The mitochondrial genome of pequi tree (*Caryocar brasiliense* Cambess.): Genome structure, gene transfers, and evolutionary insights within Malpighiales

Larissa R. Carvalho[1,2], Leonardo C. J. Corvalán[1,2], Renata O. Dias[2], Ramilla S. Braga-Ferreira[3], Cintia P. Targueta[2], José Alexandre F. Diniz-Filho[4], Mariana P. C. Telles[2,5], and Rhewter Nunes[1*]

[1] Laboratório de Bioinformática e Biodiversidade, Instituto Acadêmico de Ciências da Saúde e Biológicas (IACSB), Universidade Estadual de Goiás (UEG) - Campus Oeste - UnU de Iporá, Iporá, GO, Brasil.

[2] Laboratório de Genética & Biodiversidade, Instituto de Ciências Biológicas, Universidade Federal de Goiás, Av. Esperança, Chácaras Califórnia, Goiânia, 74.690-900 Goiás , GO, Brasil.

[3] Laboratório de Genética e Biotecnologia, Instituto de Ciências Exatas e Naturais, Universidade Federal de Rondonópolis, Av. dos estudantes, 5055, Cidade Universitária, Rondonópolis, 78736-900 Mato Grosso, MT, Brasil.

[4] Laboratório de Ecologia Teórica, Síntese, Instituto de Ciências Biológicas, Universidade Federal de Goiás, Av. Esperança, Chácaras Califórnia, Goiânia, 74.690-900 Goiás , GO, Brasil.

[5] Escola de Ciências Médicas e da Vida, Pontifícia Universidade Católica de Goiás, Av. University, 1440, Setor Leste Universitário, Goiânia, 74605-010 Goiás, GO, Brasil.

[*] Corresponding author: Rhewter Nunes; E-mail: rhewter@ueg.br

Abstract

The complete mitochondrial genome of *Caryocar brasiliense* (Caryocaraceae), an ecologically and economically important species native to the Brazilian savannas, was assembled and annotated. Using a hybrid assembly approach combining Oxford Nanopore and Illumina sequencing data, we assembled a 533,641 bp bipartite mitogenome organized into two circular chromosomes. A high density of dispersed repeats and simple sequence repeats (SSRs) was detected, along with extensive DNA transfers from the chloroplast and nuclear genomes (MTPTs and NUMTs). The variation of mitogenome size is positively correlated with the number of dispersed repeats ($R^2 = 0.88$). Genome annotation revealed 74 protein-coding genes, including sequences derived from both mitochondrial and chloroplast origins, as well as 376 predicted RNA editing sites, particularly concentrated in energy metabolism genes such as *ccm* and *nad* gene family. Comparative analysis across ten Malpighiales species identified conserved core mitochondrial genes and revealed topological differences between mitochondrial and plastid phylogenies. These findings offer new insights into the structural and evolutionary dynamics of angiosperm mitochondrial genomes and provide a foundational resource for future genetic, evolutionary, and conservation studies in *C. brasiliense* and related taxa.

Keywords: Organelle genome, Inter Organellar DNA transfer, Repetitive DNA, Genome evolution.

Introduction

Mitochondria are organelles involved in many cellular processes in plants and play key roles in development, fitness, and reproduction (Gualberto et al., 2014; Li et al., 2021). Unlike chloroplast genomes, plant mitochondrial (mtDNA) genomes show extensive variation in genome size, gene content, and genome organization (Gualberto et al., 2014; Li et al., 2021). Regarding genome size, plant mitochondrial DNA (mtDNA) ranges from 66 kb in *Viscum scurruloideum* Barlow (Skippingtona et al., 2015) to 11 Mb in *Silene conica* L. (Sloan, Alverson, Chuckalovcak, et al., 2012). mtDNA size is highly variable, even among closely related species, and this variation may be associated with the presence of repeat sequences, introns, nuclear DNA (nDNA), plastid DNA (cpDNA), as well as viral and bacterial sequences (Gualberto et al., 2014; Zardoya, 2020). Repeat sequences can promote recombination events in the mitochondrial genome, contributing to the genomic plasticity of plant mtDNA. These recombination events result in different mtDNA structures, which can be found in linear, circular, and branched-linear patterns (Chevigny et al., 2020).

Despite their large size, plant mitochondrial genomes contain a relatively small number of genes when compared to other genome compartments. Over the course of evolution, mtDNA has lost many genes due to functional redundancy and gene transfer to the nuclear (nDNA) or plastid (cpDNA) genomes, retaining only a reduced set of protein-coding genes (Mower, 2020; Zardoya, 2020). The retained genes are mainly involved in oxidative phosphorylation, as well as in the biogenesis of respiratory complexes and ribosomal proteins (Mower, 2020; Li et al., 2021). The complexity and instability of plant mitochondrial DNA (mtDNA) have contributed to a notable delay in mtDNA research compared to the cpDNA. Currently, the National Center for Biotechnology Information (NCBI) Organelle database contains approximately 465 mitochondrial genomes (NCBI, 2025) from the angiosperm clade. In contrast, more than 13,761 chloroplast genomes are available in the same database (NCBI, 2025). These numbers are continuously increasing as new organelle genomes are deposited. Expanding the availability and knowledge of plant mitochondrial genomes is essential for understanding evolutionary dynamics, adaptive processes, and biological diversity, as well as serving as a valuable tool for plant breeding. Furthermore, comparative analyses of phylogenies reconstructed

from mitochondrial, nuclear, and plastid genomes in angiosperms have revealed that each genomic compartment may display distinct evolutionary histories and topologies (Bogdanova *et al.*, 2021). Thus, exploring the congruence and differences among nuclear, mitochondrial, and plastid phylogenies provides important insights into plant evolution and lineage diversification (Hu *et al.*, 2023).

The generation of complete mitochondrial genome sequences from underrepresented lineages within Malpighiales, such as Caryocaraceae, can provide a valuable opportunity to help address existing gaps. In this context, the species *Caryocar brasiliense*, popularly known as "pequi tree", emerges as an interesting Caryocaraceae representative for comparative genomics studies. This plant is an important fruit species of the Brazilian savannas (Choubey and Rajam, 2015; Nunes, Gonçalves, and Telles, 2019; Nascimento-Silva, Mendes, and Silva, 2020; Nunes, de Lima, *et al.*, 2020; Nunes, de Souza, *et al.*, 2020) and has been identified as a priority species for conservation and sustainable use by Brazil's Ministry of the Environment (Vieira, Camillo, and Coradin, 2016). Previous efforts have made available a draft of the nuclear genome and SSR markers (Nunes, Gonçalves, and Telles, 2019), as well as the complete plastidial genome (Nunes, de Souza, *et al.*, 2020). However, the mitochondrial genome remains a scientific gap for this species and for the Caryocaraceae family as a whole.

Therefore, in this study, we present, for the first time, the complete sequence and annotation of the mitochondrial genome of *C. brasiliense*, and use it in a comparative and evolutionary analysis of mitogenomes within the order Malpighiales. Especially, we hypothesized that i) variation in mitogenome size among Malpighiales is associated with the accumulation of repetitive elements; ii) recurrent DNA transfers from the nucleus and plastid contribute substantially to the composition and expansion of the mitochondrial genome in *C. brasiliense*; iii) mitochondrial RNA editing events are predominantly concentrated in genes associated with energy metabolism; and iv) comparative analyses of mitochondrial and plastid phylogenies reveal topological differences among families within Malpighiales.

Material and methods

DNA extraction and Oxford Nanopore sequencing

Leaf tissue was collected from a *Caryocar brasiliense* individual located at the Escola de Agronomia of the Universidade Federal de Goiás, Goiânia, Goiás, Brazil (16°35'49.8" S 49°16'46.8" W). DNA extraction was performed according to the protocol by Doyle and Doyle (1987). Its integrity and quantity were verified using 1% agarose gel stained with ethidium bromide, an Agilent 4150 TapeStation with a Genomic DNA ScreenTape (Agilent Technologies, Santa Clara, CA, USA), and a Qubit® with the dsDNA High Sensitivity Kit (Thermo Fisher Scientific, United States). Library preparation was performed using the Ligation Sequencing Kit V14 – SQK-LSK114 (Oxford Nanopore Technologies, United Kingdom), according to the manufacturer's protocol. The final library was quantified and loaded into a FLO-PRO114M flow cell on a PromethION 2 at a final concentration of 1372 ng and configured for a 72 h run. After 24 hours, the flow cell was washed and the library was reapplied for a further 48 h run.

Mitochondrial genome assembly and annotation

Mitochondrial genome assembly was performed using Illumina short reads available in the NCBI database (GenBank GCA_004918865.1, Nunes, Gonçalves, and Telles, 2019) and Oxford Nanopore long-read data generated herein. First, we generated a database with 51 mitochondrial genome sequences from Malpighiales species (Supplementary File S01), available in the NCBI database. We aligned the long reads and short reads to the database using Minimap2 2.24-r1122 (Li, 2018) and extracted only the aligned reads. Then, we used Unicycler v0.5.0 (Wick *et al.*, 2017) for hybrid assembly using the filtered long reads and short reads. Subsequently, we aligned the complete set of Illumina short reads to the sequences generated using Minimap2. Afterward, we used the alignment to polish the assembly using Pilon v.1.24 (Walker *et al.*, 2014).

For genome annotation, we used the CHLOROBOX platform with the software GeSeq (Tillich *et al.*, 2017), MFannot v. October 2019 (Beck and Lang, 2010), and Mitofy v. March 22, 2012 (Alverson *et al.*, 2010), using the mitogenome of *Ricinus communis* (NC_015141.1) as a reference. For tRNA gene annotation, we used tRNAscan-SE v.2.0.7 (Lowe and Eddy, 1997) and Aragorn v.1.2.38 (Laslett and Canback, 2004). Manual curation of annotations was performed using Geneious Prime v.2023.2.1. For chloroplast genes, we retained only complete genes with $\geq$ 90% coverage, $\geq$ 80% identity, and the

presence of both start and stop codons. Putative chloroplast pseudogenes were defined by the presence of internal stop codons in the best-supported alignments, based on coverage and sequence identity. Frameshift mutations were not explicitly evaluated in the classification of pseudogenes. The genome map was drawn using OGDRAW v.1.3.1 (Greiner, Lehwark, and Bock, 2019).

Intracellular gene transfer, repeat sequences, and comparative analysis among Malpighiales mitochondrial genomes

The thresholds used to identify MTPTs (≥80% identity, ≥50 bp) and NUMTs (≥80% identity, ≥500 bp) were based on previous studies of inter-organellar transfers in plants (Wei et al. 2023; Nhat Nam et al. 2024; Kong et al. 2025). The higher threshold for NUMTs (500 bp versus 50 bp for MTPTs) was chosen to minimize spurious matches between mitochondrial and nuclear sequences, particularly considering that the nuclear genome used in this study is at draft level. These criteria represent a balance between sensitivity to detect genuine transfers and specificity to avoid false positives due to sequence similarity from ancestral homology or convergence.

To detect internal repeats in each mitochondrial chromosome of *C. brasiliense*, we used the *nucmer* command from NUCmer 4 v.4.x (Marçais *et al.*, 2018). For dispersed repeats, REPuter was employed with a hamming distance of 3 and a minimum repeat size of 30 bp (available at: https://bibiserv.cebitec.uni-bielefeld.de/reputer). Simple sequence repeats (SSRs) were identified using MISA (available at: https://webblast.ipk-gatersleben.de/misa/), with a minimum value of ten repetitions defined for mononucleotide SSR, five for dinucleotides, four for trinucleotides, and three for tetranucleotides, pentanucleotides, and hexanucleotides. The Tandem Repeat Finder v 10-2022 (Benson, 1999) program was used to identify the other tandem repeats.

