Complete chloroplast genomes and phylogeny in three *Euterpe* palms (*E. edulis*, *E. oleraceae* and *E. precatoria*) from different Brazilian biomes

Ana Flávia Francisconi[1], Luiz Augusto Cauz dos Santos[2], Jonathan Andre Morales Marroquín[1], Cássio van den Berg[3,4], Alessandro Alves-Pereira[5], Luciano Delmondes de Alencar[1], Doriane Picanço-Rodrigues[6], Cesar Augusto Zanello[1], Marcones Ferreira Costa [1,7], Maria Teresa Gomes Lopes[8], Elizabeth Ann Veasey[4], Maria Imaculada Zucchi[9]

[1] Programa de pós-gradução em Genética e Biologia Molecular Universidade Estadual de Campinas, Campinas, São Paulo, Brasil

[2] Department of Botany and Biodiversity Research, University of Vienna, Wien, Austria

[3] Departamento de Ciências Biológicas, Universidade Estadual de Feira de Santana, Feira de Santana, Bahia, Brasil

[4] Departamento de Genética, Universidade de São Paulo, Piracicaba, São Paulo, Brasil

[5] Departament de Biologia Vegetal, Universidade Estadual de Campinas, Campinas, São Paulo, Brasil

[6] Departamento de Biologia, Universidade Federal do Amazonas (ICB-UFAM), Manaus, Amazonas, Brasil

[7] Universidade Federal do Piauí, Floriano, Piauí, Brasil

[8] Departamento de Produção Animal e Vegetal, Universidade Federal do Amazonas, Manaus, Amazonas, Brasil

[9]Agência Paulista de Tecnologia dos Agronegócios, Piracicaba, São Paulo, Brasil

*Corresponding author
E-mail: mizucchi@sp.gov.br (MIZ), anaf_f@hotmail.com (AFF)

Abstract

The Brazilian palm fruits and hearts-of-palm of *Euterpe edulis, E. oleracea and E. precatoria* are an important source for agro-industrial production, but overexploitation requires conservation strategies to maintain genetic diversity. Chloroplast genomes have conserved sequences, which are useful to explore evolutionary questions. Besides the plastid DNA, genome skimming allows the identification of other genomic resources, such as single nucleotide polymorphisms (SNPs), providing information about the genetic diversity of species. We sequenced the chloroplast genome and identified gene content in the three *Euterpe* species. We performed comparative analyses, described the polymorphisms among the chloroplast genome sequences (repeats, indels and SNPs) and performed a phylogenomic inference based on 55 palm species chloroplast genomes. Finally, using the remaining data from genome skimming, the nuclear and mitochondrial reads, we identified SNPs and estimated the genetic diversity among these *Euterpe* species. The *Euterpe* chloroplast genomes varied from 159,232 to 159,275 bp and presented a conserved quadripartite structure with high synteny with other palms. In a pairwise comparison, we found a greater number of insertions/deletions (indels = 93 and 103) and SNPs (284 and 254) between *E. edulis*/*E. oleracea* and *E. edulis*/*E. precatoria* when compared to *E. oleracea*/*E. precatoria* (58 indels and 114 SNPs). Also, the phylogeny indicated a closer relationship between *E. oleracea*/*E. precatoria*. The nuclear and mitochondrial genome analyses identified 1,077 SNPs and high divergence among species (F_{ST} = 0.77), especially between *E. edulis* and *E. precatoria* (F_{ST} = 0.86). These results showed that, despite the few structural differences among the chloroplast genomes of these *Euterpe* palms, a differentiation between *E. edulis* and the other *Euterpe* species can be identified by point mutations. This study not only brings new knowledge about the

evolution of *Euterpe* chloroplast genomes, but also these new resources open the way for future phylogenomic inferences and comparative analyses within Arecaceae.

Key-Words
Arecaceae, SNPs, genome-skimming, genome comparison, polymorphism

Introduction

The palm family (Arecaceae) comprises 188 genera and ca. 2,400 species, distributed throughout tropical and subtropical regions [1,2]. Palms are often keystone species, providing ecosystem services, shaping their environment and offering many products used for fabrics, fuel, food, medicine and as ornamentals [2–4]. Within Arecaceae, the genus *Euterpe* originated in South America and includes seven species distributed throughout Central America and tropical South America [2,5,6]. In Brazil, among the native species of *Euterpe*, *E. edulis* Mart., *E. oleracea* Mart. and *E. precatoria* Mart. are the most important from an agro-industrial point of view [7].

Euterpe edulis is endemic to the Atlantic Forest and has substantial importance for the functioning of the ecosystem [8,9]. However, the overexploitation of heart-of-palm [10,11] has placed this palm species in the Brazilian Flora Red Book [12] as "vulnerable (VU)", which makes conservation strategies for it extremely necessary. Recently, as an alternative to palm heart-of-palm extraction, local communities have started to cultivate this species for fruit pulp extraction, due to the similarity with the Amazonian açaí (*E. oleracea* and *E. precatoria*) [9,10,13]. This practice has been fundamental to capture and conserve the species genetic diversity present in older forests [10].

Euterpe oleracea and *E. precatoria* are the main sources for the production of açaí fruit pulp, estimated to be 760,000 tons in 2017, mostly in the northern region of Brazil (98,36% of the national production) [14,15]. *E. oleracea* is a multiple-stemmed palm,

commonly found in the Amazon estuary, especially prevalent in the Brazilian states of Pará, Maranhão and Amapá [15–17]. However, in the last 20 years, the intense demand for açaí fruit in Pará resulted in a significant loss of local tree species richness, which presumably is occurring after decades of thinning to reduce interspecific competition with açai palm trees [18]. This resulted in a reduction of pollinators, which in turn decreased the fruit production [19]. *E. precatoria* is a single stemmed palm, predominantly found in the Brazilian states of Amazonas, Acre and Rondônia [20], in non-flooded areas [21]. This species is highly promising for açaí fruit extraction, as the state of Amazonas is already the second largest producer of Brazil [15]. Therefore, the establishment of strategies for *E. precatoria* domestication, conservation and management of populations is extremely necessary, in addition to the support of riverside communities and farmers to conserve this species based on their genetic diversity [15].

The fast growth of açaí (*E. oleracea* and *E. precatoria*) pulp production in the last decades and the threats of extinction for *E. edulis* indicate an urgent demand for new genomic resources in this genus, contributing to the understanding of the evolution and population dynamics of the species, which can result in the best planning of sustainable development and conservation [9,15]. Fortunately, high-throughput sequencing technologies have revolutionized how genomic data can be obtained for plant species and have allowed the complete assembly of the chloroplast genome for an increasing number of non-model species [22–24], while providing useful information for the study of genetic variation of native species.

The common structure of the chloroplast genomes typically has four parts: two single copy regions, one large (LSC) and one small (SSC), and a pair of inverted regions (IRs) [25] which basically consist of long circular or linear molecules (120 to 180 kb), usually conserved across palms species [26]. Apart from the conservative nature and low

120 nucleotide substitution rates of the chloroplast sequences [27], the description of this
121 organellar genome is important to identify repetitive sequences and polymorphic regions,
122 useful to further obtain molecular markers, which can be applied to assess genetic
123 structure and diversity of natural populations [28,29]. Whole chloroplast genomes and
124 chloroplast sequences have been used to infer phylogenetic relationships and investigate
125 diversification patterns [24,30–32]. The tribe Euterpeae was already analyzed with
126 combined chloroplast and nuclear sequence data [33], and whole chloroplast genomes
127 have been reported in reference for two species in the *Euterpe* genus (*E. oleracea* and *E.*
128 *edulis*) [34]. These priors research emphasize the differentiation of *E. edulis*, due to its
129 distribution in a different Brazilian biome (Atlantic Forest biome). But, the addition of
130 new chloroplast genomes can give more consistency to the evolutionary processes of the
131 genus.

132 Moreover, the amount of plastid sequences present in a data from total genomic
133 DNA extraction depends on different factors and can vary substantially, from 0.4% to
134 29.5% [23], and the recovery of these sequences can occur through the use of genome
135 skimming. Therefore, the remaining nuclear and mitochondrial sequences can provide
136 genomic resources [27,35], representing useful information to study the genetic variation
137 of native species.

138 Considering the effectiveness of genome skimming to produce genomic data, we
139 employed this approach to obtain three newly sequenced complete chloroplast genomes,
140 thereby facilitating the study of genetic diversity from *Euterpe* species. We were able to
141 investigate: i) the plastid organization in *Euterpe* and the synteny level with other
142 available Arecaceae species; ii) the repetitive sequences and polymorphic regions in the
143 chloroplast genomes; iii) the Arecaceae family evolution based on a phylogenomic study
144 with complete chloroplast genomes; and iv) the species divergence using the remaining

145 nuclear and mitochondrial data from genome-skimming.

146

147 Material and Methods

148 Leaf material

149 The leaf material from *Euterpe edulis* and *E. oleracea* were collected from the *ex*
150 *situ* collection at the Escola Superior de Agricultura "Luiz de Queiroz" (ESALQ),
151 Universidade de São Paulo (USP) in Piracicaba, São Paulo, Brazil
152 (http://www.esalq.usp.br/trilhas/palm/) and the sample of *E. precatoria* was collected at
153 the Instituto Nacional de Pesquisas da Amazônia (INPA), in Manaus, AM, Brazil. The
154 collections were registered according to the Brazilian laws (SISGEN number A411583,
155 Brazil).

156

157 Intact chloroplast isolation in sucrose gradient and chloroplast 158 DNA extraction

159 The chloroplast organelles were isolated using the sucrose gradient method [36].
160 About 20 g of fresh leaves from each species were frozen with liquid nitrogen and
161 macerated. The material was resuspended in 200 mL of isolation buffer (50 mM Tris-HCl
162 pH 8.0, 0.35 M sucrose, 7 mM EDTA, 5 mM 2-mercaptoethanol and 0.1% BSA) and
163 incubated for 10 min in the dark. The suspension was filtered through two layers of
164 Miracloth (Merck), and the filtrate was centrifuged at 1,000 × g for 10 min.

165 The pellet was resuspended in 5 mL of isolation buffer and the suspension slowly
166 laid out in the density gradients of 20/45% sucrose in 50 mM Tris-HCl (pH 8.0), 0.3 M
167 sorbitol and 7 mM EDTA. The gradients were centrifuged at 2000 × g for 30 min, and
168 the green band formed at the interface containing intact chloroplasts were collected. The

169 solution containing the chloroplasts were diluted in three volumes of buffer and
170 centrifuged at 3,000 × g for 10 minutes to obtain the purified chloroplasts in the pellet.
171 The pellet was resuspended in 2% CTAB buffer to promote lysis. The suspension
172 was incubated and stirred at 65ºC for 1 h. The supernatant was extracted twice with an
173 equal volume of chloroform: isoamyl alcohol (24: 1) and centrifuged at 10,000 × g for 20
174 min. An equal volume of isopropanol was added and incubated at 20ºC for 1h. Finally,
175 the aqueous phase was centrifuged for 10,000 × g for 20 min, and the chloroplast DNA
176 (cpDNA) pellet was washed with ethanol (70%), dried and resuspended with 40 µL TE
177 (1 M Tris-HCl, 0.5 M EDTA, pH 8).
178

179 Chloroplast genome sequencing, assembly and annotation

180 The genomic libraries were constructed using 100 ng of cpDNA and the Nextera
181 DNA Flex kit (Illumina), following the manufacturer's instructions. Paired-end
182 sequencing (2x 150 bp) was performed on the Illumina NextSeq550 platform (Fundação
183 Hemocentro de Ribeirão Preto, Brazil).
184 The assembly was conducted in three steps: First, the filtered reads from *Euterpe*
185 *oleracea* were assembled in NOVOPlasty v 4.2 [37]
186 (https://github.com/ndierckx/NOVOPlasty) using the *rbcL* gene sequence as a seed
187 (NCBI accession number: MN621452.1) and the chloroplast genome of *Acrocomia*
188 *aculeata* (NCBI accession number: NC_037084.1), a native Brazilian palm, as a reference
189 to ordinate the contigs. Subsequently, all the three chloroplast genomes (*E. precatoria, E.*
190 *edulis* and *E. oleracea*) were assembled in GetOrganelle v 1.7.3.1 [38]
191 (https://github.com/Kinggerm/GetOrganelle/) using the *E. oleracea* chloroplast sequence
192 obtained in NOVOPlasty as seed. Briefly, in GetOrganelle, the chloroplast genomes were
193 assembled using default settings, starting with the recruitment of target-associated reads

194 using Bowtie2 [39]. In this step, the seed chloroplast genome was used as target for the
195 extension in five iterations. Then, the total target-associated reads were *de novo*
196 assembled into a fasta assembly graph using SPAdes [40]. The *de novo* assemblies were
197 trimmed and used to calculate all possible paths of a complete organelle genome [38].
198 Finally, the correctness and coverage of the assembly was assessed and confirmed in
199 Geneious v2020 2.4. (https://www.geneious.com/, last assessed January, 2021). We used
200 the "Map to reference" function to map the paired-end raw data onto the final assembled
201 chloroplast genomes.

202 The chloroplast genome annotation was performed in GeSeq (Organellar Genome
203 Annotation) [41] from the Chlorobox platform, with settings for the identification of
204 protein coding sequence (CDS), rRNAs and tRNAs based on reference chloroplast
205 sequences and homologies through BLAST search. Following the automatic annotation,
206 a manual correction of start and stop codons and a verification of pseudogenes and intron
207 positions were performed using GenomeView [42]. We then obtained the chloroplast
208 circular genome maps using OGDRAW [43].

