Comparative analysis of 24 chloroplast genomes yields highly informative genetic markers for the Brazil nut family (Lecythidaceae)

Ashley M. Thomson*[1, 2], Oscar M. Vargas*[1], Christopher W. Dick[1,3]

1) Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109
2) Faculty of Natural Resources Management. Lakehead University, Thunder Bay, Ontario, Canada, P7B 5E1
3) Smithsonian Tropical Research Institute, Republic of Panama

* These authors contributed equally to this work

Abstract

The tropical tree family Lecythidaceae (order Ericales) has enormous ecological and economic importance in the Amazon basin. Lecythidaceae species can be difficult to identify without molecular data, however, and phylogenetic relationships within and among the most diverse Amazonian genera, *Lecythis* and *Eschweilera*, are unresolved. In order to develop genetic markers for ecological and evolutionary studies in the family, we used genome skimming to assemble *de novo* the full plastome of the Brazil nut tree (*Bertholletia excelsa*) and 23 other Lecythidaceae species. Indices of nucleotide diversity and phylogenetic signal were used to identify regions suitable for genetic marker development. The *B. excelsa* plastome contained 160,472 bp and was arranged in a quadripartite structure consisting of a large single copy region (85,830 bp), a small single copy region (16,670 bp), and two inverted repeats (of 27,481 bp each). The coding region *ycf1* and the spacer *rpl16-rps3* outperformed plastid DNA markers previously used for barcoding and phylogenetics. We identified 456 cpSSRs in the *B. excelsa* plastome, from which we developed 130 primer pairs. Used in a phylogenetic analysis, the matrix of 24 plastomes showed with 100% bootstrap support that *Lecythis* and *Eschweilera* are

polyphyletic, indicating the need for more detailed systematics studies of these two important Amazonian tree genera.

Keywords

DNA Barcoding, genetic markers, Amazon, tropical trees, Lecythidaceae, plastome.

Introduction

Lecythidaceae (*sensu latu*) is a pantropical family of trees that contains three subfamilies: Foetidioideae, which is restricted to Madagascar; Planchonioideae, found in the tropical forests of Asia and Africa; and the Neotropical clade Lecythidoideae (Mori *et al.* 2007). The Lecythidoideae clade contains ca. 234 (Mori 2017) of the ca. 278 known species in the broader family (Mori *et al.* 2007; Huang *et al.* 2015; Mori *et al.* 2017; Mori 2017). Neotropical Lecythidaceae are understory, canopy, or emergent trees with distinctive floral morphology and woody fruit capsules. It is the third most abundant family of trees in the Amazon forest, following Fabaceae and Sapotaceae (ter Steege *et al.* 2013). The most species-rich genus, *Eschweilera* with ca. 99 species (Mori 2017), is the most abundant tree genus in the Amazon basin, as quantified in forest inventory plots scattered across the basin (ter Steege *et al.* 2013); and *Eschweilera coriacea* (DC.) S.A.Mori is the most common tree species in much of Amazonia (ter Steege *et al.* 2013). Among its species are the iconic Brazil nut tree, *Bertholletia excelsa* Bonpl.; the oldest documented angiosperm tree, *Cariniana micrantha* Ducke (dated at >1400 years old in Manaus, Brazil; Chambers *et al.* 1998); the cauliflorous cannonball tree commonly grown in botanical gardens, *Couroupita guianensis* Aubl.; and important timber species (e.g. *Carinaria legalis* (Mart.) Kuntze). Lecythidaceae provide important ecological services such as carbon sequestration and food resource for pollinators (bats and large bees) and seed dispersers (monkeys and agouties) (Prance & Mori 1979, Mori & Prance 1990).

Species-level identification of Lecythidaceae and a robust phylogenetic hypothesis are essential for evolutionary and ecological research on Amazon tree diversity. However, despite their ease of identification at the family level, species-level identification of many Lecythidaceae (especially *Eschweilera*) is notoriously difficult when based on sterile (i.e. without fruit or floral material) herbarium specimens, and flowering specimens are often available only at multi-year intervals (Mori & Prance 1987). As a complement to other approaches, DNA barcoding (Dick & Kress 2009; Dexter *et al.* 2010) may be useful for the identification of species and clades of Lecythidaceae.

A combination of two protein-coding plastid regions (*rbcL* and *matK*) have been proposed as core plant DNA barcodes (Hollingsworth *et al.* 2009), although other coding and non-coding plastome regions (*rpoC1, rpoB, ycf5, trnL, psbA-trnH*) and the internal transcribed spacer (ITS) of nuclear ribosomal genes, have been recommended as supplemental barcodes for vascular plant identification (Kress et al. 2005; Lahaye *et al.* 2008; Li *et al.* 2011). However, an evaluation of these markers on Lecythidaceae in French Guiana (Gonzales *et al.* 2009) showed poor performance for species identification. Furthermore, the use of traditional markers (plastid *ndhF, trnL-F,* and *trnH-psbA*, and nuclear ITS) for phylogenetic analysis has produced weakly supported trees (Mori *et al.* 2007; Huang *et al.* 2015) indicating a need to develop more informative markers and/or increase molecular sampling.

The advent of high-throughput sequencing provides opportunities to obtain more informative DNA markers through the comparative analysis of full genomes. In this study, we aimed to (1) assemble, annotate, and characterize the first complete plastome sequence of Lecythidaceae, the iconic Brazil nut tree *Bertholletia excelsa*; (2) obtain a robust backbone phylogeny for the Neotropical clade using newly-assembled draft plastome sequences for an additional 23 species; and (3) develop a novel set of molecular markers for DNA barcoding, population genetics, phylogeography, and phylogenetic inference.

Methods

Plant material and DNA library preparation

We performed genomic skimming on 24 Lecythidaceae species, including 23 Lecythidoideae and one outgroup species (*Barringtonia edulis* Seem.) from the Planchonioideae. The sampling included all 10 Lecythidoideae genera (S1 Table). Silica-dried leaf tissue from herbarium-vouchered collections was collected by Scott Mori and colleagues and loaned by the New York Botanical Garden. Total genomic DNA was extracted from 20 milligrams of dried leaf tissue using the NucleoSpin Plant II extraction kit (Machery-Nagel, Bethlehem, PA, USA) with SDS lysis buffer. Prior to DNA library preparation, 5 micrograms of total DNA were fragmented using a Covaris S-series sonicator (Covaris, Inc. Woburn, MA, USA) following the manufacturer’s protocol, to obtain ca. 300 bp insert-sizes. We prepared the sequencing library using the NEBNext DNA library Prep Master Mix and Multiplex Oligos for Illumina Sets (New England BioLabs Inc. Ipswich, MA, USA) according to the manufacturer’s protocol. Size-selection was carried prior to PCR using Pippin Prep (Sage Science, Beverly, MA, USA). Molecular mass of the finished paired-end library was quantified using an Agilent 2100 Bioanalyzer (Agilent Technologies Inc., Santa Clara, CA, USA) and by qPCR using an ABI PRISM 7900HT (ThermoFisher Scientific, Waltham, MA, USA) at the University of Michigan

DNA Sequencing Core (Ann Arbor, MI, USA). We sequenced the libraries on one lane of the Illumina HiSeq 2000 (Illumina Inc., San Diego, CA, USA) with a paired-read length of 100bp.

