UNIVERSIDADE FEDERAL DO PARANÁ

RAQUEL SANTOS DA SILVA

CARACTERIZAÇÃO DO PLASTOMA DE *Bactris gasipaes* E ANÁLISES EVOLUTIVAS NA TRIBO COCOSEAE (ARECACEAE)

CURITIBA

2020

RAQUEL SANTOS DA SILVA

CARACTERIZAÇÃO DO PLASTOMA DE *Bactris gasipaes* E ANÁLISES EVOLUTIVAS NA TRIBO COCOSEAE (ARECACEAE)

Dissertação apresentada ao curso de Pós-Graduação em Botânica, Setor de Ciências Biológicas, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Mestre em Botânica.

Orientador(a): Profª. Drª. Leila do Nascimento Vieira

CURITIBA

2020

FICHA CATALOGRÁFICA

Universidade Federal do Paraná
Sistema de Bibliotecas
(Giana Mara Seniski Silva – CRB/9 1406)

Silva, Raquel Santos da
Caracterização do plastoma de *Bactris gasipaes* e análises evolutivas na tribo Cocoseae (Arecaceae). / Raquel Santos da Silva. – Curitiba, 2020.
39 p.: il.

Orientador: Leila do Nascimento Vieira

Dissertação (mestrado) - Universidade Federal do Paraná, Setor de Ciências Biológicas. Programa de Pós-Graduação em Botânica.

1. Pupunha. 2. Evolução molecular. 3. Palmeira. I. Título. II. Vieira, Leila do Nascimento, 1987-. III. Universidade Federal do Paraná. Setor de Ciências Biológicas. Programa de Pós-Graduação em Botânica.

CDD (22. ed.) 584.5

MINISTÉRIO DA EDUCAÇÃO
SETOR DE CIÊNCIAS BIOLÓGICAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO BOTÂNICA - 40001016004P9

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em BOTÂNICA da Universidade Federal do Paraná foram convocados para realizar a arguição da dissertação de Mestrado de **RAQUEL SANTOS DA SILVA** intitulada: **CARACTERIZAÇÃO DO PLASTOMA DE Bactris gasipaes E ANÁLISES EVOLUTIVAS NA TRIBO COCOSEAE (ARECACEAE)**, sob orientação da Profa. Dra. LEILA DO NASCIMENTO VIEIRA, que após terem inquirido a aluna e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 19 de Outubro de 2020.

Assinatura Eletrônica
19/10/2020 18:23:54.0
LEILA DO NASCIMENTO VIEIRA
Presidente da Banca Examinadora (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
19/10/2020 18:23:32.0
DORIANE PICANÇO RODRIGUES
Avaliador Externo (UNIVERSIDADE FEDERAL DO AMAZONAS)

Assinatura Eletrônica
19/10/2020 18:00:08.0
LILIAN DE OLIVEIRA MACHADO
Avaliador Externo (UNIVERSIDADE FEDERAL DE SANTA CATARINA)

Dedico este estudo à minha família, em especial, à minha amada Mãe Helena Machado.

AGRADECIMENTOS

A Deus, pela sabedoria, dedicação à mim concedida para realização deste estudo, por ter colocado em meu caminho pessoas maravilhosas e por todas as bênçãos conquistadas ao longo da minha vida pessoal e acadêmica, em especial, neste período que estive longe dos familiares.

Agradeço a CAPES pela bolsa de Mestrado, ao CNPq pelo investimento recebido por meio do projeto aprovado na Chamada Universal, ao Programa de Pós–graduação em Botânica, ao Instituto Nacional de Pesquisas da Amazônia (INPA), por meio do Prof. Charles Roland Clement, ao Departamento de Bioquímica e Biologia Molecular, ao Núcleo de Fixação Biológica de Nitrogênio, aos professores pelos ensinamentos ofertados neste período, sendo representado pelo Prof. Hugo Pacheco de Freitas Fraga, e agradecer especialmente, a minha orientadora Dra. Leila do Nascimento Vieira, por toda dedicação, paciência, conhecimento, compreensão e pelos momentos ímpares que compartilhamos. Leila e Hugo, vocês foram responsáveis por todo aprendizado que obtive neste período, vocês são exemplos de profissionais. Sou grata e tenho enorme admiração por vocês.

A minha família, em especial à minha mãe Helena Machado e ao meu irmão Rodrigo Santos, pelo apoio, pelo incentivo, pela força e por serem meus exemplos, pois foi por eles e para eles que consegui finalizar mais esse ciclo em minha vida profissional. Aos meus avós Maria Augusta Machado (*in memorian*), José Pereira e Raquel Félix, aos meus tios Ana Cláudia, Álvaro Carvalho, Alessandro Silva, Cristiane Silva, Ciro Rolim, Jameson Silva, que sempre me apoiaram e me deram suporte.

Aos parceiros de laboratório Jeison Magnabosco e Michelle Zavala-Paez. Aos amigos que sempre estiveram comigo Gabrielle Maciel, Guilherme Almeida e Jackeline Anjos. Levarei todos comigo sempre, graças a vocês conquistei momentos muito especiais e pude aprender que apesar da distância, das dificuldades e dos diferentes costumes foi possível desenvolver laços de amor e carinho de uma família.

RESUMO

A família Arecaceae inclui espécies de palmeiras distribuídas por regiões tropicais e subtropicais do mundo, sendo a subfamília Arecoideae sua maior representante e apresentando algumas relações intergenéricas ambíguas nas subtribos. Arecaceae possui espécies de grande importância ecológica e econômica, devido a gerarem vários produtos e por serem altamente utilizadas pela população numa escala global, com uma única espécie completamente domesticada no Neotrópico, a palmeira *Bactris gasipaes*. Genomas plastidiais são utilizados para estudos filogenéticos e evolutivos, assim como as regiões hipervariáveis, que fornecem informações para esclarecer relações filogenéticas não resolvidas. Por isso, foram realizadas análises estruturais comparativas, inferências filogenéticas e a identificação das regiões hipervariáveis nas três subtribos de Cocoseae (Attaleinae, Bactridinae e Elaeidinae), assim como sequenciado e caracterizado o plastoma de *Bactris gasipaes*. O plastoma de *Bactris gasipaes* possui 156.646 pb, com 113 genes únicos, 79 genes codificadores de proteínas, 30 genes de tRNA e 4 genes de rRNA. Dentre eles, quatro genes com o códon de iniciação alternativo (*cemA*, *rps19*, *rpl2* e *ndhD*). Os plastomas das espécies da tribo Cocoseae são bem conservados, com identidade de 97,3%, uma variação no comprimento de ~2kb e uma inversão de 4,5 kb em *Astrocaryum*. Há uma variação nas junções LSC/IR e IR/SSC com o gene *rps19* e uma sobreposição de 56 pb do gene *ndhF* no gene *ycf1*. Entre as estimativas para identificar as regiões com maior variação, a variação em sequência (SV%) e sítios parcimoniosos informativos (PIS%) apresentaram maior variação do que os marcadores moleculares plastidiais comumente utilizados em análises filogenéticas nas palmeiras. Como o esperado, em geral eles apresentaram menor variação do que os marcadores nucleares *PRK* e *RPB2*. Assim, foram descritos quatro novas regiões promissoras com base nos valores de SV% e PIS%, uma nova região baseada em SV% e três novas regiões baseadas em PIS%. As análises filogenéticas inferidas por máxima verossimilhança apresentaram topologia idêntica e consistente dentro da tribo Cocoseae, confirmando o monofiletismo das subtribos Bactridinae e Attaleinae.

Palavras-chave: Pupunha. Evolução molecular. Plastomas. Palmeiras. Filogenômica.

ABSTRACT

The family Arecaceae includes palm species distributed in tropical and subtropical regions of the world. Arecoideae is the most species-rich subfamily and presents some ambiguous intergeneric relationships in the subtribes. Arecaceae has species of great ecological and economic importance, due to the generation of various products highly used by the population on a global scale, with a single species fully domesticated in the Neotropics: the *Bactris gasipaes*. Plastid genomes are used for phylogenetic and evolutionary studies, as well as hypervariable regions, which provide information to clarify unresolved phylogenetic relationships. Therefore, we performed comparative structural analysis, phylogenetic inference, and identified hypervariable regions in the three subtribes of Cocoseae (Attaleinae, Bactridinae, and Elaeidinae), as well as sequenced and characterized the plastome of *Bactris gasipaes*. The *Bactris gasipaes* plastome is 156,646 bp in lenght, with 113 unique genes, 79 protein-coding genes, 30 tRNA genes, and 4 rRNA genes. Among them, four genes have alternative initiation codons (*cemA*, *rps19*, *rpl2* and *ndhD*). The plastome of the species from tribe Cocoseae are highly conserved, with a 97.3% identity, a variation in the length of ~ 2kb, and an inversion of 4.5 kb in *Astrocaryum*. There is a variation in the LSC/IR and IR/SSC junctions with the *rps19* gene and a 56 bp overlap of the *ndhF* gene in the *ycf1* gene. Among the estimates to identify the regions with the greatest variation, sequence variation (SV%) and parsimony informative sites (PIS%) showed greater variation than the plastid molecular markers commonly used in phylogenetic analysis in palm trees. As expected, they showed lower values than the nuclear markers *PRK* and *RPB2.* Thus, we described four new promising regions based on the values of SV and PIS, one new region based on SV, and three new regions based on PIS. The phylogenetic inferences by maximum likelihood (ML) presented identical and consistent topology within the Cocoseae tribe, confirming the monophyly of the subtribes Bactridinae and Attaleinae.

