Short Communication

Target sequence capture in the Brazil nut family (Lecythidaceae): Marker selection and in silico capture from genome skimming data

Oscar M. Vargas⁎, Myriam Heuertzb, Stephen A. Smithc, Christopher W. Dickb,c

a Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA
b Biogeco, INRA, Univ. Bordeaux, F-33610 Cestas, France
c Smithsonian Tropical Research Institute, Panama City 0843-03092, Panama

ABSTRACT

Reconstructing species trees from multi-loci datasets is becoming a standard practice in phylogenetics. Nevertheless, access to high-throughput sequencing may be costly, especially with studies of many samples. The potential high cost makes a priori assessments desirable in order to make informed decisions about sequencing. We generated twelve transcriptomes for ten species of the Brazil nut family (Lecythidaceae), identified a set of putatively orthologous nuclear loci and evaluated, in silico, their phylogenetic utility using genome skimming data of 24 species. We designed the markers using MarkerMiner, and developed a script, GoldFinder, to efficiently sub-select the best makers for sequencing. We captured, in silico, all designed 354 nuclear loci and performed a maximum likelihood phylogenetic analysis on the concatenated sequence matrix. We also calculated individual gene trees with maximum likelihood and used them for a coalescent-based species tree inference. Both analyses resulted in almost identical topologies. However, our nuclear-loci phylogenies were strongly incongruent with a published plastome phylogeny, suggesting that plastome data alone is not sufficient for species tree estimation. Our results suggest that using hundreds of nuclear markers (i.e. 354) will significantly improve the Lecythidaceae species tree. The framework described here will be useful, generally, for developing markers for species tree inference.

1. Introduction

Inferring species-level phylogenies is a pivotal step in addressing broader evolutionary questions. This task is particularly useful and difficult in tropical organisms as samples may be difficult to obtain and clades tend to be species-rich. Most plant phylogenies to date are based on sequences from few markers, mostly of plastid origin, that typically have insufficient signal to infer robust phylogenies at the species level (Gitzendanner et al., 2018). Furthermore, because plastid markers represent a single phylogenetic history, due to the non-recombinant and uniparental inheritance of plastids (chloroplasts) in plants (Birky, 1995; Ruhlman and Jansen, 2014), the plastid tree can be potentially biased in relation to the species tree. Increasing evidence of conflicting topologies between plastid and nuclear DNA suggest that plastid markers might perform especially poorly for species tree recovery in groups with high levels of recent and ancient hybridization (Rieseberg and Soltis, 1991; Sun et al., 2015; Folk et al., 2016; Pérez-Escobar et al., 2016; Bruun-Lund et al., 2017; Vargas et al., 2017; Morales-Briones et al., 2018).

Nuclear markers have been historically underused in plant phylogenetics, with the exception of the internal transcribed spacers of the nuclear ribosomal DNA (ITS). While the ITS region tends to be useful for inferring relationships among closely related species because of its high variation, it is inefficient for phylogenetics at higher levels (Hughes et al., 2006). Additionally, ITS can suffer from misleading polymorphism due to its multicopy nature and concerted evolution (Álvarez and Wendel, 2003). Numerous single and low copy nuclear markers have been proposed as useful for plant phylogenetics (Zhang et al., 2012), but these have not been widely incorporated, likely because primers have to work universally (across different plant groups), and the low or single copy nature of nuclear regions hinders their PCR amplification in degraded DNA, typically found in herbarium specimens.