To analyze the evolutionary relationships among mitochondrial genome size (bp) and number of repeats, and total repeat length (bp), respectively, we used a Phylogenetic Generalized Least Squares (PGLS), incorporating the phylogenetic covariance structure derived from the mitogenome tree. The PGLS models were constructed using mitochondrial genome size as the response variable and modelled against the two predictors (number of repeats and total repeat length) and fitted using the Caper R package (Orme *et al.*, 2023). The choice of this method is due to the potential non-independence of the

data due to the phylogenetic structure of the data. PGLS models were fitted using the caper package in R, and Pagel's λ used to model phylogenetic covariance was estimated via Maximum Likelihood for each regression. Because of the relatively small sample size, maximum likelihood estimates of Pagel's λ may be unstable, so we also fitted a standard multiple regression OLS model (which is actually a PGLS assuming Pagel's λ = 0, thus without phylogenetic structure in data).

In addition to *C. brasiliense* we used other nine Malpighiales species for comparative genomic analysis: *Banisteriopsis caapi* (Malpighiaceae, OR473419.1), *Bruguiera sexangula* (Rhizophoraceae, NC_056359.1), *Calophyllum soulattri* (Calophyllaceae, NC_079842.1), *Populus alba* (Salicaceae, NC_041085.1), *Salix brachista* (Salicaceae, NC_058733.1), *Salix dunnii* (Salicaceae, NC_058734.1), *Viola diffusa* (Violaceae, PP952082.1), *Manihot esculenta* (Euphorbiaceae, NC_045136.1) and *Ricinus communis* (Euphorbiaceae, NC_015141.1). All the selected species have their mitochondrial genomes completely sequenced and available in the NCBI RefSeq database or NCBI GenBank. They were selected for the reliability of their annotations and to minimize redundancy among genera, with the exception of two *Salix* species included to evaluate intra-genus variation..

Molecular evolution in Malpighiales mitochondrial genomes

The complete sequence for both mitochondrial chromosomes of *C. brasiliense* was used to predict RNA editing sites in protein-coding genes. The prediction was performed in the online plant RNA editing site prediction tool PREPACT3 (Lenz and Knoop, 2013) (available at: https://www.prepact.de/prepact-main.php), in BLASTX mode, using the mitogenome of *Vitis vinifera* (NC_012119.1) as a reference. The relative synonymous codon usage (RSCU) was calculated for protein-coding genes using MEGA v4 software.

We recovered a phylogenetic tree for mitochondrial and chloroplast genomes. The phylogenetic tree was recovered using the 15 species for Malpighiales and the sequence of *C. brasiliense* recovered in this study and *C. brasiliense* cpDNA available in NCBI (Nunes, et al., 2020) (Supplementary File S02), with *Delonix regia* (Fabales) as an outgroup. For both genomes (mtDNA and ptDNA), all protein-coding genes were extracted from these, and the orthologous genes were identified using the orthofinder v. 2.5.4 (Emms and Kelly, 2019). We recovered 64 orthogroups, among which 21 orthogroups were

found in all species, and only 17 are single-copy genes for the mitochondrial genome. For the chloroplast, we recovered 80 orthogroups, among which 63 were found in all species and 55 are single-copy genes. These single-copy genes were aligned using Mafft version 7.50 (Katoh, Rozewicki and Yamada, 2018) and phylogenetically informative regions were selected using Gblocks 0.91b (Castresana, 2000). The sequences were concatenated using the script catfasta2phyml.pl (https://github.com/nylander/catfasta2phyml). We performed phylogenetic inference using the maximum likelihood method with IQ-TREE3 (Wong *et al.*, 2025). The ModelFinder, which is also implemented in IQ-TREE3, was used to select the optimal nucleotide substitution model (GTR + F + R2) (Wong *et al.*, 2025). A consensus tree was generated with 1,000 bootstrap replicates.

To evaluate the two phylogenies obtained with the different regions (For both genomes (mtDNA and cpDNA), we first used standardized Robinson–Foulds (RF) distances (Robison and Foulds, 1981), which basically compare the number of partitions between them and standardize it by the total. RF values close to zero reveal a good match between the two topologies. We also compared them by means of a Mantel test (Manly, 2020) between the cophenetic pairwise distances matrices, in terms of the overall similarity between matrices (considering thus both topology and branch lengths). All analyses were performed using the R package ape (Paradis et al., 2004).

Results

General features of *Caryocar brasiliense* mitogenome and comparative analysis among Malpighiales mitochondrial genomes

A total of 866,975 long reads and 30,515 short reads were used for the assembly of the *Caryocar brasiliense* mitochondrial genome. The complete mitogenome has 533,641 bp in size, organized into two circular chromosomes: chromosome 1 (299,593 bp) and chromosome 2 (234,048 bp) (Figure 1). The non-coding regions were 62.24% of the total genome in chromosome 1 and 64.23% in chromosome 2. The GC content in *C. brasiliense* was 44% in both chromosomes. A total of 74 protein-coding genes were identified, comprising 34 in chr1 (20 mitochondrial and 14 chloroplast-derivated genes) and 40 in chr2 (18 mitochondrial and 22 chloroplast-derivated sequences). Among the chloroplast-derived sequences (MTPT genes, Supplementary file S03), nine were classified as pseudogenes. For the transfer

RNA genes (tRNA), a total of 33 were identified, of which 15 genes originated from the chloroplast genome. Regarding the ribosomal RNA (rRNA) genes, seven genes were identified, of which three originated from the chloroplast genome (Table 1).

In our analysis within species from the order Malpighiales, we identified a conserved set of core mitochondrial genes, comprising ATP synthase subunits (*atp*), cytochrome c oxidase subunits (*cox*), cytochrome c maturation genes (*ccm*), NADH dehydrogenase subunits (*nad*), as well as *mttB*, *matR*, and *cob*. Among the species analyzed, *M. esculenta* was the only one lacking the *cox1* gene. Notably, *B. sexangula* exhibited two copies of both *atp6* and *ccmB* genes, whereas *B. caapi* has two copies of *atp4* and *ccmB*. The ribosomal protein gene families (*rpl* and *rps*) displayed greater variability among the species, with only *rps3* and *rps12* consistently present across all taxa (Figure 2). Patterns of shared gene absence were observed among closely related taxa, which are consistent with the possibility of ancestral gene losses within these lineages. Based solely on the observed presence/absence patterns, the absence of *rps13* in *C. brasiliense*, *C. soulattri*, *P. alba*, *S. brachista*, *S. dunnii* and *V. diffusa* may reflect an early loss along the lineage leading to this group; however, explicit ancestral-state reconstruction would be required to formally test this hypothesis.

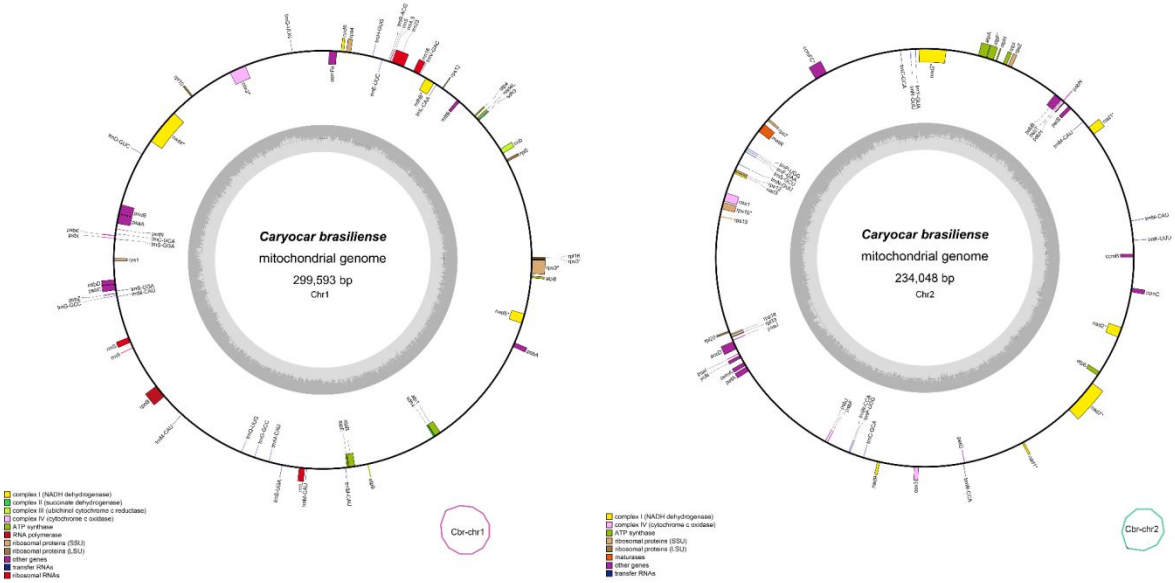

Figure 1. Genome map of the two mitochondrial chromosomes of *Caryocar brasiliense*. The genes in the inner part of the circles are transcribed clockwise, and the genes in the outer part are transcribed counterclockwise. The genes are colored according to the functional group to which they belong, as indicated in the legend. A gray bar plot on the internal circle represents the GC content distribution. The smaller pink and green circles in the image show the graph generated during assembly for each chromosome.

Table 1. Description of the genes identified in both chromosomes of the *Caryocar brasiliense* mitochondrial genome.

Group of genes	Gene
Complex I (NADH dehydrogenase)	*nad1*[2]*, *nad2*[2]*, *nad3*[2], *nad4* †, *nad4L*†*, *nad6*†, *nad7*[2]* and *nad9*[2]
Complex II (succinate dehydrogenase)	*sdh3*† and *sdh4*†
Complex III (ubiquinol cytochrome c reductase)	*cob*†
Complex IV (cytochrome c oxidase)	*cox1*‡, *cox2*†* and *cox3*‡
Complex V (ATP synthase)	*atp1*†, *atp4*†, *atp6*‡, *atp8*† and *atp9*†
Cytochrome c biogenesis	*ccmB*‡, *ccmC*‡, *ccmFC*‡* and *ccmFn*†
Ribosomal proteins (SSU)	*rps1*†, *rps3*†*, *rps4*†, *rps7*‡*, *rps10*‡*, *rps12*†‡[2#] and *rps19*‡
Ribosomal proteins (LSU)	*rpl5*†, *rpl10*† and *rpl16*†
Maturase	*matR*‡
Transport membrane proteins	*mttB*†
tRNA	*trnC-GCA*†,‡,§, *trnD-GUC*†, *trnE-UUC*†, *trnF-GAA*‡, *trnG-GCC*†,#, *trnH-GUG*†, *trnK-UUU*‡, *trnL-CAA*†, *trnM-CAU*†,‡,#,§, *trnN-GUU*‡,§, *trnP-UGG*‡,§, *trnQ-UUG*†,#, *trnR-ACG*†, *trnS-*

	GCU[†,‡,#], *trnS-GGA*[†], *trnS-UGA*[†,#], *trnV-GAC*[†], *trnW-CCA*[‡,§], *trnY-GUA*[‡]
rRNA	*rrn4.5*[†], *rrn5*[†,#], *rrn16*[†], *rrn23*[†], *rrnL*[†] and *rrnS*[†]
ptDNA genes	*accD*[‡], *atpA*[‡], *atpE*[†], *atpH*[‡], *atpI*[‡], *cemA*[‡], *petA*[‡], *petB*[‡], *petG*[‡], *petN*[‡], *psaI*[‡], *psaJ*[‡], *psbB*[‡], *psbC*[†], *psbH*[‡], *psbI*[†], *psbK*[†], *psbN*[‡], *psbT*[‡], *psbZ*[†], *rpl20*[‡], *rpl33*[‡], *rps2*[‡], *rps18*[‡], *ycf4*[‡]
pseudogenes	*atpB*[†], *atpF*[‡*], *ndhB*[†*], *psaA*[†], *psaB*[†], *psbA*[†], *psbD*[†], *psbF*[†], *rpoB*[†]

223 [†] Genes on chromosome 1; [‡] Genes on chromosome 2; [*] Genes with introns; [#] Duplicated genes in
224 chromosome 1. [§] Duplicated genes in chromosome 2.

