AMANDA DE SANTANA LOPES

EVOLUTIONARY AND GENETIC ASPECTS OF THE PLASTOME OF OLEAGINOUS SPECIES

Thesis presented to the Universidade Federal de Viçosa as part of the requirement of the Plant Physiology Graduate Program for obtention of the degree of Doctor Scientiae.

VIÇOSA
MINAS GERAIS – BRASIL
2018

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T

L864e
2018

Lopes, Amanda de Santana, 1990-
Evolutionary and genetic aspects of the plastome of oleaginous species. / Amanda de Santana Lopes. - Viçosa, MG, 2018.
vii, 165f. : il. (algumas color.) ; 29 cm.

Orientador: Marcelo Rogalski.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Plantas oleaginosas. 2. Genômica. 3. Evolução. I. Universidade Federal de Viçosa. Departamento de Biologia Vegetal. Programa de Pós-graduação em Fisiologia Vegetal. II. Título.

CDD 22 ed. 633.85

AMANDA DE SANTANA LOPES

EVOLUTIONARY AND GENETIC ASPECTS OF THE PLASTOME OF OLEAGINOUS SPECIES

Thesis presented to the Universidade Federal de Viçosa as part of the requirement of the Plant Physiology Graduate Program for obtention of the degree of *Doctor Scientiae*.

APPROVED: March 6th, 2018.

Camilo Elber Vital

Dimas Mendes Ribeiro

Marcelo Francisco Pompelli

Samuel Cordeiro Vitor Martins

Marcelo Rogalski
(Adviser)

“Remember the Lord in everything you do,

and he will show you the right way.”

Proverbs 3:6 (Bible, GNT version)

ACKNOWLEDGMENTS

Above all, I am very grateful to God for his everlasting love and grace, and to his Son, Jesus Christ, my Blessed Redeemer. My life in your protective hands is “like a tree that grows beside a stream... (psalm 1:3)”

I am very grateful to my family, my parents Edilza and Adilson who provided me home, care, and good education, and my beloved sister Dielle for all confidence and affection. I also thank my brother-in-law Tiago and the little Nina for all “confusions” rs! My sincere thanks also go to my dear longtime friends Débora and Léa for all attention and support.

I am very grateful to my adviser, Professor Marcelo Rogalski, for almost six years of dedicated mentorship and true friendship; he is a special part of my academic trajectory and I will always be grateful to him.

I thank my friends and colleagues from the Plant Physiology Graduate Program, especially from the Plant Molecular Physiology Lab for all fellowship, friendship, and coffee! Special thanks go to Túlio, my research group colleague and great friend, who spent many hours helping me to install programs and, patiently, teaching me how to use several bioinformatic tools. I also thank my dear research group colleagues Odyone, Gélia, and Gleyson for all support in the conduction of the experiments.

I am very grateful to Professors Miguel Pedro Guerra (UFSC), Rubens Onofre Nodari (UFSC), Leila do Nascimento Vieira (UFPR), Emanuel Maltempi de Souza (UFPR), and Fábio de Oliveira Pedrosa (UFPR), for supporting the conduction of the study presented here.

I thank the Professors from the Plant Physiology Graduate Program and from the Biochemistry Undergraduate Program at UFV for share knowledge, provide rich instructions, and, thus, encourage me to build my own path in the science. I also thank Dr. Camilo Elber Vital, Prof. Dimas Mendes Ribeiro, Prof. Marcelo Francisco Pompelli, and Prof. Samuel Cordeiro Vitor Martins who gently accepted to participate in my PhD committee.

To everyone, thank you very much!

TABLE OF CONTENTS

ABSTRACT

LOPES, Amanda de Santana, D.Sc., Universidade Federal de Viçosa, March, 2018. **Evolutionary and genetic aspects of the plastome of oleaginous species.** Adviser: Marcelo Rogalski.

The plastid genome (plastome) is usually a circular molecule of about 150 kb and bears about 120 genes involved mainly in photosynthesis and gene expression. Its evolutionary trajectory includes structural rearrangements, gene degenerations, gene transfer to nucleus, positive selection, and events of gain and loss of RNA editing sites. These features are subject of interest in several areas of plant science such as phylogeny, evolution, basic research, and biotechnology. Additionally, plastomes bear a set of conserved genes useful for phylogenetic studies of deep relationships, so phylogenomic approaches based on whole plastomes have been also applied to resolve intrageneric relationships. Fast-evolving regions of plastid DNA such as intergenic spacers and introns, are source of molecular markers used routinely in phylogeographic and genetic studies. Finally, plastomes are a promising platform for biotechnological applications via plastid transformation. Over the last few years with the improvement of sequencing technology, many plastomes have been sequenced. Nevertheless, plastomes of several species belonging to different families remain unknown. Therefore, the purpose of this study was the sequencing of five plastomes of oleaginous species useful for different industry demands as follows: Linum ustitatissimum (Linaceae), Crambe abyssinica (Brassicaceae), Acrocomia aculeata, Astrocaryum murumuru, and A. aculeatum (Arecaceae). Here, it is presented a detailed characterization regarding molecular markers, phylogenetic inferences, and evolution. Molecular evolution analyses of protein-coding genes in Arecaceae show that highly divergent genes seem to evolve in a species-specific manner and more than half of the genes bear signatures of positive selection. Additionally, within Arecaceae and Linaceae were identified unique plastome rearrangements. It was also found events of gain and loss of RNA editing sites in all sequenced species, indicating a relatively high evolutionary rate of the RNA editing machinery. Hundreds molecular markers were identified in the plastomes sequenced here, providing information useful for several genetic studies aiming to stablish strategies for genetic breeding, domestication, and conservation of natural resources. Moreover, the phylogenies presented in this study, based on concatenated plastid genes or whole plastomes resulted in well-supported and well-resolved trees. Finally, this study sheds light on the evolutionary trajectory of plastomes, which give rise to questions about the relationships between environment adaptation and plastid gene expression.

RESUMO

LOPES, Amanda de Santana, D.Sc., Universidade Federal de Viçosa, março de 2018. **Aspectos evolutivos e genéticos do plastoma de espécies oleaginosas.** Orientador: Marcelo Rogalski.

O genoma plastidial (plastoma), em geral, é uma molécula circular apresentando cerca de 150 kb e 120 genes, os quais atuam, principalmente, na fotossíntese e na expressão gênica. A trajetória evolutiva dos plastomas abrange rearranjos estruturais, degeneração de genes, transferência de genes para o núcleo, seleção positiva e ganhos e de perdas de edições de RNA. Tais eventos são do interesse de diversas áreas da ciência vegetal, como filogenia, evolução, pesquisa básica e biotecnologia. Somado a isto, plastomas têm um conjunto de genes conservados útil na resolução de filogenias entre taxa distantes, sendo que o uso de plastomas inteiros em abordagens filogenômicas tem sido aplicado na resolução de relações intragenéricas. Em adição, sequências plastidiais de rápida evolução, como espaços intergênicos e introns, são fontes de marcadores moleculares aplicados em filogeografia e estudos genéticos. Por fim, plastomas são promissoras plataformas para aplicações biotecnológicas por meio da transformação plastidial. Muitos plastomas têm sido sequenciados com o avanço da tecnologia de sequenciamento nos últimos anos. Porém, plastomas de diversas espécies permanecem desconhecidos. Em vista disso, este estudo propôs-se a fornecer as sequências dos plastomas de cinco espécies oleaginosas utilizadas em diferentes demandas industriais: Linum ustitatissimum (Linaceae), Crambe abyssinica (Brassicaceae), Acrocomia aculeata, Astrocaryum murumuru, and A. aculeatum (Arecaceae). Neste estudo, uma detalhada caracterização destes plastomas relativo a marcadores moleculares, filogenia e evolução foi realizada. Análises de evolução molecular de genes plastidiais em Arecaceae mostram que genes altamente divergentes parecem ser traços evolutivos distintos de determinadas espécies, e que mais da metade dos genes carregam assinaturas de seleção positiva. Ademais, rearranjos únicos foram identificados em plastomas de Arecaceae e Linaceae. Ganhos e perdas de edições de RNA foram encontrados em todas as espécies sequenciadas, sugerindo uma relativamente alta taxa evolutiva deste processo nos plastídios. Adicionalmente, centenas de marcadores moleculares plastidiais foram mapeados, constituindo informação útil para estudos genéticos visando estabelecer estratégias de melhoramento genético, domesticação e conservação dos recursos naturais. Também são apresentadas filogenias com alta resolução e suporte, baseadas em genes concatenados e plastomas inteiros. Em síntese, este estudo lança luz acerca da trajetória evolutiva dos plastomas e suscita discussões sobre a relação entre adaptação ao ambiente e expressão de genes plastidiais.

Plastids are essential organelles for plant cell viability. They can develop from proplastids located in meristematic cells to different types, such as chloroplasts, chromoplasts, amyloplasts, and elaioplasts, which are found in different specialized cells (Pyke 2007). Taken together, the different types of plastids harbor a huge number of metabolic reactions and specific functions, including photosynthesis, and biosynthesis of several compounds such as lipids, pigments, starch, vitamins and amino acids, which are specifically regulated during growth and development (Tetlow et al. 2004; Rogalski e Carrer 2011; Galili et al. 2014; Rogalski et al. 2015). Plastids contain their own genome (plastome), which, has usually a conserved structure, gene content, and gene order. Typically, plastomes of land plants are circular molecules ranging from 120 to 220 kb with a quadripartite structure [two single copy regions (SCs) separated by two inverted repeats (IRs)] and contain about 100-130 genes (Wicke et al. 2011; Tonti-Filippini et al. 2017). Based on endosymbiotic theory, the plastids were originated from an engulfment of a cyanobacterium by a heterotrophic protist. After the endosymbiosis several evolutionary changes occurred, including genome streamlining and massive gene transfer of the cyanobacterial endosymbiont to the host nucleus (Bock 2015; Bock, 2017). Nowadays, plastomes have a reduced number of genes in comparison with the genome of extant cyanobacteria (Kaneko et al. 1996; Bock 2015; Rogalski et al. 2015). The remaining plastid genes are involved in photosynthesis (subunits of the Photosystem I, Photosystem II, Cytochrome b_6f, ATP synthase, and NDH complex; large subunit of Rubisco; and a gene required to C-type cytochrome synthesis), gene expression machinery (subunits of RNA polymerase, ribosomal RNA, ribosomal proteins, tRNAs and a maturase), fatty acids biosynthesis (subunit of acetyl-CoA carboxylase), protein degradation (subunit of the protease Clp), and import of proteins (subunit of the TIC complex) (Bock 2007; Rogalski et al. 2015).

The plastome gathers features of interest in several areas of research, including functional genetics, population genetics, biotechnology, plant evolution, and phylogeny. Plastids are typically uniparental inheritance and the nonrecombinant nature of plastids have been widely explored in studies regarding to genetic structure of natural populations and in phylogeography (Wheeler et al. 2014; Rogalski et al. 2015). Fast-evolving plastid sequences, as intergenic spacers and introns, are source of molecular markers (e.g. SSRs and SNPs) that have been used to genetic population analyses (Roy et al. 2016), germplasm collection characterization (Wambulwa et al. 2016), inference of intrageneric

relationships (Li et al. 2014) and ancestry studies (Provan et al. 2001; Ebert and Peakall 2009; Wheeler et al. 2014). On the other hand, the notorious conservative nature of the plastid genes has been explored in phylogenetic studies, specially to resolve deep relationships, using a large data set composed by concatenated genes or even whole plastomes (Xi et al. 2012; Barret et al. 2016; Wei et al. 2017). Phylogenies based on whole plastome also have been applied to understand the evolutionary relationships between very close taxa. For example, relationships between cultivated rice and related wild species have been analyzed using whole plastomes as basic information to trace strategies aiming crop improvement and conservation of natural resources (Wambugo et al. 2015).

Additionally, the occurrence of plastome rearrangements are powerful phylogenetic markers due to their low level of homoplasy (Kim et al. 2005; Cosner et al. 2004; Martin et al. 2014). The rarity of occurrence of these events make them lineage-specific, providing synapomorphies useful to delimit related taxa (Martin et al. 2014; Weng et al. 2014). For example, members of the families Geraniaceae, Campanulaceae and Fabaceae bear unique and highly rearranged plastomes (Haberle at al. 2008; Guisinger et al. 2011; Cai et al. 2008). The rearrangements reported includes expansion and contraction of the IRs, IR loss, and inversion and deletions of large segments, which can result in deletion of genes, variation in the gene order, and disruption of conserved operons (Harbele et al. 2008; Guisinger et al. 2011; Wicke et al. 2011; Vieira et al. 2016; Zhu et al. 2016). Furthermore, plastome rearrangements are also a feature of interest for basic research, given that they can bring evidences to explain why and how plastomes of some lineages evolve in a different way from the common structure. The mechanism involved in inversions and/or deletions of large and short stretches of DNA segments have been explained as the result of homologous recombination between inverted and directed repeats, respectively (Milligan et al. 1989; Svab and Maliga, 1993; Rogalski et al. 2006; Rogalski et al. 2008a e b; Vieira et al. 2016). Indeed, several small dispersed repeats have been identified in highly rearranged plastomes (Harbele et al. 2008; Guisinger et al. 2011; Weng et al. 2014). High rates of nucleotide substitution are a common feature present in highly rearranged plastomes, which could be explained by a dysfunction during DNA replication, recombination and repair (DNA-RRR) (Guisinger et al. 2011; Weng et al. 2014; Zhang et al. 2016).

The plastome evolution also includes other modifications at gene level, as degeneration, transfer to nucleus, loss of introns, duplication, alternative translation initiation site, transcriptional slippage, positive selection and RNA editing (Daniell et al. 2008, 2016; Barthet et al. 2015; Lin et al. 2015a, 2015b; Xu et al. 2015; Piot et al. 2017).

However, little is understood about the role of such plastome molecular changes in the plant physiology and adaptation to different environmental cues. For example, degeneration or loss of the ndh genes (encoding the NDH complex) in many photosynthetic lineages raise questions about the adaptation required by these lineages to cope with stress conditions (Ruhlman et al. 2015; Lin et al. 2015b), since the NDH complex is supposed to act mainly in the cyclic electron transfer under different stress conditions (Horváth et al. 2000; Li et al. 2004; Yamori and Shikanai 2016). Similarly, the ycf1 gene is a highly divergent plastid gene and it was lost or is a pseudogene in several lineages (Vries et al. 2015). However, targeted disruption by reverse genetics in tobacco plastids showed that it is essential for cell viability (Drescher et al. 2000). The ycf1 gene encodes a subunit of the TIC complex involved in protein import from cytosol to plastid (Kikuchi et al. 2013) and has been hypothesized that its absence in some lineages co-evolved with modifications in the plastid protein import machinery (Nakai 2015). In some lineages ycf1 and ndh genes are absent or pseudogenes in the plastome, however the presence of functional copies of them in the nucleus were not evidenced to date. On other hand, gene transfer to nucleus and subsequent acquisition of active transcription, capture of plastid transit peptide and import of the protein into the plastids have been reported for other plastid genes such as infA, rpl32, rpl22, and rps16 in different plant lineages (Gantt et al. 1991; Millen et al. 2001; Ueda et al. 2007, 2008; Jansen et al. 2011; Keller et al. 2017). In other infrequent cases, plastid genes were replaced by a eukaryotic homologous protein imported into the plastids (Bubunenko et al. 1994; Konishi et al. 1996).

Notwithstanding the occurrence of plastid gene degenerations and losses, the evolutionary trajectory of plastomes also includes genes that bear signatures of positive selection. In grasses, across the PACMAD clade, Piot et al. (2017) reported that 25 plastid genes evolved under positive selection, including a strong positive selection during the C3-C4 photosynthetic transition in the rbcL gene (codify the large subunit of Rubisco). Positive selection of mutations at particular sites of the rbcL gene has been related to adaptive response to increased chloroplast CO_2 concentrations in C4 photosynthesis (Kapralov et al. 2011). Positive selection was also found in ten plastid genes within the family Brassicaeae, being the signatures identify in the species Cardamine resendifolia supposed be a consequence of adaptation to high altitude enviroments (Hu et al. 2015). However, except for rbcL gene, the link between adaptive mutations in plastid genes and physiology responses to environmental constraints is poorly understood (Tonti-Filippini et al. 2017).

Another interesting acquisition during the plastome evolution is the process of RNA editing, resulting in posttranscriptional nucleotide substitutions (C-to-U and U-to-C). The changes C-to-U are more frequent, and the only type reported in most flowering plants (Takenaka et al. 2013). Since no editing has been observed in any alga, it is supposed that the plastid RNA editing arose with the evolution of land plants. Among land plants, only some species of liverworts within the order Marchantiales do not have RNA editing, presumably because they lost this process secondarily (Rüdinger et al. 2008; Takenaka et al. 2013). Generally, the mainly function of RNA editing is to restore evolutionarily conserved codons and, consequently, to codify the conserved amino acids (Tillich et al. 2005). Some RNA editing sites, if do not occur, can affect severely the protein function (Schmitz-Linneweber and Barkan 2007), while other sites occur in regions more flexibles or in nonessential genes (Fiebig et al. 2004). Other functions of plastid RNA editing in protein-coding genes include the creation of initiation codon and protein variants (Takenaka et al. 2013). RNA editing does not affect only protein-coding genes but also tRNAs, introns, and untranslated regulatory regions, highlighting the role of RNA editing in the folding and stabilization of secondary structures (Takenaka et al. 2013; Chen et al. 2017). It has been identified distinct patterns of editing between different tissues, development stages, and environmental conditions (Tseng et al. 2013; Chen et al. 2017; Mirzaei et al. 2017). Additionally, several gains and losses of RNA editing sites have been identified across land plants (Freyer et al. 1997; Fiebig et al. 2004; He et al. 2016; Chen et al. 2017), indicating that this process is very dynamic and relatively fast-evolving.

Finally, the plastome has been targeted for biotechnology and basic research approaches via plastid transformation (Bock, 2015; Rogalski et al., 2015; Daniell et al., 2016). The plastid transformation offers several advantages over the nuclear transformation such as high transgene expression levels, multigene stacking in synthetic operons in a single transformation event, precisely targeted insertion of transgenes via homologous recombination, absence of epigenetic transgene silencing effects, and increased biosafety due to exclusion of transgenes transmission by pollen since the plastids are maternally inherited in most angiosperms (Bock, 2015; Jin and Daniell, 2015; Rogalski et al., 2015). The plastid transformation has been used for several applications in plant biotechnology, including resistance to biotic and abiotic stresses (Jin et al., 2011; Zhang et al., 2015), metabolic engineering of different plant metabolic pathways (Apel and Bock, 2009; Rogalski e Carrer 2011; Kumar et al., 2012; Lu et al., 2013), improve photosynthetic performance (Dhingra et al., 2004; Lin et al., 2014; Whitney et al., 2015),

production of vaccines, biopharmaceutical compounds and enzymes for biofuel production (Boyhan and Daniell, 2011; Daniell et al., 2016; Herzog et al., 2017). The availability of complete plastome sequences provides an essential tool to develop specific plastid transformation vectors, allowing the choice of target intergenic sequences within the plastome for correct transgene integration via homologous recombination. Additionally, the complete sequence permits the exact characterization of endogenous regulatory elements that should not be affected by insertion of transgenes and can be used for an optimized transgene expression and RNA stability (Bock et al. 2014; Daniell et al. 2016). Although heterologous flanking sequences have been used for chloroplast transformation (Sidorov et al., 1999; Ruf et al., 2001), studies have showed that the use of specific flanking sequences can increase the plastid transformation efficiency (Ruhlman et al., 2010; Scotti et al., 2011), probably due to the occurrence of complex homologous recombination events when heterologous flanking sequences are used (Ruhlman et al., 2010). In addition, the use of endogenous regulatory elements favors the correct interaction with the endogenous trans-acting factors maximizing the efficiency of transgene expression (Ruhlman et al., 2010; Daniell et al., 2016).

The next-generation sequencing technology has boosted the complete characterization of huge number of plastomes in the last years, supporting advances in several areas of knowledge such as phylogeny, genetic, evolution, and biotechnology applications (Tonti-Filippini et al. 2017). Here, this work presents the complete plastomes of five oleaginous species and their detailed characterization concerning breeding, genetic and evolutionary aspects.

Flax (Linum usitatissimum L.)

Flax is an annual crop widely cultivated in the world (Kvavadze et al. 2009; FAO 2017: http://www.fao.org/faostat/en/#home). The flaxseeds are rich in polyunsaturated fatty acids, lignans, proteins, and soluble fibers, making them adequate for human nutrition and industrial applications. In addition, from the stem of flax is extracted bast fibers used in textile industry (Singh et al. 2011). Flax belongs to the family Linaceae and this study represents the first plastome sequenced of a species belonging to family Linaceae.

Crambe (Crambe abyssinica Hochst. ex R.E.Fr.)

Crambe is an oilseed plant rich in erucic acid, which is an important fatty acid to industrial applications and biofuel production (Lazzeri et al. 1997; Carlsson 2009). Studies aiming to improve agronomic traits and oil profile have been carried out (Mastebroek et al. 1994; Li et al. 2012; Cheng et al. 2015). Crambe belongs to the tribe Brassiceae (Brassicaceae),

which includes other species economically important. This study represents the first plastome within the genus Crambe to be sequenced.

Macaw palm [Acrocomia aculeata (Jacq.) Lodd. ex Mart.]

Macaw palm produces oil-rich fruits that reach up to 70% of oil suitable for biofuel production (Pires et al. 2013). It is distributed in the tropical and subtropical Americas and adapted to different ecosystems (Henderson et al. 1995; Motoike and Kuki 2009). Macaw palm belongs to the tribe Cocoseae (Arecoideae: Arecaceae) along with other economically important species as Cocos nucifera and Elaeis guineensis. This study represents the first plastome within the genus Acrocomia to be sequenced.

Murumuru (Astrocaryum murumuru) and **Tucumã** (Astrocaryum aculeatum)

Murumuru and Tucumã are palm trees of the tribe Cocoseae (Arecoideae: Arecaceae) distributed throughout tropical and subtropical ecosystems (Dransfield et al. 2008), being murumuru adapted to wet forests and tucumã adapted to terra firme areas (Kahn 2008). Both species bear oil-rich fruits that are source of food, oil for cosmetic industry, and feedstock to make crafts (Clement et al. 2005; Bezerra 2012). Since the extractivism is the main form of exploitation, it is required studies aiming to improve agronomic traits and conserve the natural resources (Ramos et al. 2011, 2012, 2016; Oliveira et al. 2017). This study represents the first plastomes within the genus Astrocaryum to be sequenced.

Beyond the plastome sequencing, the data gathered here includes molecular markers mapping, phylogenetic inferences, prediction of RNA editing sites, structural analyses, and analyses of gene divergence and positive selection. Taken together, these data provide useful genetic information to conservation, crop improvement and biotechnology strategies to be applied in these species or related ones. Moreover, the present data bring new insights about plastome evolution concerning structure, gene content, positive selection, and RNA editing within the families Linaceae, Brassicaceae, and Arecaceae.

References

Apel W, Bock R (2009) Enhancement of carotenoid biosynthesis in transplastomic tomatoes by induced lycopene-to-provitamin A conversion. Plant Physiol 151: 59–66. doi: 10.1104/pp.109.140533

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. doi: 10.1111 /nph.13617

Barthet MM, Moukarzel K, Smith KN, Patel J, Hilu KW (2015) Alternative translation initiation codons for the plastid maturase MatK: unraveling the pseudogene misconception in the Orchidaceae. BMC Evol Biol 15: 210. doi: 10.1186/s12862-015-0491-1

Bezerra VS (2012) Considerações sobre a palmeira murumuruzeiro (Astrocaryum murumuru Mart.). Embrapa, Comunicado Técnico 130

Bock R (2007) Structure, function, and inheritance of plastid genomes. In: Cell and Molecular Biology of Plastids, Bock R (Ed.) Topics in Current Genetics. Springer-Verlag, Berlin Heidelberg

Bock R (2015) Engineering plastid genomes: methods, tools, and applications in basic research and biotechnology. Annu Rev Plant Biol 66: 211–241. doi: 10.1146/annurev-arplant-050213-040212

Boyhan D, Daniell H (2011) Low-cost production of proinsulin in tobacco and lettuce chloroplasts for injectable or oral delivery of functional insulin and C-peptide. Plant Biotechnol J 9: 585–598. doi: 10.1111/j.1467-7652.2010.00582.x

Bubunenko MG, Schmidt J, Subramanian AR (1994) Protein substitution in chloroplast ribosome evolution. A eukaryotic cytosolic protein has replaced its organelle homologue (L23) in spinach. J Mol Biol 240:28–41. doi:10.1006/jmbi.1994.1415

Cai Z, Guisinger M, Kim HG, Ruck E, Blazier JC, McMurtry V, Kuehl JV, Boore J, Jansen RK (2008) Extensive reorganization of the plastid genome of Trifolium subterraneum (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions. J Mol Evol 67:696–704. doi:10.1007/s00239-008-9180-7

Carlsson AS (2009) Plant oils as feedstock alternatives to petroleum – A short survey of potential oil crop platforms. Biochimie 91: 665-670. doi: 10.1016/j.biochi.2009.03.021

Chen TC, Liu YC, Wang X, Wu CH, Huang CH, Chang CC (2017) Whole plastid transcriptomes reveal abundant RNA editing sites and differential editing status in Phalaenopsis aphrodite subsp. formosana. Bot Stud 58:38. doi: 10.1186 /s405 29-017-0193 -7

Cheng J, Salentijn EM, Huang B, Denneboom C, Qi W, Dechesne AC, Krens FA, Visser RG, van Loo EN (2015) Detection of induced mutations in CaFAD2 genes by next-generation sequencing leading to the production of improved oil composition in Crambe abyssinica. Plant Biotechnol J 13 (4): 471-481. doi: 10.1111/pbi.12269

Clement CR, Lleras Pérez E, Van Leeuwen J (2005) O potencial das palmeiras tropicais no Brasil: acertos e fracassos das últimas décadas. Agrociências 9: 67-71

Cosner ME, Raubeson LA, Jansen RK (2004) Chloroplast DNA rearrangements in Campanulaceae: phylogenetic utility of highly rearranged genomes. BMC Evol Biol 4: 27. doi: 10.1186/1471-2148-4-27

Daniell H, Wurdack KJ, Kanagaraj A, Lee S-B, Saski C, Jansen RK (2008) The complete nucleotide sequence of the cassava (Manihot esculenta) chloroplast genome and the evolution of atpF in Malpighiales: RNA editing and multiple losses of a group II intron. TAG Theor Appl Genet Theor Angew Genet 116: 723–737. doi: 10.1007/s00122-007-0706-y

Daniell H, Lin C-S, Yu M, Chang W-J (2016) Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol 17: 134. doi: 10.1186/s13059-016-1004-2

Dhingra A, Portis AR, Daniell H (2004) Enhanced translation of a chloroplast-expressed RbcS gene restores small subunit levels and photosynthesis in nuclear RbcS antisense plants. Proc Natl Acad Sci USA 101: 6315–6320. doi: 10.1073/pnas.0400981101

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

Dransfield J, Uhl NW, Asmussen CB, Baker WJ, Harley M, Lewis C (2008) Genera Palmarum: the evolution and classification of palms. Kew Publishing, Royal Botanical Garden, Londres, 732

Ebert D, Peakall R (2009) Chloroplast simple sequence repeats (cpSSRs): technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant species. Mol Ecol Resour 9:673–690. doi: 10.1111/j.1755 -0998 .2008 .0231 9.x

Fiebig A, Stegemann S, Bock R (2004) Rapid evolution of RNA editing sites in a small non-essential plastid gene. Nucleic Acids Res 32:3615–3622. doi:10.1093/nar/gkh695

Freyer R, Kiefer-Meyer MC, Kössel H (1997) Occurrence of plastid RNA editing in all major lineages of land plants. Proc Natl Acad Sci USA 94: 6285–6290

Galili G, Avin-Wittenberg T, Angelovici R, Fernie AR (2014) The role of photosynthesis and amino acid metabolism in the energy status during seed development. Front Plant Sci 5:447. doi: 10.3389/fpls.2014.00447

Gantt JS, Baldauf SL, Calie PJ, Weeden NF, Palmer JD (1991) Transfer of rpl22 to the nucleus greatly preceded its loss from the chloroplast and involved the gain of an intron. EMBO J 10: 3073–3078

Guisinger MM, Kuehl JV, Boore JL, Jansen RK (2011) Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. Mol Biol Evol 28: 583–600. doi: 10.1093/molbev/msq229

Haberle RC, Fourcade HM, Boore JL, Jansen RK (2008) Extensive Rearrangements in the Chloroplast Genome of Trachelium caeruleum Are Associated with Repeats and tRNA Genes. J Mol Evol 66:350-361. doi: 10.1007/s00239-008-9086-4

He P, Huang S, Xiao G, Zhang Y, Yu J (2016) Abundant RNA editing sites of chloroplast protein-coding genes in Ginkgo biloba and an evolutionary pattern analysis. BMC Plant Biol 16(1):257. doi: 10.1186 /s128 70-016-0944 -8

Henderson A, Galeano G, Bernal R (1995) Field guide to the palms of the Americas. Princeton University Press, Princeton

Herzog RW, Nichols TC, Su J, Zhang B, Sherman A, Merricks EP, Raymer R, Perrin GQ, Häger M, Wiinberg B, Daniell H (2017) Oral Tolerance Induction in Hemophilia B Dogs Fed with Transplastomic Lettuce. Mol Ther J Am Soc Gene Ther 25: 512–522. doi: 10.1016/j.ymthe.2016.11.009

Horváth EM, Peter SO, Joel T, Rumeau D, Cournac L, Horváth G, Kavanagh TA, Schaefer C, Medgyesy P (2000) Targeted inactivation of the plastid ndhB gene in tobacco results in an enhanced sensitivity of photosynthesis to moderate stomatal closure. Plant Physiol 123:1337–1349

Jansen RK, Saski C, Lee S-B, Hansen AK, Daniell H (2011) Complete plastid genome sequences of three Rosids (Castanea, Prunus, Theobroma): evidence for at least two independent transfers of rpl22 to the nucleus. Mol Biol Evol 28: 835–847. doi: 10.1093/molbev/msq261

Jin S, Kanagaraj A, Verma D, Lange T, Daniell H (2011) Release of hormones from conjugates: chloroplast expression of β-glucosidase results in elevated phytohormone levels associated with significant increase in biomass and protection from aphids or whiteflies conferred by sucrose esters. Plant Physiol 155: 222–235. doi: 10.1104/pp.110.160754

Jin S, Daniell H (2015) The Engineered Chloroplast Genome Just Got Smarter. Trends Plant Sci 20: 622–640. doi: 10.1016/j.tplants.2015.07.004

Kahn F (2008) El género Astrocaryum (Arecaceae). Rev Peru Biol 15: 31–48

Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M, Sugiura M, Sasamoto S, et al. (1996) Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes 3: 109–136

Kapralov MV, Kubien DS, Andersson I, Filatov DA (2011). Changes in Rubisco kinetics during the evolution of C4 photosynthesis in Flaveria (Asteraceae) are associated with positive selection on genes encoding the enzyme. Mol Biol Evol 28: 1491–1503. doi: 10.1093/molbev/msq335

Keller J, Rousseau-Gueutin M, Martin GE, Morice J, Boutte J, Coissac E, Ourari M, Aïnouche M, Salmon A, Cabello-Hurtado F, Aïnouche A (2017) The evolutionary fate of the chloroplast and nuclear rps16 genes as revealed through the sequencing and comparative analyses of four novel legume chloroplast genomes from Lupinus. DNA Res Int J Rapid Publ Rep Genes Genomes 24: 343–358. doi: 10.1093/dnares/dsx006

Kikuchi S, Bédard J, Hirano M, Hirabayashi Y, Oishi M, Imai M, Takase M, Ide T, Nakai M (2013) Uncovering the protein translocon at the chloroplast inner envelope membrane. Science 339:571–574. doi:10.1126/science.1229262

Kim K-J, Choi K-S, Jansen RK (2005) Two chloroplast DNA inversions originated simultaneously during the early evolution of the sunflower family (Asteraceae). Mol Biol Evol 22: 1783–1792. doi: 10.1093/molbev/msi174

Konishi T, Shinohara K, Yamada K, Sasaki Y (1996) Acetyl-CoA carboxylase in higher plants: most plants other than gramineae have both the prokaryotic and the eukaryotic forms of this enzyme. Plant Cell Physiol 37: 117–122.

Kumar S, Hahn FM, Baidoo E, Kahlon TS, Wood DF, McMahan CM, Cornish K, Keasling JD, Daniell H, Whalen MC (2012) Remodeling the isoprenoid pathway in tobacco by expressing the cytoplasmic mevalonate pathway in chloroplasts. Metab Eng 14: 19–28. doi: 10.1016/j.ymben.2011.11.005

Kvavadze E, Bar-Yosef O, Belfer-Cohen A, Boaretto E, Jakeli N, Matskevich Z, Meshveliani T (2009) 30,000-year-old wild flax fibers. Science 325:1359. doi:10.1126/science.1175404

Lazzeri L, DeMattei F, Bucelli F, Palmieri S (1997) Crambe oil – a potential new hydraulic oil and quenchant. Ind Lubr Tribol 49: 71-77

Li XG, Duan W, Meng QW, Zou Q, Zhao SJ (2004) The function of chloroplastic NAD(P)H dehydrogenase in tobacco during chilling stress under low irradiance. Plant Cell Physiol 45:103–108. doi:10.1093/pcp/pch011

Li X, van Loo EN, Gruber J, Fan J, Guan R, Frentzen M, Stymne S, Zhu LH (2012) Development of ultra high erucic acid oil in the industrial oil crop Crambe abyssinica. Plant Biotechnol J 10 (7): 862-70. doi: 10.1111/j.1467-7652.2012.00709.x

Li P, Li Z, Liu H, Hua J (2014) Cytoplasmic diversity of the cotton genus as revealed by chloroplast microsatellite markers. Genet. Resour. Crop Evol 61: 107–119. doi: 10.1007/s10722-013-0018-9

Lin MT, Occhialini A, Andralojc PJ, Parry MAJ, Hanson MR (2014) A faster Rubisco with potential to increase photosynthesis in crops. Nature 513: 547–550. doi: 10.1038/nature13776

Lin C-P, Ko C-Y, Kuo C-I, Liu M-S, Schafleitner R, Chen L-FO (2015a) Transcriptional Slippage and RNA Editing Increase the Diversity of Transcripts in Chloroplasts: Insight from Deep Sequencing of Vigna radiata Genome and Transcriptome. PloS One 10: e0129396. doi: 10.1371/journal.pone.0129396

Lin C-S, Chen JJW, Huang Y-T, Chan M-T, Daniell H, Chang W-J, Hsu C-T, Liao D-C, Wu F-H, Lin S-Y, Liao C-F, Deyholos MK, Wong GK-S, Albert VA, Chou M-L, Chen C-Y, Shih M-C (2015b) The location and translocation of ndh genes of chloroplast origin in the Orchidaceae family. Sci Rep 5: 9040. doi: 10.1038/srep09040

Lu Y, Rijzaani H, Karcher D, Ruf S, Bock R (2013) Efficient metabolic pathway engineering in transgenic tobacco and tomato plastids with synthetic multigene operons. Proc Natl Acad Sci USA 110: E623-632. doi: 10.1073/pnas.1216898110

Martin GE, Rousseau-Gueutin M, Cordonnier S, Lima O, Michon-Coudouel S, Naquin D, de Carvalho JF, Aïnouche M, Salmon A, Aïnouche A (2014) The first complete chloroplast genome of the Genistoid legume Lupinus luteus: evidence for a novel major lineage-specific rearrangement and new insights regarding plastome evolution in the legume family. Ann Bot 113: 1197–1210. doi: 10.1093/aob/mcu050

Mastebroek HD, Wallenburg SC, van Soest LJM (1994) Variation for agronomic characteristics in crambe (Crambe abyssinica Hochst. ex Fries). Ind Crop Prod 2(2): 129-36

Millen RS, Olmstead RG, Adams KL, Palmer JD, Lao NT, Heggie L, Kavanagh TA, Hibberd JM, Gray JC, Morden CW, Calie PJ, Jermiin LS, Wolfe KH (2001) Many parallel losses of infA from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. Plant Cell 13: 645–658.

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

Mirzaei S, Mansouri M, Mohammadi-Nejad G, Sablok G (2017) Comparative assessment of chloroplast transcriptional responses highlights conserved and unique patterns across Triticeae members under salt stress. Photosynth Res 1–13. doi: 10.1007/s11120-017-0469-5

Motoike SY, Kuki KN (2009) The potential of macaw palm (Acrocomia aculeata) as source of biodiesel in Brazil. IRECHE 1:632–635

Nakai M (2015) The TIC complex uncovered: the alternative view on the molecular mechanism of protein translocation across the inner envelope membrane of chloroplasts. Biochim Biophys Acta BBA Bioenerg 1847:957–967. doi:10.1016/j.bbabio.2015.02.011

Oliveira NP, Oliveira MSP, Davide LC, Kalisz S (2017) Population genetic structure of three species in the genus Astrocaryum G. Mey. (Arecaceae). Genet Mol Res GMR 16. doi: 10.4238/gmr16039676

Piot A, Hackel J, Christin P-A, Besnard G (2018) One-third of the plastid genes evolved under positive selection in PACMAD grasses. Planta 247: 255–266. doi: 10.1007/s00425-017-2781-x

Pires TP, dos Santos Souza E, Kuki KN, Motoike SY (2013) Ecophysiological traits of the macaw palm: a contribution towards the domestication of a novel oil crop. Ind Crops Prod 44:200–210

Provan J, Powell W, Hollingsworth PM (2001) Chloroplast microsatellites: new tools for studies in plant ecology and evolution. Trends Ecol Evol 16:142–147

Pyke K (2007) Plastid biogenesis and differentiation. In: Cell and Molecular Biology of Plastids, Bock R (Ed.) Topics in Current Genetics. Springer-Verlag, Berlin Heidelberg

Ramos SLF, Lopes MTG, Lopes R, Cunha RNV da, Macêdo JLV de, Contim LAS, Clement CR, Rodrigues DP, Bernardes LG (2011) Determination of the mating system of Tucumã palm using microsatellite markers. Crop Breed Appl Biotechnol 11: 181–185. doi: 10.1590/S1984-70332011000200011

Ramos SLF, de Macêdo JLV, Lopes MTG, Batista JS, Formiga KM, da Silva PP, Saulo-Machado AC, Veasey EA (2012) Microsatellite loci for tucumã of Amazonas (Astrocaryum aculeatum) and amplification in other Arecaceae. Am J Bot 99: e508–e510. doi: 10.3732/ajb.1100607

Ramos SLF, Dequigiovanni G, Sebbenn AM, Lopes MTG, Kageyama PY, de Macêdo JLV, Kirst M, Veasey EA (2016) Spatial genetic structure, genetic diversity and pollen dispersal in a harvested population of Astrocaryum aculeatum in the Brazilian Amazon. BMC Genet 17(63). doi: 10.1186/s12863-016-0371-8

Rogalski M, Ruf S, Bock R (2006) Tobacco plastid ribosomal protein S18 is essential for cell survival. Nucleic Acids Res 34:4537–4545. doi: 10.1093/nar/gkl634

Rogalski M, Carrer H (2011) Engineering plastid fatty acid biosynthesis to improve food quality and biofuel production in higher plants: Plastid fatty acid biosynthesis. Plant Biotechnol J 9: 554–564. doi: 10.1111/j.1467-7652.2011.00621.x

Rogalski M, do Nascimento Vieira L, Fraga HP, Guerra MP (2015) Plastid genomics in horticultural species: importance and applications for plant population genetics, evolution, and biotechnology. Front Plant Sci 6: 586. doi: 10.3389/fpls.2015.00586

Roy PS, Rao GJN, Jena S, Samal R, Patnaik A, Patnaik SSC, Jambhulkar NN, Sharma S, Mohapatra T (2016) Nuclear and chloroplast DNA Variation provides insights into population structure and multiple origin of native aromatic rices of Odisha, India. PloS One 11:e0162268. doi: 10.1371 /jour nal.pone.0162 268

Rüdinger M, Polsakiewicz M, Knoop V (2008) Organellar RNA editing and plant-specific extensions of pentatricopeptide repeat proteins in jungermanniid but not in marchantiid liverworts. Mol Biol Evol 25:1405–1414. doi: 10.1093/molbev/msn084

Ruf S, Hermann M, Berger IJ, Carrer H, Bock R (2001) Stable genetic transformation of tomato plastids and expression of a foreign protein in fruit. Nat Biotechnol 19: 870–875. doi: 10.1038/nbt0901-870

Ruhlman T, Verma D, Samson N, Daniell H (2010) The role of heterologous chloroplast sequence elements in transgene integration and expression. Plant Physiol 152: 2088–2104. doi: 10.1104/pp.109.152017

Ruhlman TA, Chang W-J, Chen JJW, Huang Y-T, Chan M-T, Zhang J, Liao D-C, Blazier JC, Jin X, Shih M-C, Jansen RK, Lin C-S (2015) NDH expression marks major transitions in plant evolution and reveals coordinate intracellular gene loss. BMC Plant Biol 15: 100. doi: 10.1186/s12870-015-0484-7

Schmitz-Linneweber C, Barkan A (2007) RNA splicing and RNA editing in chloroplasts. In: Cell and Molecular Biology of Plastids, Bock R (Ed.) Topics in Current Genetics. Springer-Verlag, Berlin Heidelberg

Scotti N, Valkov VT, Cardi T (2011) Improvement of plastid transformation efficiency in potato by using vectors with homologous flanking sequences. GM Crops 2: 89–91. doi: 10.4161/gmcr.2.2.17504

Sidorov VA, Kasten D, Pang SZ, Hajdukiewicz PT, Staub JM, Nehra NS (1999) Technical Advance: Stable chloroplast transformation in potato: use of green fluorescent protein as a plastid marker. Plant J Cell Mol Biol 19: 209–216

Singh KK, Mridula D, Rehal J, Barnwal P (2011) Flaxseed: a potential source of food, feed and fiber. Crit Rev Food Sci Nutr 51:210–222. doi:10.1080/10408390903537241

Svab Z, Maliga P (1993) High-frequency plastid transformation in tobacco by selection for a chimeric aadA gene. Proc Natl Acad Sci 90: 913–917. doi: 10.1073/pnas.90.3.913

Takenaka M, Zehrmann A, Verbitskiy D, Härtel B, Brennicke A (2013) RNA editing in plants and its evolution. Annu Rev Genet 47:335–352. doi:10.1146/annurev-genet-111212-133519

Tetlow I J, Morell MK, Emes MJ (2004) Recent developments in understanding the regulation of starch metabolism in higher plants. J Exp Bot 55: 2131–2145.

Tillich M, Funk HT, Schmitz-Linneweber C, Poltnigg P, Sabater B, Martin M, Maier RM (2005) Editing of plastid RNA in Arabidopsis thaliana ecotypes. Plant J 43:708–715. doi:10.1111/j.1365-313X.2005.02484.x

Tonti-Filippini J, Nevill PG, Dixon K, Small I (2017) What can we do with 1000 plastid genomes? Plant J. Cell Mol Biol 90: 808–818. doi: 10.1111/tpj.13491

Tseng C-C, Lee C-J, Chung Y-T, Sung T-Y, Hsieh M.-H., 2013. Differential regulation of Arabidopsis plastid gene expression and RNA editing in non-photosynthetic tissues. Plant Mol Biol 82: 375–392. doi: 10.1007/s11103-013-0069-5

Ueda M, Fujimoto M, Arimura S, Murata J, Tsutsumi N, Kadowaki K (2007) Loss of the rpl32 gene from the chloroplast genome and subsequent acquisition of a preexisting transit peptide within the nuclear gene in Populus. Gene 402: 51–56. doi: 10.1016/j.gene.2007.07.019

Ueda M, Nishikawa T, Fujimoto M, Takanashi H, Arimura S-I, Tsutsumi N, Kadowaki K-I (2008) Substitution of the gene for chloroplast RPS16 was assisted by generation of a dual targeting signal. Mol Biol Evol 25: 1566–1575. doi: 10.1093/molbev/msn102

Vieira LN, Rogalski M, Faoro H, Fraga HP, Anjos KG, Picchi GFA, Nodari RO, Pedrosa FO, Souza EM, Guerra MP (2016) The plastome sequence of the endemic Amazonian conifer, Retrophyllum piresii (Silba) C.N.Page, reveals different recombination events and plastome isoforms. Tree Genet Genomes 12:10. doi: 10.1007/s11295-016-0968-0

Vries J, Sousa FL, Bölter B, Soll J, Gould SB (2015) YCF1: a green TIC? Plant Cell 27:1827–1833. doi:10.1105/tpc.114.135541

Wambugu PW, Brozynska M, Furtado A, Waters DL, Henry RJ (2015) Relationships of wild and domesticated rices (Oryza AA genome species) based upon whole chloroplast genome sequences. Sci Rep 5. doi: 10.1038/srep13957

Wambulwa MC, Meegahakumbura MK, Kamunya S, Muchugi A, Möller M, Liu J, Xu JC, Ranjitkar S, Li DZ, Gao LM (2016) Insights into the genetic relationships and breeding patterns of the African tea germplasm based on nSSR markers and cpDNA sequences. Front Plant Sci 7:1244. doi: 10.3389 /fpls.2016 .0124 4

Wei R, Yan Y-H, Harris A, Kang J-S, Shen H, Xiang Q-P, Zhang X-C (2017) Plastid Phylogenomics Resolve Deep Relationships among Eupolypod II Ferns with Rapid Radiation and Rate Heterogeneity. Genome Biol Evol 9: 1646–1657. doi: 10.1093/gbe/evx107

Weng ML, Blazier JC, Govindu M, Jansen RK (2014) Reconstruction of the Ancestral Plastid Genome in Geraniaceae Reveals a Correlation between Genome Rearrangements, Repeats, and Nucleotide Substitution Rates. Mol Biol Evol 31:645-659. doi: 10.1093/molbev/mst257

Whitney SM, Birch R, Kelso C, Beck JL, Kapralov MV (2015) Improving recombinant Rubisco biogenesis, plant photosynthesis and growth by coexpressing its ancillary RAF1 chaperone. Proc Natl Acad Sci USA 112: 3564–3569. doi: 10.1073/pnas.1420536112

Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, Quandt D (2011) The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol 76:273– 297. doi:10.1007/s11103-011-9762-4

Wheeler GL, Dorman HE, Buchanan A, Challagundla L, Wallace LE (2014) A review of the prevalence, utility, and caveats of using chloroplast simple sequence repeats for studies of plant biology. Appl Plant Sci. doi: org/10.3732 /apps .1400 059

Xi Z, Ruhfel BR, Schaefer H et al (2012) Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. Proc Natl Acad Sci USA 109:17519–17524. doi:10.1073/pnas.1205818109

Xu J-H, Liu Q, Hu W, Wang T, Xue Q, Messing J (2015) Dynamics of chloroplast genomes in green plants. Genomics 106: 221–231. doi: 10.1016/j.ygeno.2015.07.004

Yamori W, Shikanai T (2016) Physiological Functions of Cyclic Electron Transport Around Photosystem I in Sustaining Photosynthesis and Plant Growth. Annu Rev Plant Biol 67: 81–106. doi: 10.1146/annurev-arplant-043015-112002

Zhang J, Khan SA, Hasse C, Ruf S, Heckel DG, Bock R (2015) Pest control. Full crop protection from an insect pest by expression of long double-stranded RNAs in plastids. Science 347: 991–994. doi: 10.1126/science.1261680

Zhang J, Ruhlman TA, Sabir JSM, Blazier JC, Weng M-L, Park S, Jansen RK (2016) Coevolution between Nuclear-Encoded DNA Replication, Recombination, and Repair Genes and Plastid Genome Complexity. Genome Biol Evol 8: 622–634. doi: 10.1093/gbe/evw033

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

The Linum usitatissimum L. plastome reveals atypical structural evolution, new editing sites, and the phylogenetic position of Linaceae within Malpighiales

Amanda de Santana Lopes[1], Túlio Gomes Pacheco[1], Karla Gasparini dos Santos[1], Leila do Nascimento Vieira[2], Miguel Pedro Guerra[2], Rubens Onofre Nodari[2], Emanuel Maltempi de Souza[3], Fábio de Oliveira Pedrosa[3], Marcelo Rogalski[1*]

1 Laboratório de Fisiologia Molecular de Plantas, Departamento de Biologia Vegetal, Universidade Federal de Viçosa, Viçosa-MG, Brazil.
2 Laboratório de Fisiologia do Desenvolvimento e Genética Vegetal, Programa de Pós-graduação em Recursos Genéticos Vegetais, Universidade Federal de Santa Catarina, Florianópolis-SC, Brazil.
3 Departamento de Bioquímica e Biologia Molecular, Núcleo de Fixação Biológica de Nitrogênio, Universidade Federal do Paraná, Curitiba-PR, Brazil.

