Development of Single-Copy Nuclear Intron Markers for Species-Level Phylogenetics: Case Study with Paullinieae (Sapindaceae)

Authors: Joyce G. Chery, Chodon Sass, and Chelsea D. Specht
Source: Applications in Plant Sciences, 5(9)
Published By: Botanical Society of America
URL: https://doi.org/10.3732/apps.1700051
Protocol Note

Development of single-copy nuclear intron markers for species-level phylogenetics: Case study with Paullinieae (Sapindaceae)

Joyce G. Chery, Chodon Sass, and Chelsea D. Specht

Department of Integrative Biology, University of California, Berkeley, California 94720 USA; University and Jepson Herbaria, University of California, Berkeley, 1001 Valley Life Sciences Building #3140, Berkeley, California 94720 USA; Department of Plant and Microbial Biology, University of California, Berkeley, 111 Koshland Hall, Berkeley, California 94720 USA; and School of Integrative Plant Sciences, Section of Plant Biology, Cornell University, 412 Mann Library Building, Ithaca, New York 14853 USA

Premise of the study: We developed a bioinformatic pipeline that leverages a publicly available genome and published transcriptomes to design primers in conserved coding sequences flanking targeted introns of single-copy nuclear loci. Paullinieae (Sapindaceae) is used to demonstrate the pipeline.

Methods and Results: Transcriptome reads phylogenetically closer to the lineage of interest are aligned to the closest genome. Single-nucleotide polymorphisms are called, generating a "pseudoreference" closer to the lineage of interest. Several filters are applied to meet the criteria of single-copy nuclear loci with introns of a desired size. Primers are designed in conserved coding sequences flanking introns. Using this pipeline, we developed nine single-copy nuclear intron markers for Paullinieae.

Conclusions: This pipeline is highly flexible and can be used for any group with available genomic and transcriptomic resources. This pipeline led to the development of nine variable markers for phylogenetic study without generating sequence data de novo.

Keywords: introns; nuclear marker development; Paullinieae; Sapindaceae.

Rapidly evolving introns of low-copy nuclear markers have the potential to generate robust species-level phylogenetic hypotheses (Sang, 2002). With current high-throughput sequencing technologies, genomic and transcriptomic data sets are becoming available to use for the development of informative markers for phylogenetic utility. In the most fortunate of cases, the available genome or transcriptome is within the targeted lineage of interest or is closely related (Curto et al., 2012; Granados Mendoza et al., 2015; Stockenhuber et al., 2015). In other cases, authors generate transcriptome data of members of the targeted group (Tonnabel et al., 2014; Stockenhuber et al., 2015). Combining these data, authors aim to target low-copy nuclear markers for phylogenetic utility. Although this strategy is very promising, whole genomes are not typically available for nonmodel systems, and generating transcriptome data de novo is expensive and may be unnecessary given existing data (e.g., 1000 Plants [1KP] project; www.onekp.com).

Here, we present a bioinformatic pipeline that leverages a publicly available genome and published transcriptome reads to identify conserved regions in single-copy nuclear loci and to design primers for amplification of associated introns. Benefits of this pipeline include (1) reduced cost by not generating sequence data de novo and (2) targeting nuclear introns, which are expected to have high sequence variation even among closely related species. This pipeline can be powerful in cases where published transcriptomes are phylogenetically closer to the targeted lineage than the available genome, and in cases where researchers are interested in single-copy nuclear introns for phylogenetic resolution. It is useful for researchers interested in using small-scale sequencing efforts (i.e., Sanger sequencing) to identify relatively few (1–20) informative nuclear loci that can be amplified by PCR, but could be scaled up to include larger sets of loci by relaxing parameters and/or reducing the number of filtering steps in the pipeline. The final set of loci can be used to design baits for Hyb-Seq next-generation sequencing (Weitemier et al., 2014), homemade in-solution capture (Peñalba et al., 2014), or microfluidic PCR primers (Uribe-Convers et al., 2016).

We demonstrate the utility of this bioinformatic pipeline to design primers to amplify single-copy nuclear introns in the tribe Paullinieae (Sapindaceae), a Neotropical lineage of ~475 liana species (Acevedo-Rodriguez et al., 2017). A previous phylogenetic analysis of the Paullinieae tribe was strictly morphological (Acevedo-Rodriguez, 1993) and only at the generic level. Most recently, Acevedo-Rodriguez et al. (2017) aimed to resolve generic- and species-level relationships in Paullinieae...
using ITS and the trnl intron. Although important tribal relationships were resolved, these two markers resulted in a polytomy of *Serjania* Mill., *Paullinia* L., *Urvillea* Kunth, and *Cardiospermum* L. Thus, more phylogenetically informative molecular markers are needed to improve the resolution of generic- and species-level relationships in Paullinieae. Here, we describe a bioinformatic pipeline that leverages publicly available genomic data from distantly related lineages (within the order and within the family but distant to the tribe of interest) to successfully design primers for a specific tribe. We demonstrate the pipeline using the annotated *Citrus sinensis* (L.) Osbeck (Rutaceae: Sapindales) genome with two sets of transcriptome reads, *Dimocarpus longan* Lour. and *Litchi chinensis* Sonn., that are within Sapindaceae but outside of the Paullinieae tribe of interest. The estimated pairwise divergence time between the transcriptomes and genome is 94 mya (www.timetree.org). By using two sets of transcriptome reads from species in Sapindaceae, we were able to find conserved regions from which primers could be designed for amplification of single-copy nuclear introns.