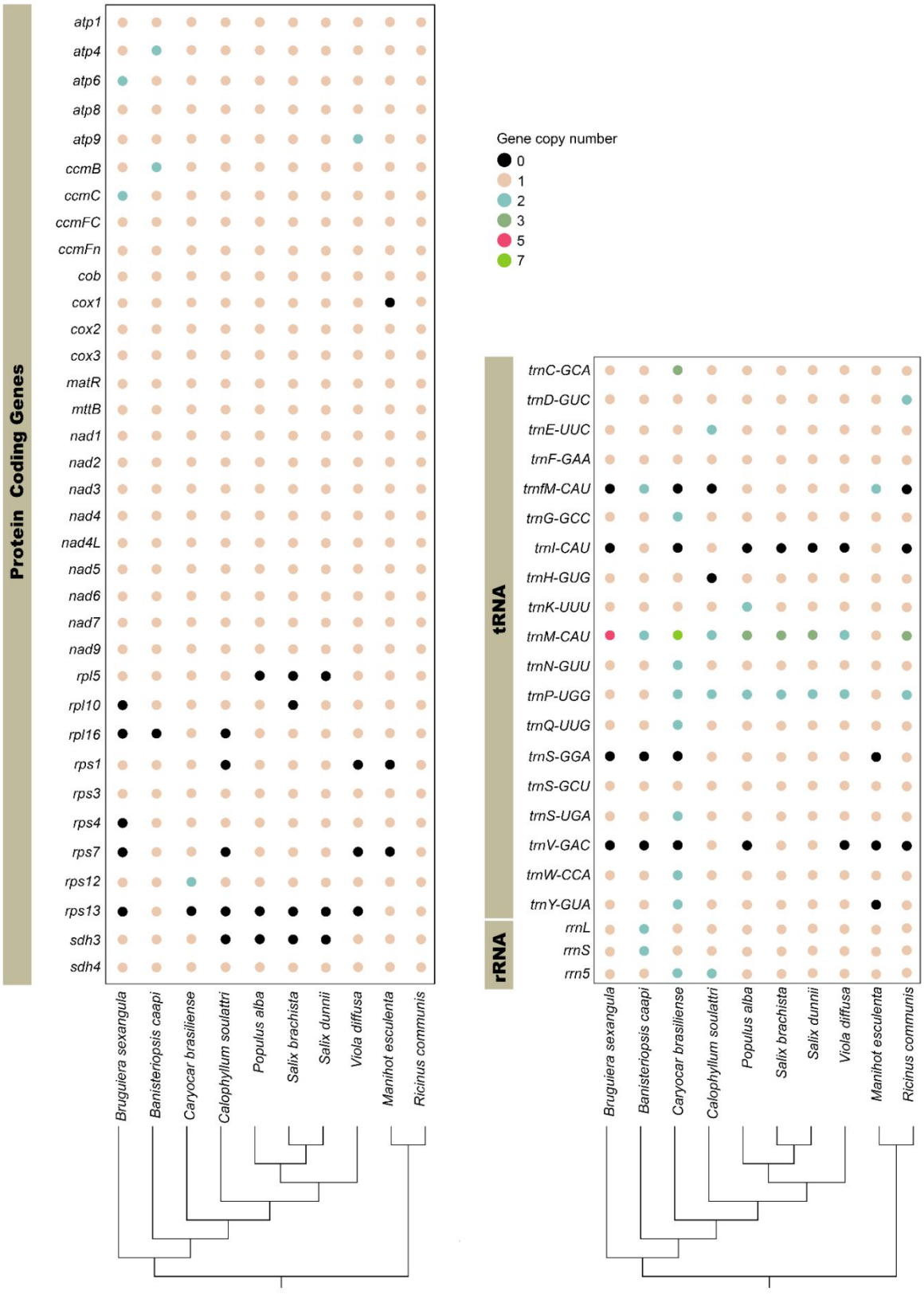

Figure 2. Gene variation across the mitogenomes of ten Malpighiales species. Circle colors indicate the number of gene copies in each species, as shown in the legend. Species were sorted according to the

mitochondrial phylogeny, inferred in IQ-TREE3 using the GTR+F+R2 substitution model with 1000 ultrafast bootstrap replicates.

Intracellular gene transfer and repeat sequences of the mitochondrial genome

We identified a total of 58 shared regions between the chloroplast and mitochondrial genomes (MTPTs) of *C. brasiliense* (Figure 3A). Of these, 28 MTPTs are located in mitochondrial chromosome one, corresponding to 50,532 bp (16.91%) of total genome size, and 30 MTPTs are located in chromosome two, corresponding to 31,803 bp (13.62%) of total genome size (Supplementary File S03). These shared regions include a total of 52 chloroplast genes: 34 protein-coding genes, 15 tRNA genes, and 3 rRNA genes. The largest MTPT was found in chromosome one, with a total length of 9,208 bp (Cbr-chr1_MTPT1). This MTPT involves five genes located in the inverted repeat (IR) region of the chloroplast (*trnV, rrn16, rrn23, rrn5,* and *trnR*). Chromosome two contains the second-largest MTPT (Cbr-chr2_MTPT1), with a total length of 6,026 bp and involving five chloroplast genes located in the large single-copy (LSC) region (*accD, psaI, ycf4, cemA,* and *petA*).

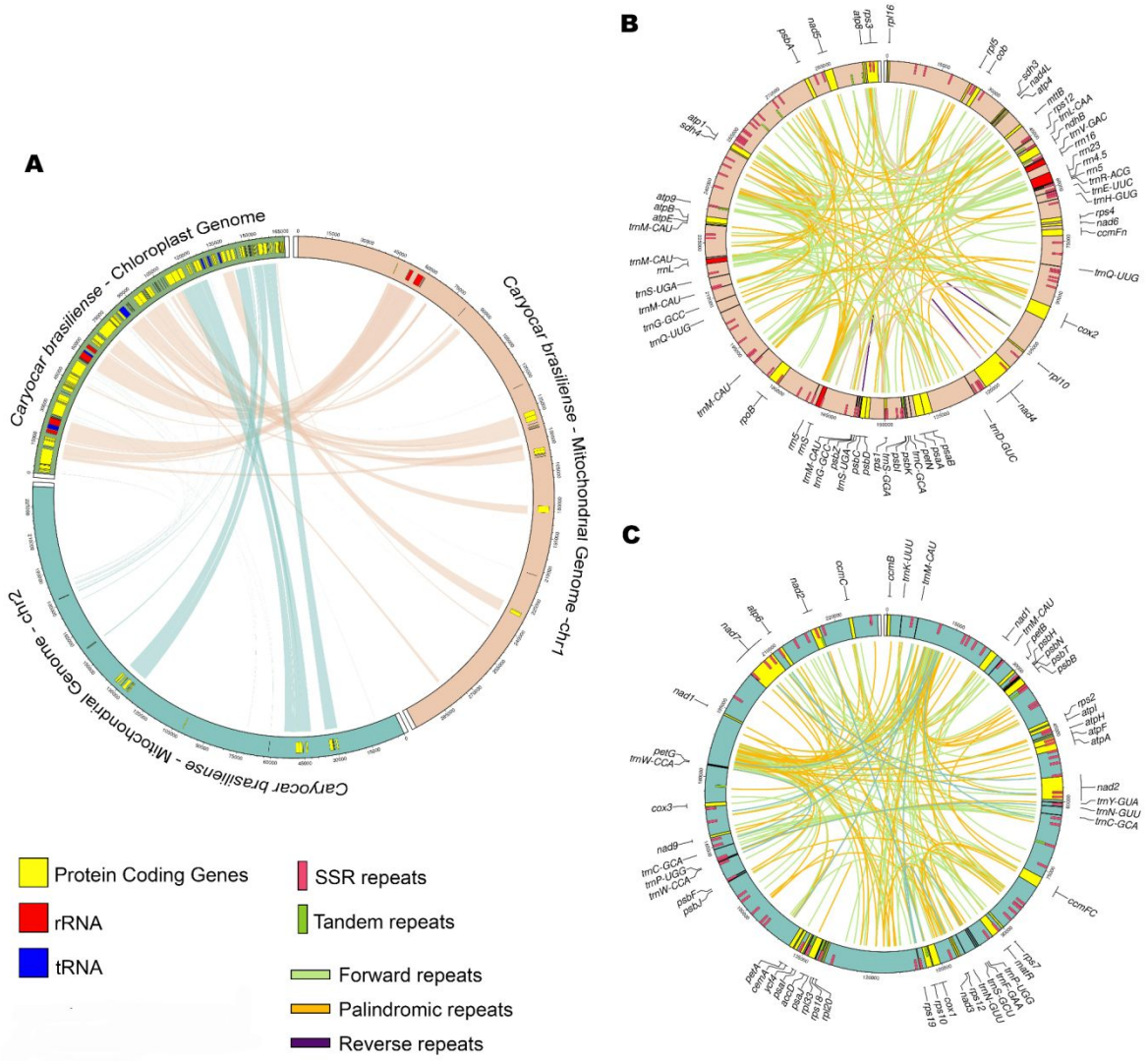

Figure 3. Schematic diagram of the *Caryocar brasiliense* genome architecture. A) Mitochondrial DNA of plastid origin (MTPTs) was identified in both mitochondrial chromosomes. B) Repetitive sequences identified in mitochondrial chromosome one. C) Repetitive sequences identified in mitochondrial chromosome two. The links between the regions in part A represent the MTPTs. The links in parts B and C represent the positions of complex repetitions, coloured according to type. The gene positions are indicated by the large boxes.

We identified 77 putative NUMT regions. Among these, 33 are located on chromosome 1, spanning 92,238 bp (30.78%), and 44 are located on chromosome 2, totaling 83,636 bp (35.72%) (Supplementary File S04). Four nuclear genes were found to be associated with these NUMT regions. In Cbr-chr1_NUMT10, we identified a segment corresponding to intron 2 of the *P-loop NTPase*

domain-containing protein LPA1 homolog gene. In Cbr-chr2_NUMT6, a segment matching intron 1 of the *Serine/threonine-protein kinase* gene was detected. In Cbr-chr2_NUMT10, the NUMT includes both exon five and intron 4 of a gene belonging to the *chaperonin (HSP60) family*. Notably, only one NUMT (Cbr-chr2_NUMT42) contains a complete nuclear gene: *ribonuclease H protein At1g65750-like*. The largest NUMT (6,905 bp) was found on chromosome 2 (Cbr-chr2_NUMT35) and is not associated with any gene region.