Plastome assembly

Illumina adaptors and barcodes were excised from raw reads using Cutadapt v.1.4.2 (Martin 2011). Reads were then quality-filtered using Prinseq v. 0.20.4 (Schmieder & Edwards 2011), which trimmed 5' and 3' sequence ends with Phred quality score < 20 and removed all trimmed sequences <50 bp in length, with >5% ambiguous bases, or with mean Phred quality score <20. A combination of *de novo* and reference-guided approaches were used to assemble the plastomes. First, chloroplast reads were separated from the raw read pool by Blast-searching all raw reads against a database consisting of all complete angiosperm plastome sequences available on GenBank (accessed in 2014). Any aligned reads with an e-value $<1^{-5}$ were retained for subsequent analysis. The filtered chloroplast reads were *de novo* assembled using Velvet v.7.0.4 (Zerbino & Birney 2008) with kmer values of 71, 81, and 91 using a low-coverage cutoff of 5 and minimum contig length of 300. The assembled contigs were then mapped to a reference genome (see below) using Geneious v. R8 (Kearse *et al*. 2012) to determine their order and direction using the reference-guided assembly tool with medium sensitivity and iterative fine-tuning options. Finally, raw reads were iteratively mapped onto the draft genome assembly to extend contigs and fill gaps using low-sensitivity reference-guided assembly in Geneious. We first assembled the draft genome of *Bertholletia excelsa* for which only one contig was obtained; the plasomes of the remaining 23 species were assembled subsequently using the plastome of *B. excelsa* as a reference. The *B. excelsa* plastome was annotated using the DOGMA (Wyman *et al.* 2004) with the default settings for chloroplast genomes. Codon start and stop positions were determined using the open reading frame finder in Geneious and by comparison with the plastome sequence of *Camellia sinensis* var. *pubilimba* Hung T. Chang (Genbank ID: KJ806280). A circular representation of the *B. excelsa* plastome was made using OGDraw V1.2 (Lohse *et al.* 2007). The complete annotated plastome of *B. excelsa* and the draft plastomes of the remaining 23 Lecythidaceae species sampled were deposited in GenBank (Table S1).

Identification of molecular markers

Chloroplast simple sequence repeats (cpSSRs) in *B. excelsa* were identified using the Phobos Tandem Repeat Finder v.3.3.12 (Mayer 2010) by searching for uninterrupted repeats of nucleotide units of 1 to 6 bp in length, with thresholds of ≥ 7 mononucleotide repeats, ≥4 dinucleotide repeats, and ≥3 tri-, tetra-, penta-, and hexanucleotide repeats. We developed primers to amplify the cpSSRs using Primer 3 v.2.3.4 (Untergasser *et al.* 2012) with the default options and setting the PCR product size range between 100 and 250 bp.

The 24 plastomes were aligned with MAFFT v.7.017 (Katoh *et al.*, 2002) and scanned for regions of high nucleotide diversity, π (Nei 1987), using a sliding window analysis implemented in DNAsp v.5.10.1 (Librado & Rozas 2009) with a window and a step size of 600 bp. Levels of nucleotide diversity were plotted using R (R core development group), and windows with values over the 95th percentile were considered of high π.

Because regions with high π do not necessarily have high phylogenetic signal (e.g. unalignable hypervariable regions), to identify phylogenetically influential regions we employed a log-likelihood approach modified from Walker *et al.* (2017). First, we inferred a phylogenetic tree with the plastome alignment (including only one inverted repeat) by performing 100 independent maximum likelihood (ML) searches using a GTRCAT model with RAxML v. 8.2.9 (Stamatakis, 2014). Those searches resulted in the same topology that was subsequently annotated with the summary from 100 bootstraps using “sumtrees.py" v.4.10 (Sukumaran & Holder 2010). Then, we calculated the site-specific log-likelihood in the alignment over the plastome phylogeny and calculated their differences site-wise to the averaged log-likelihood per site of 1000 randomly permuted trees (tips were randomly shuffled). Log-likelihood scores were calculated with RAxML. The site-wise log-likelihood differences (LD) were calculated using 600 bp non-overlapping windows with a custom R script (see below). We interpreted greater (LD) as an indication of greater phylogenetic signal, and windows with LD above the 95th percentile were considered to have exceptional phylogenetic signal.

Primers flanking the top ten regions with high π were designed using Primer 3 with default program options. We employed a maximum product size of 1300 bp because lower cutoffs values (e.g. 600 bp) made the primer design extremely challenging due to the lack of conserved regions. Primers were designed to amplify across all 23 Neotropical species without the use of degenerate bases. However, primers with a small number of degenerate bases were permitted for some regions where primer development otherwise would not have been possible due to high sequence variability in the priming sites. We investigated the potential of our markers to produce robust phylogenies by calculating individual gene trees in RAxML v.8.2.9 in an ML search with 100 rapid bootstraps (option “-f a”) using the GTRCAT model. To evaluate the number of markers needed to obtain a resolved tree with an average of ~90 bootstrap support (BS), we first concatenated the two markers with the highest π and inferred a tree; subsequently we added another marker to the matrix based on the ranking obtained from the π score. We iterated this process until we obtained a matrix with each of the 10 markers developed. For every tree obtained, we calculated its average BS and its Robinson-Foulds distance (RF) (Robinson and Foulds 1981) from the plastome phylogeny, using a custom R script employing the packages

APE (Paradis *et al*. 2004) and Phangor (Schliep 2011). Scripts and alignments used for this study can be found at https://bitbucket.org/oscarvargash/lecythidaceae_plastomes.

Results

Lecythidaceae plastome features

The sequenced plastome of *Bertholletia excelsa* contained 160,472 base pairs and 117 genes, of which 4 were rRNAs and 31 were tRNAs (Fig. 1). The arrangement of the *B. excelsa* plastome had a typical angiosperm quadripartite structure with a single copy region of 85,830 bp, a small single copy region of 16,670 bp, and two inverted of repeats of 27,481 bp each. Relative to *Camellia sinensis* var. *pubilimba*, the closest relative of Lecythidaceae with a sequenced plastome, we find no gene gain/losses in *B. excelsa*; the only main structural difference is that the inverted repeat of *B. excelsa* contained the genes *trnH-GUG*, *rps3*, *rpl22*, and *rps19* while in *C. sinensis* var. *pubilimba* these regions were located in the large single copy region. In addition to *B. excelsa*, the plastome of *Eschweilera alata* A.C.Sm. was also completely assembled; the coverage for the remaining plastomes ranged between 85% and 99.60% (S1 Table). From the non-*Bertholletia* plastomes, *Barringtonia edulis and Corythophora amapaensis* Pires ex S.A.Mori & Prance seemed to have lost *ycf15* and *psaA*, respectively.