**METHODS AND RESULTS**

**Finding single-copy nuclear markers**—To generate nuclear intron markers, the annotated genome of *C. sinensis* (Rutaceae) provided the coding sequences, intron positions, and estimated intron size while the transcriptomes of two more closely related Sapindaceae species, *D. longan* and *L. chinensis*, provided the best estimate of the gene sequence from which primers could be designed for the target lineage. By using this combination of available data, we hoped to identify single-copy gene regions to avoid amplification of unidentified paralogs, and to design primers that would have a high likelihood of amplification success across the Paullinieae (Sapindaceae).

First, the genome is processed: genome coding sequences (CDS) were downloaded and filtered for single isoform mRNA strands for ease of processing, and genes with introns of 500–1100 bp were selected so they would be easily amplified by traditional PCR and still contain a sufficient number of characters to have phylogenetic utility. The number of base pairs could be changed to include greater or fewer regions as desired. Second, the transcriptome reads are cleaned to remove adapters, low-complexity sequences, contamination, and PCR duplicates (Singhal, 2013). Third, a series of steps are applied to use the genome data to obtain homologous sequences from the more closely related transcriptome without generating a transcriptome assembly de novo as in Sass et al. (2016): (1) Cleaned transcriptome reads are aligned to the filtered genome coding sequences using NovoAlign version 3.01 (Novocraft Technologies, Petaling Jaya, Selangor, Malaysia; http://novocraft.com) with -t 480, a lenient value that allows highly divergent sequences to map. (2) Single-nucleotide polymorphisms (SNPs) are called using SAMtools version 0.1.19 (Li et al., 2009), and new consensus sequences are generated based on the SNPs called. (3) Transcriptome reads are aligned to the new consensus sequences created from the first alignment using NovoAlign -t 90 (a more stringent value). (4) SNPs are called and the final consensus sequences are created to serve as a pseudoreference for primer design. The iterative alignment and SNP calling enables more distantly related transcriptome reads to align to the genome. The following filters are applied to all pseudoreference sequences (Fig. 1) in this order: (1) Retain sequences that only BLAST to self with default settings (i.e., exclusively BLAST to the *Citrus* CDS from which the pseudoreference was generated). (2) Remove sequences that BLAST to plastid, chloroplast, ribosomal, transposon, or mitochondrial loci using MegaBLAST and the National Center for Biotechnology Information (NCBI) database (Organism: Spermatophyta; http://www.ncbi.nlm.nih.gov/). (3) Retain only genes with at least 20x average gene coverage. DNA extraction and primer testing—Several taxa were chosen that represented the phylogenetic breadth of the target lineage. These taxa were used to test the PCR primers with the aim that the primers would work in all samples in the tribe. Initial PCRs used a temperature gradient for the annealing step with a single taxon to determine the optimal annealing temperature for each primer pair. Once the annealing temperature and number of cycles were determined, the optimal PCR conditions were applied to all samples. Only primer pairs that yielded a single band across all samples were sequenced. The sequences were assembled, edited, and aligned and pairwise identity and parsimony informative scores were generated for each marker to determine phylogenetic utility.

**Primer design**—To verify that primer regions were conserved within the breadth of phylogenetic interest, a second transcriptome within the family of interest was aligned to the pseudoreference. This step increases the chances that the primers will be conserved and utilizable across the breadth of the lineage of interest. This step could be eliminated if a second transcriptome is not available; however, further primer testing would likely be necessary. *Litchi chinensis* (Sapindaceae) transcriptome reads were aligned to the *Dimocarpus* pseudoreferences using the Map to Reference tool in Geneious version 8.0.4 (Biomatters Ltd., Auckland, New Zealand; Fig. 1; Appendix 1). Primers were then designed at a conserved coding sequence flanking intron positions using Primer3 version 0.4.0 (Koressaar and Remm, 2007; Untergerstaller et al., 2012).

