers were used for the mitochondrial and chloroplast data sets.

Any differences in support values between this analysis and the previous phylogenetic context of Folk and Freudenstein (2014) could be caused by reduced taxon sampling rather than a larger data set, as analyzing fewer taxa reduces the tree search space and may lengthen short branches that would have been broken up by intermediate taxa. To evaluate this possibility, we reanalyzed a reduced version of the earlier matrix, using only DNA characters and eliminated taxon sampling rather than a larger data set, as analyzing fewer taxa reduces the tree search space and may lengthen short branches that would have been broken up by intermediate taxa. To evaluate this possibility, we reanalyzed a reduced version of the earlier matrix, using only DNA characters and eliminated all taxa not sampled in this study.

**RESULTS**

**Target enrichment**—Our target enrichment methodology resulted in between 45% and 63% on-target sequences. This range is relatively high for interspecies target enrichment (e.g., compare Weitemier et al., 2014), and may result from the particularly low divergence among species of *Heuchera*, causing high complementarity between probes and target sequences. This resulted in relatively high coverage numbers for the study loci; although coverage numbers were variable, all samples had greater than 100× average coverage (Table 1). Low coverage for target loci was extremely rare, other than for the excluded locus (see Methods). This locus appears to be a pseudogene; no well-matched short reads of this locus were obtained for about half of the samples.

The clear explanation for the difference in coverage between samples is uneven pooling, rather than on-target sequence percentage, which did not vary particularly greatly between species. Hence the success of target enrichment was not markedly different

---

**Table 1. Statistics for the target enrichment experiment.**

<table>
<thead>
<tr>
<th>Accession</th>
<th>Mean (SD)</th>
<th>Mean read-mapping alignment coverage (%)</th>
<th>Mean ratio of low-copy locus coverage (% of reads)</th>
<th>Percent chloroplast completeness (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Heuchera americana</em></td>
<td>54.7 ± 136.5</td>
<td>99.9 ± 0.6</td>
<td>99.8 ± 0.6</td>
<td>60.5</td>
</tr>
<tr>
<td><em>H. elegans</em></td>
<td>54.8 ± 321.3</td>
<td>100.0 ± 0.0</td>
<td>99.8 ± 0.6</td>
<td>60.5</td>
</tr>
<tr>
<td><em>H. grossulariifolia</em></td>
<td>52.8 ± 254.1</td>
<td>100.0 ± 0.0</td>
<td>99.7 ± 0.0</td>
<td>60.5</td>
</tr>
<tr>
<td><em>H. parishii</em></td>
<td>58.7 ± 168.1</td>
<td>100.0 ± 0.0</td>
<td>99.7 ± 0.0</td>
<td>60.5</td>
</tr>
<tr>
<td><em>H. parviflora</em></td>
<td>55.8 ± 76.6</td>
<td>100.0 ± 0.0</td>
<td>99.7 ± 0.0</td>
<td>60.5</td>
</tr>
<tr>
<td><em>H. parvifolia var. nivalis</em></td>
<td>55.7 ± 275.8</td>
<td>100.0 ± 0.0</td>
<td>99.8 ± 0.0</td>
<td>60.5</td>
</tr>
<tr>
<td><em>H. pulchella</em></td>
<td>47.8 ± 76.6</td>
<td>99.7 ± 0.0</td>
<td>99.7 ± 0.0</td>
<td>60.5</td>
</tr>
<tr>
<td><em>H. villosa</em></td>
<td>62.6 ± 361.9</td>
<td>100.0 ± 0.0</td>
<td>100.0 ± 0.0</td>
<td>60.5</td>
</tr>
<tr>
<td><em>H. wootonii</em></td>
<td>55.7 ± 275.8</td>
<td>100.0 ± 0.0</td>
<td>99.8 ± 0.0</td>
<td>60.5</td>
</tr>
</tbody>
</table>

