Molecular Phylogenetics and Evolution 135 (2019) 98–104

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Short Communication

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

Oscar M. Vargas[a,*], Myriam Heuertz[b], Stephen A. Smith[a], Christopher W. Dick[a,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

ARTICLE INFO

Keywords:
Lecythidaceae
Markers
MarkerMiner
Species Tree
Target sequencing
Transciptomes

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: oscarvargash@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/

Table 1
Summary statistics of the transcriptomes obtained in this study. The number of orthologous transcripts was calculated using Yang and Smith's (2014) pipeline and correspond to the number of orthologs of a given sample in an alignment with at least six samples using the rooting ingroups method.

Species	Platform	Raw reads	Filtered reads	Read range	Transcripts	Orthologous transcripts
Barringtonia racemosa (L.) Spreng.	Illumina	113,599,384	110,206,921	32–125	220,910	8909
Couroupita guianensis Aubl.	Illumina	61,381,502	59,277,816	36–125	52,314	9360
Eschweilera coriacea (DC.) S.A.Mori (1)	Ion Torrent	38,990,585	36,722,174	25–368	153,166	94
Eschweilera coriacea (DC.) S.A.Mori (2)	Ion Torrent	36,080,703	33,462,720	25–367	100,484	82
Eschweilera sagotiana Miers	Ion Torrent	38,441,045	36,467,981	25–367	160,630	127
Grias cauliflora L.	Illumina	120,184,636	115,758,292	26–125	172,405	9550
Gustavia augusta L.	Illumina	113,137,266	110,836,939	29–125	170,899	9679
Gustavia superba (Kunth) O.Berg (1)	Illumina	116,362,744	108,776,887	31–125	195,075	9746
Gustavia superba (Kunth) O.Berg (2)	Illumina	58,803,712	57,301,317	33–125	201,150	9607
Lecythis congestiflora Benoist	Ion Torrent	40,352,907	37,973,213	25–369	164,809	107
Lecythis persistens Sagot	Ion Torrent	35,631,041	33,490,636	25–367	124,976	86
Napoleonaea imperialis P.Beauv.	Illumina	61,324,904	59,312,901	30–125	30,942	Outgroup

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 et al., 2018), there is evidence that custom *de novo* 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 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 skimming 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/oscarvargash/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/oscarvargash/goldfinder.

2.1. Trancriptomes

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, Lihuania), 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* (Zhbannikov 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 Lecythidoideae, the New World Lecythidaceae (Mori et al., 2017): *Barringtonia racemosa* (outgroup), *Eschweilera sagotiana, 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 Lecythidaceae 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 (*Vaccinium macrocarpon* Aiton NC_023338.1), or ribosomal DNA (*Eschweilera congestiflora* JN222324.1, JN222317.1) using *bbmap.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 *bbwrap.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 (Mirarab 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 *phyparts* (Smith et al., 2015) to calculate the amount of conflict inside the nuclear markers by comparing the nuclear gene trees against the multilocus ML tree. We also calculated the number of concordant/conflicting nuclear makers against the plastome topology. *phyparts* results were visualized with *phypartspiecharts.py* (available from https://github.com/mossmatters/MJPythonNotebooks).

3. Results

We assembled *de novo* 12 transcriptomes for 10 Lecythidaceae 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 results.

MarkerMiner flagged a total of 1528 markers. We constructed a subset with *GoldFinder* of 354 transcript-markers (corresponding to 1754 exons), each marker with an average size of 1692 bp. We successfully captured *in silico* the 354 markers for the 24 species with available genome skimming data. The concatenated dataset resulted in an aligned supermatrix of 758,015 sites (including indels) with an overall occupancy of 97% (percentage of the matrix with data presence).