The drawbacks and challenges described above reveal the need to implement methods to obtain multiple and independent nuclear loci for species tree estimation. RNA-seq and genotyping by sequencing techniques have shown to be of great utility for plant phylogenetics (McVay et al., 2017; Zimmer and Wen, 2015), yet these techniques require high

⁎ Corresponding author at: University of California, Santa Cruz, USA.
E-mail address: oscarvargas@gmail.com (O.M. Vargas).

https://doi.org/10.1016/j.ympev.2019.02.020
Received 31 July 2018; Received in revised form 22 February 2019; Accepted 23 February 2019
Available online 25 February 2019
1055-7903/ © 2019 Elsevier Inc. All rights reserved.
quality tissue, making their application unfeasible in herbarium collections. Target sequence capture, which sequences regions of interest after their hybridization to probes, on the other hand, has proven to be an effective method for sequencing hundreds of nuclear loci from tissue with high quality DNA as well from herbarium specimens (Mandel et al., 2014; Weitemier et al., 2014). A typical workflow for a target enrichment study starts with mining genomic resources, typically transcriptomes, to identify the markers to be captured through probes or “baits” (custom single stranded oligonucleotides) in a DNA hybridization assay. While it has been suggested that universal baits could be used for any angiosperm taxa (Budenhagen et al., 2016; Cowan and denovo bait design produces better results—yielding longer sequences and capturing more markers per sample (Kadlec et al., 2017).

MarkerMiner (Chamala et al., 2015) is a widely used workflow to identify markers for target sequencing. MarkerMiner requires at least one transcriptome, and, by comparison to a database of low/single copy markers, produces a set of alignments from which baits are designed. The output of MarkerMiner typically contains hundreds of markers from the which the user usually has to subsample, aiming to multiplex during sequencing. With the objective of making the sub-selection task automatic and informed, we wrote GoldFinder. GoldFinder is intended to be used to identify the optimal markers for sequencing according to the five criteria: (i) marker length, (ii) percentage of short exons (relative to bait length), (iii) number of user’s sequences per marker, (iv) similarity, and (v) bait number, length, and coverage. Additionally, GoldFinder splits initial marker alignments into exon-alignments (based on a transcriptomic reference with the introns masked as N’s). This improves bait design by “fitting” the baits to the edges of the exons, increasing the efficiency of the hybridization assay by avoiding the extension of baits onto multiple exons (that might be separated by an intron).

In this study, we examined Lecythidaceae, a family of woody plants that is ecologically dominant in Amazon forests (ter Steege et al., 2006). Phylogenetic relationships of the New World Lecythidaceae, also known as the subfamily Lecythidoideae (Mori et al., 2017), have been recently examined using plastid markers, ITS, and morphology, which revealed shallow evolutionary relationships and a backbone tree with low statistical support (Huang et al., 2015). In an effort to improve the phylogeny, Thomson et al. (2018) inferred a robust backbone phylogeny using the complete plastome sequences of 24 species. Thomson's et al. tree largely agreed with that of Huang et al. (2015) adding support to the finding of Huang et al. (2015) that Eschweilera and Lecythis are nonmonophyletic and comprise the Bertholletia clade along with Bertholletia and Corythophora. Thus, the phylogenetic study of the Lecythidaceae to date has been dominated by the use of plastid markers (Huang et al., 2015; Mori et al., 2007; Thomson et al., 2018) making a case for employing nuclear DNA.

We here tested the utility of target enrichment by designing nuclear markers from transcriptomes (employing MarkerMiner and GoldFinder) and capturing those markers in silico from available genome scaffolding data using custom scripts. We used the results from these analyses to determine how well the set of nuclear markers flagged by our analysis could be used to produce a robust nuclear phylogeny, and if this phylogeny was concordant with the plastome phylogeny of Thomson et al. (2018).

2. Material and methods

Control files with commands and parameters, intermediate data files, and custom python scripts can be found at https://bitbucket.org/oscavargash/lecythidaceae_transcriptomics. GoldFinder, our newly developed python script to sub-select markers from the output of MarkerMiner (Chamala et al., 2015) can be found at https://bitbucket.org/oscavargash/goldfinder.