Identification of molecular markers

Within the plastome of *Bertholletia excelsa* we found 456 cpSSRs (Table 1). We designed 130 primers pairs for cpSSR amplification (S2 Table) for regions outside of coding regions with an acceptable product length, annealing temperature, and GC content. π for nine 600 bp windows exceeded the 95th percentile (Fig. 2A, Table 2). Similarly, 13 windows were over the 95th percentile for LD (Fig. 2B, Table 3) indicating high phylogenetic signal. While most of the informative windows were located in non-coding regions, two consecutive regions were positioned in the *ycf1* gene. Six windows contained both high π and LD. As expected, high π and greater LD largely agreed. Based on the rank of the windows obtained for nucleotide diversity we developed primers for the following regions (ordered from high to low nucleotide diversity): *ycf1, rpl16-rps3, psbM-trnD, ccsA-ndhD, trnG-psaB, petD-rpoA, psbZ-trnfM, trnE-trnT*, and *trnT-psbD* (Table 3).

Phylogenetics of the plastomes and the developed markers

The ML analysis of the plastome alignment for the Lecythidaceae (145,487 sites) yielded a fully resolved phylogeny with high BS for all clades (Fig. 3). Of the genera in which the sampling included multiples species, *Eschweilera* and *Lecythis* were polyphyletic, while *Allantoma*, *Corythophora*, *Couratari*, and *Gustavia* were monophyletic (*Bertholletia* is monospecific, and only one species of *Couroupita*, *Cariniana, and Grias* and were included in the analysis). The trees obtained from individual markers with high nucleotide diversity had an average BS of 73 throughout their nodes, while for the trees obtained from two or more concatenated regions had an average BS of 89 (Fig. 4A). None of the gene trees, single or combined, recovered the topology obtained using the complete plastome matrix (none of the gene trees obtained a RF = 0, Fig. 4B). In general, matrices with concatenated markers (mean RF = 6) outperformed single markers (mean RF = 13.8).

Discussion

Genetic markers from the Lecythidaceae plastome

We are publishing the first full plastome for Lecythidaceae, including high-depth coverage of the Brazil nut tree, and 23 draft genomes representing all Lecythoideae genera and a

Paleotropical outgroup taxon. We found no significant gene losses or major rearrangements when the plastome of *Bertholletia excelsa* was compared with that of *Camellia sinensis* var. *pubilimba*, a closely related plastome (Theaceae). However, there are likely to be some gene losses within the broader Lecythidoideae and Lecythidaceae, as indicated by the loss of *ycf15* in *Barringtonia edulis* and *psbA* in *Corythophora amapaensis*.

We inferred a robust backbone phylogeny for Lecythoideae using the 24 aligned plastomes. All nodes in our topology had 100% bootstrap support with the exception of a node that connects three closely related species of *Eschweilera*. The topology agreed with previous but weakly supported (<50% BS) Lecythidaceae phylogenies, based on chloroplast and nuclear ITS (internal transcribed spacer) sequences (Mori *et al.* 2007, Huang *et al.* 2015), indicating that *Eschweilera* and *Lecythis* are polyphyletic. Although the polyphyly of these two genera is well supported with all available data, some inferred species-level relationships may change with increased taxonomic sampling and the inclusion of nuclear genomic data.

We measured nucleotide diversity (π) and a proxy for phylogenetic signal using a log-likelihood approach (LD) modified from Walker *et al.* (2017). These calculations helped us to evaluate the performance of specific chloroplast regions as potential phylogenetic markers. The core plant DNA barcodes *matK* and *rbcL* did not exhibit high π or LD in our analysis. Of the secondary plant DNA barcodes mentioned in the literature (*rpoC1*, *rpoB, ycf5*, *trnL, psbA-trnH;* Kress et al, 2005, Lahaye *et al.* 2008, Hollingsworth 2009, Li *et al.* 2011) only *psbA-trnH* showed high LD (Table 3) although it did not exhibit exceptionally high values of π. In contrast, the regions *ycf1, rpl16-rps3, psbM-trnD, ccsA-ndhD, trnG-psaB, petD-rpoA*, *psbZ-trnfM*, *trnE-trnT*, and *trnT-psbD* displayed the highest values of π and LD and therefore outperformed all of the previously proposed plant DNA barcodes.

Phylogenetic trees calculated from concatenated marker sets (based on rank) outperformed single regions in terms of support (BS) and accuracy (RF) (Fig. 4). In fact, tree topologies using single markers deviated relatively highly from the complete plastome tree (mean RF= 13.8). The best performing concatenated matrix contained all 10 regions for which we developed primers. However, the combination of *ycf1* and *rpl16−rps3* produced an average BS ~90 (Fig. 4A) with reasonable accuracy (RF = 4, Fig. 4B); we conclude that these two regions, amplified in three PCRs (Table 3), are promising markers for DNA barcoding, phylogeny, and phylogeography in Lecythidaceae. Although barcoding efficiency in species-rich clades (i.e. *Eschweilera/Lecythis)* might decline with the addition of more samples, *ycf1* and *rpl16−rps3* effectively distinguished between three closely-related species within the *E. parvifolia* clade (see branch lengths in Fig. S1), suggesting that these markers might effectively distinguish between many other closely related species. Our results and conclusions agree with those of Dong *et al.* (2015) who proposed *ycf1* as a universal barcode for land plants.

The 130 cpSSR markers developed for noncoding portions of the *B. excelsa* plastome provide a useful resource for population genetic studies. Because of their fast stepwise mutation rate relative to SNPs, cpSSRs can also be used for finer grain phylogeographic analyses (e.g. Lemes *et al.* 2010; Twyford *et al.* 2013). This may be especially useful for species that exhibit little geographic structuring across parts of their ranges. Because they are maternally transmitted and can be variable within populations, the cpSSRs may also be used to track dispersal of seeds and seedlings relative to the maternal source trees.

Because of their high level of polymorphism and phylogenetic signal content, the cpDNA markers presented here should be useful for phylogeographic studies of widespread Lecythidaceae species. For example, *Couratari guianensis* Aubl. and *Eschweilera coriacea* range from the Amazon basin into Central America, and other species range broadly across the Amazon basin, the Guiana Shield, and the Atlantic forests.