The phylogenetic analyses performed on the complete nuclear

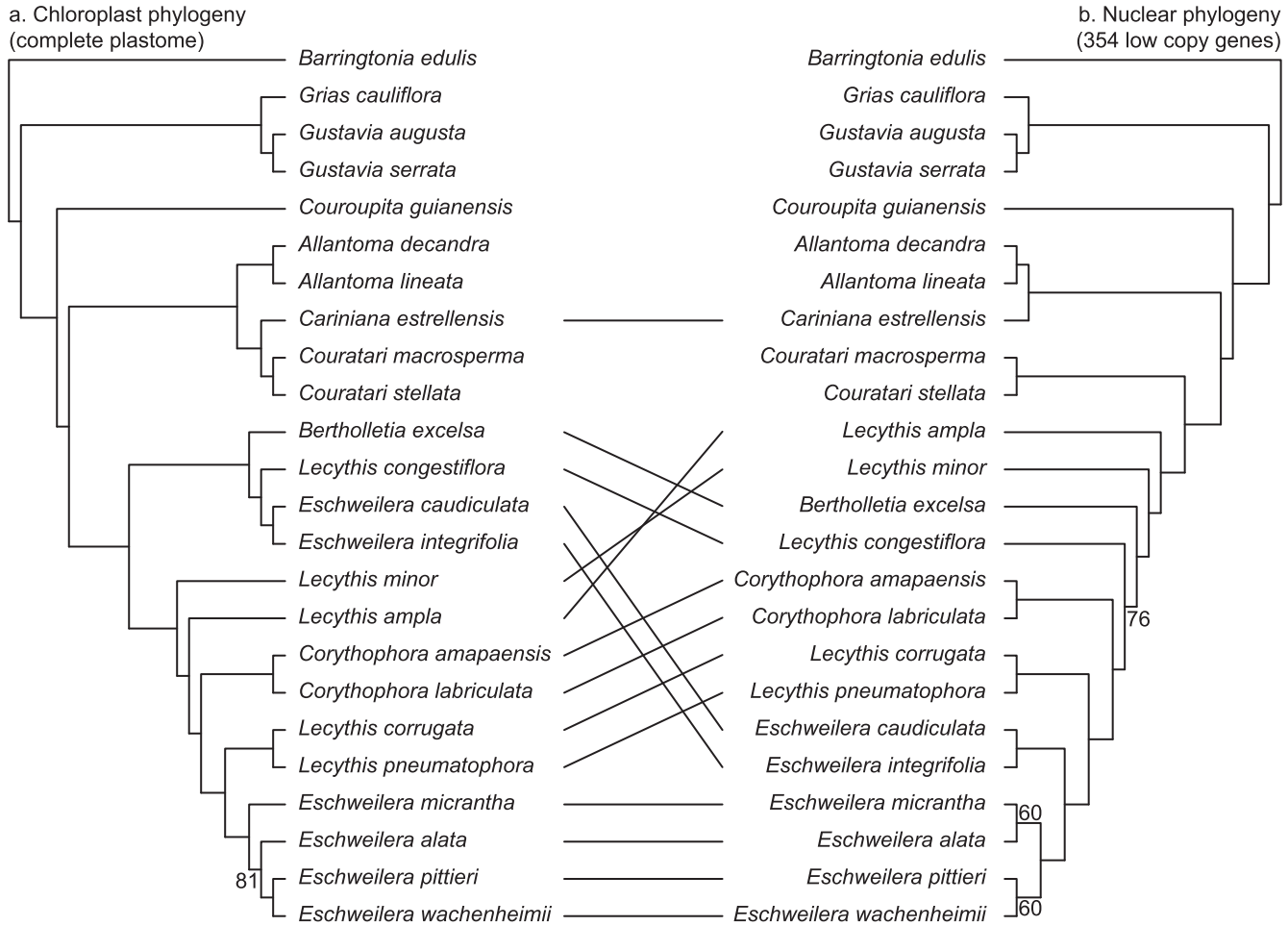

Fig. 1. Comparison between the maximum likelihood phylogenies derived from (a) the complete plastome alignment of Thomson et al. (2018) and (b) 354 nuclear markers. All nodes have a bootstrap support of 100 unless noted otherwise. Lines between taxa indicate a conflicting position between the two topologies.