Keywords: Peach palm. Molecular evolution. Plastome. Palms. Phylogenomics.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO E JUSTIFICATIVA

Esta dissertação fez parte do projeto “Obtenção de um genoma de referência da palmeira amazônica *Bactris gasipaes* como subsídio para estudos evolutivos” com financiamento aprovado pela Chamada MCTIC/CNPq n° 28/2018 – Universal/Faixa A do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). As análises foram realizadas nos laboratórios de Evolução Molecular de Algas e Plantas sob a orientação da Prof.ª Dr.ª Leila do Nascimento Vieira e no Núcleo de Fixação Biológica de Nitrogênio. A pesquisa contou com o apoio do Instituto Nacional de Pesquisa da Amazônia (INPA), por meio do pesquisador e curador do Banco Ativo de Germoplasma da Pupunha, o Dr. Charles R. Clement, para a coleta e envio do material vegetal.

A família Arecaceae é amplamente distribuída pelas regiões tropicais e subtropicais do mundo, possui cinco subfamílias, sendo Arecoideae a sua maior representante, com a tribo Cocoseae apresentando três subtribos Attaleinae, Bactridinae e Elaeidinae. Estas subtribos são amplamente estudadas, mas dentro de Attaleinae e Bactridinae há algumas relações intergenéricas ambíguas ainda não resolvidas (Baker & Dransfield, 2016). Dessa forma, a fim de verificar essas relações e confirmar a monofilia de Bactridinae, foram realizadas análises estruturais comparativas, inferências filogenéticas e identificação de regiões hipervariáveis nas três subtribos de Cocoseae, assim como sequenciar e caracterizar o plastoma de *Bactris gasipaes*, a qual é única espécie completamente domesticada no Neotrópico (Clement, 1992). Essa será a primeira sequência de plastoma desta espécie disponível nos bancos de dados.

As palmeiras fornecem vários produtos de alto valor para o processamento industrial (por exemplo, fibras, materiais de construção, óleo, compostos medicinais, palmito, frutas) e são altamente relevantes para as comunidades indígenas e tradicionais (Clement, 1992; Balslev et al., 2016). Além de possuir espécies de grande importância ecológica, devido sua interação com polinizadores (Barfod et al., 2011) e/ou com os animais frugívoros (Onstein et al., 2017; Muñoz et al., 2019).

A utilização do protocolo desenvolvido por Vieira et al. (2014) e modificada por Sakaguchi et al. (2017) possibilitou o sequenciamento do genoma plastidial completo de uma espécie de *Bactris gasipaes*, *Bactris gasipaes* var. *chichagui* Tipo 1, de

forma rápida e eficiente, possibilitando análises comparativas da estrutura do genoma plastidial dessa espécie com os outras espécies de palmeiras.

Deste modo o sequenciamento do genoma plastidial de *Bactris gasipaes* possibilitará a ampliação de estudos evolutivos, estruturais e filogenômicos em Cocoseae, pois as espécies das subtribos possuem grande importância ecológica e econômica para o Brasil.

2 CAPÍTULO I: The plastome sequence of *Bactris gasipaes* and evolutionary analysis in tribe Cocoseae (Arecaceae)

Raquel S. Silva[1]; Charles R. Clement[2]; Eduardo Balsanelli[3]; Valter A. de Baura[3]; Emanuel M. de Souza[3]; Hugo P. F. Fraga[1]; Leila N. Vieira[1*]

[1] Departamento de Botânica, Universidade Federal do Paraná, Curitiba, Paraná, Brazil

[2] Coordenação de Tecnologia e Inovação, Instituto Nacional de Pesquisas da Amazônia, Manaus, AM, Brazil

[3] Departamento de Bioquímica e Biologia Molecular, Núcleo de Fixação Biológica de Nitrogênio, Universidade Federal do Paraná, Curitiba, Paraná, Brazil

Abstract

The family Arecaceae is distributed throughout tropical and subtropical regions of the world. Among the five subfamilies, Arecoideae is the most species-rich and still contains some ambiguous inter-generic relationships, such as those within subtribes Attaleinae and Bactridinae. The hypervariable regions of plastid genomes (plastomes) are interesting tools to clarify unresolved phylogenetic relationships. We sequenced and characterized the plastome of *Bactris gasipaes* (Bactridinae) and compared it with eight species from the three Cocoseae sub-tribes (Attaleinae, Bactridinae, and Elaeidinae) to perform comparative analysis and to identify hypervariable regions. The *Bactris gasipaes* plastome has 156,646 bp, with 113 unique genes. Among them, four genes have an alternative start codon (*cemA*, *rps19*, *rpl2*, and *ndhD*). Plastomes are highly conserved within tribe Cocoseae: 97.3 % identity, length variation of ~2 kb, and a single ~4.5 kb inversion in *Astrocaryum* plastomes. The LSC/IR and IR/SSC junctions vary among the subtribes: in Bactridinae and Elaeidinae the *rps19* gene is completely contained in the IR region; in the subtribe Attaleinae the *rps19* gene is only partially contained in the IRs. The hypervariable regions selected according to sequence variation (SV%) and frequency of parsimony informative sites (PIS%) revealed plastome regions with great potential for molecular analysis. The ten regions with greatest SV% showed higher variation than the plastid molecular markers commonly used for phylogenetic analysis in palms. The phylogenetic trees based on the plastomes and the hypervariable regions (SV%) datasets had well-resolved relationships, with consistent topologies within tribe Cocoseae, and confirm the monophyly of the subtribes Bactridinae and Attaleinae.

Keywords: Peach palm; Molecular evolution; Plastome; Palms; Phylogenomics.

INTRODUCTION

The family Arecaceae comprises 181 genera and about 2,600 species distributed throughout tropical and subtropical regions of the world (Baker and Dransfield, 2016). The most recent taxonomic review in Arecaceae, published by Baker and Dransfield (2016), recognizes five subfamilies: Arecoideae, Calamoideae, Ceroxyloideae, Coryphoideae, and Nypoideae. Arecoideae is the largest subfamily of Arecaceae, with 14 tribes and 108 genera. Several Arecoideae tribes were extensively studied, but some intergeneric relationships remain ambiguous: as those within subtribes Attaleinae and Bactridinae (Baker and Dransfield, 2016). These subtribes are both from tribe Cocoseae, which includes Elaeidinae as its third subtribe.

Plastid genomes (plastomes) are a useful tool for phylogenetic and evolutionary studies (Rogalski et al., 2015). Hypervariable regions in plastomes can provide information to elucidate phylogenetic relationships that are not yet well resolved (Shaw et al., 2007; Smidt et al., 2020). However, these highly variable regions vary between clades, and their identification might be necessary for each different taxonomic level (Shaw et al., 2007). Angiosperms show structural rearrangements, loss of genes, introns, and heterogeneous nucleotide substitution rates in protein-coding genes among their plastomes (Wicke et al., 2011; Ruhlman and Jansen, 2014). Also, the fact that the palm trees have a low rate of evolution in the plastome (Wilson et al., 1990) makes the identification of clade-specific hypervariable regions singularly relevant for the group.

In addition to the phylogenetic analysis, these new or neglected hypervariable regions might also be powerful molecular markers for analysis of population genetic structure, including those among wild and domesticated populations. The family Arecaceae includes species of great ecological importance, either by their interaction with pollinators (Barfod et al., 2011) or with frugivorous animals (Onstein et al., 2017; Muñoz et al., 2019). Also, palm trees provide several high-value products for industry processing (e.g., fibers, construction materials, oil, medicinal compounds, heart of palm, fruits) and are highly relevant to indigenous and traditional communities (Clement, 1992; Balslev, 2016). Among the highly-used palm trees, the coconut tree (*Cocos nucifera* L.), the date palm (*Phoenix dactylifera* L.), and the oil palm (*Elaeis guineensis* Jacq.) stand out by their global economic

importance, and the peach palm (*Bactris gasipaes* Kunth) as the only completely domesticated palm in the Neotropics (Clement, 1992). Many Cocoseae species included in our analysis have interesting domestication history, as *Bactris gasipaes*, *Elaeis guineensis*, *Cocos nucifera (Clement, 1992; Clement et al., 2017; Muñoz-Peréz et al., 2019).* Others are from genera of species with traditional use by native populations, as *Butia*, *Astrocaryum*, *Acrocomia, and Syagrus (Rivas and Condón, 2015; Montoya et al., 2016; Levis et al., 2018; Pereira Cruz et al., 2020)*.

Thus, here we used representatives from the three sub-tribes of Cocoseae (Attaleinae, Bactridinae, and Elaeidinae) to perform comparative structural analysis, phylogenetic inference, and to identify hypervariable regions. In our study, we analyzed eight species from the tribe Cocoseae and newly sequenced and characterized the plastome of *Bactris gasipaes*.

MATERIAL AND METHODS

Species sampling

We collected fresh leaves from a wild individual of *Bactris gasipaes* (i.e., *Bactris gasipaes* var. *chichagui* type 1; accession number F0205/83) in the core collection of Peach palm Active Germplasm Bank (Cristo-Araújo et al., 2015) at the National Research Institute for Amazonia (INPA, Manaus, AM, Brazil).

Plastomes and nuclear markers sequences from seven genera of tribe Cocoseae were used, including individuals from its three subtribes: Attaleinae (3 genera; 3 species), Elaeidinae (1 species), and Bactridinae (3 genera; 4 species). The species name and the GenBank accession numbers for sequences used in the analysis are in Table S1.

Bactris gasipaes plastome sequencing

The extraction of plastid-enriched DNA was based on the methodology described by Vieira et al. (2014) and modified by Sakaguchi et al. (2017), proportionally adjusting the buffer volumes for 8 g of fresh leaves. The DNA extraction was performed with CTAB buffer, as described by Shi et al. (2012). The obtained DNA was purified with the Genomic DNA Clean & Concentrator™-10 Kit (Zymo Research, Irvine, CA, USA). The purified DNA was quantified using Qubit™

dsDNA HS Assay kit (Thermo Fisher Scientific, Carlsbad, CA, USA) in Qubit™ Fluorometer (Thermo Fisher Scientific). Libraries were prepared with Nextera XT DNA Library Preparation Kit (Illumina, San Diego, CA, USA) and sequenced in Illumina MiSeq® (Illumina), obtaining 250 bp paired-end reads.