Barcoding of tropical trees

DNA barcoding of tropical trees has been useful for several applications, including community phylogenetic analyses (Kress *et al.* 2009), inferring the species identity of the gut content (diet) of herbivores (García-Robledo *et al.* 2013), and for species identification of seedlings (Gonzalez *et al.* 2009). The power of DNA barcodes to discriminate among species should be high if the studied species are distantly related; for example, Kress *et al.* (2009) were able to discriminate 281 of 296 tree and shrub species from Barro Colorado Island (BCI) using standard DNA barcodes, but they were not able to discriminate among some congeneric species in the species-rich genera *Inga* (Fabaceae), *Ficus* (Moraceae), and *Piper* (Piperaceae). Gonzales et al (2009) encountered similar challenges with *Eschweilera* species in their study of trees and seedlings in Paracou, French Guiana. The latter study tested a wide range of putative DNA barcode regions (*rbcLa*, *rpoC1*, *rpoB*, *matK*, *ycf5*, *trnL*, *psbA-trnH*, ITS), however, they did not include the markers presented in this paper.

Limitations of plastome markers for phylogeny and species ID

The newly-identified plastome markers revealed by our study, while promising, are not free of limitations. First, plastome-based phylogenies should be interpreted with caution, as they can disagree with nuclear markers and species trees due to introgression and or lineage sorting issues (Rieseberg & Soltis 1997; Sun *et al.* 2015; Vargas *et al.* 2017). Second, hybridization and incomplete lineage sorting would also affect the performance of plastome barcodes for species identification and therefore ecological studies derived from such. For example, cpDNA haplotypes of *Nothofagus*, *Eucalyptus*, *Quercus*, *Betula*, and *Acer* were more strongly determined by geographic location than by species-identity due to the occurrence of localized introgression within these groups (Petit *et al.* 1993; Palme *et al.* 2004; Saeki *et al.* 2011; Premoli *et al.* 2012; Nevill *et al.* 2014; Thomson *et al.* 2015). The occurrence of haplotype sharing in closely-related Lecythidacae species has, to date, not been examined at a large scale and it is therefore not possible to conclude to what extent introgression or incomplete lineage sorting might affect this group. However, unique plastome sequences were retrieved for each of the 24 species sequenced in this study, including closely related *Eschweilera* and *Lecythis*, suggesting that incomplete lineage sorting was not an issue at the scale of our analysis. We suggest that future studies utilizing cpDNA barcodes for Neotropical Lecythidaceae examine species from several shared geographic localities to examine to what extent haplotypes tend to be shared among species at the same localities. Alternatively, nuclear barcode markers such as ITS could be used to examine incongruence of plastome versus nuclear markers to identify cases where introgression might have occurred.

Acknowledgements

National Science Foundation (DEB 1240869 and FESD Type I 1338694 to CD) and the University of Michigan (Associate Professor Award to CD) provided financial support for this work. We would like to thank Scott Mori, Gregory Stull, Caroline Parins-Fukuchi, and Joseph Walker for comments; and Scott Mori and the New York Botanical Garden for providing access to curated DNA samples of Lecythidaceae.

Literature cited:

Chambers, J. Q., Higuchi, N., & Schimel, J. P. (1998). Ancient trees in Amazonia. *Nature*, *391*(6663), 135–136. doi:10.1038/34325

Dexter, K. G., Pennington, T. D., & Cunningham, C. W. (2010). Using DNA to assess errors in tropical tree identifications: How often are ecologists wrong and when does it matter? *Ecological Monographs*, *80*(2), 267–286.

Dick, C. W., & Kress, W. J. (2009). Dissecting tropical plant diversity with forest plots and a molecular toolkit. *BioScience*, *59*(9), 745–755. doi:10.1525/bio.2009.59.9.6

Dong, W., Xu, C., Li, C., Sun, J., Zuo, Y., Shi, S., … Zhou, S. (2015). ycf1, the most promising plastid DNA barcode of land plants. *Scientific Reports*, *5*, 8348. doi:10.1038/srep08348

García-Robledo, C., Erickson, D. L., Staines, C. L., Erwin, T. L., & Kress, W. J. (2013). Tropical plant-herbivore networks: reconstructing species interactions using DNA barcodes. *PLoS ONE*, *8*(1), e52967. doi:10.1371/journal.pone.0052967

Gonzalez, M. A., Baraloto, C., Engel, J., Mori, S. A., Pétronelli, P., Riéra, B., … Chave, J. (2009). Identification of Amazonian trees with DNA barcodes. *PLoS ONE*, *4*(10), e7483. doi:10.1371/journal.pone.0007483

Hollingsworth, P. M., Forrest, L. L., Spouge, J. L., Hajibabaei, M., Ratnasingham, S., van der Bank, M., … Little, D. P. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(31), 12794–129797. doi:10.1073/pnas.0905845106

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*(2), 85–121. doi:10.11646/phytotaxa.203.2.1

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. doi:10.1093/nar/gkf436

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., … Drummond, A. (2012). Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*(12), 1647–1649. doi:10.1093/bioinformatics/bts199

Kress, W. J., Erickson, D. L., Jones, F. A., Swenson, N. G., Perez, R., Sanjur, O., & Bermingham, E. (2009). Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(44), 18621–6. doi:10.1073/pnas.0909820106

Kress, W. J., Wurdack, K. J., Zimmer, E. A., Weigt, L. A., & Janzen, D. H. (2005). Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(23), 8369–8374. doi:10.1073/pnas.0503123102

Lahaye, R., van der Bank, M., Bogarin, D., Warner, J., Pupulin, F., Gigot, G., … Savolainen, V. (2008). DNA barcoding the floras of biodiversity hotspots. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(8), 2923–2928. doi:10.1073/pnas.0709936105

Lemes, M. R., Dick, C. W., Navarro, C., Lowe, A. J., Cavers, S., & Gribel, R. (2010). Chloroplast DNA microsatellites reveal contrasting phylogeographic structure in mahogany (Swietenia macrophylla King, Meliaceae) from Amazonia and Central America. *Tropical Plant Biology*, *3*(1), 40–49. doi:10.1007/s12042-010-9042-5

Li, D.-Z., Gao, L.-M., Li, H.-T., Wang, H., Ge, X.-J., Liu, J.-Q., … Duan, G.-W. (2011). Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(49), 19641–19646. doi:10.1073/pnas.1104551108

Librado, P., & Rozas, J. (2009). DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, *25*(11), 1451–1452. doi:10.1093/bioinformatics/btp187

Lohse, M., Drechsel, O., & Bock, R. (2007). OrganellarGenomeDRAW (OGDRAW): A tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Current Genetics, 52(5–6), 267–274. doi:10.1007/s00294-007-0161-y

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*(1), 10–12. doi:10.14806/ej.17.1.200

Mayer, C. (2010) Phobos. Available from: http://www.rub.de/spezzoo/cm/cm_phobos.htm

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., & Prance, G. T. (1987). A Guide to Collecting Lecythidaceae. *Annals of the Missouri Botanical Garden*, *74*(2), 321–330.