2.1. Transcriptomes

We sequenced a total of twelve transcriptomes from leaf and/or flower tissue of ten Lecythidaceae species, eight of them belonging to the New World subfamily Lecythidoideae ((Mori et al., 2017), Table 1 and Supplementary Table 1). Seven of the twelve tissue samples were collected in RNAlater (Thermo Fisher Scientific, Vilnius, Lithuania), then processed with the PureLink RNA Mini Kit (Invitrogen, Carlsbad, California, USA) for RNA extraction and with the KAPA mRNA HyperPrep Kit (KAPA Biosystems, Wilmington, Massachusetts, USA) for library preparation. For the remaining five tissue samples, also collected in RNAlater, we extracted the RNA employing a CTAB-based method (Le Provost et al., 2007) and then prepared the libraries with the Ion Total RNA-Seq Kit 2 (Ion Torrent). The first seven transcriptomes were sequenced in one lane of an Illumina Hi-Seq 2500 (Illumina Inc., San Diego, CA, USA) at the DNA Sequencing Core facility of the University of Michigan, outputting paired-end sequences of 125 bp. The second group, comprised of five samples, was sequence in an Ion Proton System (Ion Torrent) at the Genome Transcriptome Facility of Bordeaux (PGTB), outputting single-end reads of variable length. We employed two different techniques because we were originally two teams, then later decided to join efforts over a common goal.

Raw reads were processed with SeqyClean (Zhambanov et al., 2017), trimming terminal nucleotides that averaged a Phred score of 10 or less (following Macmanes, 2014) on a sliding window of 10 bp, poly-A/T tails were also removed. We employed Trinity (Grabherr et al., 2011) to assemble transcripts from filtered reads. On the twelve assembled transcriptomes we applied the Yang and Smith’s (2014) pipeline using the RT method (rooted ingroups) to estimate the total number of orthologs in the dataset. The Yang and Smith (2014) method is a clustering pipeline in which orthologous sequences are identified with the help of phylogenetic trees. This method does not require an
external reference and was designed for phylogenomic analysis.

2.2. Marker development

For computational efficiency (analyses with 6 transcriptomes or more lasted longer than 4 days and were subsequently killed) and to avoid redundancy designing the baits, we employed MarkerMiner (Chamala et al., 2015) on five transcriptomes (out of 12). We selected samples with a high number of transcripts that comprise a wide phylogenetic diversity for subfamily Lecythiioideae, the New World Lecythidaceae (Mori et al., 2017): Barringtonia racemosa (outgroup), Eschweilera sagotanana, Grias cauliflora, Gustavia superba (individual #2 [two individuals were sequenced for this species]), and Lecythis congestiflora. We ran MarkerMiner using Arabidopsis thaliana (L.) Heynh. as a reference and with a minimum transcript length of 400 base pairs. To efficiently sub-select a set of markers (=transcripts) from the 1528 selected by MarkerMiner and aiming to sequencing 700 k to 1 million bp after the hybridization assay, we developed a python script named GoldFinder that selects the best markers based on the following parameters (characters in brackets indicate the argument used in the script for this study):

- Minimum length [-ml 400], markers are ranked by length, only the ones supersizing the minimum length are kept.
- Maximum percentage of short exons (relative to bait length) [-pse 30], markers might have exons that are shorter than baits, which typically are 120 bp, hindering their recovery in the hybridization assay. This parameter allows the user to filter out markers that contain excessive percentage of the sequence representing short exons.
- Number of user’s sequences per marker [-ns 2], it is recommended to include markers that are represented in at least two transcriptomes, allowing GoldFinder to assess the number of identical sites in the alignment providing a proxy for molecular divergence among samples.
- Percentage of identical sites in the alignment (excluding the reference) [-pis 50], the percentage of identical sites calculated by GoldFinder represents a conservative estimate (number columns with identical characters/length of the alignment) *100) of the overall similarity in the alignment (an overall similarity of > 75% is recommended). A higher percentage of identical sites would result in a higher probability of success in the hybridization assay. Markers under the percentage threshold provided by the user are filtered out.

We set up this parameter to 50 because a more stringent value (i.e. 75) resulted in too few markers.

Additionally, GoldFinder allows the user to modify the length of the baits (default -bl 120), the bait coverage (default -bc 2), and the total number of baits desired (default -nb 30,000). GoldFinder outputs a set of folders (mirroring the folders produced by MarkerMiner) which contain the alignments of sub-selected markers. Finally, GoldFinder splits orthologous transcript alignments into exon alignments for better bait design.