Plastome assembly was performed using CLC Genomics Workbench v.8.0 (Qiagen, Germantown, MD, USA) software in *de novo* strategy. The *Acrocomia aculeata* plastome was used as a reference for the ordering of contigs. Plastome annotation was performed using Geneious Prime® (Biomatters, Auckland, New Zealand). For all genes, manual verification was performed, adjusting the initial and terminal codons. The final plastome sequence was deposited in the GenBank database (MW054718).

Plastome structural analysis in tribe Cocoseae

The comparative analysis to identify structural rearrangements between the plastomes of the Cocoseae species was carried out using eight species (Table S1), excluding one IR from all plastomes, and using the progressive alignment on Mauve software (Darling et al., 2004). The IRScope software (Amiryousefi et al., 2018) was used to visualize and compare the plastome junctions (IRb/LSC; IRb/SSC, SSC/IRa; IRa/LSC).

Identification of hypervariable regions

We estimated the variability of the sequences with the formula proposed by Shaw et al. (2014), adapted and used in Zavala-Páez et al. (2020). First, we individually aligned each collinear coding sequence (CDS), intergenic spacers (IGS), and introns of the plastomes (list of species in Table S1) using MAFFT v.7 software (Katoh & Standley 2013). Then, the alignments were imported in the software DNAsp v6.12.03 (Rozas et al., 2017) to obtain the number of invariable sites (monomorphic), parsimony informative sites (PIS), number of substitutions, and number of InDels events. Sequence variability (SV) was calculated using the formula: *SV% = [(number of substitutions + number of InDels events)/(number of substitutions + number of InDels + invariable sites)] x 100*. The frequency of PIS was calculated using the

formula: *PIS% = [(number of parsimony informative sites/number of substitutions + number of InDels + invariable sites)] x 100.*

The ten regions with the highest SV% and PIS% values were selected to carry out the subsequent analysis. The plastid markers *matK*, *trnQ-rps16*, *rps16* intron, *trnD-trnT*, *trnL-trnF* (Asmussen et al., 2006; Eiserhardt et al., 2011) and the nuclear markers PRK and RPB2 (Baker et al., 2011; Eiserhardt et al., 2011), commonly used for phylogenetic analysis in Arecaceae, were used for comparative purposes and subjected to the same procedure to obtain the PIS% and SV% values (list of species in Table S1).

Phylogenetic inferences

Phylogenetic inferences were made including the following species of the tribe Cocoseae: *Bactris gasipaes*, *Acrocomia aculeata*, *Astrocaryum aculeatum*, *Astrocarym murumuru*, *Butia eriospatha*, *Cocos nucifera*, *Elaeis guineensis*, *Syagrus coronata*, and two species as outgroup: *Brahea brandegeei* (Purpus) H. E. Moore (subfamily Coryphoideae) and *Archontophoenix alexandrae* (F.Muell.) H.Wendl. & Drude (subfamily Arecoideae). Three data sets were used: i) the plastome alignment (one IR excluded); ii) the ten regions with the greatest SV% value; iii) the ten regions with the greatest PIS% value.

Plastome alignment was performed using progressive alignment on Mauve (Darling et al., 2004) implemented in Geneious Prime® v.2020.1.2. The Locally Collinear Blocks (LCBs) identified by Mauve were individually extracted and concatenated. The alignment of the ten regions with the highest SV% and PIS% values was carried out using MAFFT v.7.450 (Katoh and Standley, 2013) implemented in Geneious Prime® v.2020.1.2. Phylogenetic inferences were performed by Maximum Likelihood (ML) using the W-IQ-tree (Trifinopoulos et al., 2016), with 1,000 bootstrap repetitions. The choice of substitution models, including FreeRate heterogeneity model, was made according to Bayesian information criterion (BIC; Table 1). Branch support analysis was performed with 1,000 repetitions of bootstrap and single branch test SH (-aLTR, 1,000 replicates). The resulting trees were represented using Geneious Prime® software v.2020.1.2.

Table 1. Substitution models selected for the phylogenetic inferences using Maximum Likelihood (ML).

Region	Models for ML	Reference
Plastome	K3Pu+F+R2	(Kimura, 1981).
trnC-petN	HKY+F	(Hasegawa et al., 1985)
psbC-trnS	JC+I	(Jukes & Cantor, 1969)
psaC-ndhE	F81+F+G4	(Felsenstein, 1981)
ccsA-ndhD	F81+F+I	(Felsenstein, 1981)
petN-psbM	TPM2+F+I	-
accD-psaI	F81+F+I	(Felsenstein, 1981)
trnS-trnG	F81+F+I	(Felsenstein, 1981)
rps15-ycf1	HKY+F+I	(Hasegawa et al., 1985)
ndhF-rpl32	F81+F+I	(Felsenstein, 1981)
rpl16-intron	F81+F+I	(Felsenstein, 1981)
petD-rpoA	F81+F+I	(Felsenstein, 1981)
petA-psbJ	F81+F+I	(Felsenstein, 1981)
trnG-trnfM	F81+F+I	(Felsenstein, 1981)
rps8-rpl14	F81+F+I	(Felsenstein, 1981)

K3Pu: Three substitution types model and equal base freq.; HKY: Unequal transition/transversion rates and unequal base freq.; JC: Equal substitution rates and equal base frequencies; F81: Equal rates but unequal base freq.; TPM2: AC=AT, AG=CT, CG=GT and equal base freq.; +F: Empirical base frequencies. This is the default if the model has unequal base freq.; +I= Proportion of invariable sites; G= shape parameter of the gamma distribution; R: FreeRate model that generalizes the +G model by relaxing the assumption of Gamma-distributed rates.

RESULTS

Bactris gasipaes plastome

The sequencing of plastid-enriched DNA resulted in 448,600 reads with an average length of 214 bp. Of these, 47,735 were plastome reads (~10%), resulting in an average depth of coverage of 67.64 (SD = 24.32). The assembled plastome has a 21 bp gap in the IGS *trnT-UGU*/*trnL-UAA* (position 46,800 to 46,820). This gap is in an AT-rich region (sequence 22 bp upstream to 119 bp downstream is only 7.8% of

GC-content) and is, therefore, difficult to sequence. We calculated this gap length using other species of tribe Cocoseae as reference.

The *Bactris gasipaes* plastome has the quadripartite structure typically found in angiosperms (Rogalski et al., 2015), with a pair of inverted repeat (IRs), a large single-copy region (LSC), and a small single-copy region (SSC). The IRs are 27,038 bp in length (each), the LSC is 85,118 bp, and the SSC is 17,452 bp, resulting in a plastome with 156,646 bp.

Bactris gasipaes plastome has an average GC-content of 37.5%. When comparing the plastome regions, the SSC has the lowest value of GC-content, with 31.3%, followed by LSC with 35.5%. The IRs have the highest value, with 42.6% of GC-content. The rRNA and tRNA show high GC-content, with 55.3% and 53.4%, respectively. Protein-coding genes have an average GC-content of 37.9%, intergenic spacers (IGS) of 37.5%, and introns of 37.1% (Table 2). The plastome GC-content among species from tribe Cocoseae is similar, ranging from 37.40% (*Elaeis guineensis*) to 37.53% (*Acrocomia aculeata*).

Table 2. GC-content of *Bactris gasipaes* plastome.

GC content (%)	***Bactris gasipaes***
Total	37.5
SSC	31.3
LSC	35.5
IRs	42.6
Introns	37.1
IGS	37.5
CDS	37.9
tRNA	53.4
rRNA	55.3

In *Bactris gasipae*s plastome, we annotated 113 unique genes, being 79 protein-coding genes, 30 tRNA genes, and 4 rRNA genes (Table 3). Duplicate genes in IRs include 8 tRNA genes, 4 rRNA genes, and 6 protein-coding genes (Table 3). Among the 113 genes, 15 genes contain one intron (6 tRNA genes and 9 protein-

coding genes) and 3 genes contain two introns (*clpP*, *ycf3*, and *rps12*; Table 3). Among intron-containing genes, 12 are located in LSC (*trnK-UUU*, *rps16*, *trnG-UCC*, *atpF*, *rpoC1*, *ycf3*, *trnL-UAA*, *trnV-UAC*, *clpP*, *petB*, *petD*, *rpl16*), 1 in SSC (*ndhA*), 4 in IRs (*rpl2*, *ndhB*, *trnI-GAU*, *trnA-UGC*), and *rps12* is a trans-splicing gene with the first exon located in the LSC region and the second and third exons in the IRs.

Table 3. List of genes of *Bactris gasipaes* plastome organized according to their location.