209

210 Chloroplast genome structure comparison

211 To perform a comparative study and access the synteny between the obtained
212 chloroplast sequences, we used a Perl script of MUMmer4 [44]
213 (https://github.com/mummer4/mummer) to align the chloroplast genome from *E. edulis,*
214 *E. oleracea* and *E. precatoria* with the function NUCmer. This analysis enables the
215 identification of the conserved regions among sequences between species. The results
216 were visualized in dot plots created by the function MUMmerplot. Additionally, to
217 compare and align the *Euterpe* chloroplast sequences obtained with the GenBank
218 sequences [34], MAFFT v.7 [45] was used as a way to identify differences between them.

219 The location of the additional sequences found between them was also annotated in
220 GeSeq [41].

221 Taking into account the species relationships from Arecoideae subfamily, and the
222 previous evidence from palms chloroplast genome structures [28,46], we decided to
223 conduct two multiple progressive sequence alignment in Mauve v.2.4.0 [47]. The first
224 one only included chloroplast genomes from Brazilian native species of subfamily
225 Arecoideae: *Euterpe edulis*, *Euterpe oleracea*, *Syagrus coronata*, *Astrocaryum*
226 *aculeatum*, *Astrocaryum murumuru* and *Acrocomia aculeata*. The second analysis was
227 carried out using 17 chloroplast genomes from different palm species (S1 Table) available
228 in GenBank. Taking into the account the evolution of the group, we selected the species
229 that represented the five palm subfamilies: *Phytelepas aequatoriallis* and *Pseudophoenix*
230 *vinifera* (Subfamily: Ceroxyloideae); *Trachycarpus fortune* and *Caryota mitis*
231 (Subfamily: Coryphoideae); *Nypa fruticans* (Subfamily: Nypoideae); *Calamus*
232 *caryotoides* and *Eresmopatha macrocarpa* (Subfamily: Calamoideae); *Veitchia arecina*
233 (Subfamily: Arecoideae) and the Brazilian native species also from the Arecoideae
234 subfamily.

235 Expansions and contractions in the inverted repeats (IR) regions from the
236 chloroplast genome structure were also explored. The chloroplast genome junctions
237 (IRB/LSC; IRB/SSC, SSC/IRA; IRA/LSC) from the *Euterpe* species, and other Brazilian
238 palm species from subfamily Arecoideae were examined to identify differences between
239 individuals within the same genus and among subfamilies.

240

241 Identification of SSRs, dispersed repeats, indels, SNPs and 242 nucleotide divergence hotspots

243 Simple sequence repeats (SSR) consisting of 1-6 nucleotide units were carefully
244 determined using the web package MISA (available at https://webblast.ipk-
245 gatersleben.de/misa/) [48]. The criteria to search SSR motifs were: SSR of one to six
246 nucleotides long, with a minimum repeat number of 10, 5, and 4 units for mono-, di-, and
247 trinucleotide SSRs, respectively, and three units for tetra-, penta- and hexanucleotide
248 SSRs. The SSRs sequences and location were compared among the species. The
249 dispersed repeats (forward, reverse, palindrome and complement sequences) identified in
250 REPuter [49] were based on the following criteria: minimum repetition size ≥ 30 bp and
251 sequence identity ≥ 90% (Hamming distance = 3). Posteriorly, the position of the SSRs
252 and repeats were manually compared with the gene annotation of each chloroplast
253 genome.

254 MAFFT v.7 [45] was used to obtain pairwise alignments between the chloroplast
255 genomes to pinpoint small insertions/deletions (indels) in the sequences. The alignment
256 between the three species was also used to identify single nucleotide polymorphisms
257 (SNPs) and nucleotide divergence hotspots with DnaSP v.5 [50]. Specific coding genes
258 with a high number of SNPs were aligned with the software Muscle [51] in the Selecton
259 Server [52], which was also used to identify the synonymous (Ka) or non-synonymous
260 (Ks) mutations. With this analysis, according with the Ka/Ks ratios, it was possible to
261 determine positive (Ka/Ks > 1) or purifying (Ka/Ks < 1) selection, estimated under the
262 evolutionary model M8.

263 A sliding window analysis (window length of 200 bp and step size of 50 bp) was
264 conducted to find nucleotide divergence hotspots. All the positions of indels, SNPs and
265 divergence hotspots were manually identified using the annotation of the aligned
266 chloroplast genomes, performed previously in GeSeq (Organellar Genome Annotation)
267 [41].

Phylogenomic studies

Plastid sequences of 54 palm species plus one outgroup (*Daypogon bromeliifolius*, Dasypogonaceae) (S1 Table) curated annotated features in Genbank format were separated into genes and entered into a SQLITE database using a custom Python script (available on demand from the corresponding author). Sequences of all putative coding proteins were then grouped by species and aligned with MAFFT, and later assembled into an interleaved NEXUS file for phylogenetic analyses. Species with missing sequences were filled with missing data in each gene matrix. Each block of gene sequences was then manually checked for start and stop codons and evidence of non-coding behaviour. For the coding genes, we annotated first, second and third positions for each codon, but the regions *cemA* and *rpl16* presented strange start codons and non-triplet insertions and were not annotated for codon positions. We tested three *ad hoc* partitioning schemes for models of molecular evolution: i) single model for the whole matrix, ii) four partitions: 1st, 2nd, 3rd codon positions and a partition for *cemA +rpl16* (not split into codon positions), iii) five partitions: 1st, 2nd, 3rd codon positions, *cemA* (not split into codon positions), *rpl16* (not split into codon positions). The assessment of molecular evolution model for each partition was calculated in the different partition schemes with the Akaike Information Criterion (AIC) in MrModelTest 2.0 [53], but they were all GTR+I+G, probably due to the complexity of mixing different genes in each partition. The assessment of the best partition scheme was then made using stepping stone (SS) [54] which is more accurate than taking the harmonic means of the likelihoods for model comparison. Bayesian inference were carried out with MrBAYES 3.2.7 [55] in the CIPRES platform [56], with one run of two chains, and 10×10^6 generations, sampling one tree every 1,000. Marginal likelihoods of the SS runs were then compared using the standard Bayes Factors scale

293 [57]. The same partition schemes were run separately to estimate the phylogeny with two
294 runs of four chains (three hot and one cold), and 20 × 10^6 generations, sampling one tree
295 every 1,000 and discarding 25% initial generations for burn-in. The remaining trees were
296 checked for estimated sample sized (ESS) > 200 in all parameters, and the final majority-
297 rule was computed with MrBayes, rooted with the single outgroup *D. bromeliifolius* and
298 the tree was exported to FigTree 1.4 [58] for drawing, and later improved with InkScape
299 [59] for designing the figures.
300

301 Genetic variation and comparative analyses with nuclear and 302 mitochondrial sequences

303 After the assembly of the chloroplast genomes, the reads were also used to obtain
304 SNPs for preliminary comparative analysis between species. Initially, the raw reads of
305 each species were aligned in their respective assembled chloroplast genome using
306 Bowtie2 [39]. Subsequently, the non-aligned sequences (non-chloroplast) were obtained
307 with SAMTools [60], and converted to fastq file using Picard Tools program [61].
308 The reads corresponding to non-chloroplast genome (nuclear and mitochondrial
309 data) were cleaned using *process_shortreads* from the Stacks package v.1.42 [62].
310 Sequences from each species were used to build loci with a minimum depth of three and
311 up to two mismatches, applying the *ustacks* program (-m 3, -M 2) [62]. Subsequently, a
312 catalog was built from each species, allowing for up to two mismatches (-n 2, cstacks)
313 [62], and using *sstacks* [62] loci were matched with the catalog of each species. *Rxstacks*
314 [62] was used as a correction step, assuming a mean log likelihood of -10 to discard loci
315 with lower probabilities. The *populations* [62] program was administered for a final
316 filtering to retain loci with a maximum missing data of 5%, and also used to verify the

317 number of private alleles (*Ap*), observed heterozygosity (H_O) and expected heterozygosity
318 (H_E; S1 Fig).
319 Finally, the mean sequence depth and the mutation counts of the SNPs were
320 determined in vcftools [63]. The number of alleles (*A*) in each species and the genetic
321 differentiation (F_{ST}) between them were estimated with *Adegenet* v 1.3 -1 [64] and
322 *Genpop* v 1.1.7 [65] for the platform R v 4.0.3 [66] .
323

324 Results

325 Organization of the three *Euterpe* species chloroplast genomes

326 The chloroplast genomes of *E. edulis, E. oleracea* and *E. precatoria* had the
327 typical quadripartite structure (Fig 1), with the presence of two copies of inverted repeat
328 regions (Inverted Repeats A and B = IRA and IRB) and two single copy regions (Large
329 Single Copy = LSC and Small Single Copy = SSC). *E. precatoria* had the largest
330 chloroplast genome size (159,275 bp) and the largest LSC and SSC (87,282 bp and 17,756
331 bp, respectively), while *E. edulis* had the largest IR regions and *E. oleracea* a slightly
332 larger amount of GC content (37,3%) (Table1). All three chloroplast genome annotations
333 resulted in the identification of 113 unique genes, with 30 tRNAs, 4 rRNAs and 79 protein
334 coding genes (Fig 1, Table 2).
335

336 **Table 1. General features of chloroplast genomes of three *Euterpe* species.**

Species	Total cpDNA size (bp)	Length of LSC region (bp)	Length of IR region (bp)	Length of SSC region (bp)	GC content (%)	Protein coding genes (bp)	tRNA coding genes (bp)	rRNA coding genes (bp)	Introns (bp)	Intergenic Regions (bp)
Euterpe edulis	159,232	87,237	54,280	17,715	37.20	80,462	2,880	9,052	18,068	48,770
Euterpe	159,260	87,273	54,232	17,755	37.30	80,415	2,881	9,052	18,062	48,850

oleracea										
Euterpe precatoria	159,275	87,282	54,234	17,759	37.20	80,411	2,881	9,052	18,064	48,867

Table 2. Gene content in *Euterpe edulis, E. oleracea* and *E. precatoria* chloroplast genomes according to each respective category.

Category	Gene
Subunits of photosystem I	*psaA*, *psaB*, *psaC*, *psaI*, *psaJ*
Subunits of photosystem II	*psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbJ*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*, *psbZ*
Subunits of cytochrome b/f complex	*petA*, *petB*[a], *petD*[a], *petG*, *petL*, *petN*
Subunits of ATP synthase	*atpA*, *atpB*, *atpE*, *atpF*[a], *atpH*, *atpI*
Large subunit of rubisco	*rbcL*
Subunits of NADH-dehydrogenase	*ndhA*[a], *ndhB*[a,b], *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *ndhK*
Proteins of large ribosomal subunit	*rpl2*[a,b], *rpl14*, *rpl16*[a], *rpl20*, *rpl22*, *rpl23*[b], *rpl32*, *rpl33*, *rpl36*
Proteins of small ribosomal subunit	*rps2*, *rps3*, *rps4*, *rps7*[b], *rps8*, *rps11*, *rps12*[a,b,c], *rps14*, *rps15*, *rps16*[a], *rps18*, *rps19*[b]
Subunits of RNA polymerase	*rpoA*, *rpoB*, *rpoC1*[a], *rpoC2*
Maturase	*matK*
Translational initiation factor	*infA*
Protease	*clpP*[a]
Envelope membrane protein	*cemA*
Subunit of acetyl-CoA carboxylase	*accD*
Cytochrome c biogenesis	*ccsA*
Conserved hypothetical genes	*ycf3*[a], *ycf4*
Component of TIC complex	*ycf1*
Component of 2-MD heteromeric AAAATPase complex	*ycf2*[b]
Ribosomal RNAs	*rrn4.5*[b], *rrn5*[b], *rrn16*[b], *rrn23*[b]
Transfer RNAs	*trnA –UGC*[a,b]; *trnC –GCA*; *trnD –GUC*; *trnE –UUC*; *trnF –GAA*; *trnfM –CAU*; *trnG –UCC*[a] *trnG –GCC*; *trnH –GUG*[b]; *trnI –CAU*[b]; *trnI –GAU*[a,b]; *trnK –UUU*[a]; *trnL –CAA*[b]; *trnL –UAA*[a]; *trnL –UAG*; *trnM –CAU*; *trnN –GUU*[b]; *trnP –UGG*; *trnQ –UUG*; *trnR –ACG*[b]; *trnR–UCU*; *trnS –GCU*; *trnS –UGA*; *trnS –GGA*; *trnT –UGU*; *trnT –GGU*; *trnV –GAC*[b]; *trnV–UAC*[a]; *trnW –CCA*; *trnY –GUA*

[a]Intron-containing gene; [b]Two gene copies in the Irs; [c]Gene divided into two independent transcription units.

Fig 1. Gene map of the *Euterpe edulis*, *E. oleracea* and *E. precatoria* chloroplast genomes. Genes represented inside the large circle are oriented clockwise and the ones

343 outside are oriented counter clockwise. The different colors represent functional groups,
344 and the darker gray in the inner circle indicates the GC content. The quadripartite structure
345 is also reported as: LSC = Large Single Copy, SSC = Small Single Copy, IRA/IRB =
346 Inverted Repeats A and B.

347 Considering the duplicated genes in the IRs, more than 50% of the three
348 chloroplast genome sequences are from protein coding regions (86 genes, Table 1 and
349 Table 2). Among genes, *ycf2* presented a slight difference between species, 6,903 bp in
350 *Euterpe edulis* and 6,879 bp in *E. oleracea* and *E. precatoria*. Furthermore, chloroplast
351 tRNA and rRNA were conserved among species, constituting 1.8% and 5.7% of their
352 sequences, respectively. Introns constituted ca. 12% and intergenic regions represented
353 almost 31% of the chloroplast genomes (Table 1).

