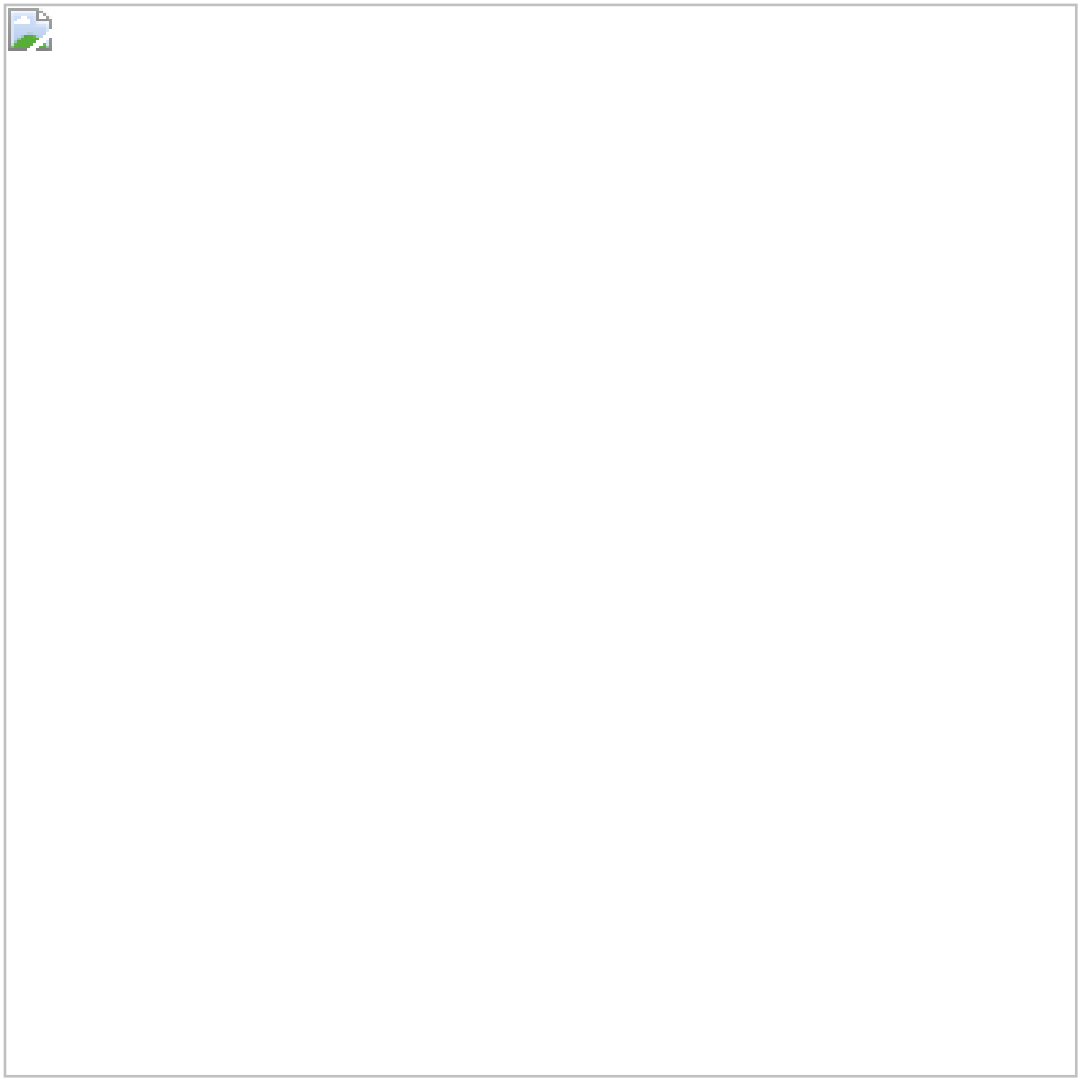

Development of Single-Copy Nuclear Intron Markers for Species-Level Phylogenetics: Case Study with Paullinieae (Sapindaceae)