Plastome region	Name of genes
Large Single Copy (LSC)	*psbA*, *trnK-UUU*[a], *matK*, *rps16*[a], *trnQ-UUG*, *psbK*, *psbI*, *trnS-GCU*, *trnG-UCC*[a], *trnR-UCU*, *atpA*, *atpF*[a], *atpH*, *atpI*, *rps2*, *rpoC2*, *rpoC1*[a], *rpoB*, *trnC-GCA*, *petN*, *psbM*, *trnD-GUC*, *trnY-GUA*, *trnE-UUC*, *trnT-GGU*, *psbD*, *psbC*, *trnS-UGA*, *psbZ*, *trnG-GCC*, *trnfM-CAU*, *rps14*, *psaB*, *psaA*, *ycf3*[b], *trnS-GGA*, *rps4*, *trnT-UGU*, *trnL-UAA*[a], *trnF-GAA*, *ndhJ*, *ndhK*, *ndhC*, *trnV-UAC*[a], *trnM-CAU*, *atpE*, *atpB*, *rbcL*, *accD*, *psaI*, *ycf4*, *cemA*, *petA*, *psbJ*, *psbL*, *psbF*, *psbE*, *petL*, *petG*, *trnW-CCA*, *trnP-UGG*, *psaJ*, *rpl33*, *rps18*, *rpl20*, *rps12*[b,c] (exon 1), *clpP*[b], *psbB*, *psbT*, *psbN*, *psbH*, *petB*[a], *petD*[a], *rpoA*, *rps11*, *rpl36*, *infA*, *rps8*, *rpl14*, *rpl16*[a], *rps3*, *rpl22*
Inverted Repeat (IR)	*rps19*, *trnH-GUG*, *rpl2*[a], *rpl23*, *trnI-CAU*, *ycf2*, *trnL-CAA*, *ndhB*[a], *rps7*, rps12[c] (exons 2 and 3), *trnV-GAC*, *rrn16*, *trnI-GAU*[a], *trnA-UGC*[a], *rrn23*, *rrn4.5*, *rrn5*, *trnR-ACG*, *trnN-GUU*
IR / SSC junction	*ndhF*, *ycf1*
Small Single Copy (SSC)	*rpl32*, *trnL-UAG*, *ccsA*, *ndhD*, *psaC*, *ndhE*, *ndhG*, *ndhI*, *ndhA*[a], *ndhH*, *rps15*

a: genes contain one intron; *b*: genes contain two intron; *c*: gene trans-splicing.

Surprisingly, the *cemA* gene exhibited an alternative start codon, as reported in species of subtribes Attaleinae and Elaeidinae, contrasting with the Bactridinae species sequenced so far (*Astrocaryum aculeatum*, *Astrocaryum murumuru*, and

Acrocomia aculeata). Three other genes have alternative initiation codons, *rps19* (GTG), *rpl2* (ACG), *ndhD* (ATC).

Comparative analysis in tribe Cocoseae

Plastomes are highly conserved (97.3% identity) within tribe Cocoseae. Species from subtribes Bactridinae and Elaeidinae have a plastome ~2 kb larger than the species from subtribe Attaleinae. This difference in length between the subtribes is mainly in the IRs and LSC regions (Table 4). The plastomes from Cocoseae species ranged from 154.048 bp (*Butia eriospatha*) to 156.937 bp (*Elaeis guineenses*).

Table 4. Plastomes of tribe Cocoseae.

Subtribe	Species	Plastome (bp)	LSC (bp)	IR (bp)	SSC (bp)
Elaeidinae	*Elaeis guineensis*	156,973	85,192	27,071	17,639
Bactridinae	*Astrocaryum aculeatum*	156,804	85,037	27,081	17,605
	Astrocaryum murumuru	156,801	85,017	27,081	17,622
	Bactris gasipaes	156,646	85,118	27,038	17,452
	Acrocomia aculeate	156,500	84,936	27,092	17,380
Attaleinae	*Syagrus coronate*	155,053	84,535	26,522	17,474
	Cocos nucifera	154,731	84,230	26,555	17,391
	Butia eriospatha	154,048	83,805	26,437	17,369

LSC: large single copy region; IR: inverted repeat; SSC: small single copy region.

The progressive alignment among species from tribe Cocoseae shows evidence for three LCBs (Fig. 1). These three LCBs are a result of the 4.5 kb inversion present in the plastomes of *Astrocaryum murumuru* and *Astrocaryum aculeatum* (Fig. 1). The set of genes that comprises this structural rearrangement is composed of *ndhC*, *ndhK*, *ndhJ*, *trnF-GAA*, and *trnL-UAA*.

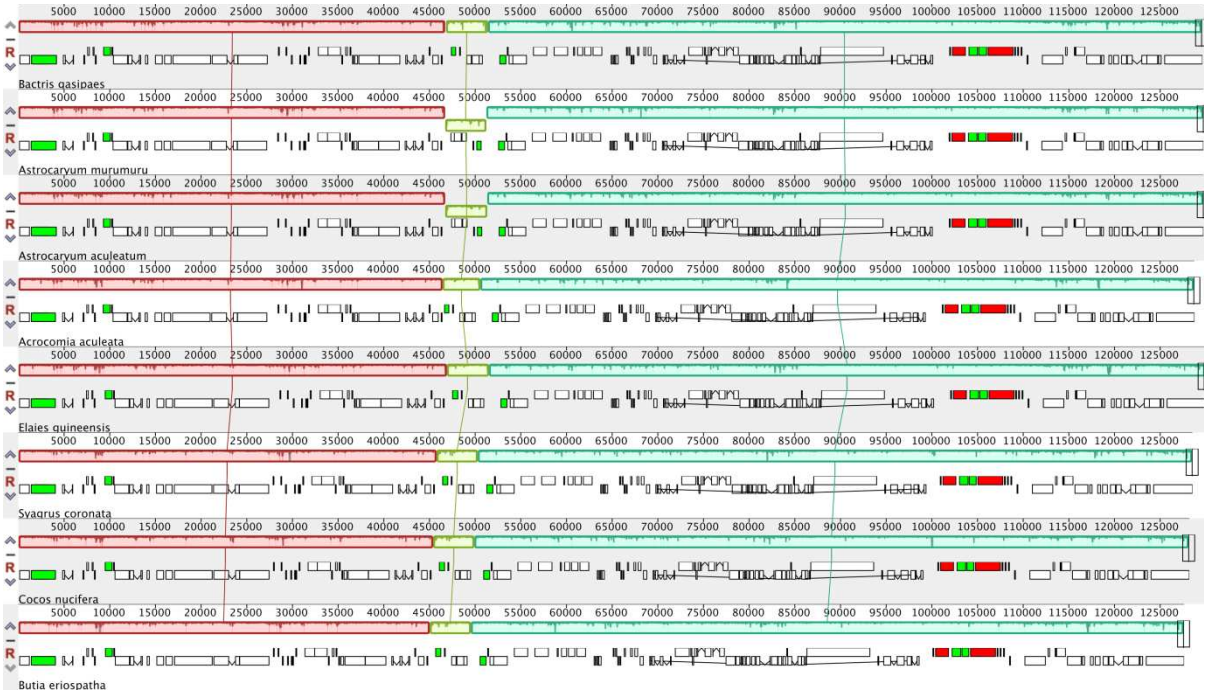


Fig. 1 – Plastome rearrangements analysis within tribe Cocoseae. Locally collinear blocks (LCBs) are indicated by colors. In green it is possible to observe the inversion of 4.5 kb in *Astrocaryum* plastomes.

In the LSC/IR and IR/SSC junctions of the plastomes, it is possible to observe differences between the subtribes (Fig. 2). In Bactridinae (*Acrocomia aculeata*, *Astrocaryum aculeatum*, *Astrocaryum murumuru*, and *Bactris gasipaes*) and Elaeidinae (*Elaeis guineensis*) the *rps19* gene is completely contained in the IR region and, therefore, two copies of the complete gene. In contrast, in the subtribe Attaleinae (*Butia eriospatha*, *Cocos nucifera*, and *Syagrus coronata*) the *rps19* gene is only partially contained in the IRs, resulting in a complete *rps19* and a partial *rps19*: the complete *rps19* gene starts at the IRb and ends at the LSC (LSC/IRb); partial *rps19* starts at IR but does not contain the final portion of the gene (Fig. 2).

Similarly, the *ycf1* gene is partially contained in IRs, with a complete *ycf1*at IRa/SSC and a partial (pseudo) *ycf1* at IRb. The *ndhF* gene has both position and length conserved at the IRb/SSC junction in tribe Cocoseae, with the portion of the gene contained in the IRb overlapping the *ycf1* gene (56 bp) (Fig. 2).

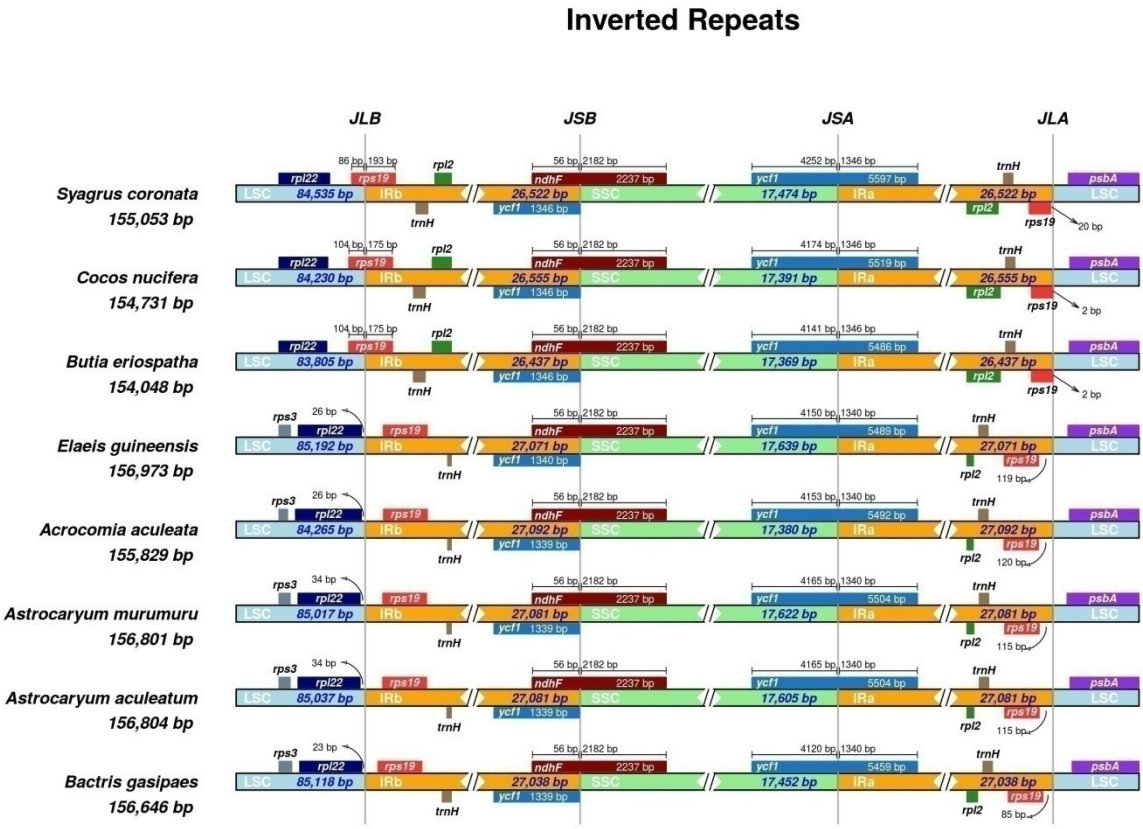


Fig. 2 - Comparison of plastome junctions (IRb/LSC; IRb/SSC, SSC/IRa; IRa/LSC) among Cocoseae species. The numbers indicate sequence length in base pairs.