2.3. In silico capture

In order to test the efficacy of the nuclear markers identified by MarkerMiner and sub-selected by GoldFinder, we performed capture in silico with genome skimming data originally used for plastome assembly of 24 Lecythiideae species (Thomson et al., 2018). First, we cleaned and trimmed the genome skimming reads with SeqyClean using the same parameters as for the transcriptomic dataset. Then, we retained only the nuclear reads by filtering out the reads mapping to chloroplast (Eschweilera congestiflora MF359937.1 [GenBank accession number]), mitochondrial (Vacinium macrocarpon Ation NC_023383.1), or ribosomal DNA (Eschweilera congestiflora JN222324.1, JN222317.1) using bbnap.sh (Bushnell, 2015) with default parameters. For each of 354 markers (see Results), we selected as a reference the longest transcriptomic sequence from the MarkerMiner alignment using the custom script longest_seq.fasta.py. We then mapped the nuclear reads of each genome skimming sample to each of our 354 marker reference sequences and built a consensus sequence per marker per sample, employing a custom script reads2sam2consensus.baits.py, which employs b unwrap.sh (Bushnell, 2015) and sam2consensus.py (available from https://github.com/edgardomortiz/sam2consensus). The resulting consensus sequences (one fasta file per sample per marker) were sorted in folders by marker (354 folders corresponding to 354 markers) using the custom script baits_file_organizer.py. Consensus sequences belonging to the same marker were combined into a single fasta file using the custom script cat_fastas_per_gene.py, these were subsequently aligned using prank_wrapper.py (Yang and Smith, 2014). phyutility_wrapper.py trimmed the alignments and concatenate matrices.py produced the supermatrix (both scripts from Yang and Smith, 2014).

2.4. Phylogenetic analysis

To infer a phylogeny with the concatenated matrix of in silico captured sequences, we searched for the best-scoring maximum likelihood (ML) tree and performed 1000 rapid bootstraps employing the option “-f a” in RAxML 8.2.11 (Stamatakis, 2014), using an independent GTRGAMMA model for each of the 354 markers. We also inferred a coalescent-based species tree from gene trees calculated with RAxML (using the same parameters described above) with ASTRAL-III (Miraраб et al., 2014; Zhang et al., 2017), which infers a species tree from gene trees accounting for the incongruence produced by incomplete lineage sorting (ILS). Because both phylogenetic analyses (species tree from the concatenated sequence matrix vs. coalescent-based species tree) produced very similar results (see Results), we arbitrarily selected the best-scoring ML tree from the concatenated sequence matrix (hereafter called the nuclear tree) to carry out our phylogenetic conflict analyses. We visually compared our nuclear tree with the plastome phylogeny of Thomson et al. (2018) using the cophyplot function of the R (R Core Team, 2018) package ape (Paradis et al., 2004). We employed phy parts (Smith et al., 2015) to calculate the amount of conflict inside the nuclear markers by comparing the nuclear gene trees against the multi-locus ML tree. We also calculated the number of concordant/conflicting nuclear markers against the plastome topology. phy parts results were visualized with phyparts piecharts (available from https://github.com/mossmatters/MJPYthonNotebooks).

3. Results

We assembled de novo 12 transcriptomes for 10 Lecythiideae species (GenBank BioSample accessions SAMN10963033–SAMN10963044). The average number of transcripts assembled by Trinity was 145,547, with Couroupita guianensis having the minimum of 52,314 and Barringtonia racemosa having the maximum of 220,910 (Table 1). After applying the Yang and Smith (2014) pipeline, we obtained a total of 10,025 orthologs, for which the average number of orthologs per sample was 5213, with Eschweilera coriacea (2) having the minimum of 82 and Gustavia superba (1) having the maximum of 9746 (Table 1). Samples sequenced using the Ion Torrent platform presented a considerate drop in orthologs in our analysis (1) having the maximum of 97% (percentage of the matrix with data present).

