Complete chloroplast genomes of six Neotropical palm tree species: genome structure comparison, identification of repeats sequences and evolutionary dynamic patterns

Ana Flávia Francisconi
Universidade Estadual de Campinas
Jonathan Andre Morales Marroquín
Universidade Estadual de Campinas
Luiz Augusto Cauz-Santos
University of Vienna
Cássio van den Berg
Universidade Estadual de Feira de Santana, Feira de Santana
Kaunne Karolina Martins
Universidade Estadual de Campinas
Marcones Ferreira Costa
Universidade Estadual de Campinas
Doriane Picanço-Rodrigues
Universidade Federal do Amazonas
Luciano Delmodes de Alencar
Universidade Estadual de Campinas
Cesar Augusto Zanello
Universidade Estadual de Campinas
Carlos Augusto Colombo
Instituto Agronômico
Brenda Gabriela Díaz Hernández
Instituto Agronômico
Danilo Trabuco Amaral
Universidade Federal do ABC
Maria Teresa Gomes Lopes
Universidade Federal do Amazonas
Elizabeth Ann Veasey
Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo
Maria Imaculada Zucchi (mizucchi@sp.gov.br)
Agência Paulista de Tecnologia Dos Agronegócios (APTA), Polo Centro Sul

Article

Keywords:

Posted Date: February 24th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2514466/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on November 23rd, 2023. See the published version at https://doi.org/10.1038/s41598-023-44631-4.

Abstract

The Arecaceae family has a worldwide distribution, specially in the tropical and subtropical regions. We sequenced the chloroplast genome of *Acrocomia intumescens and A. totai*, widely used in the food and energy industry, *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 *B. gasipaes* presented the smallest (155,078 bp). *Copernicia* spp. had two more gene copies (trnL-GAG) than the others, and *S. romanzoffiana*underwent pseudogenization in one copy of *rps19*. 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, closer phylogenetically species shared more sites. Almost all nodes of the phylogenetic analysis showed 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.

Main text

The Arecaceae family comprises 188 genera and 2,585 species. Its distribution is cosmopolitan and worldwide and is concentrated mainly in tropical and subtropical regions[1-3]. 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 endemic[4]. Palms are one of the dominant species groups of the tropical rainforest (TRF). Also, the Arecaceae family presents high species richness and is a key component on the evolution and diversification of hyper diversity in the TRF biome[5,6]. Palms are part of the TRF ecosystem 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). *Acrocomia* includes eight species, of which *Acrocomia aculeata* (Jacq.) Lodd. Mart., *A. intumescens* Drude, and *A. totai* Mart. have great economic interest considering that all vegetative structures can be used[10]. The three species have high oil content and they 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, for example *A. totai* is commonly used as food in its habitat[12].

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 natura or traditionally used in regional dishes at Amazonian markets. *B. gasipaes* is essential to supply the production chain of high quality heart-of-palm[13]. 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 America[14]. 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 potential[15]. *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 coating[18]. Both palms are endemic to Brazilian biomes, Pantanal and Caatinga, respectively. Finally, *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. The oil obtained from this species is applicable to food, cosmetics, pharmaceutical, and biodiesel industries[19].

Due to their historical use and growing economic importance, 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. There is a great difference among the genetic research addressing these six palm species. The genome-wide approach was used to obtain single nucleotide polymorphism markers (SNPs) for assessing the genomic population structure and diversity of *Acrocomia* species, suggesting the separation of three species: *A. aculeata*, *A. totai*, *A. intumescens*[10]. For *B. gasipaes*, the first chloroplast genome was assembled and used to make inferences within the tribe Cocoseae[20]. In the Brazilian Northwest, 11 natural populations of *C. prunifera* were analyzed using Inter Simple Sequence Repeats (ISSR) markers, in which genetic bottlenecks were detected due to wax extracting, land-use change for wind power expansion and agriculture decreasing natural habitat[17, 21]. For *S. romanzoffiana*, a recent study using SSR (single sequence repeats) markers detected low diversity in specific genetic parameters, emphasizing the need of increasing genetic research within this species[22]. No studies using genetic markers or gene sequences have been published for the species *C. alba*.

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-model species[23,24]. The chloroplast is a semi-autonomous organelle in plant cells, with its own genetic information and a complete genetic system, and its genome is typically uniparentally inherited with the absence of recombination[25]. 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)[26]. This genome consists basically of long circular or linear molecules (120 to 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 plants[27], including palm species[28].

For palm species occurring in Brazil, ten complete chloroplast genomes have been reported[2,20,28-31]. These genomes are used for structural comparisons, phylogenetic inferences, determination of RNA editing sites and identification of genomic regions with different repeats and polymorphism. Among the advantages of elucidating new palm chloroplast genomes are increasing the understanding of evolutionary processes in Arecaceae and deciphering phylogenetic relationships between closely related taxa[27]. Furthermore, with the available sequences, it will be possible to identify repetitive sequences and regions with high polymorphism and screen for molecular markers, which are extremely important for assessing genetic structure and diversity in natural populations[32].

Considering the importance and the necessity for increasing genomic information about these Neotropical palm species, here we report the complete chloroplast genome of *A. intumescens, A. totai, B. 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 sequences repeats and dispersed repeats; (4) RNA editing sites and their conservancy among the six palm species; (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 palms species had the typical quadripartite structure with the presence of two single copies (Large single copy – LSC, Small single copy – SSC) and two inverted repeat regions (IRA and IRB, Fig. 1). The species from the *Copernicia* genus had the largest chloroplast genome size (*C. prunifera*: 157,323 bp and *C. alba*: 157,192 bp, Fig. 1, Table 1) and *S. romanzoffiana* had the smallest size (155,078 bp, Fig. 1, Table 1). The species from the genus *Copernicia* also displayed the longest LSC (*C. alba*: 86,430 bp and *C. prunifera*: 86,264 bp, Table 1). However, *Acrocomia intumescens* presented the largest SSC (17,522 bp). Finally, the species *A. intumescens*, *A. totai* and *B. gasipaes* showed IRs with more nucleotides (bp) than the species of the genus *Copernicia* and *S. romanzoffiana*. All species had GC content (%) between 37.2–38, with the GC content concentrated in the IRs (Table 1).

Table 1
Gene features in the *Acrocomia intumescens*, *A. totai, Bactris gasipaes, Copernicia alba, C. prunifera* and *Syagrus romanzoffiana* chloroplast genomes according to each respective category.

Gene features	*A. intumescens*	*A. totai*	*B. gasipaes*	*C. alba*	*C. prunifera*	*S. romanzoffiana*
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
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	133	133	130
Protein-coding genes	85	85	85	85	85	84
rRNA genes	8	8	8	8	8	8
tRNA genes	38	38	38	40	40	38

The species presented variation in the number of genes (133–130, considering duplicated copies, Table 1). *C. alba* and *C. prunifera* had two copies of the *trnL – GAG*, a tRNA that was absent in the chloroplast genome of the species from the subfamily Arecoideae, analyzed in this study (Fig. 1, Table 1). Also, *S. romanzoffiana* presented one gene loss when compared with the other species of the subfamily Arecoideae, the *rps19*, which in this case presented just as a fragment sequenced and as considered a pseudogene. (Fig. 1).

The chloroplast genome structures and comparative analyses among species from different genera. The Brazilian palm trees chloroplast genomes annotated in this study showed a high level of synteny in their structures. Although the species are from different subfamilies (*C. alba*, *C. prunifera*, *T. brasiliensis*, subfamily Coryphoideae; *M. flexuosa*, subfamily Calamoideae; and the others from the subfamily Arecoideae), a few structural rearrangements were observed (Fig. 2). This structural conservation was also noted in the comparison of palm chloroplast genomes with five subfamilies (see Supplementary Fig. S1 online). The most notable change was in *Astrocaryum aculeatum* and *A. murumuru*, which showed flip-flop recombination (in lime green, Fig. 2)[31]. As already pointed out, a significant dissimilarity among the species of the subfamily Arecoideae that occurs in Brazil can be identified in the length of the LSC, between 40,000 and 50,000 bp[33]. 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, *Syagrus romanzoffiana* had the smallest SSC length (Table 1, Fig. 2), which was also noted in the structure of *S. coronate*.