Hypervariable regions

We carried out the SV% and PIS% estimative to identify the plastome regions with the greatest variation within tribe Cocoseae. All ten regions selected according to the highest SV% showed greater variation than the plastid molecular markers commonly used for phylogenetic analysis in palm trees (Fig. 3a). As expected, they have SV% lower than the nuclear markers PRK and RPB2 (Fig. 3a). Among the ten regions selected according to the highest PIS% values, all showed greater values than the plastid molecular markers commonly used for phylogenetic analysis in palm trees (Fig. 3b) and two of them (*trnC-petN* and *psbC-trnS*) were more variable than the nuclear marker PRK (Fig. 3b). The nuclear marker RPB2 showed the highest values for both SV% and PIS% estimative.

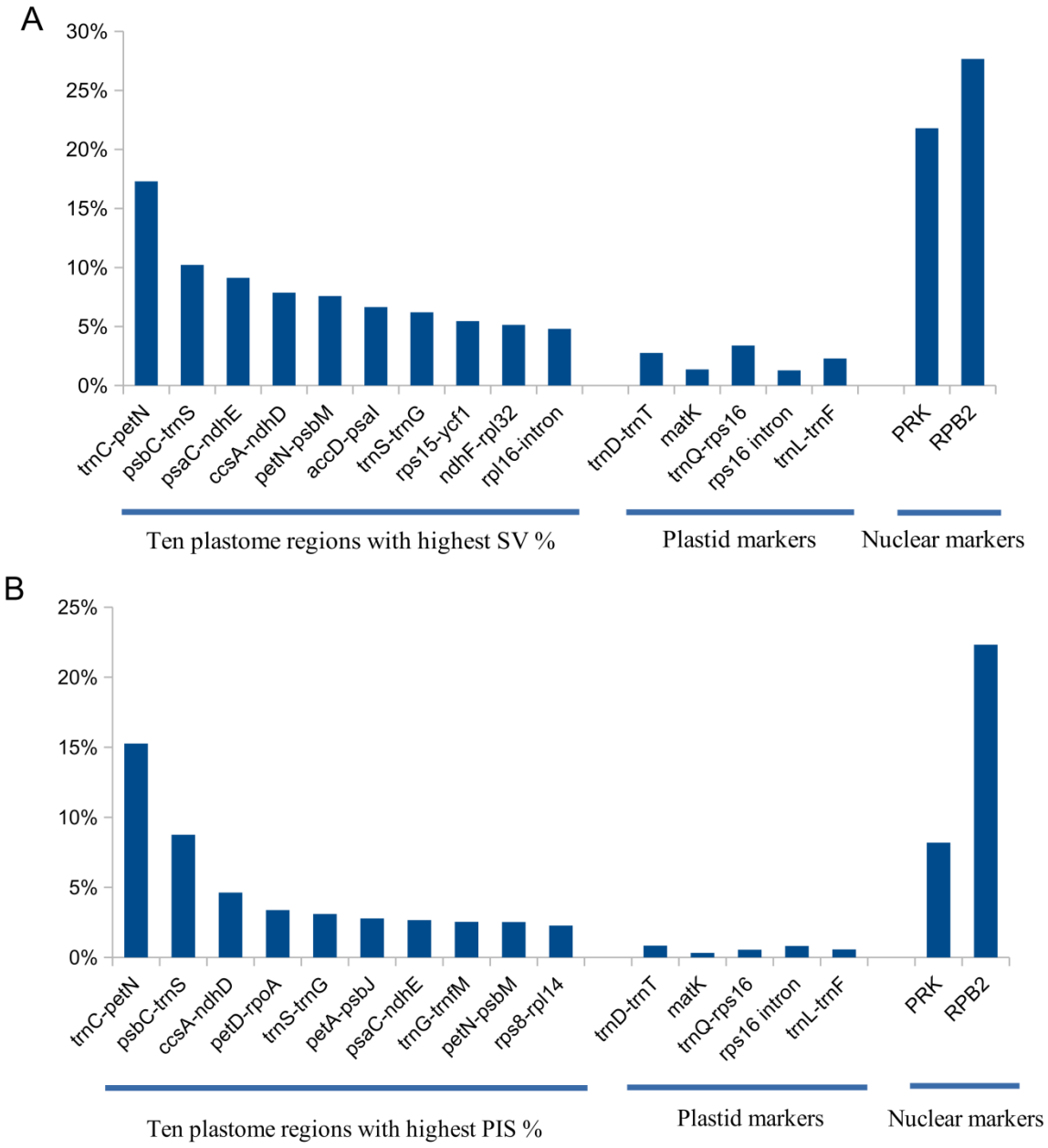


Fig. 3 - Hypervariable plastome regions compared with commonly used plastid and nuclear markers. A - The ten plastome regions with greatest sequence variation (SV%); B - The ten plastome regions with greatest frequency of parsimony informative sites (PIS%).

We also calculated the frequency of substitutions and frequency of InDels events for each plastome region. The substitution frequency was ~5x higher than InDels in plastomes (Table 5). In CDSs, substitutions are ~80x more common than InDels. In IGSs and introns, substitutions are ~4x and ~3x more frequent than InDels, respectively (Table 5). In general, these data show that in all regions of Cocoseae plastomes we have a higher frequency of substitutions than InDels.

Table 5. Frequency of substitutions/mutations and insertions/deletions (InDels) among the plastome sequence.

	Plastome	CDS	IGS	Introns
Mutations (%)	1.53	0.80	2.30	1.12
InDels (%)	0.30	0.01	0.58	0.34
Mutation/InDel ratio	5.02	79.90	3.99	3.26

Phylogenomic analysis

The complete LCB matrix of the aligned plastomes consisted of 137,452 columns, of which 134,244 are constant and 815 parsimony-informative. The SV matrix contained a total of 9614 columns, with 133 parsimony-informative, 9144 constant sites, and the PIS matrix 7538 columns, with 129 parsimony-informative and 7200 constant sites. The phylogenomic trees inferred using Maximum Likelihood (ML) and based on the plastome and the ten selected regions with greatest SV datasets showed identical topology (Fig. 4) and high bootstrap support (> 86). The phylogenetic tree inferred from the alignment of the ten regions with greatest PIS values showed a similar topology, it differed only by the presence of polytomy in the Attaleinae clade, generated by the low bootstrap support in intergeneric relationships (Supplementary Figure 1). For the topologies generated by ML in the three datasets, the monophyly of the subtribes Bactridinae and Attaleinae was confirmed. Also, the subtribe Elaeidinae appears as closest related to Bactridinae than to Attaleinae. In Bactridinae, we can see *Bactris* and *Astrocaryum* as closely related genera, and in Attaleinae, *Cocos nucifera* as sister to *Syagrus coronata*.