The repeat regions are shown in Figure 4. About the repeat regions, Chromosome 2 showed a higher abundance and wider size range of internal and dispersed repeats than chromosome 1, including the longest repeats detected in the dataset. In both chromosomes, forward and palindromic types dominated dispersed repeats, whereas reverse repeats were rare and no complementary repeats were detected. The size of dispersed repeats ranged from 30 to 222 bp in chromosome 1, and from 30 to 1000 bp in chromosome 2. Among the tandem repeats, we found eight repeats in chromosome one and seven repeats in chromosome 2. In comparison to the other Malpighiales species, *M. esculenta* exhibited the highest number of dispersed repeats, followed by *P. alba*. *B. sexangula* presented the lowest number. Regarding short tandem repeats (SSR) (1 to 6 bp), *S. dunnii* showed the highest count, while *V. diffusa* exhibited the lowest, as well as the tandem repeats (Figure 4). In all species, forward repeats are the most frequent dispersed repeat and the mononucleotide motif was the most frequent type of SSR.

Regarding the tandem repeats, we found eight in chromosome 1 and seven in chromosome 2. About the SSR repeats, we found 76 repeats in chromosome 1 and 64 in chromosome 2. The mononucleotide motifs are the most frequent, followed by the tetranucleotide motifs. In chromosome 2, we did not find hexanucleotide motifs. The most frequent repeat motifs were A/T (45 repeats) and AAAG/CTTT (20 repeats), in both chromosomes.

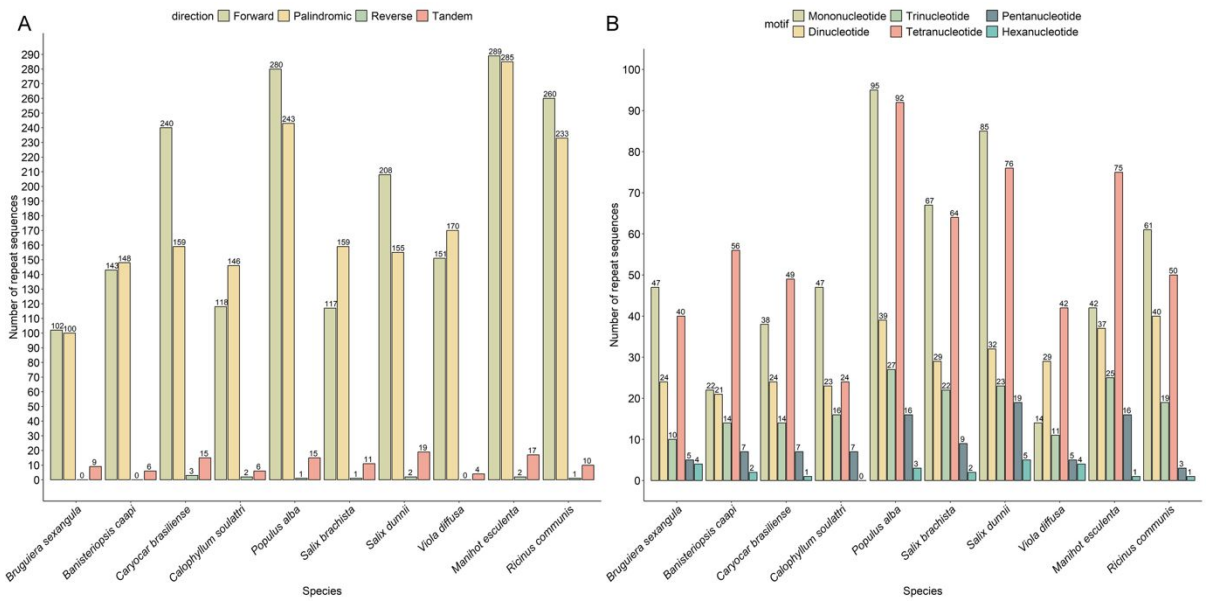

Figure 4. Repetitive elements identified in the mitochondrial genome of *Caryocar brasiliense* compared with other Malpighiales species. A) Comparison of dispersed repeats between the mitochondrial genome of *C. brasiliense* and those of nine Malpighiales species; B) SSRs identified in the mitochondrial genome of ten Malpighiales species.

Our investigation into the relationship between mitochondrial genome size and both the number of dispersed repeats and the total size of dispersed repeats, using PGLS, revealed a significant positive correlation for the number of repeats only (t = 3.17; P = 0.015). The overall adjusted R2 of the PGLS was equal to 0.85, with maximum likelihood estimate of lambda equal to 1, indicating thus a linear phylogenetic structure underlying the model, corresponding to a Brownian motion evolutionary model. The PGLS analysis confirmed a strong positive correlation between mitogenome size and the number of dispersed repeats ($\beta = 963.14$, $R^2 = 0.88$, adjusted $R^2 = 0.87$, $F_{1,8} = 61.02$, $p = 5.18 \times 10^{-5}$), with an estimated increase of approximately 963 bp in genome size per additional repeat unit (95% CI: [692, 1234] bp). However, as Pagel's λ for each variable were actually low (i.e., equal to 0.519, 0.559 and 0.746 for genome size, number of repeats and total repeat length, respectively), we also fitted a standard multiple regression (OLS) model. The results of OLS were qualitatively similar to those in PGLS, even though the relative effect of total repeat was slightly higher and marginally significant (P = 0.061), as expected due to its slightly higher Pagel's λ. The overall fit of OLS is much smaller than the PGLS (R2 = 0.62). In any case, despite the uncertainty in Pagel's λ for PGLS, the main explanations for the

mitochondrial genome size are relatively stable along the full gradient of phylogenetic structure in data. Importantly, other factors, such as number of analysed taxons, foreign DNA and variations in recombination rates, may also contribute to genome expansion. The limited number of mtDNA sequences available for Malpighiales species could restrict the robustness of this inference and could be more explored in future studies.

RNA editing sites, codon usage, and phylogenetic analysis

A total of 376 RNA editing sites were predicted across 35 protein-coding genes (Supplementary File S05). The genes *nad4* (29), *ccmFn* (27), *ccmFc* (22), and *nad7* (20) had the highest number of RNA editing sites (Figure 5). Two types of RNA editing were identified: C→U (cytosine to uracil), with 259 sites, and U→C (uracil to cytosine), with 141 sites. Regarding amino acid changes, 25 modifications were found, with the most frequent being S→L (serine to leucine), P→L (proline to leucine), S→F (serine to phenylalanine), and L→P (leucine to proline), occurring 58, 57, 38, and 31 times, respectively. The gene *ccmFc* contained an editing site in the last codon of the sequence, changing the amino acid arginine to a stop codon (R→*).

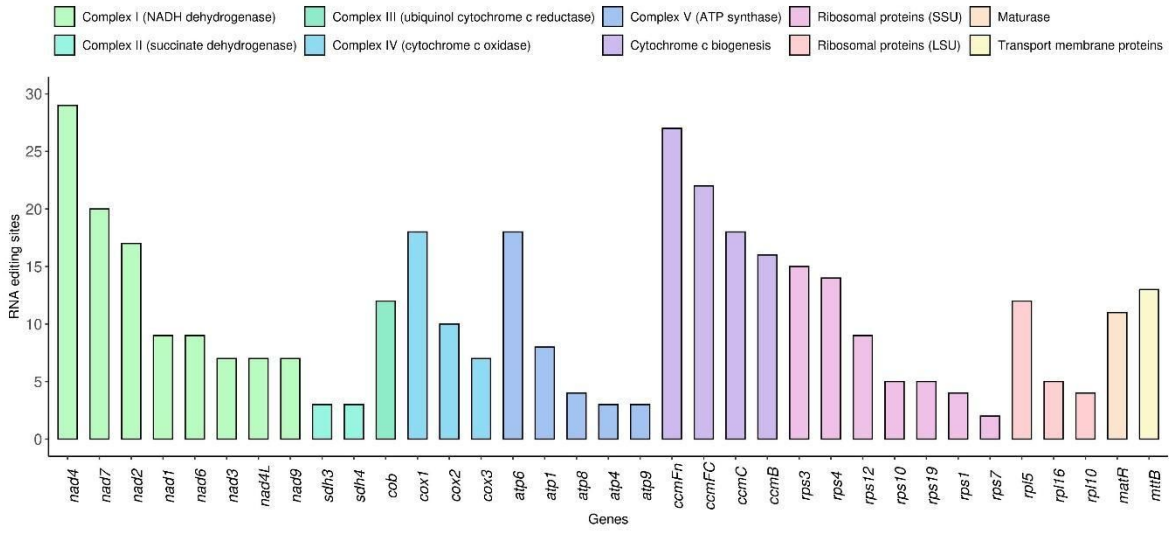

Figure 5. Total RNA editing sites for the 35 protein-coding genes present in the mitochondrial genome of *Caryocar brasiliense*. Genes are separated according to the functional group.

The RSCU values were calculated considering all protein-coding genes of *C. brasiliense* (Figure 6). A total of 32 codons showed values greater than 1, indicating a preference for these codons. The stop codon AGA exhibited the highest RSCU value (2.65), followed by GCU coding for alanine (1.61) and CAU coding for histidine (1.52). We observed a preference for the AUG codon (methionine), with RSCU values of 1.07. Additionally, there was a notable preference for codons ending with A and U.

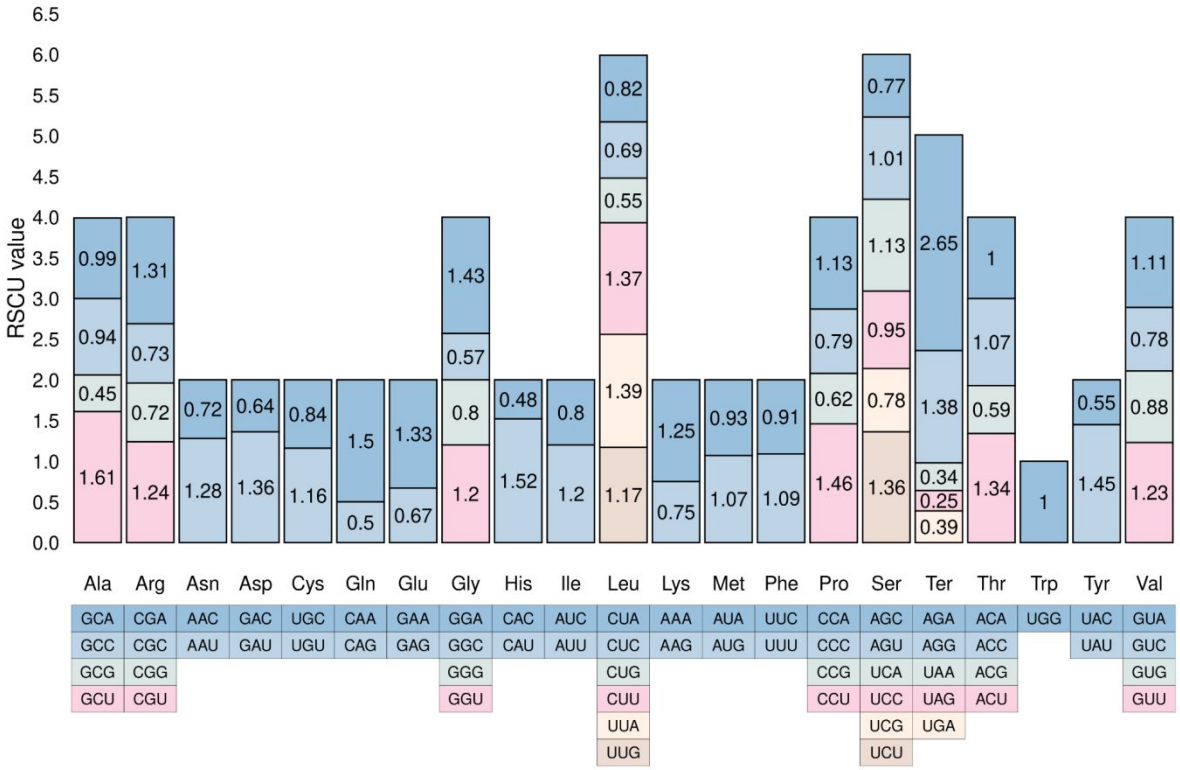

Figure 6. RSCU of the codons in the mitochondrial genome of the *Caryocar brasiliense.*