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. *C. prunifera* one 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 (Fig. 3). 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. Seventy-two and 69 SSR were identified in *A. intumescens* and *A. totai*, respectively. In *B. gasipaes*, 88 SSRs were found. A higher amount of SSRs was observed in *C. alba* and *C. prunifera*, 104 and 100 SSRs. *S. romanzoffiana* presented 83 SSR in its chloroplast genome (S2 Table). All six palms had a higher number of mononucleotides SSR type, followed by di- and tetranucleotides (Fig. 4A, see Supplementary Table S2 online), mostly concentrated in the LSC region of their chloroplast genome (Fig. 4B, see Supplementary Table S2 online). Also, all species presented the motifs A/T, AT/AT, AAAT/ATTT, AAT/ATT, AAAAT/ATTTT, AATG/ATTC, AG/CT, 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. gasiapes* (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* (41). The smallest values of these repeats were observed in *S. romanzoffiana* (35), *C. prunifera* (31) and *C. alba* (11) (S3 Table). *A.*

intumescens, A. totai and *B. gasipaes* also presented similar distribution of repeats with higher concentration of palindrome (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 eight, respectively). Inversely, *C. alba* had a higher number of forward (seven repeats) types, followed by palindrome type (two repeats, Fig. 5A). All studied species presented the highest number of dispersed repeats concentrated in the LSC region, while only *C. alba* and *C. prunifera* showed any dispersed repeats in the SSC region of the chloroplast genome (Fig. 5B). *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 (five repeats, *C. prunifera*). *S. romanzoffiana*, on the other hand, had the highest number of repeats of an intermediate size of 37 bp (seven repeats; Fig. 5C)

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. All RNA editing sites analyzed showed conversion of the nucleotides from Cytidine (C) to Uridine (U). 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) (see Supplementary Fig. S2A online). Thus, the conversions resulted in 82.35% hydrophilic to hydrophobic amino acids (see Supplementary Fig. S2B online).

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* and *S. romanzoffiana*, 19 conversions were observed (Fig. 6A). The number of RNA editing sites, considering pairwise species analyses, resulted in *S. romanzoffiana* x *B. gasipaes* sharing 78 sites, followed by *A. intumescens* x *A. totai* (77), and *C. alba* x *C. prunifera* (75) (Fig. 6B). The species that shared the lowest number of conversions belonged to *Acrocomia* and *Copernicia*: *C. prunifera* x *A. intumesces* (54), *C. prunifera* x *A. totai* (54), *C. alba* x *A. intumescens* (54), *C. alba* x *A. totai* (54) (Fig. 6B). The *ndh* genes were the ones with the highest rate of changes, 38 in total. 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 was exclusive to *Acrocomia* species and another to *Copernicia* palms. Similarly, two conversions were not exclusively detected in these species.

Phylogenomic studies. The assessment of different partition schemes indicated with large difference the 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 Factors[34], 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).

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* and the clade containing *Salacca, Metroxylon, Pigafetta*, and *Calamus* (PP = 0.92). Regarding the species with new chloroplast genome sequences generated in the current study, *Acrocomia intumescens* and *A. totai* were sister to each other and then sister to *A. aculeata*. *Bactris gasipaes* was positioned as sister to two species of *Astrocaryum*, and *Syagrus romanzoffiana* was sister to the previously sequenced *S. coronata*. The species of *Copernicia* represent the first chloroplast genomes for the genus and were sister to each other and positioned in a small clade with *Pritchardia* and *Colpothrinax*, 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 from the analyzed species were observed. The chloroplast genome of *B. gasipaes* showed a small variation in total size from the chloroplast genome already published[20], from 156.646 to 155.078 bp. This result is probably caused by differences in the assembly method, since our reads were indexed and filtered for chloroplast sequences.

Among the two species of the Coryphoideae subfamily, *Copernicia alba* and *C. prunifera*, an additional tRNA, *trnL-GAG*, was identified when compared to the other species of the Arecoideae subfamily. This gene was first noted in chloroplasts of algae species[35] and encodes the leucine amino acid. Analyzing the complete chloroplast genomes available in the GenBank, presented in the supplementary table online (S1 Table), this gene was identified only in certain palm species of the subfamilies Coryphoideae, Calamoideae and Ceroxyloideae. According to the phylogeny proposed by Barrett et al.[28] using whole chloroplast genomes, the family tend to converge to a topology of the subfamilies Calamoideae, Nypoideae, Coryphoideae, Ceroxyloideae, but in the species *Nypa fruticans*, of the subfamily Nypoideae (monotypic), this tRNA was also not reported.

Multiple alignments with 24 different palms 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, which is probably lineage-specific[31]. Regarding the patterns of IR structure, most of the expansion/contraction-related variations in the six new chloroplast genomes were identified. The most notable difference was with the species *S. romanzoffiana*. This species had one less copy of the *rps19* gene compared to the other species of the Arecoideae subfamily (Table 1), now classified as a fragment. This event was called the "*rps19* pseudogenization" and was first detected in the palm complete chloroplast genome of *Cocos nucifera*[36]. According to the authors, the palm species had two *rps19* genes and, for coconut, only one *rps19* was found at the IRB/SSC. The other copy was a *rps19*-like sequence with a 174 bp size. Compared with other *rps19* genes, of 279 bp in length, this was judged to be a pseudogene, probably caused by fluctuations, expansions/contractions, in the IR[36]. Recently, in the chloroplast genome of *Butia eriospatha*, the absence of the duplicated *rps19* gene was also observed[2]. Although not discussed, this same event was likewise found in the complete chloroplast genome of *Syagrus coronate*[29], which, like *S. romanzoffiana*, showed the pseudogene copy of *rps19* with the same size, of 193 bp. The four species are closely related, as a phylogeny based on chloroplast genome sequences identified that *C. nucifera*, *S. coronate* and *B. eriospatha* shared the same branch and can be classified as sisters, in the subtribe Attaleinae, of tribe Cocoseae[20].

Although they are considered highly conserved structures for stabilizing chloroplast genomes structure[36], fluctuations caused by contractions/expansions of the IRs have already been reported in palms[30,33]. The most surprising case was the total loss of one IR in the species *Tahina spectabilis*, instead of rearrangement, as this structure was considered canonical in monocots[28]. Since it was a minor fluctuation in IR regions, the pseudogenization 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.[28], 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-, hexanucleotides) 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[30,31,33]. Plastid SSR markers, especially combined with nuclear markers, can improve the efficiency of studies of structure, diversity and gene flow in natural populations[37]. Thus, the sequences identified are a valuable resource for knowledge leveling and sustainable planning for the species management.

Similarities in the number and distribution of dispersed repeats were also observed for *A. intumescens, A. totai* and *B. gasipaes*. These parallels could be detected in both the number and distribution of repetition types. *S. romanzoffiana* showed a higher number of reverse than forward repeats, unlike the 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 in a fewer frequency, different numbers of dispersed repeats have already been reported in other land plants. In trees of the genus *Morus* (Family: Moraceae), only 14 dispersed repeats were detected, and of these, the species *M. multicaulis* and *M. cathayana* shared only four[38]. Among seven species of the genus *Polystachya*, there were a variation of 2 (*P. dendrolliflora*) and 8 (*P. modesta*) dispersed repeats. The authors indicated that variations in these repetitions may indicate flexibility in the evolutionary process[39] and that this may be considered an indicator of diversity among species, important for the phylogeny of the genus[38].