Mori, S. A., & Prance, G. T. (1990). Lecythidaceae. Part II. The zygomorphic-flowered New World genera (*Couroupita*, *Corythophora*, *Bertholletia*, *Couratari*, *Eschweilera*, & *Lecythis*. *Flora Neotropica*, *21*, 1–376.

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. *American Journal of Botany*, *94*(3), 289–301. doi:10.3732/ajb.94.3.289

Mori, S.A. (2017). The Lecythidaceae pages. http://sweetgum.nybg.org/science/projects/lp/

Nei, M. (1987). *Molecular Evolutionary Genetics*. New York: Columbia University Press.

Nevill, P. G., Després, T., Bayly, M. J., Bossinger, G., & Ades, P. K. (2014). Shared phylogeographic patterns and widespread chloroplast haplotype sharing in Eucalyptus species with different ecological tolerances. *Tree Genetics and Genomes*, *10*(4), 1079–1092. doi:10.1007/s11295-014-0744-y

Palme, A. E., Su, Q., Palsson, S., & Lascoux, M. (2004). Extensive sharing of chloroplast haplotypes among European birches indicates hybridization among Betula pendula, B. pubescens and B. nana. *Molecular Ecology*, *13*(1), 167–178. doi:10.1046/j.1365-294X.2003.02034.x

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, *20*(2), 289–290. doi:10.1093/bioinformatics/btg412

Petit, R. J., Kremer, A., & Wagner, D. B. (1993). Geographic structure of chloroplast DNA polymorphisms in European oaks. *Theoretical and Applied Genetics: International Journal of Plant Breeding Research*, *87*(1), 122–128. doi:10.1007/BF00223755

Prance, G. T., & Mori, S. A. (1979). Lecythidaceae–Part I. The actinomorphic-flowered New World Lecythidaceae (*Asteranthos, Gustavia, Grias, Allantoma,* & *Carinaria*). *Flora Neotropica Monograph*, *21*, 1–270.

Premoli, A. C., Mathiasen, P., Cristina Acosta, M., & Ramos, V. A. (2012). Phylogeographically concordant chloroplast DNA divergence in sympatric Nothofagus s.s. How deep can it be? *New Phytologist*, *193*(1), 261–275. doi:10.1111/j.1469-8137.2011.03861.x

R Core Team. (2017) R: A Language and Environment for Statistical Computing. Available from: https://www.r-project.org/

Rieseberg, L. H., & Soltis, D. E. (1991). Phylogenetic consequences of cytoplasmic gene flow in plants. *Evolutionary Trends in Plants*, *5*(1), 64–84. doi:10.1007/s00606-006-0485-y

Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, *53*(1–2), 131–147. doi:10.1016/0025-5564(81)90043-2

Saeki, I., Dick, C. W., Barnes, B. V., & Murakami, N. (2011). Comparative phylogeography of red maple (*Acer rubrum* L.) and silver maple (*Acer saccharinum* L.): Impacts of habitat specialization, hybridization and glacial history. *Journal of Biogeography*, *38*(5), 992–1005. doi:10.1111/j.1365-2699.2010.02462.x

Schliep, K. P. (2011). phangorn: Phylogenetic analysis in R. *Bioinformatics*, *27*(4), 592–593. doi:10.1093/bioinformatics/btq706

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, *27*(6), 863–864. doi:10.1093/bioinformatics/btr026

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. doi:10.1093/bioinformatics/btu033

Sukumaran, J., & Holder, M. T. (2010). DendroPy: A Python library for phylogenetic computing. *Bioinformatics*, *26*(12), 1569–1571. doi:10.1093/bioinformatics/btq228

Sun, M., Soltis, D. E., Soltis, P. S., Zhu, X., Burleigh, J. G., & Chen, Z. (2015). Deep phylogenetic incongruence in the angiosperm Rosidae clade. *Molecular Phylogenetics and Evolution*, *83*, 156–166. doi:10.1016/j.ympev.2014.11.003

ter Steege, H., Pitman, N. C. A., Sabatier, D., Baraloto, C., Salomão, R. P., Guevara, J. E., … Silman, M. R. (2013). Hyperdominance in the Amazonian tree flora. *Science*, *342*(6156), 325–342. doi:10.1126/science.1243092

Thomson, A. M., Dick, C. W., & Dayanandan, S. (2015). A similar phylogeographical structure among sympatric North American birches (Betula) is better explained by introgression than by shared biogeographical history. *Journal of Biogeography*, *42*(2), 339–350. doi:10.1111/jbi.12394

Twyford, A. D., Kidner, C. a, Harrison, N., & Ennos, R. a. (2013). Population history and seed dispersal in widespread Central American Begonia species (Begoniaceae) inferred from plastome-derived microsatellite markers. *Botanical Journal of the Linnean Society*, *171*, 260–276.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Research*, *40*(15), 1–12. doi:10.1093/nar/gks596

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 Phytologist*, *214*, 1736–1750. doi:10.1111/nph.14530

Walker, J. F., Brown, J. W., & Smith, S. A. (2017). Analyzing contentious relationships and outlier genes in phylogenomics. *bioRxiv*. doi:http://dx.doi.org/10.1101/115774

Wyman, S. K., Jansen, R. K., & Boore, J. L. (2004). Automatic annotation of organellar genomes with DOGMA. *Bioinformatics*, *20*(17), 3252–3255. doi:10.1093/bioinformatics/bth352

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829. doi:10.1101/gr.074492.107

Data Accessibility:

DNA sequences: Genbank accessions MF359935–MF359958

Plastome alignment, gene alignments, trees, and R code: https://bitbucket.org/oscarvargash/lecythidaceae_plastomes

Tables

Table 1 Total number of perfect simple sequence repeats (SSRs) identified within the plastome of *Bertholletia excelsa*.

SSR Sequence	**Number of Repeats**															**Total**
	3	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	
A	-	-	-	-	153	70	38	22	14	5	1	2				305
C	-	-	-	-	10	1	-	-	-	-	-	-	-	-	1	12
ATC	-	-	-	-	-	-	-	-	-	-	-	-	-	-		0
AG	-	13	-	-	-	-	-	-	-	-	-	-	-	-	-	13
AT	-	23	3	-	1	-	-	-	-	-	-	-	-	-	-	27
AAC	8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8
AAG	24	-	-	-	-	-	-	-	-	-	-	-	-	-	-	24
AAT	25	2	-	-	-	-	-	-	-	-	-	-	-	-	-	27
ACC	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3
AGC	7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7
AGG	9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	9
ATC	7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7
AATC	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
AATT	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
AAAG	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3
AAAT	3	1	-	-	-	-	-	-	-	-	-	-	-	-	-	4
AAAAT	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
AACTT	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
AAAATT	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
AAACTC	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1
Total	96	39	3	0	164	71	38	22	14	5	1	2	0	0	1	456

Table 2 Regions of the chloroplast regions binned in windows of 600 sites with high (above the 95th percentile) nucleotide diversity (ND) and/or site-wise log-likelihood score differences (LD). LSC: large single copy. SSC: small single copy (see main text). Coding regions are indicated in windows that have the same 5' and 3' expressed flanking region in column 3. Notice that no regions are reported for the inverted repeat (IR). Coordinates are given on the alignment and the *Bertholletia excelsa* plastome that are assembled with the standard LSC-SSC-IR structure.