*****Corresponding author
E-mail address: rogalski@ufv.br

Published in:
Plant Cell Rep (2018) 37:307–328
DOI: 10.1007/s00299-017-2231-z

Plant Cell Rep (2018) 37:307–328
DOI 10.1007/s00299-017-2231-z

ORIGINAL ARTICLE

The *Linum usitatissimum* L. plastome reveals atypical structural evolution, new editing sites, and the phylogenetic position of Linaceae within Malpighiales

Amanda de Santana Lopes[1] · Túlio Gomes Pacheco[1] · Karla Gasparini dos Santos[1] · Leila do Nascimento Vieira[2] · Miguel Pedro Guerra[2] · Rubens Onofre Nodari[2] · Emanuel Maltempi de Souza[3] · Fábio de Oliveira Pedrosa[3] · Marcelo Rogalski[1]

Received: 9 August 2017 / Accepted: 18 October 2017 / Published online: 30 October 2017

Abstract

Key message **The plastome of *Linum usitatissimum* was completely sequenced allowing analyses of evolution of genome structure, RNA editing sites, molecular markers, and indicating the position of Linaceae within Malpighiales.**

Abstract Flax (*Linum usitatissimum* L.) is an economically important crop used as food, feed, and industrial feedstock. It belongs to the Linaceae family, which is noted by high morphological and ecological diversity. Here, we reported the complete sequence of flax plastome, the first species within Linaceae family to have the plastome sequenced, assembled and characterized in detail. The plastome of flax is a circular DNA molecule of 156,721 bp with a typical quadripartite structure including two IRs of 31,990 bp separating the LSC of 81,767 bp and the SSC of 10,974 bp. It shows two expansion events from IRB to LSC and from IRB to SSC, and a contraction event in the IRA-LSC junction, which changed significantly the size and the gene content of LSC, SSC and IRs. We identified 109 unique genes and 2 pseudogenes (*rpl23* and *ndhF*). The plastome lost the conserved introns of *clpP* gene and the complete sequence of *rps16* gene. The *clpP, ycf1*, and *ycf2* genes show high nucleotide and aminoacid divergence, but they still possibly retain the functionality. Moreover, we also identified 176 SSRs, 20 tandem repeats, and 39 dispersed repeats. We predicted in 18 genes a total of 53 RNA editing sites of which 32 were not found before in other species. The phylogenetic inference based on 63 plastid protein-coding genes of 38 taxa supports three major clades within Malpighiales order. One of these clades has flax (Linaceae) sister to Chrysobalanaceae family, differing from earlier studies that included Linaceae into the euphorbioid clade.

Keywords Extranuclear inheritance · Rearrangements · Gene duplication · Plastid microsatellites · Plastid evolution

Communicated by Teodoro Cardi.

Electronic supplementary material The online version of this article (doi:10.1007/s00299-017-2231-z) contains supplementary material, which is available to authorized users.

✉ Marcelo Rogalski
rogalski@ufv.br

[1] Laboratório de Fisiologia Molecular de Plantas, Departamento de Biologia Vegetal, Universidade Federal de Viçosa, Viçosa, MG, Brazil

[2] Laboratório de Fisiologia do Desenvolvimento e Genética Vegetal, Programa de Pós-Graduação em Recursos Genéticos Vegetais, Universidade Federal de Santa Catarina, Florianópolis, SC, Brazil

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

Introduction

Flax (*Linum usitatissimum*) is an annual plant crop with millenary use by human civilizations as source of fiber and food (Zeist and Bakker-Heeres 1975; Kvavadze et al. 2009). Nowadays, this crop is cultivated in more than 30 countries (FAO 2017: http://www.fao.org/faostat/en/#home) and has broad use as food, feed, and industrial feedstock. The bast fibers, used in textile industry, are extracted from the stem of flax, however, the most valuable product is the flaxseed, a rich source of polyunsaturated fatty acids, lignans, proteins, and soluble fibers, which makes it a functional food (Singh et al. 2011). Due to the high content of polyunsaturated fatty acids, mainly α-linolenic acid, flax oil is the noted

commercial source of plant omega-3 fatty acid, an essential fatty acid with a wide range of health benefits for humans (Simmons et al. 2011). In addition, the oil extracted from flaxseeds is also used by industry for several applications (Carlsson 2009). Moreover, flaxseed lignans have shown promising results for treatment and to prevent several types of cancers (Touré and Xueming 2010).

Flax belongs to the family Linaceae, which is included within the order Malpighiales. The family Linaceae has high morphological and ecological diversity and contains more than 270 species and 14 genera (McDill et al. 2009). The *Linum* is the main genus and contains more than 180 species, showing several morphological and physiological variations and broad biogeographic distribution (McDill et al. 2009). The order Malpighiales also exhibits remarkable morphological and ecological diversity and contains approximately 16,000 species and 42 families (Wurdack and Davis 2009), which makes this order target for extensive evolutionary and phylogenetic analyses (Wurdack and Davis 2009; Xi et al. 2012). However, just few species belonging to six families (Chrysobalanaceae, Euphorbiaceae, Erythroxylaceae, Passifloraceae, Salicaceae, and Violaceae), have the plastid genome (plastome) sequenced and available in the plastome databases (http://www.ncbi.nlm.nih.gov/genome/organelle) to date. Plastomes of species belonging to order Malpighiales have shown uncommon characteristics related to evolution such as absence of conserved ribosomal protein genes, losses of conserved introns in various genes, specific genome rearrangements, and high nucleotide divergences in some genes (Daniell et al. 2008; Asif et al. 2010; Rivarola et al. 2011; Tangphatsornruang et al. 2011; Malé et al. 2014; Wu 2016; Cheon et al. 2015; Cauz-Santos et al. 2017).

Plastome sequences are efficient markers with wide use for phylogenetic, genetic, conservation and evolutionary studies (Besnard et al. 2011; Dexter et al. 2012; López et al. 2012; Rogalski et al. 2015). The high conservation of the plastid genes between plant species has been explored for phylogenetic inferences (Jansen et al. 2008; Xi et al. 2012; Vieira et al. 2016a). On the other hand, intergenic spacers, introns, and molecular makers such as single-nucleotide polymorphisms (SNPs) and single-sequence repeats (SSRs) have a relatively higher mutation rate and they are, therefore, useful for population genetics and phylogeographycal studies (Besnard et al. 2011; Rogalski et al. 2015; Qiao et al. 2016). Plastome sequences have been used for understanding of evolutionary events in plants based on the analysis of gene content, recombination events, loss of genes, gene transfer to the nucleus and genome rearrangements (Guo et al. 2007; Jansen et al. 2011; Wicke et al. 2011; Vieira et al. 2014a, 2016b), as well as for plastid transformation aiming basic research and biotechnological applications (Rogalski et al. 2006; Rogalski and Carrer 2011; Alkatib et al. 2012; Bock 2015; Daniell et al. 2016).

Flax has long historical importance for humans with its unique nutritional characteristics and together with its morphological and ecological diversity is a species with great interest for genetic and evolutionary studies. Additionally, the Linaceae family belongs to the Malpighiales order, which is a plant lineage with exceptional mode of plastidial evolution. Given the large number of interesting features inserted in this species or even in this family, there is no complete plastome published of this lineage to date.

Therefore, we reported here the complete sequence of flax (*Linum usitatissimum* L.) plastome, the first species of the family Linaceae to have the plastome completely sequenced, assembled and characterized in detail. The plastome of flax shows two expansion events from IRB to LSC involving the *ycf1, rps15, ndhH* and *ndhA* genes and from IRB to SSC enwrapping the *trnH-GUG, psbA, trnK-UUU* and *matK* genes, and a contraction event in the IRA-LSC junction covering the *rps19, rpl2, rpl23* and *trnI-CAU* genes, which changed significantly the size and the gene content of LSC, SSC and IR regions. In addition, we identified 109 unique genes and 2 pseudogenes. Moreover, we analyzed the nucleotide and aminoacid divergences of 69 protein coding genes. Furthermore, we also identified 176 SSRs, 20 tandem repeats, 39 dispersed repeats and 53 RNA-editing sites of which 32 were not found before in other species. Finally, our phylogenetic inference supports three major clades within Malpighiales order and included flax (Linaceae) as a sister to Chrysobalanaceae family, differing from earlier studies that included Linaceae into the euphorbioid clade. Taken together, our data raise questions about the evolution of RNA editing sites, genome structure, gene content, and the phylogenetic position of the family Linaceae.

Materials and methods

Plant material and cp DNA purification

Commercial golden linseeds (*Linum usitatissimum*) were germinated on soil under greenhouse condition at Federal University of Viçosa, Viçosa-MG, Brazil. Fresh young leaves were collected and kept on dark for 96 h at 4 °C to decrease starch content. The chloroplast isolation and cp DNA extraction were carried out according to Vieira et al. (2014b).

Chloroplast genome sequencing, assembling and annotation

Approximately, 1 ng of cp DNA was used to prepare sequencing libraries with Nextera XT DNA Sample Prep Kit (Illumina Inc., San Diego, CA, USA) according to the manufacturer's instructions. The obtained library was

sequenced using Illumina MiSeq platform (Illumina Inc., San Diego, CA, USA) at the Federal University of Paraná (Curitiba-PR, Brazil) sequencing facility. The paired-end reads (2 × 250 bp) were trimmed under the threshold with probability of error < 0.05. The trimmed reads (882.077 reads) were de novo assembled using CLC Genomics Workbench 8.0.2 software (CLC Bio, Aarhus, Denmark). The lowest average coverage of the assemble contigs used for assembling the plastid genome was 298.33. Preliminarily, gene annotation of the flax plastid genome was carried out through the program Dual Organellar GenoMe Annotator (DOGMA) (Wyman et al. 2004) and BLAST searches. From this initial annotation, putative starts, stops, and intron positions were determined based on comparisons to homologous genes in other plastid genomes. All tRNA genes were further verified using tRNAscan-SE server (Lowe and Eddy 1997). The physical map of the plastid circular genome was drawn using Organellar Genome DRAW (OGDRAW) (Lohse et al. 2013). The complete plastome sequence of flax was deposited in the GenBank database under Accession Number KY849971.

Genome structure, repeat sequence and prediction of RNA-editing sites

Mauve Genome Alignment 2.3.1v software (Darling et al. 2004) and Nucleotide MUMmer (NUCmer) Perl script in MUMmer 3.0 (Kurtz et al. 2004) were used to visualize and compare the plastome structures between *L. usitatissimum* and other Malpighiales representatives, as well as *Arabidopsis thaliana*, a Brassicales and external representative.

Simple sequence repeats (SSRs) or microsatellites were detected using the MIcroSAtellite (MISA) Perl script (Thiel et al. 2003), with thresholds of eight repeat units for mononucleotide SSRs, four repeat units for di- and trinucleotide SSRs, and three repeat units for tetra-, penta- and hexanucleotide SSRs. Tandem repeats were identified using the program Tandem Repeats Finder (TRF) (Benson 1999). The parameter settings used were 2, 7 and 7 for match, mismatch, and indel, respectively. The minimum alignment score to report repeat and maximum period size were set as 80 and 500, respectively. After, they were found that the repeats were manually verified, and the nested or redundant results were removed. REPuter (Kurtz et al. 2001) was used to locate IRs by forward vs. reverse complement (palindromic) alignment. The minimal repeat size was set to 30 bp and the identity of repeats ≥ 90% (hamming distance = 3).

Potential RNA editing sites in protein-coding genes of flax cpDNA were predicted by the program Predictive RNA Editor for Plants (PREP) suite (Mower 2009), that use 35 reference genes for detecting RNA editing sites in plastid genomes. The cutoff value was set at 0.8. Reference genes: *accD, atpA, atpB, atpF, atpI, ccsA, clpP, matK, ndhA, ndhB, ndhD, ndhF, ndhG, petB, petD, petG, petL, psaB, psaI, psbB, psbE, psbF, psbL, rpl2, rpl20, rpl23, rpoA, rpoB, rpoC1, rpoC2, rps2, rps8, rps14, rps16*, and *ycf3*.

Phylogenetic inference

For the inference of the flax phylogenetic position within of Malpighiales, 63 conserved protein-coding genes were extracted from other Malpighiales families, including the Chrysobalanaceae (19 genera), Euphorbiaceae (five genera), Erythroxylaceae (one genus), Passifloracea (one genus), Salicaceae (three genera), and Violaceae (one genus). Other orders of Fabids were also used, including Rosales (two families), Cucurbitales, Fabales e Fagales (one family for each order). A malvids representative was used as outgroup. The GenBank accession number of each taxon is showed in Table S5. The protein-coding genes were extracted and aligned individually using the software Muscle (Edgar 2004) implemented in Mega 6.0 (Tamura et al. 2013). The software Sequence Matrix 1.7.8 (Vaidya et al. 2011) was used to concatenate the genes resulting in a total sequence of 48,384 bp. Partition Finder (Lanfear et al. 2012) was used to search for the best set of parameters to be optimized during the phylogenetic search for the best tree. Bayesian inference was performed using MrBayes version 3.2 (Ronquist et al. 2012), with one million generations of two runs of four Markov Chains, three hot and one cold in each run. The software Tracer 1.6 (http://tree.bio.ed.ac.uk/software/tracer/) was used to check the parameters convergence.

Pairwise distance analysis and synonymous (dS) and nonsynonymous (dN) substitution rates

Pairwise distance was calculated based on the nucleotide and amino acid sequences for 69 protein-coding genes, including putative pseudogenes, identified in the plastid genome of *L. usitatissimum*. The species used in this analysis are highlighted in the Table S5. The sequences were aligned individually by Muscle (Edgar 2004) and the matrix of pairwise distance was generated using Mega 6.0 software (Tamura et al. 2013). Pairwise deletion was set to Gaps/Missing data treatment. Further, the pairwise d*S* and d*N* values were estimated using Mega 6.0 software (Tamura et al. 2013) under the Kumar model (Kimura 2-para).

Results

Gene content and organization of plastid genome

The plastome of flax is a circular DNA molecule of 156,721 bp and has a typical quadripartite structure (Fig. 1) with a pair of inverted repeats (IRA and IRB) of 31,990 bp

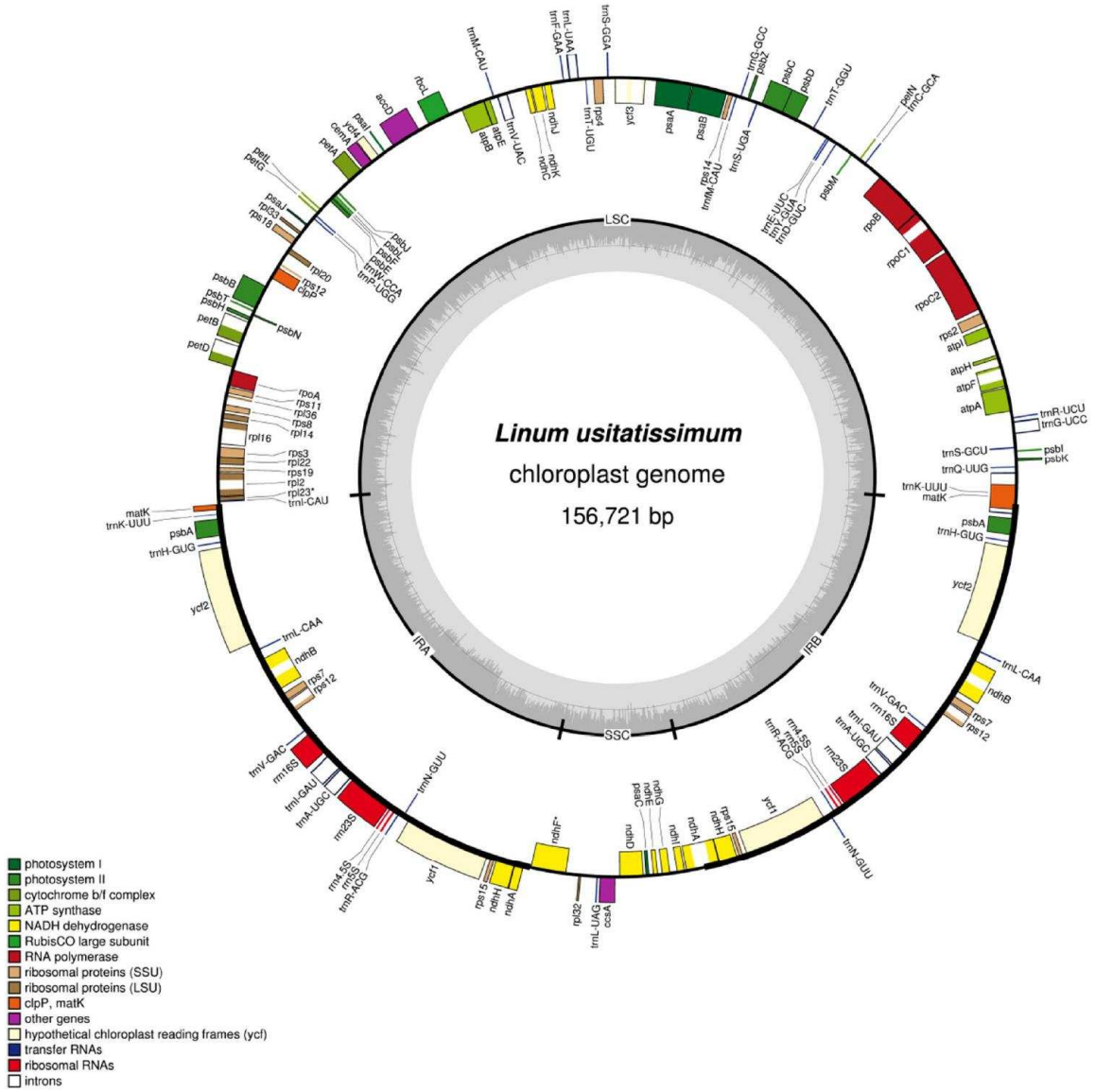


Fig. 1 Gene map of *Linum usitatissimum* chloroplast genome. Genes drawn inside the circle are transcribed in the clockwise direction, and genes drawn outside are transcribed in the counterclockwise direction. Different functional groups of genes are color coded. The darker gray in the inner circle corresponds to GC content, and the lighter gray corresponds to AT content. The genes highlighted (*) are pseudogenes. *LSC* large single copy, *SSC* small single copy, *IRA/B* inverted repeat A/B. (Color figure online)

separated by a large single copy (LSC) region of 81,767 bp and a small single copy (SSC) region containing 10,974 bp (Table 2). The GC content is 37.5%, which is similar to other angiosperms. It is predicted to encode 109 distinct genes, of which 19 are completely duplicated in the IR regions and three (*matk, trnK-UUU* and *ndhA*) partially duplicated in the IR boundaries resulting in a total of 131 genes (Table 1). The annotation revealed 75 distinct protein-coding genes (six of them completely duplicated and two partially duplicated), 30 distinct tRNAs genes (eight of them completely duplicated and one partially duplicated), and four distinct rRNA genes (all of them completely duplicated). Two pseudogenes, *rpl23* and *ndhF*, were identified due to the presence of internal stops codons. The two introns from *clpP* gene were lost

Table 1 List of genes identified in the plastome of *Linum usitatissimum*

Group of gene	Name of gene
Gene expression machinery	
Ribosomal RNA genes	*rrn16*[b]; *rrn23*[b]; *rrn5*[b]; *rrn4.5*[b]
Transfer RNA genes	*trnA* –UGC[ab]; *trnC* –GCA; *trnD* –GUC; *trnE* –UUC; *trnF* –GAA; *trnfM* –CAU; *trnG* –UCC[a]; *trnG* –GCC; *trnH* –GUG[b]; *trnI* –CAU; *trnI* –GAU[ab]; *trnK* –UUU[ac]; *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
Small subunit of ribosome	*rps2; rps3; rps4; rps7*[b]; *rps8; rps11; rps12*[ac]; *rps14; rps15*[b]; *rps18; rps19*
Large subunit of ribosome	*rpl2*[a]; *rpl14; rpl16*[a]; *rpl20; rpl22; rpl32; rpl33; rpl36*
DNA-dependent RNA polymerase	*rpoA; rpoB; rpoC1*[a]; *rpoC2*
Genes for photosynthesis	
Subunits of photosystem I (PSI)	*psaA; psaB; psaC; psaI; psaJ; ycf3*[a]; *ycf4*
Subunits of photosystem II (PSII)	*psbA*[b]; *psbB; psbC; psbD; psbE; psbF; psbH; psbI; psbJ; psbK; psbL; psbM; psbN; psbT; psbZ*
Subunits of cytochrome b_6f	*petA; petB*[a]; *petD*[a]; *petG; petL; petN*
Subunits of ATP synthase	*atpA; atpB; atpE; atpF*[a]; *atpH; atpI*
Subunits of NADH dehydrogenase	*ndhA*[ac]; *ndhB*[ab]; *ndhC; dhD; ndhE; ndhG; ndhH*[b]; *ndhI; ndhJ; ndhK*
Large subunit of Rubisco	*rbcL*
Others genes	
Maturase	*matK*[c]
Envelope membrane protein	*cemA*
Subunit of acetyl-CoA carboxylase	*accD*
C-type cytochrome synthesis gene	*ccsA*
Protease	*clpP*
Component of TIC complex	*ycf1*[b]
Genes of unknown function	*ycf2*[b]
Pseudogenes	*ndhF; rpl23*

[a]Genes containing introns

[b]Duplicated gene

[c]Partial duplicated genes

Table 2 Comparison of plastomes among six Malpighiales

Species	Family	Size (bp)	LSC (bp)	SSC (bp)	IR (bp)
Linum usitatissimum	Linaceae	156,721	81,767	10,974	31,990
Licania alba	Chrysobalanaceae	162,467	88,927	19,740	26,900
Erythroxylum novogranatense	Erythroxylaceae	163,937	91,383	18,138	27,208
Manihot esculenta	Euphorbiaceae	161,453	89,295	18,250	26,954
Viola seoulensis	Violaceae	156,507	85,691	18,008	26,404
Passiflora edulis	Passifloraceae	151,406	85,720	13,378	26,154
Populus alba	Salicaceae	156,505	84,618	16,567	27,660

in flax as observed in *Passiflora edulis* (Cauz-Santos et al. 2017), which differ from other species of Malpighiales already sequenced. The *rps16* gene is absent in flax plastome as well as observed in other families of Malpighiales, except in the family Euphorbiaceae (Daniell et al. 2008; Rivarola et al. 2011; Tangphatsornruang et al. 2011).

The multiple genome alignment analysis using plastome sequences of Malpighiales and the plastome of *Arabidopsis thaliana* as external reference revealed a rearrangement in the plastome of flax that is not present in other species of this order (Figure S1). For a more detailed exploration of this rearrangement, we compared species belonging to five families of Malpighiales and *A. thaliana* as external reference against flax plastome by dot plot analysis (Fig. 2). A set of genes commonly found in the beginning of LSC in other angiosperms was identified in the IRs in the plastome of flax (Fig. 2; number 1). It indicates an expansion event from IRB to LSC. Another IR expansion event occurred from IRB to

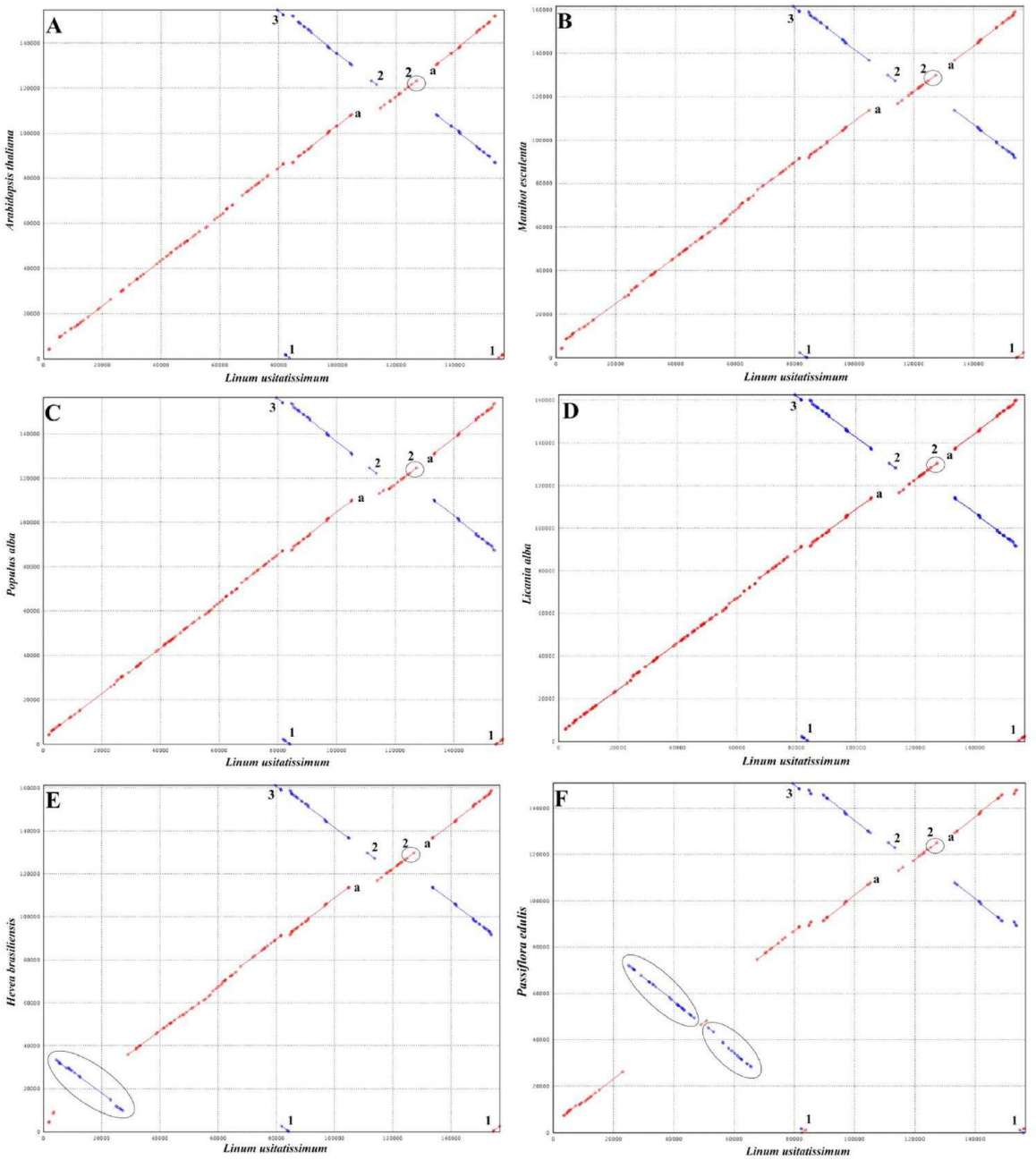


Fig. 2 Dot plot analyses of *Linum usitatissimum* plastome against five Malpighiales species and *Arabidopsis thaliana*. A positive slope denotes that the pair of sequences compared is in the same orientation. A negative slope denotes that the pair of sequences compared can be aligned, but their orientation is opposite. Sequences in the same direction are red and inversions are blue. (1) and (2) indicate a block of genes incorporated in the IRs in *L. usitatissimum* that occurred from a IR expansion event; (3) Indicate a block of genes incorporated in the end of LSC in *L. usitatissimum* that happened from a IR contraction event; **a** highlight a region of gap, which means that no similarity between the pair of genomes in this region was identified. The big circles highlight rearrangements present in *H. brasiliensis* and *P. edulis*

SSC, which shows the presence of a set of genes inside the IRs that is usually found in the end of SSC in other species (Fig. 2, number 2). In addition, a set of genes present usually in the IRs was identified in the end of LSC (Fig. 2, number 3). Low similarity was identified in the region of *ycf1* gene between flax and other Malpighiales (Fig. 2a), indicating that the *ycf1* gene evolved differently and is highly divergent in flax. These rearrangements observed in flax plastome are reported for the first time in Malpighiales, however, other types of rearrangement were already described in *H. brasiliensis* (Tangphatsornruang et al. 2011; Fig. 2e) and *Passiflora* (Cauz-Santos et al. 2017; Fig. 2f), other two species of Malpighiales belonging to families Euphorbiaceae and Passifloraceae, respectively.

The expansion and contraction of IRs, represented in the Fig. 3, highlight the set of genes involved in these events and the implication of them to the size and structure of the plastome. The expansion of the IRB to SSC involved the *ycf1* (partially with part of the coding region), *rps15, ndhH* and *ndhA* (partially with one exon and part of the intron) genes. This event decreased the size of SSC in approximately 7 kb. The other expansion, from IRB to LSC, enwrapped the *trnH*-GUG, *psbA, trnK*-UUU (partially with one exon and part of the intron) and *matK* (partially with part of the coding region) genes. This set of genes has approximately 2 kb. The contraction in the IRA-LSC junction covered the *rps19, rpl2, rpl23* and *trnI*-CAU genes, a genome region of approximately 2 kb.

Repeat sequence analysis

The occurrence, type, and distribution of SSRs in flax plastome were analyzed. In total, 176 SSRs were identified (Table S1). Homopolymers and dipolymers were the most common with, respectively, 70.4 and 19.9% of occurrence. Most homopolymers are constituted by A/T sequences (95.2%). From the dipolymers, 65.7% were also constituted of multiple A and T bases. The sequence, size and location of all SSRs are shown in the Table S2. The highest number of SSRs identified in a coding region was 12, which were found in the *ycf1* gene, being 11 of them homopolymers.

Twenty tandem repeats were identified in flax plastome (Table S3), of which 14 are located in coding regions of *ycf1* (4), *accD* (4), *rps18* (2), *ycf2* (4), and 6 distributed in the intergenic spacers. In addition, 36 direct repeats and 3 inverted repeats (≥ 30 bp) were identified in the plastome of flax (Table S4). Twenty-six directed repeats are located in coding regions, whereas the other ten are distributed in intergenic spacers, introns and pseudogenes. The three pairs of inverted repeats were found in the flax plastome. One pair is located in the intergenic spacer between *psbT* and *psbN*; the second pair is located in the intron of the *ycf3* and *petB*; and, the last one is part of the *trnS*-GCU and *trnS*-GGA.

RNA editing prediction

The RNA-editing sites predicted for plastid genes of flax occur in the first or second codon position and all nucleotide changes observed are from cytidine (C) to uridine (U), as observed in other angiosperms (Takenaka et al. 2013). The program Predictive RNA Editor for Plants found 53 putative RNA-editing sites in 18 genes (Table 3). The genes predicted to have RNA-editing sites are *ndhB* (nine sites) *ndhD* (eight sites), *matK* (six sites), *rpoC2* (five sites), *ndhA* (four sites), *ndhG* (four sites), *rpl20* (three sites), *accD* (two sites), *clpP* (two sites), *rpoA* (two sites), and the genes, *atpB, ccsA, petB, psaI, rpl2, rpoC1, rps2* and *rps14*, with only one site.

A comparison of flax with representatives of the family Chrysobalanaceae (the closest family to flax according to our phylogenetic tree shown below) shows that 22 of these 53 putative RNA-editing sites in flax are also predicted to occur in Chrysobalanaceae, being 18 sites identical to flax (Table 3; note: edited/conserved codon/conserved aa), and four with a different nucleoside in the second or third base position in the codon, although the predicted editing site recovers the conserved amino acid (Table 3; note: edited/non-conserved codon/conserved aa). Another 23 RNA-editing sites predicted in flax have deoxythymidine (T) rather

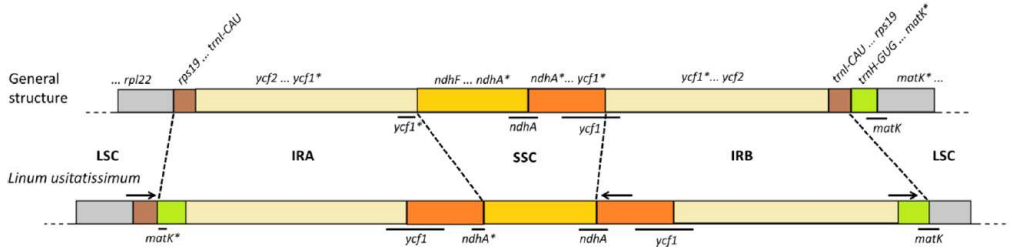


Fig. 3 Comparison of the LSC, IRs and SSC border regions between *Linum usitatissimum* and the general structure presented by other species, including Malpighiales (except *Passiflora cincinnata*) and *Arabidopsis thaliana*. The set of involved genes are color coded. Partial sequences are indicated by asterisk. (Color figure online)

Table 3 List of RNA-editing sites predicted in protein-coding genes of *Linum usitatissimum* plastome using de PREP program

Gene	Nt pos.	AA pos.	Codon change	AA change	Chrysobalanacea						Note
					K. robustus	*P. campestris*	*M. gabunensis*	*D. bellayana*	*C. icaco*	*L. alba*	
accD	1006	350	CCU-UCU	P-S	–	–	–	–	–	–	Gap
	1448	497	GCC-GUC	A-V	GUA (V)	GUA (V)	GUA (V)	GUA (V)	GUA (V)	GUA (V)	Fixed T/non-conserved codon/conserved aa
atpB	22	8	CCC-UCC	P-S	UCU (S)	UCU (S)	UCU (S)	UCU (S)	UCU (S)	UCU (S)	Fixed T/non-conserved codon/conserved aa
ccsA	896	299	ACA-AUA	T-I	AUC (I)	AUC (I)	AUC (I)	AUC (I)	AUC (I)	AUC (I)	Fixed T/non-conserved codon/conserved aa
clpP	335	112	UCG-UUG	S-L	CUU (L)	CUU (L)	CUU (L)	CUU (L)	CUU (L)	CUU (L)	Fixed T/non-conserved codon/conserved aa
	601	201	CUU-UUU	L-F	AUU (I)	UUU (F)	UUU (F)	UUU (F)	UUU (F)	UUU (F)	Fixed T/conserved codon/conserved aa
matK	304	102	CUU-UUU	L-F	UUU (F)	UUU (F)	UUU (F)	UUU (F)	UUU (F)	UUU (F)	Fixed T/conserved codon/conserved aa
	544	182	CUU-UUU	L-F	X	X	X	X	X	X	Edited/conserved codon/conserved aa
	848	283	GCU-GUU	A-V	CUU (L)	CUU (L)	CUU (L)	CUU (L)	CUU (L)	CUU (L)	Non-conserved aa
	896	299	UCA-UUA	S-L	UUG (L)	UUG (L)	UUG (L)	UUG (L)	UUG (L)	UUG (L)	Fixed T/non-conserved codon/conserved aa
	1006	336	CUU-UUU	L-F	UUU (F)	UUU (F)	UUU (F)	UUU (F)	UUU (F)	UUU (F)	Fixed T/conserved codon/conserved aa
	1205	402	UCA-UUA	S-L	X	X	X	X	X	X	Edited/conserved codon/conserved aa

Table 3 (continued)

Gene	Nt pos.	AA pos.	Codon change	AA change	Chrysobalanacea						Note
					K. robustus	*P. campestris*	*M. gabunensis*	*D. bellayana*	*C. icaco*	*L. alba*	
ndhA	341	114	UCA-UUA	S-L	GCA (A)	GCA (A)	GCA (A)	GCA (A)	GCA (A)	GCA (A)	Non-conserved aa
	566	189	UCG-UUG	S-L	UCA-UUA (S-L)	UCA-UUA (S-L)	UCA-UUA (S-L)	UCA-UUA (S-L)	UCA-UUA (S-L)	UCA-UUA (S-L)	Edited/non-conserved codon/conserved aa
	922	308	CUU-UUU	L-F	UUU (F)	UUU (F)	UUU (F)	UUU (F)	UUU (F)	UUU (F)	Fixed T/conserved codon/conserved aa
	1073	358	UCC-UUC	S-F	X	X	X	X	X	X	Edited/conserved codon/conserved aa
ndhB	149	50	UCA-UUA	S-L	X	X	X	X	X	X	Edited/conserved codon/conserved aa
	586	196	CAU-UAU	H-Y	X	X	X	X	X	X	Edited/conserved codon/conserved aa
	611	204	UCG-UUG	S-L	X	X	X	X	X	X	Edited/conserved codon/conserved aa
	737	246	CCA-CUA	P-L	X	X	X	X	X	X	Edited/conserved codon/conserved aa
	746	249	UCU-UUU	S-F	X	X	X	X	X	X	Edited/conserved codon/conserved aa
	830	277	UCA-UUA	S-L	X	X	X	X	X	X	Edited/conserved codon/conserved aa
	836	279	UCA-UUA	S-L	X	X	X	X	X	X	Edited/conserved codon/conserved aa
	1255	419	CAU-UAU	H-Y	X	X	X	X	X	X	Edited/conserved codon/conserved aa
	1481	494	CCA-CUA	P-L	X	X	X	X	X	X	Edited/conserved codon/conserved aa

Table 3 (continued)

Gene	Nt pos.	AA pos.	Codon change	AA change	Chrysobalanacea						Note
					K. robustus	*P. campestris*	*M. gabunensis*	*D. bellayana*	*C. icaco*	*L. alba*	
ndhD	2	1	ACG-AUG	T-M	ATG (M)	X	ATG (M)	ATG (M)	X	X	Fixed T/conserved codon/conserved aa
	67	23	CUU-UUU	L-F	UUU (F)	UUU (F)	UUU (F)	UUU (F)	UUU (F)	UUU (F)	Fixed T/conserved codon/conserved aa
	545	182	GCU-GUU	A-V	X	X	ACU (T)	X	X	X	Edited/conserved codon/conserved aa
	599	200	UCA-UUA	S-L	UUA (L)	UUA (L)	UUA (L)	UUA (L)	UUA (L)	UUA (L)	Fixed T/conserved codon/conserved aa
	878	293	UCA-UUA	S-L	UUA (L)	UUA (L)	UUA (L)	UUA (L)	UUA (L)	UUA (L)	Fixed T/conserved codon/conserved aa
	887	296	CCC-CUC	P-L	CUU (L)	CUU (L)	CUU (L)	CUU (L)	CUU (L)	CUU (L)	Fixed T/non-conserved codon/conserved aa
	1076	359	GCU-GUU	A-V	CUU (L)	CUU (L)	CUU (L)	CUU (L)	CUU (L)	CUU (L)	Non-conserved aa
	1405	469	CUC-UUC	L-F	CUU-UUU (L-F)	CUU-UUU (L-F)	CUU-UUU (L-F)	CUU-UUU (L-F)	CUU-UUU (L-F)	CUU-UUU (L-F)	Edited/non-conserved codon/conserved aa
ndhG	41	14	UCU-UUU	S-F	UUU (F)	UUU (F)	UUU (F)	UUU (F)	UUU (F)	UUU (F)	Fixed T/conserved codon/conserved aa
	166	56	CAC-UAC	H-Y	CAU-UAU (H-Y)	CAU-UAU (H-Y)	CAU-UAU (H-Y)	CAU-UAU (H-Y)	CAU-UAU (H-Y)	CAU-UAU (H-Y)	Edited/non-conserved codon/conserved aa
	314	105	ACA-AUA	T-I	X	X	X	X	X	X	Edited/conserved codon/conserved aa
	413	138	ACA-AUA	T-I	AUA (I)	AUA (I)	AUA (I)	AUA (I)	AUA (I)	AUA (I)	Fixed T/conserved codon/conserved aa
petB	611	204	CCA-CUA	P-L	X	X	X	X	X	X	Edited/conserved codon/conserved aa
psaI	83	28	UCU-UUU	S-F	UUU (F)	UUU (F)	UUU (F)	UUU (F)	UUU (F)	UUU (F)	Fixed T/conserved codon/conserved aa

Table 3 (continued)

Gene	Nt pos.	AA pos.	Codon change	AA change	Chrysobalanacea						Note
					K. robustus	*P. campestris*	*M. gabunensis*	*D. bellayana*	*C. icaco*	*L. alba*	
rpl2	593	198	GCG-GUG	A-V	GUG (V)	GUG (V)	GUG (V)	GUG (V)	GUG (V)	GUG (V)	Fixed T/conserved codon/conserved aa
rpl20	26	9	ACA-AUA	T-I	AUA (I)	AUA (I)	AUA (I)	AUA (I)	AUA (I)	AUA (I)	Fixed T/conserved codon/conserved aa
	59	20	UCU-UUU	S-F	UUC (F)	UUC (F)	UUC (F)	UUC (F)	UUC (F)	UUC (F)	Fixed T/non-conserved codon/conserved aa
	131	44	GCU-GUU	A-V	GUU (V)	GUU (V)	GUU (V)	GUU (V)	GUU (V)	GUU (V)	Fixed T/conserved codon/conserved aa
rpoA	824	275	UCA-UUA	S-L	X	X	X	X	X	X	Edited/conserved codon/conserved aa
	881	294	UCG-UUG	S-L	X	X	X	X	X	X	Edited/conserved codon/conserved aa
rpoC1	62	21	UCG-UUG	S-L	UCA-UUA (S-L)	UCA-UUA (S-L)	UCA-UUA (S-L)	UCA-UUA (S-L)	UCA-UUA (S-L)	UCA-UUA (S-L)	Edited/non-conserved codon/conserved aa
rpoC2	23	8	GCC-GUC	A-V	GUC (V)	GUC (V)	GUC (V)	GUC (V)	GUC (V)	GUC (V)	Fixed T/conserved codon/conserved aa
	1591	531	CUU-UUU	L-F	UUU (F)	UUU (F)	UUU (F)	UUU (F)	UUU (F)	UUU (F)	Fixed T/conserved codon/conserved aa
	2305	769	CGG-UGG	R-W	CCU (P)	CUU (L)	CUU (L)	CUU (L)	CUU (L)	CUU (L)	Non-conserved aa
	3176	1059	UCU-UUU	S-F	AUU (I)	AUU (I)	AUU (I)	AUU (I)	AUU (I)	AUU (I)	Non-conserved aa
	4010	1337	UCC-UUC	S-F	UUC (F)	UUC (F)	UUC (F)	UUC (F)	UUC (F)	UUC (F)	Fixed T/conserved codon/conserved aa
rps2	248	83	UCA-UUA	S-L	X	X	X	X	X	X	Edited/conserved codon/conserved aa
rps14	80	27	CCC-CUC	P-L	UCA-UUA (S-L)	UCA-UUA (S-L)	UCA-UUA (S-L)	UCA-UUA (S-L)	UCA-UUA (S-L)	UCA-UUA (S-L)	Edited/non-conserved codon/conserved aa

RNA editing sites predicted in flax were also investigated into the family Chrysobalanaceae for comparison

than deoxycytidine (C) at the referred sites into Chrysobalanaceae plastomes, and therefore, maintain the conserved amino acid dispensing the need for RNA editing (Table 3; notes: fixed T/conserved codon/conserved aa and fixed T/non-conserved codon/conserved aa). Five RNA editing sites show the tendency to codify non-conserved amino acids between flax and Chrysobalanaceae species (Table 3; note: non-conserved aa). Lastly, the site 350 at the *accD* gene has no comparative sequence among fax and Chrysobalanaceae species since it is located in an specific insertion present in the plastome of flax (Table 3; note: gap).

Synonymous (dS) and nonsynonymous (dN) substitution rates

The pairwise distance based on nucleotide and amino acid sequences was estimated for 69 protein-coding genes from 19 species highlighted in the Table S5, including *L. usitatissium* (Figure S2). These 69 genes include *rpl23* and *ndhF*, both identified in flax as pseudogenes due to the presence of internal stop codons. High nucleotide and amino acid substitution rates were observed in the sequences of *ycf1* and *clpP* genes in the plastome of flax in comparison with other species. The *ycf2, rps11* and *accD* genes also showed a highly significant pairwise distance in comparison with other species of Malpighiales. A high substitution rate was showed by the putative pseudogene *rpl23* if compared with other species, while the pseudogene *ndhF* did not show significant values of substitution, showing a conserved sequence except for the presence of the premature stop codon.

Synonymous (dS), nonsynonymous (dN) substitution rates and dN/dS values were also estimated for the same 69 genes (Fig. 4). The *ycf1* and *clpP* genes showed dN and dS values higher in comparison with other species. However, dN/dS values estimated for *clpP, rps11* and *accD* genes of flax were, respectively, 0.66, 0.24 and 0.63, which indicate that these genes are under negative selection (d*N*/d*S* values below 1). The d*N*/d*S* values estimated for *ycf1* and *ycf2* genes, and for the pseudogene *rpl23* in flax were 1.46, 1.06 and 1.16, respectively. Such values are higher than general means presented by other species that showed 0.68, 0.70 e 0.69, respectively. These dN/dS values suggest that *ycf1* and *ycf2* genes are under positive (d*N*/d*S* values above 1) and neutral (d*N*/d*S* values next to 1) selection, respectively. Significantly different from pseudogene *rpl23*, the pseudogene *ndhF* showed d*N*, d*S* and d*N*/d*S* values similar to a functional *ndhF* found in other species.

Phylogenetic analysis based on plastid genes

The data matrix for phylogenetic analyses included 63 protein-coding genes belonging to 38 taxa (Table S5), including 37 fabids and *Arabidopsis thaliana* as an outgroup species. The aligned sequences included 48,111 nucleotide positions. Bayesian inference (BI) analyses produced a fully resolved phylogenetic tree with a − lnL = 246740.9725. The BI posterior probability values were 100% for 22 nodes, between 80% and 99% for 12 nodes, and 58% for 1 node (Fig. 5).

Two major groups were supported within fabids, the Malpighiales clade sister to the clade composed of other fabids (Rosales, Curcubitales, Fabales and Fagales). Within Malpighiales order, the Erythroxylaceae family was sister to the clade composed of all other families sampled in this analysis. This clade splits into two sister-subclades, a subclade including the families Euphorbiaceae, Violaceae, Salicaceae and Passifloraceae and a subclade including *Linum usitatissimum* (Linaceae) sister to Chrysobalanaceae family. The branch of *L. usitatissimum* (Linaceae) was the most divergent, followed by *Passiflora edulis* (Passiloraceae), within Malpighiales clade.