354 All chloroplast genomes had 17 unique genes (11 protein- and six tRNA-coding
355 genes) with introns, five duplicated genes in the inverted repeats regions with introns, and
356 two introns in the *ycf3* and *clpP* genes. Accounting for the lengths, the largest intron was
357 identified in *trnK – UUU* (2,621 bp – *E. edulis*; 2,615 bp – *E. oleracea*; 2,618 bp – *E.*
358 *precatoria*) and the smallest in *trnL-UAA* (519 bp – *E. edulis*, 519 bp – *E. oleracea*, 521
359 bp – *E. precatoria*).

360

361 Chloroplast genome structures and comparative analyses

362 The comparative analysis enabled us to identify a high level of synteny between
363 the three *Euterpe* chloroplast genomes, with large conserved blocks. The only structural
364 difference among *Euterpe* chloroplast genomes were three small inversions and single
365 nucleotide polymorphisms (SNPs) between *E. edulis* and *E. oleracea* chloroplast
366 sequences (S2A, B and C Figs). In a local alignment we detected these divergences in the
367 region between 66,081 to 69,440 bp; the first corresponding to an intergenic region

368 between *petA* and *psbJ* genes from *E. oleracea*; the second found in the *psbJ* gene and
369 the third in the intergenic region of *trnW-CCA* and *trnP-UGG* from *E. oleracea* (S2D Fig).
370 The parallel between the chloroplast genomes of *Euterpe edulis* and *Euterpe*
371 *oleracea* with the ones available in Genbank [34] showed differences in the total number
372 of base pairs. *E. edulis* and *E. oleracea* were 835 bp and 458 bp, respectively, larger than
373 the others. The alignment revealed that these increases were due to insertions that
374 occurred mainly in intergenic spacers (83.3% of insertions in *E. edulis* and 100% in *E.*
375 *oleracea*; S2 Table).
376 Although exhibiting high synteny and only a few structural rearrangements in
377 their chloroplast genome, the most notable divergence among the three Brazilian native
378 species from subfamily Arecoideae was in the length of the LSC, between 40,000 and
379 50,000 bp (Fig 2). A high level of synteny was also observed in a comparison of palm
380 chloroplast genomes from 12 genera, with five subfamilies (S3 Fig). *Astrocaryum*
381 *aculeatum*, *A. murumuru* and *Syagrus coronata* are native palms from subfamily
382 Arecoideae with smaller LSCs compared with *Euterpe* palms (Fig 2 and S3 Fig).
383 *Acrocomia aculeata,* also a Brazilian palm, presented a reduction in the size of the region
384 comprising the genes *ndhJ*, *ndhK*, *ndhC*, *trnF-GAA* and *trnL-UAA* (in lime green Fig 2).
385 Additionally, *Astrocaryum aculeatum* and *A. murumuru* showed a flip-flop
386 recombination in this same region (in lime green, Fig 2) [67].
387 **Fig 2. Synteny and divergence in the SSC size detected in Arecaceae chloroplast**
388 **genomes using the Mauve multiple-genome alignment program**. A sample of nine
389 different chloroplast genomes is shown. Color bars indicate syntenic blocks and the lines
390 indicate the correspondence between them. Blocks on the top row are in the same
391 orientation, while blocks on the bottom row are in inverse orientation.

392 In addition, we explored in more details the chloroplast IR expansions and
393 contractions of the Brazilian palm species from the subfamily Arecoideae (Fig 3).
394 Comparing the IR borders from Genbank available for *Euterpe edulis* chloroplast genome,
395 it was possible to observe a small difference in the *rpl22-rps19* location and an increase
396 in the length of *rps19-psbA* intergenic spacers. Among the chloroplast genomes from
397 *Euterpe* genus, we identified an expansion in the *ndhF* gene at the IRB region from
398 *Euterpe precatoria*. Among the Brazilian palm species, a contraction in the IR from
399 *Syagrus coronata* influenced the position of *rps19*. Among all species analyzed, the most
400 common variations were identified around the positions of *rpl22-rps19* (IRB-SSC), the
401 *ycf1* length (IRB and IRA), and the *rps19-ycf1* (IRA-LSC) (Fig 3).

402 **Fig 3. Comparison of the IRA and IRB borders among Brazilian palms from the**
403 **Arecoideae species**. The numbers indicate the lengths of IGSs, genes, and spacers
404 between IR-LSC and IR-SSC junctions. The *ycf1** and *rps19** genes have incomplete
405 CDSs.

406

407 Sequence repeats and polymorphisms between *Euterpe* 408 chloroplast genomes

409 We identified a total of 323 SSRs (*E. edulis*: 111, *E. oleracea*: 105, *E. precatoria*:
410 107 SSRs) in the chloroplast genomes from the three species. Most of them were located
411 in intergenic spacer regions (IGS: 72.07% *E. edulis*, 72.38% *E. oleracea*, 71.96 % *E.*
412 *precatoria*; Fig 4A), especially in the IGS between the tRNAs trnS-GCU/trnG-UCC (S3
413 Table). The SSRs were more abundant in the LSC region of the chloroplast genomes
414 (78.38% *E. edulis*, 78.10% *E. oleracea*, 77.57% *E. precatoria*; S3 Table) and less frequent
415 in the IR regions. They were mostly mononucleotides (62.16% *E. edulis*, 62.86% *E.*
416 *oleracea*, 60.75% *E. precatoria*; Fig 4B) and composed with A/T motifs (60.36% *E.*

417 *edulis*, 60.00% *E. oleracea*, 59.81% *E. precatoria*; S4A Fig). From the 323 SSRs, 47 had
418 the same sequence and gene location among *E. edulis*, *E. oleracea* and *E. precatoria*. We
419 also identified 62 SSRs between *E. edulis/E. oleracea* and *E. edulis/E. precatoria,* and 68
420 between *E. oleracea and E. precatoria* (S4 Table).

421 Considering the dispersed repeats, we observed 66 in *E. edulis*, 55 in *E. oleracea*
422 and 49 in *E. precatoria*, ranging from 30 to 77 bp (S4B Fig). These repeats were
423 distributed mostly in the IGS regions (65.15% *E. edulis*, 70.91% *E. oleracea*, 73.47% *E.*
424 *precatoria*; Fig 4C). Comparing the species, in *E. precatoria*, most of the repeats were
425 found in the intergenic region of *psaC*/*ndhE*. However, in *E. edulis* and *E. oleracea* the
426 repeats were concentrated in the *ycf2* gene (S5 Table). Regarding repeat types, most of
427 the repeats are palindrome or forward sequences, 40% and 30%, respectively (Fig 4D).
428 They were also mainly found in the LSC region (48.48% *E. edulis*, 52.73% *E. oleracea*,
429 44.90% *E. precatoria*; S5 Table).

430 **Fig 4. Distribution and classification of SSR and dispersed repeats in the chloroplast**
431 **genomes of *Euterpe edulis, E. oleracea* and *E. precatoria***. (A) Proportion of coding and
432 non-coding regions containing SSRs; (B) Proportion of different types of SSR present in
433 the chloroplast genomes; (C) Proportion of regions containing repeats; (D) Frequency
434 distribution of different types of repeats: F = Forward, P = Palindrome, R = Reverse and
435 C = Complement. CDS = Coding sequence, IGS = Intergenic spacer.

436 Based on the pairwise alignment between the species chloroplast genomes, we
437 identified 93 indels between *E. edulis*/*E. oleracea*, 103 between *E. edulis/E. precatoria*
438 and 58 between *E. oleracea*/*E. precatoria*. Most indels were distributed in the IGS (Fig
439 5A), especially between the *E. oleracea/E. precatoria* chloroplast genome, located
440 mainly in the LSC region (80.65% ~ 72.41%; S6 Table). The highest occurrence of indels
441 was identified in the *trnS-GCU*/*trnG-UCC* of the three alignments (S6 Table).

442 A total of 325 SNPs were detected between the three species (Fig 5B). In a
443 pairwise comparison, 283 SNPs were detected between *E. edulis/E. oleracea* and 254
444 SNPs between *E. edulis/E. precatoria*, which were most frequent in the CDS regions
445 (50.35% and 54.72%, respectively; Fig 5C). Between *E. oleracea/E. precatoria*, a lower
446 number of SNPs was found (113 SNPs), having the greatest number concentrated in the
447 IGS (50.44%; Fig 5C). Only one SNP was found shared among the three species (Fig
448 5B).

449 **Fig 5. Indels and single nucleotide polymorphisms (SNPs) detected in comparisons**
450 **between *Euterpe* chloroplast genome**s. (A) Proportion of indels in different coding and
451 non-coding regions; (B) Comparison of the number of SNPs found in the alignment (C)
452 Proportion of SNPs in different coding and non-coding regions of the chloroplast
453 genomes.

454 The large number of SNPs in the CDS region occurs mainly in the *atpE*, *ycf1* and
455 *psbJ* genes (S7 Table). These *atpE*, *psbJ* and *ycf1* genes presented respectively 19, 18 and
456 18 SNPs for *E. edulis/E. oleracea* and 20, 19 and 21 SNPs for *E. edulis/E. precatoria*.
457 For *E. oleracea/E. precatoria* there were only 1 (*atpE*) and 7 (*ycf1*) SNPs (S7 Table).

458 The LSC region had the highest number of SNPs in the chloroplast genomes
459 (69.01% *E. edulis/E. oleracea*, 66.54% *E. edulis/E. oleracea*, 64.91% *E. oleracea/E.*
460 *precatoria*; S7 Table). For *E. edulis/E. oleracea* the greatest number of SNPs
461 corresponded to C/T - T/C (74) substitutions, the same was saw for *E. edulis/E. precatoria*
462 T/C - C/T (66), and 29 substitutions of G/A - A/G and C/T - T/C was observed for *E.*
463 *oleracea/E. precatoria* (S5 Fig).

464 Considering pi > 0.02, the sliding window analysis revealed six hotspots of high
465 nucleotide polymorphism among the *E. edulis*, *E. oleracea* and *E. precatoria* chloroplast
466 genome sequence. The six hotspots were in IGS regions (Fig 6). Among them, the two

467 hotspots with the highest polymorphism were located between the *trnM*/*atpE* (Pi > 0.06;
468 Fig 6) and *psbJ*/*psbL* (Pi > 0.06; Fig 6). Comparing with the SNP identification, the *atpE*
469 gene had the highest number of SNPs among the species *E. edulis/E. oleracea* and *E.*
470 *edulis/E. precatoria*. The *psbJ* gene had 18 SNPs in *E. edulis/E. oleracea* and 19 SNPs
471 in *E. edulis/E. precatoria*.

472 **Fig 6. Sliding window analysis of the alignment from the chloroplast genome of**
473 ***Euterpe edulis*, *E. oleracea* and *E. precatoria* chloroplast genomes**. The regions with
474 high nucleotide variability (Pi > 0.02) are indicated. Pi is the nucleotide diversity of each
475 window, and the window length was 200 bp with 50 bp step sizes.

476 Both genes, *atpE* and *psbJ*, were analyzed for positive or purifying selection
477 across the palm phylogeny using the estimation of Ka/Ks ratio. In both cases, the Ka/Ks
478 < 1 revealed that no positively selected sites were found in the corresponding protein
479 sequences, indicating purifying selection. The analysis of the amino acid sequence
480 resulted in Ka/Ks rates varying from 0.17 to 0.87 for *atpE* (S8 Table) and ranging from
481 0.50 to 0.96 for *psbJ*. Also, from the 134 amino acids of the *atpE* protein, 68 had Ka/Ks
482 < 0.19 and from the 40 amino acids, 37 had Ka/Ks < 0.55, in the case of *psbJ* (S9 Table).
483

484 Phylogenomic of *Euterpe* based on chloroplast sequences

485 In the phylogenomic analysis, the marginal mean likelihoods of partition schemes
486 for the two runs were i) -186887.39, ii) -178507.00, and iii) -178497.59. Therefore, the
487 best partition schemes were by far the ones considering codon positions, although 'iii'
488 was strongly better than 'ii' with the 2*log difference = 18,82 (>10 is strong in Kass and
489 Raftery 1995). We also compared the majority-rule consensus trees of the three partition
490 scheme analyses, which did not differ in topology, but in general the scheme 'iii'

presented better support in some nodes. For that reason, the figure presented and used in our discussion will be based on partition scheme 'iii'.

The phylogeny using coding sequences from 54 whole chloroplast genomes from palm species rooted in the outgroup species *Dasypogon bromeliifolius* revealed the same relationships in subfamilies previously reported [34], with highly supported nodes (Fig 7). According to our results, the species from *Euterpe* were placed in subfamily Arecoideae, tribe Euterpeae, and were sister to tribe Areceae. The new samples of *E. oleracea* and *E. edulis* were grouped with the respective previously published chloroplast genomes and the *E. precatoria* was placed in a branch sister to *E. oleracea*.

Fig 7. Majority-rule consensus tree of 30,000 trees obtained from a Bayesian inference analysis of chloroplast protein coding genes of 55 taxa. Posterior probabilities (PP) for each are indicated above branches. Co = Coryphoideae, Ar = Arecoideae, Ny = Nypoideae, Ce = Ceroxyloideae, Ca = Calamoideae.

Species divergence with nuclear and mitochondrial genome

From the 11,738,792 reads that were generated for *E. edulis*, 553,857 (4.71%) aligned to the assembled chloroplast genome. The remaining (ca. 95%) reads correspond to nuclear and mitochondrial regions. Similar proportions were also found for *E. oleracea* and *E. precatoria* for which 12,987,827 and 12,612,001 reads were generated, respectively. Moreover, 249,387 (1.92%) and 132,227 (1.04%) were aligned to the assembled chloroplast sequence for both species, respectively. Using Stacks, we found 1,077 SNPs in the nuclear and mitochondrial genome sequences, with 10X mean sequencing depth per locus (S6A Fig). Among *Euterpe* species, more transitions than transversions were detected, and the most frequent mutations were A-G and C-T transitions (S6C Fig).