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

Source: Applications in Plant Sciences, 5(9)

Published By: Botanical Society of America

URL: https://doi.org/10.3732/apps.1700051

Applications in Plant Sciences 2017 5(9): 1700051

Protocol Note

Development of single-copy nuclear intron markers for species-level phylogenetics: Case study with Paullinieae (Sapindaceae)[1]

Joyce G. Chery[2,3,6], Chodon Sass[3,4], and Chelsea D. Specht[5]

[2]Department of Integrative Biology, University of California, Berkeley, 3040 Valley Life Sciences Building #3140, Berkeley, California 94720 USA; [3]University and Jepson Herbaria, University of California, Berkeley, 1001 Valley Life Sciences Building #2465, Berkeley, California 94720 USA; [4]Department of Plant and Microbial Biology, University of California, Berkeley, 111 Koshland Hall, Berkeley, California 94720 USA; and [5]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.

Key words: 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-Rodríguez et al., 2017). A previous phylogenetic analysis of the Paullinieae tribe was strictly morphological (Acevedo-Rodríguez, 1993) and only at the generic level. Most recently, Acevedo-Rodríguez et al. (2017) aimed to resolve generic- and species-level relationships in Paullinieae

[1]Manuscript received 11 May 2017; revision accepted 7 August 2017.

This research was funded by the National Science Foundation (DEB 1208666 to C.D.S.) and by a National Science Foundation Graduate Student Research Grant to J.G.C.

[6]Author for correspondence: chery.joyce@berkeley.edu

doi:10.3732/apps.1700051

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 20× average read coverage (Nielsen et al., 2011). (4) Remove sequences with hits to RepeatMasker (http://repeatmasker.org/) (i.e., interspersed repeats and low-complexity DNA sequences). After the above filters are applied, these sequences fit the criteria of single-copy nuclear genes containing introns between 500–1100 bp.

