



OPEN

Complete chloroplast genomes of six neotropical palm species, structural comparison, and evolutionary dynamic patterns

Ana Flávia Francisconi^{1,11}, Jonathan Andre Morales Marroquín^{1,11}, Luiz Augusto Cauz-Santos², Cássio van den Berg³, Kauanne Karolline Moreno Martins¹, Marcones Ferreira Costa^{1,4}, Doriane Picanço-Rodrigues⁵, Luciano Delmodes de Alencar¹, Cesar Augusto Zanello¹, Carlos Augusto Colombo⁶, Brenda Gabriela Díaz Hernández⁶, Danilo Trabuco Amaral⁷, Maria Teresa Gomes Lopes⁸, Elizabeth Ann Veasey⁹ & Maria Imaculada Zucchi¹⁰✉

The Arecaceae family has a worldwide distribution, especially in tropical and subtropical regions. We sequenced the chloroplast genomes of *Acrocomia intumescens* and *A. totai*, widely used in the food and energy industries; *Bactris gasipaes*, important for palm heart; *Copernicia alba* and *C. prunifera*, worldwide known for wax utilization; and *Syagrus romanzoffiana*, of great ornamental potential. *Copernicia* spp. showed the largest chloroplast genomes (*C. prunifera*: 157,323 bp and *C. alba*: 157,192 bp), while *S. romanzoffiana* and *B. gasipaes* var. *gasipaes* presented the smallest (155,078 bp and 155,604 bp). Structurally, great synteny was detected among palms. Conservation was also observed in the distribution of single sequence repeats (SSR). *Copernicia* spp. presented less dispersed repeats, without occurrence in the small single copy (SSC). All RNA editing sites were C (cytidine) to U (uridine) conversions. Overall, closely phylogenetically related species shared more sites. Almost all nodes of the phylogenetic analysis showed a posterior probability (PP) of 1.0, reaffirming the close relationship between *Acrocomia* species. These results elucidate the conservation among palm chloroplast genomes, but point to subtle structural changes, providing support for the evolutionary dynamics of the Arecaceae family.

The Arecaceae family comprises 188 genera and 2585 species. Its distribution is cosmopolitan and worldwide, and it is concentrated mainly in tropical and subtropical regions¹⁻³. Commonly known as palms, approximately > 800 species are distributed throughout the Neotropical region and, in South America alone, 437 species occur, belonging to 50 genera, among which 18 are endemic⁴. Palms are one of the dominant species groups in the tropical rainforest (TRF). Also, the Arecaceae family presents high species richness, and is a key component in

¹Programa de Pós-Graduação em Genética e Biologia Molecular, Universidade Estadual de Campinas, R. Monteiro Lobato, 255-Barão Geraldo, Campinas, São Paulo CEP 13083-862, Brazil. ²Department of Botany and Biodiversity Research, University of Vienna, Rennweg 14, 1030 Wien, Austria. ³Departamento de Ciências Biológicas, Universidade Estadual de Feira de Santana, Av. Transnordestina S/N-Novo Horizonte, Feira de Santana/Feira de Santana, Bahia CEP 44036-900, Brazil. ⁴Universidade Federal do Piauí, BR-343 Km 3.5, Floriano, Piauí CEP 64808-605, Brazil. ⁵Departamento de Biologia, Universidade Federal do Amazonas, Avenida Gen. Rodrigo Octávio Jordão Ramos, 3000-Coroado I-Campus Universitário-Senador Arthur Virgílio Filho-Setor Sul, Bloco H, Manaus, Amazonas CEP 69077-000, Brazil. ⁶Instituto Agronômico, Av. Theodureto de Almeida Camargo, 1500, Campinas, São Paulo CEP 13075-630, Brazil. ⁷Departamento de Biologia, Centro de Ciências Humanas e Biológicas, Universidade Federal do ABC, Avenida dos Estados, 5001, Santo André, São Paulo CEP 09040-040, Brazil. ⁸Faculdade de Ciências Agrárias, Universidade Federal do Amazonas, Avenida Rodrigo Otávio Ramos, 3000-Bairro Coroado, Manaus, Amazonas CEP 69077-000, Brazil. ⁹Departamento de Genética, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Avenida Pádua Dias, 11-Bairro São Dimas, Piracicaba, São Paulo CEP 13418-900, Brazil. ¹⁰Agência Paulista de Tecnologia dos Agronegócios (APTA), Polo Centro Sul, Rodovia SP 127 Km 30, CP 28, Piracicaba, São Paulo CEP 13400-970, Brazil. ¹¹These authors contributed equally: Ana Flávia Francisconi and Jonathan Andre Morales Marroquín. ✉email: mizucchi@apta.sp.gov.br

the evolution and diversification of hyper diversity in the TRF biome^{5,6}. Palms are part of the TRF ecosystem's services. For example, they are a base resource for frugivory and pollinators, and may have influenced the diversification of dependent animal groups. Palms also have a shared history with ancestral human groups providing food, construction materials, fuel, and ornamentals^{7,8}. According to Huang et al.⁹, Arecaceae species constitute the third most economically important family in the world.

Among palm species, the genus *Acrocomia* Mart. (subtribe Bactridinae, tribe Cocoseae) is widely distributed in the Neotropics and is popularly known as Macaúba (Macaw palm) or Coyal. *Acrocomia* includes eight species, of which *Acrocomia aculeata* (Jacq.) Lodd. Ex Mart., *A. intumescens* Drude, and *A. totai* Mart. have great economic interest, considering that all vegetative structures can be used¹⁰. The three species have high oil content and are promising for the energy and food industries, thus presenting native alternative sources for food and biodiesel production^{11,12}. Besides, new products can be exploited, such as *A. totai*, for example, which is commonly used as food in its habitat¹².

Bactris gasipaes Kunth. (Bactridinae), commonly known as Pupunha (Peach palm), was an important pre-Columbian subsistence product, used for fruit consumption and its stipe woody structures for hunting, fishing, and agricultural equipment. Nowadays, the fruit is commercialized in its natural state or traditionally used in regional dishes at Amazonian markets. *B. gasipaes* is essential to supply the production chain of high-quality heart-of-palm¹³. Another Neotropical palm genus is *Copernicia* (subtribe Livistoniinae, tribe Trachycarpeae), comprising 21 species mainly distributed in North America, Greater Antilles, with only three occurring in South America¹⁴. In Brazil, *Copernicia alba* Morong ex Morong & Britton, popularly known as Carandá (Caranda palm), is economically important due to the quality of its wood, the consumption of heart-of-palm in the food industry, and the ornamental potential¹⁵. *Copernicia prunifera* (Mill.) H.E. Moore, known as Carnaúba (Carnaúba palm), is a worldwide source of quality carnauba wax^{16,17} and can also be used in cosmetics, skin care, and candy coating¹⁸. Both palms are endemic to Brazilian biomes, the Pantanal and the Caatinga, respectively. *Syagrus romanzoffiana* (Cham.) Glassman (subtribe Attaleinae, tribe Cocoseae), popularly known as Coqueiro or Jerivá (Queen palm or Cocos palm), has ornamental potential and fruits with high nutritional content, used as food sources for animals. The oil obtained from this species is applicable to the food, cosmetics, pharmaceutical, and biodiesel industries¹⁹. Finally, it is important to highlight that the six chosen Neotropical palm species also provide key ecosystem services that ensure welfare for indigenous people in South America^{8,20,21}. Due to their growing economic importance and traditional use, continued efforts to obtain genetic and genomic resources are essential to gather information about these species and thus plan their uses in a sustainable way.

The rapid development of next-generation sequencing (NGS) technology has made it cheaper and more accessible to obtain complete chloroplast sequences, even for an increasing number of non-models species^{22–25}. The chloroplast is a semi-autonomous organelle in plant cells, with its own genetic information, a complete genetic system, and its genome is typically uniparentally inherited in the absence of recombination²⁶. The chloroplast genome usually has four parts: two single copy regions, a large one (LSC), a small one (SSC), and a pair of inverted regions (IRs)²⁷. This genome consists basically of long circular or linear molecules (120–180 kb), containing 120–130 genes with functions for photosynthesis, transcription, and translation. The genome structure and gene content are, in general, highly conserved in land plants²⁸, including palm species²⁹.

For palm species occurring in Brazil, 12 complete chloroplast genomes have been reported^{2,29–36}. These genomes are used for structural comparisons, phylogenetic inferences, the determination of RNA editing sites, and the identification of genomic regions with different repeats and polymorphisms. Among the advantages of elucidating new palm chloroplast genomes are increasing our understanding of evolutionary processes in Arecaceae and deciphering phylogenetic relationships between closely related taxa²⁸. Furthermore, with the available sequences, it will be possible to identify repetitive sequences and regions with high polymorphism and screen for molecular markers. These resources are extremely important for assessing genetic structure and diversity in natural populations³⁴.

Considering the importance and necessity of increasing genomic information about these Neotropical palm species, here we report the complete chloroplast genomes of *A. intumescens*, *A. totai*, *B. gasipaes* var. *gasipaes*, *C. alba*, *C. prunifera* and *S. romanzoffiana*. With the complete chloroplast genomes, we were able to investigate: (1) gene content; (2) structural comparisons and the synteny level with other chloroplast genomes available from Arecaceae species; (3) single sequence repeats and dispersed repeats; (4) RNA editing sites and their conservancy among the six palm species; and (5) the Arecaceae family evolution based on a phylogenomic study with six new complete chloroplast genomes.

Results

Organization and gene features of the six palms chloroplast genomes

The chloroplast genomes of the six palm species had the typical quadripartite structure with the presence of two single copies (large single copy—LSC, small single copy—SSC; Fig. 1a–f), and two inverted repeat regions (IRA and IRB). The species from the *Copernicia* genus had the largest chloroplast genome size (*C. prunifera*: 157,323 bp and *C. alba*: 157,192 bp; Fig. 1d,e, Table 1), and *S. romanzoffiana* had the smallest size (155,078 bp; Fig. 1f). The species from the genus *Copernicia* also displayed the longest LSC (*C. alba*: 86,430 bp and *C. prunifera*: 86,264 bp). However, *Acrocomia intumescens* presented the largest SSC (17,522 bp). Finally, the species *A. intumescens*, *A. totai*, and *B. gasipaes* var. *gasipaes* showed IRs with more nucleotides (bp) than the species of the genus *Copernicia* and *S. romanzoffiana*. All species had GC content (%) between 37.2 and 38.0, with a similar distribution (%) of GC content in the four regions of the complete chloroplast genome (Table 1).