Discussion

Gene content and pseudogene inference

The plastome of flax encodes 30 tRNA genes and 4 rRNA genes like most angiosperms, including species of Malpighiales order. Two pseudogenes were identified in flax, *rpl23* and *ndhF*, both are reported as pseudogenes for the first time in Malpighiales. Although the *rpl23* gene (encoding the ribosomal protein L23 of the large subunit 50S) is essential for ribosome functionality and cellular viability (Fleischmann et al. 2011), the loss of the plastid *rpl23* gene has already been observed as different evolutionary events in several angiosperm families (Bubunenko et al. 1994; Schmitz-Linneweber et al. 2001; Logacheva et al. 2008; Sloan et al. 2014; Raman and Park 2015; Kim et al. 2016). In some species of Caryophyllales, the plastid *rpl23* gene (coding a prokaryotic-type ribosomal protein L23) was not transferred to nucleus and imported as a functional protein to the plastids, but it was replaced by a eukaryotic-type L23 version encoded by the nucleus and imported into the plastids. We investigated in the EST (expressed sequence tag—NCBI) and WGS (whole-genome shotgun contigs—NCBI) database of flax, using blast as a tool, the presence of similar sequences to the functional *rpl23* gene in the nuclear genome of flax, hypothesizing a possible event of transfer to nucleus. Nonetheless, we did not find evidence for a functional prokaryotic-type L23 protein in flax, suggesting that a similar substitution of the prokaryotic-type- to eukaryotic-type L23 also occurred in flax like in Caryophyllales species.

The *ndhF* gene encodes the F subunit of NAD(P)H dehydrogenase complex, which is involved in the process of chlororespiration and cyclic electron transfer (Matsubayashi et al. 1987; Bock 2007). Several *ndh* genes, including the *ndhF*,

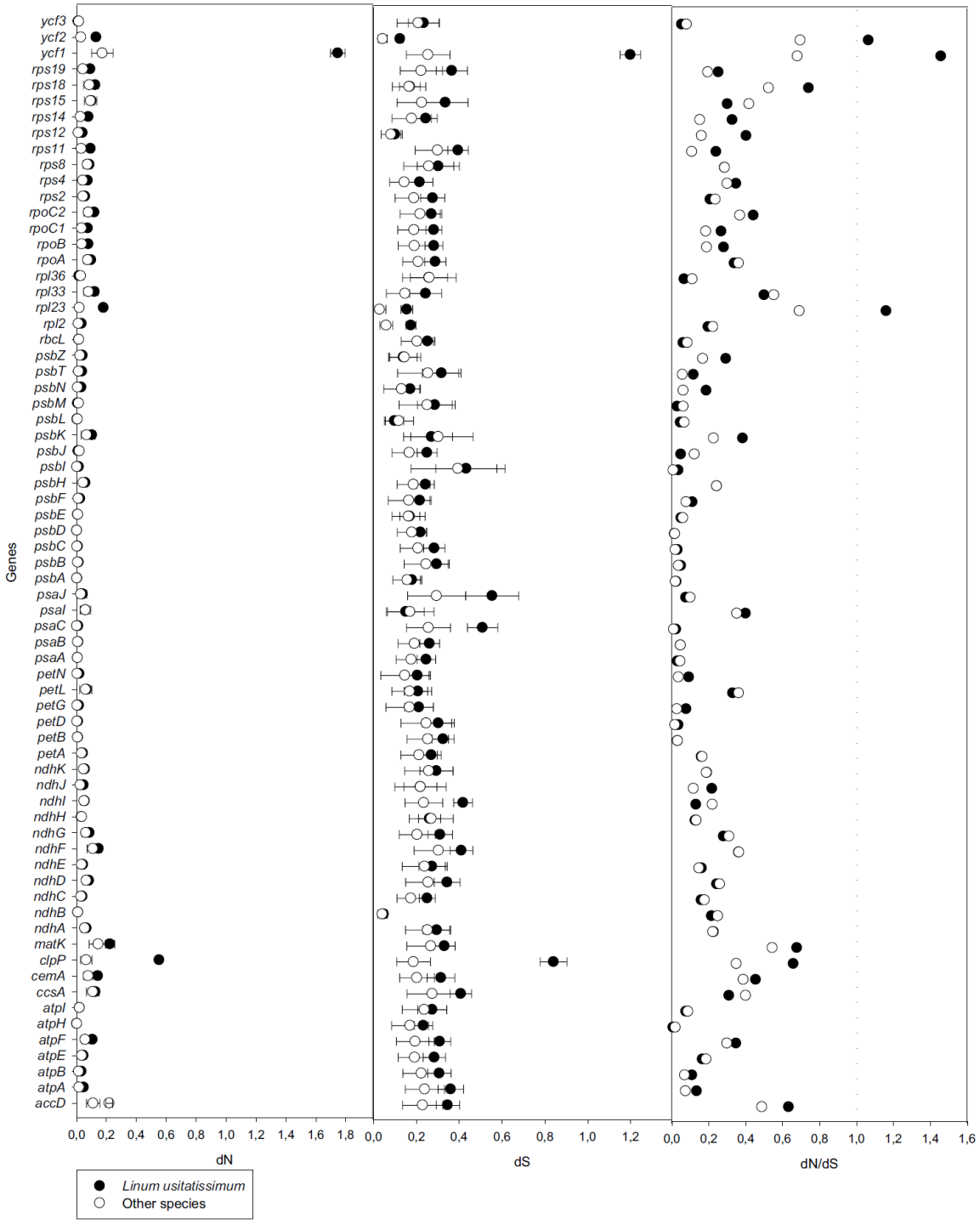


Fig. 4 Synonymous (d*S*), nonsynonymous (d*S*) substitution rates, and d*N*/d*S* values of 69 common plastid protein-coding genes. Flax is represented by black circles and the mean of the other species (see Table S5) is represented by white circles

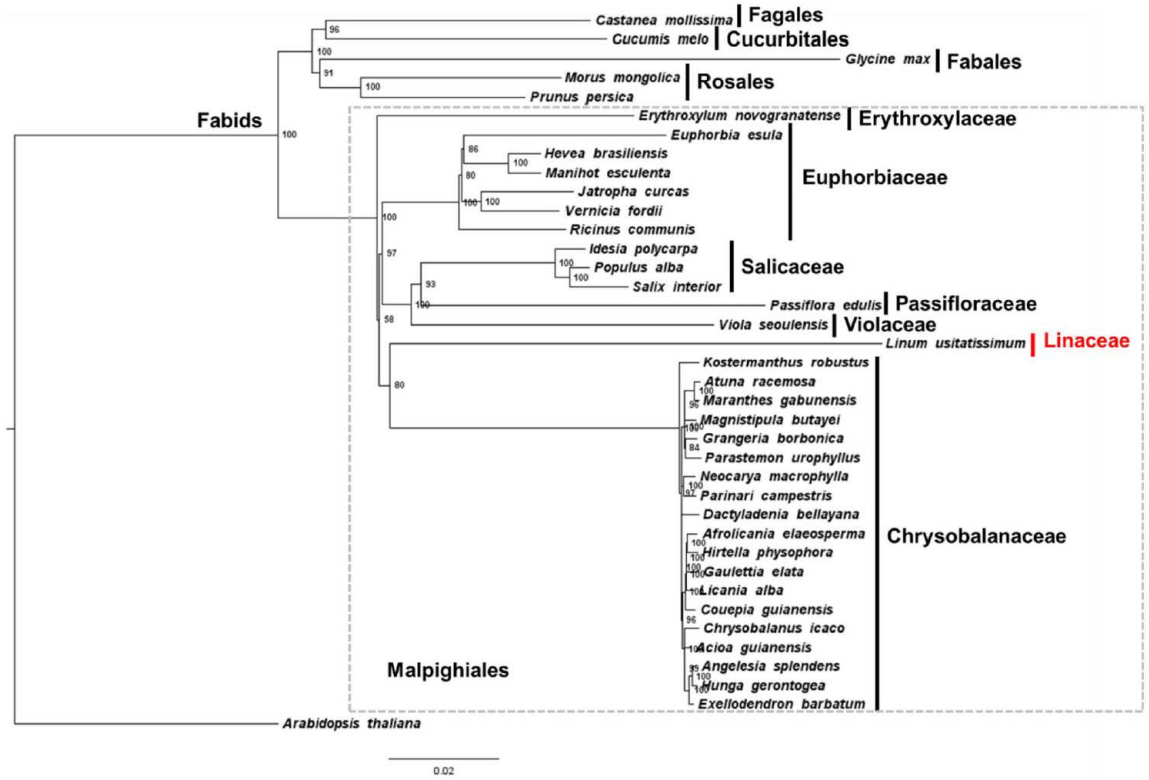


Fig. 5 Phylogenetic tree of 38 taxa based on 63 plastid protein-coding genes using bayesian inference. The numbers at the nodes are Bayesian posterior probabilities (%). The branch length is proportional to the inferred divergence level and the scale bar indicates the number of inferred nucleic acid substitutions per site

were lost or are pseudogenes in orchids (Luo et al. 2014; Chang et al. 2006; Neyland and Urbatsch 1996a), parasitic plants (Wolfe et al. 1992; McNeal et al. 2007), and some lineages of gymnosperms (Wakasugi et al. 1994; Neyland and Urbatsch 1996b; Werner et al. 2009; Xu et al. 2015). Nevertheless, the deletion of ndh genes from the tobacco plastome had no significant phenotypic consequences under standard growth conditions, but resulted in some measurable physiological effects under various stress conditions (Horváth et al. 2000; Li et al. 2004). Blast search in the WGS database of flax using the coding region of *ndhF* gene found two *ndhF*-like sequences in the scaffold1852_1 and scaffold2237_1. These two nuclear sequences of flax showed 99 and 100% of identity, respectively, with the *ndhF* pseudogene present in the flax plastome, including the presence of internal stop codon. This result suggests that the *ndhF* pseudogenization occurred before the transfer of *ndhF* gene from plastome to the nucleus in flax, which would imply that the NAD(P)H dehydrogenase complex is not functional in plastids of flax. Additionally, blast search in the EST database of flax using the coding region of the *ndhF* gene from other Malpighiales found only small sequences with 87–91% of identity and covering 4–32% of the *ndhF* sequence. However, it is very difficult to assume the loss of the NAD(P)H dehydrogenase complex given that the remaining ndh genes are highly conserved in the plastome of flax.

Most part of the intergenic spacer (IGS) between *trnK*-UUU and *trnQ*-UUG genes was lost in flax plastome, including the total loss of the *rps16* gene. The length of this IGS in flax is 292 bp, which is significantly shorter than those found in other Malpighiales that also lost the *rps16* gene, such as *Populus alba* (1210 bp) and *Parinari campestris* (1422 bp). Similar as observed for the *rpl23* gene, the *rps16* gene (encoding the ribosomal protein S16) is essential for ribosome functionality and cellular viability (Fleischmann et al. 2011). However, the *rps16* gene is one of the plastid genes with the highest parallel loss (or pseudogenization) rate and transfer from plastome to the nucleus (Xu et al. 2015). It is very probably that a functional *rps16* gene was transferred to the nucleus in flax and a functional nuclear copy makes the plastid translation viable as previously observed in other species (Guo et al. 2007; Tangphatsornruang et al. 2009).

Other special feature of flax plastome is the loss of both introns from *clpP* gene. The *clpP* gene encodes the catalytic subunit of the plastid protease Clp, which is involved in plastid protein homeostasis being essential for cellular viability (Shikanai et al. 2001; Kuroda and Maliga 2003). In addition to the loss of *clpP* introns in flax, the gene showed an accelerated synonymous and nonsynonymous nucleotide substitution rate (Figs. 4; S2), and a significantly higher d*N*/d*S* ratio if compared with other species analyzed here. However, this ratio is below 1, which means that *clpP* gene of flax is still under purifying selection. The similar situation occurs in members of the families Passifloraceae (Cauz-Santos et al. 2017), Fabaceae (Jansen et al. 2008) and Caryophyllaceae (Erixon and Oxelman 2008; Sloan et al. 2012b, 2014), which also showed a high evolutionary rate of the *clpP* gene. Despite the high divergence of the *clpP* gene in flax, the amino acid residues of the active domains are conserved (Figure S3) suggesting that it may be functional.

The *ycf1* gene of flax also showed a high nucleotide substitution rate, mainly found in nonsynonymous sites, which results in a dN/dS value above 1, suggesting that this gene is under positive selection (Figs. 4; S2). The *ycf1* gene is essential for cell viability (Drescher et al. 2000) and its function was recently characterized as a subunit of the TIC complex (Kikuchi et al. 2013). Despite the high divergence of the *ycf1* gene in flax, the putative Ycf1 protein of flax appears to have the conserved six N-terminal transmembrane domains (TMD1-6; Figure S4) and the two hydrophilic C-terminal domains (TMD7-8; Figure S4), which are largely conserved among the Chlorophyta and land plants (Kikuchi et al. 2013; Nakai 2015). The presence of these conserved domains may suggest that the *ycf1* gene is functional in flax. However, there are reports about loss or pseudogenization of the *ycf1* gene in angiosperms, including the order Poales (Vries et al. 2015) and some Malpighiales (Fajardo et al. 2013; Cauz-Santos et al. 2017). Given the essential function of *ycf1* gene for plant cell, we can deduce that it was functionally transferred to the nucleus in these species. However, it was also suggested that modifications in the plastid protein transport system occurred during the evolution to compensate the lack of Ycf1 protein in some species dispensing the essentiality of this gene (Nakai 2015).

The *ycf2* gene of flax also showed dS and dN rates higher than in other species, as well as dN/dS value next to 1, which may suggest that this gene in flax is under neutral selection as observed for the pseudogene *rpl23* (Figs. 4; S2). Within Malpighiales, the *ycf2* is a pseudogene in *Passiflora* (Cauz-Santos et al. 2017). Furthermore, phylogenetic analyses suggest that the *ycf2* gene was independently lost from the plastome in several lineages during the evolution of angiosperms (Downie et al. 1994). The function of the *ycf2* gene remains to be determined, although it is already known its essentiality for cell survival in tobacco plants (Drescher et al. 2000). However, it is important to note that the similarity among land plant *ycf2* gene is extraordinarily low if compared with other plastid genes, being less than 50% across bryophytes, ferns, and seed plants (Wicke et al. 2011). Using blast search, it was found that the putative Ycf2 protein of flax has 62% of identity and 93% of coverage in comparison with the Ycf2 protein of *Jatropha curcas*, within Malpighiales.

Repeat sequence analysis

The nonrecombinant nature and uniparental inheritance of plastomes in most seed-bearing plants makes them useful tools for evolutionary studies. Plastid microsatellites or SSRs represent useful molecular markers. Several studies have demonstrated the level of diversity and intraspecific variability in plants using of these markers (Provan et al. 2001). It is estimated that 1–2.5% of the plastome is comprised of SSRs, and their number is proportional to genome length (Raubeson et al. 2007). In the plastome of flax 176, SSRs were identified, being 124 monopolymers, which is very close to the number identified by Provan et al. (2001) in other species. Among the monopolymers and dipolymers identified, only two monopolymers and none dipolymer presented more than 15 repeats which is in accordance to the nature of plastid microsatellites (Provan et al. 2001). A large number of SSRs was identified in the *ycf1* (12) and *ycf2* (5) genes, which may be related to the high evolution rate presented by these genes in flax.

In addition to the SSRs, in the flax plastome, we also identified 36 direct repeats and three inverted repeats (≥ 30 bp), which can be considered a low number in comparison with other species of Malpighiales, such as *J. curcas* and *H. brasiliensis* (Asif et al. 2010; Tangphatsornruang et al. 2011). Among the repetitions ≥ 30 bp found in the plastome of flax, 11 of them are small dispersed repeats (SDRs) with 30–46 bp and are located in the intergenic spacers, in the introns (*ndhA, ycf3*, and *petB* genes), and in conserved regions of the *trnS* (–GCU, –UGA, and –GGA), *psaA*, and *psaB* genes. This pattern of distribution and location of SDRs were also reported in plastomes of other angiosperms. The role of these conserved repeats is not known but since many of them are shared broadly and are located in the same regions could suggest that they may have a functional role (Raubeson et al. 2007). In the plastome of flax, other 20 tandem repeats with repetition unit from 9 to 54 bp were identified using the TRF program. Some of them were found within the *rps18*- and *accD*-coding regions, contributing for the increased size of these genes in flax compared to related species.

Some RNA-editing sites in flax seem to be unique of the family Linaceae

Among the 53 predicted sites of flax (Table 3), 21 of them were validated in at least one species among angiosperm species *Jatropha curcas* (Asif et al. 2010), *Arabidopsis thaliana* (Tillich et al. 2005), *Nicotiana tabacum* (Hirose et al. 1999), *Solanum lycopersicum* (Kahlau et al. 2006), *Atropa belladonna* (Schmitz-Linneweber et al. 2002), *Cocos nucifera* (Huang et al. 2013), *Zea mays* (Maier et al. 1995; Bock et al. 1997), and *Oryza sativus* (Corneille et al. 2000; Tsudzuki et al. 2001). We also compared the sites predicted in flax with Chrysobalanaceae species, the family that formed a clade with flax (Fig. 5), to analyze the level of conservation and the evolution of RNA editing sites among related families (Table 3).

All predicted editing sites for the *ndhB* gene of flax occurred at least in five species, except for the codon position 419 that was only reported in two other species (*A. thaliana* and *C. nucifera*). The editing sites in the codon positions 196 and 277 were the most frequent, occurring in all eight species mentioned above. All *ndhB* sites are also highly conserved among Chrysobalanaceae representatives analyzed here. This indicates that the editing sites of the *ndhB* transcripts are well conserved in angiosperms.

The four editing sites predicted for the *ndhA* gene in flax, three of them (codon position 114, 189, and 358) are edited at least in two species, one of them the species *A. belladonna*. The sites 189 and 358 are also predicted to occur in Chrysobalanaceae, but the site 114 is not edited and codes another amino acid, showing that RNA editing can occur in flexible sites.

In the transcripts of the *ndhD* gene in flax, a RNA editing event was predicted to occur at the first codon ACG creating an AUG translational start codon, as previously observed in *A. thaliana, N. tabacum, A. belladonna, S. lycopersicum*, and *C. nucifera*. This editing site occurs in several angiosperms, including some Chrysobalanaceae species (Table 3), and it is suggested to have a role controlling the available pool of mature functional RNA molecules to be translated (Takenaka et al. 2013). Other editing sites predicted in the *ndhD* gene of flax and identified in other species were localized at codon positions 200 (*N. tabacum* and *S. lycopersicum*), 293 (*A. thaliana, A. belladonna, S. lycopersicum, Z. mays* and *O. sativus*), and 296 (*A. thaliana*). All these sites (200, 293, and 296) were lost in Chrysobalanaceae species through the substitution of a C to a T in the plastome.

The *rps14* gene of flax also showed a conserved editing site at the codon position 27, reported in six other species (*A. thaliana, N. tabacum, A. belladonna, C. nucifera, Z. mays* and *O. sativus*), and also predicted in Chrysobalanaceae species that we analyze. The unique site predicted to be edited in the *petB* transcripts was also conserved among Chrysobalanaceae and reported to occur in the species *N. tabacum, A. belladonna, C. nucifera*, and *Z. mays*. Similarly, the predicted editing site for the *rps2* gene of flax was also conserved in Chrysobalanaceae and reported to occur in four species (*J. curcas, N. tabacum, A. belladonna*, and *C. nucifera*). Lastly, RNA editing sites at the codon positions 275 and 21 of the *rpoA* and *rpoC1* genes, respectively, were predicted to occur in flax and Chrysobalanaceae and were validated in *N. tabacum*.

Among the remaining 32 RNA editing sites predicted in flax that we did not found in other species, only seven sites were shared between flax and Chrysobalanaceae. Twenty of them have a T instead C in the plastomes of Chrysobalanaceae species, which indicates that these sites can be unique for Linaceae lineage. The last five sites are in flexible regions that allow different amino acid or indels (*accD*, 350), suggesting that editing in those sites would not be strictly necessary in flax. However, we do not discard the possibility of editing since other studies related to the evolution of RNA editing sites indicate a trend of gain of editing sites in genes or domains of genes whose transitory loss of function can be tolerated (Corneille et al. 2000; Fiebig et al. 2004).

Interestingly, a well-conserved RNA editing site in the *atpF* transcripts is absent in flax. This site is located at nucleotide 92 (second position in the codon 31), where C-U editing occurs to correct a non-synonymous substitution and conserve an essential leucine residue. In flax, the gene *atpF* fixed a T at the nucleotide position 92, maintaining the conserved leucine residue in the protein sequence and dispensing the need of RNA editing. The specific loss of the editing site within the order Malpighiales showed a strong association with the loss of the intron in the gene *atpF* (Daniell et al. 2008). However, in the plastome of flax the *atpF* intron is kept, as well as in representatives of the family Chrysobalanaceae, which also lost the RNA editing site and kept the intron (Figure S5).

Finally, although the *rps7* gene was not identified as potentially edited gene by the PREP-Cp program, the substitution of a traditional start codon (AUG) for an ACG codon, suggests that this gene may be also edited in flax to make it functional.

Expansion and contraction events in the IRs gave rise to unusual IR sizes and gene content

Expansion and contraction events involving few hundreds of base pairs occur frequently at the IR boundaries, mainly including the *ycf1* gene at the IR-SSC junction and *rps19* or *rpl22* genes at the IR-LSC junction. However, large-scale losses or gains at the IR boundaries are rare events and have been reported only in some plant lineages (Goulding et al. 1996; Zhu et al. 2016). Here, we report that the unusual gene

order and gene content in the IRs of flax plastome, result from large-scale expansion and contraction events.

An expansion of the IRs to the SSC-end in the flax plastome incorporated a block of genes of 7 kb including the *ycf1, rps15, ndhH*, and part of *ndhA* genes in the IRs and reduced the SSC length (Fig. 3). A similar expansion was also reported in species of *Pelargonium* (Chumley et al. 2006; Weng et al. 2014). Two species of the genus *Plantago* also showed a SSC highly reduced as a result of a large expansion of the IR region, encompassing the *ycf1, rps15, ndhH, ndhA*, and *ndhI* genes in *P. maritima* and the additional *ndhG, ndhE, psaC*, and *ndhD* genes in *P. media*. The latter contains a SCC with only four genes, *ndhF, rpl32, trnL*-UAG, and *ccsA* (Zhu et al. 2016).

A set of genes found at the LSC end in the plastome of flax, including *rps19, rpl2, rpl23* and *trnI*-CAU genes, are normally located inside the IRs in other angiosperms (Zhu et al. 2016). Additionally, another set of genes, including *trnH*-GUG, *psbA, trnK*-UUU (partial) and *matK* (partial), is inside of the IRs, which is usually located at the begining of LCS in most angiosperms (Zhu et al. 2016). These structural features suggest that different contraction and expansion events occurred in the plastome of flax at the both IR-LSC junctions. Large-scale contraction of IRs was already reported in some highly rearranged plastomes of species of Campanulaceae (Haberle et al. 2008; Zhu et al. 2016). In other cases, like in some species of Geraniaceae and Fabaceae, a massive contraction of the IRs resulted in the complete loss of one of the IR copies (Cai et al. 2008; Guisinger et al. 2011; Sveinsson and Cronk 2014).

Large-scale expansion of the IR to LSC was also reported in species belonging to the genus *Pelargonium* (Chumley et al. 2006; Weng et al. 2014; Zhu et al. 2016), *Berberis* (Ma et al. 2013; Zhu et al. 2016), and within the family Trochodendraceae (Sun et al. 2013; Zhu et al. 2016). However, these events of IR expansion occurred toward the end of LSC region, while in flax the expansion occurred toward the beginning of the LSC region. In some monocots, the *trnH*-GUG (usually the first gene in the LSC region) is located into the IRs, as in *Acorus* and *Asparagus* (Zhu et al. 2016). However, no expansion of the IRs involving genes located after the *trnH*-GUG gene at the beginning of LSC was reported by Zhu et al. (2016), who analyzed the gene content of IRs in 52 families, including angiosperms, gymnosperms, and ferns.

In both genera, *Pelargonium* and *Plantago*, the location of SDRs (small dispersed repeats) at the inversion breakpoints suggest that a same mechanism may be involved in the expansion of IRs through multiple inversions promoted by such SDRs (Zhu et al. 2016; Chumley et al. 2006). The absence of such SDRs at the IR junctions in flax plastome suggests that other mechanism happened in flax, as well as in lineages of expanded IRs such as *Nicotiana acuminata* (Goulding et al. 1996), Berberidaceae (Ma et al. 2013), and Trochodendraceae (Sun et al. 2013). Goulding et al. (1996) proposed a mechanism to explain the expansion of 12 kb in the IRs of *N. acuminata*, involving double-strand DNA break and subsequent recombination between poly(A) tracts in the first intron of *clpP* gene (new IR-LSC junction after expansion) and upstream of *rps19* gene (probable IR-LSC junction before expansion). The analysis of SSRs in the plastome of flax identified several poly(A) tracts that may be involved in the expansion and contraction processes (Table S1). Three poly(A) tracts of 10, 8, and 9 bp are located inside the *rps19* gene. In the intergenic spacer between *ycf2* and *trnH*-GUG genes, a poly(A) tract of 14 bp may have been targeted of recombination with the poly(A) tracts in the *rps19* gene, as a result all genes between these recombination points were lost from the IR and fixed at the LSC end. No homopolymer constituted by A/T sequences was identified next to the IRB-LSC junction inside the *matK* gene, only three poly(A) tracts of 8 bp at the positions 356, 647 and 676 bp counted from the beginning of the LSC inside the *matK* gene (Table S1). Next to IR-SSC junction, inside the intron of *ndhA* gene, a poly(A) tract of 11 bp was also identified which may have been site for recombination with poly(A) tracts located inside the *ycf1* gene, resulting in the expansion of the IRs toward the SSC reported here for the plastome of flax. It is noteworthy to note that the *ycf1* gene of flax contain 11 poly(A) tracts of 8–13 bp.

Based on comparison of gene order inside of the IRs between the plastome of flax and the general structure found in the plastomes of most angiosperms (Fig. 3), we hypothesized that the contraction event occurred before than the expansion event at the IR-LSC junctions. First, the contraction event of the IRs at the LSC-IR_A junction removed the set of *rps19, rpl2, rpl23, trnI*-CAU genes from the IRs (normally found in the IRs in most angiosperms). Posteriorly, the expansion at the LSC-IR_B junction fixed the set of *trnH*-GUG, *psbA, trnK*-UUU, and *matK* genes in the IRs. Thus, the expansion of the IRs followed by homologous recombination events resulted in the unusual gene order constituted of *rps19, rpl2, rpl23, trnI*-CAU, *trnK*-UUU, *matK, psbA, trnH*-GUG, *ycf2* genes, which is found at the LSC-IR_A junction (Figs. 1, 3). The sequencing of other species of the genus *Linum* and other genera within the family Linaceae will be of great importance to trace the evolutionary origin of the contraction and expansion events observed in the plastome of flax. Additionally, it will be important to outline how these events occurred and to elucidate whether such events are specie-, genus- or family-specific.

Malpighiales phylogenetic tree based on conserved plastid genes supports the position of flax (Linaceae) closed to Chrysobalanaceae family

The phylogenetic tree based on 63 conserved plastid protein-coding genes using bayesian inference (Fig. 5) supports the monophyletic origin of the order Malpighiales as reported in other phylogenetic approaches using nuclear, mitochondrial, and plastidial sequences (Davis et al. 2005; Wurdack and Davis 2009; Xi et al. 2012). The relationships found within the order Malpighilaes among the families Chrysobalanaceae, Euphorbiaceae, Erythroxylaceae, Violaceae, Passifloraceae, and Salicaceae is in accordance with the most recent and well-resolved phylogeny of Malpighiales (Xi et al. 2012), which used 82 plastid genes from 58 species. However, according to Xi et al. (2012) the linoid clade (including the family Linaceae) is included within the euphorbioid clade, along with the families Euphorbiaceae, Peraceae, Phyllanthaceae, and Picrodendraceae. The euphorbioid clade was supported with 64% maximum likelihood bootstrap and 61% Bayesian posterior probability. Nevertheless, the phylogeny inferred here within Malpighiales, *L. usitatissimum* (Linaceae) is supported with 80% Bayesian posterior probability as sister to the family Chrysobalanaceae, differing clearly of the inference previously determined by Xi et al. (2012).

Conclusion

This study reported the complete plastome of flax (156,721 bp). The plastome of flax revealed unique features combining events of contraction and expansion of the IRs, which alter the size, gene content, and gene order of LSC, SSC and IR regions. The first event to occur, the contraction, moved the *rps19*/*rpl2*/*rpl23*/*trnI-CAU* genes from the IRs to the LSC. The second event, the expansion, shifted the *trnH-GUG, psbA, trnK-UUU*(partial), *matK*(partial) genes from LSC, and the *ycf1* (partial), *rps15, ndhH*, and *ndhA* (partial) genes from SSC to the IRs. The expansion changed significantly the copy number of several plastid genes. The plastome of flax also shows 32 new putative RNA editing sites, which brings new insights about loss and gain of editing sites during the evolution of angiosperms. In addition, the presence of pseudogenes (*rpl23* and *ndhF*), loss of gene (*rps16*), loss of introns (*clpP*), and the presence of high divergent genes (*ycf1, ycf2*, and *clpP*) indicate that the plastome of flax demonstrated exceptional mode of evolution with specific characteristics. As the flax is the first species within the family Linaceae to have the plastome fully sequenced and characterized in detail, such unusual features suggest the family Linaceae as an interesting lineage to study evolutionary traits in plastids. Moreover, the morphological and ecological richness of the family Linaceae raise questions about the interactions between plastome evolution and adaptation to different environmental conditions. The sequencing of other species of Linaceae will be necessary to address these questions. Furthermore, we described 176 SSR markers which can be used for different approaches related to population genetics, conservation, genetic divergence and germplasm screening. Finally, our phylogenetic analysis based on concatenated plastid genes indicates Linaceae as sister to Chrysobalanaceae, differing from previous phylogenies that position Linaceae within the euphorbioid clade. The findings showed here have important implications in the areas of genetic, evolution, conservation, breeding and biotechnology of Linaceae species.

Author contribution statement ASL, TGP, LNV, MPG, RON, EMS, FOP, and MR conceived and designed the research. ASL, TGP, KGS, LNV, EMS, FOP, and MR conducted experiments and analyzed the data. MPG, RON, EMS, FOP, and MR contributed with reagents and materials. ASL and MR wrote the manuscript. All authors read and approved the manuscript.

Acknowledgements This research was support by the National Council for Scientific and Technological Development, Brazil (CNPq, Grant 459698/2014-1). We are grateful to INCT-FBN and for the scholarships granted by the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES) to TGP and LNV, and those granted by the CNPq to ASL, KGS, MPG, RON, EMS and FOP.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Alkatib S, Scharff LB, Rogalski M, Fleischmann TT, Matthes A, Seeger S et al (2012) The conributions of wobbling and superwobbling to the reading of the genetic code. PLoS Genet 8:e1003076. doi:10.1371/journal.pgen.1003076

Asif MH, Mantri SS, Sharma A, Srivastava A, Trivedi I, Gupta P, Mohanty CS, Sawant SV, Tuli R (2010) Complete sequence and organisation of the *Jatropha curcas* (Euphorbiaceae) chloroplast genome. Tree Genet Genomes 6:941–952. doi:10.1007/s11295-010-0303-0

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573–580. doi:10.1093/nar/27.2.573

Besnard G, Hernández P, Khadari B, Dorado G, Savolainen V (2011) Genomic profiling of plastid DNA variation in the Mediterranean olive tree. BMC Plant Biol 11:1. doi:10.1186/1471-2229-11-80

Bock R (2007) Structure, function, and inheritance of plastid genomes. In: Bock R (ed) Cell and molecular biology of plastids. Springer, Berlin, pp 29–63

Bock R (2015) Engineering plastid genomes: methods, tools, and applications in basic research and biotechnology. Annu Rev Plant Biol 66:211–241. doi:10.1146/annurev-arplant-050213-040212

Bock R, Albertazzi F, Freyer R, Fuchs M, Ruf S, Zeltz P, Maier RM (1997) Transcript editing in chloroplasts of higher plants. In: Schenk HEA, Herrmann RG, Jeon KW, Müller NE, Schwemmler W (eds) Eukaryotism and symbiosis. Springer, Berlin, pp 123–137. doi:10.1007/978-3-642-60885-8_9

Bubunenko MG, Schmidt J, Subramanian AR (1994) Protein substitution in chloroplast ribosome evolution. A eukaryotic cytosolic protein has replaced its organelle homologue (L23) in spinach. J Mol Biol 240:28–41. doi:10.1006/jmbi.1994.1415

Cai Z, Guisinger M, Kim HG, Ruck E, Blazier JC, McMurtry V, Kuehl JV, Boore J, Jansen RK (2008) Extensive reorganization of the plastid genome of *Trifolium subterraneum* (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions. J Mol Evol 67:696–704. doi:10.1007/s00239-008-9180-7

Carlsson AS (2009) Plant oils as feedstock alternatives to petroleum—a short survey of potential oil crop platforms. Biochimie 91:665–670. doi:10.1016/j.biochi.2009.03.021

Cauz-Santos LA, Munhoz CF, Rodde N, Cauet S, Santos AA, Penha HA, Dornelas MC, Varani AM, Oliveira GXO, Bergès H, Vieira MLC (2017) The chloroplast genome of *Passiflora edulis* (Passifloraceae) assembled from long sequence reads: structural organization and phylogenomic studies in Malpighiales. Front Plant Sci 8:334. doi:10.3389/fpls.2017.00334

Chang CC, Lin HC, Lin IP et al (2006) The chloroplast genome of *Phalaenopsis aphrodite* (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. Mol Biol Evol 23:279–291. doi:10.1093/molbev/msj029

Cheon KS, Yang JC, Kim KA, Jang SK, Yoo KO (2015) The first complete chloroplast genome sequence from Violaceae (*Viola seoulensis*). Mitochondr DNA 1:67–68. doi:10.3109/19401736.2015.1110801

Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, Boore JL, Jansen RK (2006) The complete chloroplast genome sequence of Pelargonium × hortorum: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. Mol Biol Evol 23:2175–2190. doi:10.1093/molbev/msl089

Corneille S, Lutz K, Maliga P (2000) Conservation of RNA editing between rice and maize plastids: are most editing events dispensable? Mol Gen Genet MGG 264:419–424. doi:10.1007/s004380000295

Daniell H, Wurdack KJ, Kanagaraj A, Lee SB, Saski C, Jansen RK (2008) The complete nucleotide sequence of the cassava (*Manihot esculenta*) chloroplast genome and the evolution of atpF in Malpighiales: RNA editing and multiple losses of a group II intron. Theor Appl Genet 116:723–737. doi:10.1007/s00122-007-0706-y

Daniell H, Lin CS, Yu M, Chang WJ (2016) Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol 17:134. doi:10.1186/s13059-016-1004-2

Darling AC, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14:1394–1403. doi:10.1101/gr.2289704

Davis CC, Webb CO, Wurdack KJ, Jaramillo CA, Donoghue MJ (2005) Explosive radiation of Malpighiales supports a mid-cretaceous origin of modern tropical rain forests. Am Nat 165:E36-65. doi:10.1086/428296

Dexter KG, Terborgh JW, Cunningham CW (2012) Historical effects on beta diversity and community assembly in Amazonian trees. Proc Natl Acad Sci USA 109:7787–7792. doi:10.1073/pnas.1203523109

Downie SR, Katz-Downie DS, Wolfe KH, Calie PJ, Palmer JD (1994) Structure and evolution of the largest chloroplast gene (ORF2280): internal plasticity and multiple gene loss during angiosperm evolution. Curr Genet 25:367–378. doi:10.1007/BF00351492

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

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. doi:10.1093/nar/gkh340

Erixon P, Oxelman B (2008) Whole-gene positive selection, elevated synonymous substitution rates, duplication, and indel evolution of the chloroplast clpP1 gene. PloS One 3:e1386. doi:10.1371/journal.pone.0001386

Fajardo D, Senalik D, Ames M et al (2013) Complete plastid genome sequence of Vaccinium macrocarpon: structure, gene content, and rearrangements revealed by next generation sequencing. Tree Genet Genomes 9:489–498. doi:10.1007/s11295-012-0573-9

Fiebig A, Stegemann S, Bock R (2004) Rapid evolution of RNA editing sites in a small non-essential plastid gene. Nucleic Acids Res 32:3615–3622. doi:10.1093/nar/gkh695

Fleischmann TT, Scharff LB, Alkatib S, Hasdorf S, Schottler MA, Bock R (2011) Nonessential plastid-encoded ribosomal proteins in tobacco: a developmental role for plastid translation and implications for reductive genome evolution. Plant Cell 23:3137–3155. doi:10.1105/tpc.111.088906

Goulding SE, Olmstead RG, Morden CW, Wolfe KH (1996) Ebb and flow of the chloroplast inverted repeat. Mol Gen Genet MGG 252:195–206. doi:10.1007/BF02173220

Guisinger MM, Kuehl JV, Boore JL, Jansen RK (2011) Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. Mol Biol Evol 28:583–600. doi:10.1093/molbev/msq229

Guo X, Castillo-Ramírez S, González V, Bustos P, Fernández-Vázquez JL, Santamaría RI, Arellano J, Cevallos MA, Dávila G (2007) Rapid evolutionary change of common bean (*Phaseolus vulgaris* L) plastome, and the genomic diversification of legume chloroplasts. BMC Genom 8:228. doi:10.1186/1471-2164-8-228

Haberle RC, Fourcade HM, Boore JL, Jansen RK (2008) Extensive rearrangements in the chloroplast genome of *Trachelium caeruleum* are associated with repeats and tRNA genes. J Mol Evol 66:350–361. doi:10.1007/s00239-008-9086-4

Hirose T, Kusumegi T, Tsudzuki T, Sugiura M (1999) RNA editing sites in tobacco chloroplast transcripts: editing as a possible regulator of chloroplast RNA polymerase activity. Mol Gen Genet MGG 262:462–467

Horváth EM, Peter SO, Joel T, Rumeau D, Cournac L, Horváth G, Kavanagh TA, Schaefer C, Medgyesy P (2000) Targeted inactivation of the plastid ndhB gene in tobacco results in an enhanced sensitivity of photosynthesis to moderate stomatal closure. Plant Physiol 123:1337–1349

Huang YY, Matzke AJM, Matzke M (2013) Complete sequence and comparative analysis of the chloroplast genome of coconut palm (*Cocos nucifera*). PLoS One 8:e74736. doi:10.1371/journal.pone.0074736

Jansen RK, Wojciechowski MF, Sanniyasi E, Lee SB, Daniell H (2008) Complete plastid genome sequence of the chickpea (*Cicer arietinum*) and the phylogenetic distribution of rps12 and clpP intron losses among legumes (Leguminosae). Mol Phylogenet Evol 48:1204–1217. doi:10.1016/j.ympev.2008.06.013

Jansen RK, Saski C, Lee SB, Hansen AK, Daniell H (2011) Complete plastid genome sequences of three Rosids (Castanea, Prunus, Theobroma): evidence for at least two independent transfers of rpl22 to the nucleus. Mol Biol Evol 28:835–847. doi:10.1093/molbev/msq261

Kahlau S, Aspinall S, Gray JC, Bock R (2006) Sequence of the tomato chloroplast DNA and evolutionary comparison of solanaceous plastid genomes. J Mol Evol 63:194–207. doi:10.1007/s00239-005-0254-5

Kikuchi S, Bédard J, Hirano M, Hirabayashi Y, Oishi M, Imai M, Takase M, Ide T, Nakai M (2013) Uncovering the protein translocon at the chloroplast inner envelope membrane. Science 339:571–574. doi:10.1126/science.1229262

Kim HT, Kim JS, Kim JH (2016) The complete plastid genome sequence of *Eustrephus latifolius* (Asparagaceae: Lomandroideae). Mitochondr DNA Part DNA Mapp Seq Anal 27:1549–1551. doi:10.3109/19401736.2014.953132

Kuroda H, Maliga P (2003) The plastid clpP1 protease gene is essential for plant development. Nature 425:86–89. doi:10.1038/nature01909

Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29:4633–4642. doi:10.1093/nar/29.22.4633

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. Genome Biol 5(2):R12. doi:10.1186/gb-2004-5-2-r12

Kvavadze E, Bar-Yosef O, Belfer-Cohen A, Boaretto E, Jakeli N, Matskevich Z, Meshveliani T (2009) 30,000-year-old wild flax fibers. Science 325:1359. doi:10.1126/science.1175404

Lanfear R, Calcott B, Ho SYW, Guindon S (2012) PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol Biol Evol 29:1695–1701. doi:10.1093/molbev/mss020

Li XG, Duan W, Meng QW, Zou Q, Zhao SJ (2004) The function of chloroplastic NAD(P)H dehydrogenase in tobacco during chilling stress under low irradiance. Plant Cell Physiol 45:103–108. doi:10.1093/pcp/pch011

Logacheva MD, Samigullin TH, Dhingra A, Penin AA (2008) Comparative chloroplast genomics and phylogenetics of *Fagopyrum esculentum* ssp. ancestrale—a wild ancestor of cultivated buckwheat. BMC Plant Biol 8:59. doi:10.1186/1471-2229-8-59

Lohse M, Drechsel O, Kahlau S, Bock R (2013) OrganellarGenomeDRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res 41:W575-W581. doi:10.1093/nar/gkt289

López P, Tremetsberger K, Kohl G, Stuessy T (2012) Progenitor-derivative speciation in Pozoa (Apiaceae, Azorelloideae) of the southern Andes. Ann Bot 109:351–363. doi:10.1093/aob/mcr291

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964

Luo J, Hou B-W, Niu Z-T, Liu W, Xue Q-Y, Ding X-Y (2014) Comparative chloroplast genomes of photosynthetic orchids: insights into evolution of the Orchidaceae and development of molecular markers for phylogenetic applications. PloS One 9:e99016. doi:10.1371/journal.pone.0099016

Ma J, Yang B, Zhu W, Sun L, Tian J, Wang X (2013) The complete chloroplast genome sequence of *Mahonia bealei* (Berberidaceae) reveals a significant expansion of the inverted repeat and phylogenetic relationship with other angiosperms. Gene 528:120–131. doi:10.1016/j.gene.2013.07.037

Maier RM, Neckermann K, Igloi GL, Kössel H (1995) Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. J Mol Biol 251:614–628. doi:10.1006/jmbi.1995.0460

Malé PJG, Bardon L, Besnard G, Coissac E, Delsuc F, Engel J, Lhuillier E, Scotti-Saintagne C, Tinaut A, Chave J (2014) Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. Mol Ecol Resour 14(5):966–975. doi:10.1111/1755-0998.12246

Matsubayashi T, Wakasugi T, Shinozaki K, Yamaguchi-Shinozaki K, Zaita N, Hidaka T, Meng BY, Ohto C, Tanaka M, Kato A (1987) Six chloroplast genes (ndhA-F) homologous to human mitochondrial genes encoding components of the respiratory chain NADH dehydrogenase are actively expressed: determination of the splice sites in ndhA and ndhB pre-mRNAs. Mol Gen Genet MGG 210:385–393

McDill J, Repplinger M, Simpson BB, Kadereit JW (2009) The phylogeny of Linum and Linaceae subfamily Linoideae, with implications for their systematics, biogeography, and evolution of heterostyly. Syst Bot 34:386–405. doi:10.1600/036364409788606244

McNeal JR, Kuehl JV, Boore JL, de Pamphilis CW (2007) Complete plastid genome sequences suggest strong selection for retention of photosynthetic genes in the parasitic plant genus Cuscuta. BMC Plant Biol 7:57. doi:10.1186/1471-2229-7-57

Mower JP (2009) The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. Nucleic Acids Res 37:W253–W259. doi:10.1093/nar/gkp337

Nakai M (2015) The TIC complex uncovered: the alternative view on the molecular mechanism of protein translocation across the inner envelope membrane of chloroplasts. Biochim Biophys Acta BBA Bioenerg 1847:957–967. doi:10.1016/j.bbabio.2015.02.011

Neyland R, Urbatsch LE (1996a) Phylogeny of subfamily Epidendroideae (Orchidaceae) inferred from ndhF chloroplast gene sequences. Am J Bot 83:1195–1206

Neyland R, Urbatsch LE (1996b) The ndhF chloroplast gene detected in all vascular plant divisions. Planta 200:273–277

Provan J, Powell W, Hollingsworth PM (2001) Chloroplast microsatellites: new tools for studies in plant ecology and evolution. Trends Ecol Evol 16:142–147. doi:10.1016/S0169-5347(00)02097-8

Qiao J, Cai M, Yan G, Wang N, Li F, Chen B, Gao G, Xu K, Li J, Wu X (2016) High-throughput multiplex cpDNA resequencing clarifies the genetic diversity and genetic relationships among *Brassica napus, Brassica rapa* and *Brassica oleracea*. Plant Biotechnol J 14:409–418. doi:10.1111/pbi.12395

Raman G, Park S (2015) Analysis of the complete chloroplast genome of a medicinal plant, *Dianthus superbus* var. longicalycinus, from a comparative genomics perspective. PLOS One 10:e0141329. doi:10.1371/journal.pone.0141329

Raubeson LA, Peery R, Chumley TW, Dziubek C, Fourcade HM, Boore JL, Jansen RK (2007) Comparative chloroplast genomics: analyses including new sequences from the angiosperms *Nuphar advena* and *Ranunculus macranthus*. BMC Genom 8:174. doi:10.1186/1471-2164-8-174

Rivarola M, Foster JT, Chan AP et al (2011) Castor bean organelle genome sequencing and worldwide genetic diversity analysis. PLoS One 6:e21743. doi:10.1371/journal.pone.0021743

Rogalski M, Carrer H (2011) Engineering plastid fatty acid biosynthesis to improve food quality and biofuel production in higher plants. Plant Biotechnol J 9:554–564. doi:10.1111/j.1467-7652.2011.00621.x

Rogalski M, Ruf S, Bock R (2006) Tobacco plastid ribosomal protein S18 is essential for cell survival. Nucleic Acids Res 34:4537–4545. doi:10.1093/nar/gkl634

Rogalski M, Vieira LN, Fraga HP, Guerra MP (2015) Plastid genomics in horticultural species: importance and applications for plant population genetics, evolution, and biotechnology. Front Plant Sci 6:586. doi:10.3389/fpls.2015.00586

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP (2012) MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61:539–542. doi:10.1093/sysbio/sys029

Schmitz-Linneweber C, Maier RM, Alcaraz JP, Cottet A, Herrmann RG, Mache R (2001) The plastid chromosome of spinach (*Spinacia oleracea*): complete nucleotide sequence and gene organization. Plant Mol Biol 45:307–315. doi:10.1023/A:1006478403810

Schmitz-Linneweber C, Regel R, Du TG, Hupfer H, Herrmann RG, Maier RM (2002) The plastid chromosome of *Atropa belladonna* and its comparison with that of *Nicotiana tabacum*: the role of RNA editing in generating divergence in the process of plant speciation. Mol Biol Evol 19:1602–1612. doi:10.1093/oxfordjournals.molbev.a004222