***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; Untergasser et al., 2012).

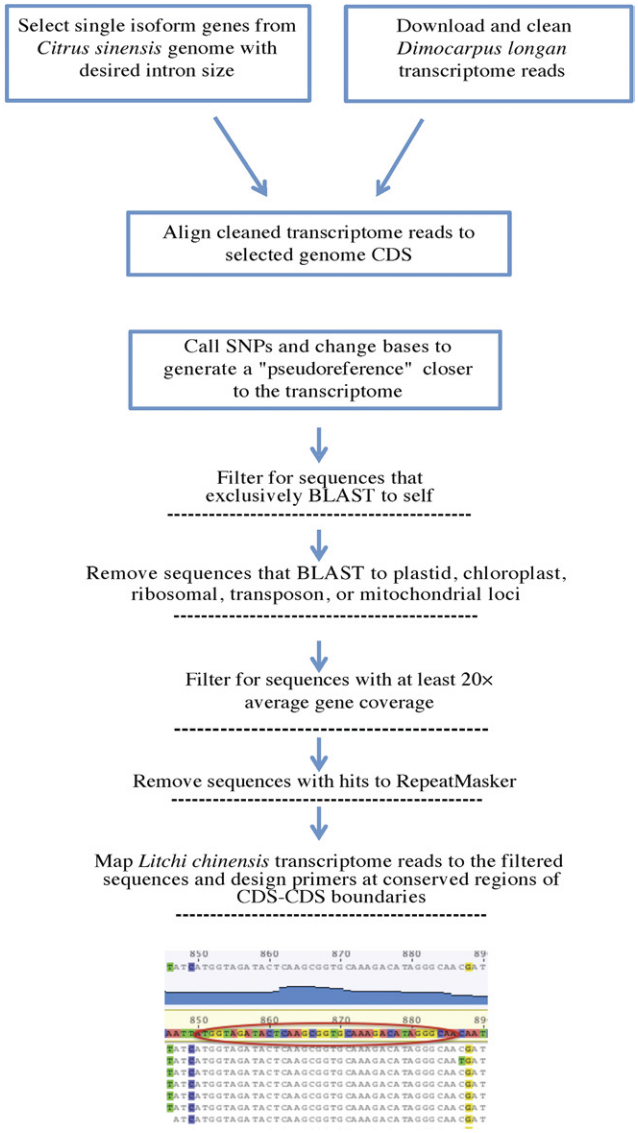

Fig. 1. Bioinformatic pipeline to target single-copy nuclear intron markers. CDS = coding sequences; NCBI = National Center for Biotechnology Information; SNP = single-nucleotide polymorphism.

***DNA extraction and primer testing*—**Several taxa were chosen that represent 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 Paullinieae (Sapindaceae)—The '*Citrus sinensis* CDS' (46,147 sequences; *C. sinensis* genome version 1.0) and genome annotation files were downloaded from the Citrus Genome Database (www.citrusgenomedb.org [accessed 22 August 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 paired-end reads (SRR412534) were downloaded from the NCBI database using the NCBI SRA Toolkit version 2.4.5-2 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, 573,149 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 20× 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 *Litchi chinensis* (NCBI: SRX258094; Li et al., 2013), was aligned to the pseudoreference using the Map to Reference tool in Geneious version 8.0.4 at low sensitivity and for two iterations (Fig. 1). Geneious version 8.04 was used at this step because the number of remaining loci made computation on a desktop computer possible and visualization manageable. Of the 53,437,444 *L. chinensis* reads, 88,779 mapped to the 46 genes of interest. Primers were designed using Primer3 version 0.4.0 (Koressaar and Remm, 2007; Untergasser et al., 2012) by randomly selecting 21 conserved Sapindaceae CDS–CDS boundaries within the Geneious mapped alignment. The *L. chinensis* reads were not filtered for our purposes, because if primers designed in conserved regions from this second alignment resulted in amplification of multiple PCR products, marker development of these loci was not pursued. However, according to best practices all transcriptome reads should be cleaned prior to mapping. See Appendix 1 for step-by-step instructions.

Taxon sampling, DNA extraction, and primer testing—DNA was extracted from silica-dried leaf material from 12 Sapindaceae samples representing nine species (Appendix 2) using the cetyltrimethylammonium bromide (CTAB) method (Doyle and Doyle, 1987) with minor modifications. Eleven ingroup samples (*Serjania paucidentata* DC., *S. pyramidata* Radlk., *S. atrolineata* C. Wright, *S. mexicana* (L.) Willd., *Paullinia turbacensis* Kunth [3 individuals], *P. bracteosa* Radlk. [2 individuals], *Paullinia* sp., and *P. glomerulosa* Radlk.) and one outgroup species (*Allophylus psilospermus* Radlk.) were used. After testing annealing temperatures over a temperature gradient (62–43°C) for each primer pair in a representative species, optimized PCR conditions were applied to all samples. Loci were amplified using Phire Hot Start II DNA Polymerase (Thermo Fisher Scientific, Pittsburgh, Pennsylvania, USA) with a 5-min initial denaturing step at 98°C; loci-specific cycles of 5 s at 98°C, 5 s at loci-specific annealing temperature, and 20 s at 72°C; and a final 1-min 72°C extension. Optimal annealing temperatures are reported in Table 1. Only primer pairs that yielded a single band across all samples were sequenced. Cycle sequencing was performed using BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, Waltham, Massachusetts, USA). Sanger sequencing was done at the UC Berkeley Evolutionary Genetics Laboratory on an Applied Biosystems 3730xl DNA analyzer. Reads were assembled, edited in Geneious version 8.0.4, and aligned using MAFFT version 7.271 (Katoh and Standley, 2013). The Paullinieae introns were consistently smaller than expected based on the *C. sinensis* intron sizes. Pairwise identity was calculated in Geneious version 8.0.4, and parsimony informative sites were calculated in PAUP* 4.0b10 (Swofford, 2002) (Table 2). Of the 21 primer sets tried, seven resulted in multiple products, two failed to amplify, and 12 resulted in single PCR products. An additional three markers were removed due to inconsistent amplification success across all samples. Amplification success of primer pairs is presented in Appendix 3, and summary statistics of each locus are presented in Table 2. To check the identity of each sequenced locus, the consensus sequence of the Sanger reads and their