The phylogenetic analyses performed on the complete nuclear
dataset with RAxML and ASTRAL resulted in a similar topology with the
only differences being the relationship of closely related species of
Eschweilera. *Eschweilera pittieri* is sister to *E. wachenheimii* in the RAxML
tree, while *E. pittieri* is sister to a clade comprised of *E. wachenheimii*
and *E. alata* + *E. micrantha* in the ASTRAL tree: both conflicting nodes
had low statistical support (BS < 80, BP < 90) whereas all other
nodes, with the exception of one in both trees, had high support
(BS ≥ 80, BP ≥ 90) (Fig. 1b, Supplementary Fig. 1). When our nuclear
(RAxML) phylogeny was compared with the plastome phylogeny ob-
tained by Thomson et al. (2018), a significant pattern of incongruence
was revealed, specifically for the relationships inside the Bertholletia
clade (Fig. 1). While the plastome phylogeny suggested that both
*Eschweilera* and *Lecythis* are polyphyletic, the nuclear phylogeny re-
covered *Eschweilera* as monophyletic (the Integrifolia and the Parvifolia
clades of Huanget al. (2015) clades are sister, the Tetrapetala clade was
not sampled by Thomson et al. (2018) and therefore is not represented
in our trees) and *Lecythis* as polyphyletic.

Both analyses of phylogenetic conflict with *phyparts* revealed a drop
in informative genes and an increase in conflict inside the Bertholletia
clade (Fig. 2) with only three clades (*Eschweilera caudiculata* + *E. in-
tegriofila*, *Corythophora amapaensis* + *C. labriculata*, and *Lecythis corrugata* + *L. pneumatophora*) being supported with more than half of the
informative markers. Ten nodes presented considerable conflict (con-
flicting markers > concordant markers) within our nuclear dataset
(Fig. 2a), nine of which are nested in the Bertholletia clade. Our results
also show that there is a strong conflict between the nuclear gene trees
and the plastome phylogeny in eleven nodes (Fig. 2b).

### 4. Discussion

#### 4.1. Utility of transcriptomics for plant systematics

High-throughput sequencing provides unprecedented opportunities
for systematists and evolutionary biologists, with data yields often at
least one order of magnitude higher than traditional sequencing tech-
niques. The sequencing of multiple transcriptomes, along with the ap-
plication of MarkerMiner for marker development in Lecythidaceae,
revealed 1528 low/single copy loci with the potential to be used for
phylogenetic analyses. An informed sub-selection of these markers with
our newly developed script GoldFinder resulted in a set of 354 loci
containing 1754 exons. The in silico captured concatenated markers
produced an aligned supermatrix of 758,015 bp. A phylogenetic analysis
carried out on the 354-marker concatenated matrix resulted in the
first available, albeit preliminary, robust Lecythidaceae backbone nuclear
phylogeny.

#### 4.2. Phylogenetic discordance

Our Lecythidaceae nuclear backbone phylogeny conflicts with the
plastome phylogeny of Thomson et al. (2018) (Fig. 1) and with that of
Huang et al. (2015) which was inferred with plastid regions (*ndhF*, *trnL-
E*, *trnH-psbA*), ITS, and morphology; the plastid phylogeny of Thomson
et al. (2018) largely agrees with that of Huang et al. (2015). Specifi-
cally, the nuclear markers strongly disagree with eleven nodes in the
plastome phylogeny of Thomson et al. (2018), nine of which are nested
inside the Bertholletia clade (Fig. 2b). Incongruence between nuclear