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[36,40]. 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 new RNA editing sites increases the stability of these trans-membrane proteins[40]. The *matK* gene is known for its rapid evolution and has been a favorite for determining phylogenetic relationships in angiosperms[41]. 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 sites[41]. The addition of new chloroplast genomes might clarify what is leading to the loss or appearance of these sites.

Despite few exceptions, RNA editing is evolutionary conserved and it is expected that more related taxa have more editing sites in common[40]. Of the 102 RNA editing sites, 54 were shared among the six palm species. Also, species from the same subfamily have more RNA editing sites in common (Coryphoideae: 20 and Arecoideae: 19). 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 genus *Acrocomia* and *Bactris* also participate in the same subtribe, Bactridinae[42]. This was reflected in the fact that they had in common 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.[43], 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*. There is no chloroplast genome available for either *Desmoncus* or *Aiphanes*, but our phylogeny indicates with PP = 1.0 that *Bactris* is sister to *Astrocaryum* and both to *Acrocomia*. This relationship is also in agreement with the whole chloroplast genome phylogeny of Silva et al.[20], which also included their sequence of the *Bactris* chloroplast genome, and the phylogenomic supertree-based study of Barret et al.[28]. The relationship among the *Acrocomia* species indicated *A. totai* as sister to *A. intumescens* and then both sister 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 that irradiate in the late Eocene before the final uplift of the Andes (late Miocene and Pliocene)[44]. In the population genomics study based on GBS 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.[43], and with leaf anatomy, by Vianna et al.[45], although some authors still considered the species as synonyms[5,46]. It is important to point out that *Acrocomia aculeata* has the widest distribution among the neotropical palms, from central Mexico to northern Argentina[10]. *Acrocomia* still have unshed phylogenetic relationships due to its hybridization process, domestication, and possible ecotypes because a complete genus phylogeny is yet not 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.[47], and the phylogenomic study of Barret et al.[28]. 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.[28].

Methods

Species, sampling and DNA extraction. In order to characterize the geographical occurrence of the six species, we surveyed the databases using BIEN[48] R package[49]. *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 online). All palms in this study were wild plants and their leaves were sampled from planted collections. The leaves were dried with silica gel and stored in a freezer at -20°C. The leaf material from the *A. intumescens, A. totai, C. alba* and *C. prunifera* were obtained in the active germplasm bank of the Plant Genetic Resources Center of the Agronomic Institute (IAC) of Campinas, Campinas – SP, Brazil (-22.8717, -47.0776). *Bactris gasipaes* and *S. romanzoffiana* were samples from the *ex situ* collection at the IAC in Piracicaba, Piracicaba-SP, Brazil (-22.6836, -47.6458). This study complies with relevant institutional, national, and international guidelines and legislation. The appropriate permissions for 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 according to the Brazilian Decree No. 8,772 (May 11, 2016) and regulated by Brazilian Law No. 13,123 (May 20, 2015; SISGEN number A411583, Brazil). In addition, because the palms species are conserved *in vivo* in the gene bank, they were not deposited in herbarium, however DNA vouchers are stored in our laboratory.

To extract the chloroplast organelles, a sucrose gradient method was used to isolate them[50]. For this, 20 g of fresh leaves from each species were frozen with liquid nitrogen and macerated. The material was resuspended with 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 1,000 × g for 10 min.

The pellet was resuspended using 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 × 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 3,000 × 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^{o}C for 1 h. The supernatant was extracted twice with an equal volume of chloroform: isoamyl alcohol (24: 1) and centrifuged at 10,000 × g (20 min). The same volume of isopropanol was added and incubated at 20^{o}C for 1h. Lastly, the aqueous phase was centrifuged for 10,000 × g (20 min), and 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 (2x 150 bp) was performed on the Illumina NextSeq550 platform (Fundação Hemocentro de Ribeirão Preto, Brazil).

The chloroplast genome assembly of *Acrocomia intumescens*, *A. totai, Copernicia alba, C. prunifera* and *Syagrus rommanzoffiana* was performed in two steps: Firstly, the filtered reads from these five palms were assembled in NOVOPlasty v 4.2[51] (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; *rcbl* of *C. prunifera*: AM110199.1; *rbcl* of *C. alba*: MK753471.1; *rbcl* of *S. romanzoffiana*: GU135249.1) and the chloroplast genome of *Acrocomia aculeata* (NCBI accession number: NC_037084.1) as a reference to ordinate the contigs[52]. 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.

NOVOPlasty was unable to assemble the chloroplast genome of *Bactris gasipaes*, so NOVOWrap v1.20[53] was used. NOVOWrap is partly based on NOVOPlasty but is more automated by choosing and testing different seeds, according to related organisms present in the literature. Also, to facilitate the NOVOWrap processing, we first obtained only chloroplast reads for genome assembly. Using BWA and SAMTools[54,55], the sequences from chloroplast genomes of other palms were indexed (S1 Table) and mapped with the raw reads of *B. gasipaes* to filter the ones present only in the organelle genome. After that, the output files .bam were converted to .fastq with the BEDTools[56]. With the filtered .fastq files, the chloroplast genome assembly was performed with the NOVOWrap[53], using the *psaC* gene as seed (NCBI accession: MH537788) and *Astrocaryum aculeatum* (NCBI accession: MH537788) as reference genome to order the contigs. Finally, the Genious 2.4. (https://www.geneious.com/, last assessed January 2022) was used, as previously described.

The annotation of the chloroplast genomes was performed in GeSeq (Organellar Genome Annotation)[57] from the Chlorobox platform, with settings for the identification of protein coding sequence (CDS), rRNAs and tRNAs based on reference chloroplast sequences and homologies through BLAST search. Following the GeSeq annotation, GenomeView[58] was used to conduct a manual correction of start and stop codons and verification of pseudogenes and intron positions. We then obtained the chloroplast circular genome maps using OGDRAW[59].

Chloroplast genome structure comparison. We conducted two multiple progressive sequence alignment in Mauve v.2.4.0[60]. The first one included the new six 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*, *E. oleracea*, *E. precatoria, Mauritia flexuosa, Syagrus coronata* and *Trithrinax brasiliensis.* The second analysis was carried out using 24 chloroplast genomes from different palm species (S1 Table). We selected species that represented the five palm subfamilies considering the evolution of the group: *Phytelepas aequatoriallis* and *Pseudophoenix vinifera* (Subfamily: Ceroxyloideae); *Copernicia alba, C. prunifera, Caryota mitis*, *Trachycarpus fortune* and *Trithrinax brasiliensis*, (Subfamily: Coryphoideae); *Nypa fruticans* (Subfamily: Nypoideae); *Calamus caryotoides*, *Eresmopatha macrocarpa* and *Maurita flexuosa* (Subfamily: Calamoideae); *Veitchia arecina* (Subfamily: Arecoideae) and the Brazilian native species also from the Arecoideae subfamily (S1 Table).

The six chloroplast genomes also presented contractions and expansions in the inverted repeat (IR) regions from the chloroplast genome analyzed. 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/)[61]. 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[62], based on the following criteria: minimum repetition size ≥ 30 bp and sequence identity ≥ 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 from *A. intumescens, A. totai, B. gasipaes, C. alba, C. prunifera* and *S. romanzoffiana* were predicted using predictive RNA editor for plants (PREP)[63]. For this, 35 coding sequences of 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 substitution.

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, and then all the putative coding regions were extracted and each region was individually aligned using MUSCLE v5.1[64]. Then, all the aligned regions were concatenated into a Nexus file, including ‘charsets’ 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 in 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 v2.4[65], and then the different partition schemes were assessed using Bayes Factors after four Stepping-Stone (SS) analyses of each scheme using MrBayes v3.2.7[66]. 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 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 and one tree sampled every 1,000 steps, and 25% burn-in. After checking convergence and ESS > 200 (in fact no parameter was under 8,000), the majority-rule consensus was used as an estimation of the phylogeny and posterior probabilities.