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 obtained 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 recovered *Eschweilera* as monophyletic (the Integrifolia and the Parvifolia clades of Huang et 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. integrifolia*, *Corythophora amapaensis* + *C. labriculata*, and *Lecythis corrugata* + *L. pneumatophora*) being supported with more than half of the informative markers. Ten nodes presented considerable conflict (conflicting 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 techniques. The sequencing of multiple transcriptomes, along with the application 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-F, trnH-psbA*), ITS, and morphology; the plastid phylogeny of Thomson et al. (2018) largely agrees with that of Huang et al. (2015). Specifically, 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

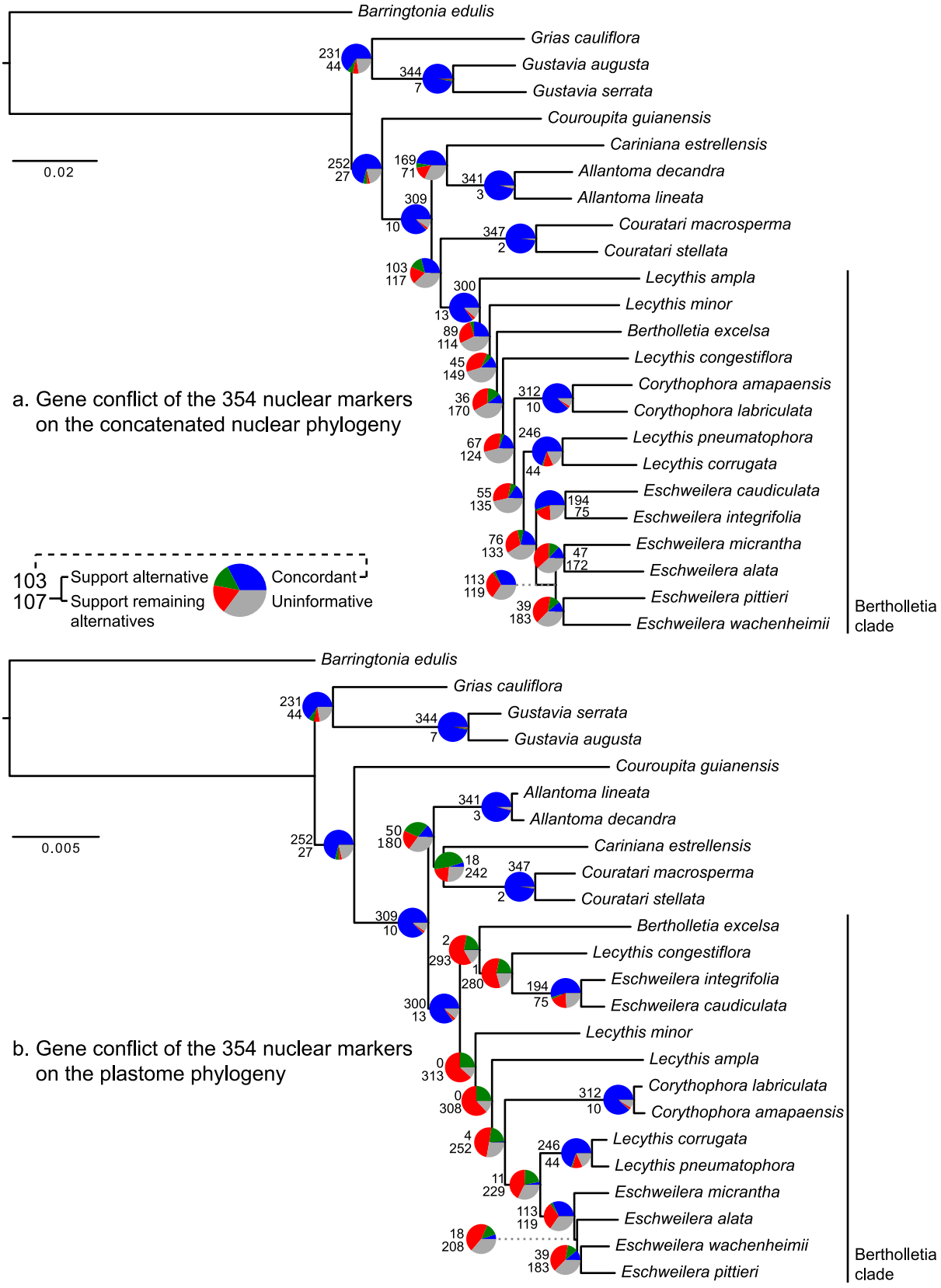

Fig. 2. Agreement and conflict of the 354 makers on the (a) maximum likelihood (ML) nuclear concatenated phylogeny and (b) plastome topology of Thomson et al. (2018). Pie charts indicate proportion of genes that agree (blue), support a main alternative topology (green), support remaining alternatives (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-Ezpeleta 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 phylograms (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 FESD 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-25-01 (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.ympev.2019.02.020.

References

Álvarez, I., Wendel, J.F., 2003. Ribosomal ITS sequences and plant phylogenetic inference. Mol. Phylogenet. Evol. 29, 417–434. https://doi.org/10.1016/S1055-7903(03)00208-2.

Birky, C.W., 1995. Uniparental inheritance of mitochondrial and chloroplast genes: mechanisms and evolution. Proc. Natl. Acad. Sci. U. S. A. 92, 11331–11338. https://doi.org/10.1073/pnas.92.25.11331.