**DNA extraction and primer testing**—Several taxa were chosen that represented the phylogenetic breadth of the target lineage. These taxa were used to test the PCR primers with the aim that the primers would work in all samples in the tribe. Initial PCRs used a temperature gradient for the annealing step with a single taxon to determine the optimal annealing temperature for each primer pair. Once the annealing temperature and number of cycles were determined, the optimal PCR conditions were applied to all samples. Only primer pairs that yielded a single band across all samples were sequenced. The sequences were assembled, edited, and aligned and pairwise identity and parsimony informative scores were generated for each marker to determine phylogenetic utility.
Case study: Targeting single-copy nuclear introns of Paulliniae (Sapindaceae)—The *C. sinensis* CDS (46,147 sequences; *C. sinensis* genome version 1.0) and genome annotation files were downloaded from the *Citrus Genomes Database* (http://citrusgenomedb.org [accessed 22 Aug 2015]; Wu et al., 2014). Sequences in this CDS file are mature mRNA strands void of introns and untranslated regions. Of these 46,147 mature mRNAs, 18,384 were single isoform mRNAs and, of those, 2159 had introns of 500–1100 bp. The *D. longan* transcriptome pair-end reads (SRR412534) were downloaded from the NCBI database using the NCBI SRA Toolkit version 2.4.2-5 and cleaned to remove adapters, low-complexity sequences, contamination, and PCR duplicates (Singhal, 2013). Of the 64,876,258 *D. longan* transcriptome reads, 39,701,810 remained after cleaning. Of the cleaned reads, 753,149 aligned to the 2159 genes from *C. sinensis* in the final alignment. In transforming the *C. sinensis* reference to a *Dimocarpus* pseudoreference, 103,088 SNP positions were changed. After removing low-coverage genes, 1547 pseudoreference sequences remained. The following filters were applied (Fig. 1) in this order: (1) retain sequences that only BLAST to self (by BLAST to both the entire Citrus CDS and against all pseudoreferences) (315 removed); (2) remove sequences that BLAST to plastid, chloroplast, ribosomal, transposon, or mitochondrial loci using MegaBLAST and the NCBI database (Organism: Spermatophyta) and compiled Sapindales plastomes (downloaded from https://www.ncbi.nlm.nih.gov/genbank) (150 removed); (3) retain only genes with at least 20x average read coverage (793 removed); (4) remove sequences with hits to RepeatMasker (http://repeatmasker.org) using default settings (i.e., interspersed repeats and low-complexity DNA sequences) (243 removed). After applying the above filters, a total of 46 sequences were isolated that fit the criteria of single-copy, single isoform nuclear genes containing introns between 500–1100 bp. To verify that primer design regions were conserved within the family, a second set of Sapindaceae transcriptome reads, from *S. pyramidata* DC., *S. atrolineata* Radlk., *P. glomerulosa* L., *P. gloriosa* L., *P. tetraptera* L., *K. biloba* (L.) Willd., *P. mexicana* (L.) Willd., *P. tigrinae* Kunth (3 individuals), *P. bracteata* (Nakai) C. Jeffrey, *P. bullata* (C. Jeffrey) C. Jeffrey, *P. dubia* C. Jeffrey, and one outgroup species (*Allophylus psilospermus* Radlk.) were used. After testing annealing temperatures over a temperature gradient (62–43 °C) to all samples. Loci were amplified using Phire Hot Start II DNA Polymerase primer pair in a representative species, optimized PCR conditions were applied to each sequenced locus, the consensus sequence of the Sanger reads and their