---

**Table 2.** Enrichment statistics.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Complete reads (%)</th>
<th>On-target (%)</th>
<th>Mean read-mapping alignment coverage (%)</th>
<th>Mean ratio of low-copy locus coverage (% of reads)</th>
<th>Percent chloroplast completeness (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Heuchera americana</em></td>
<td>54.7 ± 136.5</td>
<td>99.9 ± 0.6</td>
<td>99.8 ± 0.6</td>
<td>60.5</td>
<td></td>
</tr>
<tr>
<td><em>H. elegans</em></td>
<td>54.8 ± 321.3</td>
<td>100.0 ± 0.0</td>
<td>99.8 ± 0.6</td>
<td>60.5</td>
<td></td>
</tr>
<tr>
<td><em>H. grossulariifolia</em></td>
<td>52.8 ± 254.1</td>
<td>100.0 ± 0.0</td>
<td>99.7 ± 0.0</td>
<td>60.5</td>
<td></td>
</tr>
<tr>
<td><em>H. parishii</em></td>
<td>58.7 ± 168.1</td>
<td>100.0 ± 0.0</td>
<td>99.7 ± 0.0</td>
<td>60.5</td>
<td></td>
</tr>
<tr>
<td><em>H. parviflora var. nivalis</em></td>
<td>55.7 ± 275.8</td>
<td>100.0 ± 0.0</td>
<td>99.8 ± 0.0</td>
<td>60.5</td>
<td></td>
</tr>
<tr>
<td><em>H. pulchella</em></td>
<td>47.8 ± 76.6</td>
<td>99.7 ± 0.0</td>
<td>99.7 ± 0.0</td>
<td>60.5</td>
<td></td>
</tr>
<tr>
<td><em>H. villosa</em></td>
<td>62.6 ± 361.9</td>
<td>100.0 ± 0.0</td>
<td>100.0 ± 0.0</td>
<td>60.5</td>
<td></td>
</tr>
<tr>
<td><em>H. wootonii</em></td>
<td>55.7 ± 275.8</td>
<td>100.0 ± 0.0</td>
<td>99.8 ± 0.0</td>
<td>60.5</td>
<td></td>
</tr>
</tbody>
</table>

---

1. For all calculations, lower-case f was used to account for the reference set of gene loci, while upper-case f was used to account for the reference set of gene loci.
2. Coverage completeness figures are based on the percent of the *H. parvifolia* reference. If this calculation included the whole mitochondrial, it would be unfairly down-weighted by long regions (up to tens of thousands of bases) of uncertain origin that are not shared across species.
labels in sect. (Fig. 1) was completely resolved with strong support on all
also quite rich in indel characters, >40-fold more so than exons.
about four times the percent parsimony-informative characters.
result of introns is fourfold that of exons. Similarly, introns had
matrix (Table 1) show that for all samples (Table 1); this applies both to exons and introns, the
Overall locus assembly completeness was also nearly 100% for
ment assuming constant background prevalence of target loci
enrichment was very high (0.547/0.00057 ~960-fold enrich-
unenriched genome-skimming sample (0.057%, vs. average
unenriched genome-skimming) assemble the vast majority of intronic sequences.
Where slight signs of intron dropout are apparent (~1% lower
Comparison with the proportion of on-target reads for the unenriched genome-skimming sample (0.057%, vs. average
54.7% on target for enriched individuals) showed that the fold
enrichment was very high (0.547/0.00057 ~960-fold enrichment
assumed constant background prevalence of target loci across samples), consistent with >10 cycles increase in target
centration in the qPCR assays for all enriched individuals.
Overall locus assembly completeness was also nearly 100% for all samples (Table 1); this applies both to exons and introns, the
latter of which had negligibly lower completeness.
Percent pairwise identity calculations from the phylogenetic
matrix (Table 1) show that for Heuchera species, the percent
divergence of introns is fourfold that of exons. Similarly, introns had
about four times the percent parsimony-informative characters.
While nucleotide diversity is high in introns, as expected they are also quite rich in indel characters, >40-fold more so than exons.
Phylogenetic inference—The concatenated nuclear tree
(Fig. 1) was completely resolved with strong support on all
between samples. Even for the outgroup Mitella pentandra,
where slight signs of intron dropout are apparent (~1% lower
assembly completeness, Table 1), it remained possible to as-
semble the vast majority of intronic sequences.
Comparison with the proportion of on-target reads for the
unenriched genome-skimming sample (0.057%, vs. average
54.7% on target for enriched individuals) showed that the fold
enrichment was very high (0.547/0.00057 ~960-fold enrich-
ment assuming constant background prevalence of target loci across samples), consistent with >10 cycles increase in target
centration in the qPCR assays for all enriched individuals.
Overall locus assembly completeness was also nearly 100% for all samples (Table 1); this applies both to exons and introns, the
latter of which had negligibly lower completeness.
Percent pairwise identity calculations from the phylogenetic
matrix (Table 1) show that for Heuchera species, the percent
divergence of introns is fourfold that of exons. Similarly, introns had
about four times the percent parsimony-informative characters.
While nucleotide diversity is high in introns, as expected they are also quite rich in indel characters, >40-fold more so than exons.
Phylogenetic inference—The concatenated nuclear tree
(Fig. 1) was completely resolved with strong support on all
nodes (99% on one node in sect. Rhodoheuchera concerning
the placement of H. rubescens Torr. var. versicolor (Greene)
M. G. Stewart; otherwise 100%). Coalescent analyses in MP-
EST and STAR (Fig. 2) likewise had primarily high support
values. There was only one point of discord between these
methods; coalescent methods found H. pulchella Wooton &
Standl. and H. rubescens var. versicolor as sister to each other,
whereas concatenation inferred these as successively sister to
H. parishii, H. elegans, and H. abramsii. The phylogenetic
analyses of organellar data (chloroplast, Fig. 3; mitochondrial,
Fig. 4) likewise resulted in high support values, although these
relationships differed greatly from those suggested by nuclear
data.