Location in the alignment	***Bertholletia* cp genome location**	**Closest flanking expressed region**		**Region**	**π**	**LD**
		5'	3'			
1–600	1–490	*trnH*	*psbA*	LSC		*
5401–6000	4885–5373	*trnK-UUU*	*rps16*	LSC		*
34801–35400	30925–31450	*petN*	*trnD-GUC*	LSC		*
35401–36000	31451–31967	*psbM*	*trnD-GUC*	LSC	*	*
37201–37800	33027–33573	*trnE-UUC*	*trnT-GGU*	LSC	*	*
39601–40200	34893–35433	*trnT-GGU*	*psbD*	LSC		*
43801–44400	38798–39254	*psbZ*	*trnfM-CAU*	LSC	*	*
44401–45000	39255–39744	*trnfM-CAU*	*psaB*	LSC	*	*
61201–61800	54771–55275	*trnV-UAC*	*atpE*	LSC		*
78601–79200	70230–70771	*psaJ*	*rps18*	LSC		*
89801–90400	80536–81103	*petD*	*rpoA*	LSC	*	
95401–96000	85455–85906	*rpl16*	*rps3*	LSC	*	
131401–132000	119237–119759	*ccsA*	*ndhD*	SSC	*	
140401–141000	127827–128402	*rps15*	*ycf1*	SSC		*
144001–144600	131283–131868	*ycf1*	*ycf1*	SSC	*	*
144601–145200	131869–132446	*ycf1*	*ycf1*	SSC	*	*

Table 3 Primer sequences used to amplify the ten most polymorphic Lecythidaceae plastome regions, as sorted by decreasing nucleotide diversity (π). The product size (length) references the *Bertholletia excelsa* plastome.

Window in the alignment	**π**	**Region**	**Forward primer Sequence (5' - 3')**	**Reverse primer Sequence (5' - 3')**	**Length (bp)**
144103-145487	0.04691	*ycf1*	AGAACCTTTGATTATGTCTCGACG	AGAGACATGCTATAAAAATAGCCCA	118
95034-95741	0.03446	*rpl16-rps3*	AGAGTTTCTTCTCATCCAGCTCC	GCTTAGTGTGTGACTCGTTGG	101
35585-36413	0.02920	*psbM-trnD*	CCGTTCTTTCTTTTCTATAACCTACCC	ACGCTGGTTCAAATCCAGCT	109
143235-144102	0.02733	*ycf1*	TGATTCGAATCTTTTAGCATTAKAACT	KCGTCGAGACATAATCAAAGGT	118
131180-132054	0.02576	*ccsA-ndhD*	CCGAGTGGTTAATAATGCACGT	GCTTCTCTTGCATTACCGGG	118
44398-45132	0.02537	*trnG-psaB*	TCGATYCCCGCTATCCGCC	GCCAATTTGATTCGATGGAGAGA	88
89032-89688	0.02464	*petD-rpoA*	TGGGAGTGTGTGACTTGAACT	TGACCCATCCCTTTAGCCAA	82
43412-44397	0.02456	*psbZ-trnfM*	TCCAATTGRCTGTTTTTGCATTAATTG	CCTTGAGGTCACGGGTTCAA	70
37444-38345	0.02409	*trnE-trnT*	AGACGATGGGGGCATACTTG	CCACTTACTTTTTCTTTTGTTTGTTGA	132
38346-40085	0.02391	*trnT-psbD*	GGCGTAAGTCATCGGTTCAA	CCCAAAGCGAAATAGGCACA	171