TABLE 1. Primer sequences for the nine putative single-copy nuclear markers developed to amplify across Paullinieae (Sapindaceae).

Locus[a]	Primer sequences (5′–3′)	Expected bases (bp)	Aligned bases (bp)	T_a (°C)	Putative citrus protein homolog	GenBank accession no.
orange1.1g002083m (intron9)	F: CATATGCAGTTACAGCACTAATGA	537	358	44.3	Ap-4 complex subunit epsilon-1	KY770939–KY770948
	R: AATCTCAACAGCATGAGCATC					
orange1.1g047192m (intron7)	F: AGGTGCTTCACCTGAAATGG	755	665	44.3	NAD(P)-binding Rossmann-fold superfamily protein	KY770927–KY770938
	R: TTGGTTCACCATTTTCACC					
orange1.1g045023m (intron7)	F: AGGGCCCTTGAACTTGTTT	452	814	58.6	Sf27-adaptin, alpha/gamma/epsilon	KY770919–KY770926
	R: CAGAGAGAACCTTGAGCATCTG					
orange1.1g045023m (intron6)	F: GGGCCCTTTTACGAATAGAA	651	301	50.9	Sf27-adaptin, alpha/gamma/epsilon	KY770913–KY770918
	R: AGGGCCCTTGAACTTGTTT					
orange1.1g028997m (intron4)	F: AAAGAGTCCAAACCAACAATTC	885	852	58.6	Nuclear pore complex protein Nup50	KY770903–KY770912
	R: TAAGCAGCACTTTTCCCACA					
orange1.1g015495m (intron8)	F: CTGCTGGAAATGCCTCTAGC	151	156	49.0	Acetyl-CoA C-acetyltransferase/ acetoacetyl-CoA thiolase	KY770891–KY770902
	R: CTGAGCAGCGTCAGCATATC					
orange1.1g027952m (intron5)	F: TGGTTTTGATTGATGCAAGTG	533	495	58.6	Sf825-hydrolase, alpha/beta fold family protein	KY770879–KY770890
	R: GCATCTTCCCACCAAGGATA					
orange1.1g022777m (intron3)	F: GGAGGATTTCAATGAGGCTCT	482	781	58.6	Coatomer subunit epsilon	KY770868–KY770878
	R: TCTCAGCATAATCAGACCTGTG					
orange1.1g009973m (intron5)	F: AGTGGAACTGCTTCGCAAGT	855	470	49	Sf155-kh domain-containing RNA binding protein	KY770861–KY770867
	R: TGCATATGGGTTATAGCCTTGA					

Note: T_a = annealing temperature.
[a]Reference locus in *Citrus sinensis* v1.0 genome.

TABLE 2. Summary statistics of the nine putative single-copy nuclear markers developed to amplify across Paullinieae (Sapindaceae).[a]

Locus[b]	No. of taxa in alignment	Parsimony informative sites	Total characters in multiple sequence alignment	Pairwise identity (%)
orange1.1g002083m (intron9)	10	15	358	88.3
orange1.1g047192m (intron7)	12	45	665	84.5
orange1.1g045023m (intron7)	8	52	814	94.3
orange1.1g045023m (intron6)	6	8	301	87.7
orange1.1g028997m (intron4)	10	50	852	71.8
orange1.1g015495m (intron8)	12	11	156	94.3
orange1.1g027952m (intron5)	12	29	495	94.2
orange1.1g022777m (intron3)	11	52	781	69.8
orange1.1g009973m (intron5)	7	64	470	53.7
ITS (for comparison)	7	34	708	69.1

[a] Species code, names, and collection information are listed in Appendix 2.
[b] Reference locus in *Citrus sinensis* v1.0 genome.