The species presented variation in the number of genes (130–131, considering duplicate copies; Table 1). By comparing the structures of the six chloroplast genomes, it is possible to identify that *C. alba* and *C. prunifera* had two copies of the *trnM*—CAU and one copy of the *trnG*—UCC, instead of the two *trnI*—CAU and one

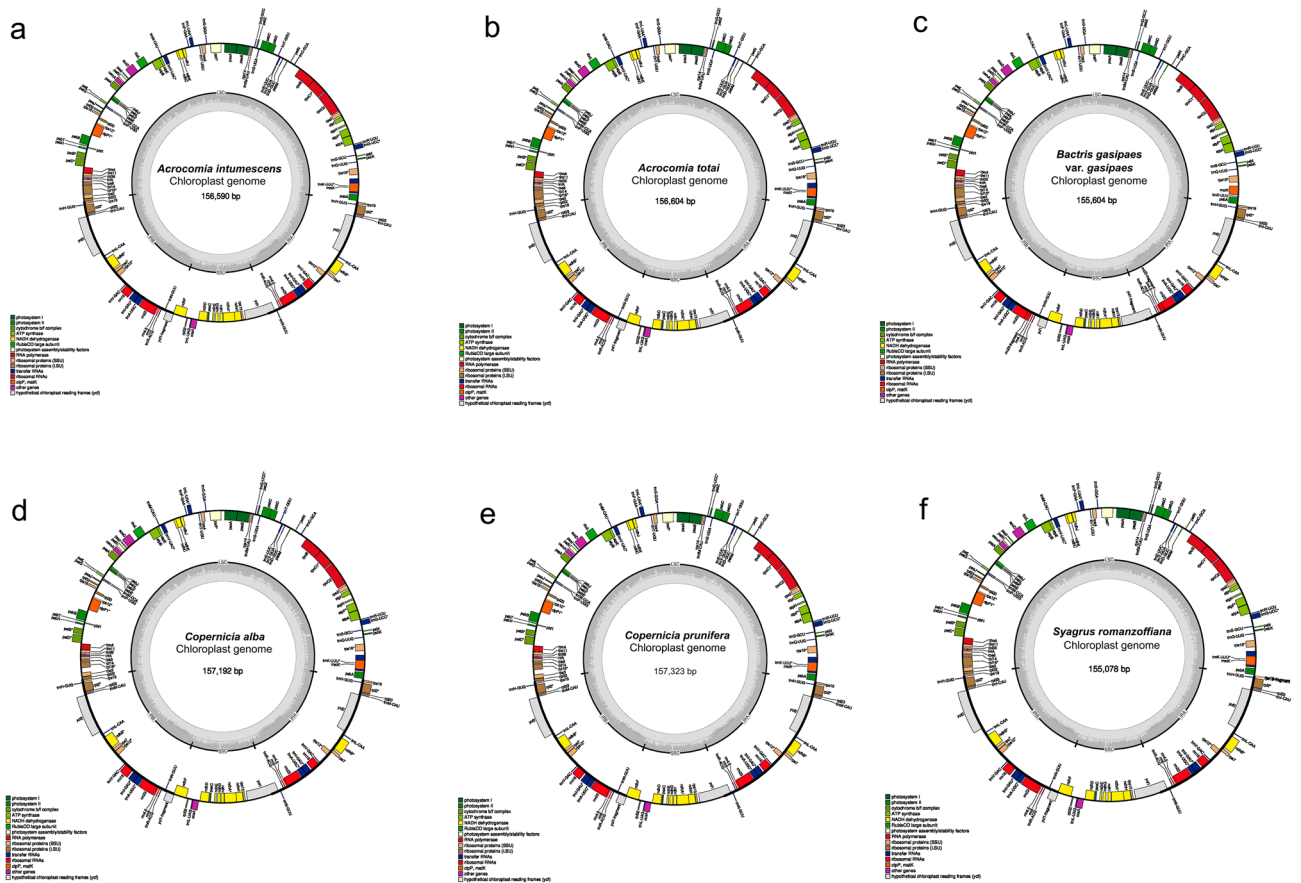


Figure 1. Gene map of (a) *Acrocomia intumescens*, (b) *A. totai*, (c) *Bactris gasipaes* var. *gasipaes*, (d) *Copernicia alba*, (e) *C. prunifera*, and (f) *Syagrus romanzoffiana* chloroplast genomes. Genes represented inside the large circle are oriented clockwise, and those outside are oriented counterclockwise. The distinct colors represent functional groups, and the darker gray in the inner circle indicates the GC content. The quadripartite structure is also reported as: *LSC* large single copy, *SSC* small single copy, *IRA* inverted repeat A, *IRB* inverted repeat B.

Gene features	<i>A. intumescens</i>	<i>A. totai</i>	<i>B. gasipaes</i> var. <i>gasipaes</i>	<i>C. alba</i>	<i>C. prunifera</i>	<i>S. romanzoffiana</i>
NCBI accession number	OQ129926	OQ129927	OQ129928	OQ129929	OQ129930	OQ129931
Total cpDNA size (bp)	156,590	156,604	155,604	157,192	157,323	155,078
LSC length (bp)	85,058	85,060	84,193	86,430	86,264	84,730
SSC length (bp)	17,522	17,360	17,362	17,440	17,475	17,473
IR length (bp)	27,068	27,092	27,024	26,661	26,792	26,437
Total GC content (%)	37.4	37.4	37.5	37.2	37.2	38.0
LSC GC content (%)	35.5	35.5	35.5	35.1	35.2	35.5
SSC GC content (%)	31.2	31.2	31.2	30.8	30.8	31.1
IR GC content (%)	42.5	42.5	42.6	42.5	42.6	42.7
Total number of genes	131	131	131	131	131	130
Protein-coding genes	85	85	85	85	85	84
rRNA genes	8	8	8m	8	8	8
tRNA genes	38	38	38	38	38	38

Table 1. Gene features in the *Acrocomia intumescens*, *A. totai*, *Bactris gasipaes* var. *gasipaes*, *Copernicia alba*, *C. prunifera*, and *Syagrus romanzoffiana* chloroplast genomes according to each respective category. *SSC* small single copy, *LSC* large single copy, *IRA* inverted repeat A, *IRB* inverted repeat B.

trnG—GCC, respectively, that were present in the four chloroplast genomes of the species from the subfamily Arecoideae, analyzed in this study (Fig. 1). Also, *S. romanzoffiana* presented one gene reduction when compared with the other species of the subfamily Arecoideae, *rps19*, which in this case was presented as a fragment (Fig. 1).

The chloroplast genome structures and comparative analyses among species from different genera

The six Brazilian palm chloroplast genomes annotated in this study showed a high level of synteny in their structures (Fig. 2). Although the species are from different subfamilies (*C. alba*, *C. prunifera*, and *Trithrinax brasiliensis* Mart., subfamily Coryphoideae; *Mauritia flexuosa* L.f., subfamily Calamoideae; and the others from the subfamily Arecoideae), few structural rearrangements were observed. This structural conservation was also noted in the comparison of palm chloroplast genomes with five subfamilies (see Supplementary Fig. S1). The most notable change was in *Astrocaryum aculeatum* G. Mey. and *A. murumuru* Mart., which showed flip-flop recombination (in lime green; Fig. 2)³². As already pointed out, a significant dissimilarity among the species of the subfamily Arecoideae that occur in Brazil can be identified in the length of the LSC, between 40,000 and 50,000 bp³⁶. Furthermore, between the species *C. alba* and *C. prunifera* of the subfamily Coryphoideae, it was also possible to observe a reduction in LSC size (Table 1, Fig. 2). Among the new six palm chloroplast genomes, *S. romanzoffiana* had the smallest SSC length, which was also noted in the structure of *S. coronata* (Mart.) Becc.

Expansions/contractions can be observed throughout all the chloroplast genome structures of *C. alba* and *C. prunifera*, even though they are species of the same genus (Fig. 3). *C. prunifera* showed a shift in *rpl22* and *rps19* genes at the LSC/IRB margin compared to *C. alba*. The same shift can also be found in *ycf1* genes between IRB/LSC and SSC/IRA, and in *rps19* and *psbA* genes between IRA/LSC. Another very divergent species with expansions/contractions in the IR was *S. romanzoffiana*. The species presented the *rps19* gene between the LSC/IRB, while in the other analyzed palms, this gene was located completely in the IRB. In addition, the copy of *rps19* is transformed into a fragment at the boundary between the IRA/LSC.

SSR and dispersed repeats in the chloroplast genome structures

Among the six palm species, a total of 516 simple sequence repeats (SSRs) were identified (Fig. 4a, Supplementary Table S1). Seventy-two and 69 SSRs were identified in *A. intumescens* and *A. totai*, respectively. In *B. gasipaes* var. *gasipaes*, 88 SSRs were found. A higher amount of SSRs was observed in *C. alba* and *C. prunifera* (104 and 100 SSRs, respectively). *S. romanzoffiana* presented 83 SSRs in its chloroplast genome. All six palms had a higher number of mononucleotides of the SSR type, followed by di- and tetranucleotides, mostly concentrated in the LSC region of their chloroplast genome (Fig. 4b, Supplementary Table S1). Also, all species presented the motifs A/T, AT/AT, AAAT/ATTT, AAT/ATT, AAAAT/ATTTT, AATG/ATTC, AG/CT, and AGAT/ATCT. Some motifs were unique to certain palm species, such as AATACT/AGTATT, AAG/CTT, and AAAG/CTT, occurring only in *S. romanzoffiana*, and AAAAAT/ATTTTT, appearing only in *B. gasipaes* (Fig. 4c).

The total number of dispersed repeats (F = forward, P = palindrome, R = reverse, and C = complement) was very similar in *A. intumescens* (43), *A. totai* (43), and *B. gasipaes* var. *gasipaes* (41; Supplementary Table S2, Fig. 5a). The smallest values of these repeats were observed in *S. romanzoffiana* (35), *C. prunifera* (31), and *C. alba* (11). *A. intumescens*, *A. totai*, and *B. gasipaes* also presented a similar distribution of repeats with a higher concentration of palindromes (P; 22–20), followed by the forward type (F; 15). Likewise, *C. prunifera* and *S. romanzoffiana* showed higher concentrations of the palindrome (15 and 20, respectively) and forward types (12 and 8, respectively). Inversely, *C. alba* had a higher number of forward (seven repeats) types, followed by palindrome types (two repeats; Fig. 5a). All studied species presented the highest number of dispersed repeats concentrated in the LSC region (Fig. 5b), while only *C. alba* and *C. prunifera* showed any dispersed repeats in the SSC region of the chloroplast genome. *A. intumescens*, *A. totai*, and *B. gasipaes* had the highest number of dispersed repeats with a size of 30 bp (14, 14, and 12, respectively; Fig. 5c). *C. alba* and *C. prunifera* had the greatest number of repeats with longer lengths, 53 bp (four repeats, *C. alba*) and 49 bp (5 repeats, *C. prunifera*). *S. romanzoffiana*, on the other hand, had the highest number of repeats of an intermediate size of 37 bp (7 repeats).

Prediction of RNA editing sites in the chloroplast genes of the six palm species

Considering all changes, such as nucleotide and amino acid positions and conversions, 102 RNA editing sites were identified and shared among the six palm species (Fig. 6a). All RNA editing sites analyzed showed conversion of the nucleotides from cytidine (C) to uridine (U; Supplementary Table S3). The conversion occurred at the second (79.41%) and first (20.59%) codon positions. From the 102 RNA editing sites, 45 corresponded to changes from serine (S) to leucine (L), followed by 15 changes from proline (P) to leucine (L), and from histidine (H) to tyrosine (Y) (Supplementary Fig. S2a). Thus, the conversions resulted in 82.35% hydrophilic to hydrophobic amino acids (Supplementary Fig. S2b).