---

**Fig. 1.** Comparison between the maximum likelihood phylogenies derived from (a) the complete plastome alignment of Thomson et al. (2018) and (b) 354 nuclear
markers. All nodes have a bootstrap support of 100 unless noted otherwise. Lines between taxa indicate a conflicting position between the two topologies.
Fig. 2. Agreement and conflict of the 354 makers on the (a) maximum likelihood (ML) nuclear concatenated phylogeny and (b) plastome topology of Thomson et al. (2018). Pie charts indicate proportion of genes that agree (blue), support a main alternative topology (green), support remaining alternatives (red), and are uninformative (gray) for a given node on the underlying topology. Number above the nodes show the number of concordant genes, while number under the nodes indicate the total number of conflicting genes (support main alternative + support remaining alternatives). Bar indicates the Bertholletia clade sensu Huang et al. (2015). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
and plastome topologies can be the result of systematic error, incomplete lineage sorting (ILS), and/or hybridization (Maddison, 1997; Rodríguez-Ezepeleta et al., 2007). While the purpose of this study was not to identify the cause of this phylogenetic conflict, we believe that ancient hybridization may best explain the conflict between the nuclear and plastome phylogenies, especially in deeper nodes representing generic relationships. The latter hypothesis is supported by the result that both nuclear trees, the concatenated RAxML and the coalescent-based ASTRAL, resulted in the same generic relationships (Fig. 1b, Supplementary Fig. 1), partially ruling out ILS for deeper nodes.

In addition to the nuclear-plastome incongruence, there is also conflict among the nuclear genes, as evidenced by ten nodes in the nuclear topology for which there is strong conflict (Fig. 2a). The fact that nine nuclear-conflicting nodes are nested in the Bertholletia clade, and that our set of nuclear markers strongly conflicts with nodes positioned in similar locations along the plastome phylogeny, suggest the presence of lower phylogenetic signal and greater conflict within the Bertholletia clade. Both plastome and nuclear phylogenoms (Fig. 2) show short branches at the base of the Bertholletia clade suggesting rapid diversification in this part of the tree. Rapid speciation in the Bertholletia clade can explain the drop in phylogenetic signal and might have involved hybridization (Wiens et al., 2006). A formal test with greater taxonomic sampling and robust coverage is needed to test the hybridization hypothesis.

4.3. Limitations of our study

While our results are encouraging, our phylogeny is still preliminary and should be taken with caution. For example, individual samples in each gene are based on consensus sequences derived from genome skimming data with low nuclear genomic coverage. Some of our consensus sequences likely suffer from problems due to missing data. Furthermore, it is unfeasible to confidently assess orthology with this dataset. Although our phylogeny contains all neotropical Lecythidaceae genera, it contains 24 species representing only ~10% of the total number of species. Finally, we noticed a drop in the number of orthologous sequences recovered with the Ion Torrent dataset when employing the pipeline of Yang and Smith (2014) (Table 1). We believe this drop was caused because the amount of data obtained with the Ion Torrent platform was lower than that obtained with Illumina (Ion Torrent reads are single vs. paired) and their length is variable, which resulted in shorter and incomplete transcripts and fewer coding regions recognized by Transdecoder (https://github.com/TransDecoder), the step in the Yang and Smith (2014) pipeline for which we observed the drop.

5. Conclusions

Our results demonstrate that Lecythidaceae nuclear and plastome phylogenies differ, suggesting the importance of gathering more nuclear data for additional taxa. The use of the 354 markers is expected to yield a more accurate Lecythidaceae species tree hypothesis, albeit with conflict in contentious nodes (i.e. nodes within the Bertholletia clade). We demonstrated the utility of transcriptomes and genome skimming data to design and test markers for species tree inference. Our framework will be valuable for others wanting to make informed decisions on planning for species-level sequencing in future projects.

Acknowledgements

This work has benefited from financial support from the National Science Foundation (grant no. DEB 1240869 and FSEI Type I 1338694 to C.W.D.), the University of Michigan (Associate Professor Award to C.W.D.), and the Investissements d’Avenir grants of the ANR (Agence Nationale de la Recherche, France), CEBA:ANR-10-LABX-23-01 (COLLEVOL project, awarded to M.H.). The authors thank Grégoire Le Provost and Christophe Boury for assistance with laboratory work and Diego Alvarado, Giorgia Auteri, and Gregory Stull for their useful comments.

Data archiving

Transcriptomes and raw reads are available in GenBank under the BioSamples SAMN10963033–SAMN10963044.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jympev.2019.02.020.

References