Figures

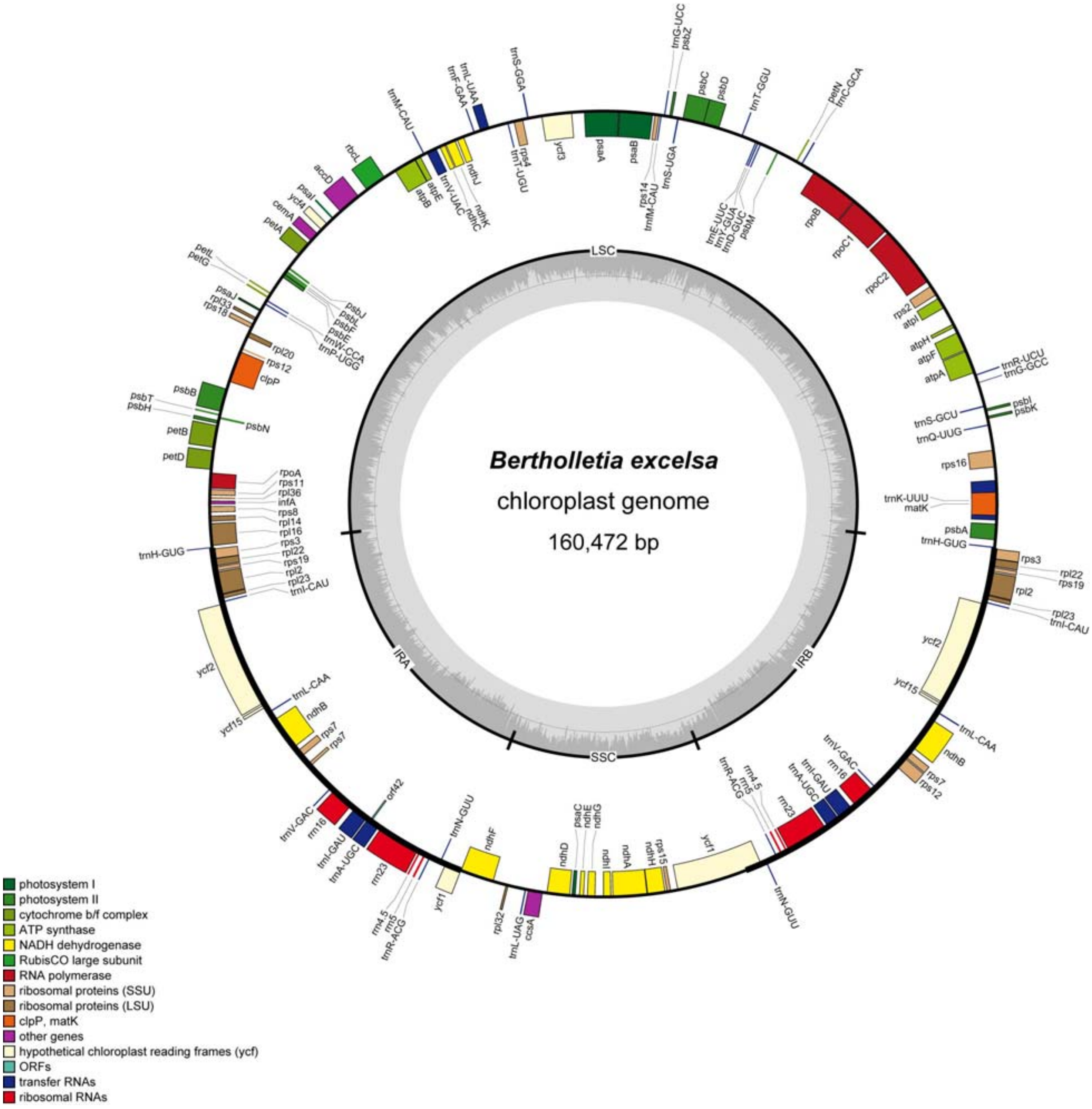
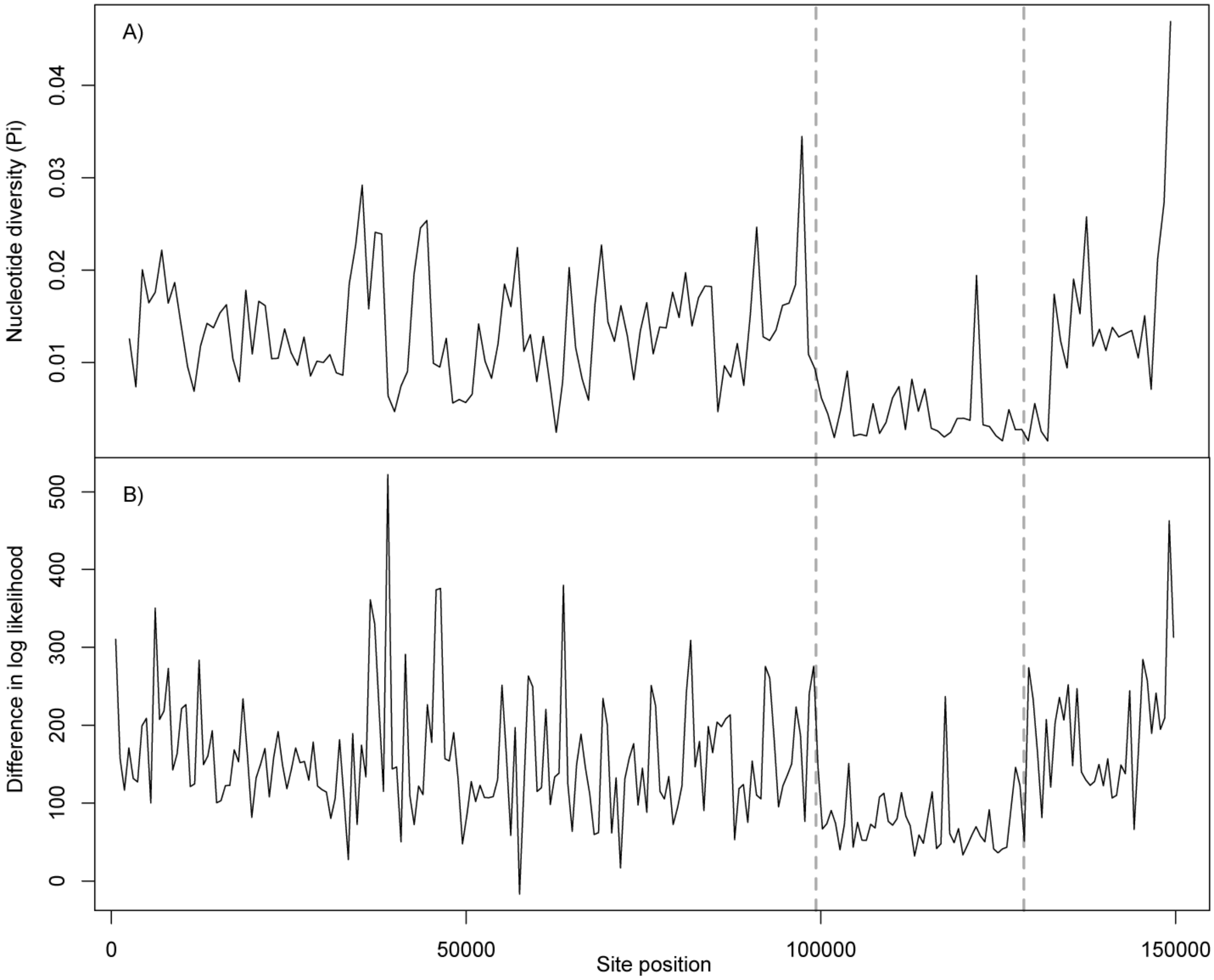


Fig. 1 Plastome map of the Brazil-nut tree *Bertholletia excelsa*. Genes outside the circle are transcribed clockwise, genes inside the circle are transcribed counter-clockwise. Gray bars in the inner ring show the GC content percentage.



456

457 **Fig. 2** A) Sliding window plot of nucleotide diversity (π) across the alignment of 24 sequenced Lecythidaceae plastomes. B) Alignment site-wise differences in log-likelihood calculated from the chloroplast topology vs. the averaged scores of 1000 random trees using a 600-site window. Regions with greater log-likelihood differences contain higher phylogenetic signal. Dashed lines indicate the boundaries, from left to right, among the large single copy, the inverted repeat, and the small single copy.