Among the 102 RNA editing sites, the six species shared 54 conversions. Between *C. alba* and *C. prunifera*, 20 conversions occurred exclusively (Fig. 6a). Also, among *A. intumescens*, *A. totai*, *B. gasipaes* var. *gasipaes*, and *S. romanzoffiana*, 19 conversions were observed. The number of RNA editing sites, considering pairwise species analyses, resulted in *S. romanzoffiana* × *B. gasipaes* sharing 78 sites, followed by *A. intumescens* × *A. totai* (77), and *C. alba* × *C. prunifera* (75; Fig. 6b). The species that shared the lowest number of conversions belonged to *Acrocomia* and *Copernicia*: *C. prunifera* × *A. intumescens* (54), *C. prunifera* × *A. totai* (54), *C. alba* × *A. intumescens* (54), and *C. alba* × *A. totai* (54). The *ndh* genes were the ones with the highest rate of changes, 38 in total (see Supplementary Table S3). However, except for the *ndhD* and *ndhF* genes, the modifications were the same for all six palms. The *matK* gene showed 11 conversions in RNA editing sites, considering the six palms. Even without considering the changes caused by alignments of the chloroplast genomes, it was observed that one conversion

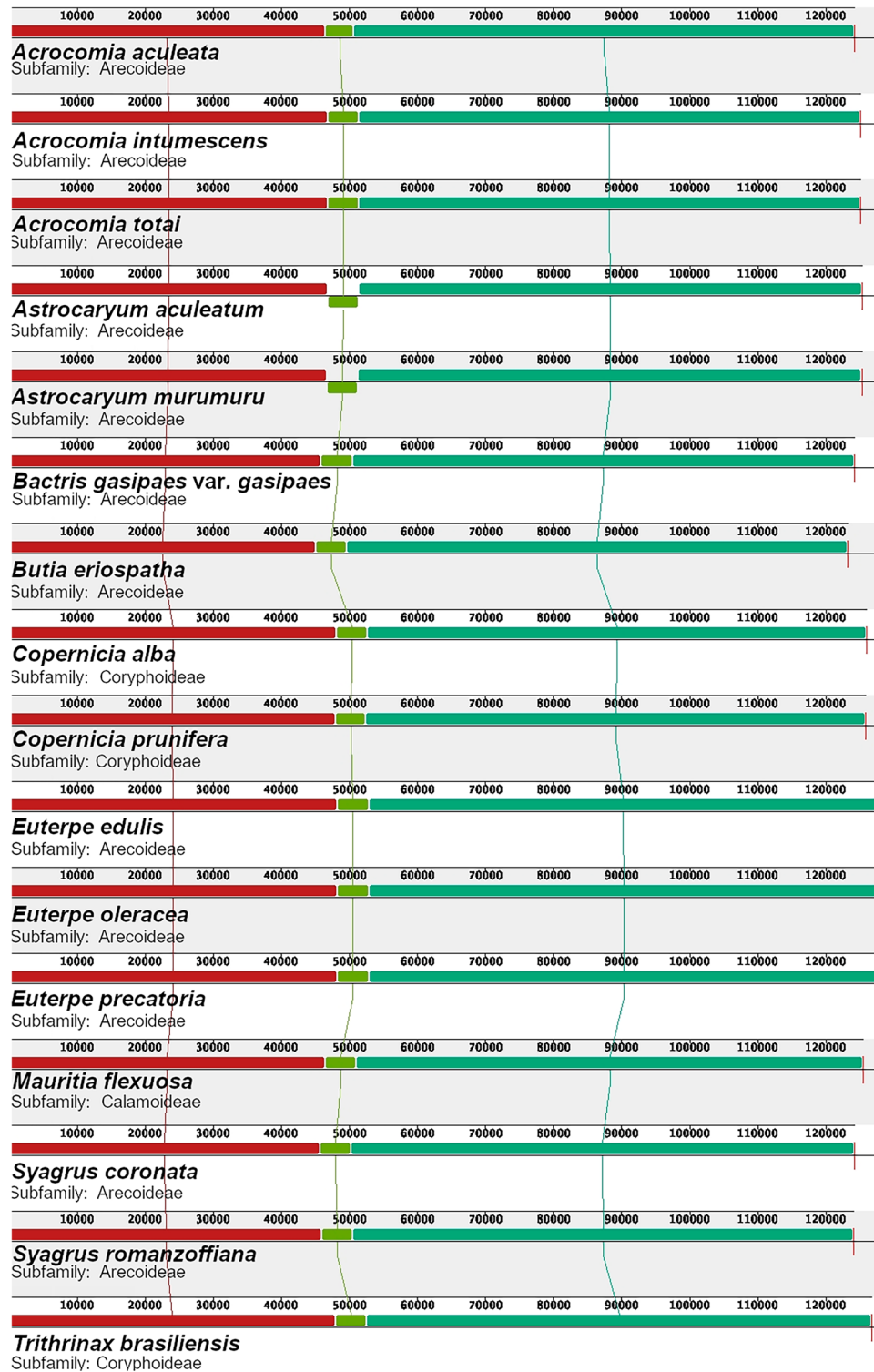


Figure 2. Synteny and divergence in the small single copy (SSC) size detected in Arecaceae chloroplast genomes using the Mauve multiple-genome alignment program. A sample of 16 different chloroplast genomes is shown. Color bars indicate syntenic blocks, and the lines indicate the correspondence between them. Blocks on the top row are in the same orientation, while blocks on the bottom row are in the opposite orientation.

was exclusive to *Acrocomia* species and another to *Copernicia* palms. Similarly, two conversions were not exclusively detected in these species.

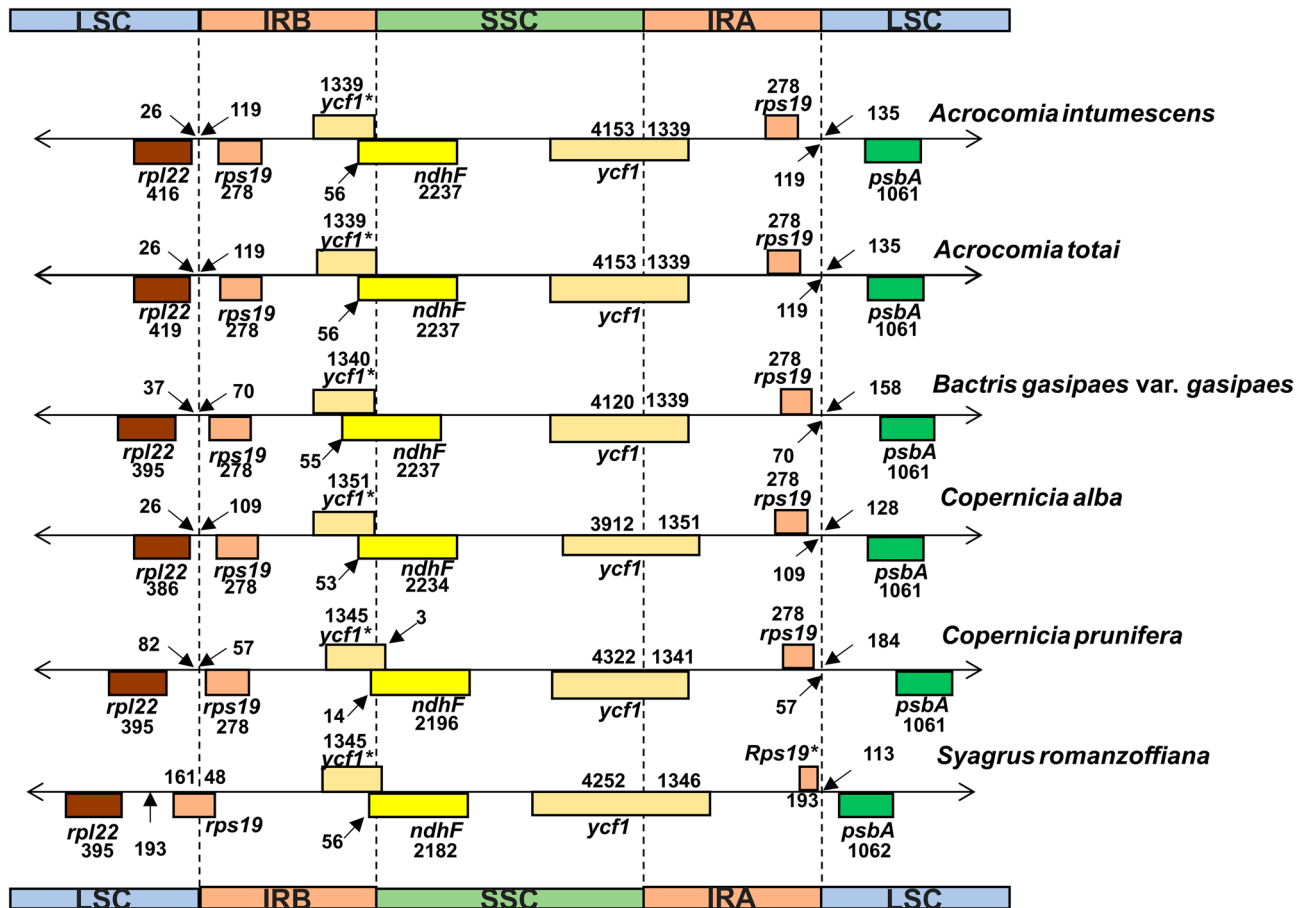


Figure 3. Comparison of the inverted repeats A and B (IRA and IRB) borders among Brazilian palms. The numbers indicate the lengths of IGSSs, genes, and spacers between IR-LSC and IR-SSC junctions. The *ycf1** and *rps19** genes have incomplete CDSs. *LSC* large single copy, *SSC* small single copy.

Phylogenomic studies

The full phylogenetic analysis with the selected partition scheme produced a tree (Fig. 7) in which all nodes have a posterior probability of 1.0 (PP = 1.0), except the node that represents the sister relationship between the clade with *Mauritia flexuosa* + *Eremospatha macrocarpa* Schaedtler and the clade containing *Salacca* Reinw., *Metroxylon* Rottb., *Pigafetta* (Blume) Becc., and *Calamus* L. (PP = 0.92). Regarding the species with new chloroplast genome sequences generated in the current study, *A. intumescens* and *A. totai* were sister to each other, and then sister to *A. aculeata*. *B. gasipaes* var. *gasipaes* was positioned as sister to two species of *Astrocaryum*, and *S. romanzoffiana* was sister to the previously sequenced *S. coronata* (Mart.) Becc. The species of *Copernicia* were sister to each other and positioned in a small clade, with *Pritchardia* Seem. & H. Wendl. and *Colpotherinax* Griseb. & H. Wendl. being more closely related to the former.

Discussion

Considering the subfamily level, significant conservation in the genetic content and genomic structures of the chloroplast genome in the analyzed species was observed. The chloroplast genome of *B. gasipaes* showed a small variation in total size from the chloroplast genome already published³⁵, from 156,646 to 155,078 bp. This result can probably be caused by differences in the origin and domestication status of the sampled individuals of *B. gasipaes*. The chloroplast genome assembled by da Silva et al.³⁵ was from a wild variety, *Bactris gasipaes* var. *chichagui*. In contrast, the individual of *B. gasipaes* var. *gasipaes* used in our research was domesticated and of Peruvian origin. It is also important to highlight that the sequencing method and assembly program applied to each of the varieties were different, which may also result in variations in the assembly size, as already observed for species from the genus *Euterpe*³⁶. Thus, more extensive sampling, together with the information provided by the assembly of chloroplast genomes, may provide support for distinguishing individuals of different origins and domestication.