Declarations

Acknowledgments

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. 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 Statement

The authors declare no competing interests.

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).

Additional information

The online version contains supplementary material of tables and figures.

Correspondence and requests for materials should be address to AFF and MIZ

References

1. 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 (2013a). https://doi.org/10.1111/j.1365-2699.2012.02794.x
2. Magnabosco, J.W., *et al.* Characterization of the complete plastid genome of *Butia eriospatha* (Arecaceae). Genet. Mol. Biol. **43**, 1–5 (2020). https://doi.org/10.1590/1678-4685-gmb-2020-0023
3. Muscarella, R., *et al.* The global abundance of tree palms. Glob. Ecol. Biogeogr. **29**, 1495–1514 (2020). https://doi.org/10.1111/GEB.13123
4. Balslev, H., *et al.* Species diversity and growth forms in tropical American palm Communities. Bot. Rev. **77**, 381–425 (2011). https://doi.org/10.1007/s12229-011-9084-x
5. Henderson, A., Galeano, G. & Bernal, R. Field guide to the palms of the Americas. Syst. Bot. **21**, 258–259 (1996). https://doi.org/10.2307/2419756
6. 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 (2017). https://doi.org/10.1111/nph.14516
7. Cámara-Leret, R., *et al.* Fundamental species traits explain provisioning services of tropical American palms. Nat. Plants **3**, 16220 (2017). https://doi.org/10.1038/nplants.2016.220
8. 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 (2014). https://doi.org/10.1016/j.foreco.2014.08.019
9. Huang, W.P., *et al.* Molecular phylogenetics and biogeography of the eastern Asian-eastern North American disjunct *Mitchella* and its close relative *Damnacanthus* (Rubiaceae, Mitchelleae). Bot. J.

Linn. Soc. **171**, 395–412 (2013). https://doi.org/10.1111/j.1095-8339.2012.01321.x

10. 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 (2021). https://doi.org/10.1371/journal.pone.0241025
11. Colombo, C.A., Chorfi Berton, L.H., Diaz, B.G. & Ferrari, R.A. Macauba: a promising tropical palm for the production of vegetable oil. OCL (Oilseeds, Crops, Fats & Lipids) 25 (2018). https://doi.org/10.1051/ocl/2017038
12. Lima, N.E. de, 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 (2018). https://doi.org/10.1007/s13127-018-0362-x
13. 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 (2004). https://doi.org/10.1023/B:AGFO.0000028999.84655.17
14. Dransfield, J., *et al.* (2008) *Genera Palmarum: the evolution and classification of palms*. (ed. Dransfield, J., *et al.*) (Kew Publishing, Royal Boranic Gardens Kew, 2008).
15. Masetto, T.E., *et al.* Germinação e armazenamento de sementes de carandá (*Copernicia alba*). Cerne **18**, 541–546 (2012). https://doi.org/10.1590/s0104-77602012000400003
16. Sousa, R.F. de, 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 (2015). https://doi.org/10.1590/01047760201521041764
17. 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 (2022). https://doi.org/10.1007/s40415-022-00801-8
18. Junio, R.F.P., *et al. Copernicia prunifera* leaf fiber: A promising new reinforcement for epoxy composites. *Polymers* **12**, 2090 (2020) https://doi.org/10.3390/POLYM12092090
19. 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 (2020). https://doi.org/10.1016/j.foodres.2020.109595
20. Silva, R.S., *et al*. (2021) The plastome sequence of *Bactris gasipaes* and evolutionary analysis in tribe Cocoseae (Arecaceae). *PLoS One* **16**, 1–15 (2021). https://doi.org/10.1371/journal.pone.0256373
21. Santos, J.R.M., *et al.* (2021) Overexploitation and anthropogenic disturbances threaten the genetic diversity of an economically important neotropical palm. *Biodivers. Conserv*. **30**, 2395–2413 (2021). https://doi.org/10.1007/s10531-021-02200-z
22. Laindorf, B.L., *et al*. Genetic diversity and structure of *Syagrus romanzoffiana* (Cham.) Glassman (Arecaceae) in Southern Brazil. Trop. Conserv. Sci. 11, 194008291879833 (2018). https://doi.org/10.1177/1940082918798330
23. Dodsworth, S. Genome skimming for next-generation biodiversity analysis. Trends Plant. Sci. **20**, 525–527 (2015). https://doi.org/10.1016/j.tplants.2015.06.012

24. Du, F.K., *et al.* An improved method for chloroplast genome sequencing in non-model forest tree species. Tree Genet. Genomes **11**, 114 (2015). https://doi.org/10.1007/s11295-015-0942-2
25. 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 (2015). https://doi.org/10.3389/fpls.2015.00586
26. Sugiura, M. The chloroplast genome. Plant Mol. Biol. **19**, 149–168 (1992). https://doi.org/10.1007/978-94-011-2656-4_10
27. Daniell, H., Lin, C.S., Yu, M. & Chang, W.J. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol. **17**, 1–29 (2016). https://doi.org/10.1186/s13059-016-1004-2
28. Barrett, C.F., *et al*. Plastid genomes reveal support for deep phylogenetic relationships and extensive rate variation among palms and other commelinid monocots. New Phytol. **209**, 855–870 (2016). https://doi.org/10.1111/nph.13617
29. 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 (2018). https://doi.org/10.1007/s40415-018-0462-z
30. 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 (2018). https://doi.org/10.1007/s00425-018-2841-x
31. 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 (2019). https://doi.org/10.1007/s00425-019-03217-z
32. Vieira, M.L.C., Santini, L., Diniz, A.L. & Munhoz, C. de F. Microsatellite markers: what they mean and why they are so useful. Genet. Mol. Biol. **39**, 312–328 (2016). https://doi.org/10.1590/1678-4685-GMB-2016-0027
33. Francisconi, A.F., *et al.* Complete chloroplast genomes and phylogeny in three Euterpe palms (*E. edulis*, *E. oleraceae* and *E. precatoria*) from different Brazilian biomes. *PLoS One* **17**, e0266304 (2022). http://doi.org/10.1371/journal.pone.0266304. eCollection 2022
34. Kass, R. E., & Raftery A. Bayes factors. J. Amer. Stat. Assoc. **90**, 773–795 (1995). https://doi.org/10.1080/01621459.1995.10476572
35. Wakasugi, T., Nagai, T., Kapoor, M. & Sigiura, M. Complete nucleotide sequence of the chloroplast genome from the green alga *Chlorella vulgaris*: The existence of genes possibly involved in chloroplast division. *Proc. Natl. Acad. Sci. USA* **94**, 5967–5972 (1997). https://doi.org/10.1073/pnas.94.11.5967
36. Aljohi, H.A. *et al*. Complete sequence and analysis of coconut palm (*Cocos nucifera*) mitochondrial genome. PLoS One **11**, 1–18 (2016). https://doi.org/10.1371/journal.pone.0163990
37. Wheeler GL, Dorman HE, 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 (2014). https://doi.org/10.3732/apps.1400059