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

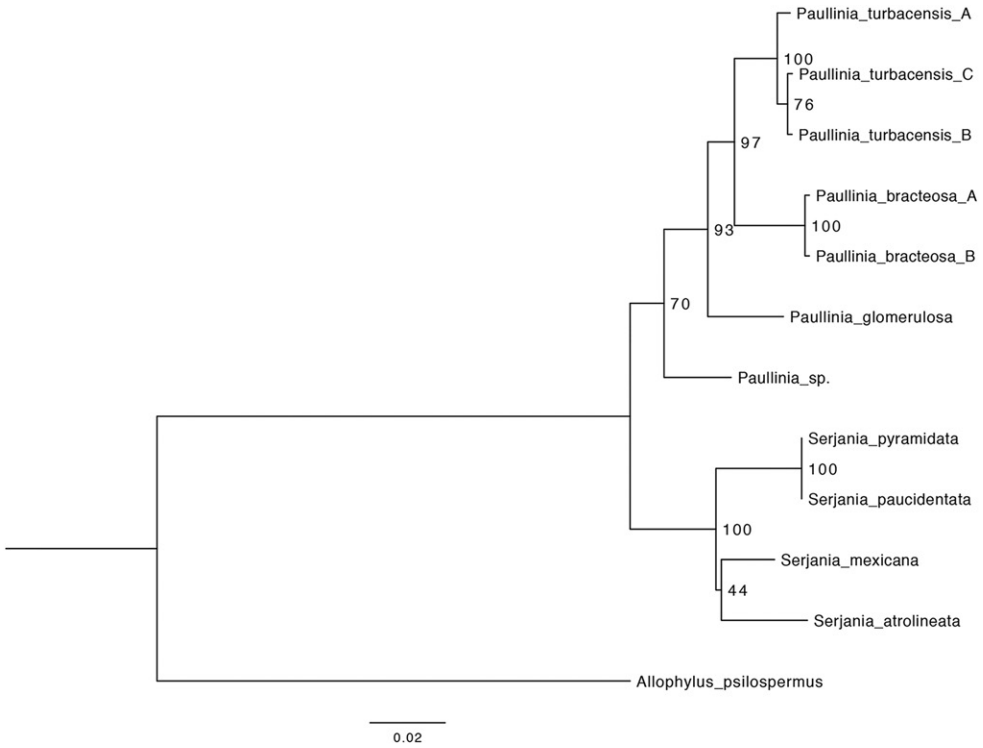

Fig. 2. Maximum likelihood phylogenetic tree of nine Sapindaceae species using concatenated alignment of nine newly developed putative single-copy nuclear markers.

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, 20× 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 Paullinieae, 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

Acevedo-Rodríguez, P. 1993. Systematics of *Serjania* (Sapindaceae). Part I: A revision of *Serjania* Sect. *Platycoccus*. Memoirs of the New York Botanical Garden, vol. 67. New York Botanical Garden Press, Bronx, New York, USA.

Acevedo-Rodríguez, P., K. J. Wurdack, M. S. Ferrucci, G. Johnson, P. Dias, R. G. Coelho, G. V. Somner, et al. 2017. Generic relationships and classification of tribe Paullinieae (Sapindaceae) with a new concept of supertribe Paulliniodae. *Systematic Botany* 42: 96–114.

Chamala, S., N. García, G. T. Godden, V. Krishnakumar, I. E. Jordon-Thaden, R. De Smet, W. B. Barbazuk, D. E. Soltis, et al. 2015. MarkerMiner 1.0: A new application for phylogenetic marker development using angiosperm transcriptomes. *Applications in Plant Sciences* 3: 1400115.