Multiple alignments with 24 different palms (Supplementary Fig. S1) from all five subfamilies did not demonstrate any major rearrangements in the chloroplast structure. The only rearrangement observed was a 4.6 kb inversion in the *Astrocaryum* chloroplast genome (Fig. 2), which is probably lineage-specific³². Regarding the patterns of IR structure, most of the expansion/contraction-related variations in the six new chloroplast genomes were identified. It has been previously suggested that chloroplast DNA comprises four equimolar isomers, with

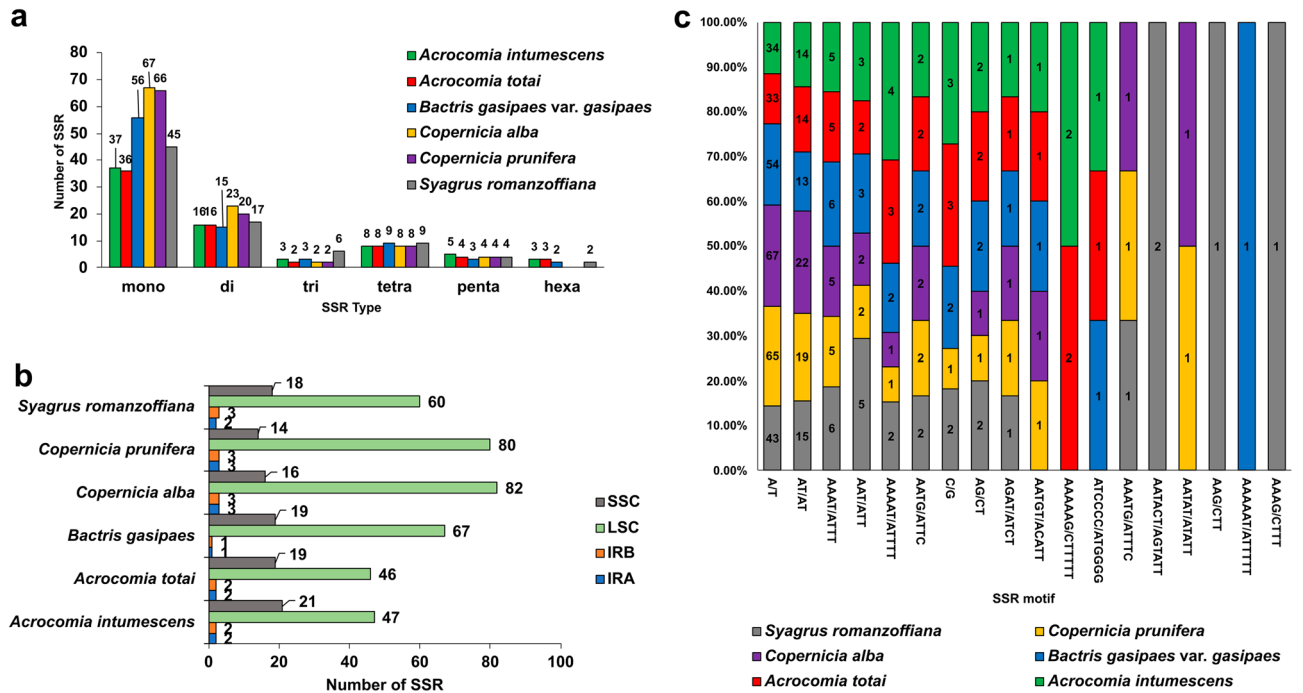


Figure 4. Distribution, classification, and motifs of single sequence repeats (SSR) in the chloroplast genomes of *Acrocomia intumescens*, *A. totai*, *Bactris gasipaes* var. *gasipaes*, *Copernicia alba*, *C. prunifera*, and *Syagrus romanzoffiana*. (a) Number of SSR types (mono-, di-, tri-, tetra-, penta-, and hexanucleotides) present in the six chloroplast genomes; (b) number of SSR in the different chloroplast genome regions; (c) number of different SSR motifs distributed in the six chloroplast genomes. SSC small single copy, LSC large single copy, IRA inverted repeat A, IRB inverted repeat B.

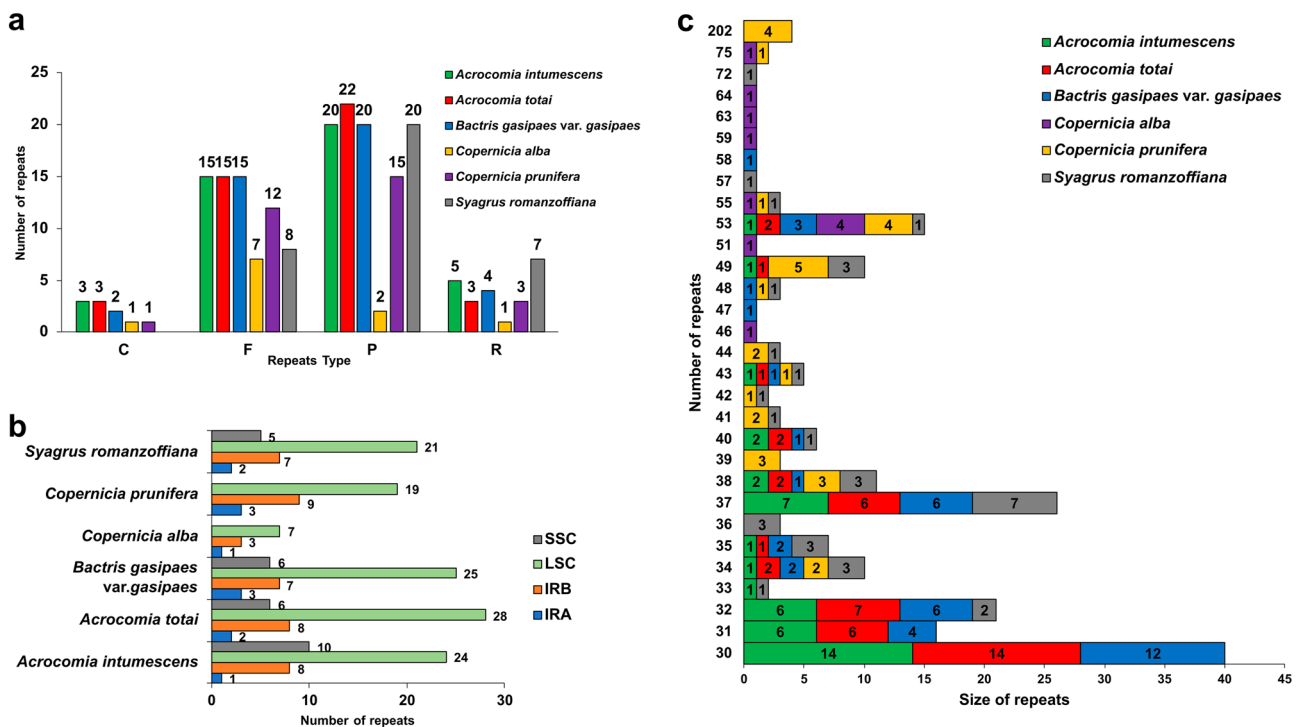


Figure 5. Distribution and classification of dispersed repeats in the chloroplast genomes of *Acrocomia intumescens*, *A. totai*, *Bactris gasipaes* var. *gasipaes*, *Copernicia alba*, *C. prunifera*, and *Syagrus romanzoffiana*. (a) Frequency distribution of different types of repeats; (b) number of dispersed repeats present in different chloroplast genome regions; (c) number of dispersed repeat sizes among the six palm species. F forward, P palindrome, R reverse, C complement, SSC small single copy, LSC large single copy, IRA inverted repeat A, IRB inverted repeat B.

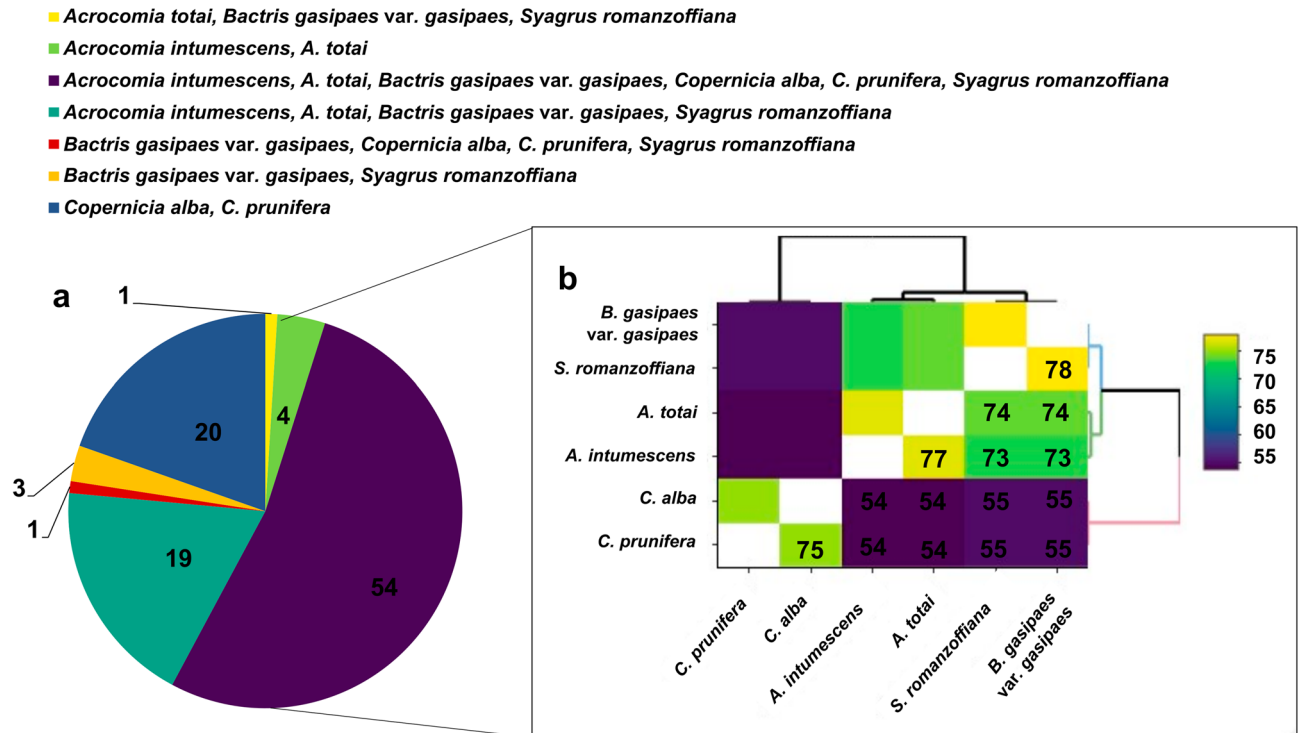


Figure 6. RNA editing sites shared by the species *Acrocomia intumescens*, *A. totai*, *Bactris gasipaes* var. *gasipaes*, *Copernicia alba*, *C. prunifera*, and *Syagrus romanzoffiana*. **(a)** Number of RNA editing sites shared among the six species; **(b)** Pairwise comparison of RNA editing sites found between the species.

the LSC potentially exhibiting either the same or opposite direction. Additionally, variations at the edges of each chloroplast genome region may arise from chloroplastial DNA replication^{37,38}. The most notable difference was with the species *S. romanzoffiana*. This species had a reduced copy of the *rps19* gene compared to the other species of the Arecoideae subfamily (Table 1, Fig. 3), now classified as a fragment. Two hypotheses can be proposed concerning this change: the first hypothesis suggests that this fragmentation event was a consequence of changes induced by the replication mechanism of the chloroplast genome³⁷, but without affecting the functional role of the *rps19* gene copy. However, an alternative hypothesis has been proposed regarding the assembly of the complete chloroplast genome of *Cocos nucifera* L.³⁵. In this species, the presence of a *rps19* fragment was considered a putative pseudogene. Recently, in the chloroplast genome of *Butia eriospatha* (Mart. ex Drude) Becc., the absence of the duplicated *rps19* gene was also observed². Although not discussed, this same event was likewise found in the complete chloroplast genome of *Syagrus coronata*³⁰, which, like *S. romanzoffiana*, showed a copy of *rps19* with the same size, 193 bp. The four species are closely related, as a phylogeny based on chloroplast genome sequences identified that *C. nucifera*, *S. coronata*, and *B. eriospatha* shared the same branch and can be classified as sisters in the subtribe Attaleinae of the tribe Cocoseae³⁵. Considering the close phylogenetic relationship between the species, such a process should be further investigated.