Organellar reference genomes—The chloroplast genome of
H. parviflora var. saurensis (154,696 bp; mean coverage
139.8×; no map shown) is conventional among angiosperms
in terms of structure and size, with complete synteny shared
with other members of the order. The H. parviflora var. sauren-
sis mitochondrial genome (541,954 bp; mean coverage 141.9×,
draft map in Fig. 5) is much less conserved in terms of struc-
ture, as is typical for plant mitochondrial genomes (Sloan, 2013),
with only a few very short regions of synteny and only short
stretches of similar sequence preserved when compared to the
closest reference, V. vinifera (Appendix S2). It has all func-
tional mitochondrial protein-coding genes that are present in Vitis,
as well as numerous interspersed sequences of chloroplast origin.
BLAST searches against the Marchantia polymorpha L. mito-
chondrial genome showed that all conserved mitochondrial genes
absent in Vitis were also absent in Heuchera. Processes of intra-
molecular recombination and multiple conformations of the
plant mitochondrial genome are thought to be mediated by moder-
ately long repeat regions (Alverson et al., 2011, and citations
therein), of which we found four. These repeat regions were in-
dependently supported by locally higher read coverage. Unusually,
the smaller three repeat regions (6428 bp, 1196 bp, 550 bp)
that we found overlap with the largest repeat (32,849 bp), but
with a single copy of each elsewhere in the genome. The issue
of multiple genome conformations is not critical for assem-
bles intended for analysis with phylogenetic methods. Using
these two organellar references, we were able to infer resolved

Table 2. Statistics for the phylogenetic alignments of low-copy nuclear loci.

<table>
<thead>
<tr>
<th>Sequence statistic</th>
<th>Entire experiment</th>
<th>Exons only</th>
<th>Introns only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent pairwise identity (pairwise divergence)</td>
<td>96.4% (3.6%)</td>
<td>98.8% (1.2%)</td>
<td>95.2% (4.8%)</td>
</tr>
<tr>
<td>Number (percent) parsimony informative characters</td>
<td>17,691 (4.6%)</td>
<td>1830 (1.5%)</td>
<td>15,861 (6.1%)</td>
</tr>
<tr>
<td>Percent invariant characters</td>
<td>82.90%</td>
<td>93.40%</td>
<td>77.80%</td>
</tr>
<tr>
<td>Undetermined sequence (N or ?)</td>
<td>0.20%</td>
<td>0.10%</td>
<td>0.30%</td>
</tr>
<tr>
<td>Gap characters (-)</td>
<td>3.00%</td>
<td>0.10%</td>
<td>4.30%</td>
</tr>
<tr>
<td>Overall alignment length</td>
<td>387,941 bp</td>
<td>126,540 bp</td>
<td>261,401 bp</td>
</tr>
<tr>
<td>Average coverage per locus</td>
<td>511.8x</td>
<td>1362 bp</td>
<td>0.06%</td>
</tr>
</tbody>
</table>