The phylogenetic analyses based on mitochondrial and chloroplast genomes, using 17 and 55 single-copy genes, respectively, grouped all species into their corresponding families (Figure 7). Bootstrap values in the mitochondrial phylogeny ranged from 53 to 100, while in the chloroplast phylogeny, they ranged from 17 to 100. However, the two trees exhibited slightly different topologies, with standardized RF equal to 0.25, but overall similarity considering the branch lengths is very high (Mantel test equal to 0.91, $P << 0.01$), due to high similarity within some genus (e.g., *Populus, Salix*), as well as common long and isolated branches in others (e.g., *Bruguiera, Banisteriopsis* and *Callophyllum*). An important point in respect of the focal species is that, in the mitochondrial tree, *C.*

brasiliense formed a sister group to the clade composed of Salicaceae, Violaceae, and Calophyllaceae, following the topology Caryocaraceae + (Calophyllaceae (Violaceae + Salicaceae)). In contrast, the chloroplast phylogeny placed Caryocaraceae as a sister clade to Euphorbiaceae (Caryocaraceae + Euphorbiaceae).

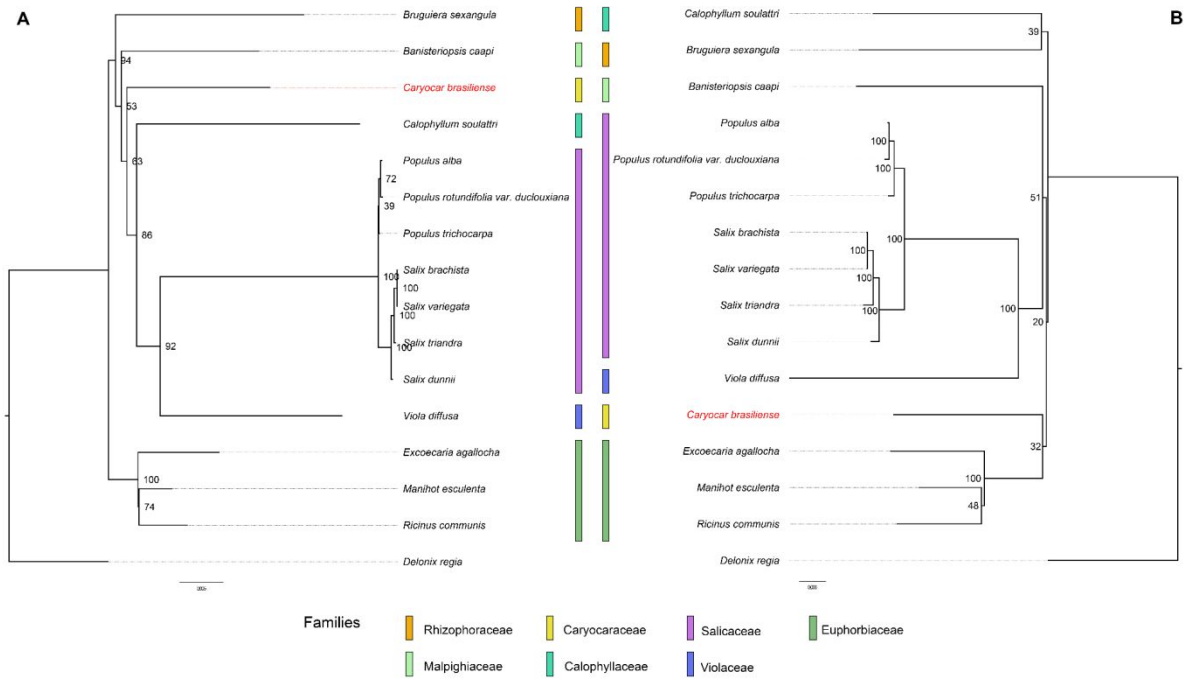

Figure 7. Comparison between Mapighiales phylogenetic trees reconstructed using the Maximum Likelihood method. A) Phylogenetic tree of 17 single-copy genes from the mitochondrial genomes. B) Phylogenetic tree of 55 single-copy genes from the chloroplast genomes. The species *Delonix regia* (Fabales, Fabaceae) was used as an outgroup to root the trees. The values on the nodes represent bootstrap values for 1,000 replicates.

Discussion

Plant mitochondrial genomes are highly dynamic, exhibiting high variation in structure, gene content, and genome size. Unlike the compact mitogenomes of animals, plant mitogenomes display frequent structural rearrangements, abundant repeats, and incorporation of foreign DNA, which complicate their genome assembly and annotation (Wei *et al*., 2022; Wang *et al*., 2024; Gui *et al*., 2025; Huang *et al*., 2025; Ma *et al*., 2025). The mitochondrial genome size among species of the Malpighiales order varied considerably, ranging from 303,026 bp in *B. sexangula* to 838,420 bp in *P. alba*. We

identified two circular contigs that together constitute the mitochondrial genome of *Caryocar brasiliense*. Within the order Malpighiales, other species have also exhibited a multi-chromosomal mitochondrial structure. For example, *Populus tomentosa* C.K. Schneid (Salicaceae) possesses four mitochondrial chromosomes (Gao, Guo and An, 2025) while *Sapria himalayana* Griff. (Rafflesiaceae) has as many as 23 (Guo *et al*., 2023).

Although circular mitogenomes are often considered the most common configuration, linear and branched structures have also been reported in plants (Z. Q. Wu *et al*., 2022). For example *Uncaria rhynchophylla* (Miq.) Miq. ex Havil., has been described as having one circular and two linear mitochondrial chromosomes (Gui *et al*., 2025). In *Caryocar brasiliense,* we confirmed the presence of two mitochondrial chromosomes using graph visualization, which resulted in a circular graph, and the annotation of genes. We additionally inspected long-read coverage across the main junction supporting chromosome circularization. For each mitochondrial chromosome, an approximate 20 kb window spanning the junction (approximate 10 kb upstream and downstream) was examined, and the region was fully covered by Nanopore reads, with high mean depth (2,349× for chromosome 1 and 2,219× for chromosome 2). These results provide additional support for the circular representation of both mitochondrial chromosomes (Supplementary file S06). Nevertheless, we did not perform a systematic assessment of read spanning or coverage across all potential recombination breakpoints. Given that recombination-mediated isoforms are common in plant mitochondrial genomes, alternative linear or branched conformations may coexist but remain unresolved or underrepresented in the current assembly.

Given the extensive structural rearrangements, recombination, and incorporation of foreign DNA characteristic of plant mitochondrial genomes, the reference-guided pre-filtering of Nanopore and Illumina reads may introduce assembly biases, potentially leading to the underrepresentation of highly divergent or lineage-specific regions in *C. brasiliense*. Although the long read (longer than 1000 bp) could allow the capture of specific regions as long as they contain a conserved region, such as protein-coding regions. This strategy was necessary to reduce interference from abundant nuclear and plastid sequences; we cannot fully exclude the possibility that extremely divergent mitochondrial sequences were not captured during reference-based read selection. While long-read sequencing substantially

improves contiguity and enables the resolution of multipartite mitogenome structures compared to short-read approaches (Wu et al., 2022), alternative recombination-mediated conformations may coexist but remain unresolved or underrepresented in the current assembly.

Although our assembly and coverage analysis supports a bipartite circular representation of the *C. brasiliense* mitochondrial genome, it is important to recognize that plant mitochondrial genomes often exist in vivo as heterogeneous populations of molecules with multiple structural conformations (Kozik et al. 2019; Hao et al. 2024). Recombination mediated by repeated sequences can generate alternative isoforms, including linear and branched configurations, that coexist with circular forms (Yang et al. 2022). Our circular representation reflects the predominant topology detected in long-read sequencing data but does not exclude the existence of lower-abundance alternative conformations that may be functionally relevant. Future studies using complementary approaches, such as fluorescence microscopy or single-molecule analysis, could more completely characterize the structural dynamics of the *C. brasiliense* mitogenome.

The bipartite organization observed in plant mitogenomes may confer adaptive advantages, particularly by enabling greater control over gene expression, as distinct chromosomes can be replicated and transcribed independently (Wu et al., 2022). In addition, the abundance of repetitive sequences promotes intragenomic recombination, fostering isoform diversity and genomic plasticity (Gui et al., 2025). Future studies integrating reference-free assembly strategies, recombination-aware analyses, and deeper long-read sequencing will be essential to further elucidate the structural dynamics and lineage-specific mitochondrial genome architecture of Caryocaraceae.

All of the core protein-coding genes have been annotated on the *C. brasiliense* mitochondrial chromosome. In contrast, the ribosomal protein gene families and succinate dehydrogenase subunit genes (*sdh3* and *sdh4*) exhibit greater variability, both present in *C. brasiliense* mitogenome. The mitochondrial genome of plants typically contains a diverse gene set; however, a conserved core of approximately 24 protein-coding genes is usually maintained (Kan *et al*., 2021; Wei *et al*., 2022). These gene families have frequently been transferred to the nuclear genome or, in some cases, substituted by homologous genes from the chloroplast genome (Mower, Sloan, and Alverson, 2012; Mower, 2020; Zardoya, 2020).

We identified 28 functional cpDNA genes transferred to the mitogenome, including 25 protein-coding genes and three rRNA genes. These transferred genes may contribute to genomic instability by facilitating recombination and structural rearrangements, previously reported in other plant species (Jackman *et al*., 2016; Zardoya, 2020; Han *et al*., 2022; Wei *et al*., 2022; Qiao *et al*., 2024; Gui *et al*., 2025). Among the genes involved in MTPTs, eight were considered pseudogenes due to the presence of multiple stop codons within their sequences. All identified genes shared at least 80% similarity with their corresponding chloroplast genes, indicating that these sequences have undergone significant evolutionary changes, leading to a gradual loss of integrity.

Furthermore, certain cpDNA regions are known to be more susceptible to transfer to the mitochondrial genome, particularly those involving the genes *trnW-CCA*, *trnA-UGC*, *rpl2*, *rpl23*, *trnI-GAU*, *rps16*, *trnQ-UUG*, *psbK*, *psbI*, *trnS-GCU*, *trnG-UCC*, *trnS-UGA*, *psbZ*, *trnG-GCC*, *ndhF*, *rpl32*, *trnL-UAG*, *ccsA*, and *ndhD* (Wang *et al*., 2007; Nam, 2024). Regarding the origin of the MTPT regions in the *C. brasiliense* mitogenome, 32 originated from the Large Single Copy (LSC) region and 26 from the Inverted Repeat (IR), both previously described as more prone to transfer. In contrast, MTPTs originating from the Small Single Copy (SSC) region were less frequent (Nam, 2024).