Shikanai T, Shimizu K, Ueda K, Nishimura Y, Kuroiwa T, Hashimoto T (2001) The chloroplast clpP gene, encoding a proteolytic subunit of ATP-dependent protease, is indispensable for chloroplast development in tobacco. Plant Cell Physiol 42:264–273

Simmons CA, Turk P, Beamer S, Jaczynski J, Semmens K, Matak KE (2011) The effect of a flaxseed oil-enhanced diet on the product quality of farmed brook trout (*Salvelinus fontinalis*) fillets. J Food Sci 76:S192-197. doi:10.1111/j.1750-3841.2011.02070.x

Singh KK, Mridula D, Rehal J, Barnwal P (2011) Flaxseed: a potential source of food, feed and fiber. Crit Rev Food Sci Nutr 51:210–222. doi:10.1080/10408390903537241

Sloan DB, Alverson AJ, Wu M, Palmer JD, Taylor DR (2012) Recent acceleration of plastid sequence and structural evolution coincides with extreme mitochondrial divergence in the angiosperm genus Silene. Genome Biol Evol 4:294–306. doi:10.1093/gbe/evs006

Sloan DB, Triant DA, Forrester NJ, Bergner LM, Wu M, Taylor DR (2014) A recurring syndrome of accelerated plastid genome evolution in the angiosperm tribe Sileneae (Caryophyllaceae). Mol Phylogenet Evol 72:82–89. doi:10.1016/j.ympev.2013.12.004

Sun Y, Moore MJ, Meng A, Soltis PS, Soltis DE, Li J, Wang H (2013) Complete plastid genome sequencing of Trochodendraceae reveals a significant expansion of the inverted repeat and suggests a Paleogene divergence between the two extant species. PLoS One 8:e60429. doi:10.1371/journal.pone.0060429

Sveinsson S, Cronk Q (2014) Evolutionary origin of highly repetitive plastid genomes within the clover genus (Trifolium). BMC Evol Biol 14:228. doi:10.1186/s12862-014-0228-6

Takenaka M, Zehrmann A, Verbitskiy D, Härtel B, Brennicke A (2013) RNA editing in plants and its evolution. Annu Rev Genet 47:335–352. doi:10.1146/annurev-genet-111212-133519

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30:2725–2729. doi:10.1093/molbev/mst197

Tangphatsornruang S, Somta P, Uthaipaisanwong P, Chanprasert J, Sangsrakru D, Seehalak W, Sommanas W, Tragoonrung S, Srinives P (2009) Characterization of microsatellites and gene contents from genome shotgun sequences of mungbean (*Vigna radiata* (L.) Wilczek). BMC Plant Biol 9(1):137. doi:10.1186/1471-2229-9-137

Tangphatsornruang S, Uthaipaisanwong P, Sangsrakru D, Chanprasert J, Yoocha T, Jomchai N, Tragoonrung S (2011) Characterization of the complete chloroplast genome of *Hevea brasiliensis* reveals genome rearrangement, RNA editing sites and phylogenetic relationships. Gene 475:104–112. doi:10.1016/j.gene.2011.01.002

Thiel T, Michalek W, Varshney R, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor Appl Genet 106:411–422. doi:10.1007/s00122-002-1031-0

Tillich M, Funk HT, Schmitz-Linneweber C, Poltnigg P, Sabater B, Martin M, Maier RM (2005) Editing of plastid RNA in *Arabidopsis thaliana* ecotypes. Plant J 43:708–715. doi:10.1111/j.1365-313X.2005.02484.x

Touré A, Xueming X (2010) Flaxseed lignans: source, biosynthesis, metabolism, antioxidant activity, bio-active components, and health benefits. Compr Rev Food Sci Food Saf 9:261–269. doi:10.1111/j.1541-4337.2009.00105.x

Tsudzuki T, Wakasugi T, Sugiura M (2001) Comparative analysis of RNA editing sites in higher plant chloroplasts. J Mol Evol 53:327–332. doi:10.1007/s002390010222

Vaidya G, Lohman DJ, Meier R (2011) SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information. Cladistics 27:171–180. doi:10.1111/j.1096-0031.2010.00329.x

Vieira LN, Faoro H, Rogalski M, Fraga HPF, Cardoso RLA, de Souza EM, Pedrosa FO, Nodari RO, Guerra MP (2014a) The complete chloroplast genome sequence of *Podocarpus lambertii*: genome structure, evolutionary aspects, gene content and SSR detection. PLoS One 9:e90618. doi:10.1371/journal.pone.0090618

Vieira LN, Faoro H, Fraga HPF, Rogalski M. de Souza EM, Pedrosa FO, Nodari RO, Guerra MP (2014b) An improved protocol for intact chloroplasts and cpDNA isolation in conifers. PLoS One 9:e84792. doi:10.1371/journal.pone.0084792

Vieira LN, Dos Anjos KG, Faoro H, Fraga HP, Greco TM, Pedrosa FO, de Souza EM, Rogalski M, de Souza RF, Guerra MP (2016a) Phylogenetic inference and SSR characterization of tropical woody bamboos tribe Bambuseae (Poaceae: Bambusoideae) based on complete plastid genome sequences. Curr Genet 62(2):443–453. doi:10.1007/s00294-015-0549-z

Vieira LN, Rogalski R, Faoro H, Fraga HPF, dos Anjos KG, Picchi GFA, Nodari RO, Pedrosa FO, de Souza EM, Guerra MP (2016b) The plastome sequence of the endemic Amazonian conifer, *Retrophyllum piresii* (Silba) C.N.Page, reveals different recombination events and plastome isoforms. Tree Genet Genomes 12(10):1–11. doi:10.1007/s11295-016-0968-0

Vries J, Sousa FL, Bölter B, Soll J, Gould SB (2015) YCF1: a green TIC? Plant Cell 27:1827–1833. doi:10.1105/tpc.114.135541

Wakasugi T, Tsudzuki J, Ito S, Nakashima K, Tsudzuki T, Sugiura M (1994) Loss of all ndh genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. Proc Natl Acad Sci USA 91:9794–9798

Weng ML, Blazier JC, Govindu M, Jansen RK (2014) Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. Mol Biol Evol 31:645–659. doi:10.1093/molbev/mst257

Werner T, Braukmann A, Kuzmina M, Stefanovic S (2009) Loss of all plastid ndh genes in Gnetales and conifers: extent and evolutionary significance for the seed plant phylogeny. Curr Genet 55:323–337. doi:10.1007/s00294-009-0249-7

Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, Quandt D (2011) The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol 76:273–297. doi:10.1007/s11103-011-9762-4

Wolfe KH, Morden CW, Palmer JD (1992) Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. Proc Natl Acad Sci USA 89:10648–10652

Wu Z (2016) The whole chloroplast genome of shrub willows (*Salix suchowensis*). Mitochondr DNA Part A 27:2153–2154. doi:10.3109/19401736.2014.982602

Wurdack KJ, Davis CC (2009) Malpighiales phylogenetics: gaining ground on one of the most recalcitrant clades in the angiosperm tree of life. Am J Bot 96:1551–1570. doi:10.3732/ajb.0800207

Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20:3252–3255. doi:10.1093/bioinformatics/bth352

Xi Z, Ruhfel BR, Schaefer H et al (2012) Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. Proc Natl Acad Sci USA 109:17519–17524. doi:10.1073/pnas.1205818109

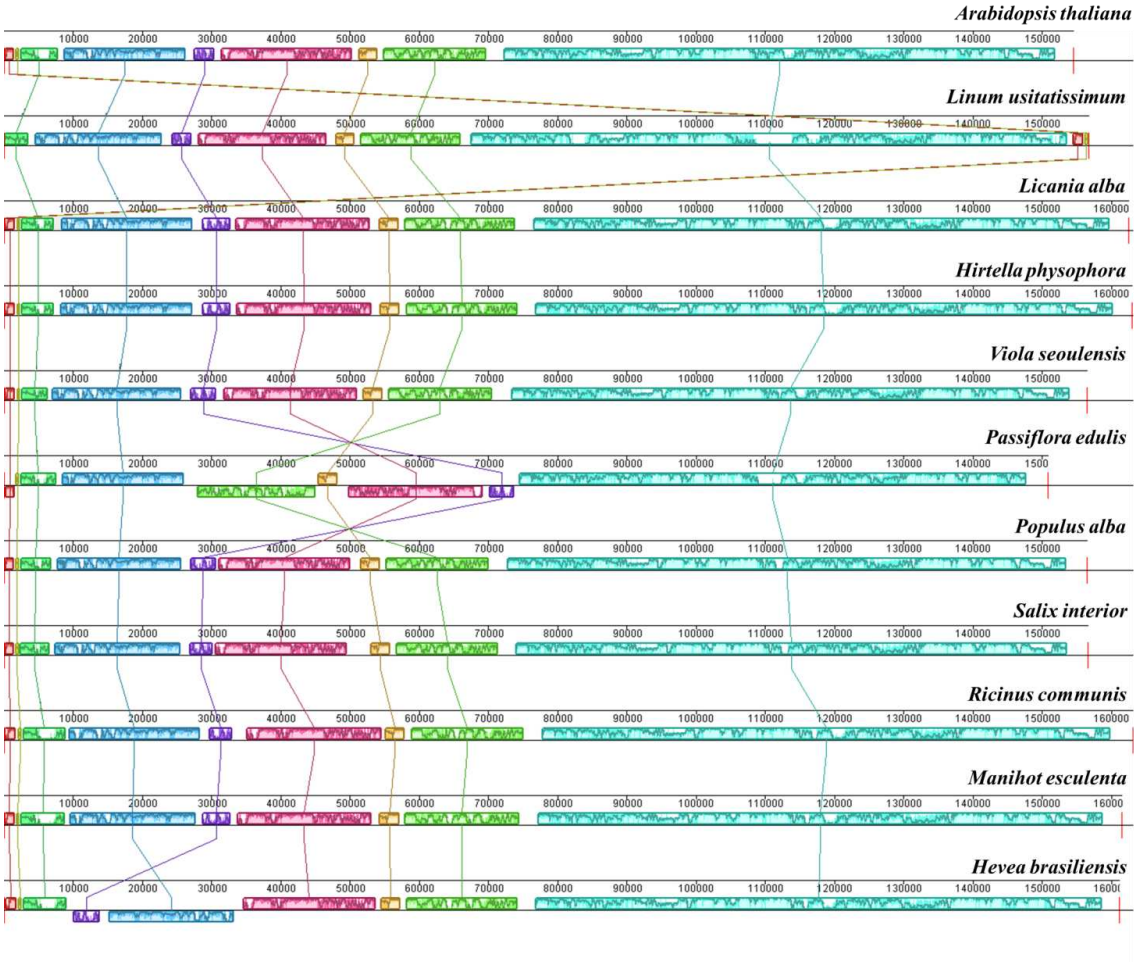
Xu JH, Liu Q, Hu W, Wang T, Xue Q, Messing J (2015) Dynamics of chloroplast genomes in green plants. Genomics 106:221–231. doi:10.1016/j.ygeno.2015.07.004

Zeist W, Bakker-Heeres JAH (1975) Evidence for linseed cultivation before 6000 bc. J Archaeol Sci 2:215–219. doi:10.1016/0305-4403(75)90059-X

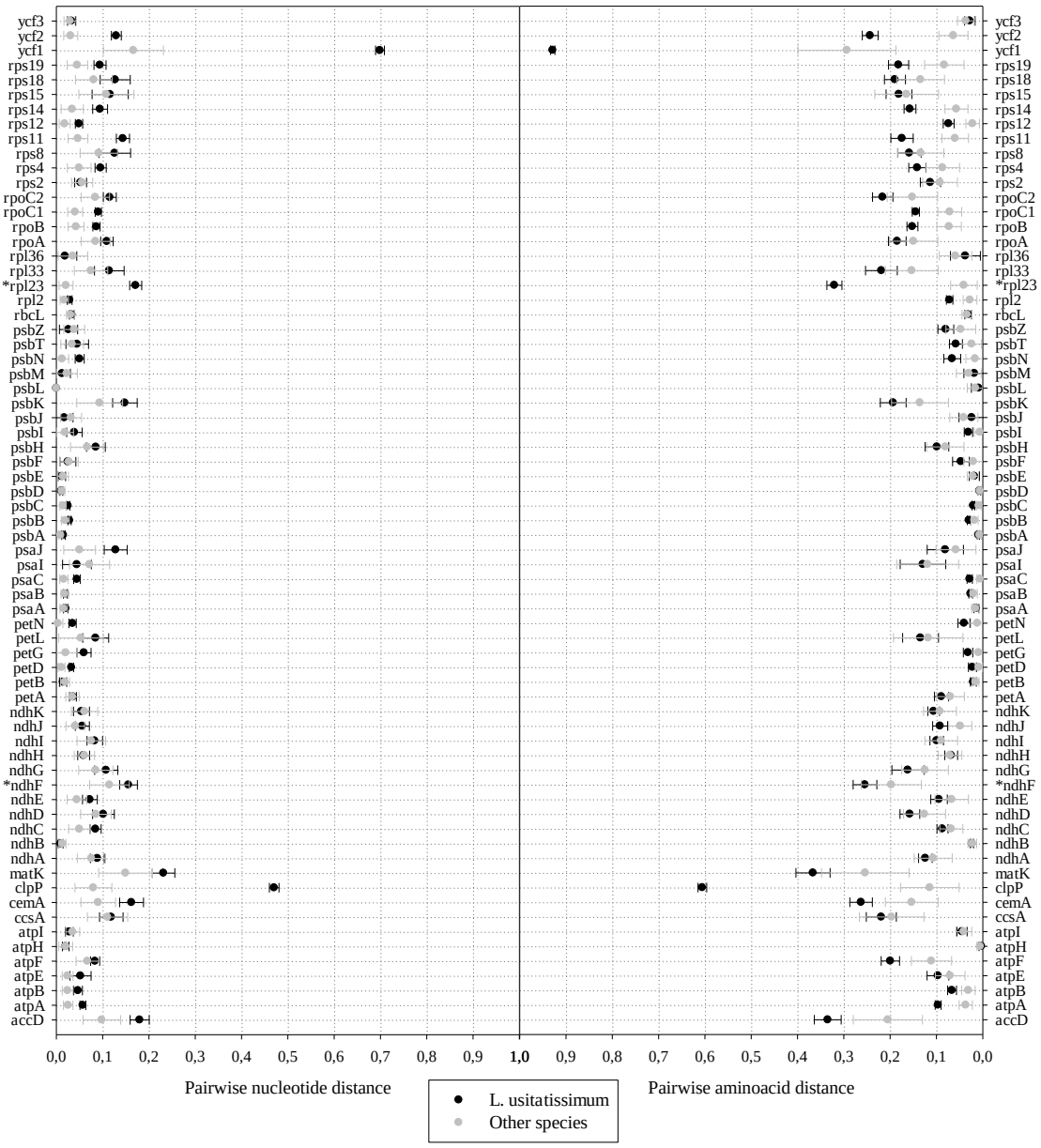
Zhu A, Guo W, Gupta S, Fan W, Mower JP (2016) Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. New Phytol 209:1747–1756. doi:10.1111/nph.13743

SUPPLEMENTARY MATERIAL

Supplementary Figures

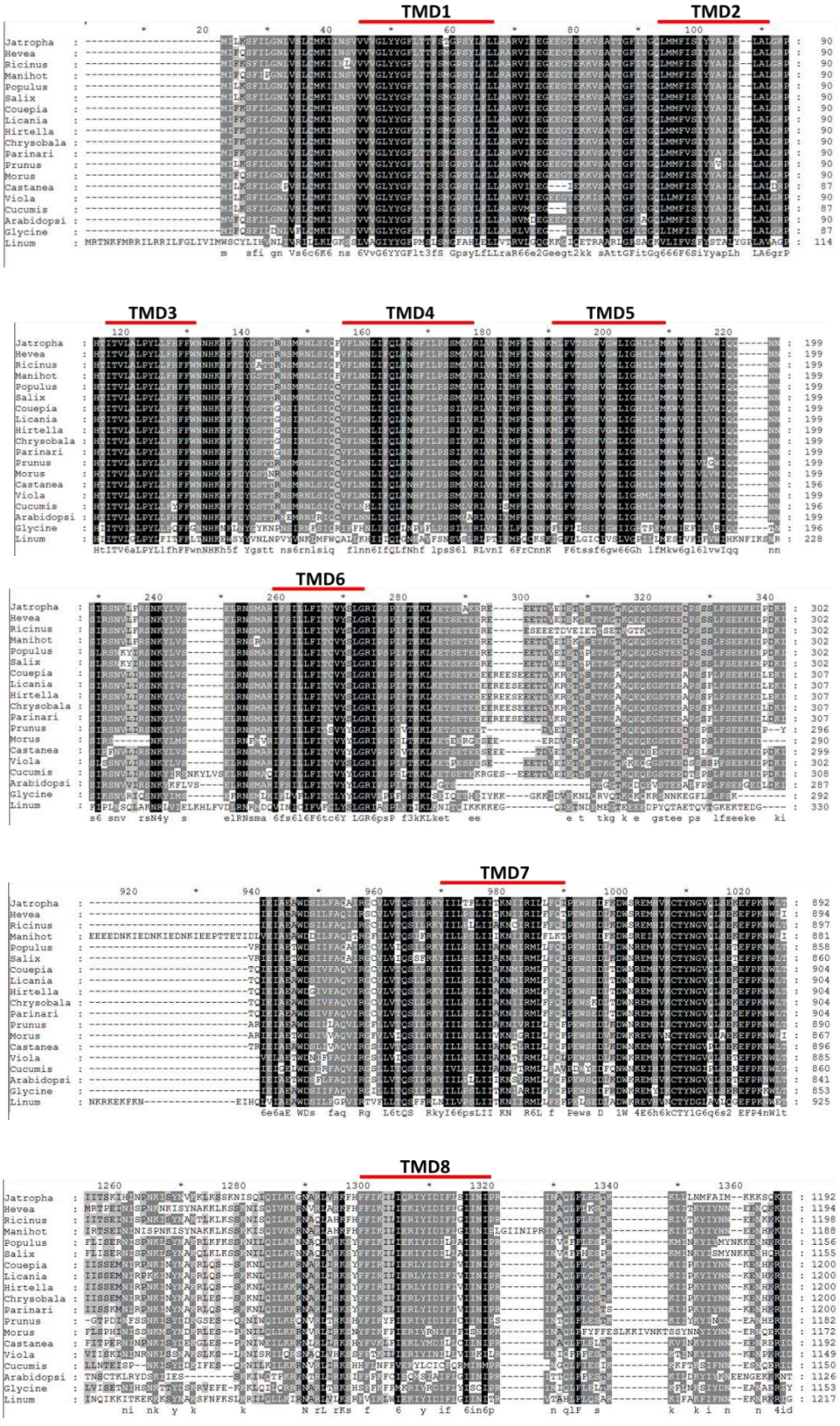


Supplementary Fig. S1. Gene order comparison of Malpighiales plastomes using Arabidopsis thaliana as external reference

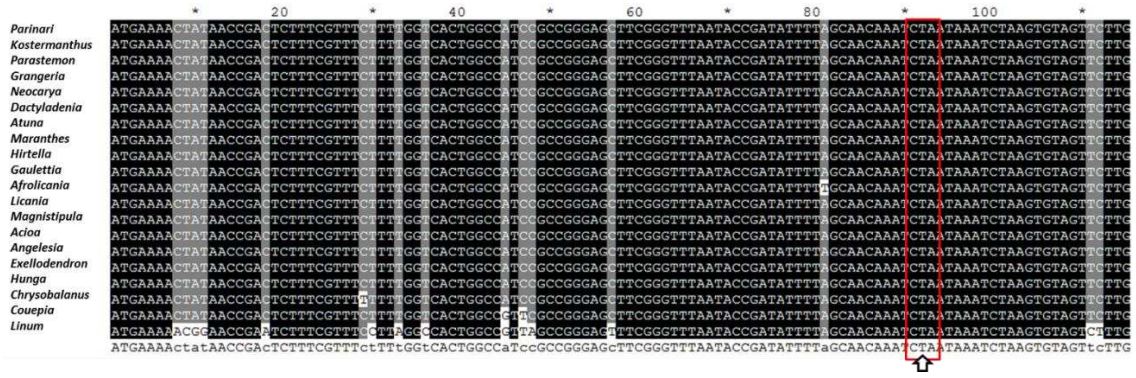


Supplementary Fig. S2. Pairwise distance based on nucleotide and amino acid sequence for 69 protein coding genes

Supplementary Fig. S3. The ClpP amino acid sequence was analyzed using Conserved Domains (NCBI). All the ClpP proteins, including flax, belong to the family S14_ClpP_2. The residues of the active site and of the oligomer interface are highlighted



Supplementary Fig. S4. Multiple sequence alignment of Ycf1 proteins from flax (Linum) and related species. The alignment was made using ClustalW. Eight conserved domains are highlighting (TMD1-8), in accordance with Kikuchi et al. (2013), six predicted transmembrane segments (TMD1-TMD6) and two moderately hydrophobic stretches (TMD7 and TMD8)



Supplementary Fig. S5. Multiple sequence alignment of the gene atpF from flax (Linum) and representatives of the family Chrysobalanaceae. The alignment was made using ClustalW. The codon highlighted in red is at position 31, and the nucleotide highlighted by arrow is the nucleotide 92. For this alignment was used one specie from each genus of the family Chrysobalanaceae present in the data base

Supplementary Tables

Supplementary Table S1. List of simple sequence repeats identified in the plastome of Linum usitatissimum

SSR Sequence	Number of repeats													Total
	3	4	5	7	8	9	10	11	12	13	14	16	20	
A/T	-	-	-	-	60	31	15	2	2	3	3	1	1	118
C/G	-	-	-	-	5	-	1	-	-	-	-	-	-	6
AC/GT	-	2	-	-	-	-	-	-	-	-	-	-	-	2
AG/CT	-	10	-	-	-	-	-	-	-	-	-	-	-	10
AT/AT	-	16	6	1	-	-	-	-	-	-	-	-	-	23
AAT/ATT	-	7	-		-	-	-	-	-	-	-	-	-	7
AAAG/CTTT	4	-	-	-	-	-	-	-	-	-	-	-	-	4
AAAT/ATTT	2	-	-	-	-	-	-	-	-	-	-	-	-	2
ACAT/ATGT	1	-	-	-	-	-	-	-	-	-	-	-	-	1
AAAAG/CTTTT	2	-	-	-	-	-	-	-	-	-	-	-	-	2
AATCCC/ATTGGG	1	-	-	-	-	-	-	-	-	-	-	-	-	1

Supplementary Table S2. Distribution of simple sequence repeats (SSRs) loci in the plastome of Linum usitatissimum

SSR type	SSR	size	start	end	Location
mono	(A)8	8	356	363	matK (CDS)
di	(AT)7	14	559	572	matK (CDS)
mono	(A)8	8	627	634	matK (CDS)
mono	(A)8	8	676	683	matK (CDS)
mono	(C)8	8	2006	2013	trnK-UUU/trnQ-UUG (IGS)
mono	(T)8	8	2159	2166	trnK-UUU/trnQ-UUG (IGS)
mono	(A)16	16	2170	2185	trnK-UUU/trnQ-UUG (IGS)
tetra	(TAAA)3	12	3486	3497	psbI/trnS-GCU (IGS)
mono	(A)9	9	3702	3710	trnS-GCU/trnG-UCC (IGS)
mono	(A)8	8	3756	3763	trnS-GCU/trnG-UCC (IGS)
mono	(T)9	9	3824	3832	trnS-GCU/trnG-UCC (IGS)
di	(AT)4	8	3925	3932	trnS-GCU/trnG-UCC (IGS)
di	(AT)4	8	5694	5701	trnR-UCU/atpA (IGS)
mono	(T)9	9	5787	5795	trnR-UCU/atpA (IGS)
mono	(A)9	9	5846	5854	trnR-UCU/atpA (IGS)
tetra	(CTTT)3	12	7466	7477	atpF (CDS)
mono	(C)8	8	7795	7802	atpF (CDS)
tri	(TAA)4	12	8060	8071	atpF (intron)
mono	(T)9	9	8182	8190	atpF (intron)
mono	(A)8	8	8646	8653	atpF (CDS)
mono	(A)9	9	8956	8964	atpF/atpH (IGS)
mono	(T)10	10	8972	8981	atpF/atpH (IGS)
mono	(T)8	8	8997	9004	atpF/atpH (IGS)
mono	(T)8	8	9136	9143	atpF/atpH (IGS)
di	(AT)4	8	9858	9865	atpH/atpI (IGS)

mono	(T)9	9	12075	12083	rps2 (CDS)
mono	(C)10	10	12312	12321	rps2/rpoC2 (IGS)
mono	(A)9	9	12322	12330	rps2/rpoC2 (IGS)
mono	(A)10	10	12342	12351	rps2/rpoC2 (IGS)
tetra	(ACAT)3	12	13452	13463	rpoC2 (CDS)
mono	(A)8	8	13970	13977	rpoC2 (CDS)
mono	(T)8	8	14043	14050	rpoC2 (CDS)
mono	(T)20	20	14559	14578	rpoC2 (CDS)
di	(AT)5	10	15953	15962	rpoC2 (CDS)
mono	(C)8	8	18535	18542	rpoC1 (intron)
mono	(A)8	8	18637	18644	rpoC1 (intron)
mono	(T)8	8	19456	19463	rpoC1 (CDS)
di	(AT)5	10	20525	20534	rpoB (CDS)
mono	(T)10	10	21567	21576	rpoB (CDS)
mono	(T)10	10	22313	22322	rpoB (CDS)
mono	(A)9	9	23413	23421	rpoB/trnC-GCA (IGS)
mono	(T)9	9	23445	23453	rpoB/trnC-GCA (IGS)
mono	(T)8	8	24475	24482	trnC-GCA/petN (IGS)
mono	(A)8	8	24872	24879	petN/psbM (IGS)
mono	(A)8	8	24975	24982	petN/psbM (IGS)
mono	(T)8	8	25096	25103	petN/psbM (IGS)
mono	(T)10	10	25276	25285	psbM/trnD-GUC (IGS)
mono	(A)8	8	25829	25836	psbM/trnD-GUC (IGS)
mono	(T)8	8	26316	26323	psbM/trnD-GUC (IGS)
mono	(A)9	9	27441	27449	trnE-UUC/trnT-GGU (IGS)
mono	(T)8	8	28698	28705	trnT-GGU/psbD (IGS)
mono	(T)8	8	30798	30805	psbC (CDS)
tetra	(TCTT)3	12	31757	31768	psbC/trnS-UGA (IGS)
mono	(T)10	10	31804	31813	psbC/trnS-UGA (IGS)
di	(GA)4	8	31921	31928	trnS-UGA (CDS)
mono	(T)9	9	32032	32040	trnS-UGA/psbZ (IGS)
di	(TA)5	10	32155	32164	trnS-UGA/psbZ (IGS)
mono	(A)8	8	32766	32773	psbZ/trnG-GCC (IGS)
mono	(A)8	8	33006	33013	psbZ/trnG-GCC (IGS)
penta	(TTTTC)3	15	33802	33816	rps14 (CDS)
mono	(A)8	8	34973	34980	psaB (CDS)
di	(AG)4	8	37806	37813	psaA (CDS)
mono	(A)10	10	38938	38947	psaA/ycf3 (IGS)
mono	(T)8	8	39925	39932	ycf3 (intron)
mono	(T)8	8	41244	41251	ycf3/trnS-GGA (IGS)
mono	(A)8	8	43279	43286	trnT-UGU/trnL-UAA (IGS)
mono	(T)8	8	44074	44081	trnL-UAA/trnF-GAA (IGS)
mono	(T)8	8	46315	46322	ndhK/ndhC (IGS)
tri	(TTA)4	12	47109	47120	ndhC/trnV-UAC (IGS)
mono	(A)14	14	47368	47381	ndhC/trnV-UAC (IGS)
mono	(T)10	10	47842	47851	ndhC/trnV-UAC (IGS)
mono	(T)8	8	47949	47956	trnV-UAC (intron)
tri	(ATA)4	12	50907	50918	atpB/rbcL (IGS)

mono	(A)12	12	50951	50962	atpB/rbcL (IGS)
di	(GA)4	8	52159	52166	rbcL (CDS)
mono	(A)9	9	53574	53582	rbcL/accD (IGS)
mono	(T)10	10	53669	53678	rbcL/accD (IGS)
mono	(T)8	8	54150	54157	accD (CDS)
di	(TA)4	8	55937	55944	accD/psaI (IGS)
di	(TA)4	8	55977	55984	accD/psaI (IGS)
di	(TA)4	8	55987	55994	accD/psaI (IGS)
mono	(A)9	9	56581	56589	psaI/ycf4 (IGS)
mono	(A)13	13	57394	57406	cemA (CDS)
di	(AT)4	8	58318	58325	petA (CDS)
di	(TG)4	8	58410	58417	petA (CDS)
mono	(C)8	8	58604	58611	petA (CDS)
mono	(A)8	8	58678	58685	petA (CDS)
mono	(T)10	10	59523	59532	petA/psbJ (IGS)
mono	(T)9	9	59611	59619	petA/psbJ (IGS)
mono	(A)8	8	59763	59770	petA/psbJ (IGS)
mono	(A)9	9	61435	61443	psbE/petL (IGS)
mono	(A)8	8	61693	61700	psbE/petL (IGS)
mono	(T)8	8	62108	62115	petG/trnW-CCA (IGS)
di	(TG)4	8	62369	62376	trnW-CCA/trnP-UGG (IGS)
mono	(A)8	8	62771	62778	trnP-UGG/psaJ (IGS)
tri	(TTA)4	12	62938	62949	trnP-UGG/psaJ (IGS)
di	(AT)5	10	63870	63879	rpl33/rps18 (IGS)
di	(AT)5	10	63905	63914	rpl33/rps18 (IGS)
tri	(TAA)4	12	64381	64392	rps18 (CDS)
mono	(T)11	11	64693	64703	rps18/rpl20 (IGS)
mono	(A)9	9	65274	65282	rpl20/rps12 (IGS)
mono	(A)8	8	65387	65394	rpl20/rps12 (IGS)
mono	(T)8	8	65832	65839	rpl20/rps12 (IGS)
di	(CT)4	8	65915	65922	rpl20/rps12 (IGS)
di	(AT)4	8	66787	66794	rps12/psbB (IGS)
mono	(G)8	8	67572	67579	rps12/psbB (IGS)
di	(AT)4	8	69493	69500	psbT/psbN (IGS)
mono	(T)9	9	70071	70079	psbH/petB (IGS)
mono	(A)8	8	72342	72349	petD (intron)
di	(AT)4	8	73316	73323	rpoA (CDS)
mono	(T)9	9	73552	73560	rpoA (CDS)
mono	(T)8	8	75883	75890	rps8/rpl14 (IGS)
mono	(T)8	8	75973	75980	rps8/rpl14 (IGS)
mono	(A)9	9	76432	76440	rpl14/rpl16 (IGS)
mono	(T)8	8	77148	77155	rpl16 (intron)
mono	(T)9	9	77572	77580	rpl16 (intron)
mono	(T)9	9	77915	77923	rpl16 (intron)
mono	(T)8	8	78003	78010	rpl16/rps3 (IGS)
mono	(T)8	8	78832	78839	rpl22 (CDS)
tri	(ATT)4	12	79425	79436	rpl22/rps19 (IGS)
mono	(T)10	10	79475	79484	rps19 (CDS)

mono	(T)8	8	79680	79687	rps19 (CDS)
mono	(T)9	9	79712	79720	rps19 (CDS)
tetra	(TTTC)3	12	83877	83888	psbA/trnH-GUG (IGS)
di	(AT)5	10	84188	84197	trnH-GUG/ycf2 (IGS)
mono	(T)14	14	84415	84428	trnH-GUG/ycf2 (IGS)
mono	(A)8	8	85865	85872	ycf2 (CDS)
hexa	(GATTGG)3	18	86394	86411	ycf2 (CDS)
mono	(A)9	9	86913	86921	ycf2 (CDS)
di	(GA)4	8	86934	86941	ycf2 (CDS)
mono	(A)12	12	88139	88150	ycf2 (CDS)
di	(AT)4	8	91431	91438	trnL-CAA/ndhB (IGS)
mono	(A)8	8	91444	91451	trnL-CAA/ndhB (IGS)
di	(AT)4	8	91530	91537	trnL-CAA/ndhB (IGS)
di	(AG)4	8	99069	99076	rrn16/trnI-GAU (IGS)
di	(CT)4	8	102965	102972	rrn23 (CDS)
di	(AG)4	8	104716	104723	rrn5/trnR-ACG (IGS)
tetra	(AAGA)3	12	105035	105046	trnR-ACG/trnN-GUU (IGS)
mono	(T)9	9	105455	105463	trnN-GUU/ycf1 (IGS)
mono	(A)13	13	106460	106472	ycf1 (CDS)
mono	(T)8	8	106778	106785	ycf1 (CDS)
mono	(A)8	8	106942	106949	ycf1 (CDS)
mono	(T)9	9	107873	107881	ycf1 (CDS)
penta	(AAAAG)3	15	108072	108086	ycf1 (CDS)
mono	(A)10	10	108582	108591	ycf1 (CDS)
mono	(A)8	8	108603	108610	ycf1 (CDS)
mono	(A)10	10	108717	108726	ycf1 (CDS)
mono	(A)9	9	108764	108772	ycf1 (CDS)
mono	(A)8	8	109242	109249	ycf1 (CDS)
mono	(A)8	8	110650	110657	ycf1 (CDS)
mono	(A)8	8	110834	110841	ycf1 (CDS)
di	(AT)4	8	111011	111018	ycf1/rps15 (IGS)
mono	(A)8	8	111168	111175	ycf1/rps15 (IGS)
mono	(A)8	8	111194	111201	ycf1/rps15 (IGS)
di	(AG)4	8	113723	113730	ndhA (intron)
mono	(A)11	11	113732	113742	ndhA (intron)
mono	(A)8	8	114425	114432	ndhF (pseudo)
mono	(T)13	13	114579	114591	ndhF (pseudo)
mono	(T)9	9	114643	114651	ndhF (pseudo)
di	(AT)4	8	115829	115836	ndhF (pseudo)
di	(AT)4	8	116224	116231	ndhF/rpl32 (IGS)
tri	(TAA)4	12	116433	116444	ndhF/rpl32 (IGS)
mono	(T)8	8	116754	116761	ndhF/rpl32 (IGS)
mono	(T)8	8	116808	116815	ndhF/rpl32 (IGS)
mono	(A)10	10	116877	116886	rpl32 (CDS)
mono	(A)9	9	117565	117573	rpl32/trnL-UAG (IGS)
mono	(T)8	8	117753	117760	rpl32/trnL-UAG (IGS)
mono	(T)10	10	117862	117871	rpl32/trnL-UAG (IGS)
mono	(T)14	14	119276	119289	ccsA/ndhD (IGS)

mono	(T)8	8	119344	119351	ccsA/ndhD (IGS)
tetra	(AATA)3	12	119586	119597	ndhD (CDS)
di	(AT)4	8	120769	120776	ndhD (CDS)
mono	(T)9	9	120941	120949	ndhD/psaC (IGS)
di	(AG)4	8	122522	122529	ndhG (CDS)
mono	(A)9	9	122753	122761	ndhG/ndhI (IGS)
mono	(A)8	8	122956	122963	ndhG/ndhI (IGS)

Supplementary Table S3. Distribution of tandem repeats in the plastome of Linum usitatissimum

Copy number	**Consensus length**	**Start**	**End**	**Location**
2.0	45	32128	32220	trnS-UGA/psbZ (IGS)
3.0	54	54380	54532	accD (CDS)
2.0	33	54912	54976	accD (CDS)
3.0	22	54979	55040	accD (CDS)
2.0	49	55021	55117	accD (CDS)
2.0	35	63854	63923	rpl33/rps18 (IGS)
3.0	21	64132	64194	rps18 (CDS)
5.0	27	64383	64508	rps18 (CDS)
3.0	24	66877	66948	rps12/psbB (IGS)
4.0	18	66933	67004	rps12/psbB (IGS)
2.0	23	83827	83873	psbA/trnH-GUG (IGS)
4.0	21	86087	86165	ycf2 (CDS)
2.0	21	86334	86374	ycf2 (CDS)
2.0	42	88705	88788	ycf2 (CDS)
3.0	28	90120	90188	ycf2 (CDS)
5.0	18	107915	108004	ycf1 (CDS)
6.0	9	109786	109839	ycf1 (CDS)
3.0	24	110075	110146	ycf1 (CDS)
4.0	24	110496	110586	ycf1 (CDS)
2.0	36	116333	116404	ndhF/rpl32 (IGS)

Supplementary Table S4. Distribution of direct (D) and inverted (I) sequence repeats loci in the plastome of Linum usitatissimum

Type	Size (bp)	Start			Location		
		repeat 1	repeat 2	repeat 3	repeat 1	repeat 2	repeat 3
D	119	87016	87169		ycf2 (CDS)	ycf2 (CDS)	
D	46	67221	81279		rps12/psbB (IGS)	rpl23 (pseudo)	
D	45	64409	64463		rps18 (CDS)	rps18 (CDS)	
D	41	54316	54433		accD (CDS)	accD (CDS)	
D	40	95750	124140		rps12/trnV-GAC (IGS)	ndhA (intron)	
D	39	110499	110547		ycf1 (CDS)	ycf1 (CDS)	
D	38	54443	54497		accD (CDS)	accD (CDS)	
D	38	40250	124142		ycf3 (intron)	ndhA (intron)	
D	38	90069	90132		ycf2 (CDS)	ycf2 (CDS)	
D	37	40252	95754		ycf3 (intron)	rps12/trnV-GAC (IGS)	
D	36	107914	107968		ycf1 (CDS)	ycf1 (CDS)	
D	35	54405	54450		accD (CDS)	accD (CDS)	
D	35	55029	55077		accD (CDS)	accD (CDS)	
D	35	64382	64433	64460	rps18 (CDS)	rps18 (CDS)	rps18 (CDS)
D	34	86064	86340		ycf2 (CDS)	ycf2 (CDS)	
D	34	116082	116194		ndhF/rpl32 (IGS)	ndhF/rpl32 (IGS)	
D	32	110074	110122		ycf1 (CDS)	ycf1 (CDS)	
D	32	35237	37461		psaB (CDS)	psaA (CDS)	
D	32	54325	54442		accD (CDS)	accD (CDS)	
D	32	54911	55082		accD (CDS)	accD (CDS)	
D	32	66876	66912		rps12/psbB (IGS)	rps12/psbB (IGS)	
D	32	86078	86354		ycf2 (CDS)	ycf2 (CDS)	
D	31	107920	107956		ycf1 (CDS)	ycf1 (CDS)	
D	31	54326	54398	54497	accD (CDS)	accD (CDS)	accD (CDS)
D	31	66855	66933		rps12/psbB (IGS)	rps12/psbB (IGS)	
D	31	67116	78050		rps12/psbB (IGS)	rpl16/rps3 (IGS)	
D	30	67082	78005		rps12/psbB (IGS)	rpl16/rps3 (IGS)	
D	30	66945	66981		rps12/psbB (IGS)	rps12/psbB (IGS)	
D	30	35311	37535		psaB (CDS)	psaA (CDS)	
D	30	54938	55076		accD (CDS)	accD (CDS)	
D	30	54911	55034		accD (CDS)	accD (CDS)	
D	30	3546	31914		trnS-GCU (CDS)	trnS-UGA (CDS)	
D	30	90060	90150		ycf2 (CDS)	ycf2 (CDS)	
D	30	86033	86258		ycf2 (CDS)	ycf2 (CDS)	
D	30	107938	107974		ycf1 (CDS)	ycf1 (CDS)	
D	30	110510	110558		ycf1 (CDS)	ycf1 (CDS)	
I	30	3549	41486		trnS-GCU (CDS)	trnS-GGA (CDS)	
I	31	69459	69491		psbT/psbN (IGS)	psbT/psbN (IGS)	
I	30	40251	70854		ycf3 (intron)	petB (intron)	

Supplementary Table S5. Species included in the phylogenetic inference of the position of Linum usitatissimum (Linaceae)

Species	Clade	Order	Family	GenBank
Linum usitatissimum[+]	fabids	Malpighiales	Linaceae	KY849971
Acioa guianensis	fabids	Malpighiales	Chrysobalanaceae	NC_030534.1
Afrolicania elaeosperma	fabids	Malpighiales	Chrysobalanaceae	NC_030544.1
Angelesia splendens	fabids	Malpighiales	Chrysobalanaceae	NC_030545.1
Atuna racemosa	fabids	Malpighiales	Chrysobalanaceae	NC_030546.1
Chrysobalanus icaco[+]	fabids	Malpighiales	Chrysobalanaceae	NC_024061
Couepia guianensis[+]	fabids	Malpighiales	Chrysobalanaceae	NC_024063
Dactyladenia bellayana	fabids	Malpighiales	Chrysobalanaceae	NC_030555.1
Exellodendron barbatum	fabids	Malpighiales	Chrysobalanaceae	NC_030558.1
Gaulettia elata	fabids	Malpighiales	Chrysobalanaceae	NC_030559.1
Grangeria borbonica	fabids	Malpighiales	Chrysobalanaceae	NC_030560.1
Hunga gerontogea	fabids	Malpighiales	Chrysobalanaceae	NC_030564.1
Hirtella physophora[+]	fabids	Malpighiales	Chrysobalanaceae	NC_024066
Kostermanthus robustus	fabids	Malpighiales	Chrysobalanaceae	NC_030565.1
Licania alba[+]	fabids	Malpighiales	Chrysobalanaceae	NC_024064
Magnistipula butayei	fabids	Malpighiales	Chrysobalanaceae	NC_030576.1
Maranthes gabunensis	fabids	Malpighiales	Chrysobalanaceae	NC_030577.1
Neocarya macrophylla	fabids	Malpighiales	Chrysobalanaceae	NC_030580.1
Parastemon urophyllus	fabids	Malpighiales	Chrysobalanaceae	NC_030517.1
Parinari campestris[+]	fabids	Malpighiales	Chrysobalanaceae	NC_024067
Euphorbia esula	fabids	Malpighiales	Euphorbiaceae	NC_033910.1
Hevea brasiliensis[+]	fabids	Malpighiales	Euphorbiaceae	NC_015308
Jatropha curcas[+]	fabids	Malpighiales	Euphorbiaceae	NC_012224
Manihot esculenta[+]	fabids	Malpighiales	Euphorbiaceae	NC_010433
Ricinus communis[+]	fabids	Malpighiales	Euphorbiaceae	NC_016736
Vernicia fordii	fabids	Malpighiales	Euphorbiaceae	NC_034803.1
Erythroxylum novogranatense	fabids	Malpighiales	Erythroxylaceae	NC_030601.1
Passiflora edulis	fabids	Malpighiales	Passifloracea	NC_034285.1
Idesia polycarpa	fabids	Malpighiales	Salicaceae	NC_032060.1
Populus alba[+]	fabids	Malpighiales	Salicaceae	NC_008235
Salix interior[+]	fabids	Malpighiales	Salicaceae	NC_024681
Viola seoulensis[+]	fabids	Malpighiales	Violaceae	NC_026986
Morus mongolica[+]	fabids	Rosales	Moraceae	NC_025772
Prunus persica[+]	fabids	Rosales	Rosaceae	NC_014697
Cucumis melo[+]	fabids	Cucurbitales	Cucurbitaceae	NC_015983
Glycine max[+]	fabids	Fabales	Fabaceae	NC_007942
Castanea mollissima[+]	fabids	Fagales	Fagaceae	NC_014674
Arabidopsis thaliana*[+]	malvids	Brassicales	Brassicaceae	NC_000932

*****Out-group; [+]Species used in the Pairwise distance and dN/dS analyses.

The Crambe abyssinica plastome: Brassicaceae phylogenomic, evolution of RNA editing sites, hotspot and microsatellite analyses of the tribe Brassiceae

Amanda de Santana Lopes[1], Túlio Gomes Pacheco[1], Leila do Nascimento Vieira[2], Miguel Pedro Guerra[2], Rubens Onofre Nodari[2], Emanuel Maltempi de Souza[3], Fábio de Oliveira Pedrosa[3], Marcelo Rogalski[1*]

1 Laboratório de Fisiologia Molecular de Plantas, Departamento de Biologia Vegetal, Universidade Federal de Viçosa, Viçosa-MG, Brazil.
2 Laboratório de Fisiologia do Desenvolvimento e Genética Vegetal, Programa de Pós-graduação em Recursos Genéticos Vegetais, Universidade Federal de Santa Catarina, Florianópolis-SC, Brazil.
3 Departamento de Bioquímica e Biologia Molecular, Núcleo de Fixação Biológica de Nitrogênio, Universidade Federal do Paraná, Curitiba-PR, Brazil.

*****Corresponding author
E-mail address: rogalski@ufv.br

Manuscript submitted to **Gene**

Key message:

The evolution of Crambe abyssinica plastome: characterization of molecular markers and phylogenomic of Brassicaceae

Abstract

Crambe abyssinica is an important oilseed crop that accumulates high levels of erucic acid being recognized as a potential oil platform for several industrial purposes. It belongs to the Brassicaceae family and it is assigned within the tribe Brassiceae. Both, family and tribe have been targeted for several phylogenetic studies, but the relationship between some lineages and genera remains unclear. Here, we report the complete sequencing and characterization of Crambe abyssinica plastome. Plastome structure, gene order, and gene content of C. abyssinica are similar to other species of the family Brassicaceae. The only exception is the rps16 gene, which is absent in many genera within family Brassicaceae, but seems to be functional in the tribe Brassiceae, including C. abyssinica. However, the analysis of gene divergence shows that the rps16 is the most divergent gene in C. abyssinica and whitin the tribe Brassiceae. In addition, species of the tribe Brassiceae also show similar SSR loci distribution, with some regions containing a high number of SSRs, which are located mainly at the single copy regions. Six hotspots of nucleotide divergence among Brassiceae species were located in the regions of single copy by sliding window analysis. Brassicaceae phylogenomic, based on whole-plastome of 72 taxa, resulted in a well-supported and well-resolved tree. The genus Crambe is positioned within the Brassiceae clade together with genera Brassica, Raphanus, Sinapis, Cakile, Orychophragmus and Sinalliaria. Moreover, species of the tribe Brassiceae showed several events of losses and gains of RNA editing sites during the evolution.

Keywords: Plastid molecular markers; genetic divergence; extranuclear inheritance; organellar DNA; gene function; oilseeds.

1. Introduction

Crambe abyssinica is an oilseed plant that naturally accumulates up to 60 % of erucic acid, which is a useful source for industrial feedstock and several other applications (Lazzeri et al., 1997). Therefore, C. abyssinica has been recognized as a potential oil platform for sustainable industrial feedstock production and an alternative to petroleum (Carlsson, 2009). Several studies have been carried out to improve agronomic traits and oil profile in C. abyssinica by traditional breeding (Mastebroek et al., 1994) and nuclear genetic engineering (Li et al., 2012; Cheng et al., 2015; Zhu et al., 2016b).

The plastid genome (plastome) transformation has been proved to be a great tool for metabolic engineering (Apel and Bock, 2009; Kumar et al., 2012; Lu et al., 2013) and it is a viable alternative to manipulate plastid fatty acid biosynthesis (Rogalski and Carrer, 2011). The plastome sequencing and characterization is essential to reveal target intergenic sequences aiming the insertion of transgenes and the development of transplastomic plants with new desired characteristics (Rogalski and Carrer, 2011; Bock, 2015; Daniell et al., 2016).