38. 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 (2017). https://doi.org/10.7717/peerj.3037
39. 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 (2022). https://doi.org/10.1186/s12870-022-03529-5
40. 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 (2011). https://doi.org/10.1111/j.1744-7909.2011.01082.x
41. Tillich, M., *et al.* Loss of matK RNA editing in seed plant chloroplasts. BMC Evol. Biol. **9**, 201 (2009). https://doi.org/10.1186/1471-2148-9-201
42. 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 (2016). https://doi.org/10.1016/j.ympev.2016.03.002
43. Meerow, A.W., *et al.* Phylogeny and historical biogeography of the cocosoid palms (*A. recaceae*, *A. recoideae*, *C. ocoseae*) inferred from sequences of six WRKY gene family loci. *Cladistics* **31**, 509–534 (2015). https://doi.org/10.1111/cla.12100
44. 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 (2013b). https://doi.org/10.1111/j.1365-2699.2012.02795.x
45. 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 (2017). https://doi.org/10.1007/s00606-016-1369-4
46. Govaerts, R., & Dransfield, J. *World checklist of palms* (eds. Govaerts, R., Dransfield, J.) 1-223 (Richmond, 2005)
47. Meerow, A. W., Noblick, L., Borrone, J. W., Couvreur, T. L., Mauro-Herrera, M., Hahn, W. J., *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 (2009). https://doi.org/10.1371/journal.pone.0007353
48. 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 (2018). https://doi.org/10.1111/2041-210X.12861
49. R Core Team. R: A language and environment for statistical computing. https://www.R-project.org/ (2021)
50. 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 (2018). https://doi.org/10.3389/fpls.2018.00266

51. Dierckxsens, N., Mardulyn, P. & Smits, G. NOVOPlasty: De novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. **45**, e18 (2017) https://doi.org/10.1093/nar/gkw955
52. Jin, J.J., *et al.* GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol. **21**, 241 (2020). https://doi.org/10.1101/256479
53. Wu, P., *et al.* NOVOWrap: an automated solution for plastid genome assembly and structure standardization. Mol. Ecol. Resour. **21**, 2177–2186 (2021). https://doi.org/10.1111/1755-0998.13410
54. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**, 1754–1760 (2009). https://doi.org/10.1093/bioinformatics/btp324
55. Li, H., *et al.* The sequence alignment/map format and SAMtools. Bioinformatics **25**, 2078–2079 (2009). https://doi.org/10.1093/bioinformatics/btp352
56. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**, 841–842 (2010). https://doi.org/10.1093/bioinformatics/btq033
57. Tillich, M., *et al.* GeSeq - Versatile and accurate annotation of organelle genomes. Nucleic Acids Res. **45**, W6-W11 (2017). https://doi.org/10.1093/nar/gkx391
58. 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 (2012). https://doi.org/10.1093/nar/gkr995
59. 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 (2019). https://doi.org/10.1093/nar/gkz238
60. 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 (2004). https://doi.org/10.1101/gr.2289704
61. Beier, S., Thiel, T., Münch, T., Scholz, U. & Mascher, M. MISA-web: a web server for microsatellite prediction. Bioinformatics **33**, 2583–2585 (2017). https://doi.org/10.1093/bioinformatics/btx198
62. Kurtz, S. & Schleiermacher, C. REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics **15**, 426–427 (1999). https://doi.org/10.1093/bioinformatics/15.5.426
63. 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 (2009). https://doi.org/10.1093/nar/gkp337
64. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucl. Acids Res. **32**, 1792–1797 (2004). https://doi.org/10.1093/nar/gkh340
65. Nylander, J.A.A. Mr.Modeltest v2. https://github.com/nylander/MrModeltest2 (2004)
66. Ronquist, F. *et al.* MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. **61**, 539–542 (2012). https://doi.org/10.1093/sysbio/sys029

Figures

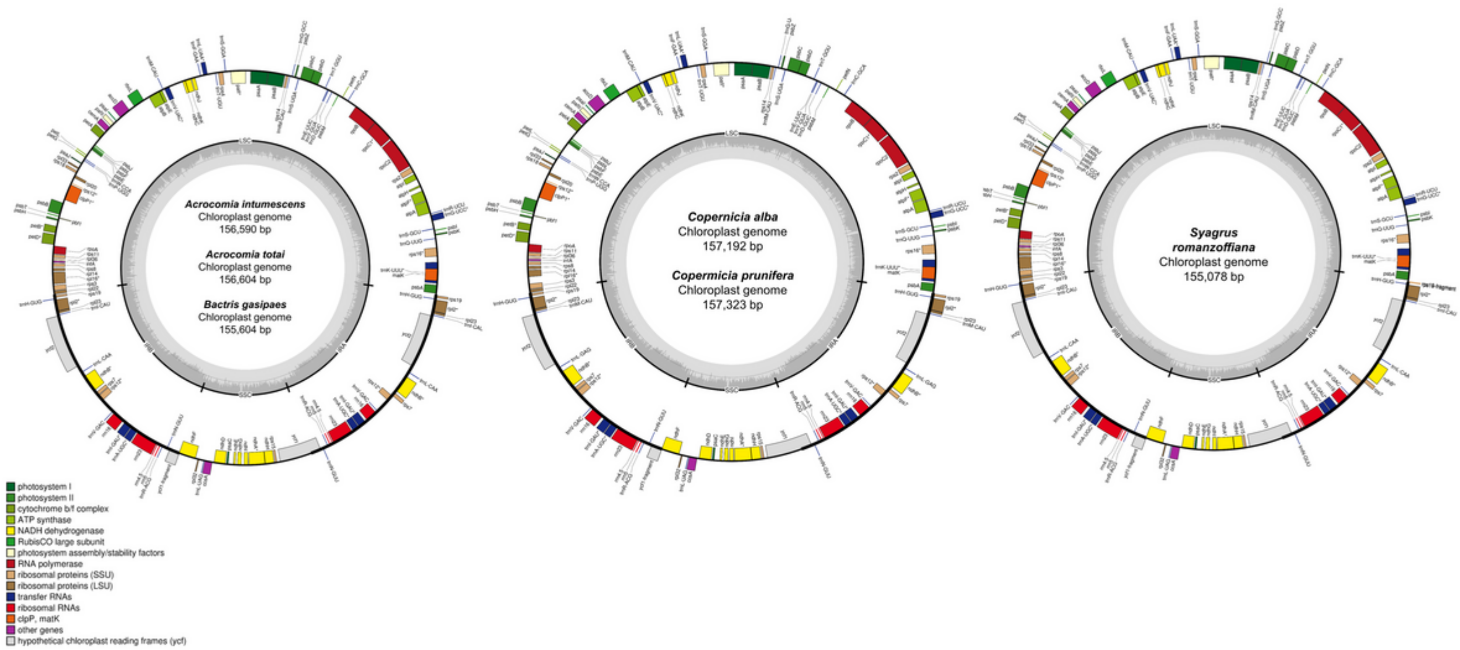

Figure 1

Gene map of *Acrocomia intumescens*, *A. totai, Bactris gasipaes, Copernicia alba, C. prunifera* and *Syagrus romanzoffiana* chloroplast genomes. Genes represented inside the large circle are oriented clockwise and the ones outside are oriented counter clockwise. 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/IRB = Inverted Repeats A and B.