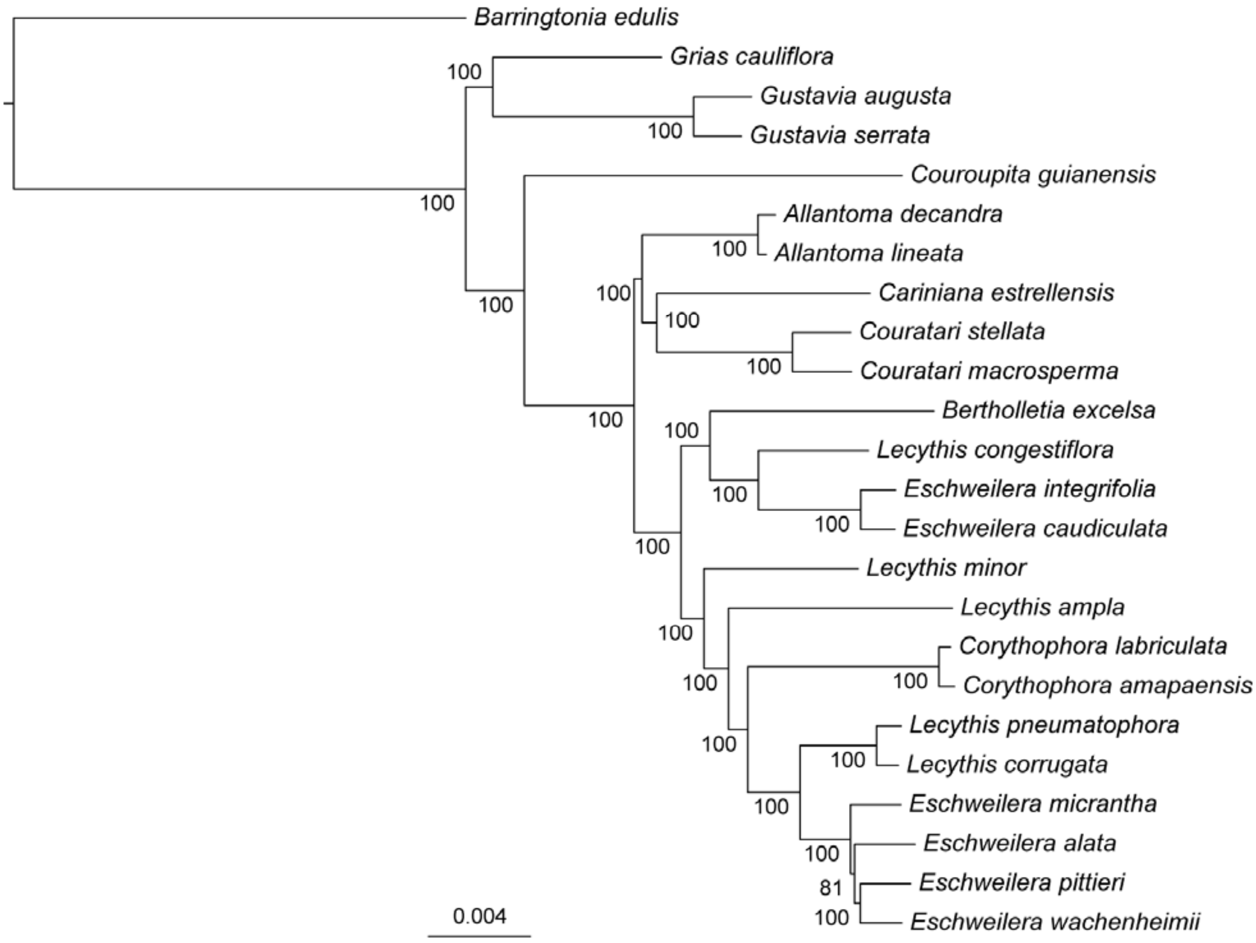


Fig. 3 Maximum likelihood phylogeny inferred from plastomes of Neotropical Lecythidaceae. Numbers at nodes indicate bootstrap support.

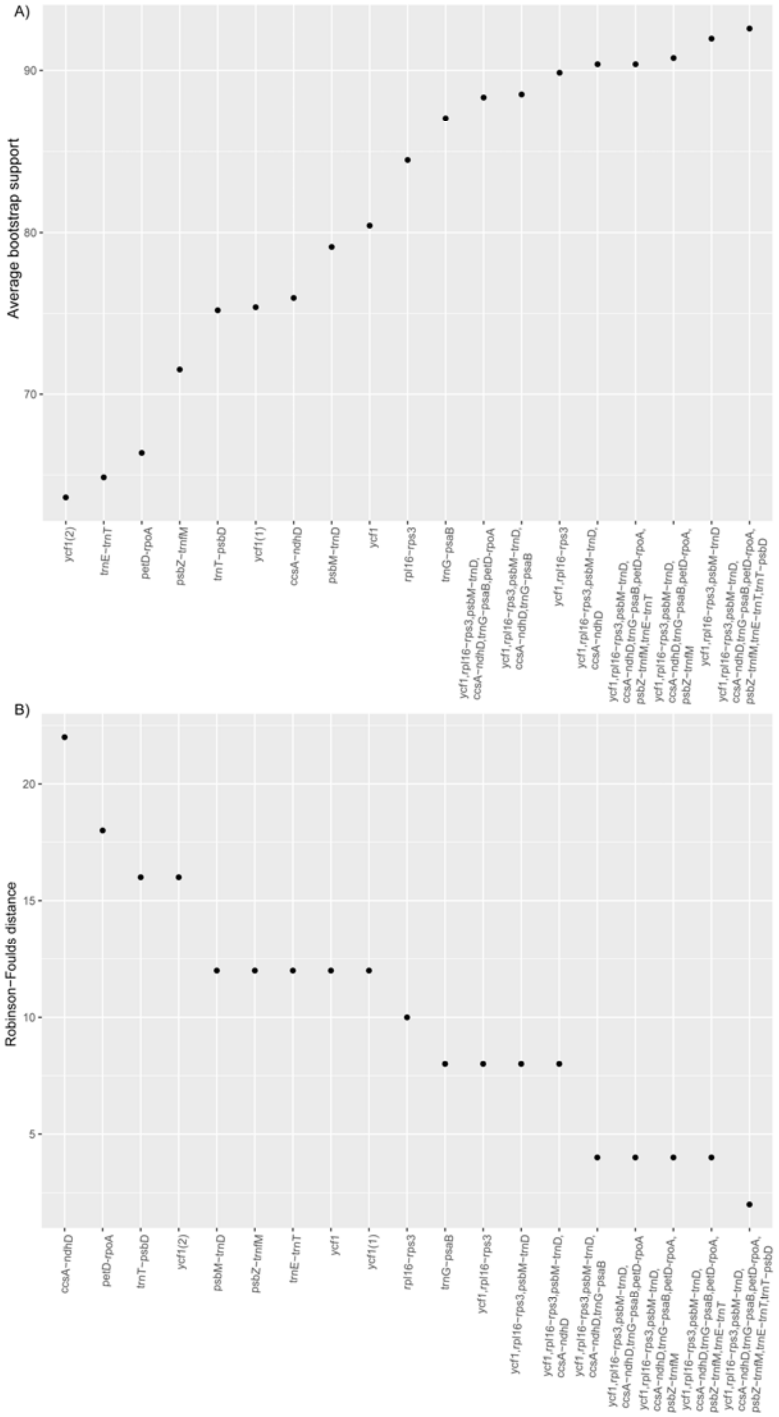


Fig. 4 A) Average bootstrap support for trees inferred from matrices of concatenated regions with relatively high nucleotide diversity sorted in ascending order; and B) Robinson-Foulds distance (RF) sorted in descending order. Lower RF distances, which measures the number of different bipartitions from the complete plastome topology, indicate better accuracy.

Supporting information

Fig. S1 Trees obtained from single and combined markers with high nucleotide diversity.

Table S1 Lecythidaceae species sequenced with their voucher, assembly information, and GenBank accession number. All voucher specimens are deposited at herbarium of the New York Botanical Garden (NY).

Table S2 Primers for the amplification of simple sequence repeats in the plastome of *Bertholletia excelsa*.