516 The highest number of alleles (A = 1,309) and highest observed heterozygosity
517 (H_O = 0.162) was identified in *E. oleracea*, while the highest number of private alleles
518 was observed in *E. precatoria* (Ap = 421). All the three species had a higher value of H_O
519 compared with the expected heterozygosity (H_E, Table 3).

520 **Table 3. Parameters of genetic diversity using 1,077 SNPs found the nuclear and**
521 **mitochondrial genome sequences from three *Euterpe* species.**

	E. edulis	*E. oleracea*	*E. precatoria*
A	1,251	1,309	1,263
Ap	292	260	421
H_O	0.162 (±0.011)	0.215 (± 0.013)	0.173 (±0.012)
H_E	0.081 (±0.006)	0.108 (±0.006)	0.086 (±0.006)

522 A = number of alleles, Ap = number of private alleles, H_O = observed heterozygosity, H_E
523 = expected heterozygosity.

524 The pairwise F_{ST} showed high and significant divergence between species, with
525 values higher than 0.770. Although all the F_{ST} values were high, the species *E. edulis* and
526 *E. precatoria* presented the greatest genetic divergence (0.860, S7 Fig) among the three
527 pairs of species, and *E. edulis* and *E. oleracea* had the lowest divergence (0.779, S7 Fig).

528

529 # Discussion

530 ## High level of conservation between Brazilian palms
531 ## chloroplast genomes

532 Among the chloroplast genomes from the *Euterpe* genus, we found a conserved
533 organization and gene content, with small differences between species, as in the length of
534 the duplicated *ycf2* gene. The *ycf2* is the largest plastid gene reported in angiosperms [68],
535 and is part of the 2-MD heteromeric AAA-ATPase complex directly interacting with
536 various translocating preproteins [69]. Particularly, *ycf2* has an insertion in the *E. edulis*
537 chloroplast genome that is absent in *E. oleracea* and *E. precatoria*, a feature previously

538 observed in *Euterpe* chloroplast genomes [34]. Across palm species, the chloroplast
539 genome structure is highly conserved as reported in other studies [28,34,70]. The multiple
540 alignments with 15 different palms from five subfamilies did not exhibit significant
541 rearrangements in the chloroplast structure. The only rearrangement observed was an
542 inversion of 4.6 kb in the *Astrocaryum* chloroplast genome. It was possibly to have a
543 lineage-specific structural variation, creating a isoform from a flip-flop recombination
544 between inverted repeats [67]. However, differences in the total size of chloroplast
545 genomes were observed in the same species, such as in *E. edulis* and *E. oleracea*, which
546 may be due to the assembly method used. Different from the approach of all possible
547 paths' calculation to complete target organelle genome, as applied by GetOrganelle, other
548 widely used pipelines utilize reference genomes to select and filter scaffolds/contigs for
549 further concatenation or post-assembly gap filling and closing [38].

550 Besides these conserved general features, similar occurrences of SSRs were
551 identified among the three *Euterpe* species. Currently, SSR markers are very useful in
552 studies of population structure, genetic mapping, and evolutionary processes [71]. The
553 SSR identified here can be a valuable resource for studies aiming at the conservation of
554 the species or their sustainable exploitation. These SSR are mostly composed of A/T
555 motifs, and are more frequently located in the IGS of the single-copy regions from the
556 chloroplast genomes [1,72,73], which is expected since these regions evolve faster than
557 the CDS [74]. Also, the limited occurrence of SSRs in the IRs is related to a lower
558 mutation rate observed in these regions, caused by an efficient mechanism of gene copy-
559 correction, as observed in other palm chloroplast genomes [34,75].

560 The dispersed repeats (forward, palindrome, reverse and complement) showed the
561 same pattern of the SSRs and were frequently distributed in the IGS of the single copy
562 regions, differing from a predominant occurrence of reverse repeats (18,37%) in *E.*

563 *precatoria*. Actually, the number of repeats found in the *Euterpe* chloroplast genomes
564 were not considered high [76]. According to Milligan et al. [77], repetitive sequences are
565 substrates for recombination and chloroplast genome rearrangements. Since non-
566 structural rearrangements were detected in the *Euterpe* chloroplast genomes, low
567 frequency of repetitive elements was also expected.

568 Regarding the structure patterns of the IRs, we could identify most of the variation
569 related to expansion/contractions when comparing only Brazilian palm species from
570 subfamily Arecoideae. Events of expansion and contraction at the IR boundaries was
571 previously observed in the *ycf1* gene at IR-SSC/LSC junction and *rps19* intergenic spacer
572 with *rpl22*/*psbA* in IR-SSC/LSC junctions of other species [28,75].

573

574 Indels and single nucleotide polymorphisms clearly 575 demonstrate differentiation among the three *Euterpe* species, 576 which was reflected in the phylogeny

577 We identified six hotspots of high nucleotide polymorphism with the largest
578 polymorphism (Pi > 0.06) in intergenic regions composed mainly by genes with a great
579 number of SNPs. The genes *atpE* and *psbJ* had the greatest number of SNPs, especially
580 when considering the nucleotide substitutions between *E. edulis* and the *Euterpe*
581 Amazonian species, *E. oleracea* and *E. precatoria.* This divergence between *E. edulis* in
582 relation to the two Amazonian species was also observed in the number of indels and
583 plastid SNPs and was almost 50% higher than the number of polymorphisms detected in
584 a pairwise analysis with *E. oleracea* and *E. precatoria.*

585 Indels display no ambiguity in complex mutation patterns and SNPs are the most
586 abundant type of markers [78], giving us a clear pattern of divergence between *E. edulis*
587 in relation to *E. oleracea* and *E. precatoria*. In our study, these mutations were frequently

588 found in the CDS than in the intergenic spacers. This high number of indels and SNPs
589 between *E. edulis/E. oleracea* had already been identified in a previous study [34], but
590 they were mainly found in non-coding regions, which presumably may be related to a
591 more recent event of divergence.

592 Regarding the phylogenetic relationships among the palm species, we identified
593 a clear distinction among subfamilies using the informative chloroplast coding sequences,
594 corroborating previous studies [28,34,67,79]. Our phylogeny suggests that *E. oleracea*
595 and *E. precatoria* are sister species with an early split of *E. edulis* from these two,
596 differently from what was observed previously, having four nuclear low copy markers
597 and one plastid region [33]. Our result reveals the possibility of incongruences between
598 nuclear and plastid phylogenies in *Euterpe*, since the previous study using nuclear
599 markers reported paraphyly in *E. precatoria* and showed one of the *E. precatoria* varieties
600 as sister to *E. edulis* [33].

601 The closest relationship between *E. oleracea* and *E. precatoria* could reflect their
602 environmental occupation since both species occur in the Brazilian Amazon. *E. edulis*,
603 nonetheless, evolved via vicariant split from its common ancestor (around 1.4 Mya), and
604 the speciation was intensified by the savanna barrier composed from the Cerrado and
605 Caatinga biomes [33,34,80]. Furthermore, the phenotypic plasticity and local adaptation
606 may have favored the successful expansion of *E. edulis* throughout the Atlantic Forest
607 [11,80].

608 Regarding the coding sequences, the *rpl22*, *rpl32*, *rpl14*, *rps14, matK, accD, rbcl,*
609 *ccsA, ycf2 and ycf1* genes have been reported to harbor positive selection in an analysis
610 using 41 species of the Arecaceae family, including the *E. oleracea* and *E. edulis*
611 chloroplast genomes [34]. This was highlighted as indicative of convergent evolution,
612 associated with environmental adaptation in *E. edulis*. However, using 52 palm species,

613 we could only identify a tendency of purifying selection for *psbJ* and *atpE*, the genes with
614 higher SNPs and polymorphisms hotspots. This may reflect the typically conservative
615 nature of the chloroplast genome sequences across most angiosperms [81,82]. These
616 genes are essential for plant development, since *atpE* encodes a subunit of the ATP
617 synthase complex that participates in the photosynthetic phosphorylation [83] and *psbJ* is
618 associated with photosystem II (PSII) [84].
619

620 A comparison using nuclear and mitochondrial genome 621 sequences detected high divergence between *Euterpe* species

622 The reads from nuclear and mitochondrial DNA obtained from the sequencing
623 were used in this first-time comparison of the genetic diversity and genomic divergence
624 among the three species from the *Euterpe* genus. The advances in next-generation
625 sequencing (NGS) is allowing the discovery of a large number of single nucleotide
626 polymorphism (SNPs), even for non-model species [85–88]. This generates a valuable
627 source of bi-allelic genetic markers that are appropriate to evaluate the genetic diversity
628 even when sample sizes are small [89]. Additionally, SNPs occurs at higher density across
629 the genome, with lower genotyping error rates than other markers [90], enabling robust
630 estimates of genetic diversity and structuring [9].
631 The three *Euterpe* species had similar number of alleles and low observed
632 heterozygosity. Yet, they are higher than the expected heterozygosity, with *E. oleracea*
633 presenting slightly higher results. However, despite this enlightening information
634 obtained with the remaining nuclear and mitochondrial sequences, we highlight that these
635 findings could be biased because they were based on only one individual per species.
636 Considering assessments of genome-wide diversity, previous studies identifying SNPs in
637 palms were only performed for *E. edulis* [9,80]. Furthermore, the genetic diversity

638 estimates from our *E. edulis* sample, that was also from the Atlantic rainforest of São
639 Paulo state, were in agreement with previously published studies. This observed genetic
640 variability could be influenced by the type of vegetation and the altitude where the
641 samples were collected [9,80].

642 In summary, our results open the way for future genome-skimming studies in
643 palms. Also, the addition of larger sample sizes from different sampling locations will
644 provide a better understanding on the evolution and diversification of this important group
645 of plants. However, the high differentiation found between the three *Euterpe* species may
646 indicate that they are evolving independently and that this differentiation is caused by a
647 private pool of alleles from each species, which is in accordance with the phylogenetic
648 analyses.

649

650 # Acknowledgements

651 The authors would like to thank Dr. Charles R. Clement for sending samples of *Euterpe*
652 *precatoria* from Instituto Nacional de Pesquisas da Amazônia (INPA), Manaus, Brazil
653 for the genomic extraction and for reading the manuscript.

654

655 # Supporting information

656 **S1 Fig. Flowchart of the SNPs identification methodology using nuclear and**
657 **mitochondrial genomes sequences.**
658 (TIFF)

659

660 **S2 Fig. Dot-plot analyses comparing three *Euterpe* chloroplast genomes**. (A) *Euterpe*
661 *edulis* (X – axis) x *Euterpe Oleracea* (Y-axis), (B) *E. edulis* (X-axis) x *Euterpe precatoria*

662 (Y-axis), (C) *E. oleracea* (X – axis) x *E. precatoria* (Y-axis). The positive slope, in purple,
663 represents the pair of sequences aligned and in the same orientation. The negative slope,
664 in blue, represents the pair of sequence aligned, but in opposite orientation. The blue
665 arrow in A highlights the region with inversions and SNPs between *E. edulis* and *E.*
666 *oleracea*. (D) Local alignment with the chloroplast genomes of *E. edulis*, *E. oleracea* and
667 *E. precatoria* in the region where inversions and SNPs were detected in A.
668 (TIFF)
669
670 **S3 Fig. Synteny and divergence in the size of the SSC detected in Arecaceae**
671 **chloroplast genome sequences**. A sample of 20 different chloroplast genomes is shown.
672 Color bars indicate syntenic blocks and the lines indicate the correspondence between
673 them. Blocks on the top row are in the same orientation, while blocks on the bottom row
674 are in inverse orientation.
675 (TIFF)
676
677 **Fig S4. SSR and dispersed repeats in the chloroplast genomes of *Euterpe edulis, E.***
678 ***oleracea* and *E. precatoria***. (A) The number of SSR motifs found in the three
679 chloroplast genomes, considering sequence complementarities; (B) Number of
680 dispersed repeats present in different size classes.
681 (TIFF)
682
683 **S5 Fig. Number and type of SNPs found in the alignment of three *Euterpe***
684 **chloroplast genomes**
685 (TIFF)
686

S6 Fig. Description of 1,077 SNPs found among the nuclear and mitochondrial genome sequences from three *Euterpe* species. A) Mean depth per loci; B) Mean depth per species sequenced; C) Mutation count in the present in the 1,077 SNPS.
(TIFF)

S7 Fig. Pairwise F_{ST} based on 1,077 SNPs of nuclear and mitochondrial genomes sequences from three *Euterpe* specie
(TIFF)

S1 Table. List of complete chloroplast genomes used in comparative analysis. Synteny and rearrangement, border comparison and phylogenomics.
(XSLX)

S2 Table. Insertions identified in the alignment of *Euterpe edulis* and *E. oleracea* chloroplast genomes with sequences from GenBank (de Santana Lopes et al. 2021). CDS = Coding sequence; IGS = intergenic spacer; LSC = Large single copy; SSC = Small single copy.
(XSLX)

S3 Table. Simple sequence repeats (SSR) identified in the complete chloroplast genomes of three *Euterpe* species. CDS = Coding sequence; IGS = intergenic spacer; LSC = Large single copy; SSC = Small single copy; IRs = Inverted Repeats.
(XSLX)

711 **S4 Table. Simple sequence repeats (SSR) among the chloroplast genomes of**

712 ***Euterpe edulis*, *E. oleracea* and *E. precatoria* according to their location.**

713 (XSLX)

714

715 **S5 Table. Repeats identified in the chloroplast genomes of *Euterpe edulis, E.***

716 ***oleracea* and *E. precatoria*.** F = Forward; P = Palindrome; R = Reverse; C =

717 Complement; CDS= Coding sequence; IGS = intergenic spacer; LSC = Large single

718 copy; SSC = Small single copy; IRs = Inverted Repeats.