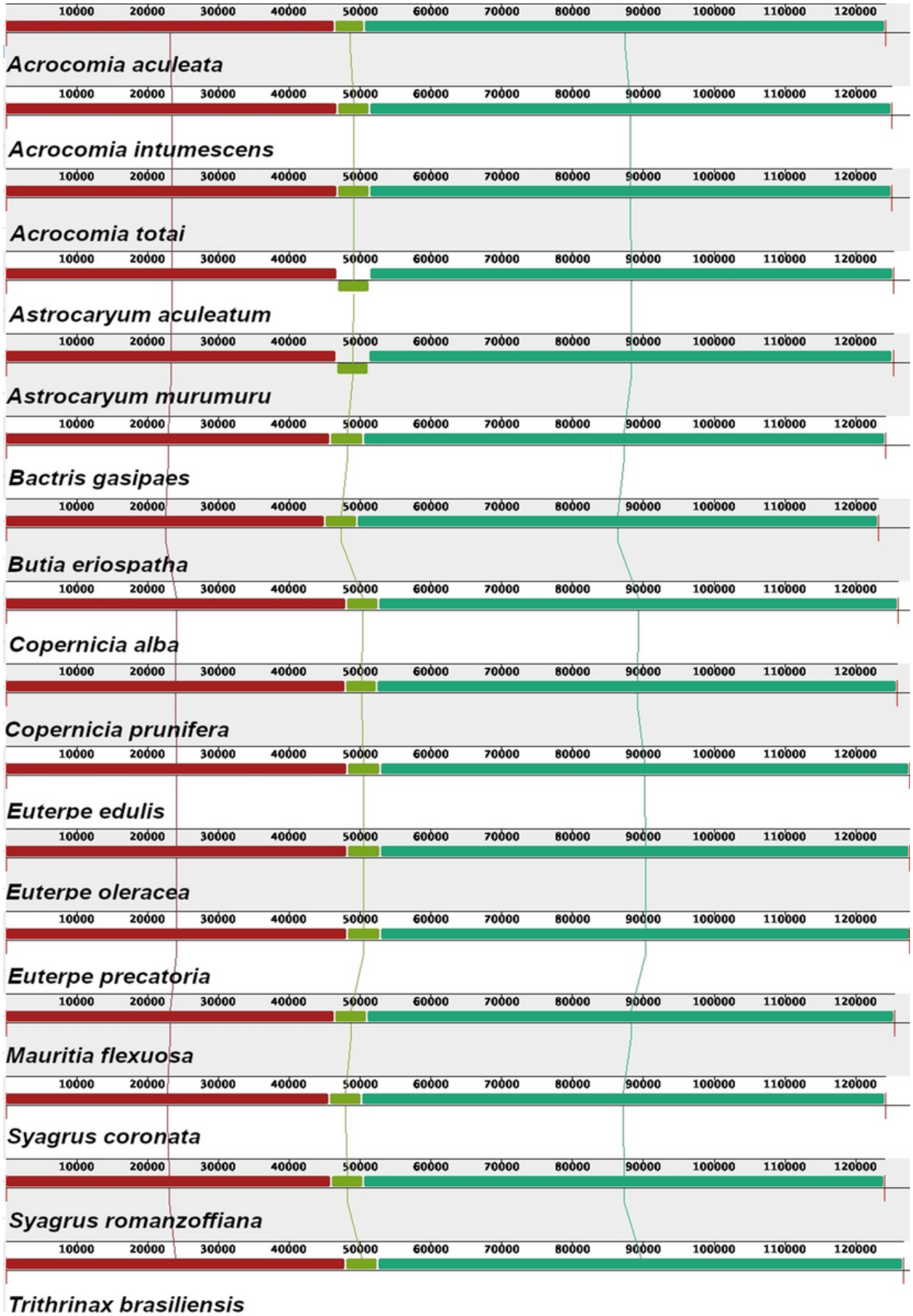

Figure 2

Synteny and divergence in the SSC size detected in Arecaceae chloroplast genomes using the Mauve multiple-genome alignment program. A sample of nine 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 inverse orientation.

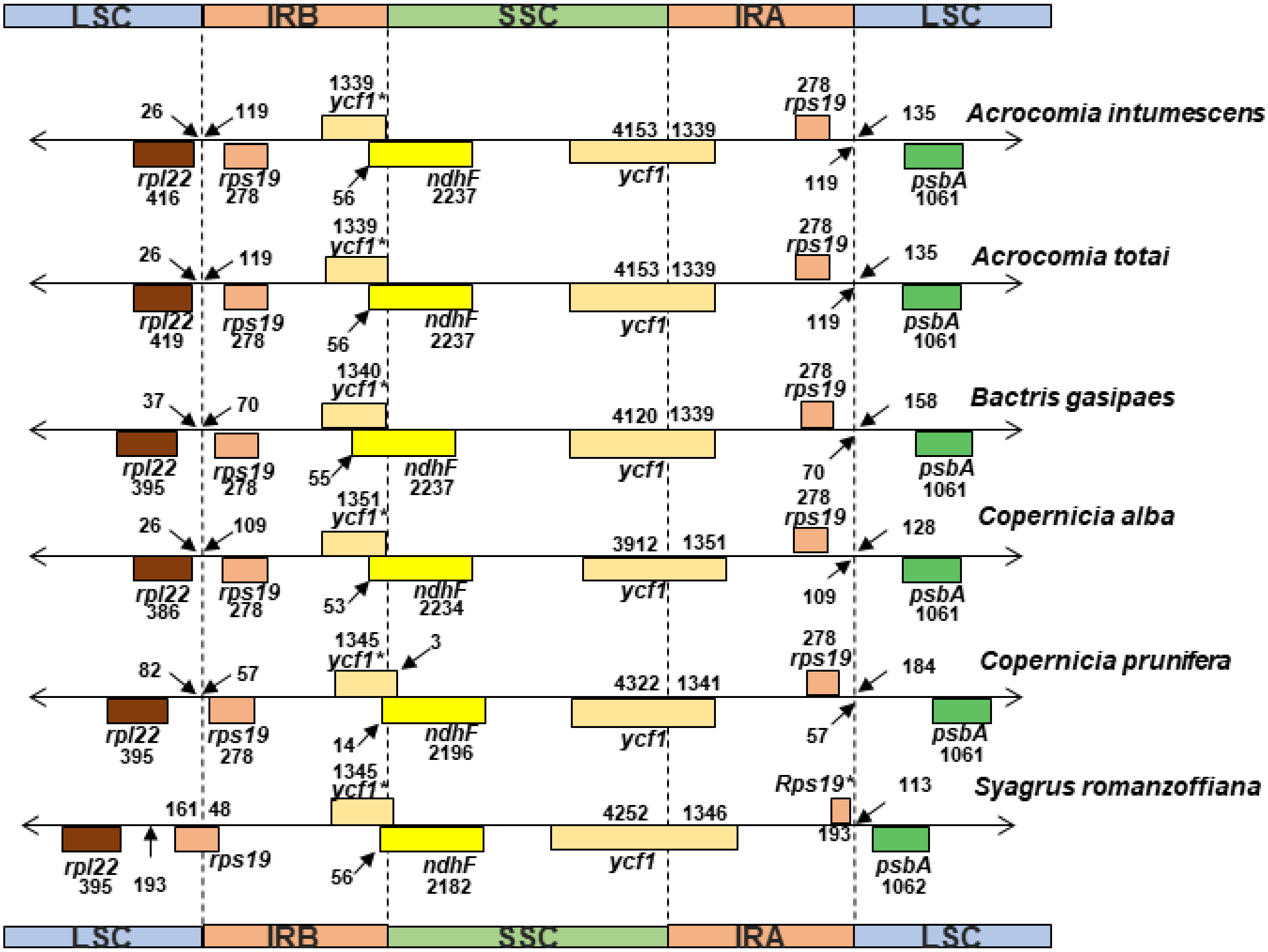

Figure 3

Comparison of the IRA and IRB borders among Brazilian palms. The numbers indicate the lengths of IGSs, genes, and spacers between IR-LSC and IR-SSC junctions. The *ycf1** and *rps19** genes have incomplete CDSs.