Fig. 1. Concatenated maximum likelihood (ML) tree based on 277 low-copy nuclear loci; branch lengths in the figure represent ML branch length estimates. Coloring of branches represents the taxonomic sections recognized in Folk and Freudenstein (2014); orange = sect. Holochloa; blue = sect. Heuchera; green = sect. Bracteatae; magenta = sect. Rhodoheuchera. Values plotted on branches are support values based on 5000 bootstrap replicates. Varetial assignments in sections Heuchera and Rhodoheuchera reflect the taxonomy of the treatments by Rosendahl et al. (1936) and Wells (1984); taxon labels in sect. Holochloa reflect Folk and Freudenstein (in press).
However, the increase in support was remarkable in several instances. The improvement in support over earlier work was most remarkable for section *Rhodoheuchera* (Fig. 1, marked in magenta), the largest in the genus (~17 species). In Folk and Freudenstein (2014), this section was the most problematic group for phylogenetic inference; there was weak or no support for all relationships in the section, and at best moderate support for the monophyly of the section as a whole. The new data set confidently confirms the monophyly of sect. *Rhodoheuchera* and resolves relationships in this group.

The analysis of the reduced Folk and Freudenstein (2014) matrix (Appendix S3) showed that higher support values were due to a mixture of reduced taxon sampling and more informative data. The average nodal bootstrap (BS) proportion was 100% for the entire phylogenomic data set and 78.4% for the Sanger data set; several nodes in the Sanger data set were higher than in earlier work. Nevertheless, the improvement caused by greater locus sampling is apparent; the addition of further taxa would inflate the tree search space, likely causing a greater disparity between the two data sets. It is important to note that the Sanger tree was incongruent in several places with the phylogenomic data set; particularly, section *Heuchera* (blue, Appendix S3) is sister to the rest of the genus (BS 76). Likewise, the positions of *H. rubescens* var. *versicolor* (BS 45) and *H. parviflora* var. *parviflora* (BS 99) differed. The position of section *Heuchera* phylogenies with high support values as well (chloroplast, Fig. 4; mitochondrion, Fig. 5).

**DISCUSSION**

**Low-copy locus design**—Most eukaryotic nuclear genes have introns; this splice site structure is highly conserved (The Arabidopsis Initiative, 2000; Roy et al., 2003), making these regions a practically inexhaustible source of phylogenetic signal. Fully assembled genomes are not required for marker development. At minimum, however, intronic locus design for targeted sequencing requires the researcher to obtain a large number of contigs at least as long as the desired marker length, as well as transcriptomic sequences to identify splice sites. Finally, some method must be used to select putative low-copy genes, such as curated genome drafts from the study organism (de Sousa et al., 2014; Weitemier et al., 2014) or, when this information is absent, panels of loci derived from model organism genomes (Mandel et al., 2014; this study). In the latter case, if genome skimming data are available from one of the study organisms, coverage data are effective as an independent source of evidence for low-copy marker status.

**Concatenated nuclear analysis**—The relationships recovered were broadly congruent with results based on Sanger sequencing (Folk and Freudenstein, 2014); however, the increase in support was remarkable in several instances. The improvement in support over earlier work was most remarkable for section *Rhodoheuchera* (Fig. 1, marked in magenta), the largest in the genus (~17 species). In Folk and Freudenstein (2014), this section was the most problematic group for phylogenetic inference; there was weak or no support for all relationships in the section, and at best moderate support for the monophyly of the section as a whole. The new data set confidently confirms the monophyly of sect. *Rhodoheuchera* and resolves relationships in this group.

The analysis of the reduced Folk and Freudenstein (2014) matrix (Appendix S3) showed that higher support values were due to a mixture of reduced taxon sampling and more informative data. The average nodal bootstrap (BS) proportion was 100% for the entire phylogenomic data set and 78.4% for the Sanger data set; several nodes in the Sanger data set were higher than in earlier work. Nevertheless, the improvement caused by greater locus sampling is apparent; the addition of further taxa would inflate the tree search space, likely causing a greater disparity between the two data sets. It is important to note that the Sanger tree was incongruent in several places with the phylogenomic data set; particularly, section *Heuchera* (blue, Appendix S3) is sister to the rest of the genus (BS 76). Likewise, the positions of *H. rubescens* var. *versicolor* (BS 45) and *H. parviflora* var. *parviflora* (BS 99) differed. The position of section *Heuchera*
is congruent with the earlier study, while the positions of the other two were equivocal. The relatively high bootstrap for two of these anomalous Sanger relationships is unlikely to have been caused by reduced taxon sampling; it is more likely that the smaller data set inadequately sampled the genome for conflicting phylogenetic signals for the positions of these taxa.