<table>
<thead>
<tr>
<th>Locus</th>
<th>Primer sequences (5′-3′)</th>
<th>Expected bases (bp)</th>
<th>Positive citrus DNA homology</th>
<th>Aligned bases (bp)</th>
<th>Tm (°C)</th>
<th>Tm a = annealing temperature.</th>
<th>Ref. locus</th>
<th>Reference primer identification</th>
<th>Reference primer database</th>
</tr>
</thead>
<tbody>
<tr>
<td>orange 1</td>
<td>g1009830m (intron9)</td>
<td>537</td>
<td>558</td>
<td>Ap-4 complex subunit epsilon-1</td>
<td>44.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>orange 2</td>
<td>g004520m (intron7)</td>
<td>575</td>
<td>665</td>
<td>NAD(P)-binding Rossmann-fold superfamily protein</td>
<td>44.3</td>
<td>885</td>
<td>58.6</td>
<td>Acidocitrate lyase/4-hydroxybutyrate dehydrogenase</td>
<td>NCBI SRA Toolkit v.2.4.2-5</td>
</tr>
<tr>
<td>orange 3</td>
<td>g104492m (intron5)</td>
<td>482</td>
<td>522</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>52.5</td>
<td>533</td>
<td>58.6</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>NCBI SRA Toolkit v.2.4.2-5</td>
</tr>
<tr>
<td>orange 4</td>
<td>g104502m (intron7)</td>
<td>641</td>
<td>651</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>48.2</td>
<td>482</td>
<td>58.6</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>NCBI SRA Toolkit v.2.4.2-5</td>
</tr>
<tr>
<td>orange 5</td>
<td>g104502m (intron7)</td>
<td>641</td>
<td>651</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>48.2</td>
<td>482</td>
<td>58.6</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>NCBI SRA Toolkit v.2.4.2-5</td>
</tr>
<tr>
<td>orange 6</td>
<td>g104502m (intron7)</td>
<td>641</td>
<td>651</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>48.2</td>
<td>482</td>
<td>58.6</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>NCBI SRA Toolkit v.2.4.2-5</td>
</tr>
<tr>
<td>orange 7</td>
<td>g104502m (intron7)</td>
<td>641</td>
<td>651</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>48.2</td>
<td>482</td>
<td>58.6</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>NCBI SRA Toolkit v.2.4.2-5</td>
</tr>
<tr>
<td>orange 8</td>
<td>g104502m (intron7)</td>
<td>641</td>
<td>651</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>48.2</td>
<td>482</td>
<td>58.6</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>NCBI SRA Toolkit v.2.4.2-5</td>
</tr>
<tr>
<td>orange 9</td>
<td>g104502m (intron7)</td>
<td>641</td>
<td>651</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>48.2</td>
<td>482</td>
<td>58.6</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>NCBI SRA Toolkit v.2.4.2-5</td>
</tr>
<tr>
<td>orange 10</td>
<td>g104502m (intron7)</td>
<td>641</td>
<td>651</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>48.2</td>
<td>482</td>
<td>58.6</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>NCBI SRA Toolkit v.2.4.2-5</td>
</tr>
<tr>
<td>orange 11</td>
<td>g104502m (intron7)</td>
<td>641</td>
<td>651</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>48.2</td>
<td>482</td>
<td>58.6</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>NCBI SRA Toolkit v.2.4.2-5</td>
</tr>
<tr>
<td>orange 12</td>
<td>g104502m (intron7)</td>
<td>641</td>
<td>651</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>48.2</td>
<td>482</td>
<td>58.6</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>NCBI SRA Toolkit v.2.4.2-5</td>
</tr>
<tr>
<td>orange 13</td>
<td>g104502m (intron7)</td>
<td>641</td>
<td>651</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>48.2</td>
<td>482</td>
<td>58.6</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>NCBI SRA Toolkit v.2.4.2-5</td>
</tr>
<tr>
<td>orange 14</td>
<td>g104502m (intron7)</td>
<td>641</td>
<td>651</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>48.2</td>
<td>482</td>
<td>58.6</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>NCBI SRA Toolkit v.2.4.2-5</td>
</tr>
<tr>
<td>orange 15</td>
<td>g104502m (intron7)</td>
<td>641</td>
<td>651</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>48.2</td>
<td>482</td>
<td>58.6</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>NCBI SRA Toolkit v.2.4.2-5</td>
</tr>
<tr>
<td>orange 16</td>
<td>g104502m (intron7)</td>
<td>641</td>
<td>651</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>48.2</td>
<td>482</td>
<td>58.6</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>NCBI SRA Toolkit v.2.4.2-5</td>
</tr>
<tr>
<td>orange 17</td>
<td>g104502m (intron7)</td>
<td>641</td>
<td>651</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>48.2</td>
<td>482</td>
<td>58.6</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>NCBI SRA Toolkit v.2.4.2-5</td>
</tr>
<tr>
<td>orange 18</td>
<td>g104502m (intron7)</td>
<td>641</td>
<td>651</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>48.2</td>
<td>482</td>
<td>58.6</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>NCBI SRA Toolkit v.2.4.2-5</td>
</tr>
<tr>
<td>orange 19</td>
<td>g104502m (intron7)</td>
<td>641</td>
<td>651</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>48.2</td>
<td>482</td>
<td>58.6</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>NCBI SRA Toolkit v.2.4.2-5</td>
</tr>
<tr>
<td>orange 20</td>
<td>g104502m (intron7)</td>
<td>641</td>
<td>651</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>48.2</td>
<td>482</td>
<td>58.6</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>NCBI SRA Toolkit v.2.4.2-5</td>
</tr>
<tr>
<td>orange 21</td>
<td>g104502m (intron7)</td>
<td>641</td>
<td>651</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>48.2</td>
<td>482</td>
<td>58.6</td>
<td>Nuclear pore complex protein Nap2/3</td>
<td>NCBI SRA Toolkit v.2.4.2-5</td>
</tr>
</tbody>
</table>

Note: Tm = annealing temperature. **Ref. locus in* Citrus sinensis* v1.0 genome.

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

Terms of Use: https://bioone.org/terms-of-use
Primers were mapped back to their respective pseudoreference and the original *Citrus* gene. For loci with more internally designed primers (i.e., almost entirely intronic), the primers were mapped to both references. In all cases, either the CDS region of the Sanger sequence read itself or the primers aligned to their respective references.

**Phylogenetic analysis**—The concatenated alignment of nine markers with a total of 4892 aligned characters was used to generate phylogenetic hypotheses under maximum likelihood. Gaps and the ends of shorter sequences were treated as missing data. Maximum likelihood trees were generated with the general time-reversible (GTR) sequence evolution model in RAxML-HPC on XSEDE 8.2.8 (Stamatakis, 2014) using the CIPRES Scientific Gateway (Miller et al., 2010) (Fig. 2). Support was evaluated with 100 bootstrap replicates. Additionally, gene trees were generated using the GTRGAMMA model in RAxML with 1000 bootstraps on XSEDE 8.2.8 (Stamatakis, 2014). These gene trees were used as input into statistical binning (Mirarab et al., 2014), after which final gene trees were run under a partitioned RAxML run and were used as input in Astral-II (Appendix S1). The two liana genera (*Paullinia* and *Serjania*) each form monophyletic groups with moderate to high bootstrap support. Multiple individuals of *P. turbacensis* and *P. bracteosa* were included in the tree and formed monophyletic groups. The long branch of the *Paullinia*-*Serjania* group is explained by the relatively distantly related outgroup *A. psilospermus*. The Astral-II optimal tree differed from the RAxML tree only in the order of the three *P. turbacensis* specimens (Fig. 2), providing evidence that these markers are of appropriate length to be informative. Given the sequence variation, moderate to high bootstrap support across nodes, and recovery of monophyly of major groups, we expect these markers to be highly informative with more inclusive sampling.