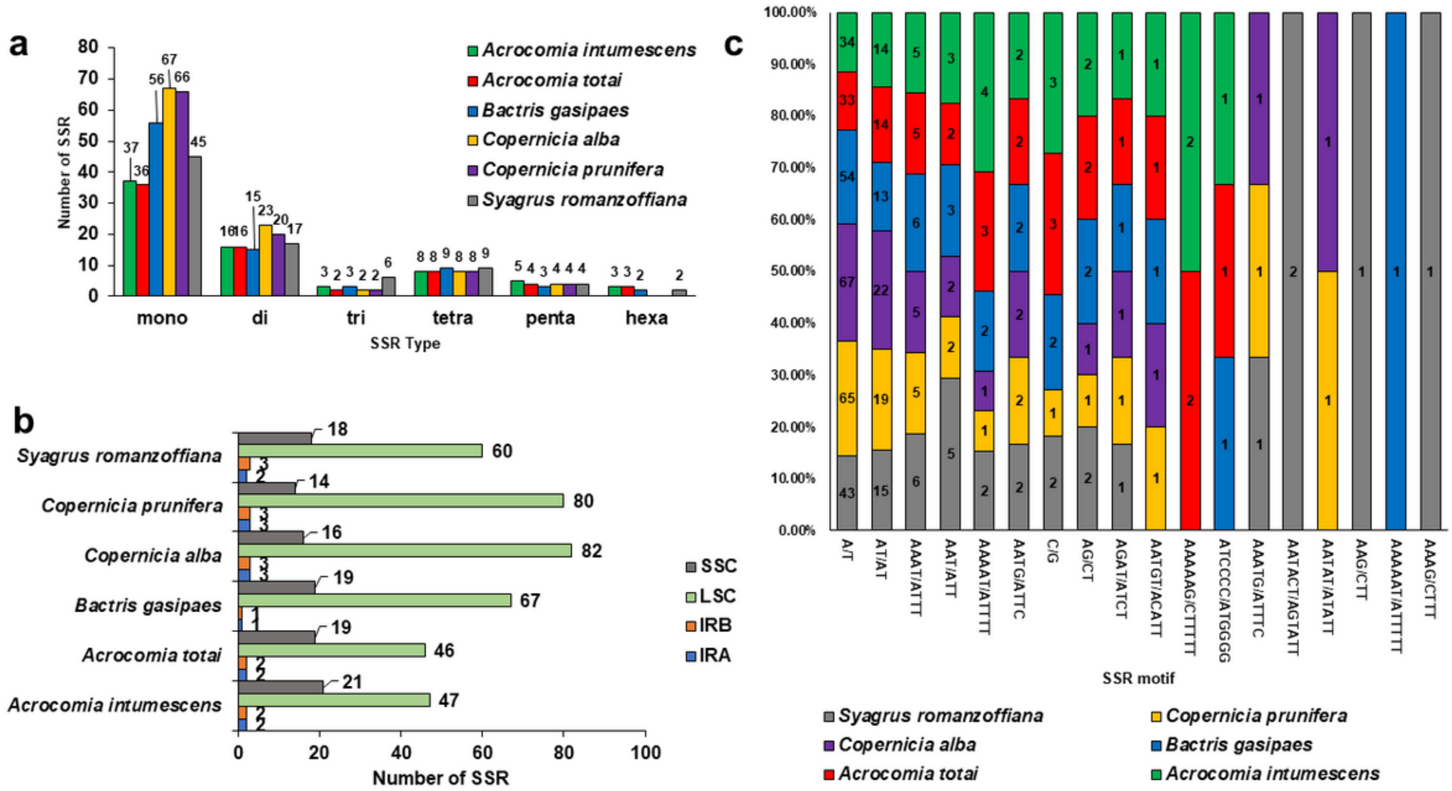

Figure 4

Distribution, classification and motifs of SSR repeats in the chloroplast genomes of *Acrocomia intumensces, A. totai, Bactris gasipaes, Copernicia alba, C. prunifera and Syagrus romanzoffiana* (A) Number of SSR type (mono-, di-, tri-, tetra-, penta- and hexanucleotides) present in the six chloroplast genomes (B) Number of SSR in the different chloroplast genomes regions; (C) Number of different SSRs motifs distributed in the six chloroplast genomes. SSC = Small single copy, LSC = Large single copy, IRA = Inverted repeat A and IRB = Inverted repeat B.

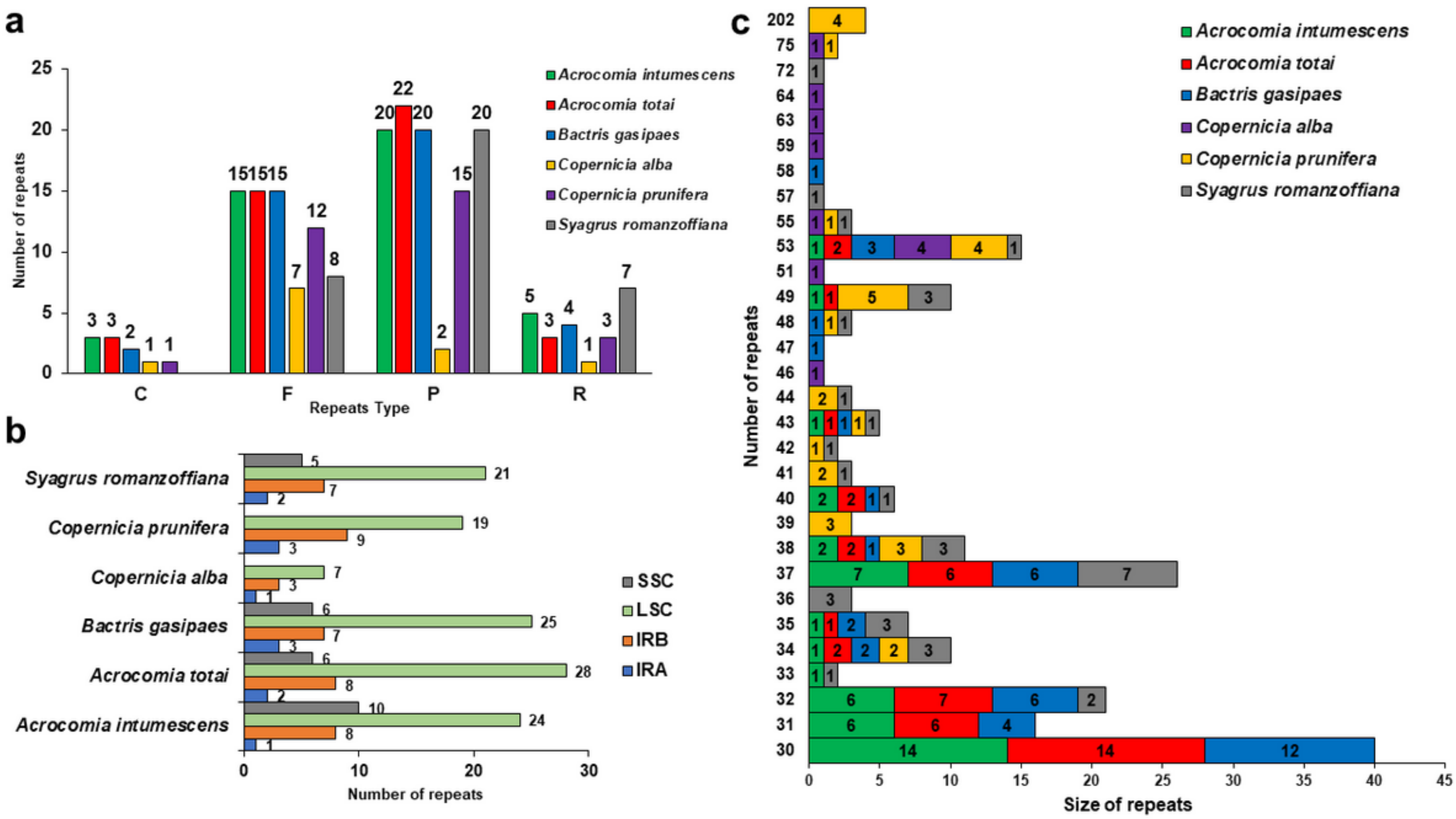

Figure 5

Distribution and classification of dispersed repeats in the chloroplast genomes of *Acrocomia intumesces, A. totai, Bactris 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 genomes regions; (C) Number of dispersed repeats size among the six palm species; (F = Forward, P = Palindrome, R = Reverse and C = Complement, SSC = Small single copy, LSC = Large single copy, IRA = Inverted repeat A and IRB = Inverted repeat B.