The transfer of sequences between chloroplast and mitochondrial genomes is a frequent evolutionary event, contributing between 1% and 10% to the total size of the mitogenome (Wei *et al*., 2022; Jo *et al*., 2024). These transferred regions have been associated with genome expansion, recombination, and structural rearrangement events within plant mitogenomes (Sloan, Alverson, Wu, *et al*., 2012; Gualberto and Newton, 2017).

The proportion of plastid-derived DNA in the *C. brasiliense* mitogenome (~17% of the mitogenome) is notably high but falls within the wide variation observed in angiosperms. Recent studies have documented proportions ranging from ~1.5% in *Viburnum chinshanense* (Ma et al. 2022) to exceptional cases where plastid transfers represent 28-42% of the plastid genome incorporated into the mitogenome (*Paphiopedilum micranthum*, Yang et al. 2023; *Vitis vinifera*, Fang et al. 2012). This variation reflects lineage-specific differences in transfer rate, retention of transferred sequences, and recombination dynamics (Park et al. 2020). The high proportion in *C. brasiliense* may be related to recent transfer events or a reduced rate of loss of incorporated plastid sequences. Additional

comparative studies within Caryocaraceae and Malpighiales will be necessary to determine whether this feature is specific to *C. brasiliense* or represents a broader pattern in the family.

The presence of MTPTs highlights the complexity of plant mitochondrial genomes. This genomic complexity, driven by frequent inter-organellar DNA transfer events, distinguishes plant mitogenomes from those of other eukaryotic lineages, where such transfers are less common (Wang *et al*., 2007; Mower, Sloan, and Alverson, 2012; Mower, 2020; Zardoya, 2020). The high number of pseudogenes among the transferred cpDNA sequences also suggests that, although these transfers occur recurrently, many of the incorporated sequences gradually lose functionality over time due to the accumulation of mutations and structural rearrangements. Moreover, the preferential transfer of LSC and IR regions, along with the clustering of MTPTs in specific genomic areas, supports the hypothesis that recombination hotspots play a significant role in shaping the mitogenome structure throughout plant evolution.

The mechanisms underlying the transfer of mitochondrial DNA segments to the nuclear genome remain unclear. The complexity of both nuclear and mitochondrial genomes, along with limitations in available genomic data, makes the identification of these transferred segments particularly challenging. In *C. brasiliense*, the presence of 77 NUMT regions in the nuclear genome suggests that mitochondrial DNA transfer is a relatively frequent evolutionary event. Notably, most of the NUMTs identified, which are not associated with any annotated gene, support the hypothesis that many of these insertions may be functionally neutral and retained passively over time. Previous studies have shown that the number of NUMTs is often positively correlated with nuclear genome size and that these insertions tend to accumulate preferentially in intergenic regions (Ko and Kim, 2016; Park *et al*., 2025). Although we detected significant evidence of DNA sequence transfer between the nucleus and the chloroplast, our findings likely underestimate the true number of NUMTs. This is primarily due to the preliminary stage of the *C. brasiliense* draft nuclear genome, with only an estimated 49% of the genome currently assembled (Nunes, Gonçalves, and Telles, 2019). We therefore encourage further genomic and transcriptomic studies to improve the completeness and quality of the reference genome.

When compared to other Malpighiales species, the number of dispersed repeats and SSRs identified in *C. brasiliense* is intermediate, with higher counts than *B. sexangula* but lower than *M.*

esculenta and *P. alba*. Notably, in *C. brasiliense*, forward repeats were the most frequent type of dispersed repeat, and mononucleotide motifs predominated among SSRs, a pattern consistently observed across all analyzed species. This predominance of forward and mononucleotide repeats has also been widely reported in other angiosperm mitogenomes (Cai *et al*., 2024; Wang *et al*., 2024), suggesting that these repetitive elements can be used in the recombination process, contributing to mitogenome structural rearrangements and size variation (Gui *et al*., 2025; Huang *et al*., 2025; Ma *et al*., 2025). The chloroplast genome of *C. brasiliense* contains 49 dispersed repeats, including 18 forward, 30 palindromic, and one reverse repeat. Additionally, 85 SSRs were identified, with mononucleotide repeats being the most abundant (Nunes *et al.,* 2020). In comparison, the mitochondrial genome of *C. brasiliense* exhibits a higher repeat density than the chloroplast genome. Together with MTPT transfers, the accumulation and recombination of repeat sequences represent key drivers of plant mitogenome size evolution, shaping its dynamic and complex architecture.

RNA editing is an essential post-transcriptional process that ensures the accuracy of gene expression and the production of functional proteins. In plants, RNA editing plays crucial roles in various biological processes, including signal transduction, environmental adaptability, and the biosynthesis of essential components within mitochondria and chloroplasts (Chen *et al*., 2025). The number of RNA editing sites can vary across different plant tissues and species, with cytosine-to-uridine conversions being the most predominant type of editing event (Ouyang *et al*., 2025). Among mitochondrial genes, members of the cytochrome c maturation (*ccm*) gene family consistently exhibit the highest number of RNA editing sites, followed by *nad4* and *nad7* (Cai *et al*., 2024; Wang *et al*., 2024; Zhang, Liu, and Wei, 2024). In this work, the *nad4*, *ccmFn*, *ccmFc*, and *nad7* genes exhibited the highest number of RNA editing sites. These genes play essential roles in the mitochondrial respiratory chain and cytochrome c maturation, suggesting that RNA editing may modulate bioenergetic efficiency. Notably, an editing event in *ccmFc* converts an arginine codon into a stop codon (R→*), potentially regulating translation termination and influencing protein structure and function. The computational prediction of 376 C-to-U RNA editing sites provides testable hypotheses about post-transcriptional regulation in the *C. brasiliense* mitogenome. However, it is essential to recognize the limitations of predictive approaches: sequence-based methods do not fully capture the biological determinants of

editing, including trans-acting factors (PPR and auxiliary proteins) and transcript structural contexts (Gutmann et al. 2021; Brehme et al. 2015). Validation studies in other species show that while computational prediction is useful, confirmation rates vary. For example, Zheng et al. (2020) validated 87.5% (14 of 16) of predicted sites in *Salvia miltiorrhiza* using RT-PCR and Sanger sequencing, while other studies show variable rates depending on method and organism.

The potential functional consequences of predicted editing sites in energy metabolism genes, including possible changes in protein structure and translational efficiency, should be considered hypothetical until confirmed by transcriptomic analysis (strand-specific RNA-seq), cDNA sequencing, or proteomic analysis. Future studies combining RNA sequencing with functional analysis of mutants will be necessary to determine which predicted sites are actually edited in vivo and their consequences for mitochondrial function and plant fitness.

The RSCU values ranged from 0.25 (UAG - stop codon) to 2.65 (AGA - stop codon). RSCU values greater than 1 indicate a preference for specific codons under random usage conditions, and in this study, we observed a preference for codons ending with A and U, consistent with previous reports (Zhang, Liu and Wei, 2024; Hu *et al*., 2025; Ma *et al*., 2025). The amino acids leucine and serine were the most frequently used in *C. brasiliense*, and a preference for these codons in plant mitogenomes has also been reported in species of the genus *Scaevola* (Asterales) (Meng *et al*., 2025) and the genus *Cymbaria* (Orobanchaceae) (Ma *et al*., 2025). Understanding codon usage preference is essential for several biological processes, including protein structure, gene expression, genome evolution, and adaptability.

The position of the Caryocaraceae family in the phylogenetic tree for the order Malpighiales is not well defined. Studies based on chloroplast genomes have shown Caryocaraceae as a sister clade to the other families of the order (Nunes, de Souza, *et al*., 2020). However, the low support values observed in the chloroplast-based tree may be related to the limited number of species and genes analyzed. Increasing both taxa and gene sampling may help improve support for phylogenetic relationships more clearly. In our mitochondrial phylogenetic tree, the position of the Caryocaraceae family is supported by moderate bootstrap values. The complexity of the mitochondrial genome, combined with the lower availability of mitochondrial data, makes phylogenetic reconstruction particularly challenging.

In this study, we report the first complete mitochondrial genome of *Caryocar brasiliense*, revealing a bipartite structure and providing a detailed annotation and comparative analysis within Malpighiales. Our findings highlight the dynamic nature of plant mitochondrial genomes, including high levels of intergenomic transfer and extensive RNA editing. The results provide valuable insights into mitochondrial genome evolution and offer a new genomic resource for the conservation, phylogenetics, and genetic characterization of *C. brasiliense*. This type of study is significant for *C. brasiliense*, a species that remains underexplored despite its significant socio-economic importance.

Competing interests

The authors declare there are no competing interests.

Author contributions

L.R.C, R.N and M.P.C.T conceptualized the study. L.R.C and C.P.T sampled the biological material and made the wet laboratory procedures. L.R.C, L.C.J.C, R.N and J.A.F.D.F made bioinformatic analysis and prepared figures. R.O.D and R.S.B.F Writing – review & editing. R.N. and M.P.C.T. supervised the work. L.R.C wrote the manuscript. All authors reviewed the manuscript.

Funding

This work was supported by the National Institute of Science and Technology (INCT) in Ecology, Evolution, and Biodiversity Conservation funded by CNPq (grant 465610/2014-5/ 409197/2024-6) and FAPEG (grant 201810267000023). We would also like to express our gratitude to the PPBio Araguaia project, which is funded by CNPq (proc. 441114/2023-7) and to the Projeto Universal funded by CNPq (proc. 406597/2023-5), for their financial support and the Araguaia vivo project, carried out by the Tropical Alliance Water Research (TWRA) and financed by FAPEG (proc. 202210267000536). LRC was a scholarship supported by Programa de Demanda Social (DS-CAPES), and LCJC received a doctoral scholarship from FAPEG. JAFD-F and MPCT have been continuously supported by productivity grants from CNPq.

Data availability

All DNA sequence data generated in this study have been deposited in the NCBI. The assembly mtDNA was deposited at the GenBank database under the accession numbers PX396084 (Cbr_mt-Chr1) and PX396085 (Cbr_mt-Chr2). The raw reads generated in this study were deposited at the SRA database under the accession numbers SRR37028966 (Cbr_chr1) and SRR37028965 (Cbr_chr2).

References

Alverson, A.J. *et al.* (2010) 'Insights into the Evolution of Mitochondrial Genome Size from Complete Sequences of Citrullus lanatus and Cucurbita pepo (Cucurbitaceae)', *Molecular Biology and Evolution*, 27(6), p. 1436. Available at: https://doi.org/10.1093/MOLBEV/MSQ029.

Ballesteros-Mejia, L., Lima, J.S. and Collevatti, R.G. (2018) 'Spatially-explicit analyses reveal the distribution of genetic diversity and plant conservation status in Cerrado biome', *Biodiversity and Conservation*, pp. 1–18. Available at: https://doi.org/10.1007/s10531-018-1588-9.

BECK, N.; LANG, B. (2010) 'MFannot, organelle genome annotation websever', *Canada: Université de Montréal QC* [Preprint].

Benson, G. (1999) 'Tandem repeats finder: a program to analyze DNA sequences', *Nucleic Acids Research*, 27(2), pp. 573–580. Available at: https://doi.org/10.1093/nar/27.2.573.

Bogdanova, V.S. *et al.* (2021) 'Discordant evolution of organellar genomes in peas (Pisum L.)', *Molecular Phylogenetics and Evolution*, 160, p. 107136. Available at: https://doi.org/10.1016/j.ympev.2021.107136.