Bruun-Lund, S., Clement, W.L., Kjellberg, F., Rønsted, N., 2017. First plastid phylogenomic study reveals potential cyto-nuclear discordance in the evolutionary history of *Ficus* L. (Moraceae). Mol. Phylogenet. Evol. 109, 93–104. https://doi.org/10.1016/j.ympev.2016.12.031.

Budenhagen, C., Lemmon, A.R., Lemmon, E.M., Bruhl, J., Cappa, J., Clement, W.L., Donoghue, M., Edwards, E.J., Hipp, A.L., Kortyna, M., Mitchell, N., Moore, A., Prychid, C.J., Segovia-Salcedo, M.C., Simmons, M.P., Soltis, P.S., Wanke, S., Mast, A., 2016. Anchored phylogenomics of angiosperms I: assessing the robustness of phylogenetic estimates. bioRxiv 086298. https://doi.org/10.1002/art.

Bushnell, B., 2015. BBMap short read aligner, and other bioinformatics tools. https://sourceforge.net/projects/bbmap/ (accessed Feb 2018).

Chamala, S., García, N., Godden, G.T., Krishnakumar, V., Jordon-Thaden, I.E., Smet, R. De, Barbazuk, W.B., Soltis, D.E., Soltis, P.S., 2015. MarkerMiner 1.0: a new application for phylogenetic marker development using Angiosperm transcriptomes. Appl. Plant Sci. 3, 1400115. https://doi.org/10.3732/apps.1400115.

Cowan, S., Devault, A., Eiserhardt, W.L., Epitawalage, N., Forest, F., Jan, T., 2018. A universal flowering plant probe set. bioRxiv 361618. https://doi.org/10.1101/361618.

Folk, R.A., Mandel, J.R., Freudenstein, J.V., 2016. Ancestral gene flow and parallel organellar genome capture result in extreme phylogenomic discord in a lineage of angiosperms. Syst. Biol. 3, 320–337. https://doi.org/10.1093/sysbio/syw083.

Gitzendanner, M.A., Soltis, P.S., Yi, T., Li, D.Z., Soltis, D.E., 2018. Plastome phylogenetics: 30 years of inferences into plant evolution. In: Chaw, S.-M., Jansen, R.K. (Eds.), Advances in Botanical Research. Elsevier Ltd., London, pp. 293–313. https://doi.org/10.1016/bs.abr.2017.11.016.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652. https://doi.org/10.1038/nbt.1883.