In addition to the biotechnological interest, plastomes are sources of information for several genetic studies. The use of conserved plastid genes and also whole-plastomes have generated highly precise phylogenetic inferences (Jansen et al., 2007; Xi et al., 2012; Ruhfel et al., 2014; Barret et al., 2016; Vieira et al., 2016a). Plastid intergenic spacers, introns, and molecular makers such as single nucleotide polymorphisms (SNPs) and single sequence repeats (SSRs) have been used for genetic analyses in natural plant populations and phylogeographycal studies (Besnard et al., 2011; Rogalski et al., 2015; Qiao et al., 2016). Gene content, loss of genes, gene transfer to the nucleus, recombination events, and genome rearrangements of plastomes are also explored to understand evolutionary events in plants (Guo et al., 2007; Jansen et al., 2011; Wicke et al., 2011; Vieira et al., 2014a, 2016b). Additionally, comparison between plastome and plastid transcript sequences is important to study the evolution of RNA editing, widely spread among land plants organelles (Takenaka et al., 2013), and the implication of this process to the nuclear-plastome incompatibility upon hybridization (Greiner and Bock, 2013).

C. abyssinica belongs to the family Brassicaceae, which comprises 3660 species classified within 321 genera, according to BrassiBase database (https://brassibase.cos.uni-heidelberg.de/; Al-Shehbaz, 2012). Most Brassicaceae genera are assigned to 49 tribes (Al-Shehbaz, 2012; BrassiBase, 2017). The genus Crambe is assigned within the tribe Brassiceae, which contains approximately 50 genera and 240 species, including the genera Brassica and Raphanus (Warwick and Sauder, 2005). The

family Brassicaceae is characterized by high speciation rate (Couvreus et al., 2010; Karl and Koch, 2013). Part of this high diversification has been correlated with ancient polyploidization events (Lysak and Koch, 2011; Kagale et al., 2014; Hohmann et al., 2015). Moreover, gene flow and incomplete lineage sorting in Brassicaceae have been associated with nonbifurcating speciation and phylogenetic inconsistences between plastid and nuclear datasets (Hu et al., 2016; Novikova et al., 2016). Furthermore, data from phylogenetic studies within the tribe Brassiceae also suggest polyphyletic origins for some genera, including Brassica, and show incongruence between plastid and nuclear data (Warwick and Sauder, 2005; Hall et al., 2011; Hu et al., 2016).

Until the present date, 71 Brassicaceae complete plastomes are fully sequenced and available at the organelle genome database (https://www.ncbi.nlm.nih.gov/genome/organelle/), encompassing 36 genera. In order to increase the resolution of previous Brassicaceae phylogenetic studies based on plastid sequences (Hohmann et al., 2015; Guo et al., 2017) we present here a plastid phylogenomic covering all plastomes available and the inclusion of Crambe abyssinica, which we sequenced the whole plastome and analyzed in detail. Finally, we also present here a characterization of plastid SSRs, identification of hotspots, analysis of gene divergence, and a prediction of RNA editing sites in the tribe Brassiceae, which may be useful for data selection in future phylogenetic and evolutionary studies of the tribe.

2. Materials and Methods

2.1. Chloroplast isolation, DNA extraction, sequencing, assembling, and annotation

Fresh young leaves of Crambe abysinica were collected at the Federal University of Viçosa, MG, Brazil, and kept for 96 hours at 4°C to decrease starch level. The chloroplast isolation and cp DNA extraction were carried out according to Vieira et al. (2014b). Approximately 1 ng of cp DNA was used to prepare sequencing libraries with Nextera XT DNA Sample Prep Kit (Illumina Inc., San Diego, CA, USA) according to the manufacturer's instructions. The obtained library was sequenced using Illumina MiSeq platform (Illumina Inc., San Diego, CA, USA) at the Federal University of Paraná, PR, Brazil. The paired-end reads (2 x 300 bp) were trimmed under the threshold with probability of error <0.05. The trimmed reads (813,731) were de novo assembled in contigs using CLC Genomics Workbench 8.0.2 software (CLC Bio, Aarhus, Denmark). The average coverage of the contigs used for assembling of plastome ranged from 709.67 to 101.37. The program Dual Organellar GenoMe Annotator (DOGMA) (Wyman et al., 2004) and BLAST were used for preliminarily gene annotation. From this initial

annotation, putative start codons, stop codons, and intron positions were determined based on comparisons to homologous genes in other plastid genomes at the GenBank database. All tRNA genes were further verified by using tRNAscan-SE server (Lowe and Eddy, 1997). The physical circular map of the plastome was drawn using Organellar Genome DRAW (OGDRAW) (Lohse et al., 2013). The complete plastid genome of C. abyssinica was deposited in the GenBank database under accession number KY883663.

2.2. Genome structure analysis

Nucleotide MUMmer (NUCmer) Perl script in MUMmer 3.0 (Kurtz et al., 2004) was used to visualize and compare the plastome structures between C. abyssinica and other Brassicaceae representatives, as well as Carica papaya belonging to family Caricaceae as external group.

2.3. Repeat sequences identification

Simple sequence repeats (SSRs) were detected using the MIcroSAtellite (MISA) Perl script (Thiel et al., 2003). The thresholds were set to eight repeat units for mononucleotide SSRs, four repeat units for di- and trinucleotide SSRs, and three repeat units for tetra-, penta- and hexanucleotide SSRs. Tandem repeats were identified using the program Tandem Repeats Finder (TRF) (Benson, 1999). The setting parameters were 2, 7 and 7 for match, mismatch, and indel, respectively. The minimum alignment score to report repeat and maximum period size were set to 80 and 500, respectively. After, the repeats found were manually verified, and the nested or redundant results were removed. REPuter (Kurtz et al., 2001) was used to locate inverted and directed dispersed repeats. The minimal repeat size was set to 30 bp and the identity of repeats ≥ 90% (hamming distance = 3).

2.4. Sliding window analysis of the tribe Brassiceae

The hotspots of sequence divergence in Brassiceae tribe were investigated by sliding window analysis. The complete plastomes of all Brassiceae species available in the organelle genome database (NCBI), including C. abyssinica characterized here, were aligned using ClustalW implemented in Mega 7.0 (Tamura et al., 2013). Posteriorly, the sliding window analysis was conducted by using the DnaSP v.5 software (Librado and Rozas, 2009). The window length and the step size were set as 400 bp and 100 bp, respectively.

2.5. RNA editing prediction analysis

Potential RNA editing sites in plastid protein-coding genes of species belonging to tribe Brassiceae were predicted by the program Predictive RNA Editor for Plants (PREP) suite (Mower, 2009). The program PREP uses 35 reference genes for detecting

of possible RNA editing sites in plastomes. The cutoff value was set to 0.8. The analyzed genes were defined as follows: accD, atpA, atpB, atpF, atpI, ccsA, clpP, matK, ndhA, ndhB, ndhD, ndhF, ndhG, petB, petD, petG, petL, psaB, psaI, psbB, psbE, psbF, psbL, rpl2, rpl20, rpl23, rpoA, rpoB, rpoC1, rpoC2, rps2, rps8, rps14, rps16, and ycf3.

2.6. Phylogenetic inference

The phylogenetic inference of the family Brassicaceae were carried out using whole plastid genomes. The GenBank accession number of each taxon used in both approaches is shown in the Supplementary Table S1.

Whole plastid genomes (IR_B was omitted to prevent overrepresentation of the IR sequences) were extracted and aligned using MAFFT v.7 (Katoh and Standley, 2013). The best substitution model (GTR+I+G) was selected by using jModelTest v.2.1.7 (Darriba et al., 2012). Bayesian inference analysis was performed using MrBayes version 3.2 (Ronquist et al., 2012), with one million generations of two runs of four Markov Chains, with three hot and one cold in each run. The software Tracer 1.6 (http://tree.bio.ed.ac.uk/software/tracer/) was used to check the parameters convergence and the software FigTree 1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/) was used to visualize the consensus tree.

2.7. Estimation of gene divergence in the tribe Brassiceae

The seventy-eight protein coding genes present in the plastomes of Brassiceae tribe and A. thaliana (as external group) were extracted and aligned individually (codon alignment) using the software Muscle (Edgar, 2004) implemented in Mega 7.0 (Tamura et al., 2013). The substitution models for each gene were selected using jModelTest v.2.1.7 (Darriba et al., 2012). Bayesian inference analyses and visualization of the trees were performed as described in the topic 2.6. The gene divergence was estimated by the sum of total branch lengths that link the operational taxonomical units to the common ancestor among the species of Brassiceae sampled.

3. Results

3.1. Size, gene content and organization of Crambe abyssinica plastome

Crambe abyssinica complete plastid genome (plastome) is 153,771 bp in length (Fig. 1), and includes a pair of inverted repeats (IR_A and IR_B) of 26,195 bp long, separated by a large single copy (LSC) and a small single copy (SSC) regions of 83,600 bp and 17,782 bp, respectively. The gene content of the C. abyssinica plastome consists of 112 unique genes, which are 78 protein-coding genes, 30 tRNA genes, and 4 rRNA genes (Table 1). There are 19 duplicated genes, all of them located in the IR_A and IR_B, two of

them, the ycf1 and rps19 genes, are partially duplicated in the boundary of IRs. Six tRNA genes and 10 protein-coding genes contain one intron and the two protein-coding genes, clpP and ycf3, possess two introns. Comparison of these general features between the C. abyssinica plastome and other twelve plastomes of the tribe Brassiceae available in the organelle genome database (NCBI) (Hu et al., 2011; Jeong et al., 2014; Hu et al., 2016; Prabhudas et al., 2016; Zeng et al., 2016; Seol et al., 2017) shows a high similarity to the genome content, GC percentage, total size, and the structure of LSC, SSC, and IR regions (Table 2). The rps16 gene, which is absent or pseudogene in several genera of family Brassicaceae (Guo et al., 2017), is functional in all plastomes already sequenced from species belonging to the tribe Brassiceae. In addition, comparison by dot plot analyses among plastomes of C. abyssinica and other species of Brassicaceae, including the tribe Brassiceae, shows a conserved structure and gene order in the family (Supplementary Fig. S1), as previously observed (Hu et al., 2015; Guo et al., 2017).

3.2. Estimation of gene divergence in C. abyssinica and within the tribe Brassiceae

Trees based on plastid protein coding genes sequences were constructed and the sum of total branch lengths for each species/gene was used to evaluate the gene divergence within the tribe Brassiceae. Arabidopsis thaliana was used to root all trees. The Fig. 2 shows that most genes are highly conserved in the tribe, presenting branch lengths below 0,02. The rps16 gene is the most divergent one in C. abyssinica and in the tribe Brassiceae (Fig. 2 and Supplementary Fig. S2, and S3). Analyzing the average branch lengths of the tribe Brassiceae, the genes rps16 and ycf1 present the highest values, as well as in C. abyssinica plastid genes (Supplementary Fig.S2 and S3). The genes rpl22, matK, ndhG, and ndhF are also among the most divergent genes in C. abyssinica as well as in the tribe Brassiceae.

3.3. RNA editing sites predicted in plastid protein-coding genes in the tribe Brassiceae

As in other angiosperms (Takenaka et al., 2013), the RNA editing sites predicted in the tribe Brassiceae occur in the first or second position of the codon and all base changes observed here were from cytidine (C) to uridine (U) (Table 3). Among the 35 plastid genes analyzed by cp-PREP program, 18 are predicted to harbor RNA editing sites, accounting for a total of 51 sites. From the total number of editing sites, 40 are conserved among all species of Brassiceae tribe analyzed here, while the accD (472), matK (45), ndhD (16, 225), ndhF (97), psbE (72), rpoB (293, 754), rpoC2 (765, 850), and rps14 (27) editing sites are not present in one or more species of Brassiceae.

3.4. Repeat sequences identification and sliding window analysis of the tribe Brassiceae

The occurrence, type, and distribution of SSRs in C. abyssinica and other Brassiceae plastomes were analyzed. In total, 273 SSRs were identified in C. abyssinica (Supplementary Table S2), close to number of SSRs identified in other species of the tribe Brassiceae (Fig. 3) that ranged from 252 in Sinapis arvensis to 294 in Sinalliaria limprichtiana (Supplementary Table S3-S13). Monopolymers and dipolymers constitute the most part of SSR identified (Fig. 3), accounting for over 94% of the total with 92-94% constituted by A/T sequences. Among the monopolymers and dipolymers identified in C. abyssinica plastome, only six monopolimers and none dipolymer presented more than 15 repeats which is in accordance to the nature of plastid microsatellites, which occur generally with less than 15 mononucleotide repeats (Provan et al., 2001).

The sequence, size, and location of all SSRs in C. abyssinica plastome are shown in the Supplementary Table S14. Among the 273 SSRs identified, 157 are located in the intergenic spacers (IGSs), 76 in coding sequences (CDSs), and 40 in introns. The 76 SSRs identified in CDSs are distributed among 23 genes, of which the ycf1 (28 SSRs), rpoC2 (7 SSRs), and ycf2 (6 SSRs) genes harbor the highest number of SSRs. The introns harboring SSRs are located in 12 genes, of which the rpl16 gene with 7 SSRs harbors the highest number of markers (Supplementary Table S14).

The distribution of SSRs among Brassiceae plastomes shows a great similarity with points of high numbers of SSRs located within the LSC and SSC regions and minor occurrence of SSRs within the IR regions (Fig. 4). This result is in accordance with previous studies that showed greater gene conservation of IRs in comparison with LSC and SSC regions (Zhu et al., 2016a). In addition to that, through sliding window analysis, we find six hotspots of sequence divergence in Brassiceae plastomes, all of them located within LSC and SSC regions (Fig. 5). Several SSRs in C. abyssinica are located within these hotspots, except the hotspot trnL-UAA/trnF-GAA where none was identified. The hotspot rps15/ycf1, whose sequence covers part of the CDSs of rps15 and ycf1, presents 25 SSRs in the C. abyssinica plastome.

In addition to SSRs, eight tandem repeats were identified in C. abyssinica plastome, of which one is located in the CDS of ycf2, one is located in the clpP intron, and six are located in IGSs (Supplementary Table S15). Eight small dispersed repeats (SDRs) ≥ 30 bp were also identified distributed among IGSs, introns, and CDSs, of which five are direct and three are inverted (Supplementary Table S16). Four of them were classified due to the similarity shared among the tRNA genes, trnS-GCU, trnS-GGA, and

trnS-UGA, as well as between the genes psaA and psaB. These types of SDRs have also been reported in plastomes of other angiosperms (Raubeson et al., 2007).

3.5. Brassicaceae phylogenetic inference

The Brassicaceae phylogenetic inference was carried using whole plastid genomes of 74 taxa (Supplementary Table S1), including 72 species of Brassicacea (37 genera, 20 tribes) and two species, Carica papaya (Caricacea: Brassicales) and Tarenaya hassleriana (Cleomaceae: Brassicales), as out-group. Bayesian inference (BI) analysis produced a phylogenetic tree with a –lnL = 829998.5617 (Fig. 6). The BI posterior probability values were 100% for all nodes, except one that was 51% (within the genus Arabidopsis).

Phylogenetic relationships among the genera are in accordance to previous tribe and lineage classifications of Brassicaceae family (Al-Shehbaz, 2012; BrassiBase, 2017). Two major clades were well supported within Brassicaceae, a clade including only the genus Aethionema (tribe Aethionemeae) and a clade including all other tribes. The second clade is divided in two subclades, one including the tribes of the lineage I (Alyssopsideae, Camelineae, Cardamineae Crucihimalayeae, Microlepidieae, and Lepidieae), and other including the linage II (Alysseae, Anastaticeae, Arabideae, Brassiceae, Biscutelleae, Cochlearieae, Eutremeae, Isatideae, Megacarpaeeae, and Thalaspideae) sister to lineage III (Anchonieae, Euclidieae, and Hesperideae).

In our phylogeny tree, Crambe is positioned in the Brassiceae clade which also includes the genera Brassica, Cakile, Orychophragmus, Raphanus, Sinalliaria, and Sinapis. The genera Orychophragmus and Sinalliaria formed a clade sister to all other genera within the Brassiceae clade. Lastly, the species C. abyssinica, Brassica nigra, and Sinapis arvensis formed a group sister to B. juncea, B. napus and Raphanus sativus.

4. Discussion

4.1. The plastome structure, gene content and evolution of plastid genes within the tribe Brassiceae

The C. abyssinica plastome is divided in four major segments (LSC, SSC, IRA, and IRB), showing the common quadripartite structure found in most angiosperms (Wicke et al., 2011; Zhu et al., 2016a), including all published plastomes of the family Brassicaceae (Guo et al., 2017). The size, gene order, and gene content of the LSC, SSC, and IRs of the C. abyssinica plastome are highly conserved among the other species of the tribe Brassiceae (Hu et al., 2011; Jeong et al., 2014; Hu et al., 2016; Prabhudas et al., 2016; Zhou et al., 2016; Seol et al., 2017), which include the presence of a functional

rps16 gene, that encodes an essential ribosomal protein for plastid ribosome biogenesis and cellular viability (Fleischmann et al., 2011). In several genera of the family Brassicaceae and other families the plastid rps16 gene is absent or is a pseudogene (Xu et al., 2015; Guo et al., 2017). The rps16 is one of the plastid genes with highest parallel loss (or pseudogenization) rate and transference from plastome to the nucleus (Xu et al., 2015). Although the rps16 gene is present in all Brassiceae plastomes published so far, our gene divergence inference shows that it is the most divergent gene of the tribe. It suggests a possible transference of the gene to the nucleus in this tribe and the maintenance of a plastid copy in process of degeneration.

The second more divergent gene of the Brassiceae tribe is the ycf1. The ycf1 gene is essential for cell viability (Drescher et al., 2000) and its function was already characterized as a subunit (Tic214) of the TIC complex, a protein translocon located at the inner membrane of plastids (Kikuchi et al., 2013). Nevertheless, this gene is described as one the most divergent gene present in angiosperm plastomes with several reports of gene loss, pseudogenization and transfer to the nucleus (Wicke et al., 2011; Vries et al., 2015).

4.2. The evolution of RNA editing sites within the tribe Brassiceae

It is suggested that evolution of RNA editing in plant organelles (plastid and mitochondria) is correlated with the evolution of land plant, since this process occurs in all land plants, with the notable exception of a bryophyte lineage, the Marchantiales (Rüdinger et al., 2008; Fujii and Small, 2011; Takenaka et al., 2013), which has no RNA editing sites in organelles. Hundreds of editing sites are present in plastomes of some bryophytes lineages, lycopods and ferns, while among angiosperms only approximately 30-40 editing sites are still present (Kahlau et al., 2006; Mower, 2009; Ruwe et al., 2013; Takenaka et al., 2013). Nevertheless, although the number of sites is similar in different species, a high diversity of editing sites has been found among species of flowering plant (Freyer et al., 1997; Fiebig et al., 2004). The reduction of the number and even complete loss of editing sites in some land plant lineages open several questions about the selection and maintenance of these sites. RNA editing can restore conserved codons and create start and stop codons. However, RNA editing seems also to act in regulatory functions and in the creation of variant proteins to better deal with different physiological needs (Takenaka et al., 2013).

Although high interspecific diversity of RNA editing sites has been reported (Freyer et al., 1997; Fiebig et al., 2004), a high conservation of RNA editing sites is predicted within the tribe Brassiceae as observed in our data. The RNA editing data are

in accordance with the low gene divergence observed also within this tribe. Among the 51 sites predicted, 25 of them occur in A. thaliana and were validated by Tillich et al. (2005), including the sites in accD (269), atpF (31), clpP (188), ndhA (114), ndhB (50, 156, 196, 204, 249, 277, 419, and 494), ndhD (1, 128, 225, 293, and 296), ndhF (97), psbE (72), psbF (26), rpoB (113, 184, 811), and rps14 (27 and 50) genes. These data suggest that some sites may be common in the family Brassicaceae and a complete and detailed description of RNA editing sites among Brassicaceae species may be useful to evolutionary studies of the family.

Of the total of 51 RNA editing sites predicted, 11 are not present in all species of the tribe Brassiceae, confirming a diversity of RNA editing in different species even belonging to the same tribe. In 8 of these 11 sites, the species that do not undergo RNA editing have a conserved T replacing the C at the edition site in the DNA sequence. Correlating the distribution of these 8 sites and phylogenetic relationships among the Brassiceae species based on our plastome data, we found possible events of losses and gains of RNA editing sites (Fig. 7A-B). The accD (472) editing site is not present in B. nigra and S. arvensis (clade I) and the ndhF (97) site is not present in B. nigra, S. arvensis, and C. abyssinica (clade H). In accordance with our phylogenetic, these data suggest that the common ancestral I and H lost, respectively, the sites accD (472) and ndhF (97) in the course of molecular evolution. The accD (16) editing site seems to have been lost in S. grandifolia, while is kept in other species of Brassiceae such as in its close relative S. limprichtiana, suggesting a species-specific loss. Similarly, the ndhD (225) editing site seems to have been lost solely in C. abyssinica, but this hypothesis need to be proved by analyses of other species in the genus Crambe. The rps14 (27) editing site seems to have been lost in the ancestral K, since this site is not present in the sisters B. juncea and B. napus. Lastly, two parallel losses seem to have been happened at the psbE (72) site, in the C. arabica branch and in the clade J, which includes the species R. sativus, B. napus, and B. juncea.

On the other hand, the distribution of the rpoB (754) and rpoC2 (765) sites among Brassiceae species suggests that these RNA editing sites arose within this tribe (Fig. 7B). The rpoB (754) mRNA editing is predicted to happen only in C. abyssinica, suggesting that a new site of editing may be arisen in the genus Crambe. This RNA editing site occurs in a highly conserved region of the rpoB gene (Conserved Domains search; https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml). The rpoC2 (765) editing site is predicted to happen in the species R. sativus, B. napus, and B. juncea, which are grouped in the same clade of the common ancestral (J). Although this site occurs in a conserved

region of rpoC2 gene among the Brassiceae species studied here, this region is slightly variable when compared to other species by Conserved Domains search. Both, rpoB and rpoC2 genes encode subunits of the E. coli-like plastid RNA polymerase (PEP) and the inactivation of them produce plants with reduced transcription and photosynthetic capacity (Allison et al., 1996; Serino and Maliga, 1998). Detection of several new RNA editing sites was reported by Fiebig et al. (2004) in the petL gene of different species belonging to various plant taxa. Several RNA editing data show that the RNA editing site evolution occurs more readily in genes or domains of genes whose transitory loss of function can be tolerated (Fiebig et al., 2004). Moreover, the high protein diversity of the editosome complex and the relative flexibility to bind to cis-elements may facilitate the fixation of new RNA editing site (Sun et al., 2016).

The last three RNA editing sites predicted to occur only in some Brassiceae species are matK (45), rpoB (293), and rpoC2 (850). We correlated the relationships among different species of the tribe with the codon usage and RNA editing data (Fig. 7C). Our correlation for the matK (45) editing site would suggest that the distant common ancestral (A) carried the codon UCU and a possible mutation in the first nucleotide position created later the CCU codon and the eventual need for RNA editing in the ancestral J. However, it was lost by a sequential mutation on the second position of the codon creating the codon CUU and further loss the editing site in R. sativus branch. However, it is important to note that this site flanks a highly variable region in MatK protein analyzed by Conserved Domain search and, consequently, it could dispense the essentiality of RNA editing in this site. Similarly, codon diversification is present among rpoC2 mRNAs in the tribe Brassiceae. The editing at site position 850 of rpoC2 transcript is predicted to be edited only in C. abyssinica, which creates the codon UCU (serine) in comparison with the sequence observed in O. diffusus, Raphanus, Sinapis, Sinalliaria and Brassica species, where a TCT is already fixed in the DNA sequence. Nevertheless, at the same amino acid position a tyrosine is present in C. arabica (codon UAU), and a cysteine in O. hupehensis and O. taibaiensis (codon UGU), suggesting that some variability is possible. Indeed, the amino acid at position 850 in RpoC2 protein is highly polymorphic (by Conserved Domain search) and the predicted editing site in C. abyssinica may not be necessary for a functional RpoC2 protein. Lastly, the change CCC to CUC at rpoB (293) transcript is predicted to occur only in B. napus, while in all other species a codon UUG (leucine) is found. In this case, due to genetic code degeneration, both codons codify the same amino acid (leucine). The amino acid located in this position is well conserved (by Conserved Domain search), but some divergences located next to

this site in B. napus (Supplementary Fig. S4) suggest that other modifications may have driven during rpoB gene evolution without the need for RNA editing in this stretch.

Future experiments will be necessary to validate the RNA editing sites predicted here within the tribe Brassiceae, mainly for non-conserved sites, which are interesting sources of information regarding to the evolution of RNA editing and useful to outline relationships among species of the tribe.

4.3. Nucleotide divergence hotspots as a tool for phylogenetic analyses of the tribe Brassiceae

Plastome sequences are valuable source of molecular markers for resolving phylogenetic relationships between closely related taxa (Rogalski et al., 2015; Daniell et al., 2016) as well as for unresolved relationships among related genera in the tribe Brassiceae.

Most part of the genera within tribe Brassiceae is distributed in seven major lineages: Cakile, Crambe, Nigra, Savignya, Rapa/Oleraceae, Vella, and Zilla, while the genera Henophyton, Pseuderucaria, and Orychophragmus remain unrelated (Warwick and Sauder, 2005; Hall et al., 2011). Phylogenetic studies of the tribe Brassiceae based on nuclear sequences (PHYA and ITS) and plastid sequences (matK, trnL intron, and restriction site polymorphism) support the monophyletic origin of the tribe and the inclusion of some controversial genera, but the limits and the relationships among lineages and genera remain unclear, with some genera presenting polyphyletic origins (Warwick and Sauder, 2005; Hall et al., 2011; BrassiBase, 2017).

Six plastid nucleotide divergence hotspots were identified in this study based on sliding window analysis, none of them used before for Brassiceae phylogenetic studies. In addition, a mapping of SSRs loci, within Brassiceae plastomes, showed here suggests some regions in the LSC and SSC as promising molecular markers, complementing previous studies that identified loci of plastid SSRs in Brassica (Hu et al., 2011) and Raphanus (Jeong et al., 2014) plastomes. Furthermore, the most divergent plastid genes identified here, rps16, ycf1, rpl22, and ccsA, together with matK database used in previous analyses (Hall et al., 2011), are attracting sources of useful genetic information to improve the phylogenetic resolution of the tribe Brassiceae.

4.4. Phylogenomic analysis based on whole plastomes of different genera within the family Brassicaceae are well supported

Several phylogenies based on nuclear and plastid nucleotide sequences have been used to resolve the relationships among tribes, genera, and species within the family Brassicaceae (Warwick et al., 2010; Couvreur et al., 2010; Hall et al., 2011; Hohmann et

al., 2015; Huang et al., 2015; Hu et al., 2016; Novikova et al., 2016). The early radiation of the major clades and high speciation rates associated to polyploidization, gene flow, and incomplete lineage sorting events (Baker et al., 2009; Hohmann et al., 2015; Novikova et al., 2016; Guo et al., 2017) make the evolutionary history of Brassicaceae a challenging task.

The most recent phylogeny of Brassicaceae based on plastome sequences was done by Guo et al. (2017) using 51 taxa and a partitioned supermatrix of 77 protein coding genes. Our Brassicaceae phylogenomic represents a modified (based on whole plastomes) and expanded phylogeny of Brassicaceae, including 72 taxa that resulted in enrichment of three genera (*Arabidopsis*, *Brassica*, and *Lobularia*) and inclusion of 13 new genera in comparison with previously work from Guo et al. (2017). Generally, several phylogenetic studies agreed each other to divide the family Brassicaceae in four major lineages: the lineages I, II, and III (i.e., core Brassicaceae), and a basal lineage composed by Aethionemeae tribe (Warwick et al., 2010; Couvreur et al., 2010; Al-Shehbaz, 2012; Guo et al., 2017; BrassiBase, 2017). The phylogeny presented here based on whole-plastomes includes several genera belonging to all of these lineages (Fig. 6) and in accordance with those previous classifications, in our Brassicaceae phylogenomic the genera *Aethionema* (Aethionemeae tribe) is the most basal lineage (Warwick et al., 2010; Couvreur et al., 2010; Al-Shehbaz, 2012) and the lineage I form a sister-group with a clade including the lineage II sister to lineage III (Guo et al., 2017).

Previous phylogeny about the relationships within the lineage I showed that the Lepidideae tribe is sister to Cardamineae tribe (Homann et al., 2015), while in our tree the Cardamineae tribe is well supported as sister to a clade composed by all the other tribes of the lineage I. Guo et al. (2017) showed no distinct bifurcating divergence among the tribes Lepidieae and others of the lineage I, when all mutation sites of the 77 concatenated plastid genes were used. However, when they excluded the third codon position or used only high conserved genes with slow evolving rates located at the IRs, the resulting phylogenies were efficient to resolve the relationships in the lineage I (Guo et al., 2017). When the nucleotide of third position of the codon was excluded, the Cardamineae tribe formed a clade sister to all other tribes corroborating our phylogenomic inference within the lineage I. On the other hand, when they used only IR genes, the tribe Lepidieae formed a clade sister to all other tribes instead. To solve this incongruence, future phylogenies need to be inferred including more taxa within the lineage I. The relationships that we found within the lineage II and III using whole-plastomes, are in accordance to previous phylogenies based on plastid sequence dataset

(Guo et al., 2017), just with the difference that our phylogeny has two new tribes, Anastaticeae (lineage II) and Hesperideae (lineage III), which changed slightly the topology of the tree.

It is worth to note that several incongruences occur in relationships within the core Brassicaceae (lineages I, II, and III) based on plastid (our phylogeny, Hohmann et al., 2015; Guo et al., 2017) and nuclear sequences (Warwick et al., 2010; Huang et al., 2016), indicating complex evolutionary events, as non-bifurcating radiation in the genera Arabidopsis (Novikova et al., 2016) and presumably in Orychophragmus (Hu et al., 2016). In the Brassiceae tribe the phylogenetic inconsistences are also observed between plastid and nuclear sequences (Warwick and Sauder, 2005; Hall et al., 2011) and some genera seem to have polyphyletic origins, for example the genus Brassica (Warwick and Sauder, 2005; Hall et al., 2011), whose polyphyletic origins are corroborated by the phylogeny presented here. Within the tribe Brassiceae, according to our data, three Brassica species (B. juncea, B. rapa, and B. napus) formed a clade with Rhaphanus sativus that is sister to other clade formed by Crambe abyssinica sister to Brassica nigra and Sinapis arvensis. The first clade belongs to the lineage Rapa/Oleraceae, which has different polyphyletic origins, and the second includes the lineages Crambe and Nigra (Warwick and Sauder, 2005; Hall et al., 2011). The monophyly of Crambe lineage is well-resolved, but some phylogenies indicate polyphyly of Nigra lineage (Warwick and Sauder, 2005; Hall et al., 2011).

Lastly, the small genus Sinalliaria, recently described by Zhou et al. (2014), includes only two species and is kept unassigned in tribes according BrassiBase (2017). Our analysis phylogenetic shows Sinalliaria sister to Orychophragmus with high support, forming a clade sister to the other genera of Brassiceae sampled here. Thereby, our data strongly suggest that the genus Sinalliaria belongs to the tribe Brassiceae. In addition, previous works have showed that Sinalliaria is a distinct group related to Orychophragmus based on morphological characters and molecular sequences from nucleus and plastid (Zhou et al., 2014; Zeng et al., 2016). Because of these close relationships among the plastomes of Sinalliaria and the genera of Brassiceae tribe, we included the two species of Sinalliaria in all analyses showed here about the tribe Brassiceae.

4.5 Conclusion

The plastome organization of C. abyssinica is quite similar to other species of the family Brassicaceae. However, analysis of gene divergence among species of the tribe

Brassicea showed high divergent genes such as rps16 and ycf1, being the first lost in several species in other tribes of Brassicaceae family. The data suggest a possible process of degeneration of rps16 gene in the tribe Brassiceae, particularly in C. abyssinica which shows higher divergence value. The RNA editing predictions revealed a highly conservation of RNA editing sites within the tribe Brassiceae and some of them shared with A. thaliana. Nevertheless, among the non-conservative sites, events of losses and gains are suggested. Among Brassiceae species, the locations of hotspots of nucleotide divergence and SSR loci prevail in the regions of single copy. Lastly, our Brassicaceae phylogenomic resulted in a well-supported and well-resolved tree, including several genera belonging to the core Brassicaceae (lineages I, II, and III) and the basal lineage (tribe Aethionemeae), and represents an update of the plastomes available in the organelle genome database (NCBI). Nevertheless, further phylogenetic studies should be performed as the Brassicaceae plastome database is enriched with plastome sequences of other species.

Conflict of Interest

The authors declare that they have no conflict of interest.

Author contribution statement

ASL, TGP, LNV, MPG, RON, EMS, FOP, and MR conceived and designed the research. ASL, TGP, LNV, EMS, FOP, and MR conducted experiments and analyzed the data. MPG, RON, EMS, FOP, and MR contributed with reagents and materials. ASL and MR wrote the manuscript. All authors read and approved the manuscript.

Acknowledgments

This research was supported by the National Council for Scientific and Technological Development, Brazil (CNPq, Grant 459698/2014-1). We are grateful for the scholarships granted by the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES) to TGP and LNV, and those granted by the CNPq to ASL, RON, MPG, FOP and EMS.

References

Allison, L.A., Simon, L.D., Maliga, P., 1996. Deletion of rpoB reveals a second distinct transcription system in plastids of higher plants. EMBO J. 15, 2802-2809.

Al-Shehbaz, I.A., 2012. A generic and tribal synopsis of the Brassicaceae (Cruciferae). Taxon 61, 931-954.

Apel, W., Bock, R., 2009. Enhancement of carotenoid biosynthesis in transplastomic tomatoes by induced lycopene-to-provitamin A conversion. Plant Physiol. 151, 59-66. doi: 10.1104/pp.109.140533

Barker, M.S., Vogel, H., Schranz, M.E., 2009. Paleopolyploidy in the Brassicales: analyses of the Cleome transcriptome elucidate the history of genome duplications in Arabidopsis and other Brassicales. Genome Biol. Evol. 1:391-399. doi: 10.1093/gbe/evp040

Barrett, C.F., Baker, W.J., Comer, J.R., Conran, J.G., Lahmeyer, S.C., Leebens-Mack, J.H., Li, J., Lim, G.S., Mayfield-Jones, D.R., Perez, L., Medina, J., Pires, J.C., Santos, C., Wm Stevenson, D., Zomlefer, W.B., Davis, J.I., 2016. Plastid genomes reveal support for deep phylogenetic relationships and extensive rate variation among palms and other commelinid monocots. New Phytol. 209, 855-870. doi: 10.1111/nph.13617

Benson, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27, 573-580. doi: 10.1093/nar/27.2.573

Besnard, G., Hernández, P., Khadari, B., Dorado, G., Savolainen, V., 2011. Genomic profiling of plastid DNA variation in the Mediterranean olive tree. BMC Plant Biol. 11, 80. doi: 10.1186/1471-2229-11-80

Bock, R., 2015. Engineering Plastid Genomes: Methods, Tools, and Applications in Basic Research and Biotechnology. Annu Rev Plant Biol 66, 211-241. doi: 10.1146/annurev-arplant-050213-040212

Carlsson, A.S., 2009. Plant oils as feedstock alternatives to petroleum – A short survey of potential oil crop platforms. Biochimie 91, 665-670. doi: 10.1016/j.biochi.2009.03.021

Cheng, J., Salentijn, E.M., Huang, B., Denneboom, C., Qi, W., Dechesne, A.C., Krens, F.A., Visser, R.G., van Loo, E.N., 2015. Detection of induced mutations in CaFAD2 genes by next-generation sequencing leading to the production of improved oil composition in Crambe abyssinica. Plant Biotechnol. J. 13 (4), 471-481. doi: 10.1111/pbi.12269

Couvreur, T.L.P., Franzke, A., Al-Shehbaz, I.A., Bakker F.T., Koch, M.A., Mummenhoff, K., 2010. Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae). Mol. Biol. Evol. 27 (1), 55-71. doi: 10.1093/molbev/msp202

Daniell, H., Lin, C.S., Yu, M., Chang, W.J., 2016. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol. 17 (1), 134. doi: 10.1186/s13059-016-1004-2

Darriba, D., Taboada, G.L., Doallo, R., Posada, D., 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat. Methods 9 (8), 772-772. doi: 10.1038/nmeth.2109

Drescher, A., Ruf, S., Calsa, T., Carrer, H., Bock, R., 2000. The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. Plant J. 22 (2), 97-104. doi: 10.1046/j.1365-313x.2000.00722.x

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic. Acids. Res. 32, 1792-1797. doi: 10.1093/nar/gkh340

Fiebig, A., Stegemann, S., Bock, R., 2004. Rapid evolution of RNA editing sites in a small non-essential plastid gene. Nucleic Acids Res. 32, 3615-3622. doi: 10.1093/nar/gkh695

Fleischmann, T.T., Scharff, L.B., Alkatib, S., Hasdorf, S., Schöttler, M.A., Bock, R., 2011. Nonessential Plastid-Encoded Ribosomal Proteins in Tobacco: A Developmental Role for Plastid Translation and Implications for Reductive Genome Evolution. Plant Cell 23, 3137-3155. doi: 10.1105/tpc.111.088906

Freyer, R., Kiefer-Meyer, M.C., Kössel, H., 1997. Occurrence of plastid RNA editing in all major lineages of land plants. Proc. Natl. Acad. Sci. 94, 6285-6290.

Fujii, S., Small, I., 2011. The evolution of RNA editing and pentatricopeptide repeat genes. New Phytol. 191, 37-47. doi: 10.1111/j.1469-8137.2011.03746.x

Greiner, S., Bock, R., 2013. Tuning a ménage à trois: co-evolution and co-adaptation of nuclear and organellar genomes in plants. BioEssays 35, 354-365. doi: 10.1002/bies.201200137

Guo, X., Castillo-Ramírez, S., González, V., Bustos, P., Fernández-Vázquez, J.L., Santamaría, R.I., Arellano, J., Cevallos, M.A., Dávila, G., 2007. Rapid evolutionary change of common bean (Phaseolus vulgaris L) plastome, and the genomic diversification of legume chloroplasts. BMC Genomics 8, 228. doi: 10.1186/1471-2164-8-228

Guo, X., Liu, J., Hao, G., Zhang, L., Mao, K., Zhang, D., Ma, T., Hu, Q., Al-Shehbaz, I.A., Koch, M.A., 2017. Plastome phylogeny and early diversification of Brassicaceae. BMC Genomics 18, 176. doi: 10.1186/s12864-017-3555-3

Hall, J.C., Tisdale, T.E., Donohue, K., Wheeler, A., Al-Yahya, M.A., Kramer, E.M., 2011. Convergent evolution of a complex fruit structure in the tribe Brassiceae (Brassicaceae). Am. J. Bot. 98(12), 1989-2003. doi: 10.3732/ajb.1100203

Hohmann, N., Wolf, E.M., Lysak, M.A., Koch, M.A., 2015. A Time-Calibrated Road Map of Brassicaceae Species Radiation and Evolutionary History. Plant Cell 27(10), 2770-2784. doi: 10.1105/tpc.15.00482

Hu, H., Hu, Q., Al-Shehbaz, I.A., Luo, X., Zeng, T., Guo, X., Liu, J., 2016. Species Delimitation and Interspecific Relationships of the Genus Orychophragmus (Brassicaceae) Inferred from Whole Chloroplast Genomes. Front. Plant Sci. 7, 1826. doi: 10.3389/fpls.2016.01826

Hu, S., Sablok, G., Wang, B., Qu, D., Barbaro, E., Viola, R., Li, M., Varotto, C., 2015. Plastome organization and evolution of chloroplast genes in Cardamine species adapted to contrasting habitats. BMC Genomics 16, 306. doi: 10.1186/s12864-015-1498-0

Hu, Z.Y., Hua, W., Huang, S.M., Wang, H.Z., 2011. Complete chloroplast genome sequence of rapeseed (Brassica napus L.) and its evolutionary implications. Genet. Resour. Crop Evol. 58, 875–887. doi: 10.1007/s10722-010-9626-9

Huang, C.H., Sun, R., Hu, Y., Zeng, L., Zhang, N., Cai, L., Zhang. Q., Koch, M.A., Al-Shehbaz, I., Edger, P.P., Pires, J.C., Tan, D.Y., Zhong, Y., Ma, H., 2016. Resolution of Brassicaceae Phylogeny Using Nuclear Genes Uncovers Nested Radiations and Supports Convergent Morphological Evolution. Mol. Biol. Evol. 33 (2), 394–412. doi: 10.1093/molbev/msv226

Jansen, R.K., Cai, Z., Raubeson, L.A., Daniell, H., Depamphilis, C.W., Leebens-Mack, J., Müller, K.F., Guisinger-Bellian, M., Haberle, R.C., Hansen, A.K., Chumley, T.W., Lee, S.B., Peery, R., McNeal, J.R., Kuehl, J.V., Boore, J.L., 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc. Natl. Acad. Sci. 104 (49), 19369-19374. doi: 10.1073/pnas.0709121104

Jansen, R.K., Saski, C., Lee, S.B., Hansen, A.K., Daniell, H., 2011. Complete plastid genome sequences of three Rosids (Castanea, Prunus, Theobroma): evidence for at least two independent transfers of rpl22 to the nucleus. Mol. Biol. Evol. 28 (1), 835-847. doi: 10.1093/molbev/msq261

Jeong, Y.M., Chung, W.H., Mun, J.H., Kim, N., Yu, H.J., 2014. De novo assembly and characterization of the complete chloroplast genome of radish (Raphanus sativus L.). Gene 551 (1), 39-48. doi: 10.1016/j.gene.2014.08.038

Kagale, S., Robinson, S.J., Nixon, J., Xiao, R., Huebert, T., Condie, J., Kessler, D., Clarke, W.E., Edger, P.P., Links, M.G., Sharpe, A.G., Parkin, I.A.P., 2014. Polyploid evolution of the Brassicaceae during the Cenozoic era. Plant Cell 26, 2777-2791. doi: 10.1105/tpc.114.126391

Kahlau, S., Aspinall, S., Gray, J.C., Bock, R., (2006) Sequence of the tomato chloroplast DNA and evolutionary comparison of solanaceous plastid genomes. J. Mol. Evol. 63, 194-207. doi: 10.1007/s00239-005-0254-5

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol. Biol. Evol. 30, 772-780. doi: 10.1093/molbev/mst010

Karl, R., Koch, M.A., 2013. A world-wide perspective on crucifer speciation and evolution: phylogenetics, biogeography and trait evolution in tribe Arabideae. Ann. Bot. 112, 983-1001. doi: 10.1093/aob/mct165

Kikuchi, S., Bédard, J., Hirano, M., Hirabayashi, Y., Oishi, M., Imai, M., Takase, M., Ide, T., Nakai, M., 2013. Uncovering the protein translocon at the chloroplast inner envelope membrane. Science 339, 571-574. doi: 10.1126/science.1229262

Kumar, S., Hahn, F.M., Baidoo, E., Kahlon, T.S., Wood, D.F., McMahan, C.M., Cornish, K., Keasling, J.D., Daniell, H., Whalen, M.C., 2012. Remodeling the isoprenoid pathway in tobacco by expressing the cytoplasmic mevalonate pathway in chloroplasts. Metab. Eng. 14, 19-28. doi: 10.1016/j.ymben.2011.11.005

Kurtz, S., Choudhuri, J.V., Ohlebusch, E., Schleiermacher, C., Stoye, J., Giegerich, R., 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic. Acids Res. 29, 4633-4642. doi: 10.1093/nar/29.22.4633

Kurtz. S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., Salzberg, S.L., 2004. Versatile and open software for comparing large genomes. Genome Biol. 5, R12. doi: 10.1186/gb-2004-5-2-r12

Lanfear, R., Calcott, B., Ho, S.Y.W., Guindon, S., 2012. PartitionFinder: Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses. Mol. Biol. Evol. 29, 1695-1701. doi: 10.1093/molbev/mss020

Lazzeri, L., DeMattei, F., Bucelli, F., Palmieri, S., 1997. Crambe oil – a potential new hydraulic oil and quenchant. Ind Lubr Tribol 49, 71-77.

Li, X., van Loo, E.N., Gruber, J., Fan, J., Guan, R., Frentzen, M., Stymne, S., Zhu, L.H., 2012. Development of ultra high erucic acid oil in the industrial oil crop Crambe abyssinica. Plant Biotechnol. J. 10 (7), 862-70. doi: 10.1111/j.1467-7652.2012.00709.x

Librado, P., Rozas, J., 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinforma. Oxf. Engl. 25, 1451-1452. doi: 10.1093/bioinformatics/btp187

Lohse, M., Drechsel, O., Kahlau, S., Bock, R., 2013. OrganellarGenomeDRAW--a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res. 41, W575-W581. doi: 10.1093/nar/gkt289

Lowe, T.M., Eddy, S.R., 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25, 955-964.

Lu, Y., Rijzaani, H., Karcher, D., Ruf, S., Bock, R., 2013. Efficient metabolic pathway engineering in transgenic tobacco and tomato plastids with synthetic multigene operons. Proc. Natl. Acad. Sci. 110, E623–E632. doi: 10.1073/pnas.1216898110

Lysak, M.A., Koch, M.A., 2011. Phylogeny, genome, and karyotype evolution of crucifers (Brassicaceae). In: Schmidt, R., Bancroft, I. (Eds.), Genetics and Genomics of the Brassicaceae, Springer, New York, pp. 1-31.

Mastebroek, H.D., Wallenburg, S.C., van Soest, L.J.M., 1994. Variation for agronomic characteristics in crambe (Crambe abyssinica Hochst. ex Fries). Ind. Crop. Prod. 2(2), 129-36.