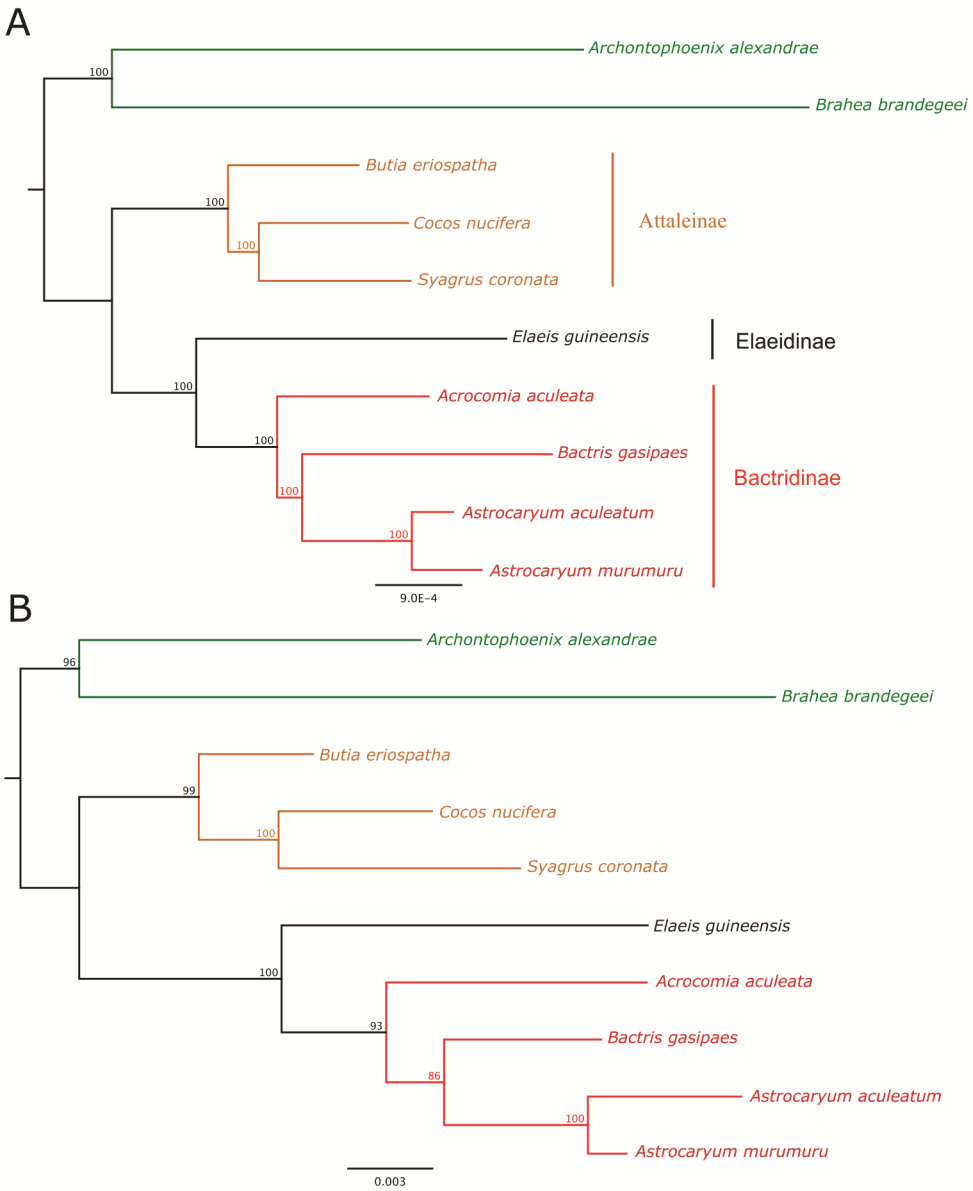


Fig. 4 - Phylogenetic trees based on maximum likelihood inference. A – Phylogenetic inference using plastome sequences (one IR removed); B - Phylogenetic inference using the ten plastome regions with greatest sequence variability (SV%) values. The numbers above the branches are maximum likelihood bootstrap values (1000 replicates).

DISCUSSION

Bactris gasipaes plastome

Here we sequenced the plastome of *Bactris gasipaes*, which has a typical quadripartite structure and gene content similar to those described in other Cocoseae species (Wicke et al., 2011; Uthaipalanwong et al., 2012; Áquila et al., 2018; Lopes et al., 2018; 2019; Magnabosco et al., 2020). The IRs are identical, which probably occur due to mechanisms of replication and repair through recombination-dependent

replication (RDR), as previously reported in plant plastomes (Gonçalves et al., 2020). The GC content in the plastomes of *Bactris gasipaes* and other species of Cocoseae corroborate the mean value described by Kwon et al. (2020) for angiosperms (37.71; SD 1.10). Also, the tendency of a higher GC content in the IRs than in the LSC and SSC was previously reported in bryophytes, ferns, lycophytes, and angiosperms (Li et al., 2016; Kwon et al., 2020).

The gene content is conserved among species of Bactridinae (e.g., *Acrocomia aculeata*, *Astrocaryum murumuru*, *Astrocaryum aculeatum*, and *Bactris gasipaes*) and Elaeidinae (*Elaeis guineensis*) (Uthaipaisanwong et al., 2012; Lopes et al., 2018; 2019). Attaleinae species present one pseudogenized *rps19*, and thus, one fewer CDS (Magnabosco et al., 2020).

The *cemA* gene has an unconventional start codon in *Bactris gasipaes*, what was also previously described in species of subtribes Attaleinae and Elaeidinae, in *Podococcus barteri* (NC_027276.1), *Phoenix dactylifera* 'Khalas', and other monocots. However, it is still not clear if this gene, with the unconventional start codon, is translatable to protein (Yang et al. 2010). Although most of the genes encoding proteins have ATG initiation codons (Wilson 2014; Wu et al., 2018), some alternative initiation codons are not new in plants (Huang et al. 2013), such as the GTG in the *rps19* gene, ACG in *rpl2* and ATC in *ndhD*, which were also reported in *Lilium longiflorum*, *Phoenix dactylifera* 'Khalas' and *Amomum compactum*, respectively (Yang et al., 2010; Kim and Kim, 2013; Wu et al., 2018).

Comparative plastome analysis within Cocoseae

Comparative studies using plastomes of species at different taxonomic levels can bring insights into plastome evolution, phylogenetic relationships, and evolutionary rates substitutions (Young et al., 2011; Rogalski et al., 2015). Plastomes of the three subtribes analyzed (Attaleinae, Bactridinae, and Elaeidinae), represented here by eight species, provided information to compare sequence variations in tribe Cocoseae. We identified slight differences in the plastome size (~2 kb) and an inversion of 4.5 kp that occurs in the *ndh* complex (LSC region) of *Astrocaryum plastome* (relative to any other Cocoseae species in this study). Similarly, Barrett et al. (2016) reported a highly conserved structure of Arecaceae plastomes, describing

only one 1.9 kb inversion located between the *rps16* and *trnG-UUC* genes in *Tahina spectabilis*.

Also, the variability among Attaleinae, Bactridinae, and Elaeidinae in the LSC/IR junctions, mainly regarding the *rps19* gene, was previouly described in *Acrocomia aculeata* (Lopes et al., 2018), *Butia eriosphata* (Magnabosco et al. 2020) and *Phoenix dactylifera* (Yang et al., 2010; Khan et al., 2012). Similarly, the *ndhF* gene overlapping *ycf1* in ~25 bp is commonly observed in palms (Yang et al., 2010; Khan et al., 2012; Lopes et al., 2018; Magnabosco et al., 2020). Despite the differences in the IR junction, the IR structure and gene content is conserved among palms, corroborating the hypothesis that the IR regions offer an isolation mechanism that stabilizes the structure of the genome (Yue et al., 2008).

Hypervariable regions

Plastomes have several non-coding regions, but not all of them have been explored for phylogenetic studies (Shaw et al., 2007; Smidt et al., 2020; Zavala-Páez et al., 2020). Among the ten regions with the greatest SV% identified in our study, only four IGSs (e.g., *accD-psaI*, *ndhF-rpl32*, *trnS-trnG, psaC-ndhE*) and one intron (*rpl16*) were previously used and/orhighlighted in angiosperms studies (Shaw et al., 2007; Clement et al., 2017, Lopes et al., 2018). Among the ten regions with the greatest PIS%, only three IGS (*trnS-trnG*, *petA-psbJ*, and *psaC-ndhE*) were identified in studies carried out by Shaw et al. (2007; 2014) and Lopes et al. (2018). Thus, in our study, we described four new promising regions based on both SV and PIS values (*trnC-petN*, *psbC-trnS, ccsA-ndhD, petN-psbM*), one new region based on SV (*rps15-ycf1*), and three new regions based on PIS (*petD-rpoA*, *trnG-trnfM*, *rps8-rpl14*). As expected, the nuclear genes PRK and RPB2 showed greater variation than most plastidial regions. These nuclear markers produce substantial informational characters, with well-resolved topologies (Baker et al., 2011). The combined use of the plastidial regions described here and the nuclear markers PRK and RPB2 have a great potential for phylogenetic studies in tribe Cocoseae.

Phylogenomic analysis

The phylogenetic inferences confirmed that the ten regions with the greatest SV% values are suitable for phylogenetic inferences and produces phylogenetic trees with well-resolved and the expected topologies. In ML analysis, all datasets tested (plastome, ten SV regions, and ten PIS regions), results in subtribe Bactridinae well-resolved as monophyletic. The monophyly Bactridinae was previouly described by Eiserhardt et al. (2011), as well as the sister relationship between the subtribes Elaeidinae and Bactridinae (Baker et al., 2011, Merrow et al., 2015; Lopes et al., 2018). Our results are in contrastwith those verified by Gunn (2004), in which the sister relationship between *Astrocaryum* and *Bactris* are weakly supported. In all of our datasets this sister relationship is well-supported. In addition, the monophyly of subtribe Cocoseae was also verified in plastid DNA analysis (Hahn 2002a), in the supertree method (Baker et al. 2009) and by the combined analysis with the PRK and RPB2 genes (Baker et al., 2011). The sister relationship between *Cocos nucifera* and *Syagrus coronata* was also previously described (Merrow et al., 2015), corroborating our results. Thus, both plastome and the ten regions with greatest SV values were able to produce well-resolved phylogenetic trees and with consistent topologies within tribe Cocoseae.

ACKNOWLEDGMENTS

We thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for grants 435200/2018-6. The Brazilian Program of National Institutes of Science and Technology-INCT/Brazilian Research Council-CNPq/MCT and Embrapa (Portfólio Recursos Genéticos). Jeison W. de S. Magnabosco and Michelle E. Zavala-Páez is thanked for her help in the analyses.

Online Resource 1

Table S1. GenBank accession numbers of the nucleotide sequences used in our analysis.