719 (XSLX)

720

721 **S6 Table. Indels identified in the pairwise comparison of three *Euterpe* species**

722 **complete chloroplast genomes**. CDS = Coding sequence; IGS = intergenic spacer;

723 LSC = Large single copy; SSC = Small single copy; IRs = Inverted Repeats.

724 (XSLX)

725

726 **S7 Table. Single nucleotide polymorphisms (SNPs) identified in the pairwise**

727 **comparison of three *Euterpe* species complete chloroplast genomes.** CDS = Coding

728 sequence; IGS = intergenic spacer; LSC = Large single copy; SSC = Small single copy;

729 IRs = Inverted Repeats.

730 (XSLX)

731

732 **S8 Table. Bayesian selection results based on synonymous (Ka)/non-synonymous**

733 **(Ks) ratio of amino acid substitutions from the *atpE* gene in 54 palm species**.

734 (XSLX)

735

S9 Table. Bayesian selection results based on synonymous (Ka)/non-synonymous (Ks) ratio of amino acid substitutions from the *psbJ* gene in 54 palm species.

(XSLX)

Author Contributions

Conceptualization: Maria Imaculada Zucchi

Data curation: Ana Flávia Francisconi, Jonathan Andre Morales Marroquín, Cássio van den Berg, Luciano Delmondes de Alencar, Alessandro Alves-Pereira

Formal Analysis: Ana Flávia Francisconi, Luiz Augusto Cauz dos Santos, Cássio van den Berg, Alessandro Alves-Pereira, Luciano Delmondes de Alencar

Funding Acquisition: Maria Imaculada Zucchi

Investigation: Ana Flávia Francisconi, Luiz Augusto Cauz dos Santos

Methodology: Luiz Augusto Cauz dos Santos, Jonathan Andre Morales Marroquín, Doriane Picanço-Rodrigues, Luciano Delmondes de Alencar, Cesar Augusto Zanello, Marcones Ferreira Costa

Resources: Maria Imaculada Zucchi, Doriane Picanço-Rodrigues

Software: Cássio van den Berg

Supervision: Maria Imaculada Zucchi

Writing – Original Draft Preparation: Ana Flávia Francisconi, Luiz Augusto Cauz dos Santos

Writing – Review & Editing: Ana Flávia Francisconi, Luiz Augusto Cauz dos Santos, Jonathan Andre Morales Marroquín, Cássio van den Berg, Alessandro Alves-Pereira, Doriane Picanço-Rodrigues, Luciano Delmondes de Alencar, Cesar Augusto Zanello, Marcones Ferreira Costa, Maria Teresa Gomes Lopes, Elizabeth Ann Veasey, Maria Imaculada Zucchi

References

1. Uthaipaisanwong P, Chanprasert J, Shearman JR, Sangsrakru D, Yoocha T, Jomchai N, et al. Characterization of the chloroplast genome sequence of oil palm (*Elaeis guineensi*s Jacq.). Gene. 2012;500: 172–180. doi:10.1016/j.gene.2012.03.061
2. Baker WJ, Couvreur TLP. Global biogeography and diversification of palms sheds light on the evolution of tropical lineages. I. Historical biogeography. J Biogeogr. 2013;40: 274–285. doi:10.1111/j.1365-2699.2012.02795.x
3. Brokamp G, Valderrama N, Mittelbach M, Grandez CA, Barfod AS, Weigend M. Trade in palm products in North-Western South America. Bot Rev. 2011;77: 571–606. doi:10.1007/s
4. Eiserhardt WL, Pintaud J, Asmussen-lange C, Hahn J, Bernal R, Balslev H, et al. Phytogeny and divergence times of Bactridinae (Arecaceae, Palmae) based on plastid and nuclear DNA sequences. 2011;60: 485–498.
5. Baker WJ, Couvreur TLP. Global biogeography and diversification of palms sheds light on the evolution of tropical lineages. II. Diversification history and origin of regional assemblages. J Biogeogr. 2013;40: 286–298. doi:10.1111/j.1365-2699.2012.02794.x
6. Henderson A, Galeano G. Euterpe, Prestoea, and Neonicholsonia (Palmae). Flora Neotrop. 1996;72: 1–89.
7. Mart E, Bussmann RW, Zambrana NYP. Acta Societatis Botanicorum Poloniae Facing global markets – usage changes in Western Amazonian plants: the example of *Euterpe precatoria* Mart. and *E. oleracea* Mart. 2012;81: 257–261. doi:10.5586/asbp.2012.032
8. Galetti M, Guevara R, Côrtes MC, Fadini R, Matter S Von, Leite AB, et al. Functional extinction of birds drives rapid evolutionary changes in seed size. Science (80-). 2013;340: 1086–1091. doi:DOI: 10.1126/science.1233774
9. Alves-Pereira A, Novello M, Dequigiovanni G, Pinheiro JB, Brancalion PHS, Veasey EA, et al. Genomic diversity of three Brazilian native food crops based on double-digest restriction site-associated DNA sequencing. Trop Plant Biol. 2019;12: 268–281. doi:10.1007/s12042-019-09229-z
10. Novello M, Viana JPG, Alves-Pereira A, de Aguiar Silvestre E, Nunes HF, Pinheiro JB, et al. Genetic conservation of a threatened Neotropical palm through community-management of fruits in agroforests and second-growth forests. For Ecol Manage. 2017;407: 200–209. doi:10.1016/j.foreco.2017.06.059
11. Coelho GM, Santos AS, de Menezes IPP, Tarazi R, Souza FMO, Silva M das GCPC, et al. Genetic structure among morphotypes of the endangered Brazilian palm *Euterpe edulis* Mart (Arecaceae). Ecol Evol. 2020;10: 6039–6048. doi:10.1002/ece3.6348
12. Martinelli G, Moraes MA. Livro vermelho da flora do Brasil. 1st ed. Rio de Janeiro: CNCFlora, Centro Nacional de Conservação da Flora Rio de Janeiro; 2013.
13. Ball AA, Brancalion PHS. Governance challenges for commercial exploitation of a non-timber forest product by marginalized rural communities. Environ Conserv. 2016;43: 208–220. doi:10.1017/S0376892916000072
14. IBGE. Censo agropecuário 2017: resultados definitivos. Censo agropecuário. 2019;8: 1–105. Available from: https://biblioteca.ibge.gov.br/visualizacao/periodicos/3096/agro_2017_resultados_definitivos.pdf

15. Ramos SLF, Dequigiovanni G, Lopes MTG, Aguiar AV de, Lopes R, Veasey EA, et al. Genetic structure in populations of *Euterpe precatoria* Mart. in the Brazilian Amazon. Front Ecol Evol. 2021;8: 1–11. doi:10.3389/fevo.2020.603448
16. Shanley P, Medina G. Frutíferas e plantas úteis na vida amazônica. 1st ed. Belém-PA: Cifor, Imazon; 2005.
17. Vedel-sørensen M, Tovaranonte J, Bøcher PK, Balslev H, Barfod AS. Spatial distribution and environmental preferences of 10 economically important forest palms in western South America. For Ecol Manage. 2013;307: 284–292. doi:10.1016/j.foreco.2013.07.005
18. Freitas MAB, Vieira ICG, Albernaz ALKM, Magalhães JLL, Lees AC. Floristic impoverishment of Amazonian floodplain forests managed for açaí fruit production. For Ecol Manage. 2015;351: 20–27. doi:10.1016/j.foreco.2015.05.008
19. Campbell AJ, Carvalheiro LG, Maués MM, Jaffé R, Giannini TC, Freitas MAB, et al. Anthropogenic disturbance of tropical forests threatens pollination services to açaí palm in the Amazon river delta. J Appl Ecol. 2018; 1725–1736. doi:10.1111/1365-2664.13086
20. Kahn F. Palms as key swamp forest resources in Amazonia. For Ecol Manage. 1991;38: 133–142. doi:https://doi.org/10.1016/0378-1127(91)90139-M
21. Ramos SLF, Dequigiovanni G, Magno A, Pimentel P, Teresa M, Lopes G, et al. Paternity analysis, pollen flow, and spatial genetic structure of a natural population of *Euterpe precatoria* in the Brazilian Amazon. Ecol Evol. 2018;8: 11143–11157. doi:10.1002/ece3.4582
22. Du FK, Lang T, Lu S, Wang Y, Li J, Yin K. An improved method for chloroplast genome sequencing in non-model forest tree species. Tree Genet Genomes. 2015;11. doi:10.1007/s11295-015-0942-2
23. Dodsworth S. Genome skimming for next-generation biodiversity analysis. Trends Plant Sci. 2015;20: 525–527. doi:10.1016/j.tplants.2015.06.012
24. Sobreiro MB, Vieira LD, Nunes R, Novaes E, Coissac E, Silva-Junior OB, et al. Chloroplast genome assembly of *Handroanthus impetiginosus*: Comparative analysis and molecular evolution in Bignoniaceae. Planta. 2020;252: 1–16. doi:10.1007/s00425-020-03498-9
25. Sugiura M. The chloroplast genome. Plant Mol Biol. 1992;19: 149–168. doi:10.1007/978-94-011-2656-4_10
26. Barrett CF, Bacon CD, Antonelli A, Cano Á, Hofmann T. An introduction to plant phylogenomics with a focus on palms. Bot J Linn Soc. 2016;182: 234–255. doi:10.1111/boj.12399
27. Dabral A, Shamoon A, Meena RK, Kant R, Pandey S, Ginwal HS, et al. Genome skimming-based simple sequence repeat (SSR) marker discovery and characterization in *Grevillea robusta*. Physiol Mol Biol Plants. 2021;27: 1623–1638. doi:10.1007/s12298-021-01035-w
28. de Santana Lopes A, Gomes Pacheco T, Nimz T, do Nascimento Vieira L, Guerra MP, Nodari RO, 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. 2018;247: 1011–1030. doi:10.1007/s00425-018-2841-x
29. Aljohi HA, Liu W, Lin Q, Zhao Y, Zeng J, Alamer A, et al. Complete sequence and analysis of coconut palm (*Cocos nucifera*) mitochondrial genome. PLoS One. 2016;11: 1–18. doi:10.1371/journal.pone.0163990

30. Yang Y, Zhou T, Duan D, Yang J, Feng L, Zhao G. Comparative analysis of the complete chloroplast genomes of five quercus species. Front Plant Sci. 2016;7: 1–13. doi:10.3389/fpls.2016.00959
31. Cauz-santos LA, Munhoz CF, Rodde N, Cauet S, Santos AA, Penha HA, et al. The chloroplast genome of *Passiflora edulis* (Passifloraceae) assembled from long sequence reads: structural organization and phylogenomic studies in malpighiales. 2017;8: 1–17. doi:10.3389/fpls.2017.00334
32. Machado L de O, Vieira L do N, Stefenon VM, Oliveira Pedrosa F de, Souza EM de, Guerra MP, et al. Phylogenomic relationship of feijoa (*Acca sellowiana* (O.Berg) Burret) with other Myrtaceae based on complete chloroplast genome sequences. Genetica. 2017;145: 163–174. doi:10.1007/s10709-017-9954-1
33. Pichardo-Marcano FJ, Nieto-Blázquez ME, MacDonald AN, Galeano G, Roncal J. Phylogeny, historical biogeography and diversification rates in an economically important group of Neotropical palms: Tribe Euterpeae. Mol Phylogenet Evol. 2019;133: 67–81. doi:10.1016/j.ympev.2018.12.030
34. de Santana Lopes A, Gomes Pacheco T, Nascimento da Silva O, do Nascimento Vieira L, Guerra MP, Pacca Luna Mattar E, et al. Plastid genome evolution in Amazonian açaí palm (*Euterpe oleracea* Mart.) and Atlantic forest açaí palm (*Euterpe edulis* Mart.). Plant Mol Biol. 2021;105: 559–574. doi:10.1007/s11103-020-01109-5
35. Liu L, Zhang Y, Li P. Development of genomic resources for the genus Celtis (Cannabaceae) based on genome skimming data. Plant Divers. 2021;43: 43–53. doi:10.1016/j.pld.2020.09.005
36. Takamatsu T, Baslam M, Inomata T, Oikawa K. Optimized method of extracting rice chloroplast DNA for high-quality plastome resequencing and de novo assembly. 2018;9: 1–13. doi:10.3389/fpls.2018.00266
37. Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: De novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. 2017;45. doi:10.1093/nar/gkw955
38. Jin JJ, Yu W Bin, Yang JB, Song Y, DePamphilis CW, Yi TS, et al. GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome. 2020;21: 1–31. doi:10.1101/256479
39. Langdon WB. Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. BioData Min. 2015;8: 1–7. doi:10.1186/s13040-014-0034-0
40. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19: 455–477. doi:10.1089/cmb.2012.0021
41. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, et al. GeSeq - Versatile and accurate annotation of organelle genomes. Nucleic Acids Res. 2017;45: W6–W11. doi:10.1093/nar/gkx391
42. Abeel T, Van Parys T, Saeys Y, Galagan J, Van De Peer Y. GenomeView: A next-generation genome browser. Nucleic Acids Res. 2012;40: 1–10. doi:10.1093/nar/gkr995
43. Greiner S, Lehwark P, Bock R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: Expanded toolkit for the graphical visualization of organellar genomes. Nucleic Acids Res. 2019;47: W59–W64. doi:10.1093/nar/gkz238
44. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. PLoS Comput Biol. 2018;14: 1–14. doi:10.1371/journal.pcbi.1005944