Curto, M. A., P. Puppo, D. Ferreira, M. Nogueria, and H. Meimberg. 2012. Development of phylogenetic markers from single-copy nuclear genes for multi locus, species level analyses in the mint family (Lamiaceae). *Molecular Phylogenetics and Evolution* 63: 758–767.

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

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

Granados Mendoza, C., J. Naumann, M.-S. Samain, P. Goetghebeur, Y. De Smet, and S. Wanke. 2015. A genome-scale mining strategy for recovering novel rapidly-evolving nuclear single-copy genes for addressing shallow scale phylogenetics in *Hydrangea*. *BMC Evolutionary Biology* 15: 132.

Katoh, K., and D. M. Standley. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* 30: 772–780.

Koressaar, T., and M. Remm. 2007. Enhancement and modifications of primer design program Primer3. *Bioinformatics (Oxford, England)* 23: 1289–1291.

Li, C., Y. Wang, X. Huang, J. Li, H. Want, and J. Li. 2013. De novo assembly and characterization of fruit transcriptome in *Litchi chinensis* Sonn and analysis of differentially regulated genes in fruit in response to shading. *BMC Genomics* 14: 552.

Li, H., B. Handsaker, A. Wysoker, T. Fennel, J. Ruan, N. Homer, G. Marth, et al. (1000 Genome Project Data Processing SubGroup). 2009. The sequence alignment/map format and SAMtools. *Bioinformatics (Oxford, England)* 25: 2078–2079.

Miller, M. A., W. Pfeiffer, and T. Schwartz. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees, 1–8. Proceedings of the Gateway Computing Environments Workshop (GCE), 14 November 2010, New Orleans, Louisiana, USA.

Mirarab, S., M. S. Bayzid, B. Boussau, and T. Warnow. 2014. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science* 346: 1250463.

Nielsen, R., J. S. Paul, A. Albrechtsen, and Y. S. Song. 2011. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews. Genetics* 12: 443–451.

Peñalba, J. V., L. L. Smith, M. A. Tonione, C. Sass, S. M. Hykin, P. L. Skipwith, J. A. McGuire, et al. 2014. Sequence capture using PCR-generated probes: A cost-effective method of targeted high-throughput sequencing for nonmodel organisms. *Molecular Ecology Resources* 14: 1000–1010.

Sang, T. 2002. Utility of low-copy nuclear gene sequences in plant phylogenetics. *Critical Reviews in Biochemistry and Molecular Biology* 37: 121–147.

Sass, C., W. J. D. Iles, C. F. Barrett, S. Y. Smith, and C. D. Specht. 2016. Revisiting the Zingiberales: Using multiplexed exon capture to resolve ancient and recent phylogenetic splits in a charismatic plant lineage. *PeerJ* 4: e1584.

Singhal, S. 2013. De novo transcriptomic analyses for non-model organisms: An evaluation of methods across a multi-species data set. *Molecular Ecology Resources* 13: 403–416.

Stamatakis, A. 2014. RAxML Version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)* 30: 1312–1313.

Stockenhuber, R., S. Zoller, R. Shimizu-Inatsugi, F. Gugerli, K. K. Shimizu, A. Widmer, and M. C. Fischer. 2015. Efficient detection of novel nuclear markers for Brassicaceae by transcriptome sequencing. *PLoS ONE* 10: e0128181.

Swofford, D. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Sinauer, Sunderland, Massachusetts, USA.

Tonnabel, J., I. Oliveieri, A. Mignot, A. Rebelo, F. Justy, S. Santoni, S. Caroli, et al. 2014. Developing nuclear DNA phylogenetic markers in the angiosperm genus *Leucadendron* (Proteaceae): A next generation sequencing transcriptomic approach. *Molecular Phylogenetics and Evolution* 70: 37–46.

Untergasser, A., I. Cutcutache, T. Koressaar, J. Ye, B. C. Faircloth, M. Remm, and S. G. Rozen. 2012. Primer3–New capabilities and interfaces. *Nucleic Acids Research* 40: e115.