### CONCLUSIONS

This bioinformatic pipeline utilizes publicly available genomic and transcriptomic resources to design primers in coding sequences...
flanking targeted introns of single-copy nuclear loci without generating sequence data de novo. An annotated genome provides the information to determine approximate intron size and location, and two more closely related transcriptomes provide the best estimate of gene sequence to optimize primer design. By using this combination of data and through primer validation for exclusively single-band PCR products, we increase the chances of targeting and successfully amplifying orthologous single-copy nuclear introns in the lineage of interest. Although many steps were taken to obtain orthologous loci (i.e., only included genes that BLAST exclusively to self, only proceeded with markers that yielded single PCR products, only proceeded with markers that sequenced a single PCR product, and finally mapped Sanger sequences and primers back to their respective pseudoreference and original Citrus gene), orthology could also be further assessed by constructing and analyzing gene trees with greater taxon sampling.

It is important to note that the number of potential target loci can easily be increased by customizing and relaxing parameters or filters throughout the pipeline (e.g., filters for single isoform genes, average gene coverage, desired intron length). By using the closest sequenced genome available, we had the best estimate of intron size and location. This method also worked in Zingiberales (Sass et al., 2016) for designing markers from individuals distant from the genome by approximately 100 mya. Together with increasing transcriptome availability from publicly available sequencing projects like the 1KP project, this pipeline might be increasingly available for many different plant groups. However, if reasonably close genomic and/or transcriptomic resources are unavailable, MarkerMiner (Chamala et al., 2015) is an easy-to-use alternative tool to develop primers spanning introns without the need for a closely related genome. Key differences between MarkerMiner and the bioinformatic pipeline presented here include our use of transcriptome reads rather than a full transcriptome assembly, explicitly targeting introns of a desired size, and the intentional use of the closest available genome.

Several authors present methods to target low-copy nuclear markers for phylogenetic utility (Duarte et al., 2010; Curto et al., 2012; Tonnabel et al., 2014; Stockenhuber et al., 2015; Sass et al., 2016). Using published genomes from across angiosperms, Duarte et al. (2010) recovered 959 genes that were determined to be single copy across Arabidopsis thaliana (L.) Heynh., Populus trichocarpa Torr. & A. Gray, Oryza sativa L., and Vitis vinifera L. (APOV). Interestingly, of these 959 APOV genes, only 201 were found in the C. sinensis genome and, of those, only 24 were determined to be single copy under our criteria. When specifically looking for those 24 remaining single-copy genes following our developed pipeline, we determined that 23 of the 24 APOV genes were removed through various steps specific to our filtering pipeline (i.e., intron size-specific filters, 20x coverage filter, RepeatMasker filter). Considering copy number in A. thaliana as a reference, Curto et al. (2012) successfully developed nuclear markers in Lamiaceae; however, given the low number of single-copy APOVs found in C. sinensis due to lineage-specific gene loss and duplication, the APOV markers were not appropriate for our purposes. By using phylogenetically closer genomic and transcriptomic data, we were able to test all markers for copy number prior to including them in phylogenetic analysis. Stockenhuber et al. (2015) and Tonnabel et al. (2014) efficiently detected low-copy nuclear markers for Brassicaceae and Proteaceae; however, these authors generated transcriptome sequences of members of the lineage of interest. Sass et al. (2016) detected low-copy nuclear markers using publicly available data but targeted exons. The pipeline presented here is cost efficient in that it does not generate sequence data de novo. Rather, it utilizes publicly available genomic and transcriptomic resources spanning the breadth of the plant order Sapindales to design intron markers at conserved coding sequence boundaries. Using the presented pipeline, amplification of nine novel primer pairs was successful in generating phylogenetically informative markers from nine Sapindaceae species, including amplification in the designated outgroup. Sequence variation within these markers ranges from 53.7–94.3% pairwise identity, making them promising for generating a robust data matrix to resolve species-level phylogenetic relationships within Paululliniae, especially when combined with other highly variable markers (e.g., ITS). This flexible marker development pipeline could be applied to any group with appropriate genomic resources. Identified regions of interests can be used in a variety of ways—amplified by PCR and sequenced using Sanger sequencing or as baits for a Hyb-Seq next-generation sequencing approach.