Huang, Y.Y., Mori, S.A., Kelly, L.M., 2015. Toward a phylogenetic-based generic classification of neotropical Lecythidaceae–I. Status of *Bertholletia*, *Corythophora, Eschweilera and Lecythis*. Phytotaxa 203, 85–121. https://doi.org/10.11646/phytotaxa.203.2.1.

Hughes, C.E., Eastwood, R.J., Donovan Bailey, C., Hughest, C.E., Eastwood, R.J., Bailey, C.D., 2006. From famine to feast? Selecting nuclear DNA sequence loci for plant species-level phylogeny reconstruction. Philos. Trans. R. Soc. B Biol. Sci. 361, 211–225. https://doi.org/10.1098/rstb.2005.1735.

Kadlec, M., Bellstedt, D.U., Le Maitre, N.C., Pirie, M.D., 2017. Targeted NGS for species level phylogenomics: "made to measure" or "one size fits all"? PeerJ 5, e3569. https://doi.org/10.7717/peerj.3569.

Le Provost, G., Herrera, R., Paiva, J.A., Chaumeil, P., Salin, F., Plomion, C., 2007. A micromethod for high throughput RNA extraction in forest trees. Biol. Res. 40, 291–297. https://doi.org/10.4067/S0716-97602007000400003.

Macmanes, M.D., 2014. On the optimal trimming of high-throughput mRNA sequence data. Front. Genet. 5, 1–7. https://doi.org/10.3389/fgene.2014.00013.

Maddison, W.P., 1997. Gene trees in species trees. Syst. Biol. 46, 523–536.

Mandel, J.R., Dikow, R.B., Funk, V.A., Masalia, R.R., Staton, S.E., Kozik, A., Michelmore, R.W., Rieseberg, L.H., Burke, J.M., 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: an example from the Compositae. Appl. Plant Sci. 2, 1300085. https://doi.org/10.3732/apps.1300085.

McVay, J.D., Hauser, D., Hipp, A.L., Manos, P.S., 2017. Phylogenomics reveals a complex evolutionary history of lobed-leaf white oaks in western North America. Genome 60, 733–742. https://doi.org/10.1139/gen-2016-0206.

Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T.S., Swenson, M., Warnow, T., 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30, 541–548. https://doi.org/10.1093/bioinformatics/btu462.

Morales-Briones, D.F., Liston, A., Tank, D.C., 2018. Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus *Lachemilla*