Uribe-Convers, S., M. L. Settles, and D. C. Tank. 2016. A phylogenomic approach based on PCR target enrichment and high throughput sequencing: Resolving the diversity within the South American species of *Bartsia* L. (Orobanchaceae). *PLoS ONE* 11: e0148203.

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

Wu, G. A., S. Prochnik, J. Jenkins, J. Salse, U. Hellsten, F. Murat, X. Perrier, et al. 2014. Sequencing of diverse mandarin, pummel and orange genomes reveals complex history of admixture during citrus domestication. *Nature Biotechnology* 32: 656–662.

Appendix 1. Description of the bioinformatic pipeline used to isolate single-copy nuclear introns of a desired size by use of a genome and two sets of transcriptome reads, followed by the primer design and primer testing protocols. All scripts are included in the corresponding directory on https://github.com/joycechery/Sapindaceae or the link is provided in the appendix text. In addition, command syntax for external software, a bash command file, and a description of a process are given.

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/Parent=//'|
sed 's/m.g/m/' > mrnaNames
```

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 mrnaNames intronsize > mrnaIntronSize
```

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

```
sort mrnaNames | uniq -c | sort | awk '$1<2' | 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<1000' 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}' > singlemRNAintrons_small_uniq
```

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(-fh => \*STDIN, -format => 'fasta'); while (my $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 '{FS ="|"}{print $1}' Csinensis_v1.0_cds_seqs.fa > Csinensis_v1.0_cds_orangeNames.fa
```

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

`-selectSeqs.pl` (available at: http://raven.iab.alaska.edu/~ntakebay/teaching/programming/perl-scripts/perl-scripts.html)
```
perl selectSeqs.pl -in Csinensis_v1.0_cds_orangeNames.fa -out SelectedCitrusSeqs.fa -idfile singlemRNAintrons_
small_uniq
mv InSelectedCitrusSeqs.fa SelectedCitrusSeqs.fa
```

2. Transcriptome read processing:

The following commands and scripts clean the transcriptome reads.

a. Download the paired-end reads of the phylogenetically closest transcriptome to the Paullinieae (*Dimocarpus longan*) from the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/) database and split using the NCBI SRA Toolkit version 2.4.5-2.

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

```
sed '/^@SRR412534/ s/SRR412534.[0-9]* //' SRR412534_1.fastq | sed'/^@A807H/ s/length=90/\/1/g' | sed
'/^+SRR412534/ s/.*/\+/' > JCSR_R1.fq
sed '/^@SRR412534/ s/SRR412534.[0-9]* //' SRR412534_2.fastq |sed'/^@A807H/ s/length=90/\/2/g' | sed
'/^+SRR412534/ s/.*/\+/' > JCSR_R2.fq
```

c. 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)
`-2-scrubReads.pl` (available at: https://github.com/MVZSEQ/SCPP/blob/master/2-scrubReads.pl)

3. 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 3.01 (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 MarkDuplicates 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 consensuses 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 pseudoreference from which primers are designed.

–AlignFix.sh (available at: https://github.com/joycechery/Sapindaceae/blob/master/AlignReadstoCDS/AlignFix.sh)
–AlignFix.pl (available at: https://github.com/joycechery/Sapindaceae/blob/master/AlignReadstoCDS/AlignFix.pl)

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:

–BlastToRemove_Paralogs (available at: https://github.com/joycechery/Sapindaceae/blob/master/FilterSequences/BlastToRemove_Paralogs)

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

–BlastToRemove_nonnuclear (available at: https://github.com/joycechery/Sapindaceae/blob/master/FilterSequences/BlastToRemove_nonnuclear)

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

i. Create a list of pseudoreferences names:

```
grep '>' LocalBlastRemoveNCBIRemove.fa | sed 's/>//' > LocalBlastRemoveNCBIRemove_Names
```

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

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

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