Intron- and exon-only phylogenetic analyses (Appendices S4, S5) revealed further differences in these data types. Support values for the two analyses were only somewhat different (the intron-only analysis had 96.8% average nodal support; the exon-only analysis had 94.8% average support). However, only the intron-only analysis was congruent with the full data set result. For the exon-only analysis, *H. rubescens* var. *versicolor* and *H. acutifolia* exchanged topological positions (BS 78). Because the full analysis also addressed gene tree heterogeneity through coalescent analysis, the exon-only result does not appear to be optimal.

Coalescent nuclear analysis—While we are aware of the issues associated with using long genetic loci with multispecies coalescent methods and the importance of incorporating gene tree estimation error (Gatesy and Springer, 2014), nevertheless we contend that a coalescent analysis retains heuristic value, particularly in concert with the more standard concatenation methods. Especially when only well-supported nodes are considered, for this data set the difference between coalescent and concatenated estimates of the species tree is not particularly great. It is interesting that the only point of discord between concatenated and coalescent approaches was one of only two nodes in the coalescent analysis below 95% (= strong support; however, at 88% [MP-EST] or 78% [STAR] it is still moderate); Folk and Freudenstein (2014) also noted that points of discord between coalescence and concatenation estimates tend to be associated with lower nodal support in the coalescent tree. On the other hand, MP-EST and STAR were remarkably consistent with one another.

Chloroplast phylogeny—*Heuchera* is known as a particularly dramatic example of chloroplast capture (Soltis et al., 1991), although in the time that has passed since this system was examined it has become possible to sequence essentially entire chloroplast genomes to explore this signal further. In the current analysis (Fig. 4), the chloroplast genome assemblies are practically complete; the few missing regions tend to correspond to long A-T repeats and other types of microsatellite-like regions (total aligned length 158,903 bp; 1.8% missing data). This study serves as the first sequence-based confirmation of the discord between nuclear and chloroplast phylogenies (see Soltis et al., 1991; Soltis and Kuzoff, 1995) and uncovers new discord beyond what has previously been observed. The clade of *Heuchera* in this analysis that resolves as sister to *Mitella pentandra* corresponds to a clade hypothesized to have captured the *Tiarella* plastome (Soltis et al., 1991), while the clade consisting of *H. parvifolia*, *H. acutifolia*, and *H. pulchella* corresponds to the clade hypothesized to possess the ancestral plastid genome. However, the position of *H. parishii*, *H. elegans*, and *H. abramsii* (species unsampled in the earlier study) as a clade sister to the rest of the *Heuchera* group of genera represents a new lineage not uncovered previously.

Mitochondrial phylogeny—We acknowledge that, due to the lower coverage of the mitochondrial data in many samples (mostly caused by the lower background prevalence and larger size of mitochondrial genomes as compared to plastomes), our mitochondrial phylogeny should be viewed as experimental. The lack of conservation of basic structural features and even large stretches of sequence further hinders assembly. Nevertheless, it is remarkable that our mitochondrial analysis (total aligned length 421,260 bp; 43.9% missing data) resulted in a resolved topology, largely with strong support values, in a genus for which mitochondrial phylogenetics has never been performed.