Subtribe	Species	Genbank Accession		
		PRK	***RPB2***	**Plastome**
Elaeidinae	*Elaeis guineensis*	AY601219	HQ265661	NC_017602
Bactridinae	*Astrocaryum aculeatum*	JQ821944	JQ821977	MH537788
	Astrocaryum murumuru	HQ265590	HQ265637	MH537787
	Bactris gasipaes	KP218842	HQ265650	MW054718
	Acrocomia aculeata	HQ265574	HQ265620	NC_037084
Attaleinae	*Cocos nucifera*	HQ265608	EF491150	NC_022417
	Butia eriospatha	-	-	MN329806
	Butia capitata	AY601252	EF491157	-
	Syagrus coronata	-	-	NC_029241
	Syagrus smithii	AY601263	HQ265666	-

Online Resource 2

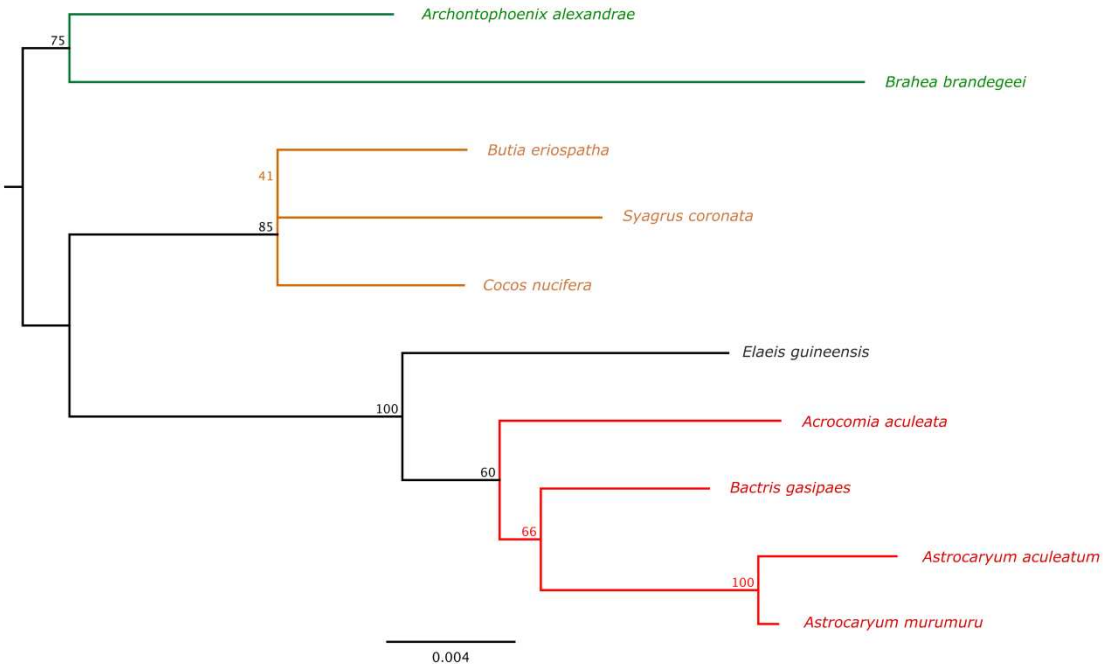


Fig. S1 – Phylogenetic tree based on maximum likelihood inference using the ten plastome regions with greatest PIS% values. The numbers above the branches are maximum likelihood bootstrap values (1000 replicates).

3. CONSIDERAÇÕES FINAIS

Neste trabalho foi sequenciado o primeiro genoma plastidial completo *Bactris gasipaes* e foi verificado que este possui uma estrutura quadripartida típica e conteúdo gênico semelhante aos descritos em outras espécies de Cocoseae. Foram descritas novas regiões promissoras, quatro com base nos valores de SV e PIS (*trnC-petN, psbC-trnS, ccsA-ndhD, petN-psbM*), uma baseada em SV (*rps15-ycf1*) e três baseadas em PIS (*petD-rpoA*, *trnG-trnfM*, *rps18-rpl14*). Verificou-se também que, tanto o plastoma de *Bactris gasipaes* quanto as dez regiões com maiores valores de SV foram capazes de produzir árvores filogenéticas bem resolvidas e com topologias consistentes dentro da tribo Cocoseae.

Com a caracterização destas novas regiões, será possível realizar novos estudos comparativos com outras espécies de palmeiras e até mesmo entre *Bactris gasipaes*, já que esta espécie possui grande variação, pois apresenta diversas raças crioulas com características morfológicas diferentes, devido ao seu processo de domesticação e diversificação.

REFERÊNCIAS

Amiryousefi A, Hyvonen J and Poczai P (2018) IRscope: An online program to visualize the junction sites of chloroplast genomes. Bioinformatics, 34:3030–3031.

Áquila R, Silva J, Silva JV and Almeida C (2018) The first complete chloroplast genome sequence of *Syagrus coronata*: comparative analysis in Arecoideae (Arecaceae). Revista Brasileira de Botânica, 41:393–401.

Asmussen CB, Dransfield J, Deickmann V, Barfod AS, Pintaud JC and Baker WJ (2006) A new subfamily classification of the palm family (Arecaceae): evidence from plastid DNA phylogeny. Botanical Journal of the Linnean Society, 151:15–38.

Baker WJ and Dransfield J (2016) Beyond Genera Palmarum: progress and prospects in palm systematic. Botanical Journal of the Linnean Society, 2016, 182:207–233.

Baker WJ, Norup MV, Clarkson JJ, Couvreur TLP, Dowe JL, Lewis CE, Pintaud JC, Savolainen V, Wilmot T and Chase MW (2011) Phylogenetic relationships among arecoid palms (Arecaceae: Arecoideae). Annals of Botany, 108:1417–1432.

Baker WJ, Savolainen V, Asmussen-Lange CB, Chase MW, Dransfield J, Forest F, Harley MM, Uhl NW and Wilkinson M (2009) Complete generic-level phylogenetic analyses of palms (Arecaceae) with comparisons of supertree and supermatrix approaches. Systematic Biology, 58:240–256.

Balslev H, Bernal R and Fay MF (2016) Palms – emblems of tropical forests. Botanical Journal of the Linnean Society, 182:195–200.

Barfod AS, Hagen M and Borchsenius F (2011) Twenty-five years of progress in understanding pollination mechanisms in palms (Arecaceae). Annals of Botany 108: 1503–1516.

Barrett CF, Baker WJ, Comer JR, Conran JG, Lahmeyer SC, Leebens-Mack JH, Li J, Lim GS, Mayfield-Jones DR, Perez L et al (2016) Plastid genomes reveal support for deep phylogenetic relationships and extensive rate variation among palms and other commelinid monocots. New Phytol, 209:855–870.

Clement CR (1992) Domesticated palms. Principes, 36:70–78.

Clement CR, Cristo-Araújo M, Coppens d'Eeckenbrugge G, Reis VM, Lehnebach R and Picanço-Rodrigues D (2017) Origin and Dispersal of Domesticated Peach Palm. Frontiers in Ecology and Evolution, 5:1–19.

Cristo-Araújo MR, Doriane P, Astolfi-Filho S and Clement CR (2015) Peach palm core collection in Brazilian Amazonia. Crop Breeding and Applied Biotechnology, 15:18–25.

Darling AC, Mau B, Blattner FR and Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome research, 14:1394–1403.

Eiserhardt WL, Pintaud JC, Asmussen-Lange C, Hahn WJ, Bernal R, Balslev H and Borchsenius F (2011) Phylogeny and divergence times of Bactridinae (Arecaceae, Palmae) based on plastid and nuclear DNA sequences. Taxon, 60:485–498.

Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution, 17:368–376.

Gonçalves DJP, Jansen RK, Ruhlman TA, Mandel JR (2020) Under the rug: Abandoning persistent misconceptions that obfuscate organelle evolution. Molecular Phylogenetics and Evolution, 151:106903.

Gunn B (2004) The phylogeny of the Cocoeae (Arecaceae) with emphasis on *Cocos nucifera*. Annals Missouri Botanical Garden, 91:505–522.

Hahn WJ (2002a) A phylogenetic analysis of the arecoid line of palms based on plastid DNA sequence data. Molecular Phylogenetics and Evolution, 23:189–204.

Hasegawa M, Kishino H, Yano T-A (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution, 22:160–174.

Huang YY, Matzke AJM and Matzke M (2013) Complete Sequence and Comparative Analysis of the Chloroplast Genome of Coconut Palm (*Cocos nucifera*). PLoS ONE, 8:74736.

Jukes TH & Cantor CR (1969) CHAPTER 24-Evolution of Protein Molecules. Mammalian Protein Metabolism, 3:21–132.

Katoh K and Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution, 30:772–780.

Kim JS and Kim JH (2013) Comparative genome analysis and phylogenetic relationship of order Liliales insight from the complete plastid genome sequences of two Lilies (*Lilium longiflorum* and *Alstroemeria aurea*). PLoS ONE, 8:1–11.

Kimura M (1981) Estimation of evolutionary distances between homologous nucleotide sequences. Proceedings of the National Academy of Sciences, 78:454–458.

Khan A, Khan IA, Heinze B and Azim MK (2012) The Chloroplast Genome Sequence of Date Palm (*Phoenix dactylifera* L. cv. ‘Aseel’). Plant Molecular Biology Rep, 30:666–678.

Kwon EC, Kim JH and Kim NS (2020) Comprehensive genomic analyses with 115 plastomes from algae to seed plants: structure, gene contents, GC contents, and introns. Genes & Genomics, 42:553–570.

Li FW, Kuo LY, Pryer KM and Rothfels CJ (2016) Genes translocated into the plastid inverted repeat show decelerated substitution rates and elevated GC content. Genome Biology Evolution, 8:2452–2458.

Levis C, Flores BM, Moreira PA, Luize BG, Alves RP, Franco-Moraes J and Costa FR (2018) How people domesticated Amazonian forests. Frontiers in Ecology and Evolution, 5:171.

Lopes AS, Pacheco TG, Silva ON, Cruz LM, Balsanelli E, Souza EM, Pedrosa FO and Rogalski M (2019) 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.

Lopes AS, Pacheco TG, Nimz T, Vieira LN, Guerra MP, Nodari RO, Souza EM, Pedrosa FO and Rogalski (2018) 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.

Magnabosco JWS, Fraga HPF, Silva RS, Rogalski M, Souza EM, Guerra MP and Vieira, LN (2020) Characterization of the complete chloroplast genome of *Butia eriospatha* (Arecaceae). Genetics and Molecular Biology, 43: e20200023.

Meerow AW, Noblick L, Dayana ES-L, Sanchez V, Javier F-O, Jestrow B, Nakamura K (2015) Phylogeny and historical biogeography of the cocosoid palms (Arecaceae, Arecoideae, Cocoseae) inferred from sequences of six WRKY gene family loci. Cladistics, 31:509–534.