(Rosaceae). New Phytol. https://doi.org/10.1111/nph.15099.
Mori, S.A., Kiernan, E.A., Smith, N.P., Kelley, L.M., Huang, Y.-Y., Prance, G.T., Thiers, B., 2017. Observations on the phytogeography of the Lecythidaceae clade (Brazil nut family). Phytoneuron 30, 1–85.
Mori, S.A., Tsou, C.H., Wu, C.C., Cronholm, B., Anderberg, A.A., 2007. Evolution of Lecythidaceae with an emphasis on the circumscription of neotropical genera: information from combined ndhF and trnL-F sequence data. Am. J. Bot. 94, 289–301. https://doi.org/10.3732/ajb.94.3.289.
Paradis, E., Claude, J., Strimmer, K., 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20, 289–290. https://doi.org/10.1093/bioinformatics/btg412.
Pérez-Escobar, O.A., Balbuena, J.A., Gottschling, M., 2016. Rumbling orchids: how to assess divergent evolution between chloroplast endosymbionts and the nuclear host. Syst. Biol. 65, 51–65. https://doi.org/10.1093/sysbio/syv070.
R Core Team, 2018. R: A Language and Environment for Statistical Computing.
Rieseberg, L.H., Soltis, D.E., 1991. Phylogenetic consequences of cytoplasmic gene flow in plants. Evol. Trends Plants 5, 64–84. https://doi.org/10.1007/s00606-006-0485-y.
Rodríguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B.F., Philippe, H., 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol. 56, 389–399. https://doi.org/10.1080/10635150701397643.
Ruhlman, T.A., Jansen, R.K., 2014. The plastid genomes of flowering plants. In: Maliga, P. (Ed.), Chloroplast Biotechnology Methods and Protocols. Springer, New York, pp. 3–38.
Smith, S.A., Moore, M.J., Brown, J.W., Yang, Y., 2015. Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. BMC Evol. Biol. 15, 150. https://doi.org/10.1186/s12862-015-0423-0.
Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313. https://doi.org/10.1093/bioinformatics/btu033.
Sun, M., Soltis, D.E., Soltis, P.S., Zhu, X., Burleigh, J.G., Chen, Z., 2015. Deep phylogenetic incongruence in the angiosperm Rosidae clade. Mol. Phylogenet. Evol. 83, 156–166. https://doi.org/10.1016/j.ympev.2014.11.003.
ter Steege, H., Pitman, N.C.A., Phillips, O.L., Chave, J., Sabatier, D., Duque, A., Molino, J.-F., Prévost, M.-F., Spichiger, R., Castellanos, H., von Hildebrand, P., Vásquez, R., 2006. Continental-scale patterns of canopy tree composition and function across Amazonia. Nature 443, 444–447. https://doi.org/10.1038/nature05134.
Thomson, A.M., Vargas, O.M., Dick, C.W., 2018. Complete plastome sequences from Bertholletia excelsa and 23 related species yield informative markers for Lecythidaceae. Appl. Plant Sci. 6, e1151. https://doi.org/10.1002/aps3.1151.
Vargas, O.M., Ortiz, E.M., Simpson, B.B., 2017. Conflicting phylogenomic signals reveal a pattern of reticulate evolution in a recent high-Andean diversification (Asteraceae: Astereae: *Diplostephium*). New Phytol. 214, 1736–1750. https://doi.org/10.1111/nph.14530.
Weitemier, K., Straub, S.C.K., Cronn, R.C., Fishbein, M., Schmickl, R., McDonnell, A., Liston, A., 2014. Hyb-Seq: combining target enrichment and genome skimming for plant phylogenomics. Appl. Plant Sci. 2, 1400042. https://doi.org/10.3732/apps.1400042.
Wiens, J.J., Engstrom, T.N., Chippindale, P.T., 2006. Rapid diversification, incomplete isolation, and the "speciation clock" in North American salamanders (genus *Plethodon*): testing the hybrid swarm hypothesis of rapid radiation. Evolution 60, 2585–2603. https://doi.org/10.1111/j.0014-3820.2006.tb01892.x.
Yang, Y., Smith, S.A., 2014. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. Mol. Biol. Evol. 31, 3081–3092. https://doi.org/10.1093/molbev/msu245.
Zhang, C., Sayyari, E., Mirarab, S., 2017. ASTRAL-III: increased scalability and impacts of contracting low support branches. In: Meidanis, J., Nakhleh, L. (Eds.), Comparative Genomics. Springer International Publishing, Cham, pp. 53–75.
Zhang, N., Zeng, L., Shan, H., Ma, H., 2012. Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. New Phytol. 195, 923–937. https://doi.org/10.1111/j.1469-8137.2012.04212.x.
Zhbannikov, I.Y., Hunter, S.S., Foster, J.A., Settles, M.L., 2017. SeqyClean: a pipeline for high-throughput sequence data preprocessing. In: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, ACM-BCB '17. ACM, New York, pp. 407–416.
Zimmer, E.A., Wen, J., 2015. Using nuclear gene data for plant phylogenetics: progress and prospects II. Next-gen approaches. J. Syst. Evol. 53, 371–379. https://doi.org/10.1111/jse.12174.