Although they are considered highly conserved structures for stabilizing chloroplast genome structure³⁹, fluctuations caused by contractions/expansions of the IRs have already been reported in palms^{31,36}. The most surprising case was the total loss of one IR in the species *Tahina spectabilis* J.Dransf. & Rakotoarin., instead of rearrangement, as this structure was considered canonical in monocots²⁹. Since it was a minor fluctuation in IR regions, the reduction of *rps19* in *C. nucifera*, *B. eriospatha*, and *Syagrus* species can be considered moderate in evolutionary terms. However, given that this was a deviation from a highly conserved structure, this information helps to reinforce the perspective raised by Barret et al.²⁹, that these changes may be more prevalent than was previously hypothesized, given that the same change was detected in different species of the same subtribe.

As for the number of SSRs, they were found to be quite conserved among species of the same genus. Besides the number, the distribution of SSR types (mono-, di-, tri-, tetra-, penta-, and hexanucleotides; Fig. 4) was highly similar, especially the mono- and dinucleotides. In all species, the mononucleotide A/T was the most frequent motif. This characteristic was also reported in other palm complete chloroplast genomes^{31,32,36}. Plastid SSR markers, especially when combined with nuclear markers, can improve the efficiency of studies of structure, diversity, and gene flow in natural populations⁴⁰. Furthermore, plastid SSRs are efficient markers for differentiate highly related species. Variations in the numbers of SSRs based on the chloroplast genomes of *Cerriops* species (Rhizophoraceae) reinforced the differences between *C. tagal*, *C. decandra*, and *C. zippeliana*²⁵. Thus, the sequences identified are a valuable resource for knowledge leveling and sustainable planning for species management.

Similarities in the number and distribution of dispersed repeats were also observed for *A. intumescens*, *A. totai*, and *B. gasipaes* var. *gasipaes*. These parallels could be detected in both the number and distribution of repetition types (Fig. 5). *S. romanzoffiana* showed a higher number of reverse than forward repeats, unlike the

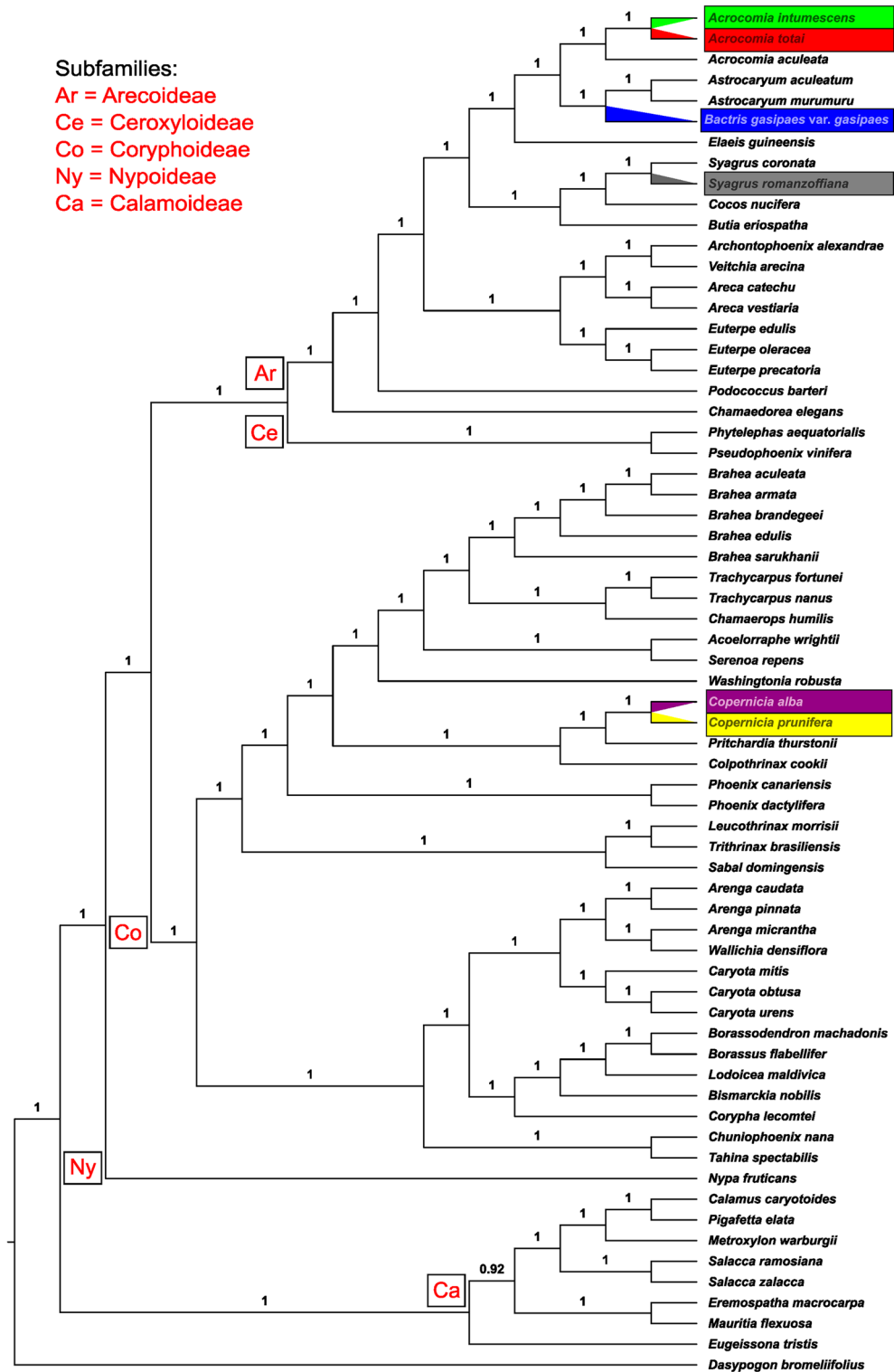


Figure 7. Majority-rule consensus tree of 30,000 trees obtained from a Bayesian inference analysis of chloroplast protein-coding genes of 66 taxa. Posterior probabilities (PP) for each are indicated above the branches.

first three species. In contrast, the species that differed the most were those from the genus *Copernicia*. The two species differed in both the distribution and size of the dispersed repeats compared to the other species belonging to the subfamily Arecoideae. In particular, the two species of *Copernicia* also differed from each other.

Although at a lower frequency, different numbers of dispersed repeats have already been reported in other land plants. In trees of the genus *Morus* L. (Family: Moraceae), only 14 dispersed repeats were detected, and of these, the species *M. multicaulis* Perr. and *M. cathayana* Hemsl. shared only four⁴¹. Among seven species of the genus *Polystachya* Hook., there were a variation of two (*P. dendrolliflora* Rchb.f.) and eight (*P. modesta* Rchb.f.) dispersed repeats. Variations in these repetitions may indicate flexibility in the evolutionary process⁴², and this may be considered an indicator of diversity among species, which is important for the phylogeny of the genus⁴¹.

The post-transcriptional process originated from RNA editing may induce the occurrence of substitutions or indels, which can result in transcript alterations. Editing events usually increase the hydrophobicity of encoded amino acids. In general, the conversions are from serine to leucine/phenylalanine (hydrophilic to hydrophobic) and frequently appear in interfaces to benefit interactions^{39,43}. As identified in the six palms, an example of a hydrophobic benefit was observed in the *ndh* complex genes, which encode membrane-binding polypeptides responsible for the transfer of NADH to plastoquinone. The resulting level of hydrophobicity of the new RNA editing sites increases the stability of these trans-membrane proteins⁴³. The *matK* gene is known for its rapid evolution and has been a favorite for determining phylogenetic relationships in angiosperms⁴⁴. With the chloroplast genome of the six palms, it was possible to observe the dynamics of the RNA editing sites in this gene and identify the appearance of unique conversions in certain species. Some authors have already reported that the rapid evolution of *matK* creates a selective pressure favoring C-to-T mutations, which reflects in the loss of RNA editing sites⁴⁴. The addition of new chloroplast genomes might clarify what is leading to the loss or appearance of these sites.

Despite a few exceptions, RNA editing is evolutionary conserved, and it is expected that more related taxa will have more editing sites in common⁴³. Of the 102 RNA editing sites, 54 were shared among the six palm species (Fig. 6a). Also, species from the same subfamily have more RNA editing sites in common (Coryphoideae: 20 and Arecoideae: 19; Fig. 6b). In addition to the subfamily hierarchy, the species *A. intumescens*, *A. totai*, *B. gasipaes*, and *S. romanzoffiana* are from the same tribe, Cocoseae, and the species from the genera *Acrocomia* and *Bactris* also participate in the same subtribe, Bactridinae⁴¹. This was reflected in the fact that they shared more than 70 RNA editing sites. Similarly, the species of the genus *Copernicia*, being phylogenetically close, shared 75 RNA editing sites.

In the phylogeny by Meerow et al.⁴⁶, based on WRKY nuclear genes, *Acrocomia* was sister to *Astrocaryum* and then both to *Bactris* + *Desmoncus*, however with low support, and both clades in a polytomy with *Aiphanes* Willd. There is no chloroplast genome available for either *Desmoncus* Mart. or *Aiphanes*, but our phylogeny indicates with PP = 1.0 that *Bactris* is sister to *Astrocaryum* and both to *Acrocomia* (Fig. 7). This relationship is also in agreement with the whole chloroplast genome phylogeny of Silva et al.³⁵, which included their sequence of the *Bactris* chloroplast genome, and the phylogenomic supertree-based study of Barret et al.²⁹. The relationship among the *Acrocomia* species indicated *A. totai* as sister to *A. intumescens* and then both sisters to *A. aculeata*. This is also supported by the phylogenetic and biogeographic studies with different evolutionary models showing that *Acrocomia* and *Aiphanes* were the first Bactridinae genera to irradiate in the late Eocene before the final uplift of the Andes (late Miocene and Pliocene)⁴⁷. In the population genomics study based on Genotyping by Sequencing data from Díaz et al.¹⁰, *A. totai* was closer to *A. aculeata* (in a genetic distance dendrogram), whereas *A. intumescens* was more divergent in relation to this pair. On the other hand, our results reaffirm the close relationship between *A. aculeata* and *A. intumescens*, also found by Meerow et al.⁴⁶, and with leaf anatomy by Vianna et al.⁴⁸, although some authors still considered the species synonyms^{5,49}. It is important to point out that *A. aculeata* has the widest distribution among the Neotropical palms, from central Mexico to northern Argentina¹⁰. *Acrocomia* still has unshed phylogenetic relationships due to its hybridization process, domestication, and possible ecotypes because a complete genus phylogeny is not yet available, including a significant number of botanical holotypes and natural samples comprising all its distribution.