```
grep -f LocalBlastRemoveNCBIRemove_Names JCSR.positioncoverage > LocalBlastRemoveNCBIRemove_PositionCoverage
```

iv. Create an average gene coverage file:

–AverageGeneCoverage.pl (available at: https://github.com/joycechery/Sapindaceae/blob/master/FilterSequences/AverageGeneCoverage.pl)

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

```
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 InLocalBlastRemoveNCBIRemove20XCov.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:

```
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 clean 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:
–parse_gff_to_bed.pl (available at: https://github.com/chodon/zingiberales/blob/master/2MusaCDSprocess/parse_gff_to_bed.pl)

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 cycles of 5 s at 98°C, 5 s at loci-specific annealing temperature, and 20 s at 72°C; and a final 1-min 72°C extension.
c. Apply optimal PCR conditions to all samples.
d. 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.
–ASTRAL-II-Commands (available at: https://github.com/joycechery/Sapindaceae/blob/master/ASTRAL-II/ASTRAL-II-Commands)

Appendix 2. Silica-dried specimens used in this study.

Sample ID[a]	Species name	Collection no.[a]	Locality	Latitude	Longitude
P0	*Paullinia* sp.	UCBG no. 92.0509	Provenance: Chiapas State, Mexico, North America	—	—
P1	*Paullinia bracteosa*_A	NTBG no. 760259	Provenance: Venezuela	—	—
P3	*Serjania mexicana*	Chery 23	Barro Colorado Island, Panama	09°09.914′N	079°50.213′W
P5	*Serjania paucidentata*	Chery 34	Barro Colorado Island, Panama	09°09.905′N	079°50.202′W
P6	*Serjania pyramidata*	Chery 29	Barro Colorado Island, Panama	09°10.867′N	079°49.444′W
P7	*Paullinia glomerulosa*	Chery 20	Barro Colorado Island, Panama	09°09.896′N	079°50.294′W
P8	*Paullinia turbacensis*_A	Chery 13	Barro Colorado Island, Panama	09°09.924′N	079°50.188′W
P9	*Paullinia turbacensis*_B	Chery 24	Barro Colorado Island, Panama	09°09.830′N	079°50.270′W
P10	*Paullinia turbacensis*_C	Chery 10	Barro Colorado Island, Panama; Donata Trail	—	—
P11	*Serjania atrolineata*	Chery 42	Barro Colorado Island, Panama	09°10.705′N	079°50.819′W
P12	*Paullinia bracteosa*_B	Chery 26	Barro Colorado Island, Panama; Lake #4 Trail Marking	—	—
P13	*Allophylus psilospermus*	Chery 19	Barro Colorado Island, Panama	09°09.910′N	079°50.287′W

[a] 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.

Locus[a]	P0	P1	P3	P5	P6	P7	P8	P9	P10	P11	P12	P13
orange1.1g002083m (intron9)	+	+	+	—	—	+	+	+	+	+	+	+
orange1.1g047192m (intron7)	+	+	+	+	+	+	+	+	+	+	+	+
orange1.1g045023m (intron7)	+	+	—	+	+	+	+	+	+	—	—	—
orange1.1g045023m (intron6)	+	+	—	+	+	—	—	—	—	+	—	+
orange1.1g028997m (intron4)	+	+	+	—	+	—	+	+	+	+	+	+
orange1.1g015495m (intron8)	+	+	+	+	+	+	+	+	+	+	+	+
orange1.1g027952m (intron5)	+	+	+	+	+	+	+	+	+	+	+	+
orange1.1g022777m (intron3)	+	—	+	+	+	+	+	+	+	+	+	+
orange1.1g009973m (intron5)	+	+	+	+	+	—	—	—	—	—	+	+

Note: + = successful amplification and sequencing; — = no amplification or unsuccessful sequencing.
[a]Reference locus in *Citrus sinensis* v1.0 genome.