Mower, J.P., 2009. The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. Nucleic Acids Res. 37, W253-W259. doi: 10.1093/nar/gkp337

Novikova, P.Y., Hohmann, N., Nizhynska, V., Tsuchimatsu,T., Ali, J., Muir, G., Guggisberg, A., Paape, T., Schmid, K., Fedorenko, O.M., Holm, S., Säll, T., Schlötterer, C., Marhold, K., Widmer, A., Sese, J., Shimizu, K.K., Weigel, D., Krämer, U., Koch, M.A., Nordborg, M., 2016. Sequencing of the genus Arabidopsis identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. Nat. Genet. 48, 1077-1082. doi: 10.1038/ng.3617

Prabhudas, S.K., Raju, B., Kannan Thodi, S., Parani, M., Natarajan, P., 2016. The complete chloroplast genome sequence of Indian mustard (Brassica juncea L.). Mitochondrial DNA Part. DNA Mapp. Seq. Anal. 27, 4622-4623. doi: 10.3109/19401736.2015.1101586

Provan, J., Powell, W., Hollingsworth, P.M., 2001. Chloroplast microsatellites: new tools for studies in plant ecology and evolution. Trends Ecol. Evol. 16, 142-147. doi: 10.1016/S0169-5347(00)02097-8

Qiao, J., Cai, M., Yan, G., Wang, N., Li, F., Chen, B., Gao, G., Xu, K., Li, J., Wu, X., 2016. High-throughput multiplex cpDNA resequencing clarifies the genetic diversity and genetic relationships among Brassica napus, Brassica rapa and Brassica oleracea. Plant Biotechnol J. 14, 409-418. doi: 10.1111/pbi.12395

Raubeson, L.A., Peery, R., Chumley, T.W., Dziubek, C., Fourcade, H.M., Boore, J.L., Jansen, R.K., 2007. Comparative chloroplast genomics: analyses including new sequences from the angiosperms Nuphar advena and Ranunculus macranthus. BMC Genomics 8, 174. doi: 10.1186/1471-2164-8-174

Rogalski, M., Carrer, H., 2011. Engineering plastid fatty acid biosynthesis to improve food quality and biofuel production in higher plants: Plastid fatty acid biosynthesis. Plant Biotechnol. J. 9, 554-564. doi: 10.1111/j.1467-7652.2011.00621.x

Rogalski, M., do Nascimento Vieira L., Fraga, H.P., Guerra, M.P., 2015. Plastid genomics in horticultural species: importance and applications for plant population genetics, evolution, and biotechnology. Front. Plant Sci. 6, 586. doi: 10.3389/fpls.2015.00586

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Hohna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P., 2012. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Syst. Biol. 61, 539-542. doi: 10.1093/sysbio/sys029

Rüdinger, M., Polsakiewicz, M., Knoop, V., 2008. Organellar RNA editing and plant-specific extensions of pentatricopeptide repeat proteins in jungermanniid but not in marchantiid liverworts. Mol. Biol. Evol. 25, 1405-1414. doi: 10.1093/molbev/msn084

Ruhfel, B.R., Gitzendanner, M.A., Soltis, P.S., Soltis, D.E., Burleigh, J.G., 2014. From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. BMC Evol. Biol. 14, 23. doi: 10.1186/1471-2148-14-23

Ruwe, H., Castandet, B., Schmitz-Linneweber, C., Stern, D.B., 2013. Arabidopsis chloroplast quantitative editotype. FEBS Lett. 587, 1429-1433. doi: 10.1016/j.febslet.2013.03.022

Seol, Y.J., Kim, K., Kang, S.H., Perumal, S., Lee, J., Kim, C.K., 2017. The complete chloroplast genome of two Brassica species, Brassica nigra and B. Oleracea. Mitochondrial DNA Part DNA Mapp. Seq. Anal. 28, 167-168. doi: 10.3109/19401736.2015.1115493

Serino, G., Maliga, P., 1998. RNA polymerase subunits encoded by the plastid rpo genes are not shared with the nucleus-encoded plastid enzyme. Plant Physiol. 117, 1165-1170.

Sun, T., Bentolila, S., Hanson, M.R., 2016. The Unexpected Diversity of Plant Organelle RNA Editosomes. Trends Plant Sci. 21, 962-973. doi: 10.1016/j.tplants.2016.07.005

Takenaka, M., Zehrmann, A., Verbitskiy, D., Härtel, B., Brennicke, A., 2013. RNA Editing in Plants and Its Evolution. Annu. Rev. Genet. 47, 335-352. doi: 10.1146/annurev-genet-111212-133519

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Mol. Biol. Evol. 30, 2725-2729. doi: 10.1093/molbev/mst197

Tillich, M., Funk, H.T., Schmitz-Linneweber, C., Poltnigg, P., Sabater, B., Martin, M., Maier, R.M., 2005. Editing of plastid RNA in Arabidopsis thaliana ecotypes. Plant J. 43, 708-715. doi: 10.1111/j.1365-313X.2005.02484.x

Thiel, T., Michalek, W., Varshney, R., Graner, A., 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). Theor. Appl. Genet. 106, 411-422. doi: 10.1007/s00122-002-1031-0

Vaidya, G., Lohman, D.J., Meier, R., 2011. SequenceMatrix: concatenation software for the fast assembly of multi-gene datasets with character set and codon information. Cladistics 27, 171-180. doi: 10.1111/j.1096-0031.2010.00329.x

Vieira, L.N., Faoro, H., Rogalski, M., Fraga, H.P., Cardoso, R.L.A., Souza, E.M., Pedrosa F.O., Nodari, R.O., Guerra, M.P., 2014a. The Complete Chloroplast Genome Sequence of Podocarpus lambertii: Genome

Structure, Evolutionary Aspects, Gene Content and SSR Detection. PLoS ONE 9 (3), e90618. doi: 10.1371/journal.pone.0090618

Vieira, L.N., Faoro, H., Fraga, H.P., Rogalski, M., de Souza, E.M., de Oliveira Pedrosa, F., Nodari, R.O., Guerra, M.P., 2014b. An Improved Protocol for Intact Chloroplasts and cpDNA Isolation in Conifers. PLoS ONE 9 (1), e84792. doi: 10.1371/journal.pone.0084792

Vieira, L.N, dos Anjos, K.G., Faoro, H., Fraga, H.P., Greco, T.M., Pedrosa, F.O., de Souza, E.M., Rogalski, M., de Souza, R.F., Guerra, M.P., 2016a. Phylogenetic inference and SSR characterization of tropical woody bamboos tribe Bambuseae (Poaceae: Bambusoideae) based on complete plastid genome sequences. Curr. Genet. 62, 443-453. doi: 10.1007/s00294-015-0549-z

Vieira, L.N., Rogalski, M., Faoro, H., Fraga, H.P., dos Anjos, K.G., Picchi, G.F.A., Nodari, R.O., Pedrosa, F.O., de Souza, E.M., Guerra, M.P., 2016b. The plastome sequence of the endemic Amazonian conifer, Retrophyllum piresii (Silba) C.N.Page, reveals different recombination events and plastome isoforms. Tree Genet. Genomes 12,10. doi: 10.1007/s11295-016-0968-0

Vries, J., Sousa, F.L., Bölter, B., Soll, J., Gould, S.B., 2015. YCF1: A Green TIC? Plant Cell 27, 1827-1833. doi: 10.1105/tpc.114.135541

Warwick, S.I., Sauder, C.A., 2005. Phylogeny of tribe Brassiceae (Brassicaceae) based on chloroplast restriction site polymorphisms and nuclear ribosomal internal transcribed spacer and chloroplast trnL intron sequences. Can. J. Bot. 83, 467-483. doi: 10.1139/b05-021

Warwick, S.I., Mummenhoff, K., Sauder, C.A., Koch, M.A., Al-Shehbaz, I.A., 2010. Closing the gaps: phylogenetic relationships in the Brassicaceae based on DNA sequence data of nuclear ribosomal ITS region. Plant Syst. Evol. 285, 209-232. doi: 10.1007/s00606-010-0271-8

Wicke, S., Schneeweiss, G.M., de Pamphilis, C.W., Müller, K.F., Quandt, D., 2011. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol. Biol. 76, 273-297. doi: 10.1007/s11103-011-9762-4

Wyman, S.K., Jansen, R.K., Boore, J.L., 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20, 3252-3255. doi: 10.1093/bioinformatics/bth352

Xi, Z., Ruhfel, B.R., Schaefer, H., Amorim, A.M., Sugumaran, M., Wurdack, K.J., Endress, P.K., Matthews, M.L., Stevens, P.F., Mathews, S., Davis, C.C., 2012. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. Proc. Natl. Acad. Sci. 109, 17519-17524. doi: 10.1073/pnas.1205818109

Xu, J.H., Liu, Q., Hu, W., Wang, T., Xue, Q., Messing, J., 2015. Dynamics of chloroplast genomes in green plants. Genomics 106, 221-231. doi: 10.1016/j.ygeno.2015.07.004

Zeng, T., Hu, H., Guo, X., Hu, Q., 2016. The complete chloroplast genomes of two Sinalliaria species and species delimitation (Brassicaceae). Conservation Genet. Resour. 8, 379-381. doi: 0.1007/s12686-016-0563-6

Zhou, Y.Y., Zhang, H.W., Hu, J.Q., Jin, X.F., 2014. Sinalliaria, a new genus of Brassicaceae from eastern China, based on morphological and molecular data. Phytotaxa 186 (4). doi: 10.11646/phytotaxa.186.4.2

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

Zhu, L.H., Krens, F., Smith, M.A. Li, X., Qi, W., van Loo, E.N., Iven, T., Feussner, I., Nazarenus, T.J., Huai, D., Taylor, D.C., Zhou, X.R., Green, A.G., Shockey, J., Klasson, K.T., Mullen, R.T., Huang, B., Dyer, J.M., Cahoon, E.B., 2016b. Dedicated Industrial Oilseed Crops as Metabolic Engineering Platforms for Sustainable Industrial Feedstock Production. Sci. Rep. 6, 22181. doi: 10.1038/srep22181. doi: 10.1038/srep22181

FIGURES

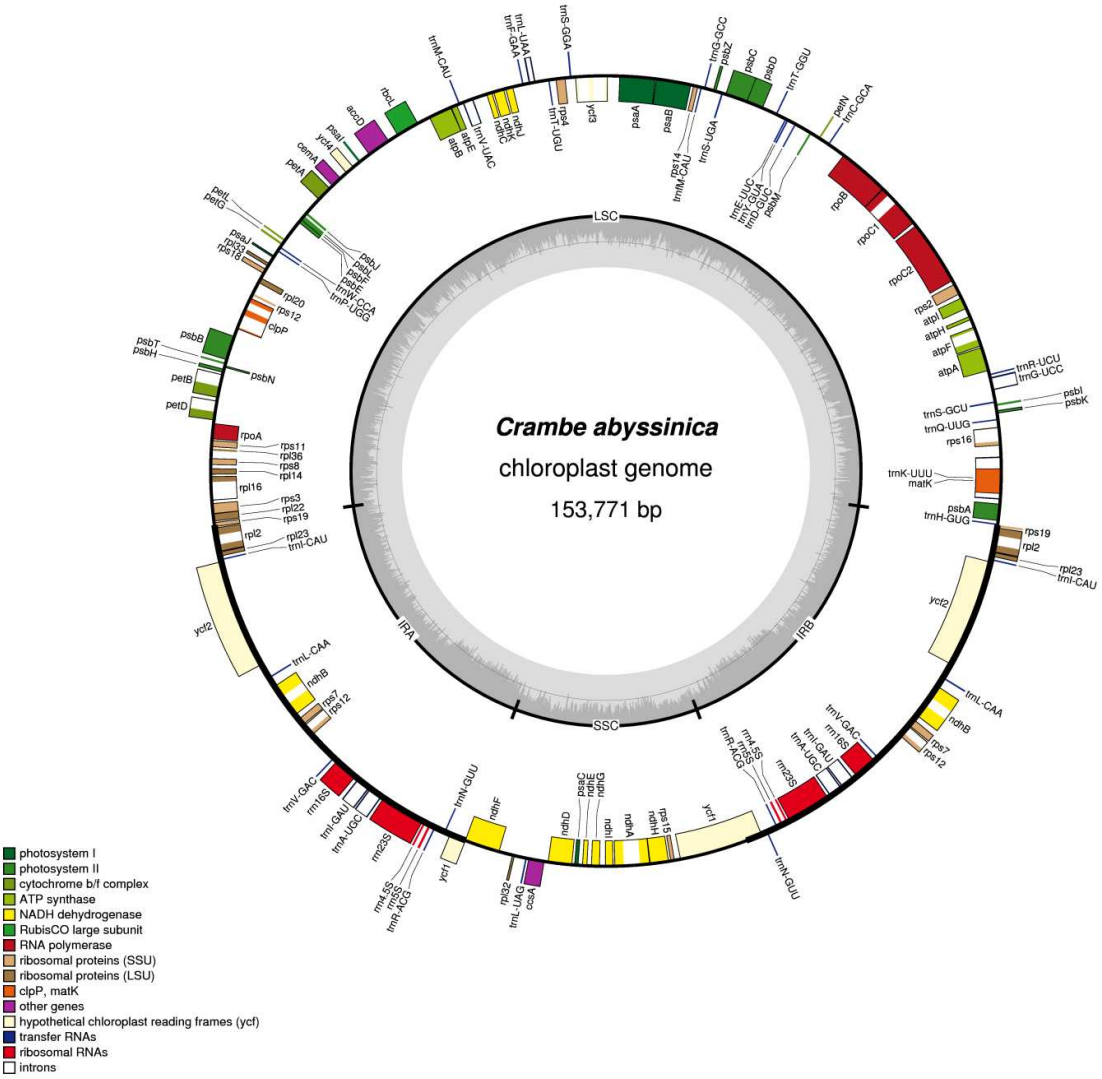


Fig. 1 Gene map of Crambe abyssinica plastome. Genes drawn inside the circle are transcribed in the clockwise direction, and genes drawn outside are transcribed in the counterclockwise direction. Different functional groups of genes are color-coded. The darker gray in the inner circle corresponds to GC content, and the lighter gray corresponds to AT content. LSC, Large Single Copy; SSC, Small Single Copy; IRA/B, Inverted Repeat A/B

Fig. 2 Divergence of the plastid protein coding genes among species of the tribe Brassiceae. The gene divergence was estimated by the sum of total branch lengths in each gene tree inferred

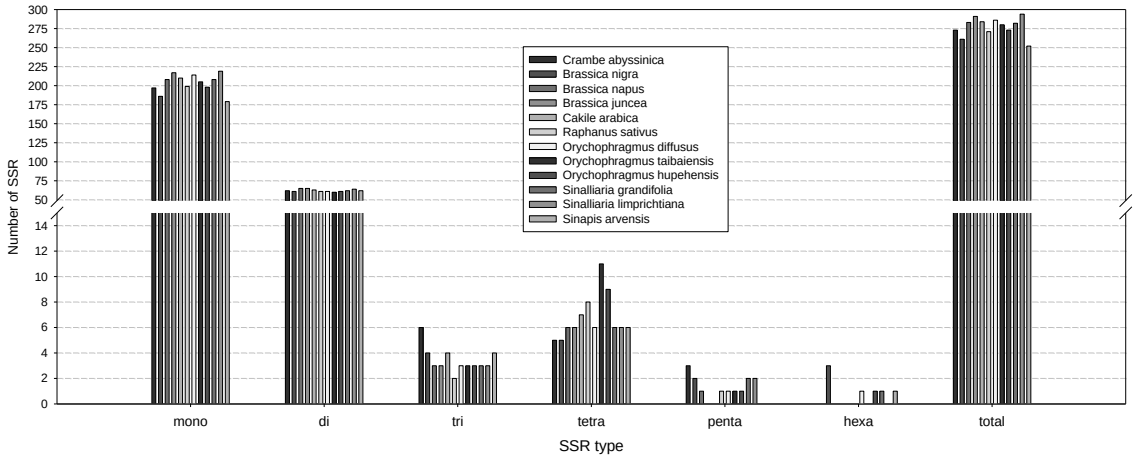


Fig. 3 Number of SSR loci among Brassiceae species. It was set eight repeat units for mononucleotide SSRs, four repeat units for di- and trinucleotide SSRs, and three repeat units for tetra-, penta- and hexanucleotide SSRs

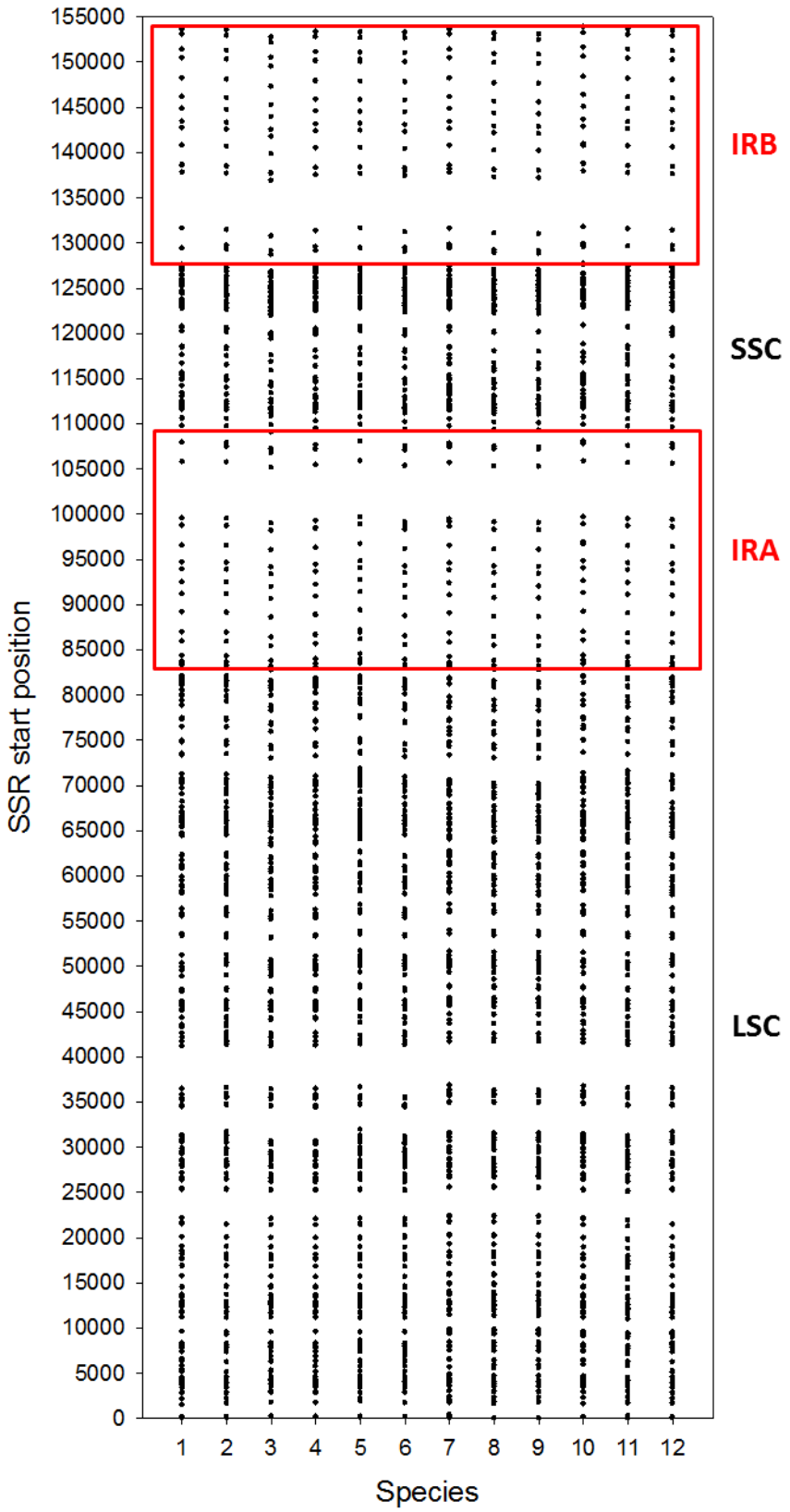


Fig. 4 Distribution of SSRs into plastomes of the tribe Brassiceae. Species sampled: 1) Crambe abyssinica; 2) Brassica nigra; 3) B. napus; 4) B. juncea; 5) Cakile arabica; 6) Raphanus sativus; 7) Orychophragmus diffusus; 8) O. taibaiensis; 9) O. hupehensis; 10) Sinalliaria grandifolia; 11) S. limprichtiana; 12) Sinapis arvensis

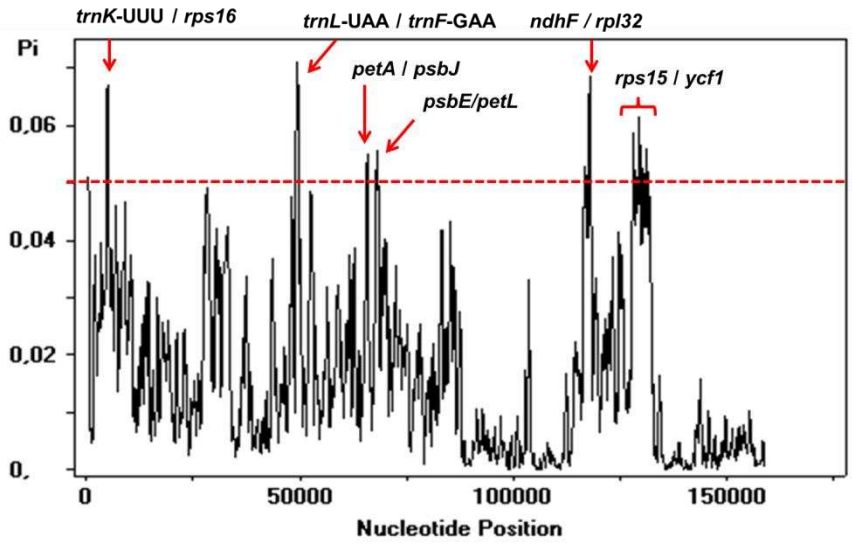


Fig. 5 Sliding window analysis of aligned whole-plastomes of the tribe Brassiceae. The regions with high nucleotide variability (Pi > 0.050) are indicated. Pi, nucleotide diversity of each window. Window length, 400 pb. Step size, 100 pb

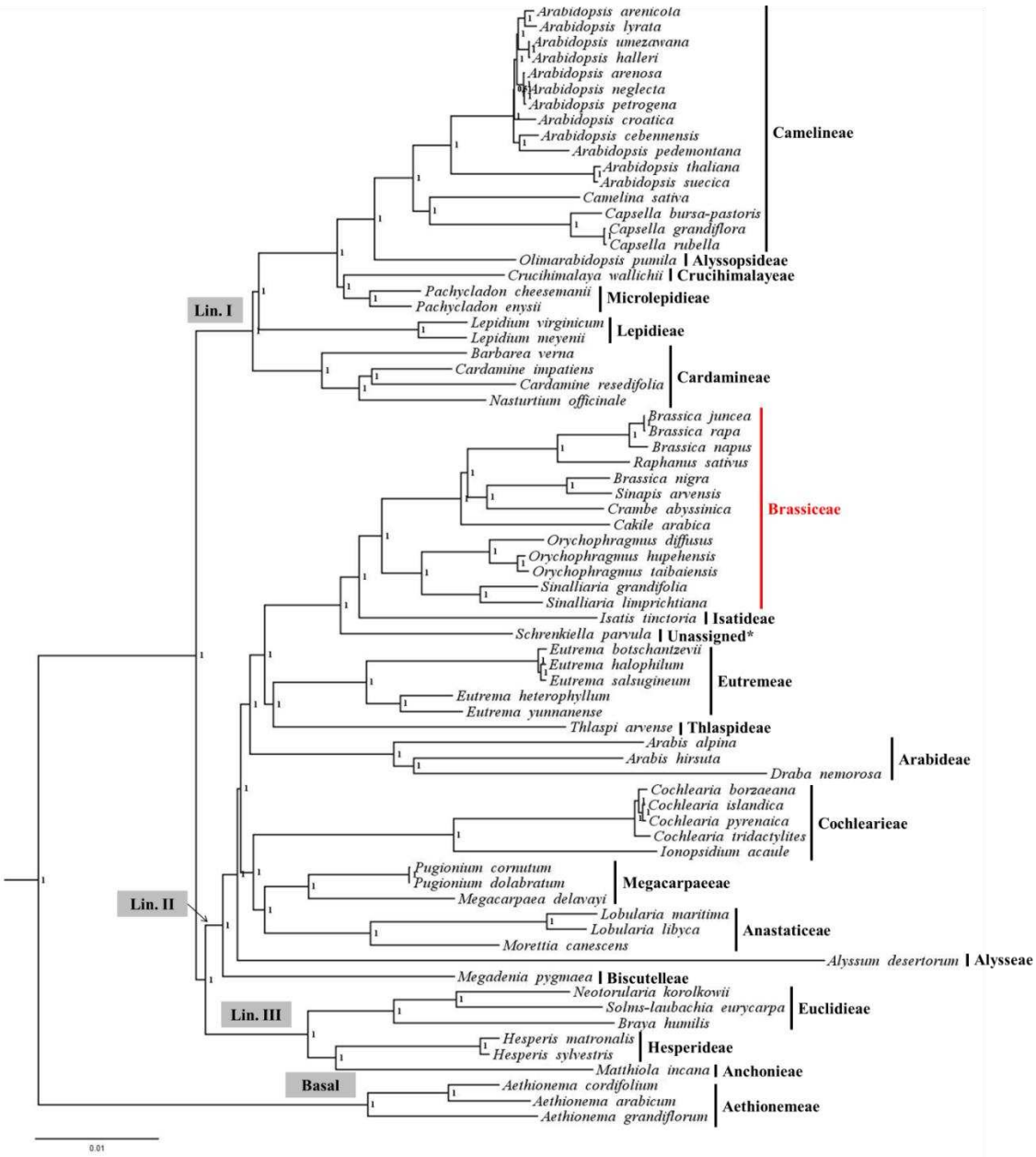


Fig. 6 Brassicaceae phylogenomic tree of 72 taxa based on whole-plastomes using bayesian inference. The numbers at the nodes are Bayesian posterior probabilities. The branch length is proportional to the inferred divergence level and the scale bar indicates the number of inferred nucleic acids substitutions per site. The Brassicales Carica papaya (Caricaceae) and Tarenaya hassleriana (Cleomaceae) were used as out-group to root the tree

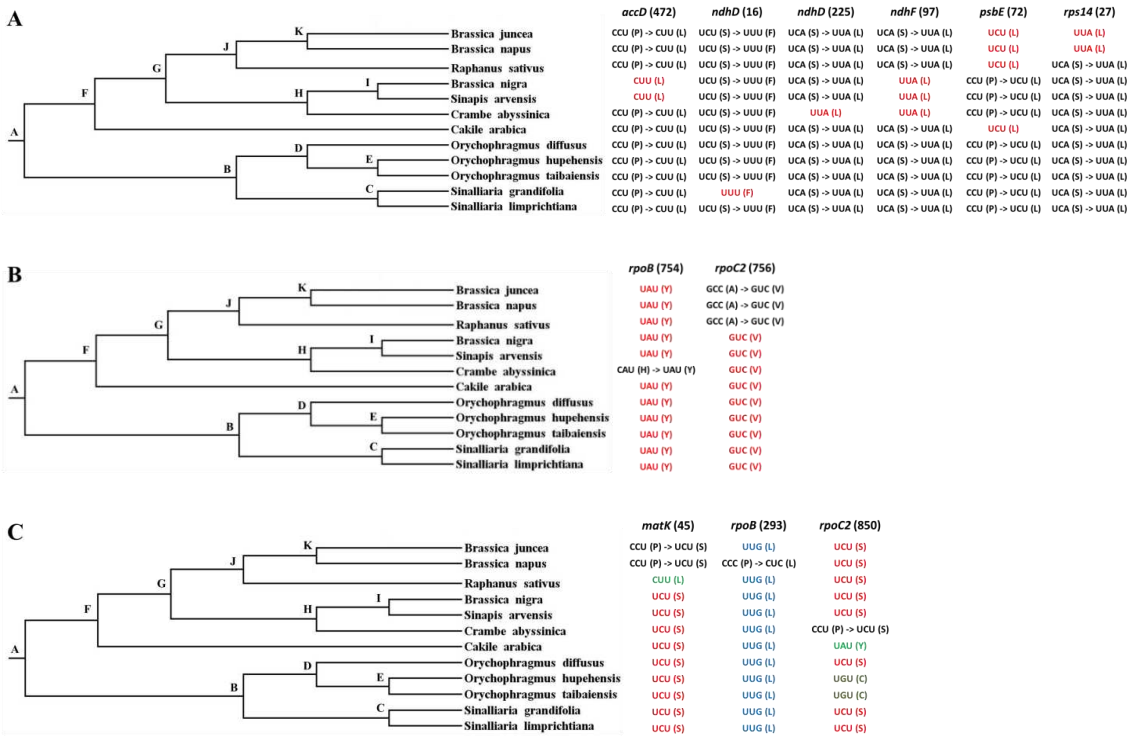


Fig. 7 Correlation between plastidial phylogenomic and the distribution of the RNA editing sites predicted that were not unanimous among the species of Brassiceae tribe. The codons highlighted in colors are codons without editing sites; red for codons that have a T fixed in the plastome, blue for codon variation but conserved amino acid, and light/dark green for non-conserved codons and amino acids. The letters on the nodes of the tree represent the ancestors. A) Possible events of losses of RNA editing sites; B) Possible events of gains of RNA editing sites; C) RNA editing sites in positions without codon conservation among all species of the tribe.

TABLES

Table 1. List of genes identified in Crambe abyssinica plastome

Group of gene	Name of gene
Gene expression. Machinery	
Ribosomal RNA genes	rrn16[b]; rrn23[b]; rrn5[b]; rrn4.5[b]
Transfer RNA genes	trnA –UGC[ab]; trnC –GCA; trnD –GUC; trnE –UUC; trnF –GAA; trnfM –CAU; trnG –UCC[a]; trnG –GCC; trnH –GUG; trnI –CAU[b]; trnI –GAU[ab]; 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
Small subunit of ribosome	rps2; rps3; rps4; rps7[b]; rps8; rps11; rps12[ab]; rps14; rps15; rps16[a]; rps18; rps19[c]
Large subunit of ribosome	rpl2[ab]; rpl14; rpl16[a]; rpl20; rpl22; rpl23[b]; rpl32; rpl33; rpl36
DNA-dependent RNA polymerase	rpoA; rpoB; rpoC1[a]; rpoC2
Genes for photosynthesis	
Subunits of photosystem I (PSI)	psaA; psaB; psaC; psaI; psaJ; ycf3[a]; ycf4
Subunits of photosystem II (PSII)	psbA; psbB; psbC; psbD; psbE; psbF; psbH; psbI; psbJ; psbK; psbL; psbM; psbN; psbT; psbZ
Subunits of cytochrome b_6f	petA; petB[a]; petD[a]; petG; petL; petN
Subunits of ATP synthase	atpA; atpB; atpE; atpF[a]; atpH; atpI
Subunits of NADH dehydrogenase	ndhA[a]; ndhB[ab]; ndhC; ndhD; ndhE; ndhF; ndhG; ndhH; ndhI; ndhJ; ndhK
Large subunit of Rubisco	rbcL
Others genes	
Maturase	matK
Envelope membrane protein	cemA
Subunit of acetyl-CoA carboxylase	accD
C-type cytochrome synthesis gene	ccsA
Protease	clpP[a]
Component of TIC complex	ycf1[c]
Genes of unknown function	ycf2[b]

[a]Genes containing introns; [b]Duplicated gene; [c]Partial duplicated genes

Table 2. Summary of the plastome characteristics among species within the tribe Brassiceae

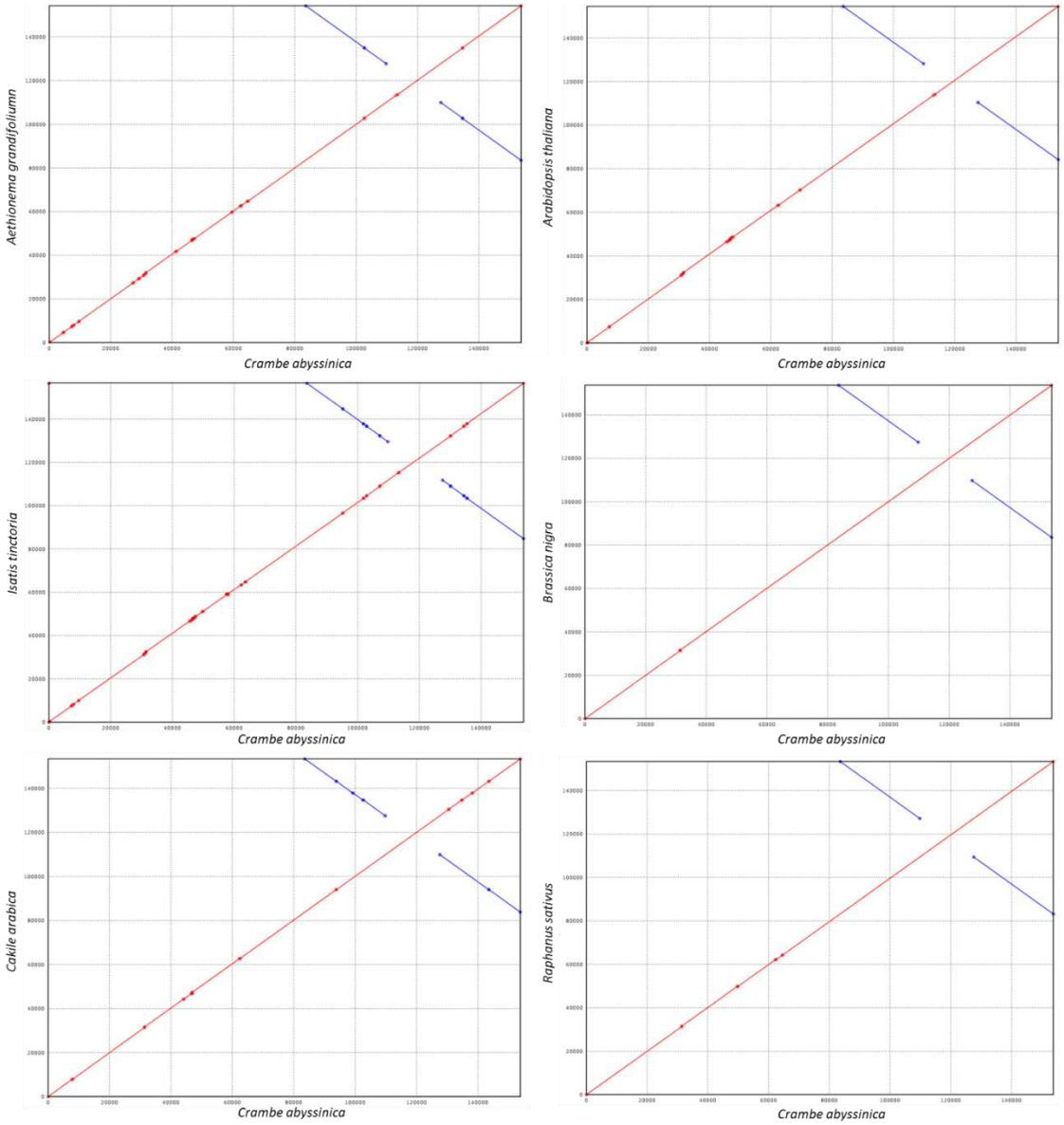
	Size (Kb)	LSC (Kb)	IRs (Kb)	SSC (Kb)	GC%
Crambe abyssinica	153,771	83,600	26,195	17,782	36.37
Brassica juncea	153,483	83,285	26,211	17,776	36.36
Brassica napus	152,860	83,030	26,035	17,760	36.32
Brassica nigra	153,633	83,552	26,193	17,695	36.39
Brassica rapa	153,482	83,281	26,213	17,775	36.36
Cakile arabica	153,378	83,800	25,819	17,940	36.41
Orychophragmus diffusus	153,777	83,456	26,255	17,811	36.29
Orychophragmus hupehensis	153,184	83,060	26,224	17,676	36.35
Orychophragmus taibaiensis	153,255	83,106	26,233	17,683	36.34
Raphanus sativus	153,368	83,170	26,217	17,764	36.34
Sinalliaria grandifolia	154,113	83,820	26,253	17,788	36.01
Sinalliaria limprichtiana	154,060	83,749	26,274	17,768	36.05
Sinapis arvensis	153,590	83,391	26,240	17,719	36.31

Table 3. List of RNA editing sites predicted by PREP program. All species analyzed belong to the tribe Brassiceae. Orychophragmus diffusus (OD), O. hupehensis (OH), O. taibaiensis (OT), Sinalliaria grandifolia (SG), S. limprichtiana (SL), Cakile arabica (CAr), Crambe abyssinica (CAb), Brassica nigra (BNi), Sinapis arvensis (SA), B. juncea (BJ), B. napus (BNa)

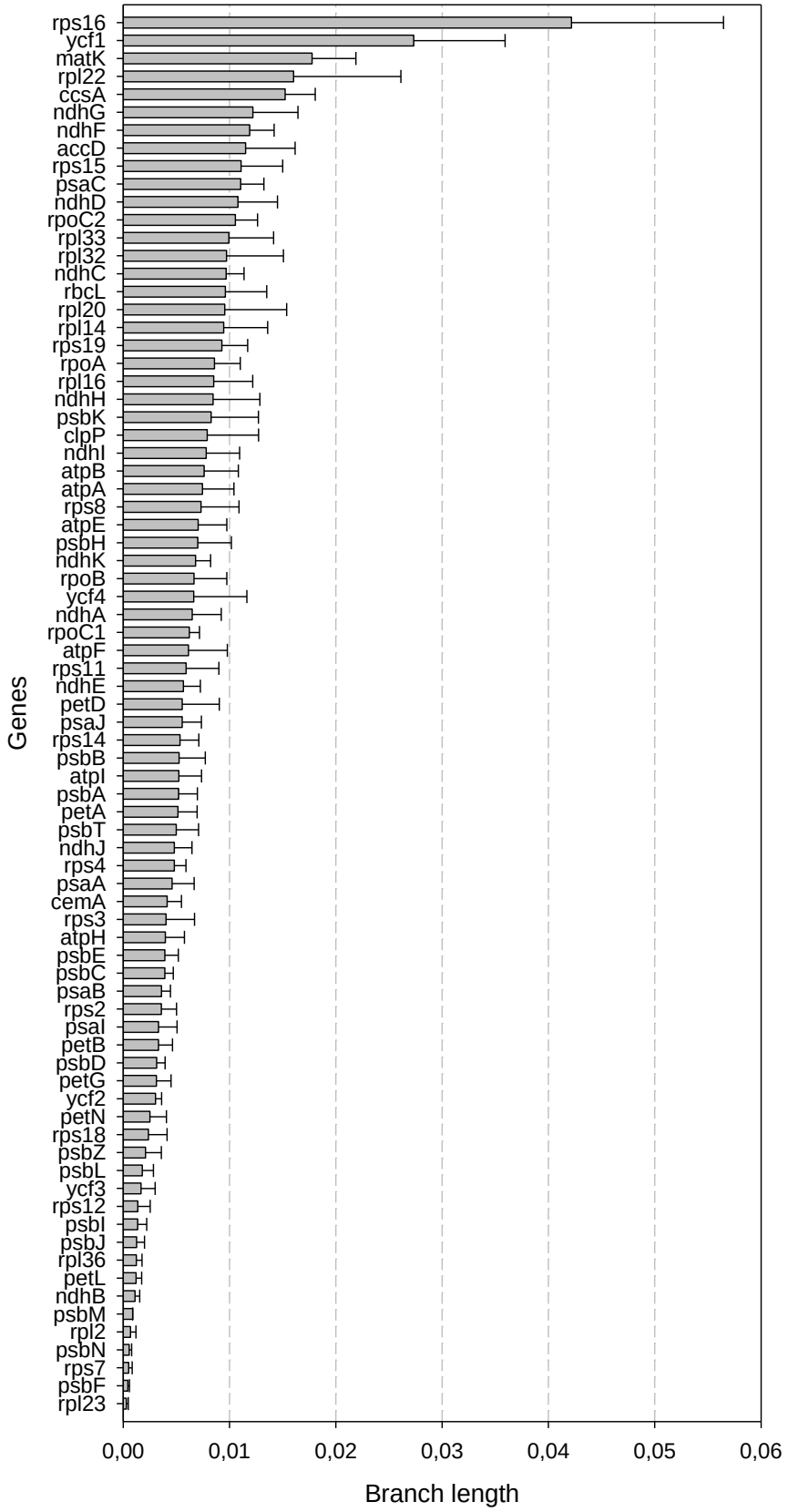
				AA Position											
Gene	**Align position**	**Codon position**	**Effect**	**OD**	**OH**	**OT**	**SG**	**SL**	**CAr**	**CAb**	**BNi**	**SA**	**RS**	**BJ**	**BNa**
accD	256	1	CAT (H) => TAT (Y)	248	248	248	248	248	248	250	250	248	252	248	250
	269	2	TCG (S) => TTG (L)	261	261	261	261	261	261	263	263	261	265	261	263
	472	1	CCT (P) => CTT (L)	464	464	464	464	464	464	466	-	-	468	464	466
atpF	31	2	CCA (P) => CTA (L)	31	31	31	31	31	31	31	31	31	31	31	31
clpP	188	1	CAT (H) => TAT (Y)	187	187	187	187	187	188	188	187	187	187	188	187
matK	45	1	CCT (P) => TCT (S)	-	-	-	-	-	-	-	-		-	45	45
	212	1	CAT (H) => TAT (Y)	212	212	212	212	212	212	212	212	212	212	212	212
	393	2	TCA (S) => TTA (L)	393	393	393	393	393	393	393	393	393	393	393	393
ndhA	42	2	ACA (T) => ATA (I)	42	42	42	42	42	42	42	42	42	42	42	42
	114	2	TCA (S) => TTA (L)	114	114	114	114	114	114	114	114	114	114	114	114
ndhB	50	2	TCA (S) => TTA (L)	50	50	50	50	50	50	50	50	50	50	50	50
	156	2	CCA (P) => CTA (L)	156	156	156	156	156	156	156	156	156	156	156	156
	196	1	CAT (H) => TAT (Y)	196	196	196	196	196	196	196	196	196	196	196	196
	204	2	TCA (S) => TTA (L)	204	204	204	204	204	204	204	204	204	204	204	204
	249	2	TCT (S) => TTT (F)	249	249	249	249	249	249	249	249	249	249	249	249
	277	2	TCA (S) => TTA (L)	277	277	277	277	277	277	277	277	277	277	277	277
	419	1	CAT (H) => TAT (Y)	419	419	419	419	419	419	419	419	419	419	419	419
	494	2	CCA (P) => CTA (L)	494	494	494	494	494	494	494	494	494	494	494	494
ndhD	1	2	ACG (T) => ATG (M)	1	1	1	1	1	1	1	1	1	1	1	1
	16	2	TCT (S) => TTT (F)	16	16	16	-	16	16	16	16	16	16	16	16
	128	2	TCG (S) => TTG (L)	128	128	128	128	128	128	128	128	128	128	128	128
	225	2	TCA (S) => TTA (L)	225	225	225	225	225	225	-	225	225	225	225	225
	293	2	TCA (S) => TTA (L)	293	293	293	293	293	293	293	293	293	293	293	293
	296	2	CCC (P) => CTC (L)	296	296	296	296	296	296	296	296	296	296	296	296
	437	2	TCA (S) => TTA (L)	437	437	437	437	437	437	437	437	437	437	437	437
	469	1	CTT (L) => TTT (F)	469	469	469	469	469	469	469	469	469	469	469	469
ndhF	69	1	CAT (H) => TAT (Y)	69	69	69	69	69	69	69	69	69	69	69	69
	97	2	TCA (S) => TTA (L)	97	97	97	97	97	97	-	-	-	97	97	97
	196	1	CTT (L) => TTT (F)	196	196	196	196	196	196	196	196	196	196	196	196
	712	2	ACA (T) => ATA (I)	712	712	712	712	712	712	712	712	712	712	712	712
ndhG	56	1	CAT (H) => TAT (Y)	56	56	56	56	56	56	56	56	56	56	56	56
	105	2	ACA (T) => ATA (I)	105	105	105	105	105	105	105	105	105	105	105	105
petD	102	2	GCT (A) => GTT (V)	102	102	102	102	102	102	102	102	102	102	102	102
petG	5	2	TCT (S) => TTT (F)	5	5	5	5	5	5	5	5	5	5	5	5
psbE	72	1	CCT (P) => TCT (S)	72	72	72	72	72	-	72	72	72	-	-	-
psbF	26	2	TCT (S) => TTT (F)	26	26	26	26	26	26	26	26	26	26	26	26
rpoB	113	2	TCT (S) => TTT (F)	113	113	113	113	113	113	113	113	113	113	113	113
	184	2	TCA (S) => TTA (L)	184	184	184	184	184	184	184	184	184	184	184	184
	189	2	TCG (S) => TTG (L)	189	189	189	189	189	189	189	189	189	189	189	189
	293	2	CCC (P) => CTC (L)	-	-	-	-	-	-	-	-	-	-	-	293
	325	1	CTC (L) => TTC (F)	325	325	325	325	325	325	325	325	325	325	325	325
	754	1	CAT (H) => TAT (Y)	-	-	-	-	-	-	754	-	-	-	-	-
	811	2	TCA (S) => TTA (L)	811	811	811	811	811	811	811	811	811	811	811	811
rpoC1	651	2	ACT (T) => ATT (I)	647	647	647	647	647	651	647	647	647	647	651	647
rpoC2	765	2	GCC (A) => GTC (V)	-	-	-	-	-	-	-	-	-	763	763	763
	767	1	CGG (R) => TGG (W)	765	765	765	765	765	765	765	767	765	765	765	765
	783	1	GCC (A) => GTC (V)	781	781	781	781	781	781	781	783	781	781	781	781
	850	1	CCT (P) => TCT (S)	-	-	-	-	-	-	848	-	-	-	-	-
rps14	27	2	TCA (S) => TTA (L)	27	27	27	27	27	27	27	27	27	27	-	-
	50	2	CCA (P) => CTA (L)	50	50	50	50	50	50	50	50	50	50	50	50
rps16	71	2	TCA (S) => TTA (L)	71	71	71	71	71	71	71	71	71	71	71	71