Finally, the position of *Syagrus romanzoffiana* with *S. coronata* is expected, and its sister relationship to *Cocos* and then both to *Butia* agrees with the Cocosoid Palms by Meerow et al.⁴⁶, and the phylogenomic study of Barret et al.²⁹. The position of *Copernicia* is quite distant from the previous terminals since they belong to the subfamily Coryphoideae. Our phylogeny indicated a sister relationship between *Copernicia* spp. and *Pritchardia*, and between both and *Colpothrinax*, similar to Barret et al.²⁹.

Conclusion

Among the six Neotropical palms studied, gene content and chloroplast genome structure were highly conserved. Some punctual changes were observed, such as different tRNAs in *Copernicia* palm species in relation to the other four palms from the subfamily Arecoideae. Multiple alignments with palms from all Arecaceae subfamilies revealed no major rearrangements in chloroplast structure. Regarding the IRs, most of the variations were expansion/contraction. A reduction of the *rps19* gene copy was observed in *S. romanzoffiana*, a moderate but prevalent change in evolutionary terms, as it was identified in other species of the tribe Cocoseae. Conservation in the number of SSRs was also detected in palms of the same genus. However, some motifs occurred only in *B. gasipaes* var. *gasispaes* and *S. romanzoffiana*. The sequences identified are valuable for obtaining genetic markers. Similarities in the number and distribution of dispersed repeats were observed in the subfamily Arecoideae. Nevertheless, these repeats presented a distinct pattern in *Copernicia* palms, considered an indicator of diversity. For RNA editing, it was observed that related taxa had more sites in common, as expected. Finally, our phylogeny presented high support, and the positioning of the six palms was like those found in previous studies. Especially in the case of the genus *Acrocomia*, our results restated the close relationship between *A. aculeata* and *A. intumescens*. Overall, a highly consolidated pattern was identified among palm chloroplast genomes, with subtle structural changes. Resources were provided for evolutionary analyses of the Arecaceae family as well as for species conservation studies.

Methods

Species, sampling, and DNA extraction

To characterize the geographical occurrence of the six palm species, we surveyed the databases using the BIEN⁵¹ R package⁵². *Bactris gasipaes* is dispersed in Central and South America, while *Acrocomia totai*, *Copernicia alba*, and *Syagrus romanzoffiana* occur in South America. *Acrocomia intumescens* and *C. prunifera* have records only in the Brazilian Northeast Region (see Supplementary Fig. S3). All palms in this study were wild plants, except for *B. gasipaes*. This palm was an introduction of domesticated seed from Yurimáguas, Peru, in the 1970s, also known as *Bactris gasipaes* var. *gasipaes*.

The leaves were dried with silica gel and stored in a freezer at $-20\text{ }^{\circ}\text{C}$. The leaf material from *A. intumescens*, *A. totai*, *B. gasipaes* var. *gasipaes*, *C. alba*, and *C. prunifera* was obtained in the active germplasm bank of the Plant Genetic Resources Center of the Agronomic Institute of Campinas (IAC), Campinas, SP, Brazil (geographical coordinates: -22.8717 , -47.0776). *S. romanzoffiana* was sampled from the ex situ collection at the IAC in Piracicaba, Piracicaba-SP, Brazil (geographical coordinates: -22.6836 , -47.6458). This study complies with relevant institutional, national, and international guidelines and legislation. The appropriate permissions for the collection of plant material were taken, and the collections were registered according to the National System for the Management of Genetic Heritage and Associated Traditional Knowledge (SISGEN), as stated by Brazilian Decree No. 8,772 (May 11, 2016) and regulated by Brazilian Law No. 13,123 (May 20, 2015; SISGEN numbers: A411583 and A9BEE40, Brazil). The voucher numbers are shown in Supplementary Table S4 and are available at the IAC Herbarium.

To extract the chloroplast organelles, a sucrose gradient method was used to isolate them⁵³. For this, 20 g of fresh leaves from each species were frozen with liquid nitrogen and macerated. The material was resuspended in 200 mL of isolation buffer (50 mM Tris-HCl pH 8.0, 0.35 M sucrose, 7 mM EDTA, 5 mM 2-mercaptoethanol, and 0.1% BSA) and incubated for 10 min in the dark. The suspension was filtered using two layers of Miracloth (Merck), and then the filtrate was centrifuged at $1000\times g$ for 10 min.

The pellet was resuspended in 5 mL of isolation buffer. The suspension was placed in the density gradient column of 20/45% sucrose in 50 mM Tris-HCl (pH 8.0), 0.3 M sorbitol, and 7 mM EDTA. After the centrifugation at $2000\times g$ (30 min), the green band formed at the interface containing intact chloroplasts was collected. The solution with the chloroplasts was then diluted in three volumes of buffer and centrifuged at $3000\times g$ (10 min) to obtain a pellet with purified chloroplasts.

The pellet was then resuspended in 2% CTAB buffer to initiate lysis. The suspension was incubated and stirred at $65\text{ }^{\circ}\text{C}$ for 1 h. The supernatant was extracted twice with an equal volume of chloroform: isoamyl alcohol (24:1) and centrifuged at $10,000\times g$ (20 min). The same volume of isopropanol was added and incubated at $20\text{ }^{\circ}\text{C}$ for 1 h. Lastly, the aqueous phase was centrifuged at $10,000\times g$ (20 min). The chloroplast DNA (cpDNA) pellet was washed with ethanol (70%), dried, and resuspended with 40 μL TE (1 M Tris-HCl, 0.5 M EDTA, pH 8).

Chloroplast genome sequencing, assembly, and annotation

The genomic libraries were constructed using 100 ng of cpDNA and the Nextera DNA Flex kit (Illumina), following the manufacturer's instructions. Paired-end sequencing (2×150 bp) was performed on the Illumina NextSeq550 platform (Fundação Hemocentro de Ribeirão Preto, Brazil).

Two programs were used to assemble the complete chloroplast genomes of the six Neotropical palms, as they presented different benefits. First, NOVOPlasty was applied to all palms, as its main advantage was the fast de novo assembly of organelle genomes (chloroplasts and mitochondria) from unfiltered whole genome sequence (WGS)⁵⁴. This resulted in high-quality genomes in terms of coverage and accuracy⁵⁴. However, raw WGS-based assembly is not always successful. Many sequencing technologies can result in error-prone, such as highly repetitive regions. With this, NOVOPlasty can generate multiple contigs⁵⁴, making it difficult to circularize the chloroplast genome, this was the case of *B. gasipaes*. Therefore, we adopted another strategy as a way to reduce the ratio of nuclear reads in relation to organelle reads. Before assembly, we mapped them onto a reference genome and performed their filtering⁵⁴. This process required more pre-assemble steps, however, with the partial extraction of data from the original files, it was possible to use NOVOWrap⁵⁵. The program is partly based on NOVOPlasty, but its main advantage is a more automated process with the identification and testing of different seeds according to related organisms present in the literature. Testing different seeds and reference genomes provides additional opportunities for successful assembly⁵⁵. However, this kind of assembly requires high computational memory, which can be a limiting factor.

In this way, the chloroplast genome assembly of *Acrocomia intumescens*, *A. totai*, *Copernicia alba*, *C. prunifera*, and *Syagrus romanzoffiana* was performed in two steps: Firstly, the paired-end reads from these five palms were assembled in NOVOPlasty v4.2⁵⁴ (<https://github.com/ndierckx/NOVOPlasty>) using the *rbcl* gene sequence as a seed (NCBI accession numbers: for *A. intumescens* and *A. totai*, *rbcl* from *A. aculeata*: AY044625.1; *rbcl* of *C. prunifera*: AM110199.1; *rbcl* of *C. alba*: MK753471.1; and *rbcl* of *S. romanzoffiana*: GU135249.1), and the chloroplast genome of *Acrocomia aculeata* (NCBI accession number: NC_037084.1) as a reference to order the contigs⁵⁶. Secondly, the confirmation of quality, correctness, and coverage of the assembly was carried out using Geneious v2020 2.4 (<https://www.geneious.com/>, last assessed January 2022). We applied the “Map to reference” function to map the paired-end raw data onto the final assembled chloroplast genomes.

Using BWA and SAMTools^{57,58}, the sequences from the chloroplast genomes of other palms were indexed (Supplementary Table S4) and mapped with the raw reads of *B. gasipaes* var. *gasiapes* to filter out the ones present only in the organelle genome. After that, the output files .bam were converted to .fastq with BEDTools⁵⁹. With the filtered.fastq files, the chloroplast genome assembly was performed with NOVOWrap v1.20⁵⁵, using the *psaC* gene as seed (NCBI accession number: MH537788) and *Astrocaryum aculeatum* (NCBI accession number: MH537788) as reference genome to order the contigs. Finally, Geneious v2020 2.4 ([Scientific Reports | \(2023\) 13:20635 |](https://www.genei</p>
</div>
<div data-bbox=)

ous.com/, last assessed January 2022) was used, as previously described. The raw reads for each species, as well as the coverage resulting from the assembly of each chloroplast genome, are available in Supplementary Table S4.

The annotation of the chloroplast genomes was performed in GeSeq (Organellar Genome Annotation)⁶⁰ from the Chlorobox platform, with settings for the identification of protein coding sequences (CDS), rRNAs, and tRNAs based on reference chloroplast sequences and homologies through BLAST search. Following the GeSeq annotation, GenomeView⁶¹ was used to conduct a manual correction of start and stop codons and verify pseudogene and intron positions. We then obtained the chloroplast circular genome maps using OGDRAW⁶².

Chloroplast genome structure comparison

We conducted two multiple progressive sequence alignments in Mauve v2.4.0⁶³. The first one included the six new chloroplast genomes and those that were available in GenBank from species that occur in Brazil: *Acrocomia aculeata*, *Astrocaryum aculeatum*, *A. murumuru*, *Butia eriospatha*, *Euterpe edulis* Mart., *E. oleracea* Mart., *E. precatória* Mart., *Mauritia flexuosa*, *Syagrus coronata*, and *Trithrinax brasiliensis*. The second analysis was carried out using 24 chloroplast genomes from different palm species (Supplementary Table S5). We selected species that represented the five palm subfamilies, considering the evolution of the group: *Phytelephas aequatorialis* Spruce and *Pseudophoenix vinifera* (Mart.) Becc. (Subfamily: Ceroxyloideae); *Copernicia alba*, *C. prunifera*, *Caryota mitis* Lour., *Trachycarpus fortunei* (Hook.) H. Wendl., and *Trithrinax brasiliensis* (Subfamily: Coryphoideae); *Nypa fruticans* (Subfamily: Nypoideae); *Calamus caryotoides* A. Cunn. ex Mart., *Eremospatha macrocarpa* Schaedtler, and *Mauritia flexuosa* (Subfamily: Calamoideae); *Veitchia arecina* Becc. (Subfamily: Arecoideae); and the Brazilian native species also from the Arecoideae subfamily (Supplementary Table S5).