LITERATURE CITED


1. Genome data preparation:

The following commands and scripts select mRNAs that contain introns of a desired size from the genome coding sequence (CDS) file.

a. Download CDS FASTA files and genome annotation files from the Citrus Genome Database (www.citrusgenomedb.org).

b. Select mRNA names from the genome annotation file:

   grep 'mRNA' Csinensis_v1.0_gene.gff3 | awk '{print $9}' | awk 'BEGIN{FS= "\""}{print $4}' | sed 's/\'/\'/g' > mrnNames

c. Extract intron sizes by using the genome annotation file:

   perl get_intron_size.pl Csinensis_v1.0_gene.gff3 > intronsize

   -get_intron_size.pl (available at https://gist.github.com/baihezimu/4163862)

d. Associate mRNA names with their intron sizes:

   paste mrnNames intronsize > mrnaIntronSize

e. Select single isoform genes from the "mrnaIntronSize":

   sort mrnaNames | uniq -c | sort | awk '{print $2}' > singlemRNANames

grep -f singlemRNANames mrnaIntronSize > singlemRNAintrons

f. Count the number of columns (i.e., each represents an intron size):

   awk '{print NF}' singlemRNAintrons | sort -nu | tail -n 1

   g. Search each column for introns between 500 and 1100 bp in length:

   awk '$4>500&&$4<1100' singlemRNAintrons >> singlemRNAintrons_small
   awk '$5>500&&$5<1100' singlemRNAintrons >> singlemRNAintrons_small...for as many columns as necessary

h. Sort through the list of mRNAs with introns of 500–1100 bp to generate a list of sequence names to align transcriptome reads to in the next steps:

   sort singlemRNAintrons_small | uniq | awk '{print $1}' > singleRNAintrons_small

   i. Convert the genome CDS FASTA file from interleaved to sequential to be in the right format to run scripts in the next steps:

      perl -MBio::SeqIO -e 'my $seqIn = Bio::SeqIO->new(-f => "*STDIN", -format => "fasta")'; while ($seq = $seqIn->next_seq) { print "$seq->id","\n$seq->seq","\n" } Csinensis_v1.0_cds.fa > Csinensis_v1.0_cds_seqs.fa

j. Alter mRNA names for ease of processing downstream:

   awk '{print NF}' Csinensis_v1.0_cds_seqs.f > Csinensis_v1.0_cds_orangeNames.fa

   k. Select sequences from the list of single isoform mRNAs with introns between 500 and 1100 bp:

      perl selectSeqs.pl -in Csinensis_v1.0_cds_orangeNames.fa -out SelectedCitrusSeqs.fa -idfile singlemRNAintrons_small

      mv InSelectedCitrusSeqs.fa SelectedCitrusSeqs.fa

      perl selectSeqs.pl -in Csinensis_v1.0_cds_orangeNames.fa -out SelectedCitrusSeqs.fa -idfile singlemRNAintrons_small

   l. Alter mRNA names to run 2-scrubReads.pl in the next step:

      awk '{FS = "\""}{print $1} Csinensis_v1.0_cds_seqs.f > Csinensis_v1.0_cds_orangeNames.fa

   m. Fix read names to run 2-scrubReads.pl in the next step:

      sed '/^@SRR412534/ s/SRR412534.\[0-9\]* //' SRR412534_2.fastq | sed'/^@A807H/ s/length=90/\\/2/g' | sed 's/\'/\'/g' > JCSR_R1.fq

      sed '/^@SRR412534/ s/SRR412534.\[0-9\]* //' SRR412534_1.fastq | sed'/^@A807H/ s/length=90/\\/1/g' | sed 's/\'/\'/g' > JCSR_R2.fq

   n. Clean all reads to remove adapters, low-complexity sequences, contamination, and PCR duplicates using the protocol in Singhal (2013). This script removes PCR duplicates by removing all reads with identical sequences. Download NC_012947.1 Escherichia coli from GenBank as a negative filter. Required files:

      -adapters.fa and library.txt (available at: https://github.com/joycechery/Sapindaceae/tree/master/TranscriptomeProcessing)
      -scrubReads.sh (available at: https://github.com/joycechery/Sapindaceae/blob/master/transcriptomeprocessing/scrubReads.sh)

   o. Align transcriptome reads to genome:

      This script is used to align cleaned transcriptome reads to the selected Citrus sinensis mRNAs (single isoform mRNAs with introns of 500–1100 bp) using NovoAlign version 3.0.h (Novocraft Technologies, Petaling Jaya, Selangor, Malaysia; http://novocraft.com) with *t* = 480 to allow highly divergent sequences to map. PCR duplicates are additionally removed by identical mapping location using MarkDuplicated in Picard (Picard, Broad Institute, Cambridge, Massachusetts, USA; http://broadinstitute.github.io/picard/). Single-nucleotide polymorphisms (SNPs) are called using SAMtools version 0.1.19 (Li et al., 2009), and new consensus sequences are generated based on SNPs called. A second iteration aligned reads to the first alignment consensus sequences using NovoAlign -t 90. The rate of multiple mapping is 0.5% in the first alignment and 0.2% in the second alignment with these data. SNPs are called again to generate a set of final consensus sequences; these serve as a pseudo-reference from which primers are designed.

   http://www.bioone.org/loi/apps
4. Filter pseudoreference sequences:

The following commands and scripts filter the pseudoreferences to (1) retain only sequences that BLAST to self, (2) have at least 20× average gene coverage, and (3) remove sequences with hits to RepeatMasker.

a. Retain only sequences that exclusively BLAST to self:

```bash
-BlastToRemove_Paralogs
```

b. Remove sequences that have BLAST hits to ribosomal, plastid, chloroplast, mitochondrial, or transposons:

```bash
-BlastToRemove_nonnuclear
```

c. Remove genes with less than 20× average gene coverage:

i. Create a list of pseudoreferences names:

```bash
grep '>' LocalBlastRemoveNCBIREmove.fa | sed 's/>//' > LocalBlastRemoveNCBIREmove_Names
```

ii. Create a coverage-per-position file from the .bam file produced by “Alignfix.pl”:

```bash
module load bedtools/2.22.1
bedtools genom cov -ibam JCSR.new2.bam -d -split > JCSR.positioncoverage
```

iii. Create a coverage-per-position file for filtered pseudoreference sequences:

```bash
grep -f LocalBlastRemoveNCBIREmove_Names JCSR.positioncoverage > LocalBlastRemoveNCBIREmove_PositionCoverage
```

iv. Create an average gene coverage file:

```bash
-AverageGeneCoverage.pl
```

v. Select genes with more than 20× average gene coverage:

```bash
awk '$2> 20' CoverageFile_PerGene > CoverageFile_MoreThan20x
awk '{print $1}' CoverageFile_MoreThan20x > 20Xgenes
perl selectSeqs.pl -in LocalBlastRemoveNCBIREmove.fa -out LocalBlastRemoveNCBIREmove20XCov.fa -idfile 20Xgenes
mv LocalBlastRemoveNCBIREmove20XCov.fa LocalBlastRemoveNCBIREmove20XCov.fa
```

d. Run all sequences through RepeatMasker using default settings except search engine (rmblast available at: http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker). Compile a list of gene names with hits “RepeatMaskerGenestoRemove” and remove them:

```bash
perl selectSeqs.pl -in LocalBlastRemoveNCBIREmove20XCov.fa -out LocalBlastRemoveNCBIREmove20XCovRM.fa -idfile RepeatMaskerGenestoRemove
mv NotLocalBlastRemoveNCBIREmove20XCovRM.fa LocalBlastRemoveNCBIREmove20XCovRM.fa
```

5. Primer design:

To verify that primer design regions are conserved across the desired phylogenetic breadth, *Litchi chinensis* transcriptome reads are mapped to the filtered pseudoreferences using the Map to Reference tool in Geneious version 8.0.4 (low sensitivity and two iterations) (Biomatters Ltd., Auckland, New Zealand). This second alignment is optional, and if users choose to do so, it is recommended to clip in the transcriptome reads prior to mapping. Primers are designed using Primer3 version 0.4.0 (http://bioinfo.ut.ee/primer3-0.4.0; Koressaar and Remm, 2007; Untergasser et al., 2012) at conserved CDS–CDS boundaries to target introns. The input sequences consist of at least 30 nucleotides from the first CDS followed by 100 Xs (representing the unknown intron sequence), followed by at least 30 nucleotides of the second CDS. The Xs are excluded by use of the “Excluded Region” option. To design primers at the CDS–CDS boundaries, intron sizes from “singlemRNAintrons_small” were used and CDS sizes within each mRNA are extracted:

```
mv InLocalBlastRemoveNCBIRemove20XCov.fa LocalBlastRemoveNCBIRemove20XCov.fa
```

```bash
perl selectSeqs.pl -in LocalBlastRemoveNCBIREmove20XCov.fa -out LocalBlastRemoveNCBIREmove20XCovRM.fa -idfile RepeatMaskerGenestoRemove
mv NotLocalBlastRemoveNCBIREmove20XCovRM.fa LocalBlastRemoveNCBIREmove20XCovRM.fa
```

6. Primer testing:

a. Extract DNA from silica material using cetyltrimethylammonium bromide (CTAB; Doyle and Doyle, 1987).

b. Conduct PCR temperature gradients (62–43°C) with a representative species to determine the optimal annealing temperature for each primer pair. PCR conditions:

- Phire Hot Start II DNA Polymerase (Thermo Fisher Scientific, Pittsburgh, Pennsylvania, USA; 5-min initial denaturing step at 98°C; loci-specific annealing temperature, and 20 s at 72°C; and a final 1-min 72°C extension.

- Extract optimal PCR conditions to all samples.

- Only sequence PCR products that yield a single band across all samples. Sequencing was done on an Applied Biosystems 3730xl DNA analyzer (Applied Biosystems, Waltham, Massachusetts, USA).

e. Assemble and edit Sanger reads in Geneious version 8.0.4 and align using MAFFT version 7.271.

f. Calculate pairwise identity in Geneious version 8.0.4 and calculate parsimony informative sites in PAUP* 4.0b10 (Swofford, 2002).