45. Katoh K, Standley DM. MAFFT: Iterative refinement and additional methods. Methods Mol Biol. 2014;1079: 131–146. doi:10.1007/978-1-62703-646-7_8
46. Da Silva RS, Clementi CR, Balsanelli E, de Baura VA, de Souza EM, de Freitas Fraga HP, et al. The plastome sequence of *Bactris gasipaes* and evolutionary analysis in tribe Cocoseae (Arecaceae). PLoS One. 2021;16: 1–15. doi:10.1371/journal.pone.0256373
47. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004;14: 1394–1403. doi:10.1101/gr.2289704
48. Beier S, Thiel T, Münch T, Scholz U, Mascher M. MISA-web: A web server for microsatellite prediction. Bioinformatics. 2017;33: 2583–2585. doi:10.1093/bioinformatics/btx198
49. Kurtz S, Schleiermacher C. REPuter: fast computation of maximal repeats in complete genomes. Bioinforma Appl Note. 1999;15: 426–427. doi:10.1093/bioinformatics/15.5.426
50. Librado P, Rozas J. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009;25: 1451–1452. doi:10.1093/bioinformatics/btp187
51. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32: 1792–1797. doi:10.1093/nar/gkh340
52. Stern A, Doron-Faigenboim A, Erez E, Martz E, Bacharach E, Pupko T. Selecton 2007: Advanced models for detecting positive and purifying selection using a Bayesian inference approach. Nucleic Acids Res. 2007;35: 506–511. doi:10.1093/nar/gkm382
53. Nylander JAA. MrModeltest v2. Program distributed by the author. Evolutionary Biology Center, Uppsala University; 2004.
54. Xie W, Lewis PO, Fan Y, Kuo L, Chen MH. Improving marginal likelihood estimation for bayesian phylogenetic model selection. Syst Biol. 2011;60: 150–160. doi:10.1093/sysbio/syq085
55. Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, et al. Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 2012;61: 539–542. doi:10.1093/sysbio/sys029
56. Miller MA, Pfeiffer W, Schwartz T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. 2010 Gatew Comput Environ Work GCE 2010. 2010. doi:10.1109/GCE.2010.5676129
57. Kass RE, Raftery AE. Bayes Factors. J Am Stat Assoc. 1995;90: 773–795.
58. Rambaut A. FigTree v1. 3.1. 2012. Available from: http://tree.bio.ed.ac.uk/software/figtree
59. Inkscape. Inkscape Project. 2021. Available from: https://inkscape.org/pt-br/
60. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079. doi:10.1093/bioinformatics/btp352
61. Broad Institute. Picard Tools. Broad Institute GitHub Repository. 2018. Available from: http://broadinstitute.github.io/picard/
62. Catchen J, Hohenlohe PA, Bassham S, Amores A. Stacks: an analysis tool set for population genomics. 2013; 3124–3140. doi:10.1111/mec.12354
63. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27: 2156–2158. doi:10.1093/bioinformatics/btr330
64. Jombart T, Ahmed I. adegenet 1.3-1: New tools for the analysis of genome-wide

SNP data. Bioinformatics. 2011;27: 3070–3071. doi:10.1093/bioinformatics/btr521

65. Rousset F, Lopez L, Belkhir K. R package: genepop. 2020. p. 16. Available from: http://kimura.univ-montp2.fr/~rousset/Genepop.htm

66. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021. Available from: https://www.r-project.org/

67. de Santana Lopes A, Gomes Pacheco T, Nascimento da Silva O, Magalhães Cruz L, Balsanelli E, Maltempi de Souza E, et al. The plastomes of *Astrocaryum aculeatum* G. Mey. and *A. murumuru* Mart. show a flip-flop recombination between two short inverted repeats. Planta. 2019;250: 1229–1246. doi:10.1007/s00425-019-03217-z

68. Drescher A, Stephanie R, Calsa T, Carrer H, Bock R. The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. Plant J. 2000;22: 97–104. doi:10.1046/j.1365-313X.2000.00722.x

69. Kikuchi S, Asakura Y, Imai M, Nakahira Y, Kotani Y, Hashiguchi Y, et al. A Ycf2-FtsHi heteromeric AAA-ATPase complex is required for chloroplast protein import. Plant Cell. 2018;30: 2677–2703. doi:10.1105/tpc.18.00357

70. Barrett CF, Davis JI, Leebens-Mack J, Conran JG, Stevenson DW. Plastid genomes and deep relationships among the commelinid monocot angiosperms. Cladistics. 2013;29: 65–87. doi:10.1111/j.1096-0031.2012.00418.x

71. Vieira MLC, Santini L, Diniz AL, Munhoz C de F. Microsatellite markers: What they mean and why they are so useful. Genet Mol Biol. 2016;39: 312–328. doi:10.1590/1678-4685-GMB-2016-0027

72. Eguiluz M, Rodrigues NF, Guzman F, Yuyama P, Margis R. The chloroplast genome sequence from *Eugenia uniflora*, a Myrtaceae from Neotropics. Plant Syst Evol. 2017;303: 1199–1212. doi:10.1007/s00606-017-1431-x

73. Ping J, Feng P, Li J, Zhang R, Su Y, Wang T. Molecular evolution and SSRs analysis based on the chloroplast genome of *Callitropsis funebri*s. Ecol Evol. 2021;11: 4786–4802. doi:10.1002/ece3.7381

74. Rogalski M, Vieira LDN, Fraga HP, Guerra MP. Plastid genomics in horticultural species: Importance and applications for plant population genetics, evolution, and biotechnology. Front Plant Sci. 2015;6: 1–17. doi:10.3389/fpls.2015.00586

75. Zhu A, Guo W, Gupta S, Fan W, Mower JP. Evolutionary dynamics of the plastid inverted repeat: The effects of expansion, contraction, and loss on substitution rates. New Phytol. 2016;209: 1747–1756. doi:10.1111/nph.13743

76. Zeng S, Zhou T, Han K, Yang Y, Zhao J, Liu ZL. The complete chloroplast genome sequences of six rehmannia species. Genes (Basel). 2017;8. doi:10.3390/genes8030103

77. Milligan BG, Hampton JN, Palmer JD. Dispersed repeats and structural reorganization in subclover chloroplast DNA. Mol Biol Evol. 1989;6: 355–368. doi:10.1093/oxfordjournals.molbev.a040558

78. Song Y, Dong W, Liu B, Xu C, Yao X, Gao J, et al. Comparative analysis of complete chloroplast genome sequences of two tropical trees *Machilus yunnanensis* and *Machilus balansae* in the family Lauraceae. Front Plant Sci. 2015;6: 1–8. doi:10.3389/fpls.2015.00662

79. Barrett CF, Baker WJ, Comer JR, Conran JG, Lahmeyer SC, Leebens-Mack JH, et al. Plastid genomes reveal support for deep phylogenetic relationships and extensive rate variation among palms and other commelinid monocots. New

Phytol. 2016;209: 855–870. doi:10.1111/nph.13617
80. Brancalion PHS, Oliveira GCX, Zucchi MI, Novello M, van Melis J, Zocchi SS, et al. Phenotypic plasticity and local adaptation favor range expansion of a Neotropical palm. Ecol Evol. 2018;8: 7462–7475. doi:10.1002/ece3.4248
81. Du X, Zeng T, Feng Q, Hu L, Luo X, Weng Q, et al. The complete chloroplast genome sequence of yellow mustard (*Sinapis alba* L.) and its phylogenetic relationship to other Brassicaceae species. Gene. 2020;731. doi:10.1016/j.gene.2020.144340
82. Liu Q, Li X, Li M, Xu W, Schwarzacher T, Heslop-Harrison JS. Comparative chloroplast genome analyses of Avena: Insights into evolutionary dynamics and phylogeny. BMC Plant Biol. 2020;20: 1–20. doi:10.1186/s12870-020-02621-y
83. Huang Y, Yang Z, Huang S, An W, Li J, Zheng X. Comprehensive analysis of *Rhodomyrtus tomentosa* chloroplast genome. Plants. 2019;8. doi:10.3390/plants8040089
84. Williams-Carrier R, Brewster C, Belcher SE, Rojas M, Chotewutmontri P, Ljungdahl S, et al. The *Arabidopsis pentatricopeptide* repeat protein LPE1 and its maize ortholog are required for translation of the chloroplast psbJ RNA. Plant J. 2019;99: 56–66. doi:10.1111/tpj.14308
85. Díaz BG, Zucchi MI, Alves-Pereira A, de Almeida CP, Moraes ACL, Vianna SA, et al. Genome-wide SNP analysis to assess the genetic population structure and diversity of Acrocomia species. PLoS One. 2021;16: 1–24. doi:10.1371/journal.pone.0241025
86. Viana JPG, Siqueira MVBM, Araujo FL, Grando C, Sujii PS, De Aguiar Silvestre E, et al. Genomic diversity is similar between Atlantic Forest restorations and natural remnants for the native tree *Casearia sylvestris* Sw. PLoS One. 2019;13: 1–14. doi:10.1371/journal.pone.0192165
87. Cordeiro EMG, Macrini CM, Sujii PS, Schwarcz KD, Pinheiro JB, Rodrigues RR, et al. Diversity, genetic structure, and population genomics of the tropical tree *Centrolobium tomentosum* in remnant and restored Atlantic forests. Conserv Genet. 2019;20: 1073–1085. doi:10.1007/s10592-019-01195-z
88. Allendorf FW, Hohenlohe PA, Luikart G. Genomics and the future of conservation genetics. Nat Rev Genet. 2010;11: 697–709. doi:10.1038/nrg2844
89. Willing EM, Dreyer C, van Oosterhout C. Estimates of genetic differentiation measured by fst do not necessarily require large sample sizes when using many snp markers. PLoS One. 2012;7: 1–7. doi:10.1371/journal.pone.0042649
90. Hamblin MT, Warburton ML, Buckler ES. Empirical comparison of simple sequence repeats and single nucleotide polymorphisms in assesment of maize diversity and relatedness. PLoS One. 2007;2. doi:10.1371/journal.pone.0001367

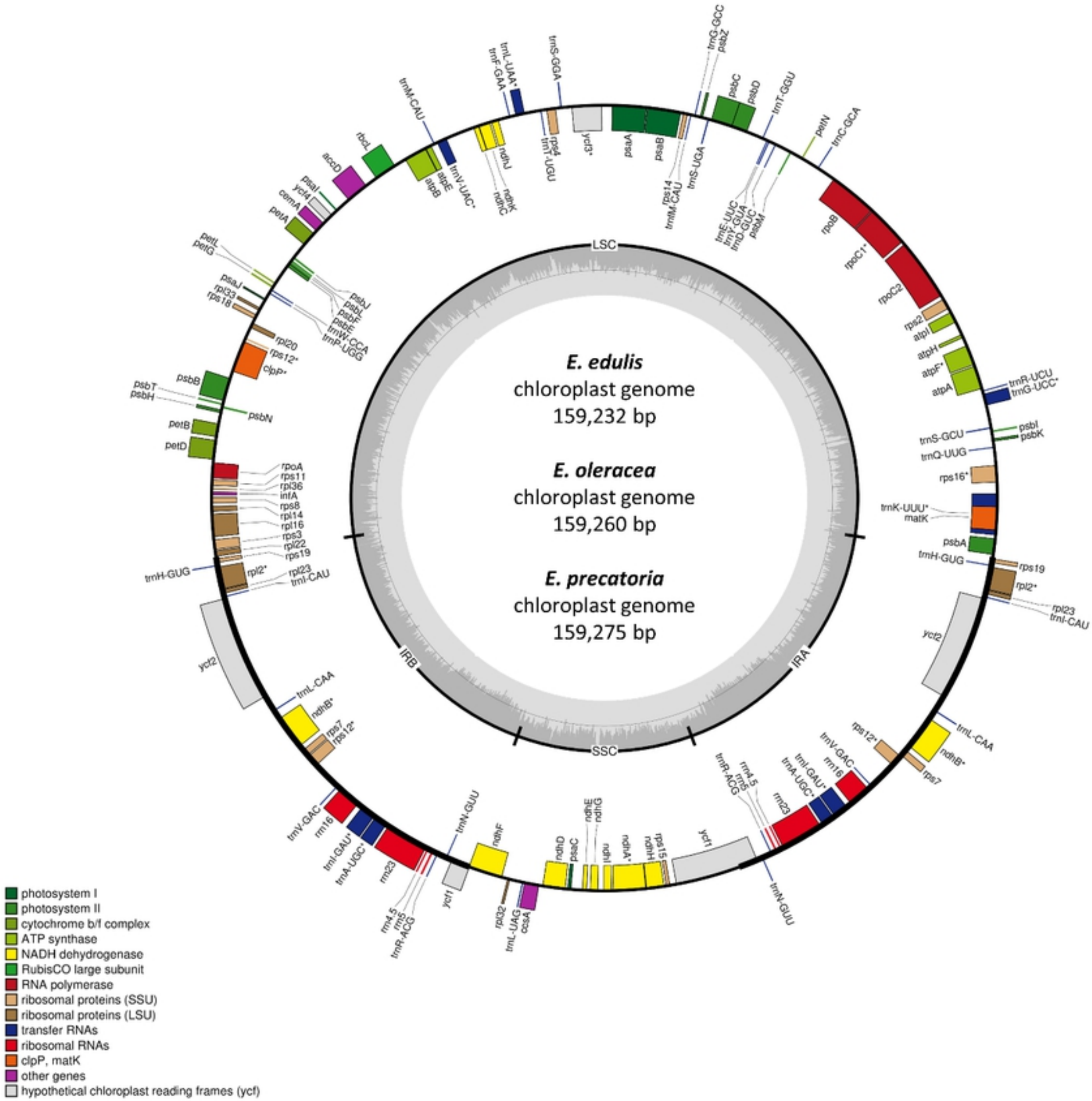
E. edulis
chloroplast genome
159,232 bp
E. oleracea
chloroplast genome
159,260 bp
E. precatoria
chloroplast genome
159,275 bp
LSC
SSC
IRA
IRB
photosystem I
photosystem II
cytochrome b/f complex
ATP synthase
NADH dehydrogenase
RubisCO large subunit
RNA polymerase
ribosomal proteins (SSU)
ribosomal proteins (LSU)
transfer RNAs
ribosomal RNAs
clpP, matK
other genes
hypothetical chloroplast reading frames (ycf)