SUPPLEMENTARY MATERIAL

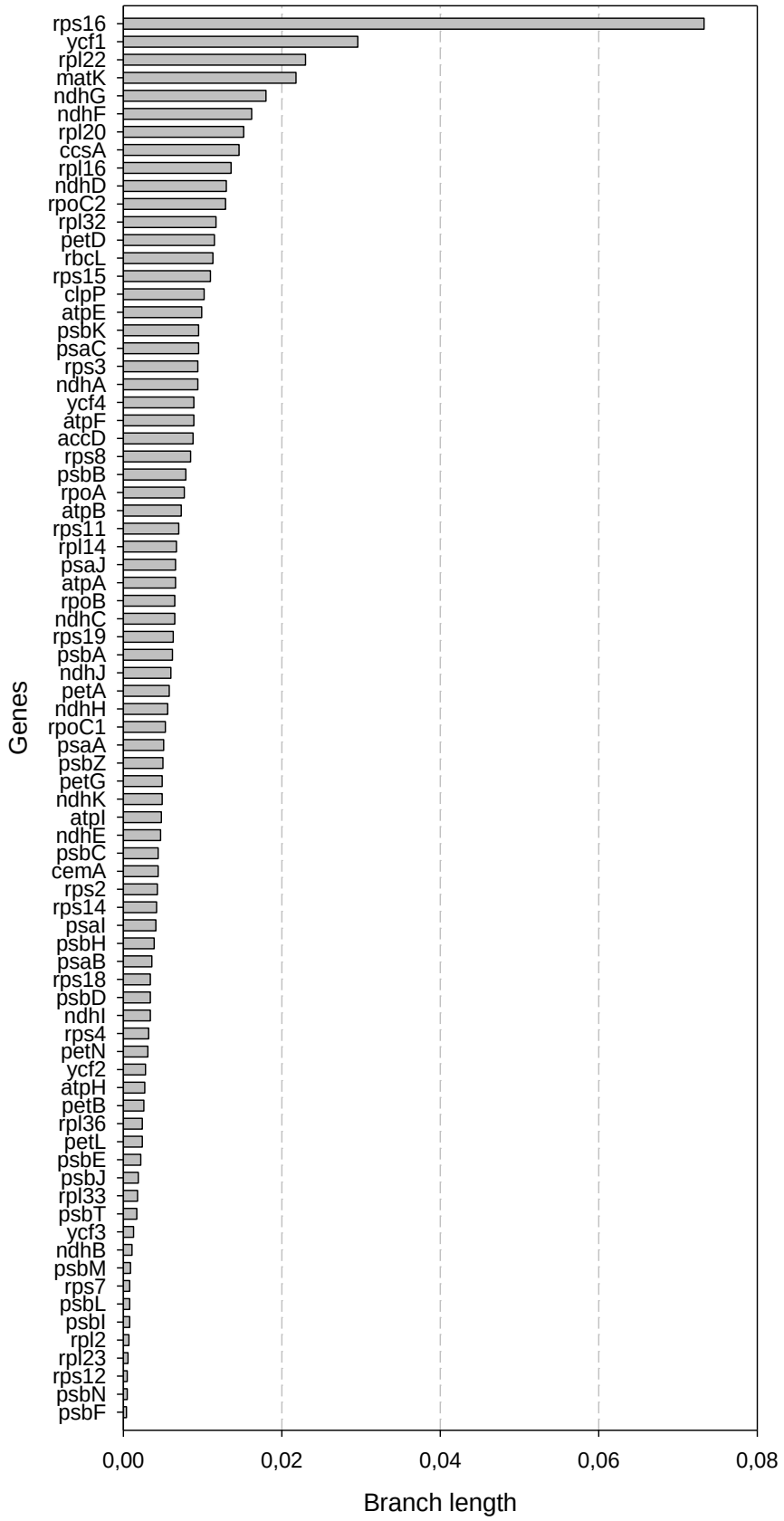
Supplementary Figures



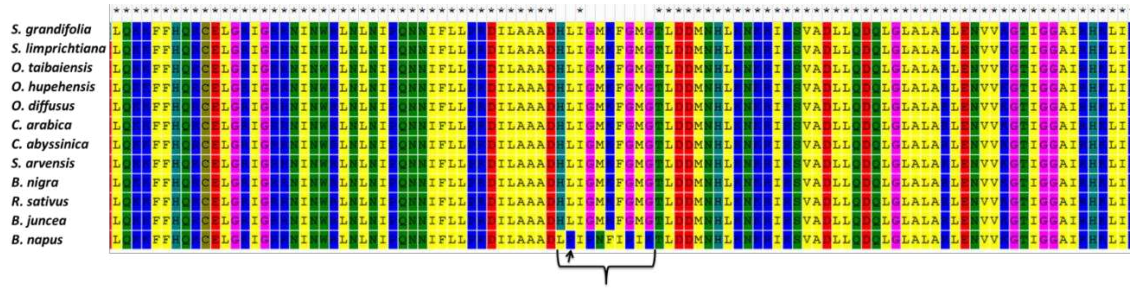
Supplementary Fig. S1. Dot-plot analyses of Crambe abyssinica (X-axis) plastome against selected species (Y-axis) within Brassicaceae family. A positive slope denotes that the pair of sequences compared is in the same orientation. A negative slope denotes that the pair of sequences compared can be aligned, but their orientation is opposite. Sequences in the same direction are red and inversions are blue



Supplementary Fig. S2. Divergence of the plastid protein coding genes within the tribe Brassiceae. The gene divergence was estimated by the sum of total branch lengths in each gene tree inferred. Mean ± SD



Supplementary Fig. S3. Divergence of the plastid protein coding genes of C. abyssinica. The gene divergence was estimated by the sum of total branch lengths in each gene tree inferred



Supplementary Fig. S4. Alignment among RpoB proteins of Brassiceae species. The arrow points to the site 293 (without RNA editing). The region highlighted is very variable in B. napus and highly conserved in the other Brassiceae species

Supplementary Tables

Supplementary Table S1. List of species included in the Brassicaceae phylogenomic

Specie	Genus	Tribe	Family	GenBank
*Carica papaya**	*Carica*	-	Caricaceae	NC_010323.1
*Tarenaya hassleriana**	*Tarenaya*	-	Cleomaceae	NC_034364.1
Aethionema cordifolium	*Aethionema*	Aethionemeae	Brassicaceae	NC_009265.1
Aethionema grandiflorum	*Aethionema*	Aethionemeae	Brassicaceae	NC_009266.1
Aethionema arabicum	*Aethionema*	Aethionemeae	Brassicaceae	NC_034367.1
Alyssum desertorum	*Alyssum*	Alysseae	Brassicaceae	NC_034299.1
Olimarabidopsis pumila	*Olimarabidopsis*	Alyssopsideae	Brassicaceae	NC_009267.1
Lobularia marítima	*Lobularia*	Anastaticeae	Brassicaceae	NC_009274.1
Lobularia libyca	*Lobularia*	Anastaticeae	Brassicaceae	NC_035513.1
Morettia canescens	*Morettia*	Anastaticeae	Brassicaceae	NC_035514.1
Matthiola incana	*Matthiola*	Anchonieae	Brassicaceae	NC_034358.1
Arabis alpina	*Arabis*	Arabideae	Brassicaceae	NC_023367.1
Arabis hirsuta	*Arabis*	Arabideae	Brassicaceae	NC_009268.1
Draba nemorosa	*Draba*	Arabideae	Brassicaceae	NC_009272.1
Megadenia pygmaea	*Megadenia*	Biscutelleae	Brassicaceae	NC_034357.1
Crambe abyssinica	*Crambe*	Brassiceae	Brassicaceae	KY883663
Brassica juncea	*Brassica*	Brassiceae	Brassicaceae	NC_028272.1
Brassica napus	*Brassica*	Brassiceae	Brassicaceae	NC_016734.1
Brassica nigra	*Brassica*	Brassiceae	Brassicaceae	NC_030450.1
Brassica rapa	*Brassica*	Brassiceae	Brassicaceae	NC_015139.1
Cakile arabica	*Cakile*	Brassiceae	Brassicaceae	NC_030775.1
Orychophragmus diffusus	*Orychophragmus*	Brassiceae	Brassicaceae	NC_033498.1
Orychophragmus hupehensis	*Orychophragmus*	Brassiceae	Brassicaceae	NC_033500.1
Orychophragmus taibaiensis	*Orychophragmus*	Brassiceae	Brassicaceae	NC_033499.1
Raphanus sativus	*Raphanus*	Brassiceae	Brassicaceae	NC_024469.1
Sinapis arvensis	*Sinapis*	Brassiceae	Brassicaceae	NC_035303.1
Sinalliaria limprichtiana	*Sinalliaria*	Brassiceae	Brassicaceae	NC_034287.1
Sinalliaria grandifolia	*Sinalliaria*	Brassiceae	Brassicaceae	NC_034286.1
Arabidopsis arenicola	*Arabidopsis*	Camelineae	Brassicaceae	NC_030346.1
Arabidopsis arenosa	*Arabidopsis*	Camelineae	Brassicaceae	NC_029334.1
Arabidopsis cebennensis	*Arabidopsis*	Camelineae	Brassicaceae	NC_029335.1
Arabidopsis croatica	*Arabidopsis*	Camelineae	Brassicaceae	NC_030347.1
Arabidopsis neglecta	*Arabidopsis*	Camelineae	Brassicaceae	NC_030348.1
Arabidopsis pedemontana	*Arabidopsis*	Camelineae	Brassicaceae	NC_029336.1
Arabidopsis petrogena	*Arabidopsis*	Camelineae	Brassicaceae	NC_030349.1
Arabidopsis suecica	*Arabidopsis*	Camelineae	Brassicaceae	NC_030350.1
Arabidopsis thaliana	*Arabidopsis*	Camelineae	Brassicaceae	NC_000932.1
Arabidopsis umezawana	*Arabidopsis*	Camelineae	Brassicaceae	NC_030351.1
Arabidopsis halleri	*Arabidopsis*	Camelineae	Brassicaceae	NC_034366.1
Arabidopsis lyrata	*Arabidopsis*	Camelineae	Brassicaceae	NC_034365.1
Camelina sativa	*Camelina*	Camelineae	Brassicaceae	NC_029337.1
Capsella bursa-pastoris	*Capsella*	Camelineae	Brassicaceae	NC_009270.1
Capsella grandiflora	*Capsella*	Camelineae	Brassicaceae	NC_028517.1
Capsella rubella	*Capsella*	Camelineae	Brassicaceae	NC_027693.1
Barbarea verna	*Barbarea*	Cardamineae	Brassicaceae	NC_009269.1
Cardamine impatiens	*Cardamine*	Cardamineae	Brassicaceae	NC_026445.1
Cardamine resedifolia	*Cardamine*	Cardamineae	Brassicaceae	NC_026446.1
Nasturtium officinale	*Nasturtium*	Cardamineae	Brassicaceae	NC_009275.1
Cochlearia borzaeana	*Cochlearia*	Cochlearieae	Brassicaceae	NC_029253.1
Cochlearia islandica	*Cochlearia*	Cochlearieae	Brassicaceae	NC_029254.1
Cochlearia pyrenaica	*Cochlearia*	Cochlearieae	Brassicaceae	NC_029331.1
Cochlearia tridactylites	*Cochlearia*	Cochlearieae	Brassicaceae	NC_029332.1
Ionopsidium acaule	*Ionopsidium*	Cochlearieae	Brassicaceae	NC_029333.1

Crucihimalaya wallichii	*Crucihimalaya*	Crucihimalayeae	Brassicaceae	NC_009271.1
Braya humilis	*Braya*	Euclidieae	Brassicaceae	NC_035515.1
Neotorularia korolkowii	*Neotorularia*	Euclidieae	Brassicaceae	NC_034361.1
Solms-laubachia eurycarpa	*Solms-laubachia*	Euclidieae	Brassicaceae	NC_034359.1
Eutrema botschantzevii	*Eutrema*	Eutremeae	Brassicaceae	NC_029379.1
Eutrema halophilum	*Eutrema*	Eutremeae	Brassicaceae	NC_029378.1
Eutrema heterophyllum	*Eutrema*	Eutremeae	Brassicaceae	NC_028728.1
Eutrema salsugineum	*Eutrema*	Eutremeae	Brassicaceae	NC_028170.1
Eutrema yunnanense	*Eutrema*	Eutremeae	Brassicaceae	NC_028727.1
Hesperis sylvestris	*Hesperis*	Hesperideae	Brassicaceae	NC_035512.1
Hesperis matronalis	*Hesperis matronalis*	Hesperideae	Brassicaceae	NC_035511.1
Isatis tinctoria	*Isatis*	Isatideae	Brassicaceae	NC_028415.1
Lepidium virginicum	*Lepidium*	Lepidieae	Brassicaceae	NC_009273.1
Lepidium meyenii	*Lepidium*	Lepidieae	Brassicaceae	NC_034363.1
Pugionium cornutum	*Pugionium*	Megacarpaeeae	Brassicaceae	NC_030516.1
Pugionium dolabratum	*Pugionium*	Megacarpaeeae	Brassicaceae	NC_030515.1
Megacarpaea delavayi	*Megacarpaea*	Megacarpaeeae	Brassicaceae	NC_034360.1
Pachycladon cheesemanii	*Pachycladon*	Microlepidieae	Brassicaceae	NC_021102.1
Pachycladon enysii	*Pachycladon*	Microlepidieae	Brassicaceae	NC_018565.1
Thlaspi arvense	*Thlaspi*	Thlaspideae	Brassicaceae	NC_034362.1
Schrenkiella parvula	*Schrenkiella*	Unassigned	Brassicaceae	NC_028726.1

(*) out-group

Supplementary Table S2. List of SSRs identified in Crambe abyssinica plastome

Repeats	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**total**
A/T	-	-	-	-	-	94	43	20	10	9	4	3	1	2		2			1	189
C/G	-	-	-	-	-	6	2													8
AC/GT	-	2																		2
AG/CT	-	11																		11
AT/AT	-	33	10	4	1		1													49
AAG/CTT	-	1																		1
AAT/ATT	-	4																		4
ACC/GGT	-	1																		1
AAAC/GTTT	1																			1
AAAT/ATTT	2																			2
AAGT/ACTT	1																			1
AGAT/ATCT	1																			1
AACAT/ATGTT	1																			1
AATAG/ATTCT	1																			1
AATAT/ATATT	1																			1
Total																				**273**

Supplementary Table S3. List of SSRs identified in Brassica nigra plastome

Repeats	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	total
A/T	-	-	-	-	-	87	41	20	10	6	11	2		2		1		2	182
C/G	-	-	-	-	-	2	1	1											4
AC/GT	-	1																	1
AG/CT	-	9		1															10
AT/AT	-	34	9	5	1		1												50
AAG/CTT	-	1																	1
AAT/ATT	-	2	1																3
AAAC/GTTT	1																		1
AAAT/ATTT	3																		3
AGAT/ATCT	1																		1
AAATC/ATTTG	1																		1
AATAT/ATATT	1																		1
AAAAGT/ACTTTT	1																		1
AAGTAG/ACTTCT	1																		1
AATACC/ATTGGT	1																		1
Total																			**261**

Supplementary Table S4. List of SSRs identified in Brassica napus plastome

Repeats	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	total
A/T	-	-	-	-	-	86	58	22	16	5	6	2	3	1		1	1	201
C/G	-	-	-	-	-	3	2	2										7
AC/GT	-	2																2
AG/CT	-	10																10
AT/AT	-	37	10	3	1	2												53
AAT/ATT	-	3																3
AAAC/GTTT	1																	1
AAAG/CTTT	1																	1
AAAT/ATTT	2																	2
AGAT/ATCT	2																	2
AAGAT/ATCTT	1																	1
Total																		**283**

Supplementary Table S5. List of SSRs identified in Brassica juncea plastome

Repeats	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	total
A/T	-	-	-	-	-	90	61	21	16	6	7	4	1	1					1	208
C/G	-	-	-	-	-	4	4		1											9
AC/GT	-	2																		2
AG/CT	-	10																		10
AT/AT	-	38	10	2	2	1														53
AAT/ATT	-	3																		3
AAAC/GTTT	1																			1
AAAG/CTTT	1																			1
AAAT/ATTT	2																			2
AGAT/ATCT	2																			2
Total																				**291**

Supplementary Table S6. List of SSRs identified in Cakile arabica plastome

Repeats	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	total
A/T	-	-	-	-	-	83	50	29	19	6	2	6	2	1	2	1			1	202
C/G	-	-	-	-	-	4	2	2												8
AC/GT	-	2																		2
AG/CT	-	9																		9
AT/AT	-	33	9	2	5	2	1													52
AAT/ATT	-	4																		4
AAAC/GTTT	1																			1
AAAT/ATTT	4																			4
AGAT/ATCT	2																			2
Total																				**284**

Supplementary Table S7. List of SSRs identified in Raphanus sativus plastome

Repeats	3	4	5	6	7	8	9	10	11	12	13	14	15	16	total
A/T	-	-	-	-	-	86	52	26	8	5	8	5		2	192
C/G	-	-	-	-	-	5	1	1							7
AC/GT	-	2													2
AG/CT	-	10													10
AT/AT	-	30	9	5	4	1									49
AAT/ATT	-	2													2
AAAG/CTTT	1														1
AAAT/ATTT	4														4
AATT/AATT	2														2
AGAT/ATCT	1														1
AACAC/GTGTT	1														1
Total															**271**

Supplementary Table S8. List of SSRs identified in Orychophragmus diffusus plastome

Repeats	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	total
A/T	-	-	-	-	-	86	57	25	20	6	6	3	1	1		1	1	1	208
C/G	-	-	-	-	-	2	3		1										6
AC/GT	-	2																	2
AG/CT	-	10																	10
AT/AT	-	32	6	5	1	1	3		1										49
AAG/CTT	-	1																	1
AAT/ATT	-	2																	2
AAAT/ATTT	3																		3
AATT/AATT	1	1																	2
AGAT/ATCT	1																		1
AAATC/ATTTG	1																		1
AATACT/AGTATT	1																		1
Total																			**286**

Supplementary Table S9. List of SSRs identified in Orychophragmus taibaiensis plastome

Repeats	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	total
A/T	-	-	-	-	-	100	48	25	11	5	5	3	2	2				1	202
C/G	-	-	-	-	-	3													3
AC/GT	-	2																	2
AG/CT	-	10																	10
AT/AT	-	28	6	11	1	1	1												48
AAG/CTT	-	1																	1
AAT/ATT	-	2																	2
AAAG/CTTT	1																		1
AAAT/ATTT	5																		5
AACC/GGTT	1																		1
AATT/AATT	3																		3
AGAT/ATCT	1																		1
AATAT/ATATT	1																		1
Total																			**280**

Supplementary Table S10. List of SSRs identified in Orychophragmus hupehensis plastome

Repeats	3	4	5	6	7	8	9	10	11	12	13	14	15	total
A/T	-	-	-	-	-	93	53	22	10	3	5	5	4	195
C/G	-	-	-	-	-	2	1							3
AC/GT	-	2												2
AG/CT	-	10												10
AT/AT	-	31	3	12	1	1	1							49
AAG/CTT	-	1												1
AAT/ATT	-	2												2
AAAG/CTTT	1													1
AAAT/ATTT	4	1												5
AATT/AATT	2													2
AGAT/ATCT	1													1
AATAT/ATATT	1													1
AAATAT/ATATTT	1													1
Total														**273**

Supplementary Table S11. List of SSRs identified in Sinalliaria grandifolia plastome

Repeats	3	4	5	6	7	8	9	10	11	12	13	14	15	16	total
A/T	-	-	-	-	-	90	50	24	10	4	8	7	3	5	201
C/G	-	-	-	-	-	5	1			1					7
AC/GT	-	2													2
AG/CT	-	10													10
AT/AT	-	31	9	6	2	2									50
AAG/CTT	-	1													1
AAT/ATT	-	2													2
AAAT/ATTT	4														4
AGAT/ATCT	2														2
AAAAT/ATTTT	1														1
AATGG/ATTCC	1														1
AAAAAG/CTTTTT	1														1
Total															**282**

Supplementary Table S12. List of SSRs identified in Sinalliaria limprichtiana plastome

Repeats	3	4	5	6	7	8	9	10	11	12	13	14	15	16	total
A/T	-	-	-	-	-	104	49	32	8	6	7	3	1	2	212
C/G	-	-	-	-	-	3	1	2		1					7
AC/GT	-	2													2
AG/CT	-	9													9
AT/AT	-	34	8	10			1								53
AAG/CTT	-	1													1
AAT/ATT	-	2													2
AAAT/ATTT	4														4
AATT/AATT	1														1
AGAT/ATCT	1														1
AAAAT/ATTTT	1														1
AATAT/ATATT	1														1
Total															**294**

Supplementary Table S13. List of SSRs identified in Sinapis arvensis plastome

Repeats	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	total
						7	5	1													
A/T	-	-	-	-	-	7	3	5	8	7	4	5	1	1		1		1		1	174
C/G	-	-	-	-	-	3	2														5
AC/GT	-	1																			1
AG/CT	-	9		1																	10
		2	1																		
AT/AT	-	9	2	7	2		1														51
AAG/CTT	-	1																			1
AAT/ATT	-	2	1																		3
AAAC/GTTT	1																				1
AAAT/ATTT	4																				4
AGAT/ATCT	1																				1
AAGTAG/ACTTC T	1																				1
Total																					**252**

Supplementary Table S14. Location of SSRs in Crambe abyssinica plastome

SSR type	SSR	size	start	end	Location
mono	(A)12	12	113	124	*trnH*-GUG/*psbA* (IGS)
mono	(A)8	8	176	183	*trnH*-GUG/*psbA* (IGS)
mono	(T)9	9	1491	1499	*psbA*/*trnK*-UUU (IGS)
mono	(T)10	10	2190	2199	*matK* (CDS)
mono	(A)9	9	2784	2792	*matK* (CDS)
mono	(T)9	9	2859	2867	*matK* (CDS)
mono	(A)8	8	3378	3385	*matK* (CDS)
di	(AT)5	10	3703	3712	*trnK*-UUU (intron)
mono	(T)8	8	3716	3723	*trnK*-UUU (intron)
mono	(T)15	15	3923	3937	*trnK*-UUU (intron)
mono	(T)14	14	4008	4021	*trnK*-UUU (intron)
mono	(C)8	8	4245	4252	*trnK*-UUU/rps16 (IGS)
mono	(A)9	9	4253	4261	*trnK*-UUU/rps16 (IGS)
di	(TA)4	8	4517	4524	*trnK*-UUU/rps16 (IGS)
di	(AT)4	8	4549	4556	*trnK*-UUU/rps16 (IGS)
mono	(C)9	9	5189	5197	*rps16* (intron)
mono	(A)9	9	5279	5287	*rps16* (intron)
mono	(T)8	8	5345	5352	*rps16* (intron)

mono	(T)8	8	5462	5469	*rps16* (intron)
mono	(T)10	10	5783	5792	*rps16* (intron)
di	(CA)4	8	6342	6349	*rps16*/*trnQ*-UUG (IGS)
mono	(T)8	8	6590	6597	*trnQ*-UUG/*psbK* (IGS)
mono	(A)8	8	7325	7332	*psbK*/*psbI* (IGS)
mono	(T)10	10	7372	7381	*psbK*/*psbI* (IGS)
mono	(A)10	10	7731	7740	*trnS*-GCU/*trnG*-UCC (IGS)
di	(AT)5	10	7840	7849	*trnS*-GCU/*trnG*-UCC (IGS)
mono	(T)9	9	7849	7857	*trnS*-GCU/*trnG*-UCC (IGS)
di	(TA)5	10	7874	7883	*trnS*-GCU/*trnG*-UCC (IGS)
di	(TA)4	8	7953	7960	*trnS*-GCU/*trnG*-UCC (IGS)
di	(AT)7	14	7965	7978	*trnS*-GCU/*trnG*-UCC (IGS)
di	(TA)6	12	7985	7996	*trnS*-GCU/*trnG*-UCC (IGS)
mono	(T)10	10	8174	8183	*trnS*-GCU/*trnG*-UCC (IGS)
mono	(T)9	9	8336	8344	*trnS*-GCU/*trnG*-UCC (IGS)
di	(TA)4	8	9599	9606	*trnR*-UCU/*atpA* (IGS)
mono	(A)11	11	11183	11193	*atpA*/*atpF* (IGS)
mono	(A)10	10	11661	11670	*atpF* (intron)
di	(AT)4	8	11835	11842	*atpF* (intron)
di	(CA)4	8	12354	12361	*atpF* (intron)
mono	(T)9	9	12547	12555	*atpF*/*atpH* (IGS)
penta	(CTATT)3	15	12632	12646	*atpF*/*atpH* (IGS)
tri	(AAT)4	12	12679	12690	*atpF*/*atpH* (IGS)
mono	(A)8	8	12742	12749	*atpF*/*atpH* (IGS)
mono	(A)8	8	12780	12787	*atpF*/*atpH* (IGS)
mono	(A)9	9	12922	12930	*atpF*/*atpH* (IGS)
mono	(A)9	9	12936	12944	*atpF*/*atpH* (IGS)
di	(AT)5	10	13418	13427	*atpH*/*atpI* (IGS)
mono	(A)9	9	13706	13714	atpH/*atpI* (IGS)
mono	(T)9	9	14522	14530	*atpI*/*rps2* (IGS)
mono	(T)9	9	15739	15747	*rpoC2* (CDS)
mono	(T)9	9	16878	16886	*rpoC2* (CDS)
mono	(T)11	11	17632	17642	*rpoC2* (CDS)
mono	(A)8	8	17775	17782	*rpoC2* (CDS)
mono	(T)9	9	18167	18175	*rpoC2* (CDS)
di	(CT)4	8	18463	18470	*rpoC2* (CDS)
di	(TA)5	10	19004	19013	*rpoC2* (CDS)
di	(AT)4	8	20045	20052	*rpoC1* (CDS)
mono	(A)8	8	21463	21470	*rpoC1* (CDS)
mono	(C)8	8	21595	21602	*rpoC1* (intron)
mono	(A)8	8	22115	22122	*rpoC1* (intron)
mono	(T)11	11	22164	22174	*rpoC1* (intron)
mono	(T)8	8	25277	25284	*rpoB* (CDS)
mono	(T)10	10	25381	25390	*rpoB* (CDS)
mono	(A)9	9	26305	26313	*rpoB*/*trnC*-GCA (IGS)
mono	(T)10	10	26428	26437	*rpoB*/*trnC*-GCA (IGS)
di	(AT)5	10	26621	26630	*rpoB*/*trnC*-GCA (IGS)
mono	(A)8	8	27089	27096	*rpoB*/*trnC*-GCA (IGS)
di	(CT)4	8	27857	27864	*petN* (CDS)
mono	(T)8	8	27970	27977	*petN*/*psbM* (IGS)
mono	(T)9	9	27982	27990	*petN*/*psbM* (IGS)
tetra	(CAAA)3	12	28133	28144	*petN*/*psbM* (IGS)
mono	(T)8	8	28390	28397	*petN*/*psbM* (IGS)
mono	(T)9	9	28549	28557	*psbM/trnD*-GUC (IGS)
di	(TA)4	8	29222	29229	*psbM/trnD*-GUC (IGS)
tri	(AAT)4	12	29355	29366	*psbM/trnD*-GUC (IGS)
mono	(A)12	12	29430	29441	*psbM/trnD*-GUC (IGS)
mono	(A)10	10	29486	29495	*psbM/trnD*-GUC (IGS)
mono	(T)10	10	29511	29520	*psbM/trnD*-GUC (IGS)

mono	(A)12	12	29832	29843	*trnD-GUC/trnY*-GUA (IGS)
mono	(A)9	9	30499	30507	*trnE-UUC/trnT*-GGU (IGS)
mono	(A)9	9	30559	30567	*trnE-UUC/trnT*-GGU (IGS)
mono	(T)18	18	30590	30607	*trnE-UUC/trnT*-GGU (IGS)
di	(TA)4	8	30791	30798	*trnE-UUC/trnT*-GGU (IGS)
di	(AT)6	12	30803	30814	*trnE-UUC/trnT*-GGU (IGS)
di	(AT)4	8	31141	31148	*trnT*-GGU/*psbD* (IGS)
mono	(T)8	8	31285	31292	*trnT*-GGU/*psbD* (IGS)
di	(GA)4	8	34451	34458	*trnS*-UGS (CDS)
mono	(T)8	8	34563	34570	*trnS*-UGA/*psbZ* (IGS)
mono	(G)8	8	35198	35205	*psbZ*/*trnG*-GCC (IGS)
mono	(A)9	9	35299	35307	*psbZ*/*trnG*-GCC (IGS)
di	(AT)9	18	35434	35451	*psbZ*/*trnG*-GCC (IGS)
mono	(A)9	9	35795	35803	*trnG*-GCC/*trnfM*-CAU (IGS)
mono	(T)9	9	36428	36436	*rps14/psaB* (IGS)
mono	(A)10	18	36439	36448	*rps14/psaB* (IGS)
di	(TA)4	8	36449	36456	*rps14/psaB* (IGS)
c	(C)8	18	41189	41196	*psaA/ycf3* (IGS)
	(T)10		41197	41206	*psaA/ycf3* (IGS)
mono	(A)11	11	41668	41678	*psaA/ycf3* (IGS)
mono	(A)8	8	42127	42134	*ycf3* (intron)
mono	(T)12	12	42562	42573	*ycf3* (intron)
tri	(AAG)4	12	43245	43256	*ycf3* (intron)
mono	(A)9	9	43554	43562	*ycf3* (intron)
mono	(T)8	8	44301	44308	*trnS*-GGA/*rps4* (IGS)
mono	(T)8	8	45074	45081	*rps4*/*trnT*-UGU (IGS)
di	(AT)4	8	45094	45101	*rps4*/*trnT*-UGU (IGS)
di	(AT)4	8	45173	45180	*rps4*/*trnT*-UGU (IGS)
tetra	(TAAA)3	12	45533	45544	*trnT*-UGU/*trnL*-UAA (IGS)
mono	(T)9	9	45644	45652	*trnT*-UGU/*trnL*-UAA (IGS)
mono	(T)8	8	45684	45691	*trnT*-UGU/*trnL*-UAA (IGS)
tri	(TAT)4	12	45707	45718	*trnT*-UGU/*trnL*-UAA (IGS)
mono	(A)9	9	45915	45923	*trnT*-UGU/*trnL*-UAA (IGS)
di	(AT)4	8	45982	45989	*trnT*-UGU/*trnL*-UAA (IGS)
mono	(A)8	8	46076	46083	*trnT*-UGU/*trnL*-UAA (IGS)
mono	(A)8	8	47209	47216	*trnF*-GAA/*ndhJ* (IGS)
mono	(T)8	8	47425	47432	*trnF*-GAA/*ndhJ* (IGS)
mono	(T)8	8	48887	48894	*ndhK*/*ndhC* (IGS)
mono	(T)8	8	48897	48904	*ndhK*/*ndhC* (IGS)
tetra	(ACTT)3	12	49512	49523	*ndhC/trnV*-UAC (IGS)
mono	(T)8	8	49598	49605	*ndhC/trnV*-UAC (IGS)
mono	(A)8	8	49890	49897	*ndhC/trnV*-UAC (IGS)
mono	(T)11	11	50301	50311	*trnV*-UAC (intron)
mono	(A)10	10	50370	50379	*trnV*-UAC (intron)
mono	(T)11	11	51192	51202	*trnM*-CAU/*atpE* (IGS)
mono	(T)8	8	51234	51241	*trnM*-CAU/*atpE* (IGS)
mono	(A)8	8	53385	53392	*atpB/rbcL* (IGS)
mono	(T)11	11	53502	53512	*atpB/rbcL* (IGS)
mono	(C)9	9	53560	53568	*atpB/rbcL* (IGS)
mono	(A)9	9	55522	55530	*rbcL/accD* (IGS)
mono	(T)9	9	55717	55725	*rbcL/accD* (IGS)
mono	(T)8	8	55798	55805	*rbcL/accD* (IGS)
mono	(T)9	9	56384	56392	*accD* (CDS)
di	(AT)4	8	58062	58069	*accD/psaI* (IGS)
di	(AT)4	8	58100	58107	*accD/psaI* (IGS)
mono	(A)10	10	58350	58359	*psaI/ycf4* (IGS)
di	(AT)4	8	58763	58770	*psaI/ycf4* (IGS)
mono	(T)8	8	58941	58948	*ycf4* (CDS)
mono	(T)8	8	59438	59445	*ycf4/cemA* (IGS)

mono	(T)8	8	59847	59854	*cemA* (CDS)
di	(TC)4	8	59888	59895	*cemA* (CDS)
di	(AT)4	8	60805	60812	*petA* (CDS)
mono	(C)8	8	61091	61098	*petA* (CDS)
mono	(A)8	8	61165	61172	*petA* (CDS)
mono	(A)8	8	61742	61749	*petA/psbJ* (IGS)
di	(AT)4	8	62219	62226	*petA/psbJ* (IGS)
di	(TA)5	10	62236	62245	*petA/psbJ* (IGS)
di	(AT)5	10	62254	62263	*petA/psbJ* (IGS)
di	(AT)6	12	62286	62297	*petA/psbJ* (IGS)
di	(TA)4	8	62310	62317	*petA/psbJ* (IGS)
tetra	(AAAT)3	12	64384	64395	*psbE/petL* (IGS)
mono	(A)12	12	64560	64571	*psbE/petL* (IGS)
mono	(T)8	8	64655	64662	*psbE/petL* (IGS)
mono	(T)8	8	64668	64675	*psbE/petL* (IGS)
mono	(T)8	8	65363	65370	*petG/trnW*-CCA (IGS)
mono	(A)9	9	65567	65575	*trnW*-CCA/*trnP*-UGG (IGS)
mono	(T)8	8	65872	65879	*trnP*-UGG/*psaJ* (IGS)
di	(AT)4	8	65897	65904	*trnP*-UGG/*psaJ* (IGS)
mono	(T)8	8	65920	65927	*trnP*-UGG/*psaJ* (IGS)
di	(TA)4	8	66101	66108	*trnP*-UGG/*psaJ* (IGS)
di	(TA)4	8	66113	66120	*trnP*-UGG/*psaJ* (IGS)
mono	(T)8	8	66277	66284	*psaJ* (CDS)
mono	(T)8	8	66371	66378	*psaJ/rpl33* (IGS)
mono	(G)8	8	66403	66410	*psaJ/rpl33* (IGS)
mono	(T)8	8	66561	66568	*psaJ/rpl33* (IGS)
mono	(T)8	8	66611	66618	*psaJ/rpl33* (IGS)
di	(AT)4	8	66985	66992	*rpl33/rps18* (IGS)
mono	(A)8	8	67072	67079	*rpl33/rps18* (IGS)
mono	(A)10	10	67559	67568	*rps18/rpl20* (IGS)
mono	(A)8	8	68187	68194	*rpl20/rps12* (IGS)
mono	(T)16	16	69043	69058	*rps12/clpP* (IGS)
mono	(T)21	21	69687	69707	*clpP* (intron)
mono	(A)8	8	70210	70217	*clpP* (CDS)
mono	(T)9	9	70519	70527	*clpP* (intron)
mono	(T)8	8	70691	70698	*clpP* (intron)
mono	(T)10	10	70710	70719	*clpP* (intron)
mono	(A)8	8	71246	71253	*clpP/psbB* (IGS)
mono	(T)9	9	73299	73307	*psbB/psbT* (IGS)
mono	(T)8	8	73521	73528	*psbT* (CDS)
penta	(AACAT)3	15	74778	74792	*petB* (intron)
di	(GA)4	8	74886	74893	*petB* (intron)
mono	(T)8	8	74941	74948	*petB* (intron)
di	(AT)4	8	76484	76491	*petD* (intron)
mono	(T)8	8	77272	77279	*petD/rpoA* (IGS)
mono	(T)8	8	77281	77288	*rpoA* (CDS)
mono	(T)13	13	77467	77479	*rpoA* (CDS)
mono	(T)9	9	78804	78812	*rps11/rpl36* (IGS)
tri	(GGT)4	12	79375	79386	*rpl36/rps8* (IGS)
mono	(T)9	9	79401	79409	*rpl36/rps8* (IGS)
di	(TA)4	8	79925	79932	*rps8/rpl14* (IGS)
mono	(T)8	8	80046	80053	*rps8/rpl14* (IGS)
mono	(A)12	12	80513	80524	*rpl14/rpl16* (IGS)
mono	(T)8	8	81119	81126	*rpl16* (intron)
mono	(T)11	11	81294	81304	*rpl16* (intron)
mono	(A)10	10	81330	81339	*rpl16* (intron)
mono	(T)12	12	81596	81607	*rpl16* (intron)
mono	(T)12	12	81861	81872	*rpl16* (intron)
mono	(T)8	8	82007	82014	*rpl16* (intron)

mono	(T)8	8	82016	82023	*rpl16* (intron)
mono	(T)8	8	82125	82132	*rpl16/rps3* (IGS)
di	(AT)4	8	83309	83316	*rpl22* (CDS)
mono	(A)10	10	83387	83396	*rpl22/rps19* (IGS)
mono	(T)8	8	83414	83421	*rps19* (CDS)
mono	(T)9	9	83689	83697	*rsp19* (CDS)
mono	(T)8	8	83728	83735	*rps19/rpl2* (IGS)
di	(TA)4	8	84313	84320	*rpl2* (intron)
di	(GA)4	8	85953	85960	*ycf2* (CDS)
di	(GA)4	8	86940	86947	*ycf2* (CDS)
mono	(A)9	9	89143	89151	*ycf2* (CDS)
di	(GA)4	8	89164	89171	*ycf2* (CDS)
di	(CT)4	8	91197	91204	*ycf2* (CDS)
di	(TA)4	8	92511	92518	*ycf2* (CDS)
di	(TA)5	10	93915	93924	*trnL*-CAA/*ndhB* (IGS)
di	(AG)4	8	94675	94682	*ndhB* (CDS)
mono	(A)8	8	96570	96577	*ndhB/rps7* (IGS)
mono	(A)14	14	98741	98754	*rps12/trnV*-GAC (IGS)
mono	(T)9	9	98774	98782	*rps12/trnV*-GAC (IGS)
mono	(T)13	13	99546	99558	*rps12/trnV*-GAC (IGS)
di	(CT)4	8	105757	105764	*rrn23S*
mono	(T)8	8	107935	107942	*trnR*-ACG/*trnN*-GUU (IGS)
mono	(A)10	10	109775	109784	*ycf1* (CDS)
mono	(T)8	8	110591	110598	*ndhF* (CDS)
mono	(A)8	8	111548	111555	*ndhF* (CDS)
tetra	(ATAG)3	12	111820	111831	*ndhF* (CDS)
di	(AT)4	8	112166	112173	*ndhF/rpl32* (IGS)
mono	(A)8	8	112222	112229	*ndhF/rpl32* (IGS)
mono	(T)18	18	112239	112256	*ndhF/rpl32* (IGS)
mono	(A)8	8	112396	112403	*ndhF/rpl32* (IGS)
mono	(T)8	8	112502	112509	*ndhF/rpl32* (IGS)
mono	(T)8	8	112536	112543	*ndhF/rpl32* (IGS)
di	(TA)4	8	113069	113076	*rpl32*/*trnL*-UAG (IGS)
mono	(T)8	8	113082	113089	*rpl32/trnL*-UAG (IGS)
tri	(ATA)4	12	113351	113362	*rpl32/trnL*-UAG (IGS)
mono	(T)11	11	113390	113400	*rpl32/trnL*-UAG (IGS)
mono	(A)9	9	114149	114157	*ccsA* (CDS)
mono	(A)8	8	114886	114893	*ccsA/ndhD* (IGS)
mono	(T)9	9	115044	115052	*ccsA/ndhD* (IGS)
mono	(A)8	8	115362	115369	*ndhD* (CDS)
mono	(T)8	8	115658	115665	*ndhD* (CDS)
mono	(A)8	8	116668	116675	*ndhD/psaC* (IGS)
di	(AT)4	8	117642	117649	*ndhE/ndhG* (IGS)
mono	(T)8	8	118448	118455	*ndhG/ndhI* (IGS)
penta	(ATATA)3	15	118574	118588	*ndhG/ndhI* (IGS)
mono	(A)8	8	120253	120260	*ndhA* (intron)
mono	(T)8	8	120565	120572	*ndhA* (intron)
di	(AT)5	10	120784	120793	*ndhA* (intron)
mono	(T)9	9	122758	122766	*rps15* (CDS)
mono	(T)8	8	122925	122932	*rps15* (CDS)
di	(TA)6	12	123261	123272	*rps15/ycf1* (IGS)
mono	(T)8	8	123447	123454	*ycf1* (CDS)
mono	(T)12	12	123546	123557	*ycf1* (CDS)
mono	(T)8	8	123639	123646	*ycf1* (CDS)
mono	(T)11	11	123658	123668	*ycf1* (CDS)
di	(TA)4	8	123724	123731	*ycf1* (CDS)
mono	(A)8	8	124353	124360	*ycf1* (CDS)
mono	(T)9	9	124499	124507	*ycf1* (CDS)
di	(TA)4	8	124711	124718	*ycf1* (CDS)

mono	(T)9	9	125172	125180	*ycf1* (CDS)
mono	(T)16	16	125220	125235	*ycf1* (CDS)
mono	(T)8	8	125275	125282	*ycf1* (CDS)
mono	(A)8	8	125291	125298	*ycf1* (CDS)
mono	(A)8	8	125375	125382	*ycf1* (CDS)
mono	(T)8	8	125431	125438	*ycf1* (CDS)
mono	(T)8	8	125597	125604	*ycf1* (CDS)
mono	(T)13	13	125718	125730	*ycf1* (CDS)
mono	(T)8	8	125857	125864	*ycf1* (CDS)
mono	(T)8	8	125898	125905	*ycf1* (CDS)
mono	(T)14	14	125974	125987	*ycf1* (CDS)
mono	(T)9	9	126386	126394	*ycf1* (CDS)
mono	(T)10	10	126416	126425	*ycf1* (CDS)
mono	(A)13	13	126445	126457	*ycf1* (CDS)
mono	(A)8	8	126473	126480	*ycf1* (CDS)
mono	(A)8	8	126937	126944	*ycf1* (CDS)
mono	(T)8	8	127068	127075	*ycf1* (CDS)
mono	(T)8	8	127336	127343	*ycf1* (CDS)
mono	(A)8	8	127447	127454	*ycf1* (CDS)

Supplementary Table S15. Distribution of tandem repeats in Crambe abyssinica plastome

Copy number	Consensus size	Start	End	Location
8	20	7866	8009	*trnS*-GCU/*trnG*-UCC (IGS)
2	42	35364	35448	*psbZ*/*trnG*-GCC (IGS)
3	21	58438	58504	*psaI*/*ycf4* (IGS)
2	25	59504	59555	*ycf4*/*cemA* (IGS)
2	43	62306	62386	*petA*/*psbJ* (IGS)
2	21	70848	70890	*clpP* (intron)
4	22	88500	88582	*ycf2* (CDS)
2	32	107131	107191	*rrn4.5S*/*rrn5S* (IGS)

Supplementary Table S16. Distribution of direct (D) and inverted (I) sequence repeats loci in Crambe abyssinica plastome

Type	Size (bp)	Start		Location	
		repeat 1	repeat 2	repeat 1	repeat 2
D	41	37812	40036	*psaB* (CDS)	*psaA* (CDS)
D	39	42901	98211	*ycf3* (intron)	*rps12*/*trnV*-GAC (IGS)
D	37	98214	119791	*rps12*/*trnV*-GAC (IGS)	*ndhA* (intron)
D	31	7606	34446	*trnS*-GCU	*trnS*-UGA
D	30	58434	58476	*psaI*/*ycf4* (IGS)	*psaI*/*ycf4* (IGS)
I	30	7607	44009	*trnS*-GCU	*trnS*-GGA
I	30	34514	43947	*trnS*-UGA	*trnS*-GGA
I	30	34447	44009	*trnS*-UGA	*trnS*-GGA

The complete plastome of macaw palm [Acrocomia aculeata (Jacq.) Lodd. ex Mart.] and extensive molecular analyses of the evolution of plastid genes in Arecaceae

Amanda de Santana Lopes[1], Túlio Gomes Pacheco[1], Tabea Nimz[1], Leila do Nascimento Vieira[2], Miguel Pedro Guerra[2], Rubens Onofre Nodari[2], Emanuel Maltempi de Souza[3], Fábio de Oliveira Pedrosa[3], Marcelo Rogalski[1*]

1 Laboratório de Fisiologia Molecular de Plantas, Departamento de Biologia Vegetal, Universidade Federal de Viçosa, Viçosa-MG, Brazil.
2 Laboratório de Fisiologia do Desenvolvimento e Genética Vegetal, Programa de Pós-graduação em Recursos Genéticos Vegetais, Universidade Federal de Santa Catarina, Florianópolis-SC, Brazil.
3 Departamento de Bioquímica e Biologia Molecular, Núcleo de Fixação Biológica de Nitrogênio, Universidade Federal do Paraná, Curitiba-PR, Brazil.

*****Corresponding author
E-mail address: rogalski@ufv.br

Published in:
Planta (2018) in press
DOI: 10.1007/s00425-018-2841-x

Planta
https://doi.org/10.1007/s00425-018-2841-x

ORIGINAL ARTICLE

The complete plastome of macaw palm [*Acrocomia aculeata* (Jacq.) Lodd. ex Mart.] and extensive molecular analyses of the evolution of plastid genes in Arecaceae

Amanda de Santana Lopes[1] · Túlio Gomes Pacheco[1] · Tabea Nimz[1] · Leila do Nascimento Vieira[2] · Miguel P. Guerra[2] · Rubens O. Nodari[2] · Emanuel Maltempi de Souza[3] · Fábio de Oliveira Pedrosa[3] · Marcelo Rogalski[1]

Received: 1 November 2017 / Accepted: 10 January 2018

Abstract

Main conclusion **The plastome of macaw palm was sequenced allowing analyses of evolution and molecular markers. Additionally, we demonstrated that more than half of plastid protein-coding genes in Arecaceae underwent positive selection.**

Macaw palm is a native species from tropical and subtropical Americas. It shows high production of oil per hectare reaching up to 70% of oil content in fruits and an interesting plasticity to grow in different ecosystems. Its domestication and breeding are still in the beginning, which makes the development of molecular markers essential to assess natural populations and germplasm collections. Therefore, we sequenced and characterized in detail the plastome of macaw palm. A total of 221 SSR loci were identified in the plastome of macaw palm. Additionally, eight polymorphism hotspots were characterized at level of subfamily and tribe. Moreover, several events of gain and loss of RNA editing sites were found within the subfamily Arecoideae. Aiming to uncover evolutionary events in Arecaceae, we also analyzed extensively the evolution of plastid genes. The analyses show that highly divergent genes seem to evolve in a species-specific manner, suggesting that gene degeneration events may be occurring within Arecaceae at the level of genus or species. Unexpectedly, we found that more than half of plastid protein-coding genes are under positive selection, including genes for photosynthesis, gene expression machinery and other essential plastid functions. Furthermore, we performed a phylogenomic analysis using whole plastomes of 40 taxa, representing all subfamilies of Arecaceae, which placed the macaw palm within the tribe Cocoseae. Finally, the data showed here are important for genetic studies in macaw palm and provide new insights into the evolution of plastid genes and environmental adaptation in Arecaceae.

Keywords Palm tree · Plastid genome · Plastid SSRs · Polymorphism hotspots · Gene divergence · Positive selection

Introduction

Macaw palm [*Acrocomia aculeata* (Jacq.) Lodd. Ex Mart.] is a native species distributed in the tropical and subtropical Americas, with center of origin in Brazil (Henderson et al. 1995; Lanes et al. 2015). Its fruits are oil-rich, accumulating up to 70% of oil (dry-weight) and yielding approximately 6200 kg of oil per hectare (Pires et al. 2013). If compared with current oil crops the macaw palm can reach oil production similar to the oil crops with the highest productivity such as oil palm (*Elaeis guineensis* Jacq.) (Motoike and Kuki 2009). In addition to the high oil productivity its oil profile is suitable for biofuel production (Pires et al. 2013; Lanes et al. 2014). The integral use of the fruits still provides

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00425-018-2841-x) contains supplementary material, which is available to authorized users.

✉ Marcelo Rogalski
rogalski@ufv.br

[1] Laboratório de Fisiologia Molecular de Plantas, Departamento de Biologia Vegetal, Universidade Federal de Viçosa, Viçosa, MG, Brazil

[2] Laboratório de Fisiologia do Desenvolvimento e Genética Vegetal, Programa de Pós-Graduação em Recursos Genéticos Vegetais, Universidade Federal de Santa Catarina, Florianópolis, SC, Brazil

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

Published online: 16 January 2018

biomass feedstock for several industrial applications focusing on sustainable energy such as ethanol and charcoal (Vilas-Boas et al. 2010; Gonçalves et al. 2013; Pires et al. 2013). The macaw palm has interesting agronomic and ecological features because it can occupy degraded areas or agroforestry systems given that it has high plasticity to grow in different ecosystems (Henderson et al. 1995; Motoike and Kuki 2009; Pires et al. 2013) avoiding conflict with areas of food production.

Due to its attractive features, the macaw palm domestication, genetic breeding and development of commercial plantations have encouraged financial investments in basic research, biotechnology and industry. Studies in several areas have been carried out with this purpose, including ecophysiology (Pires et al. 2013), mating system (Lanes et al. 2016), vegetative development (Berton et al. 2013; Machado et al. 2016), fruit development (Montoya et al. 2016), phenotypic diversity (Ciconini et al. 2013; Lanes et al. 2015; Conceição et al. 2015; Coser et al. 2016) and biotechnological applications (Moura et al. 2009; Luis and Scherwinski-Pereira 2014; Padilha et al. 2015).