7. Phylogenetic analysis:

a. A concatenated alignment of all loci was used to generate phylogenetic hypotheses under a maximum likelihood general time-reversible (GTR) sequence evolution model in RAxML-HPC on XSEDE 8.2.8 (https://www.phylo.org/portal2/home.action; Stamatakis, 2014). Support was evaluated with 100 bootstrap replicates.

b. An ASTRAL-II analysis was run using statistically binned gene trees generated in RAxML using the GTRGAMMA model and 1000 bootstraps.

---

http://www.bioone.org/loi/apps
## APPENDIX 2. Silica-dried specimens used in this study.

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Species name</th>
<th>Collection no.</th>
<th>Locality</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>P0</td>
<td><em>Paullinia</em> sp.</td>
<td>UCBG no. 92.0509</td>
<td>Provenance: Chiapas State, Mexico, North America</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>P1</td>
<td><em>Paullinia bracteosa</em> A</td>
<td>NTBG no. 760259</td>
<td>Provenance: Venezuela</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>P3</td>
<td><em>Serjania mexicana</em></td>
<td>Chery 23</td>
<td>Barro Colorado Island, Panama</td>
<td>09°09.914′N</td>
<td>079°50.213′W</td>
</tr>
<tr>
<td>P5</td>
<td><em>Serjania pseudocentata</em></td>
<td>Chery 34</td>
<td>Barro Colorado Island, Panama</td>
<td>09°09.05′N</td>
<td>079°50.202′W</td>
</tr>
<tr>
<td>P6</td>
<td><em>Serjania pyramidata</em></td>
<td>Chery 29</td>
<td>Barro Colorado Island, Panama</td>
<td>09°10.867′N</td>
<td>079°49.444′W</td>
</tr>
<tr>
<td>P7</td>
<td><em>Paullinia glomerulosa</em></td>
<td>Chery 20</td>
<td>Barro Colorado Island, Panama</td>
<td>09°09.896′N</td>
<td>079°50.294′W</td>
</tr>
<tr>
<td>P8</td>
<td><em>Paullinia turbacensis</em> A</td>
<td>Chery 13</td>
<td>Barro Colorado Island, Panama</td>
<td>09°09.924′N</td>
<td>079°50.188′W</td>
</tr>
<tr>
<td>P9</td>
<td><em>Paullinia turbacensis</em> B</td>
<td>Chery 24</td>
<td>Barro Colorado Island, Panama</td>
<td>09°09.830′N</td>
<td>079°50.270′W</td>
</tr>
<tr>
<td>P10</td>
<td><em>Paullinia turbacensis</em> C</td>
<td>Chery 10</td>
<td>Barro Colorado Island, Panama; Donata Trail</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>P11</td>
<td><em>Serjania atrolineata</em></td>
<td>Chery 42</td>
<td>Barro Colorado Island, Panama</td>
<td>09°10.705′N</td>
<td>079°50.819′W</td>
</tr>
<tr>
<td>P12</td>
<td><em>Paullinia bracteosa</em> B</td>
<td>Chery 26</td>
<td>Barro Colorado Island, Panama; Lake #4 Trail Marking</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>P13</td>
<td><em>Allophylus psilospermus</em></td>
<td>Chery 19</td>
<td>Barro Colorado Island, Panama</td>
<td>09°09.910′N</td>
<td>079°50.287′W</td>
</tr>
</tbody>
</table>

*Sample IDs P3–P13 are deposited in the University of Panama Herbarium (PMA), Panama City, Panama. Sample IDs P3–P13 represent personal collection numbers by Joyce G. Chery.*
### APPENDIX 3. Report of amplification success for each marker for each sample.

<table>
<thead>
<tr>
<th>Locus</th>
<th>P0</th>
<th>P1</th>
<th>P3</th>
<th>P5</th>
<th>P6</th>
<th>P7</th>
<th>P8</th>
<th>P9</th>
<th>P10</th>
<th>P11</th>
<th>P12</th>
<th>P13</th>
</tr>
</thead>
<tbody>
<tr>
<td>orange1.1g002083m (intron9)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>—</td>
<td>—</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>orange1.1g047192m (intron7)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>—</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>orange1.1g045023m (intron7)</td>
<td>+</td>
<td>+</td>
<td>—</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>orange1.1g045023m (intron6)</td>
<td>+</td>
<td>+</td>
<td>—</td>
<td>+</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>+</td>
<td>—</td>
<td>—</td>
<td>+</td>
</tr>
<tr>
<td>orange1.1g028997m (intron4)</td>
<td>+</td>
<td>+</td>
<td>—</td>
<td>+</td>
<td>—</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>orange1.1g015495m (intron8)</td>
<td>+</td>
<td>+</td>
<td>—</td>
<td>+</td>
<td>+</td>
<td>—</td>
<td>—</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>orange1.1g027952m (intron5)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>orange1.1g022777m (intron3)</td>
<td>+</td>
<td>—</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>orange1.1g009973m (intron5)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>—</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

Note: + = successful amplification and sequencing; — = no amplification or unsuccessful sequencing.

*Reference locus in *Citrus sinensis v1.0 genome.*