Fig 1

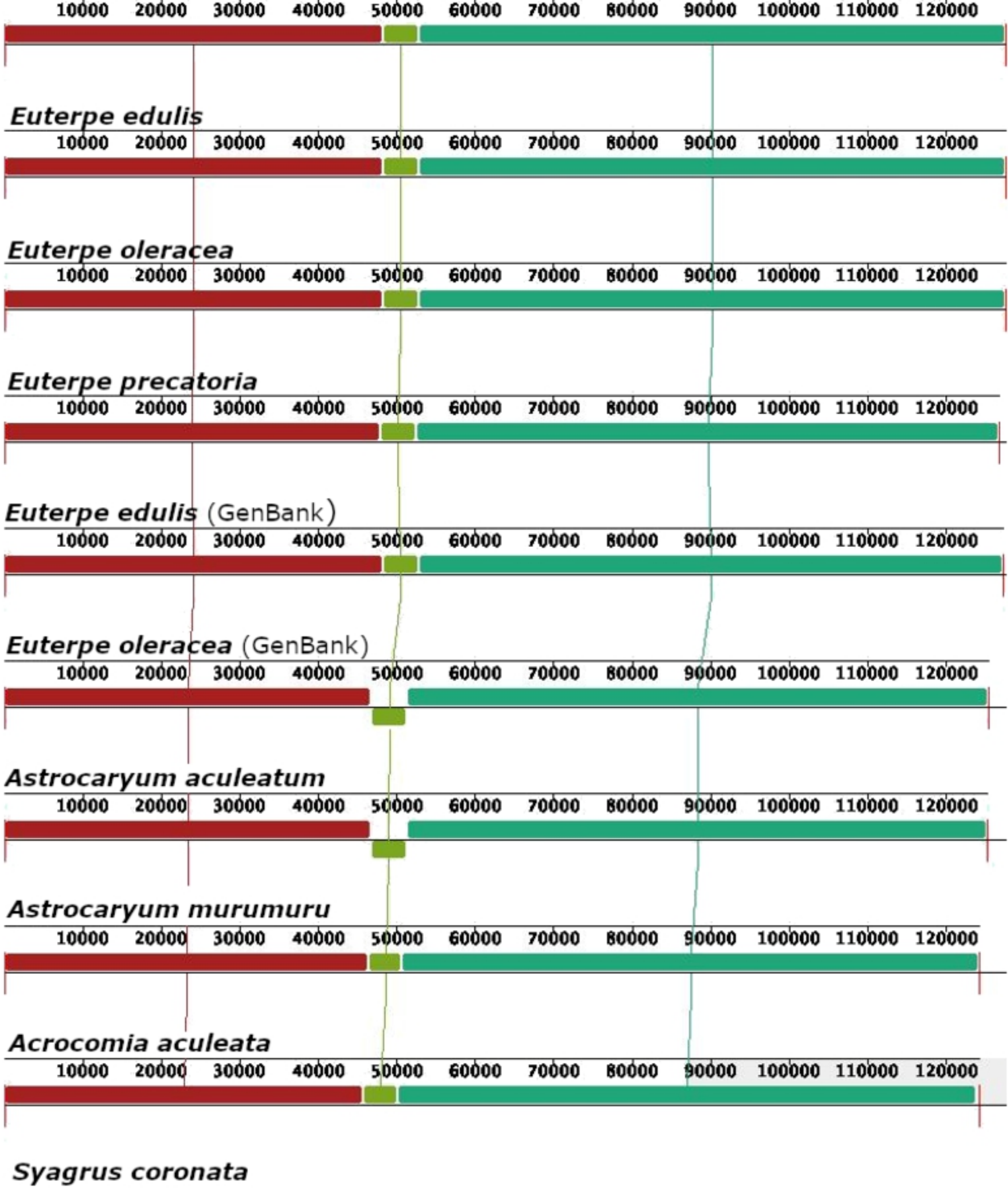
10000 20000 30000 40000 50000 60000 70000 80000 90000 100000 110000 120000
Euterpe edulis
Euterpe oleracea
Euterpe precatoria
Euterpe edulis (GenBank)
Euterpe oleracea (GenBank)
Astrocaryum aculeatum
Astrocaryum murumuru
Acrocomia aculeata
Syagrus coronata

Fig 2

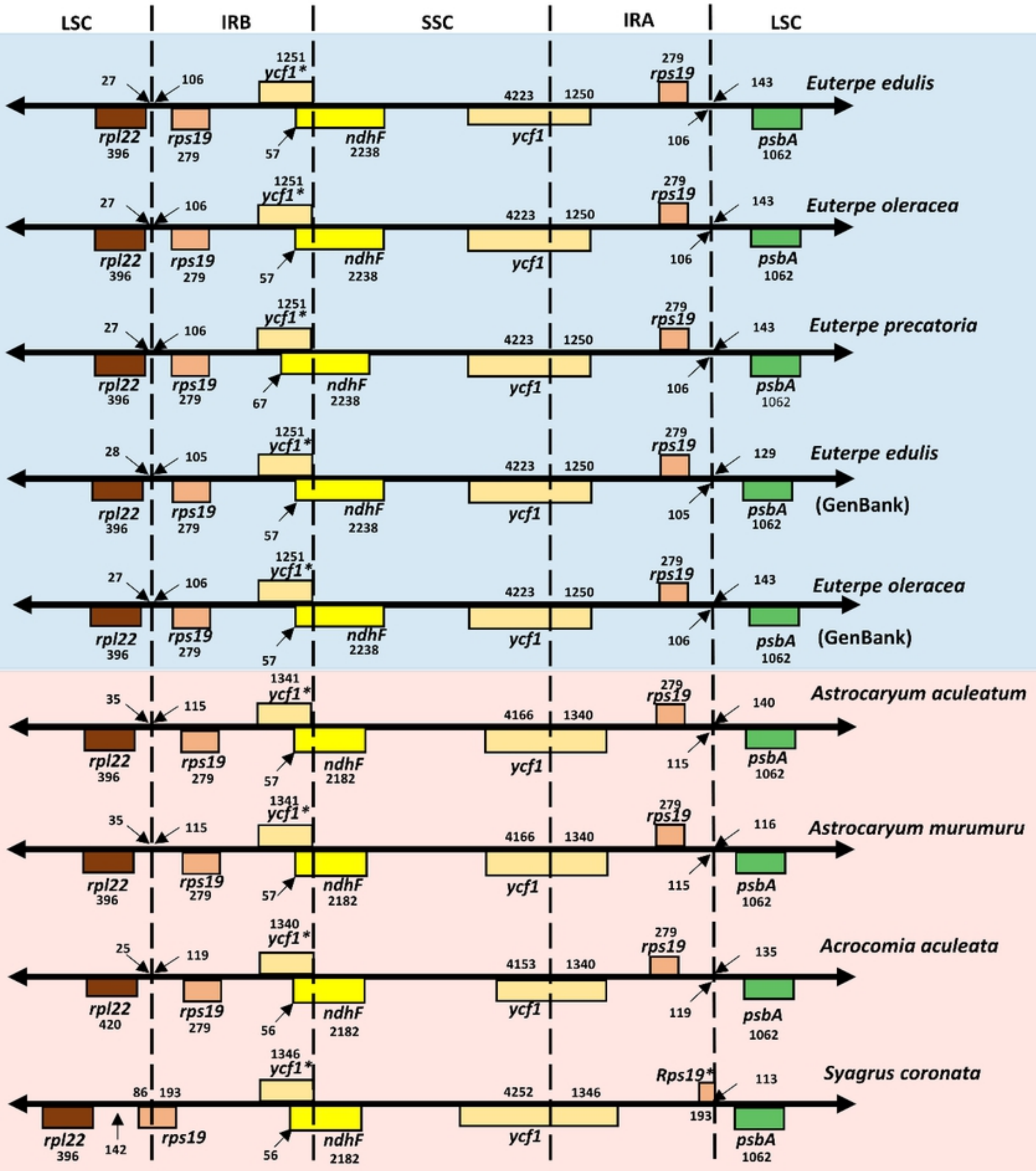
LSC
IRB
SSC
IRA
LSC
Euterpe edulis
Euterpe oleracea
Euterpe precatoria
Euterpe edulis (GenBank)
Euterpe oleracea (GenBank)
Astrocaryum aculeatum
Astrocaryum murumuru
Acrocomia aculeata
Syagrus coronata
rpl22
rps19
ycf1*
ndhF
ycf1
psbA
Rps19*

Fig 3

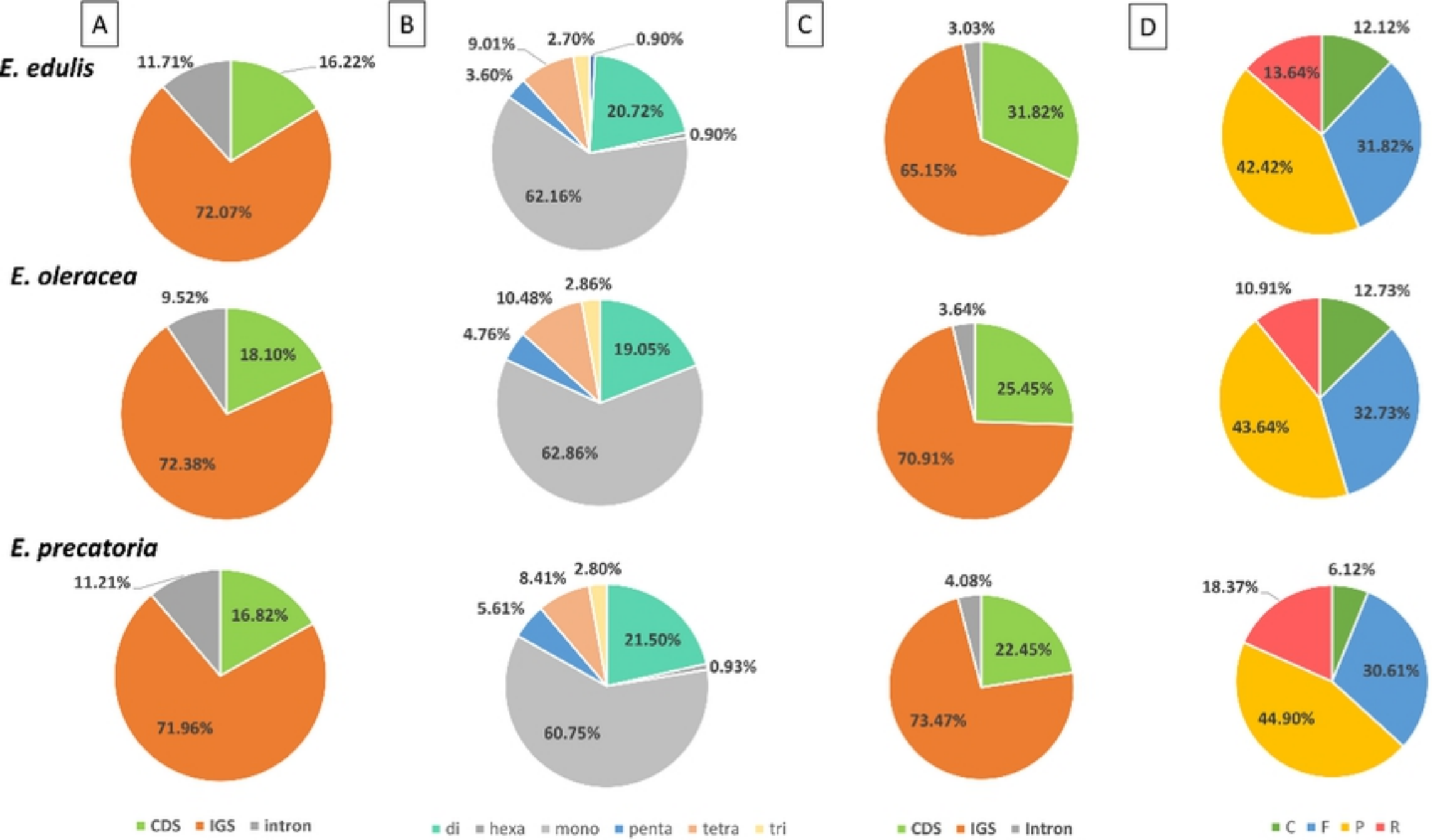
A
B
C
D
E. edulis
11.71%
16.22%
72.07%
9.01%
2.70%
0.90%
3.60%
20.72%
0.90%
62.16%
3.03%
31.82%
65.15%
12.12%
13.64%
31.82%
42.42%
E. oleracea
9.52%
18.10%
72.38%
2.86%
10.48%
4.76%
19.05%
62.86%
3.64%
25.45%
70.91%
10.91%
12.73%
32.73%
43.64%
E. precatoria
11.21%
16.82%
71.96%
8.41%
2.80%
5.61%
21.50%
0.93%
60.75%
4.08%
22.45%
73.47%
18.37%
6.12%
30.61%
44.90%
CDS
IGS
intron
di
hexa
mono
penta
tetra
tri
CDS
IGS
Intron
C
F
P
R