The six chloroplast genomes also presented contractions and expansions in the inverted repeat (IR) regions. Since these regions may show structural differences, it is expected to identify variability among species and within palm subfamilies (Arecoideae: *Acrocomia intumescens*, *A. totai*, *Bactris gasipaes*, and *Syagrus romanzoffiana*; Coryphoideae: *Copernicia alba* and *C. prunifera*).

Identification of SSRs and dispersed repeats

Single sequence repeats (SSR) containing 1–6 nucleotides were identified using the MISA web package (available at <https://webblast.ipk-gatersleben.de/misa/>)⁶⁴. To search for SSR motifs, the following configuration was considered: SSR of one to six nucleotides long, with a minimum repeat number of 10, 5, and 4 units for mono-, di-, and trinucleotide SSRs, respectively, and three units for tetra-, penta-, and hexanucleotide SSRs. The determination of dispersed repeats (forward, reverse, palindrome, and complement sequences) was performed in REPuter (available at: <https://bibiserv.cebitec.uni-bielefeld.de/reputer/>)⁶⁵ based on the following criteria: minimum repetition size ≥ 30 bp and sequence identity $\geq 90\%$ (Hamming distance = 3). The composition and position of the SSRs and dispersed repeats were manually compared among each chloroplast genome.

RNA editing sites

The RNA editing sites of the chloroplast genomes of *A. intumescens*, *A. totai*, *B. gasipaes* var. *gasipaes*, *C. alba*, *C. prunifera*, and *S. romanzoffiana* were predicted using predictive RNA editor for plants (PREP)⁶⁶. For this, 35 coding sequences from each chloroplast genome and a cutoff value of 0.8 were used. RNA editing sites were compared between species, considering their positions as well as their amino acid substitutions.

Phylogenomic studies

The annotated GenBank files of each species were imported using a set of Python scripts developed in our group (available under request from CvdB) into a SQLite database. All the putative coding regions were extracted, and each region was individually aligned using MUSCLE v5.1⁶⁷. Then, all the aligned regions were concatenated into a Nexus file, including 'charset' for each individual region and for each codon position in each region. The regions *cemA*, *ndhD*, *petD*, and *rps12* did not produce alignments compatible with codon assignment (missing start codons, non-ternary indels, etc.) and were therefore separated from the canonical coding regions for further testing. Evolutionary models were assessed using four different partition schemes: (i) two partitions, with one model for all coding regions versus a different model for all non-canonical regions (*cemA*, *ndhD*, *petD*, and *rps12*); (ii) five partitions, with one model for all coding regions and four different models for *cemA*, *ndhD*, *petD*, and *rps12*; (iii) four partitions, with three different models for each codon position (1st, 2nd, and 3rd) versus one model for all non-canonical regions; and iv) seven partitions, with three different models for each codon position (1st, 2nd, and 3rd) and four different models for *cemA*, *ndhD*, *petD*, and *rps12*. The evolutionary models for each partition in the partition schemes were estimated with MrModeltest v2.4⁶⁸, and then the different partition schemes were assessed using Bayes Factors⁶⁹ after four Stepping-Stone (SS) analyses of each scheme using MrBayes v3.2.7⁷⁰. Each analysis consisted of two runs with four chains each (one cold and three hot chains) for 20 million generations and a burn-in of 25%. The marginal likelihood of each analysis was estimated using a stepping-stone sampling of 50 steps with 196,000 generations. For phylogeny estimation of the best-selected partition scheme, MrBayes was then run with the same number of runs and chains, 20 million generations, one tree sampled every 1000 steps, and 25% burn-in. After checking convergence and ESS > 200 (in fact, no parameter was under 8000), the majority-rule consensus was used as an estimation of the phylogeny and posterior probabilities.

The assessment of the different partition schemes indicated a large difference in scheme (iii), with an average marginal log-likelihood (AML) of $-189,272.26$, against model (i) = $-191,019.93$, model (ii) = $-191,092.66$, and model (iv) = $-189,485.41$. The difference between the two best models (iii–iv) was 213.15, which indicated very strong evidence under the criterion of Bayes Factors⁶⁹, for the scheme with four partitions (a model for

each codon position and a single model for the regions that could not have codons properly assigned, putatively non-coding).

Data availability

The datasets generated and analysed during the current study are available in the NCBI's GenBank repository (<https://www.ncbi.nlm.nih.gov/>), with the accession numbers and direct links: *Acrocomia intumescens* (OQ129926; <https://www.ncbi.nlm.nih.gov/nuccore/OQ129926>), *Acrocomia totai* (OQ129927; <https://www.ncbi.nlm.nih.gov/nuccore/OQ129927>), *Bactris gasipaes* (OQ129928; <https://www.ncbi.nlm.nih.gov/nuccore/OQ129928>), *Copernicia alba* (OQ129929; <https://www.ncbi.nlm.nih.gov/nuccore/OQ129929>), *Copernicia prunifera* (OQ129930; <https://www.ncbi.nlm.nih.gov/nuccore/OQ129930>) and *Syagrus romanzoffiana* (OQ129931; <https://www.ncbi.nlm.nih.gov/nuccore/OQ129931>).

Received: 25 January 2023; Accepted: 10 October 2023

Published online: 23 November 2023

References

- Baker, W. J. & Couvreur, T. L. P. Global biogeography and diversification of palms sheds light on the evolution of tropical lineages. II. Diversification history and origin of regional assemblages. *J. Biogeogr.* **40**, 286–298. <https://doi.org/10.1111/j.1365-2699.2012.02794.x> (2013).
- Magnabosco, J. W. *et al.* Characterization of the complete plastid genome of *Butia eriospatha* (Arecaceae). *Genet. Mol. Biol.* **43**, 1–5. <https://doi.org/10.1590/1678-4685-gmb-2020-0023> (2020).
- Muscarella, R. *et al.* The global abundance of tree palms. *Glob. Ecol. Biogeogr.* **29**, 1495–1514. <https://doi.org/10.1111/GEB.13123> (2020).
- Balslev, H. *et al.* Species diversity and growth forms in tropical American palm Communities. *Bot. Rev.* **77**, 381–425. <https://doi.org/10.1007/s12229-011-9084-x> (2011).
- Henderson, A., Galeano, G. & Bernal, R. Field guide to the palms of the Americas. *Syst. Bot.* **21**, 258–259. <https://doi.org/10.2307/2419756> (1996).
- Eiserhardt, W. L., Couvreur, T. L. P. & Baker, W. J. Plant phylogeny as a window on the evolution of hyperdiversity in the tropical rainforest biome. *New Phytol.* **214**, 1408–1422. <https://doi.org/10.1111/nph.14516> (2017).
- Cámara-Leret, R. *et al.* Fundamental species traits explain provisioning services of tropical American palms. *Nat. Plants* **3**, 16220. <https://doi.org/10.1038/nplants.2016.220> (2017).
- Cámara-Leret, R. *et al.* Ecological community traits and traditional knowledge shape palm ecosystem services in northwestern South America. *For. Ecol. Manag.* **334**, 28–42. <https://doi.org/10.1016/j.foreco.2014.08.019> (2014).
- Huang, W. P. *et al.* Molecular phylogenetics and biogeography of the eastern Asian-eastern North American disjunct *Mitchella* and its close relative *Dammacanthus* (Rubiaceae, Mitchelleae). *Bot. J. Linn. Soc.* **171**, 395–412. <https://doi.org/10.1111/j.1095-8339.2012.01321.x> (2013).
- Díaz, B. G. *et al.* Genome-wide SNP analysis to assess the genetic population structure and diversity of *Acrocomia* species. *PLoS One* **16**, 1–24. <https://doi.org/10.1371/journal.pone.0241025> (2021).
- Colombo, C. A., Chorfi Berton, L. H., Díaz, B. G. & Ferrari, R. A. Macauba: A promising tropical palm for the production of vegetable oil. *OCL (Oilseeds, Crops, Fats & Lipids)* **25**. <https://doi.org/10.1051/ocl/2017038> (2018).
- de Lima, N. E., Carvalho, A. A., Meerow, A. W. & Manfrin, M. H. A review of the palm genus *Acrocomia*: Neotropical green gold. *Org. Divers. Evol.* **18**, 151–161. <https://doi.org/10.1007/s13127-018-0362-x> (2018).
- Clement, C. R. *et al.* Why extensive research and development did not promote use of peach palm fruit in Latin America. *Agrofor. Syst.* **61**, 195–206. <https://doi.org/10.1023/B:AGFO.0000028999.84655.17> (2004).
- Dransfield, J., *et al.* *Genera Palmarum: The evolution and classification of palms*. (ed. Dransfield, J., *et al.*) (Kew Publishing, Royal Botanic Gardens Kew, 2008).
- Masetto, T. E. *et al.* Germinação e armazenamento de sementes de carandá (*Copernicia alba*). *Cerne* **18**, 541–546. <https://doi.org/10.1590/s0104-77602012000400003> (2012).
- de Sousa, R. F., Silva, R. A. R., Rocha, T. G. F., Santana, J. A. S. & Vieira, F. A. Etnoecologia e etnobotânica da palmeira carnaúba no semiárido Brasileiro. *Cerne* **21**, 587–594. <https://doi.org/10.1590/01047760201521041764> (2015).
- Costa, M. F., Francisconi, A. F., Vancine, M. H. & Zucchi, M. I. Climate change impacts on the *Copernicia alba* and *Copernicia prunifera* (Arecaceae) distribution in South America. *Rev. Bras. Bot.* **45**, 807–818. <https://doi.org/10.1007/s40415-022-00801-8> (2022).
- Junio, R. F. P. *et al.* *Copernicia prunifera* leaf fiber: A promising new reinforcement for epoxy composites. *Polymers* **12**, 2090. <https://doi.org/10.3390/POLYM12092090> (2020).
- Andrade, A. C. *et al.* Prebiotic potential of pulp and kernel cake from Jerivá (*Syagrus romanzoffiana*) and Macaúba palm fruits (*Acrocomia aculeata*). *Food Res. Int.* **136**, 109595. <https://doi.org/10.1016/j.foodres.2020.109595> (2020).
- Cámara-Leret, R. *et al.* Fundamental species traits explain provisioning services of tropical American palms. *Nat. Plants* **3**(2), 16220. <https://doi.org/10.1038/nplants.2016.220> (2017).
- Cámara-Leret, R., Fortuna, M. A. & Bascompte, J. Indigenous knowledge networks in the face of global change. *Proc. Natl. Acad. Sci. USA* **116**(20), 9913–9918. <https://doi.org/10.1073/pnas.1821843116> (2019).
- Dodsworth, S. Genome skimming for next-generation biodiversity analysis. *Trends Plant. Sci.* **20**, 525–527. <https://doi.org/10.1016/j.tplants.2015.06.012> (2015).
- Du, F. K. *et al.* An improved method for chloroplast genome sequencing in non-model forest tree species. *Tree Genet. Genomes* **11**, 114. <https://doi.org/10.1007/s12955-015-0942-2> (2015).
- Ruang-areerate, P. *et al.* Complete chloroplast genome sequences of five *Bruguiera* species (Rhizophoraceae): Comparative analysis and phylogenetic relationships. *PeerJ* **9**, e12268. <https://doi.org/10.7717/peerj.12268> (2021).
- Ruang-areerate, P. *et al.* Comparative analysis and phylogenetic relationships of *Ceriops* species (Rhizophoraceae) and *Avicennia lanata* (Acanthaceae): Insight into the chloroplast genome evolution between middle and seaward zones of Mangrove Forests. *Biology* **11**(3), 383. <https://doi.org/10.3390/biology11030383> (2022).
- Rogalski, M., Vieira, L. D. N., Fraga, H. P. & Guerra, M. P. Plastid genomics in horticultural species: Importance and applications for plant population genetics, evolution, and biotechnology. *Front. Plant Sci.* **6**, 1–17. <https://doi.org/10.3389/fpls.2015.00586> (2015).
- Sugiura, M. The chloroplast genome. *Plant Mol. Biol.* **19**, 149–168. https://doi.org/10.1007/978-94-011-2656-4_10 (1992).
- Daniell, H., Lin, C. S., Yu, M. & Chang, W. J. Chloroplast genomes: Diversity, evolution, and applications in genetic engineering. *Genome Biol.* **17**, 1–29. <https://doi.org/10.1186/s13059-016-1004-2> (2016).