Montoya SG, Motoike SY, Kuki KN and Couto AD (2016). Fruit development, growth, and stored reserves in macauba palm (*Acrocomia aculeata*), an alternative bioenergy crop. Planta, 244:927–938.

Muñoz G, Trøjelsgaard K and Kissling WD (2019) A synthesis of animal-mediated seed dispersal of palms reveals distinct biogeographical differences in species interactions. Journal of Biogeography, 46:466–484.

Muñoz-Pérez JM, Cañas GP, López L and Arias T (2019). Genome-wide diversity of northern South America cultivated Coconut (*Cocus nucifera* L.) uncovers diversification times and targets of domestication of coconut globally. Bio Rxiv, 825398.

Onstein RE, Baker WJ, Couvreur TLP, Faurby S, Svenning JC and Kissling WD (2017) Frugivory-related traits promote speciation of tropical palms. Nature Ecology & Evolution, 12:1903–1911.

Pereira Cruz A, Giehl ELH, Levis C, Machado JS, Bueno L, and Peroni N (2020). Pre-colonial Amerindian legacies in forest composition of southern Brazil. Plos one, 15:e0235819.

Rivas M and Condón F (2015) Plant domestication and utilization: The case of the Pampa biome. In Advances in plant breeding strategies: Breeding, biotechnology and molecular tools, 3:24.

Rogalski M, Vieira LN, Fraga HP and Guerra MP (2015) Plastid Genomics in Horticultural Species: Importance and Applications for Plant Population Genetics, Evolution and Biotechnology. Frontiers in Plant Science, 6:1–17.

Rozas J, Ferrer-Mata A, Sánchez-Delbarrio JC, Guirão-Rico S, Librado P, Ramos-Onsins S and Sánchez-Gracia A (2017) DnaSP 6: DNA Sequence Polymorphism Analysis of Large Datasets. Molecular Biology and Evolution, 34**:**3299–3302.

Ruhlman TA and Jansen RK (2014) The plastid genomes of flowering plants in Chloroplast Biotechnology: Methods and Protocols, ed. M. Pal Totowa, NJ:Humana Press, 1132:3–38.

Sakaguchi S, Ueno S, Tsumura Y, Setoguchi H, Ito M, Hattori C, Nozoe S, Takahashi D, Nakamasu R, Sakagami T et al (2017) Application of a simplified method of chloroplast enrichment to small amounts of tissue for chloroplast genome sequencing. Applications in Plant Sciences, 5:1-8.

Shaw J, Lickey EB, Schuilling EE and Small RL (2007) Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. American Journal of Botany, 94:275–288.

Shaw J, Shafer HL, Rayne LO, Kovach MJ, Schorr M and Morris AB (2014) Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: The tortoise and the hare IV. American Journal of Botany, 101:1987–2004.

Shi C, Hu N, Huang H, Gao J, Zhao YJ and Gao LZ (2012) An improved chloroplast DNA extraction procedure for whole plastid genome sequencing. PLoS ONE, 7:1–7.

Smidt EdC, Páez MZ, Vieira LdN, Viruel J, de Baura VA, Balsanelli E, et al. (2020) Characterization of sequence variability hotspots in Cranichideae plastomes (Orchidaceae, Orchidoideae). PLoS ONE, 15:e0227991.

Trifinopoulos J, Nguyen LT, von Haeseler A and Minh BQ (2016) W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Research, 44:W232-W235.

Uthaipaisanwong P, Chanprasert J, Shearman JR, Sangsrakru D, Yoo-cha T, Jomchai N, Jantasuriyarat C, Tragoonrung S and Tangphat- sornruang S (2012) Characterization of the chloroplast genome sequence of oil palm (*Elaeis guineensis* Jacq.). Gene, 500:172–180.

Vieira LDN, Fraga HPF, Rogalski M, Souza EM, Pedrosa FO and Guerra MP (2014) An improved protocol for intact chloroplasts and cpDNA isolation in conifers. PLoS ONE, 9:1–8.

Wu M-L, Li Q, Xu J, Li X-W (2018) Complete chloroplast genome of the medicinal plant *Amomum compactum*: gene organization, comparative analysis, and phylogenetic relationships within Zingiberales. Chinese Medicine, 13:10.

Wicke S, Schneeweiss GM, Depamphilis CW, Muller KF and Quandat D (2011) The evolution of the plastid chromosome in land plants: Gene content, gene order, gene function. Plant Molecular Biology, 76:273–297.

Wilson MA, Gaut B and Clegg MT (1990) Chloroplast DNA evolves slowly in the palm family (Arecaceae). Molecular biology and evolution, 7:303-314.

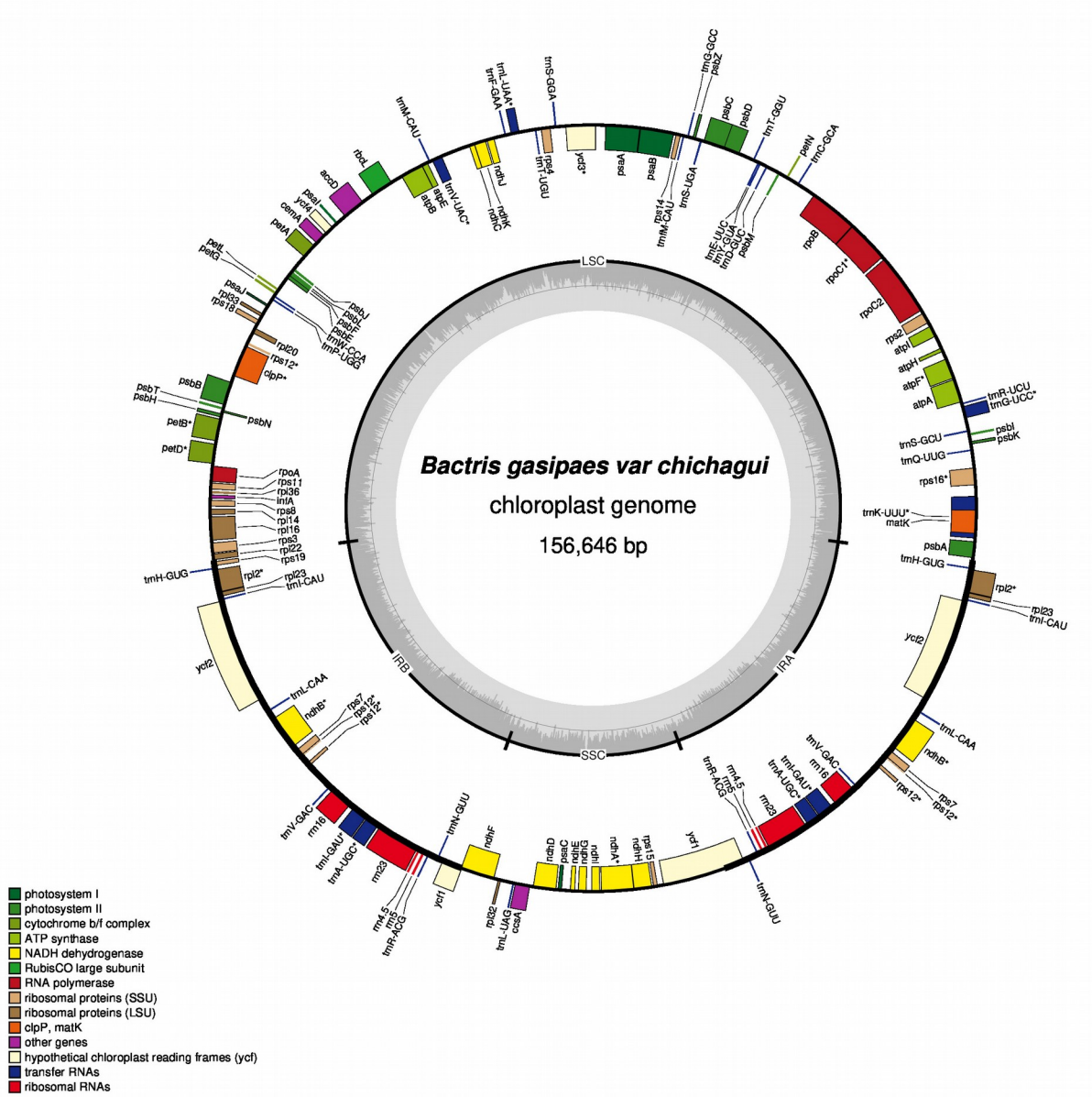
Yang M, Zhang X, Liu G, Yin Y, Chen K, Yun Q, Zhao D, Al-Mssallem IS and Yu J (2010) The complete chloroplast genome sequence of date palm (*Phoenix dactylifera* L.). PLoS One, 5:e12762.

Young HA, Lanzatella CL, Sarah G, Tobias CM (2011) Chloroplast Genome Variation in Upland and Lowland Switchgrass. PloS ONE, 6:e23980.

Yue F, Cui L, dePamphilis CW, Moret BME and Tang J (2008) Gene rearrangement analysis and ancestral order inference from chloroplast genomes with inverted repeat. BMC Genomics, 9:S25.

Zavala-Páez M, Vieira LN, Baura VA, Balsanelli E, Souza EM, Cerna M, Chase MW and Smidt EC (2020) Comparative plastid genomics of Neotropical Bulbophyllum (Orchidaceae; Epidendroideae). Frontiers in Plant Science, 11:799.

ANEXO 1 – DESENHO ESTRUTURAL DO GENOMA DE *Bactris gasipaes* var. *chichagui* Tipo 1



Os genes desenhados no interior do círculo são transcritos no sentido horário e os genes do exterior são transcritos no sentido anti-horário. Os genes pertencentes a diferentes grupos funcionais são codificados por cores. O cinza mais escuro no círculo interno corresponde ao conteúdo GC e o cinza mais claro ao conteúdo AT.