Fig 4

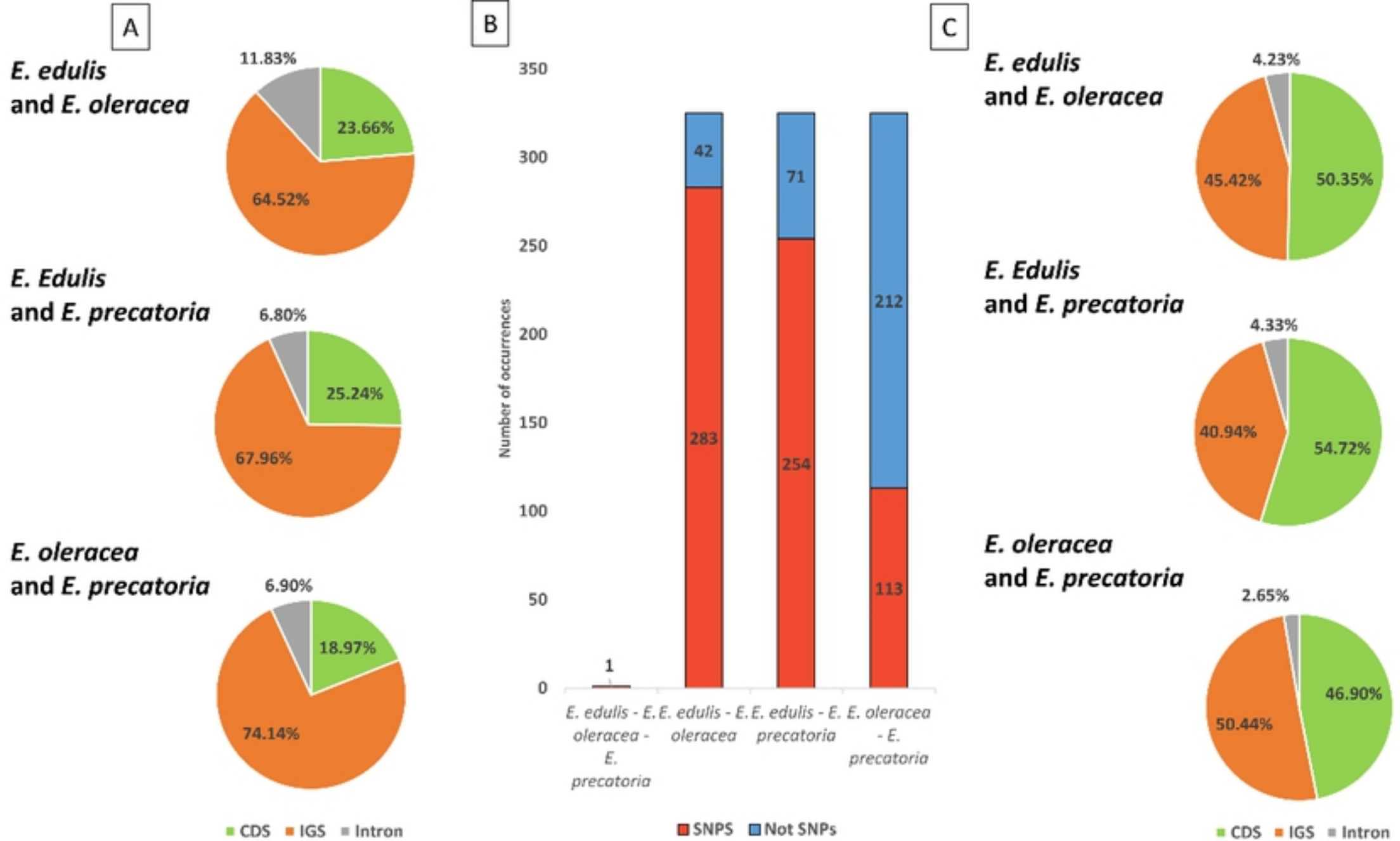
A
E. edulis
and E. oleracea
11.83%
23.66%
64.52%
E. Edulis
and E. precatoria
6.80%
25.24%
67.96%
E. oleracea
and E. precatoria
6.90%
18.97%
74.14%
CDS
IGS
Intron
B
Number of occurrences
350
300
250
200
150
100
50
0
1
42
283
71
254
212
113
E. edulis - E. oleracea - E. precatoria
E. edulis - E. oleracea
E. edulis - E. precatoria
E. oleracea - E. precatoria
SNPS
Not SNPs
C
E. edulis
and E. oleracea
4.23%
45.42%
50.35%
E. Edulis
and E. precatoria
4.33%
40.94%
54.72%
E. oleracea
and E. precatoria
2.65%
50.44%
46.90%
CDS
IGS
Intron

Fig 5

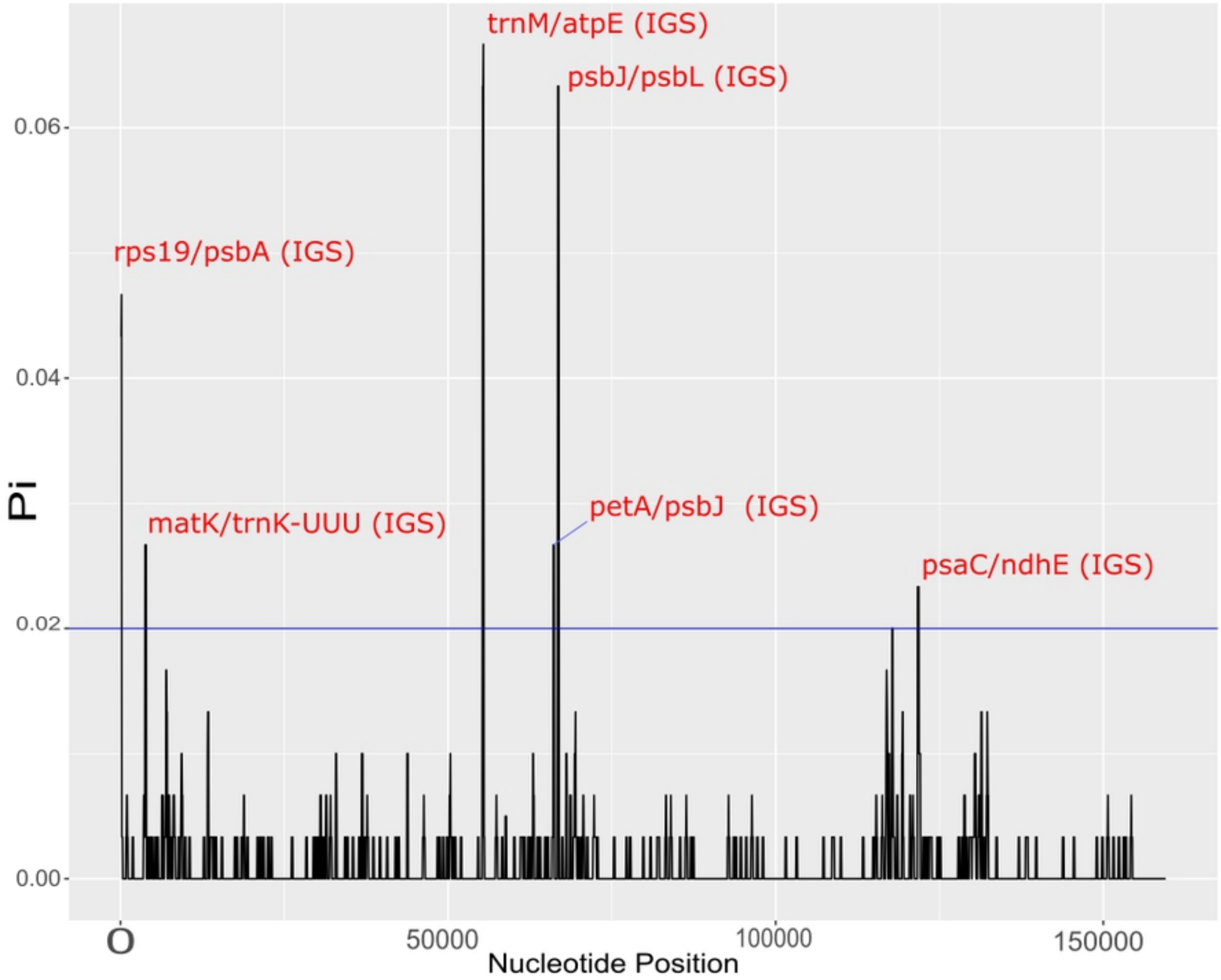
trnM/atpE (IGS)
psbJ/psbL (IGS)
rps19/psbA (IGS)
matK/trnK-UUU (IGS)
petA/psbJ (IGS)
psaC/ndhE (IGS)
0.06
0.04
0.02
0.00
Pi
0
50000
100000
150000
Nucleotide Position

Fig 6

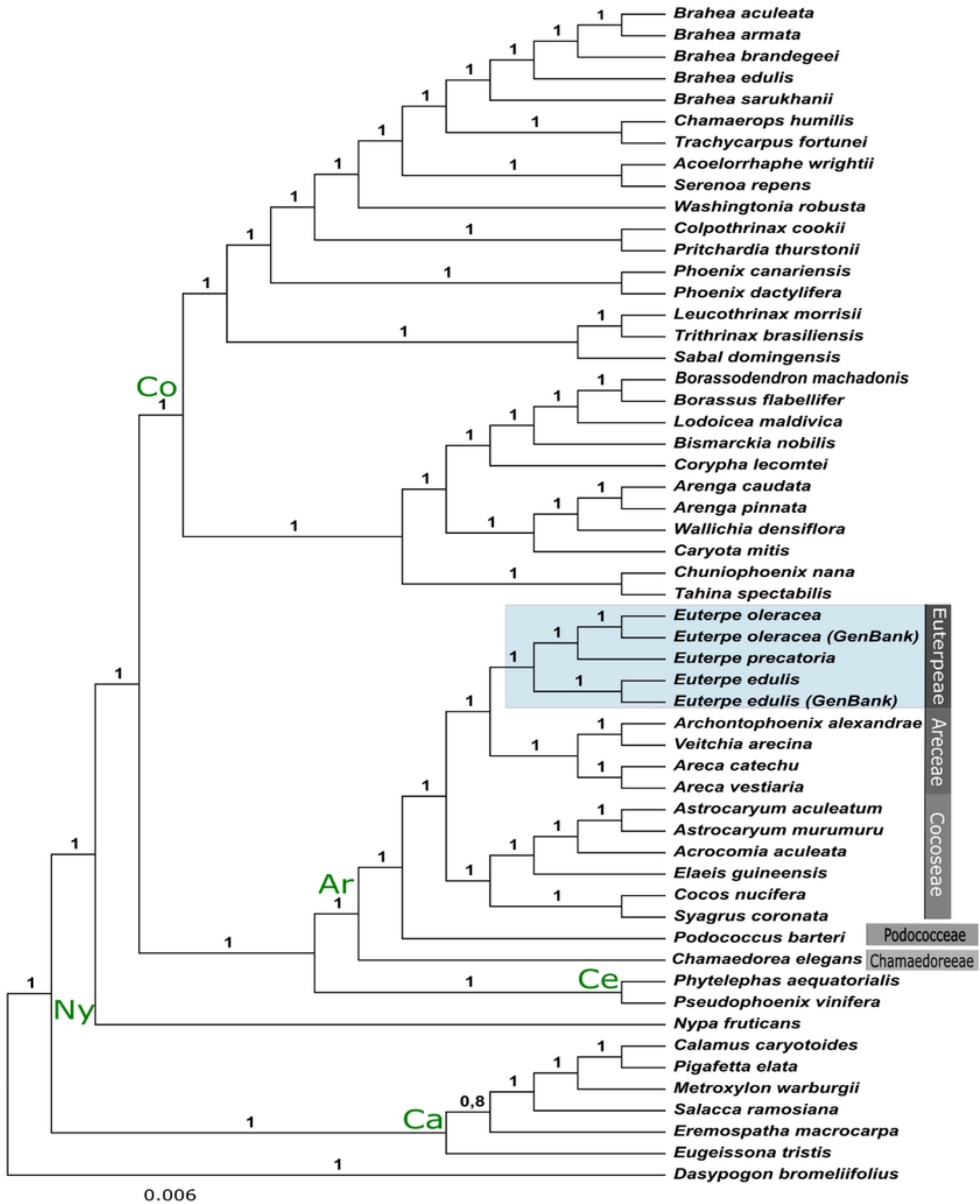
Brahea aculeata
Brahea armata
Brahea brandegeei
Brahea edulis
Brahea sarukhanii
Chamaerops humilis
Trachycarpus fortunei
Acoelorrhaphe wrightii
Serenoa repens
Washingtonia robusta
Colpothrinax cookii
Pritchardia thurstonii
Phoenix canariensis
Phoenix dactylifera
Leucothrinax morrisii
Trithrinax brasiliensis
Sabal domingensis
Borassodendron machadonis
Borassus flabellifer
Lodoicea maldivica
Bismarckia nobilis
Corypha lecomtei
Arenga caudata
Arenga pinnata
Wallichia densiflora
Caryota mitis
Chuniophoenix nana
Tahina spectabilis
Euterpe oleracea
Euterpe oleracea (GenBank)
Euterpe precatoria
Euterpe edulis
Euterpe edulis (GenBank)
Archontophoenix alexandrae
Veitchia arecina
Areca catechu
Areca vestiaria
Astrocaryum aculeatum
Astrocaryum murumuru
Acrocomia aculeata
Elaeis guineensis
Cocos nucifera
Syagrus coronata
Podococcus barteri
Chamaedorea elegans
Phytelephas aequatorialis
Pseudophoenix vinifera
Nypa fruticans
Calamus caryotoides
Pigafetta elata
Metroxylon warburgii
Salacca ramosiana
Eremospatha macrocarpa
Eugeissona tristis
Dasypogon bromeliifolius
Co
Ar
Ce
Ny
Ca
Euterpeae
Areceae
Cocoseae
Podococceae
Chamaedoreeae
0,8
0.006

Fig 7