The mitochondrial phylogeny (Fig. 5) shows a similar degree of conflict with the nuclear topology as does the chloroplast tree. In fact, there are some parallel points of conflict between the data sets. Particularly, the position of *H. parishii*, *H. elegans*, and *H. abramsii* as sister to all other members of the *Heuchera* group of genera is exactly congruent with the chloroplast analysis, although the topology within this clade differs somewhat. Such mitochondrial skimming results have broad relevance for plant systematists. Mitochondrial markers have been under-sampled in plant phylogenetics as compared to nuclear and chloroplast markers; in plants, mitochondrial loci have seen use mainly at higher levels (e.g., Chaw et al., 2000; Jian et al., 2008), and among conifer species (Gugerli et al., 2001), and among angiosperm species (Gugerli et al., 2001; Chaw et al., 2000; Jian et al., 2008), and among conifer species (Gugerli et al., 2001), and among angiosperm species (Gugerli et al., 2001; Chaw et al., 2000; Jian et al., 2008), and among conifer species (Gugerli et al., 2001), and among angiosperm species (Gugerli et al., 2001; Chaw et al., 2000; Jian et al., 2008), and among conifer species (Gugerli et al., 2001), and among angiosperm species (Gugerli et al., 2001; Chaw et al., 2000; Jian et al., 2008), and among conifer species (Gugerli et al., 2001), and among angiosperm species (Gugerli et al., 2001; Chaw et al., 2000; Jian et al., 2008), and among conifer species (Gugerli et al., 2001), and among angiosperm species (Gugerli et al., 2001; Chaw et al., 2000; Jian et al., 2008), and among conifer species (Gugerli et al., 2001), and among angiosperm species (Gugerli et al., 2001; Chaw et al., 2000; Jian et al., 2008), and among conifer species (Gugerli et al., 2001), and among angiosperm species (Gugerli et al., 2001; Chaw et al., 2000; Jian et al., 2008), and among conifer species (Gugerli et al., 2001), and among angiosperm species (Gugerli et al., 2001; Chaw et al., 2000; Jian et al., 2008), and among conifer species (Gugerli et al., 2001), and among angiosperm species (Gugerli et al., 2001; Chaw et al., 2000; Jian et al., 2008), and among conifer species (Gugerli et al., 2001), and among angiosperm species (Gugerli et al., 2001; Chaw et al., 2000; Jian et al., 2008), and among conifer species (Gugerli et al., 2001), and among angiosperm species (Gugerli et al., 2001; Chaw et al.
independently resolve past hybridization events (see also Govindarajulu et al., 2015), but this remains less common than the use of chloroplast data. Mitochondrial capture is well-known in animals (Good et al., 2008, and citations therein); hence in plant groups, particularly those that are believed to have undergone chloroplast capture, mitochondrial phylogenies promise to further and independently resolve past hybridization events (see also Govindarajulu et al., 2015). The current paucity of good reference mitochondrial genomes among angiosperms may be among the major setbacks preventing resolved mitochondrial phylogenetic hypotheses in many plant groups. Given their unusual patterns of evolution, it will be interesting to explore whether

Fig. 5. Map of regions of interest in the mitochondrial genome of *Heuchera parviflora* var. *saurensis*, prepared using OrganellarGenomeDRAW (Lohse et al., 2013). The outer circle depicts protein-coding genes, rRNAs, and tRNAs; the inner circle depicts the large repeat regions. Genes annotated on the inner face of the circle indicate genes transcribed in a clockwise orientation, and genes on the outer face are transcribed counterclockwise. The orientation of the repeat regions is arbitrary; the green and blue regions are direct repeats, while the yellow and lavender regions exist both as direct repeats (where they overlap the green direct repeat) and inverted repeats (elsewhere). Annotated gene regions include both introns and exons. Plant mitochondrial genomes have several trans-spliced genes (e.g., *nad5*); only the cistronic portions are annotated to avoid overlap.
mitochondrial phylogenies show capture events as frequently as has been observed for chloroplast markers.

Conclusions—Even given large phylogenomic data sets, sequencing methods and marker choice must be no less carefully optimized for particular research questions. Recent radiations straddle the boundary of phylogenetic and population-level processes, lending themselves to targeted enrichment of intron markers. For these systems, where coalescent branch lengths are short and barriers to gene flow often poorly developed, gene tree discord is likely to be prevalent. Examining conflicting histories necessitates sequences long enough to infer resolved gene trees. The ability to assemble large amounts of off-target organellar data with such methods further results in a robust sampling of alternative genomic histories. Likewise, sequence divergence will usually be low in recent radiations, with the result that targeting more conserved portions of the genome may waste sequencing effort. Compounding these difficulties for any inferential method is the effect of missing data; we have shown that this method results in essentially no missing data for targeted nuclear markers and chloroplast data (in contrast to a typical RNA-seq or RAD-seq experiment; see also Weitemier et al., 2014).

In the increasingly data-rich field of systematics, phylogenetic conflict is becoming the rule rather than the exception, with the result that reporting a single total-evidence tree without examining the underlying data is becoming less satisfactory. As systematists grapple with gene tree heterogeneity in the most difficult systems, targeted methods optimized for gene tree resolution are poised to address all of these concerns. Intracolling low-copy sequence markers, long-favored for Sanger sequencing (Folk and Freudenstein, 2014, and citations therein), will be increasingly important in genera and species complexes for a robust sampling not only for a total-evidence phylogeny, but more critically for individual gene histories.

LITERATURE CITED


APPENDIX 1. Voucher information for accesses used in this study.