Cai, L. *et al.* (2021) 'The Perfect Storm: Gene Tree Estimation Error, Incomplete Lineage Sorting, and Ancient Gene Flow Explain the Most Recalcitrant Ancient Angiosperm Clade, Malpighiales', *Systematic Biology*. Edited by M. Fishbein, 70(3), pp. 491–507. Available at: https://doi.org/10.1093/sysbio/syaa083.

Cai, Y. *et al.* (2024) 'Repeat-mediated recombination results in Complex DNA structure of the mitochondrial genome of Trachelospermum jasminoides', *BMC Plant Biology*, 24(1), p. 966. Available at: https://doi.org/10.1186/s12870-024-05568-6.

Cao, P.-J. et al. (2023) 'De Novo Assembly and Comparative Analysis of the Complete Mitochondrial Genome of Chaenomeles speciosa (Sweet) Nakai Revealed the Existence of Two Structural Isomers',

564 *Genes*, 14(2), p. 526. Available at: https://doi.org/10.3390/genes14020526.

565 Castresana, J. (2000) 'Selection of Conserved Blocks from Multiple Alignments for Their Use in
566 Phylogenetic Analysis', *Molecular Biology and Evolution*, 17(4), pp. 540–552. Available at:
567 https://doi.org/10.1093/oxfordjournals.molbev.a026334.

568 Chen, Z. *et al.* (2025) 'Mitochondrial genome of Lonicera macranthoides: features, RNA editing, and
569 insights into male sterility', *Frontiers in Plant Science*, 15. Available at:
570 https://doi.org/10.3389/fpls.2024.1520251.

571 Chevigny, N. *et al.* (2020) 'DNA repair and the stability of the plant mitochondrial genome',
572 *International Journal of Molecular Sciences*, 21(1). Available at:
573 https://doi.org/10.3390/ijms21010328.

574 Choubey, A. and Rajam, M.V. (2015) 'Organellar genomes of flowering plants', in *Plant Biology and*
575 *Biotechnology: Volume II: Plant Genomics and Biotechnology*. New Delhi: Springer India, pp. 179–
576 204. Available at: https://doi.org/10.1007/978-81-322-2283-5_8.

577 Collevatti, R.G. *et al.* (2010) 'Spatial genetic structure and life history traits in cerrado tree species:
578 Inferences for conservation', *Natureza a Conservacao*, 8(1), pp. 54–59. Available at:
579 https://doi.org/10.4322/natcon.00801008.

580 Collevatti, R.G., Brondani, R.V. and Grattapaglia, D. (1999) 'Development and characterization of
581 microsatellite markers for genetic analysis of a Brazilian endangered tree species Caryocar brasiliense',
582 *Heredity*, 83(6), pp. 748–756. Available at: https://doi.org/10.1046/j.1365-2540.1999.00638.x.

583 Collevatti, R.G., Grattapaglia, D. and Hay, J.D. (2001) 'High resolution microsatellite based analysis
584 of the mating system allows the detection of significant biparental inbreeding in Caryocar brasiliense,
585 an endangered tropical tree species', *Heredity*, 86(1), pp. 60–67. Available at:
586 https://doi.org/10.1046/j.1365-2540.2001.00801.x.

587 Collevatti, R.G., Grattapaglia, D. and Hay, J.D. (2003) 'Evidences for multiple maternal lineages of
588 Caryocar brasiliense populations in the Brazilian Cerrado based on the analysis of chloroplast DNA
589 sequences and microsatellite haplotype variation', *Molecular Ecology*, 12(1), pp. 105–115. Available
590 at: https://doi.org/10.1046/j.1365-294X.2003.01701.x.

591 Collevatti, R.G., Nabout, J.C. and Diniz-Filho, J.A.F. (2011) 'Range shift and loss of genetic diversity

592 under climate change in Caryocar brasiliense, a Neotropical tree species', *Tree Genetics and Genomes*,
593 7(6), pp. 1237–1247. Available at: https://doi.org/10.1007/s11295-011-0409-z.
594 Diniz-Filho, J.A.F. *et al.* (2009) 'Niche modelling and landscape genetics of Caryocar brasiliense
595 ("Pequi" tree: Caryocaraceae) in Brazilian Cerrado: An integrative approach for evaluating central-
596 peripheral population patterns', *Tree Genetics and Genomes*, 5(4), pp. 617–627. Available at:
597 https://doi.org/10.1007/s11295-009-0214-0.
598 Doyle, J.J., & and Doyle, J.L. (1987) 'A Rapid DNA Isolation Procedure for Small Quantities of Fresh
599 Leaf Tissue', *Phytochemical Bulletin*, pp. 11–15. Available at:
600 https://webpages.uncc.edu/~jweller2/pages/BINF8350f2011/BINF8350_Readings/Doyle_plantDNAe
601 xtractCTAB_1987.pdf (Accessed: 1 June 2019).
602 Emms, D.M. and Kelly, S. (2019) 'OrthoFinder: phylogenetic orthology inference for comparative
603 genomics', *Genome Biology*, 20(1), p. 238. Available at: https://doi.org/10.1186/s13059-019-1832-y.
604 Gao, K., Guo, T. and An, X. (2025) 'Comprehensive analysis of the multi-rings mitochondrial genome
605 of Populus tomentosa', *BMC Genomics*, 26(1), p. 23. Available at: https://doi.org/10.1186/s12864-024-
606 11184-3.
607 Fang, Y. et al. (2012) 'A Complete Sequence and Transcriptomic Analyses of Date Palm (Phoenix
608 dactylifera L.) Mitochondrial Genome', *PLOS ONE*, 7(5), p. e37164. Available at:
609 https://doi.org/10.1371/JOURNAL.PONE.0037164.
610 Greiner, S., Lehwark, P. and Bock, R. (2019) 'OrganellarGenomeDRAW (OGDRAW) version 1.3.1:
611 Expanded toolkit for the graphical visualization of organellar genomes', *Nucleic Acids Research*,
612 47(W1), pp. W59–W64. Available at: https://doi.org/10.1093/nar/gkz238.
613 Gualberto, J.M. *et al.* (2014) 'The plant mitochondrial genome: Dynamics and maintenance',
614 *Biochimie*, 100(1), pp. 107–120. Available at: https://doi.org/10.1016/j.biochi.2013.09.016.
615 Gualberto, J.M. and Newton, K.J. (2017) 'Plant Mitochondrial Genomes: Dynamics and Mechanisms
616 of Mutation', *Annual Review of Plant Biology*, 68(1), pp. 225–252. Available at:
617 https://doi.org/10.1146/annurev-arplant-043015-112232.
618 Gui, L. *et al.* (2025) 'Mitogenome of Uncaria rhynchophylla: genome structure, characterization, and
619 phylogenetic relationships', *BMC Genomics*, 26(1), p. 199. Available at:

https://doi.org/10.1186/s12864-025-11372-9.

Guo, X. *et al.* (2023) 'The Sapria himalayana genome provides new insights into the lifestyle of endoparasitic plants', *BMC Biology*, 21(1), p. 134. Available at: https://doi.org/10.1186/s12915-023-01620-3.

Han, F. *et al.* (2022) 'Assembly and comparative analysis of the complete mitochondrial genome of Salix wilsonii using PacBio HiFi sequencing', *Frontiers in Plant Science*, 13(November), pp. 1–12. Available at: https://doi.org/10.3389/fpls.2022.1031769.

Hao, Z. et al. (2024) 'Complete mitochondrial genome of Melia azedarach L., reveals two conformations generated by the repeat sequence mediated recombination', *BMC Plant Biology,* 24(1), p. 635. Available at: https://doi.org/10.1186/s12870-024-05319-7.

Hu, H. *et al.* (2023) 'Genome-scale angiosperm phylogenies based on nuclear, plastome, and mitochondrial datasets', *Journal of Integrative Plant Biology*, 65(6), pp. 1479–1489. Available at: https://doi.org/10.1111/jipb.13455.

Hu, Q. *et al.* (2025) 'Assembly and analysis of the complete mitochondrial genome of Leonurus japonicus (Lamiaceae)', *Scientific Reports*, 15(1), p. 13372. Available at: https://doi.org/10.1038/s41598-025-97594-z.

Huang, Yutong *et al.* (2025) 'The first complete mitochondrial genome assembly and comparative analysis of the fern Blechnaceae family: Blechnopsis orientalis', *Frontiers in Plant Science*, 16. Available at: https://doi.org/10.3389/fpls.2025.1534171.

Jackman, S.D. *et al.* (2016) 'Organellar Genomes of White Spruce (Picea glauca): Assembly and Annotation', *Genome Biology and Evolution*, 8(1), pp. 29–41. Available at: https://doi.org/10.1093/gbe/evv244.

Jo, S. *et al.* (2024) 'Intracellular gene transfer (IGT) events from the mitochondrial genome to the plastid genome of the subtribe Ferulinae Drude (Apiaceae) and their implications', *BMC Plant Biology*, 24(1), p. 1172. Available at: https://doi.org/10.1186/s12870-024-05891-y.

Kan, S.-L.L. *et al.* (2021) 'Both Conifer II and Gnetales are characterized by a high frequency of ancient mitochondrial gene transfer to the nuclear genome', *BMC Biology*, 19(1), p. 146. Available at: https://doi.org/10.1186/s12915-021-01096-z.

Katoh, K., Rozewicki, J. and Yamada, K.D. (2018) 'MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization', *Briefings in Bioinformatics*, 20(4), pp. 1160–1166. Available at: https://doi.org/10.1093/bib/bbx108.

Ko, Y.-J. and Kim, S. (2016) 'Analysis of Nuclear Mitochondrial DNA Segments of Nine Plant Species: Size, Distribution, and Insertion Loci', *Genomics & Informatics*, 14(3), p. 90. Available at: https://doi.org/10.5808/GI.2016.14.3.90.

Kozik, A. et al. (2019a) 'The alternative reality of plant mitochondrial DNA: one ring does not rule them all', bioRxiv. Available at: https://doi.org/10.1101/564278.

Kong, J., Wang, J., Nie, L. *et al.* (2025). Evolutionary dynamics of mitochondrial genomes and intracellular transfers among diploid and allopolyploid cotton species. *BMC biology*, v. 23, n. 1, p. 9. Available at: https://doi.org/10.1186/s12915-025-02115-z

Laslett, D. and Canback, B. (2004) 'ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences', *Nucleic Acids Research*, 32(1), p. 11. Available at: https://doi.org/10.1093/NAR/GKH152.

Lenz, H. and Knoop, V. (2013) 'PREPACT 2.0: Predicting C-to-U and U-to-C RNA editing in organelle genome sequences with multiple references and curated RNA editing annotation', *Bioinformatics and Biology Insights*, 7. Available at: https://doi.org/10.4137/BBI.S11059.

Leu, K.-C. *et al.* (2016) 'Distinct role of Arabidopsis mitochondrial P-type pentatricopeptide repeat protein-modulating editing protein, PPME, in nad1 RNA editing', *RNA Biology*, 13(6), pp. 593–604. Available at: https://doi.org/10.1080/15476286.2016.1184384.

Li, H. (2018) 'Minimap2: pairwise alignment for nucleotide sequences', *Bioinformatics*. Edited by I. Birol, 34(18), pp. 3094–3100. Available at: https://doi.org/10.1093/bioinformatics/bty191.

Li, J. *et al.* (2021) 'Assembly of the complete mitochondrial genome of an endemic plant, Scutellaria tsinyunensis, revealed the existence of two conformations generated by a repeat-mediated recombination', *Planta*, 254(2), p. 36. Available at: https://doi.org/10.1007/s00425-021-03684-3.