29. Barrett, C. F., *et al.* Plastid genomes reveal support for deep phylogenetic relationships and extensive rate variation among palms and other commelinid 28. Sugiura, M. The chloroplast genome. *Plant Mol. Biol.* **19**, 149–168. https://doi.org/10.1007/978-94-011-2656-4_10 (1992).
30. Áquila, R., Silva, J., Silva, J. V. & Almeida, C. The first complete chloroplast genome sequence of *Syagrus coronata*: comparative analysis in Arecoideae (Arecaceae). *Rev. Bras. Bot.* **41**, 393–401. <https://doi.org/10.1007/s40415-018-0462-z> (2018).
31. Lopes, A. S., *et al.* The complete plastome of macaw palm [*Acrocomia aculeata* (Jacq.) Lodd. ex Mart.] and extensive molecular analyses of the evolution of plastid genes in Arecaceae. *Planta* **247**, 1011–1030. <https://doi.org/10.1007/s00425-018-2841-x> (2018).
32. Lopes, A. S., *et al.* The plastomes of *Astrocaryum aculeatum* G. Mey. and *A. murumuru* Mart. show a flip-flop recombination between two short inverted repeats. *Planta* **250**, 1229–1246. <https://doi.org/10.1007/s00425-019-03217-z> (2019).
33. Lopes, A. S., *et al.* Plastid genome evolution in Amazonian açai palm (*Euterpe oleracea* Mart.) and Atlantic Forest açai palm (*Euterpe edulis* Mart.). *Plant Mol. Biol.* **105**, 559–574. <https://doi.org/10.1007/s11103-020-01109-5> (2021).
34. Vieira, M. L. C., Santini, L., Diniz, A. L. & Munhoz, C. F. Microsatellite markers: What they mean and why they are so useful. *Genet. Mol. Biol.* **39**, 312–328. <https://doi.org/10.1590/1678-4685-GMB-2016-0027> (2016).
35. Silva, R. S. *et al.* The plastome sequence of *Bactris gasipaes* and evolutionary analysis in tribe Cocoseae (Arecaceae). *PLoS One* **16**, 1–15. <https://doi.org/10.1371/journal.pone.0256373> (2021).
36. Francisconi, A. F., *et al.* Complete chloroplast genomes and phylogeny in three *Euterpe* palms (*E. edulis*, *E. oleracea* and *E. precatoria*) from different Brazilian biomes. *PLoS One* **17**, e0266304. <https://doi.org/10.1371/journal.pone.0266304> (2022).
37. Walker, J. F., Jansen, R. K., Zanis, M. J. & Emery, N. C. Sources of inversion variation in the small single copy (SSC) region of chloroplast genomes. *Am. J. Bot.* **102**(11), 1751–1752. <https://doi.org/10.3732/ajb.1500299> (2015).
38. Oldenburg, D. J. & Bendich, A. J. Most chloroplast DNA of maize seedlings in linear molecules with defined ends and branched forms. *J. Mol. Biol.* **335**, 953–970. <https://doi.org/10.1016/j.jmb.2003.11.02040> (2004).
39. Aljohi, H. A. *et al.* Complete sequence and analysis of coconut palm (*Cocos nucifera*) mitochondrial genome. *PLoS One* **11**, 1–18. <https://doi.org/10.1371/journal.pone.0163990> (2016).
40. Wheeler, G. L., Dorman, H. E., Buchanan, A., Challagundla, L. & Wallace, L. E. A review of the prevalence, utility, and caveats of using chloroplast simple sequence repeats for studies of plant biology. *Appl. Plant Sci.* **2**, 1400059. <https://doi.org/10.3732/apps.1400059> (2014).
41. Kong, W. Q. & Yang, J. H. The complete chloroplast genome sequence of *Morus cathayana* and *Morus multicaulis*, and comparative analysis within genus *Morus* L. *PeerJ* **5**, e3037. <https://doi.org/10.7717/peerj.3037> (2017).
42. Jiang, H. *et al.* Comparative and phylogenetic analyses of six *Kenya polystachya* (Orchidaceae) species based on the complete chloroplast genome sequences. *BMC Plant Biol.* **22**, 177. <https://doi.org/10.1186/s12870-022-03529-5> (2022).
43. Chen, H., Deng, L., Jiang, Y., Lu, P. & Yu, J. RNA editing sites exist in protein-coding genes in the chloroplast genome of *Cycas taitungensis*. *J. Integr. Plant Biol.* **53**, 961–970. <https://doi.org/10.1111/j.1744-7909.2011.01082.x> (2011).
44. Tillich, M. *et al.* Loss of matK RNA editing in seed plant chloroplasts. *BMC Evol. Biol.* **9**, 201. <https://doi.org/10.1186/1471-2148-9-201> (2009).
45. Faurby, S., Eiserhardt, W. L., Baker, W. J. & Svenning, J. C. An all-evidence species-level supertree for the palms (Arecaceae). *Mol. Phylogenet. Evol.* **100**, 57–69. <https://doi.org/10.1016/j.ympev.2016.03.002> (2016).
46. Meerow, A. W., *et al.* Phylogeny and historical biogeography of the coccosoid palms (*A. recaceae*, *A. recoideae*, *C. ocoseae*) inferred from sequences of six WRKY gene family loci. *Cladistics* **31**, 509–534. <https://doi.org/10.1111/cla.12100> (2015).
47. Baker, W. J. & Couvreur, T. L. P. Global biogeography and diversification of palms sheds light on the evolution of tropical lineages I. Historical biogeography. *J. Biogeogr.* **40**, 274–285. <https://doi.org/10.1111/j.1365-2699.2012.02795.x> (2013).
48. Vianna, S. A., Carmelo-Guerreiro, S. M., Noblick, L. R. & Colombo, C. A. Leaf anatomy of *Acrocomia* (Arecaceae): An additional contribution to the taxonomic resolution of a genus with great economic potential. *Plant Syst. Evol.* **303**, 233–248. <https://doi.org/10.1007/s00606-016-1369-4> (2017).
49. Govaerts, R., & Dransfield, J. *World checklist of palms* (eds. Govaerts, R., Dransfield, J.) 1–223 (Richmond, 2005).
50. Meerow, A. W. *et al.* Phylogenetic analysis of seven WRKY genes across the palm subtribe Attaleinae (Arecaceae) identifies *Syagrus* as sister group of the coconut. *PLoS One* **4**, e7353. <https://doi.org/10.1371/journal.pone.0007353> (2009).
51. Maitner, B. S. *et al.* The bien r package: A tool to access the botanical information and ecology network (BIEN) database. *Methods Ecol. Evol.* **9**, 373–379. <https://doi.org/10.1111/2041-210X.12861> (2018).
52. R Core Team. R: A language and environment for statistical computing. <https://www.R-project.org/> (2021).
53. Takamatsu, T. *et al.* Optimized method of extracting rice chloroplast DNA for high-quality plastome resequencing and de Novo assembly. *Front. Plant Sci.* **9**, 266. <https://doi.org/10.3389/fpls.2018.00266> (2018).
54. Dierckxens, N., Mardulyn, P. & Smits, G. NOVOPlasty: De novo assembly of organelle genomes from whole genome data. *Nucleic Acids Res.* **45**, e18. <https://doi.org/10.1093/nar/gkw955> (2017).
55. Wu, P. *et al.* NOVOWrap: an automated solution for plastid genome assembly and structure standardization. *Mol. Ecol. Resour.* **21**, 2177–2186. <https://doi.org/10.1111/1755-0998.13410> (2021).
56. Jin, J. J. *et al.* GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* **21**, 241. <https://doi.org/10.1101/256479> (2020).
57. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324> (2009).
58. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352> (2009).
59. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842. <https://doi.org/10.1093/bioinformatics/btq033> (2010).
60. Tillich, M. *et al.* GeSeq—Versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* **45**, W6–W11. <https://doi.org/10.1093/nar/gkx391> (2017).
61. Abeel, T., Van Parys, T., Saeys, Y., Galagan, J. E. & Peer, Y. V. Genome view: A next-generation genome browser. *Nucleic Acids Res.* **40**, 1–10. <https://doi.org/10.1093/nar/gkr995> (2012).
62. Greiner, S., Lehwark, P. & Bock, R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: Expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* **47**, W59–W64. <https://doi.org/10.1093/nar/gkz238> (2019).
63. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**, 1394–1403. <https://doi.org/10.1101/gr.2289704> (2004).
64. Beier, S., Thiel, T., Münch, T., Scholz, U. & Mascher, M. MISA-web: A web server for microsatellite prediction. *Bioinformatics* **33**, 2583–2585. <https://doi.org/10.1093/bioinformatics/btx198> (2017).
65. Kurtz, S. & Schleiermacher, C. REPuter: Fast computation of maximal repeats in complete genomes. *Bioinformatics* **15**, 426–427. <https://doi.org/10.1093/bioinformatics/15.5.426> (1999).
66. Mower, J. P. The PREP suite: Predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. *Nucleic Acids Res.* **37**, W253–W259. <https://doi.org/10.1093/nar/gkp337> (2009).
67. Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.* **32**, 1792–1797. <https://doi.org/10.1093/nar/gkh340> (2004).
68. Nylander, J. A. A. MrModeltest v2. <https://github.com/nylander/MrModeltest2> (2004).
69. Kass, R. E. & Raftery, A. Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795. <https://doi.org/10.1080/01621459.1995.10476572> (1995).

70. Ronquist, F. *et al.* MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542. <https://doi.org/10.1093/sysbio/sys029> (2012).

Acknowledgements

This work was supported by the São Paulo Research Foundation (FAPESP, Portuguese: Fundação de Amparo à Pesquisa do Estado de São Paulo) (21/10319-0). The sponsor has no role in the study design, data collection, and data analysis, or manuscript preparation.

Author contributions

A.F.F. and J.A.M.M. are both first authors of the paper, performed the experiment, analysed the results, and wrote the manuscript. L.A.C.S. designed the experiment, analysed, and reviewed the results. C.vd.B analysed the results. K.K.M.M. collected the samples and analysed the results. M.F.C. and D.P.R. collected the samples and performed the experiment. L.D.A. and C.A.Z. performed the experiment. C.A.C. and B.G.D. collected the samples. D.T.A. analysed the results. E.A.V. and M.T.G.L. reviewed and formatted the manuscript. M.I.Z. designed the experiment, reviewed the manuscript and funding support. All authors read the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-44631-4>.

Correspondence and requests for materials should be addressed to M.I.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023