<table>
<thead>
<tr>
<th>Species</th>
<th>Herbarium voucher no. (Herbarium)</th>
<th>Collection locality</th>
<th>Geographic coordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Heuchera abramsii</em> Rydb.</td>
<td>Folk 45 (OS)</td>
<td>Mount Baldy, CA, USA</td>
<td>34°17.423′N, 117°38.738′W</td>
</tr>
<tr>
<td><em>H. acutifolia</em> Rose (herbaceous specimen)</td>
<td>Matrix 14560 (TEX)</td>
<td>Rancho Poza, VER, Mexico</td>
<td>20°32′10.38″N, 98°28′51.58″W</td>
</tr>
<tr>
<td><em>H. americana</em> L. var. <em>americana</em></td>
<td>Folk 70 (OS)</td>
<td>Berry Road, Wayne National Forest, OH, USA</td>
<td>38°36′9.11″N, 82°20′22.16″W</td>
</tr>
<tr>
<td><em>H. elegans</em> Abrams</td>
<td>Folk 44 (OS)</td>
<td>Talus above road to Crystal Lake, 0.5 mi. from Hwy. 39, CA, USA</td>
<td>43°18.940′N, 117°50.537′W</td>
</tr>
<tr>
<td><em>H. grossulariifolia</em> Rydb. var. <em>grossulariifolia</em> (tetraploid)</td>
<td>Folk 160 (OS)</td>
<td>Cliff above roadside parking lot outside of Riggins, ID, USA</td>
<td>45°24′00″N, 116°19′35″W</td>
</tr>
<tr>
<td><em>H. missouriensis</em> Rosend.</td>
<td>Folk I-69 (OS)</td>
<td>Little Grand Canyon near Carbondale, IL, USA</td>
<td>37°41′4.46″N, 89°24′38.92″W</td>
</tr>
<tr>
<td><em>H. parvifolia</em> Rydb.</td>
<td>Folk 43 (OS)</td>
<td>Sugarloaf Mountain, CA, USA</td>
<td>34°12.003′N, 116°47.62′W</td>
</tr>
<tr>
<td><em>H. parviflora</em> Bartl. var. <em>parviflora</em></td>
<td>Folk 72 (OS)</td>
<td>Natural Bridge State Park, KY, USA</td>
<td>37°46′26.33′N, 83°40′51.39′W</td>
</tr>
<tr>
<td><em>H. parviflora</em> var. <em>nivalis</em> (Rosend., Butters &amp; Lakela) Á. Löve, D. Löve &amp; B. M. Kapoor</td>
<td>Folk 55 (OS)</td>
<td>Along Rock Creek Hills Road, Lost Park, CO, USA</td>
<td>39°20′59″N, 105°41′25″W</td>
</tr>
<tr>
<td><em>H. puberula</em> Mack. &amp; Bush</td>
<td>Folk 69 (OS)</td>
<td>Pea Vine Road near Big Springs, Ozark National Riversway, MO, USA</td>
<td>36.94794′N, 90.99215′W</td>
</tr>
<tr>
<td><em>H. pulchella</em> Wooton &amp; Standl.</td>
<td>Folk 20 (OS)</td>
<td>Sandia Crest, NM, USA</td>
<td>35°12.662′N, 106°27.029′W</td>
</tr>
<tr>
<td><em>H. rubescens</em> Torr. var. <em>versicolor</em> (Greene) M. G. Stewart</td>
<td>Folk 26 (OS)</td>
<td>Outcrop above Hwy. 159, NM, USA</td>
<td>33°22′14″N, 108°41′34″W</td>
</tr>
<tr>
<td><em>H. villosa</em> Michx. var. <em>villosa</em></td>
<td>Folk I-1 (OS)</td>
<td>Devil’s Courthouse, NC, USA</td>
<td>35°18′9.44″N, 82°53′44.39″W</td>
</tr>
<tr>
<td><em>H. wootoni</em> Rydb.</td>
<td>Folk 22 (OS)</td>
<td>Dry bank, Capitan Peak, NM, USA</td>
<td>33°36′45″N, 105°14′45″W</td>
</tr>
<tr>
<td><em>Mitella pentandra</em> Hook.</td>
<td>Folk 128 (OS)</td>
<td>Pacific Crest Trail access from Hwy. 90, WA, USA</td>
<td>47°25′40″N, 121°24′48″W</td>
</tr>
</tbody>
</table>

*OS = Ohio State University Herbarium; TEX = University of Texas at Austin Herbarium.

http://www.bioone.org/loi/apps