Lowe, T.M. and Eddy, S.R. (1997) 'tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence', *Nucleic Acids Research*, 25(5), pp. 955–964. Available at:

https://doi.org/10.1093/nar/25.5.955.

Ma, Q. et al. (2022) 'Assembly and comparative analysis of the first complete mitochondrial genome of Acer truncatum Bunge: a woody oil-tree species producing nervonic acid', *BMC Plant Biology*, 22(1), p. 29. Available at: https://doi.org/10.1186/s12870-021-03416-5.

Ma, Y. *et al.* (2025) 'Complete mitochondrial genomes of the hemiparasitic genus Cymbaria (Orobanchaceae): insights into repeat-mediated recombination, phylogenetic relationships, and horizontal gene transfer', *BMC Genomics*, 26(1), p. 314. Available at: https://doi.org/10.1186/s12864-025-11474-4.

Manly, B. F. J. (2020). *Randomization, Bootstrap and Monte Carlo Methods in Biology*. 2nd ed. Chapman & Hall, London.

Marçais, G. *et al.* (2018) 'MUMmer4: A fast and versatile genome alignment system', *PLOS Computational Biology*. Edited by A.E. Darling, 14(1), p. e1005944. Available at: https://doi.org/10.1371/journal.pcbi.1005944.

Meng, D. *et al.* (2025) 'Organelle genomes of two Scaevola species, S. taccada and S. hainanensis, provide new insights into evolutionary divergence between Scaevola and its related species', *Frontiers in Plant Science*, 16. Available at: https://doi.org/10.3389/fpls.2025.1587750.

Mower, J.P. (2020) 'Variation in protein gene and intron content among land plant mitogenomes', *Mitochondrion*, 53(June), pp. 203–213. Available at: https://doi.org/10.1016/j.mito.2020.06.002.

Mower, J.P., Sloan, D.B. and Alverson, A.J. (2012) 'Plant Mitochondrial Genome Diversity: The Genomics Revolution', in *Plant Genome Diversity Volume 1*. Vienna: Springer Vienna, pp. 123–144. Available at: https://doi.org/10.1007/978-3-7091-1130-7_9.

Nascimento-Silva, N.R.R., Mendes, N.S.R. and Silva, F.A. (2020) 'Nutritional composition and total phenolic compounds content of pequi pulp (Caryocar brasiliense Cambess.)', *Journal of Bioenergy and Food Science*, pp. 1–10. Available at: https://doi.org/10.18067/jbfs.v7i2.281.

NCBI (2025) *Organelle - NCBI - NLM*. Available at: https://www.ncbi.nlm.nih.gov/datasets/organelle/?taxon=2759 (Accessed: 9 June 2025).

Nhat Nam, N., Pham Anh Thi, N. and Do, H.D.K. (2024) 'New Insights into the Diversity of Mitochondrial Plastid DNA', *Genome Biology and Evolution*. Edited by G. Piganeau, 16(9). Available

703 at: https://doi.org/10.1093/gbe/evae184.

704 Nunes, R., de Lima, N.E., *et al.* (2020) 'Caryocaraceae Voigt (Malpighiales): a Synthesis Based on
705 Science Mapping and Systematic Review', *The Botanical Review*, 86(3–4), pp. 338–358. Available at:
706 https://doi.org/10.1007/s12229-020-09233-z.

707 Nunes, R., de Souza, U.J.B., *et al.* (2020) 'Complete chloroplast genome sequence of caryocar
708 brasiliense camb. (caryocaraceae) and comparative analysis brings new insights into the plastome
709 evolution of Malpighiales', *Genetics and Molecular Biology*, 43(2), pp. 1–7. Available at:
710 https://doi.org/10.1590/1678-4685-GMB-2019-0161.

711 Nunes, R., Gonçalves, A.R. and Pires de Campos Telles, M. (2019) 'Data on the draft genome sequence
712 of Caryocar brasiliense Camb. (Caryocaraceae): An important genetic resource from Brazilian
713 savannas', *Data in Brief*, 26, p. 104543. Available at: https://doi.org/10.1016/j.dib.2019.104543.

714 Orme, D. *et al.* (2023) 'Comparative Analyses of Phylogenetics and Evolution in R [R package caper
715 version 1.0.3]', *CRAN: Contributed Packages* [Preprint]. Available at:
716 https://doi.org/10.32614/CRAN.PACKAGE.CAPER.

717 Ouyang, L. *et al.* (2025) 'Comprehensive analysis of the mitochondrial genome of Iris domestica
718 emphasizing multichromosomal organization and repeat-mediated homologous recombination',
719 *Frontiers in Plant Science*, 15. Available at: https://doi.org/10.3389/fpls.2024.1520033.

720 Paradis E, Claude J, and Strimmer K. (2004). APE: Analyses of Phylogenetics and Evolution in R
721 language. *Bioinformatics* 20(2), pp. 289-90. Available at: https//doi: 10.1093/bioinformatics/btg412.
722 PMID: 14734327.

723 Park, S. *et al.* (2025) 'Insights into the nuclear-organelle DNA integration in Cicuta virosa (Apiaceae)
724 provided by complete plastid and mitochondrial genomes', *BMC Genomics*, 26(1), p. 102. Available
725 at: https://doi.org/10.1186/s12864-025-11230-8.

726 Qiao, H. *et al.* (2024) 'Assembly and comparative analysis of the first complete mitochondrial genome
727 of Salix psammophila, a good windbreak and sand fixation shrub', *Frontiers in Plant Science*, 15.
728 Available at: https://doi.org/10.3389/fpls.2024.1411289.

729 Robinson, D. F. and Foulds, R. L. (1981) Comparison of phylogenetic trees. *Mathematical Biosciences*,
730 53, pp. 131–147. Available at: https://doi:10.1016/0025-5564(81)90043-2

731 Sagan, L (1967) 'On the origin of mitosing cells'. *Journal of Theoretical Biology*, 14, pp. 225-IN6.
732 Available at: https://doi.org/10.1016/0022-5193(67)90079-3.
733 Shan, Y. et al. (2023) 'Assembly of the Complete Mitochondrial Genome of Pereskia aculeata Revealed
734 That Two Pairs of Repetitive Elements Mediated the Recombination of the Genome', *International*
735 *Journal of Molecular Sciences*, 24(9), p. 8366. Available at: https://doi.org/10.3390/ijms24098366.
736 Skippingtona, E. *et al.* (2015) 'Miniaturized mitogenome of the parasitic plant viscum scurruloideum
737 is extremely divergent and dynamic and has lost all nad genes', *Proceedings of the National Academy*
738 *of Sciences of the United States of America*, 112(27), pp. E3515–E3524. Available at:
739 https://doi.org/10.1073/pnas.1504491112.
740 Sloan, D.B., Alverson, A.J., Chuckalovcak, J.P., *et al.* (2012) 'Rapid evolution of enormous,
741 multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates',
742 *PLoS Biology*, 10(1). Available at: https://doi.org/10.1371/journal.pbio.1001241.
743 Sloan, D.B., Alverson, A.J., Wu, M., *et al.* (2012) 'Recent Acceleration of Plastid Sequence and
744 Structural Evolution Coincides with Extreme Mitochondrial Divergence in the Angiosperm Genus
745 Silene', *Genome Biology and Evolution*, 4(3), pp. 294–306. Available at:
746 https://doi.org/10.1093/gbe/evs006.
747 Souza, H.A.V. e *et al.* (2017) 'A large historical refugium explains spatial patterns of genetic diversity
748 in a Neotropical savanna tree species', *Annals of Botany*, 119(2), pp. 239–252. Available at:
749 https://doi.org/10.1093/aob/mcw096.
750 Tillich, M. *et al.* (2017) 'GeSeq - Versatile and accurate annotation of organelle genomes', *Nucleic*
751 *Acids Research*, 45(W1), pp. W6–W11. Available at: https://doi.org/10.1093/nar/gkx391.
752 Vieira, R.F., Camillo, J. and Coradin, L. (2016) *Espécies nativas da flora brasileira de valor econômico*
753 *atual ou potencial: plantas para o future – Região centro-oeste.* Available at: http://www.b4fn.org/the-
754 countries/brazil.
755 Walker, B.J. *et al.* (2014) 'Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection
756 and Genome Assembly Improvement', *PLoS ONE*. Edited by J. Wang, 9(11), p. e112963. Available at:
757 https://doi.org/10.1371/journal.pone.0112963.
758 Wang, D. *et al.* (2007) 'Transfer of Chloroplast Genomic DNA to Mitochondrial Genome Occurred At

759 Least 300 MYA', *Molecular Biology and Evolution*, 24(9), pp. 2040–2048. Available at:
760 https://doi.org/10.1093/molbev/msm133.
761 Wang, L. *et al.* (2024) 'Comparative analysis of the mitochondrial genomes of four Dendrobium species
762 (Orchidaceae) reveals heterogeneity in structure, synteny, intercellular gene transfer, and RNA editing',
763 *Frontiers in Plant Science*, 15. Available at: https://doi.org/10.3389/fpls.2024.1429545.
764 Wei, L. *et al.* (2022) 'Comparative analyses of three complete Primula mitogenomes with insights into
765 mitogenome size variation in Ericales', *BMC Genomics*, 23(1), p. 770. Available at:
766 https://doi.org/10.1186/s12864-022-08983-x.
767 Wick, R.R. *et al.* (2017) 'Unicycler: Resolving bacterial genome assemblies from short and long
768 sequencing reads', *PLOS Computational Biology*. Edited by A.M. Phillippy, 13(6), p. e1005595.
769 Available at: https://doi.org/10.1371/journal.pcbi.1005595.
770 Wong, T. *et al.* (2025) 'IQ-TREE 3: Phylogenomic Inference Software using Complex Evolutionary
771 Models'. Available at: https://doi.org/10.32942/X2P62N.
772 Wu, Z. *et al.* (2022) 'Genomic architectural variation of plant mitochondria—A review of
773 multichromosomal structuring', *Journal of Systematics and Evolution*, 60(1), pp. 160–168. Available
774 at: https://doi.org/10.1111/jse.12655.
775 Wu, Z.Q. *et al.* (2022) 'Genomic architectural variation of plant mitochondria—A review of
776 multichromosomal structuring', *Journal of Systematics and Evolution*, 60(1), pp. 160–168. Available
777 at: https://doi.org/10.1111/jse.12655.
778 Yang, J.-X. et al. (2023) 'Multichromosomal Mitochondrial Genome of Paphiopedilum micranthum:
779 Compact and Fragmented Genome, and Rampant Intracellular Gene Transfer', *International Journal of
780 Molecular Sciences*, 24(4), p. 3976. Available at: https://doi.org/10.3390/ijms24043976.
781 Yang, Z. et al. (2022) 'De novo assembly of the complete mitochondrial genome of sweet potato
782 (Ipomoea batatas [L.] Lam) revealed the existence of homologous conformations generated by the
783 repeat-mediated recombination', BMC Plant Biology, 22(1), p. 285. Available at:
784 https://doi.org/10.1186/s12870-022-03665-y.
785 Zardoya, R. (2020) 'Recent advances in understanding mitochondrial genome diversity',
786 *F1000Research*, 9, p. 270. Available at: https://doi.org/10.12688/f1000research.21490.1.

787 Zhang, J., Liu, G. and Wei, J. (2024) ‘Assembly and comparative analysis of the first complete
788 mitochondrial genome of Setaria italica’, *Planta*, 260(1), p. 23. Available at:
789 https://doi.org/10.1007/s00425-024-04386-2.