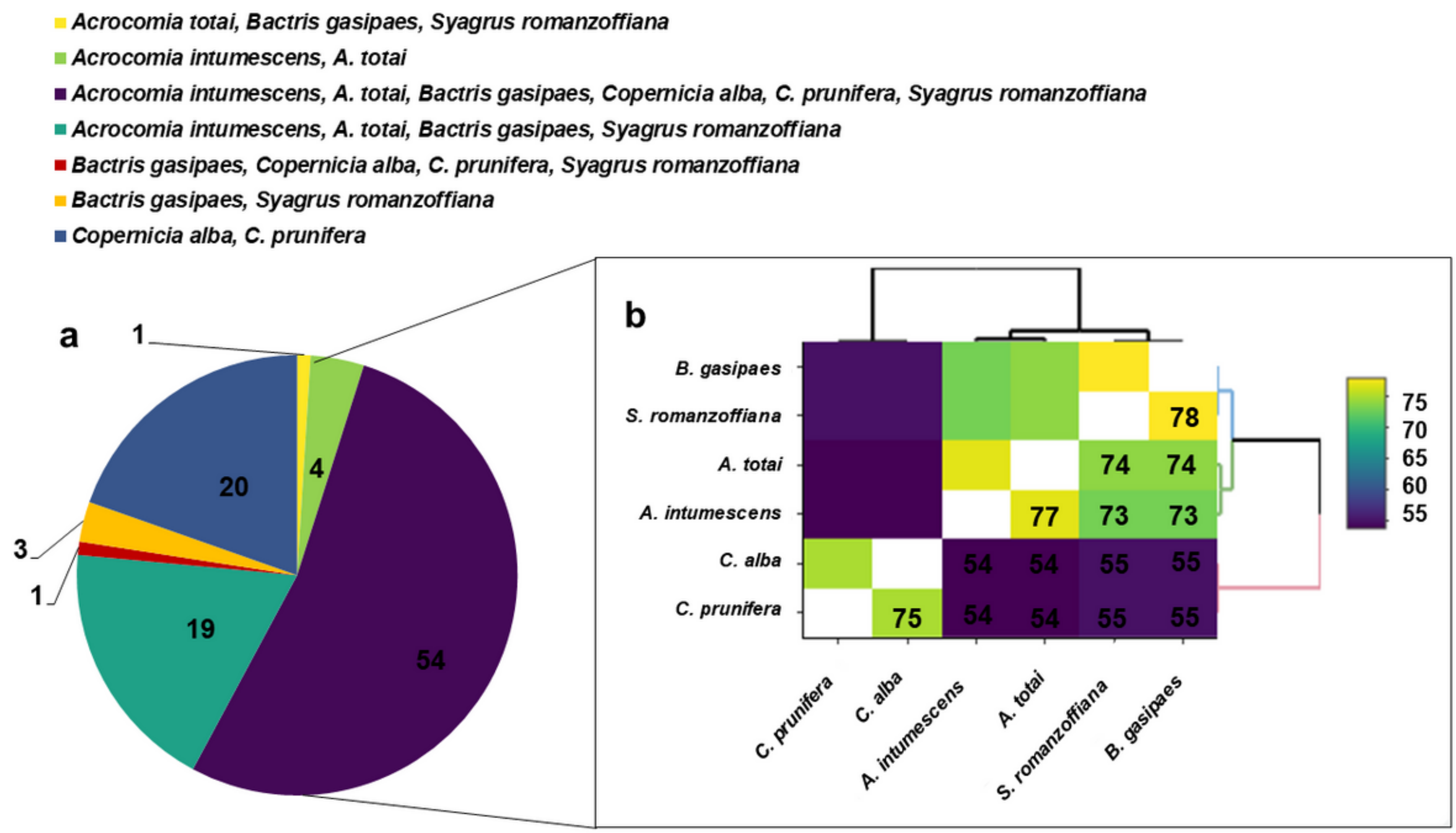

Figure 6

RNA editing sites shared by the species *Acrocomia intumescens, A. totai, Bactris 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.

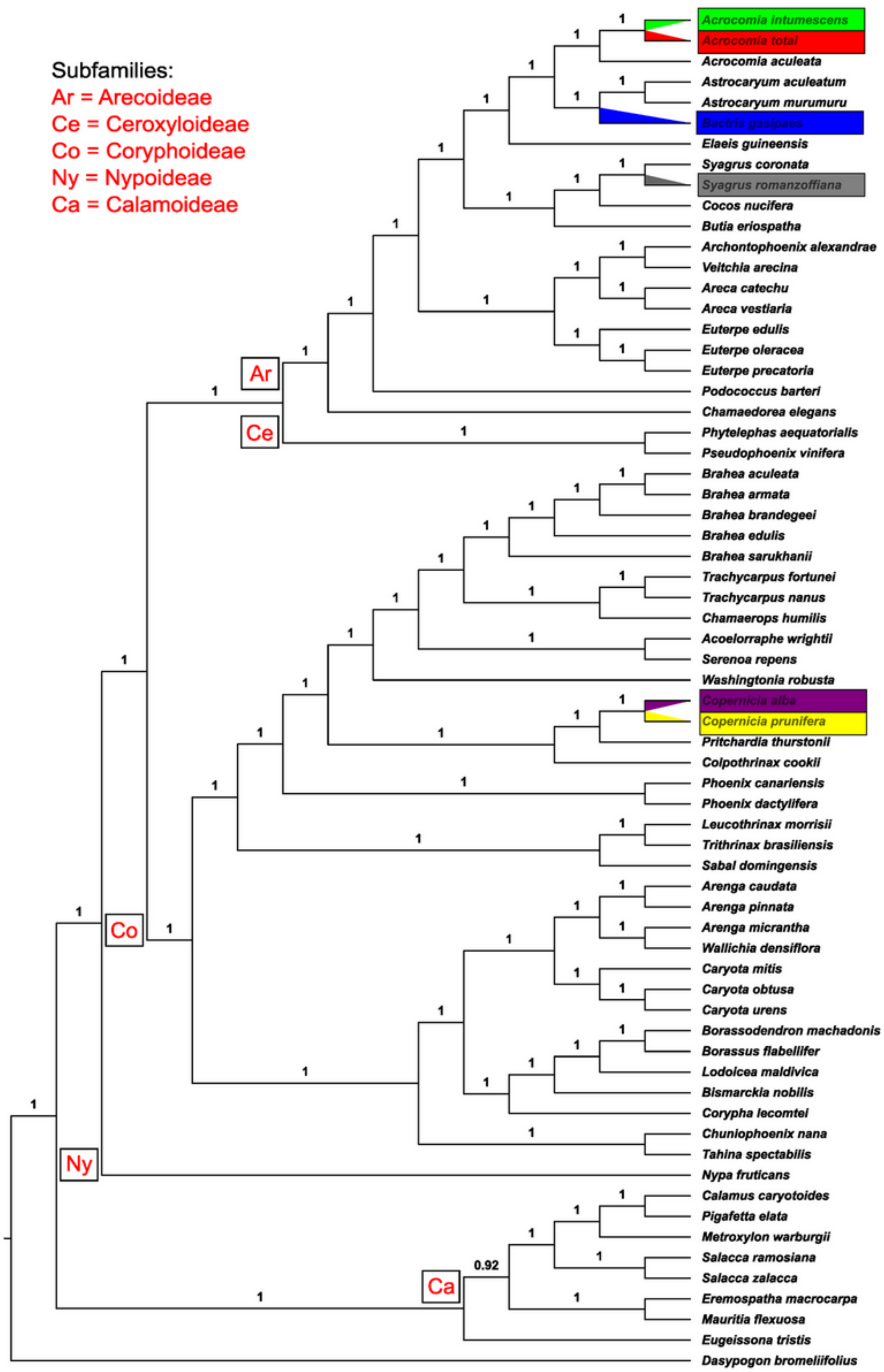

Figure 7

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 branches.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFigures.docx
- TableS1.xlsx
- TableS2.xlsx
- TableS3.xlsx
- TableS4.xlsx