Considering the interest to make macaw palm an alternative platform for the production of renewable energy, the characterization of molecular markers has essential importance to assess the genetic diversity of natural populations aiming the establishment of suitable strategies for conservation, germplasm characterization, domestication and genetic breeding. However, a small number of molecular markers were developed for macaw palm (Mengistu et al. 2016a, b; Nucci et al. 2008) and all of them are based on nuclear microsatellites (SSRs). The nonrecombinant and uniparentally inherited nature of plastid genome (plastome) makes it a great source of molecular markers, particularly intergenic spacers (IGSs) and introns, where the mutation rates are higher in comparison with coding sequences (Rogalski et al. 2015; Vieira et al. 2016a). Plastid SSRs have been used in several genetic studies including phyllogeography, population genetic and gene flow analyses (Provan et al. 2001; Ebert and Peakall 2009; Wheeler et al. 2014). Plastid sequences with high polymorphism are also applied to assess the genetic structure and diversity/divergence of natural populations and germplasm collections (Tsai et al. 2015; Wambulwa et al. 2016; Roy et al. 2016).

Plastome sequences are also of great importance to understand evolutionary events in plants based on the knowledge of gene content, recombination events, loss of genes, gene transfer to the nucleus and genome rearrangements (Vieira et al. 2016b; Bock 2017; Lopes et al. 2017; Park et al. 2017; Ruhlman et al. 2017), as well as for plastid transformation aiming basic research (Rogalski et al. 2006, 2008; Alkatib et al. 2012) and biotechnological applications (Daniell et al. 2016; Zhang et al. 2017) since the complete sequence of plastome is a prerequisite to choose target intergenic regions for insertion of transgenes and development of transplastomic plants (Daniell et al. 2016; Fuentes et al. 2017). The plastid transformation has been used efficiently for metabolic engineering (Daniell et al. 2016; Fuentes et al. 2017) and it is a viable alternative to manipulate plastid fatty acid biosynthesis (Rogalski and Carrer 2011) given that several efficient protocols of plant regeneration based on somatic embryogenesis are available for macaw palm (Moura et al. 2009; Luis and Scherwinski-Pereira 2014; Padilha et al. 2015).

Macaw palm belongs to the family Arecaceae, which contains about 2450 species distributed in five subfamilies, Calamoideae, Nypoideae, Arecoideae, Coryphoideae, and Ceroxyloideae (Dransfield et al. 2005; Asmussen et al. 2006; Barfod et al. 2011). The Arecoideae is the largest subfamily of Arecaceae, including *A. aculeata*, *Cocos nucifera* L., and *E. guineensis*, which are species of great economic importance (Baker et al. 2009; Comer et al. 2015). The genus *Acrocomia* is Neotropical occurring from north Mexico to south Argentina. The number of species is not taxonomically well resolved. The first classification included only two species to the genus, *A. aculeata* and *A. hassleri* (Henderson et al. 1995). The last classification recognizes eight species including *A. aculeata*, *A. crispa*, *A. emensis*, *A. glaucescens*, *A. hassleri*, *A. intumescens*, *A. media,* and *A. totai* (Lorenzi et al. 2010; The Plant List 2013). Recently, a new species was described, *Acrocomia corumbaensis*, showing an arborescent habit similar to *A. aculeata*, *A. crispa*, *A. intumescens,* and *A. totai* (Vianna 2017). The new data have demonstrated an increasing number of new species, which makes the plastid genomics a useful tool to classify correctly them. According to Barrett et al. (2016) the plastomes of Arecaceae have a low rate of variation compared with other commelinids (e.g. grasses). Nevertheless, the diversification rate over time within palm genera seems to increase and a convergent evolution has been reported among palm species adapted to shaded areas (Ma et al. 2015; Faurby et al. 2016). Despite most plastid genes are conserved, several evolutionary events have been described such as new RNA editing sites, loss of introns, high divergence of genes, and positive signatures (Sen et al. 2012; Williams et al. 2015; He et al. 2016; Chen et al. 2017). In grasses, one-third of the plastid genes underwent positive selection, which shows a relationship between gene evolution and environmental conditions regarding the photosynthetic apparatus (Piot et al. 2017). The recent diversification of palm species may be related to some adaptive changes in plastid genes. However, the magnitude of the changes in plastid genes that underwent positive selection and how it could be related to adaptation to different environmental conditions remain unknown in Arecaceae.

Here, we reported the complete plastome of macaw palm, which was molecularly characterized in details. The description and location of all SSR loci and polymorphism hotspots

within macaw palm plastome were shown. Among the 221 SSR loci identified, 157 are located in fast-evolving regions (IGSs and introns). In addition, we performed a phylogenomic analysis using whole plastomes of 40 taxa, including 37 species representing all five subfamilies of Arecaceae, which placed the macaw palm within tribe Cocoseae. Moreover, some IGSs at the level of subfamily and tribe were identified as polymorphism hotspots, especially the *trnC*-GCA/*petN* and *psaC*/*ndhE*. Furthermore, 100 putative RNA editing sites were found within Arecoideae, including some possible events of gain and loss of editing sites. Finally, we investigated extensively the molecular evolution of all protein-coding genes from 37 palm plastomes. We identified some highly divergent genes in a species-specific manner suggesting that gene degeneration processes may be occurring within Arecaceae at the level of genus or species. We investigated if the plastid genes of Areacaceae underwent positive selection and the analysis indicated that more than half of the plastid protein-coding genes have one or more positive signatures, which can significantly affect essential plastid functions. Taken together, our data bring new molecular markers useful for genetic studies in macaw palm natural populations and raise questions about the evolution of plastome within Arecaceae and the relationship between positive selection and evolution of plastid genes under different environmental conditions.

Materials and methods

Chloroplast isolation, DNA extraction, sequencing, assembling, annotation, and data archiving statement

Fresh leaves from a young macaw palm plant were collected and kept for 1 week at 4 °C to decrease starch level. The young plant was obtained from seeds of macaw palm plants belonging to the ex situ plant collection, Macaúba Active Germplasm Bank (BAG-Macaúba repository: 084/2013/CGEN/MMA) located in the experimental farm of the Universidade Federal de Viçosa (208400100S, 4283101500W), State of Minas Gerais, Brazil (Coser et al. 2016; Lanes et al. 2015; Lanes et al. 2016; Mengistu et al. 2016a, b; Montoya et al. 2016).

The chloroplast isolation and cp DNA extraction were carried out according to Vieira et al. (2014). Approximately 1 ng of plastid DNA was used to prepare sequencing libraries with Nextera XT DNA Sample Prep Kit (Illumina Inc., San Diego, CA, USA) according to the manufacturer's instructions. The obtained library was sequenced using Illumina MiSeq platform (Illumina Inc., San Diego, CA, USA) at the Federal University of Paraná, State of Paraná, Brazil. The paired-end reads (total of 907,088 reads), with average length of 280.9 bp, were trimmed under the threshold with probability of error < 0.05. The trimmed reads (901,031 reads, average length of 218.4 bp) were de novo assembled in contigs using CLC Genomics Workbench 8.0.2 software (CLC Bio, Aarhus, Denmark). The contigs used for assembling of macaw palm plastome ranged from 1323.96 to 315.12 of average coverage.

The program Dual Organellar GenoMe Annotator (DOGMA) (Wyman et al. 2004) and BLAST were used for preliminarily gene annotation. From this initial annotation, putative start codons, stop codons, and intron positions were determined based on comparisons to homologous genes of other plastomes at the GenBank database. All tRNA genes were further verified by using tRNAscan-SE server (Lowe and Eddy 1997). The physical circular map of the plastome was drawn using Organellar Genome DRAW (OGDRAW) (Lohse et al. 2013). The complete nucleotide sequence of macaw palm plastome sequenced in this study was deposited in the GenBank database under accession number MG020488.

SSR identification, sliding window analysis and RNA editing sites prediction

Simple sequence repeats (SSRs) loci were detected in the macaw palm plastome and in other six Arecoideae plastomes available in the GenBank database (Supplementary Table S1) using the MIcroSAtellite (MISA) Perl script (Thiel et al. 2003). The thresholds were set to eight repeat units for mononucleotide SSRs, four repeat units for di- and trinucleotide SSRs, and three repeat units for tetra-, penta-, and hexanucleotide SSRs.

To identify the hotspots of sequence divergence in the subfamily Arecoideae and tribe Cocoseae, we performed the sliding window analysis. First, complete plastomes were aligned using MAFFT v.7 (Katoh and Standley 2013), and posteriorly, the sliding window analysis was conducted by using the DnaSP v.5 software (Librado and Rozas 2009). The window length and the step size were set as 200 and 50 bp, respectively.

Potential RNA editing sites in plastid protein-coding genes of species belonging to subfamily Arecoideae were predicted by the program predictive RNA editor for plants (PREP) suite (Mower 2009). The program PREP uses 35 reference genes for detecting of possible RNA editing sites in plastomes. The cutoff value was set to 0.8. Additional RNA editing sites were predicted by comparison with Huang et al. (2013) that based on RT-PCR and sequencing data identified several RNA editing sites in transcripts of plastid genes of *C. nucifera*.

Comparative analysis of plastome structure

To characterize the general structure of the subfamily Arecoideae, nucleotide MUMmer (NUCmer) Perl script in MUMmer 3.0 (Kurtz et al. 2004) was used to visualize and compare the plastome structures between *A. aculeata* and other Arecoideae representatives (Supplementary Table S1).

Phylogenomic reconstruction of the family Arecaceae

The phylogenetic reconstruction of the family Arecaceae was carried out using whole plastomes. The GenBank accession number of each taxon used here is shown in the Supplementary Table S1. The species *Hanguana malayana* (Hanguanaceae: Commelinales), *Baxteria australis*, and *Dasypogon bromeliifolius* (Dasypogonaceae: Arecales) were used as outgroups. First, whole plastomes were extracted from GenBank and the IRB was withdrawn to prevent overrepresentation of the IR sequences. The alignment of plastomes was done using MAFFT v.7 (Katoh and Standley 2013) and the best substitution model (GTR + I+ G) was selected by using jModelTest v.2.1.7 (Darriba et al. 2012). Last, Bayesian inference analysis was performed using MrBayes version 3.2 (Ronquist et al. 2012), with one million generations of two runs of four Markov Chains, with three hot and one cold in each run. To check the parameter convergence we used the software Tracer 1.6 (http://tree.bio.ed.ac.uk/software/tracer/). The software FigTree 1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/) was used to visualize the consensus tree.

Molecular evolution analysis of protein-coding genes in Arecaceae plastomes

The 79 protein-coding genes present in Arecaceae plastomes (Supplementary Table S1) were extracted and the codon alignment was done using the software Muscle (Edgar 2004) implemented in Mega 7.0 (Tamura et al. 2013). Phylogenetic reconstruction was performed to assess the gene divergence. First, substitution models for each gene were selected using jModelTest v.2.1.7 (Darriba et al. 2012), which are listed in the Supplementary Table S2. Then, a Bayesian inference analysis was performed using MrBayes version 3.2 (Ronquist et al. 2012), with two million generations of two runs of four Markov Chains, with three hot and one cold in each run and the software FigTree 1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/) was used to visualize the consensus tree. The gene divergence was estimated by the sum of total branch lengths that link the operational taxonomical units to the common ancestor of Arecaceae species sampled here. The convergence parameters of each tree were checked using the software Tracer 1.6 (http://tree.bio.ed.ac.uk/software/tracer/).

Finally, to investigate the presence of positive signatures (positive selection), the genes were aligned as described before. The positive signatures were analyzed using Selecton version 2.4 (http://selecton.tau.ac.il/index.html; Stern et al. 2007), performing the comparison M8 (allows positive selection) against M8a (null model) with one degree of freedom and cutoff (ε) of 0.1. We consider in our analysis only sites where reliable positive selection was inferred (lower bound > 1 and test with probability < 0.01).

Results

General features of macaw palm plastome and comparative analyses within the subfamily Arecoideae

The macaw palm plastome is a circular molecule of 155,829 bp in length with a typical quadripartite structure found in most plastomes of angiosperms (Wicke et al. 2011; Zhu et al. 2016), which include a pair of inverted repeats (IRs) between two regions of single copy, the large single copy (LSC) and the small single copy (SSC) (Fig. 1). The macaw palm plastome contains 113 unique genes, being 79 protein-coding genes, 30 tRNA genes, and four rRNA genes (Table 1). Some of these genes are duplicated in the IRs (Fig. 1); they are eight tRNA genes, all rRNA genes, and nine protein-coding genes (one of them, the *ycf1*, is partially duplicated). Among the 113 unique genes, 16 possess one intron (six tRNA genes and ten protein-coding genes) and two contain two introns (*clpP* and *ycf3* genes).

The size of LSC, SSC, and IRs of macaw palm plastome is compared with other plastomes from species belonging to the subfamily Arecoideae in the Table 2. They have similar dimensions of the quadripartite structure. The only exception, *Areca vestiaria*, contains a very small IR region, which is compensated by a larger LSC region. These general structural features were also explored by dot-plot analyses (Supplementary Fig. S1), which show high similarity and absence of rearrangements between macaw palm and other Arecoideae plastomes. The only exception is *A. vestiaria* that lost most part of the IR region, which corresponds to IRB in the plastome of macaw palm. A more detailed view of the IRA and IRB borders of Arecoideae plastomes (Fig. 2), except for *A. vestiaria* plastome, reveals slight differences within this subfamily. Most plastomes, including the macaw palm, present the *rps19* gene completely duplicated in the IR region, while in *C. nucifera* and *Syagrus coronata* it is partially duplicated in the IR borders and encodes a functional protein only in the LSC-IRA junction. A variability of nucleotide number in the *rps19-rpl22* and *rps19-psbA* intergenic spacers (IGS) is observed in the LSC-IRA and LSC-IRB junctions, respectively. Some variability is also

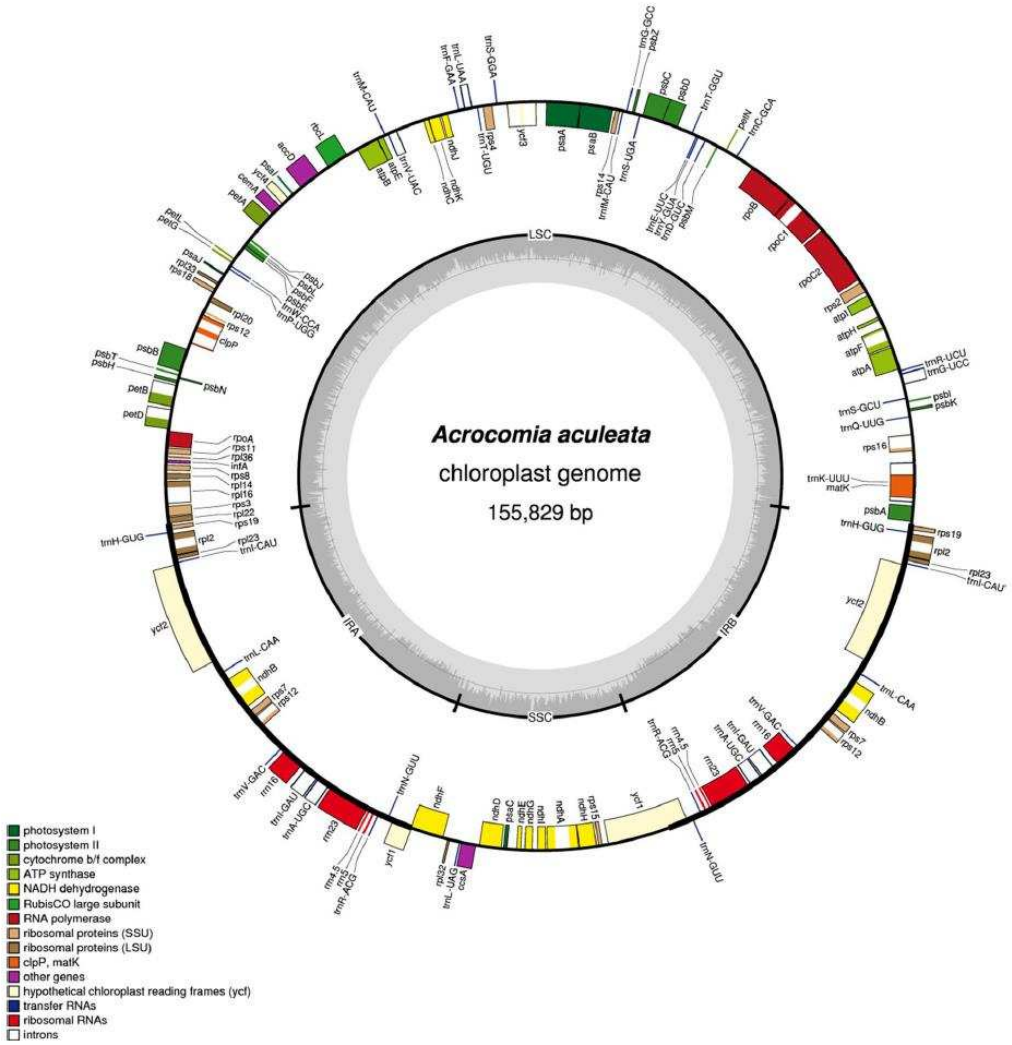


Fig. 1 Gene map of macaw palm (*Acrocomia aculeata*) plastome. Genes drawn inside the circle are transcribed in the clockwise direction, and genes drawn outside are transcribed in the counterclockwise direction. Different functional groups of genes are color-coded. The darker gray in the inner circle corresponds to GC content, and the lighter gray corresponds to AT content. *LSC* large single copy, *SSC* small single copy, *IRA/B* inverted repeat A/B

present in the IR-SSC junctions, mainly in the *ycf1* gene. Part of the *ndhF* gene is located within the IRs (overlapping part of the *ycf1* gene), which represents a stretch of 56 bp highly conserved among all species analyzed here (Fig. 2).

SSR content and nucleotide divergence analysis in macaw palm plastome and other Arecoideae species

The SSR loci number and distribution among Arecoideae plastomes (Fig. 3a), including *A. aculeata*, are very similar, with the majority composed of mononucleotide repeats (Fig. 3b). The mononucleotide repeats in Arecoideae plastomes are basically constituted of A/T sequences (ranging from 94.8 to 96.8% among the species sampled here). The density of SSR loci (SSR number/kilobase) was higher in the SSC (mean of 2.28 ± 0.11), followed by LSC (mean of 1.93 ± 0.07), and lower in the IR (mean of 0.75 ± 0.04). The total mean density, considering only one IR, was 1.75 (± 0.05). In the species *A. vestiaria* we found the same pattern of SSR distribution in the SSC, LSC, and IR regions as delimited in the other Arecoideae species analyzed here.

Table 1 List of genes identified in the plastome of *Acrocomia aculeata*

Group of gene	Name of gene
Gene expression machinery	
Ribosomal RNA genes	*rrn16*[b]; *rrn23*[b]; *rrn5*[b]; *rrn4.5*[b]
Transfer RNA genes	*trnA* –UGC[ab]; *trnC* –GCA; *trnD* –GUC; *trnE* –UUC; *trnF* –GAA; *trnfM* –CAU; *trnG* –UCC[a]; *trnG* –GCC; *trnH* –GUG[b]; *trnI* –CAU[b]; *trnI* –GAU[ab]; *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
Small subunit of ribosome	*rps2; rps3; rps4; rps7*[b]; *rps8; rps11; rps12*[ab]; *rps14; rps15*; *rps16*[a]; *rps18; rps19*[b]
Large subunit of ribosome	*rpl2*[ab]; *rpl14; rpl16*[a]; *rpl20; rpl22; rpl23*[b]; *rpl32; rpl33; rpl36*
DNA-dependent RNA polymerase	*rpoA; rpoB; rpoC1*[a]; *rpoC2*
Translational initiation factor	*infA*
Intron maturase	*matK*
Genes for photosynthesis	
Subunits of photosystem I (PSI)	*psaA; psaB; psaC; psaI; psaJ; ycf3*[a]; *ycf4*
Subunits of photosystem II (PSII)	*psbA*; *psbB; psbC; psbD; psbE; psbF; psbH; psbI; psbJ; psbK; psbL; psbM; psbN; psbT; psbZ*
Subunits of cytochrome b_6f	*petA; petB*[a]; *petD*[a]; *petG; petL; petN*
Subunits of ATP synthase	*atpA; atpB; atpE; atpF*[a]; *atpH; atpI*
Subunits of NADH dehydrogenase	*ndhA*[a]; *ndhB*[ab]; *ndhC; ndhD; ndhE; ndhF; ndhG; ndhH*; *ndhI; ndhJ; ndhK*
Large subunit of Rubisco	*rbcL*
Other functions	
Envelope membrane protein	*cemA*
Subunit of acetyl-CoA carboxylase	*accD*
C-type cytochrome synthesis	*ccsA*
Subunit of protease Clp	*clpP*[a]
Component of TIC complex	*ycf1*[c]
Unknown function	*ycf2*[b]

[a]Genes containing introns

[b]Duplicated genes

[c]Partially duplicated genes

Table 2 General features of plastid genomes within the subfamily Arecoideae

	Acrocomia aculeata	*Areca vestiaria*	*Cocos nucifera*	*Elaeis guineensis*	*Pododoccus barteri*[a]	*Syagrus coronata*	*Veitchia arecina*[a]
Total cpDNA size	155,829	130,807	154,731	156,973	157,683	155,053	157,194
Length of LSC region	84,265	110,702	84,230	85,192	85,522	84,535	85,647
Length of IR region	27,092	1426	26,555	27,071	27,220	26,522	27,050
Length of SSC	17,380	17,253	17,391	17,639	17,721	17,474	17,447
GC content (%)	37.53	36.58	37.44	37.40	37.66	37.46	37.47

[a]The values are an approximation because the nucleotide sequences of these plastomes are not complete in the GenBank database

A complete description and location of all SSRs identified in the plastome of macaw palm is shown in Supplementary Table S3. Among the 221 SSR loci identified, 123 are located in intergenic spacers (IGSs), 61 in coding sequences (CDSs), 34 in introns, and three in tRNA genes. The CDSs with higher number of SSRs are *ycf1* (18 SSRs), *ycf2* (7 SSRs), and *rpoC2* (7 SSRs) genes. In addition, the *ndhC*/*trnV*-UAC (10 SSRs) and *ndhF*/*rpl32* (6 SSRs) IGSs, and the introns of *clpP* (8 SSRs) and *trnK*-UUU (7 SSRs) genes contain a high number of SSRs. A comparison of SSRs found here with SSRs identified in the plastome of *E. guineensis* (the closest taxon sampled to *A. aculeata*,

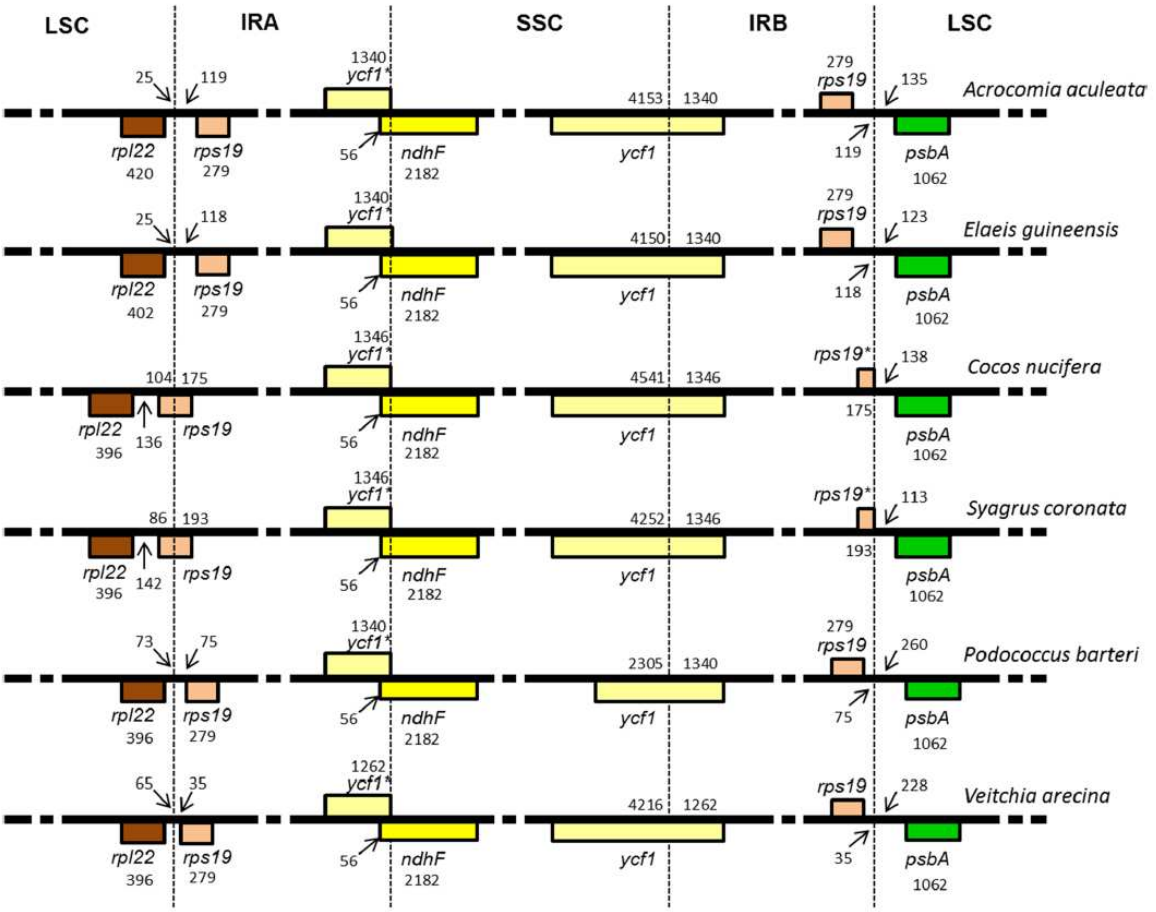


Fig. 2 Comparison of the IRA and IRB borders among Arecoideae species. The numbers indicate the lengths of IGSs, genes, and spacers between IR-LSC and IR-SSC junctions. The *ycf1** and *rps19** genes have incomplete CDSs

according to our phylogenomic tree showed below) showed that 49% of the SSRs located in the IGSs and introns are polymorphic. Only four polymorphic SSRs (all located in the *ycf1* gene) are located in CDS (Supplementary Table S3).

Moreover, we performed a nucleotide divergence analysis to identify the regions with higher polymorphism within the subfamily Arecoideae and tribe Cocoseae, in order to select putative fast-evolving plastid sequences in macaw palm plastome (Fig. 4). The *trnC*-GCA/*petN* and *psaC*/*ndhE* IGSs, and the CDS of *ycf1* gene (part included in the SSC region) are shared between the subfamily Arecoideae and the tribe Cocoseae as the regions with the highest nucleotide divergence. Within the subfamily Arecoideae, other three regions of high polymorphism are located in the junctions IRB-LSC [*rps19*/*psbA* (IGS)] and LSC/IRA [*rpl22*/*rps19* (IGS)], and in the IGS between the *accD* and *psaI* genes. Furthermore, aiming to detect more specific polymorphism hotspots, we also performed the sliding window analysis aligning the plastomes of macaw palm and *E. guineensis* (Fig. 4). As the previous analyses revealed for the subfamily Arecoideae and tribe Cocoseae, the plastome regions *trnC*-GCA/*petN* (IGS), *psaC*/*ndhE* (IGS), and the CDS of *ycf1* gene, show again the highest nucleotide divergence. In addition to them, other two regions (the *psbC*/*trnS*-UGA and *trnL*-UAG/*ccsA* IGSs) were found to be exclusive between *A. aculeata* and *E. guineensis*.

Identification of putative RNA editing sites in plastid protein-coding genes of macaw palm and other species within Arecoideae

The prediction of RNA editing sites in the subfamily Arecoideae was carried out based on PREP program and comparison with RNA editing sites validated by Huang et al.

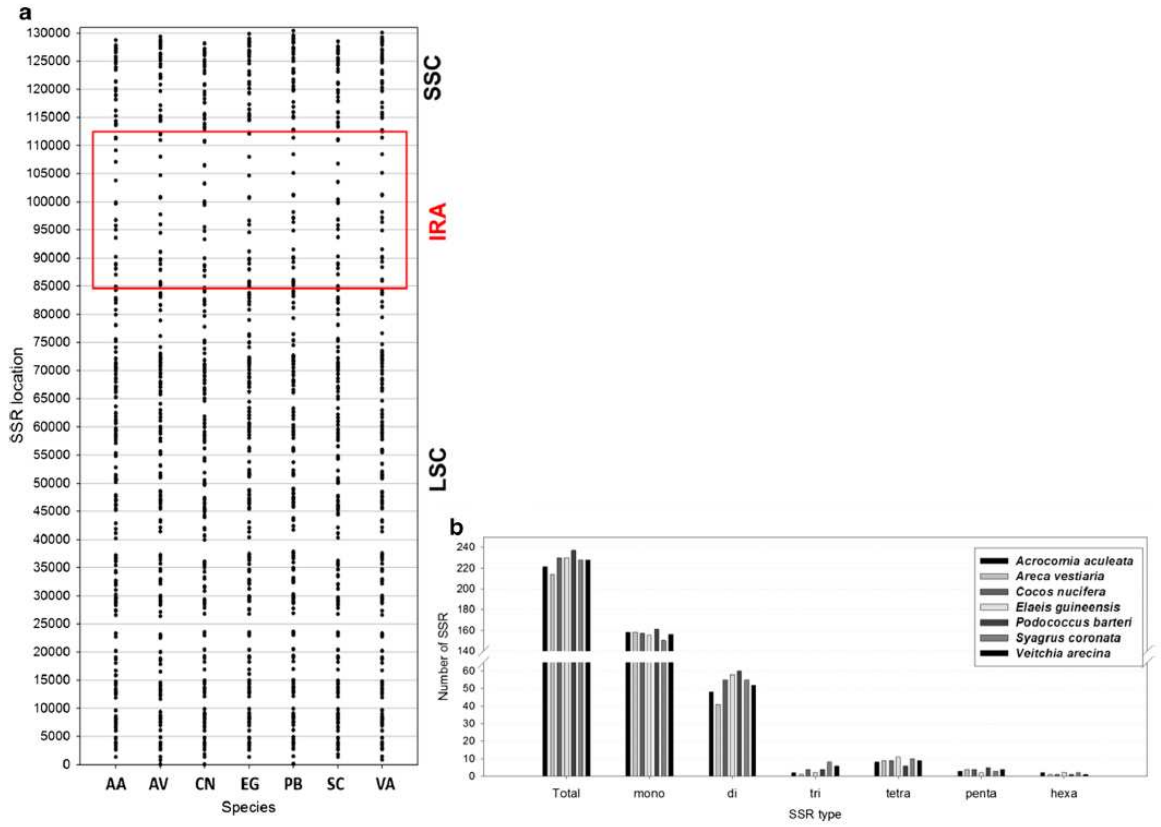


Fig. 3 SSR loci identification in Arecoideae plastomes (IRB omitted). **a** Distribuition of SSR loci. **b** Number of SSR loci. It was set eight repeat units for mononucleotide SSRs, four repeat units for di- and trinucleotide SSRs, and three repeat units for tetra-, penta- and hexanucleotide SSRs. Species sampled: *Acrocomia aculeata* (AA), *Areca vestiaria* (AV), *Cocos nucifera* (CN), *Elaeis guineensis* (EG), *Podococcus barteri* (PB), *Syagrus coronate* (SC), and *Veitchia arecina* (VA)

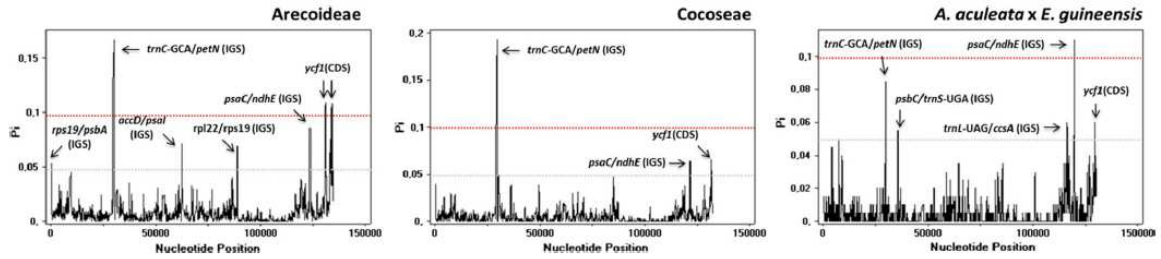


Fig. 4 Sliding window analyses of aligned whole plastomes (IRB omitted) within the subfamily Arecoideae, the tribe Cocoseae, and between *Acrocomia aculeata* and *Elaeis guineesis*. The regions with high nucleotide variability ($Pi > 0.05$) are indicated. Pi, nucleotide diversity of each window. Window length 200 pb. Step size 50 pb

(2013) for some plastid protein-coding genes in *C. nucifera* (Supplementary Tables S4 and S5). All RNA editing sites identified are C-to-U conversions, at the first (24%) or second (76%) positions of the codons. A total of 100 RNA editing sites (distributed among 30 protein-coding genes) were predicted here, being 87 of them shared among all species sampled in this study. According to Huang et al. (2013), 74 out of 100 sites identified here are completely or partially edited in *C. nucifera*. Most RNA editing sites predicted here change the encoded amino acid from polar to apolar (62 out

of 100). Only one putative RNA editing site changes the amino acid from apolar to polar (proline-serine), while the remaining do not change de polarity (20, polar-polar; 17, apolar-apolar).

Despite the high conservation of RNA editing sites among Arecoideae species, we also identified 13 sites that are not edited in one or more species. In addition, the RNA editing sites shared between this lineage show codon variants in 6 sites in which the RNA editing recovers the same conserved amino acid [e.g. in the site (97) of *ndhF* gene the RNA editing changes different codons TCA (S-L) and CCA (P-L) generating the same conserved amino acid]. All those codons (edited or unedited) that diverge in one or more species are shown in the Table 3. In order to identify if there are correlation between the Arecoideae classification and the evolution of RNA editing sites among the species analyzed here, we performed a Bayesian analysis using the divergent codons concatenated from different genes. The reconstructed tree (Fig. 5a) distinguishes the three tribes represented in the analysis, showing that the tribe Cocoseae is sister to the clade composed by the sister tribes Podococceae and Areceae.

Analyzing the 13 putative RNA editing sites, that are not edited in one or more species of Arecoideae, in the light of the Arecoideae phylogenetic tree (based on our phylogenomic analysis showed below), it is possible to suggest some events of gain and loss of RNA editing sites within the subfamily Arecoideae (Fig. 5b). The *accD* (429) and *ndhF* (131) sites were supposedly gained and lost, respectively, in the tribe Areceae given that these features are not shared with the tribes Podococceae (sister group) and Cocoseae. However, species-specific gains and losses are more frequent. In our analyses we count five gains [the *accD* (241), *accD* (487), *clpP* (118), matK (312), and *ndhF* (607) sites] and four losses [the *matK* (219), *ndhF* (148), *ndhH* (182), and *rpoB* (665) sites]. Lastly, the distribution pattern of editing at *accD* (389) and *ccsA* (274) sites among the species sampled here do not allow hypotheses about gains or losses in the subfamily Arecoideae.

Arecaceae phylogenomic reconstruction based on whole plastomes

The Arecaceae phylogenomic was carried out using whole plastomes of 40 taxa (Supplementary Table S1), including 37 species representing all five subfamilies of Arecaceae: Calamoideae, Nypoideae, Arecoideae, Coryphoideae, and Ceroxyloideae (Dransfield et al. 2005; Asmussen et al. 2006). The three remaining taxa are two species of family Dasypogonaceae (most related to Arecaceae) and *H. malayana* (Hanguanaceae: Commelinales), which were used as outgroup. Bayesian inference (BI) analysis produced a

Table 3 List of RNA editing sites whose codons diverge in one or more species of Arecoideae subfamily

Gene	AA position	*Acrocomia aculeata*	*Elaeis guineensis*	*Syagrus coronata*	*Cocos nucifera*	*Podococcus barteri*	*Veitchia arecina*	*Areca vestiaria*
accD	241	UUU (F)	UUU (F)	CUU (L-F)	UUU (F)	UUU (F)	UGU (C)	UUU (F)
accD	389	CAU (H-Y)	CAU (H-Y)	CAU (H-Y)	CAU (H-Y)	UAU (Y)	UAU (Y)	UAU (Y)
accD	429	UAC (Y)	UAC (Y)	UAC (Y)	UAC (Y)	UAC (Y)	CAC (H-Y)	CAU (H-Y)
accD	487	UUG (L)	UCG (S-L)	UUG (L)	UUG (L)	UUG (L)	UUG (L)	UUG (L)
ccsA	274	UUA (L)	UUA (L)	UUA (L)	UUA (L)	UCA (S-L)	UCA (S-L)	UCA (S-L)
clpP	118	UUC (F)	CUC (L-F)	UUC (F)	UUC (F)	UUC (F)	UUC (F)	UUC (F)
matK	219	ACA (U)	CCA (P-L)	CCA (P-L)	CCA (P-L)	CCA (P-L)	CCA (P-L)	CCA (P-L)
matK	310	CAU (H-Y)	CAU (H-Y)	CAU (H-Y)	CAU (H-Y)	CAC (H-Y)	CAC (H-Y)	CAC (H-Y)
matK	312	GCC (A-V)	GUC (V)	GUC (V)	GUC (V)	GUC (V)	GUC (V)	GUC (V)
matK	426	CAC (H-Y)	CAC (H-Y)	CAC (H-Y)	CAC (H-Y)	CAC (H-Y)	CAC (H-Y)	CAU (H-Y)
ndhD	225	UCG (S-L)	UCG (S-L)	UCG (S-L)	UCG (S-L)	UCG (S-L)	UCA (S-L)	UCG (S-L)
ndhD	398	UCA (S-L)	UCA (S-L)	UCA (S-L)	UCA (S-L)	UCA (S-L)	UCG (S-L)	UCG (S-L)
ndhF	97	UCA (S-L)	UCA (S-L)	UCA (S-L)	UCA (S-L)	CCA (P-L)	UCA (S-L)	UCA (S-L)
ndhF	131	UCC (S-F)	UCC (S-F)	UCC (S-F)	UCC (S-F)	UCC (S-F)	UUC (F)	UUC (F)
ndhF	148	CAU (H-Y)	CAU (H-Y)	CAU (H-Y)	CAU (H-Y)	CAU (H-Y)	UAU (Y)	CAU (H-Y)
ndhF	607	UUG (L)	UUU (F)	UUC (F)	UUC (F)	CUC (L-F)	UUC (F)	UUC (F)
ndhF	698	UCU (S-F)	UCC (S-F)	UCC (S-F)	UCC (S-F)	UCC (S-F)	UCC (S-F)	UCC (S-F)
ndhH	182	UCU (S-F)	UCU (S-F)	UCU (S-F)	UCU (S-F)	UCU (S-F)	UCU (S-F)	UUU (F)
rpoB	665	UCU (S-F)	UCU (S-F)	UUU (F)	UCU (S-F)	UCU (S-F)	UCU (S-F)	UCU (S-F)

The cytidines marked are putatively edited to uridines. The letters in parentheses represent the encoded amino acids. The encoded amino acid after RNA editing is also showed

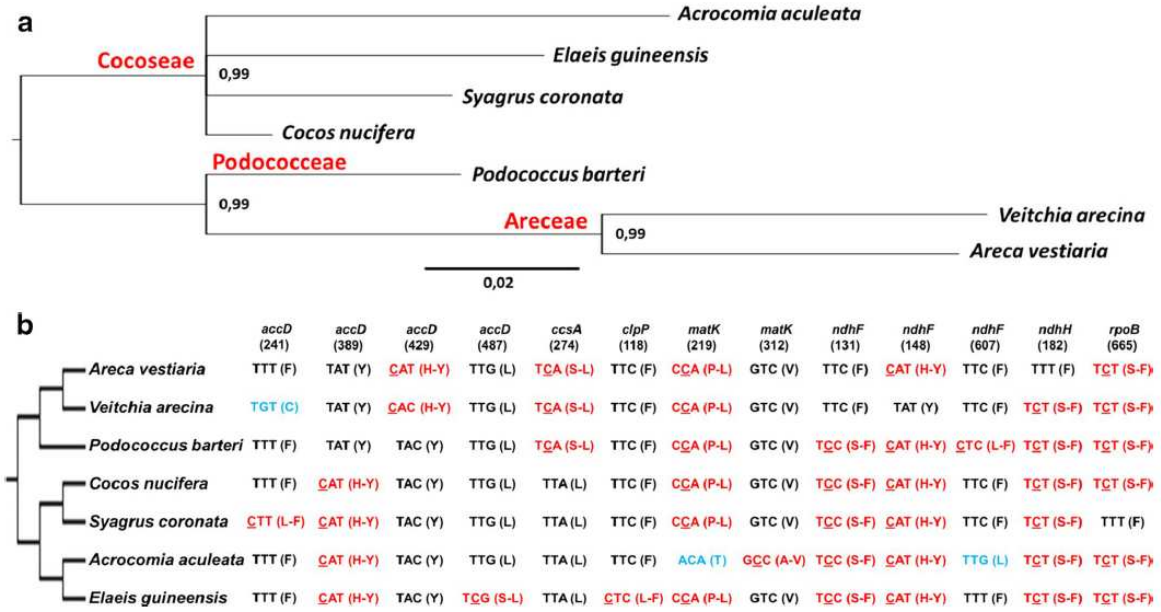


Fig. 5 RNA editing analysis. **a** Relationships within the subfamily Arecoideae based on concatenated codons using bayesian inference. The codons used were extracted from putative RNA editing sites that diverge in one or more species of Arecoideae. The tribes are highlighted in red. The bayesian posterior probability of all nodes is 0.99. The branch length is proportional to the inferred divergence level and the scale bar indicates the number of inferred nucleic acids substitutions per site. **b** Distribution of putative RNA editing sites across the Arecoideae phylogeny based on whole plastome. The codons highlighted in red are edited, and the codons in blue codify non-conserved amino acids

phylogenetic tree with a − lnL = 510,541.335 (Fig. 6) and a high branch support (BI posterior probability value of 1 for all nodes). Regarding the relationships among the subfamilies within Arecaceae, the subfamily Calaimoideae is sister to a clade composed by other four subfamilies. This clade contains the subfamily Nypoideae sister to a subclade where the subfamily Coryphoideae forms a sister-group with the subfamilies Arecoideae and Ceroxyloideae. Among the Arecoideae species sampled here, the tribe Cocoseae forms a sister-group with the tribes Podococceae and Areceae. The macaw palm (*A. aculeata*) is placed within the tribe Cocoseae and it is sister to *E. guineensis*.

Gene divergence analysis and identification of positive signatures in plastid protein-coding genes within Arecaceae

Overall, the gene divergence analysis indicates that the plastid protein-coding genes in Arecaceae are well conserved (Fig. 7a). The *ycf1* gene is the most divergent gene, followed by *rpl32* and *rps16* genes. These genes show extensive variation of branch length among the species given that some of them exhibit high divergent branches such as *Podococcus barteri* (*ycf1* and *rpl32*) and *Salacca ramosiana* (*rpl16*) (Supplementary Fig. S2). Other genes, such as *ccsA*, *rpl22*, *psaJ*, and *ndhK*, also occur in one or two species with a highlighted long branch in comparison with the remaining species with small branches (Supplementary Fig. S2).

To investigate whether any plastid gene of Arecaceae underwent positive selection we used the Selecton program. We identified a total of 283 putative positive signatures distributed in more than half of plastid genes (40 out of 79 protein-coding genes) (Fig. 7b). The positive selection shows a tendency to be related to gene divergence rate given that most divergent genes have more positive signatures. Positive signatures were identified in several genes involved in different essential functions such as photosynthesis [including subunits of PSI (*ycf3*, *ycf4* and *psa* genes), PSII (*psb* genes), ATP synthase (*atp* genes), cytochrome b_6f (*petD* gene), and ndh complex (*ndh* genes)], plastid gene expression [including subunits of RNA polymerase (*rpo* genes), RNA maturation (*matK* gene), and ribosomal proteins (*rpl* and *rps* genes)], fatty acid biosynthesis (*accD* gene), cytochrome biosynthesis (*ccsA* gene), import of protein (*ycf1* gene), uptake of inorganic carbon (*cemA* gene), and unknown function (*ycf2* gene).

In order to identify the evolutionary patterns across the palm family, we plotted the sites under positive selection against the phylogeny inferred from whole plastomes (Fig. 8; Supplementary Figs. S3–S9). Three evolutionary types are

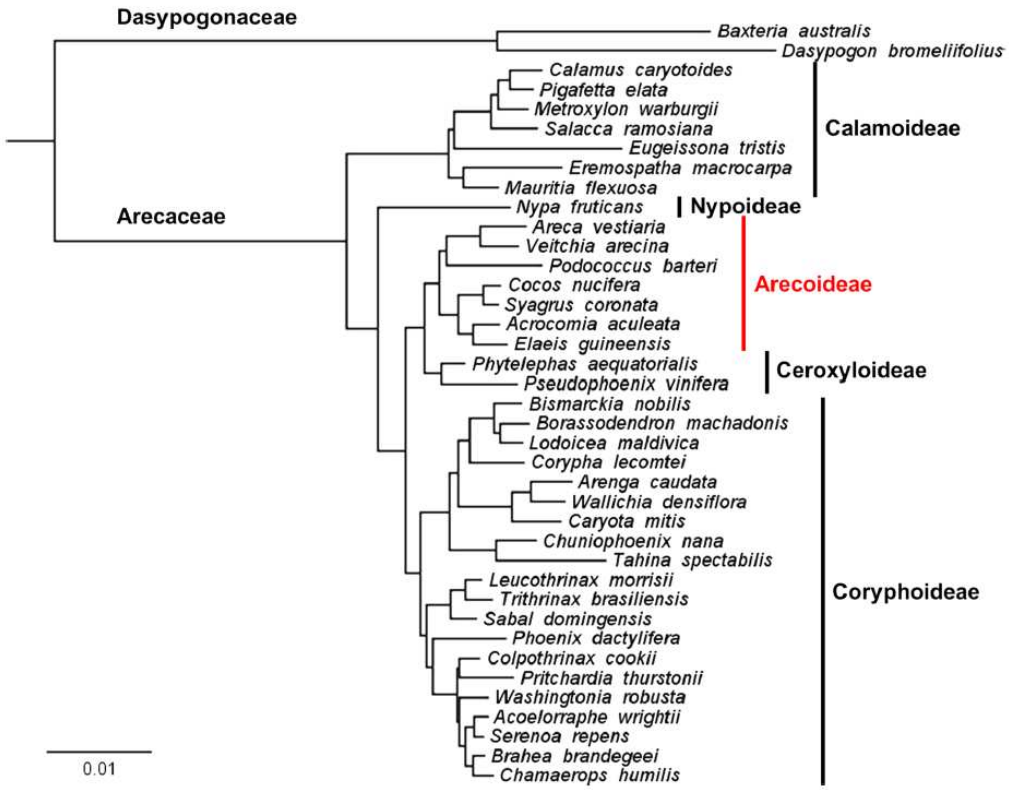


Fig. 6 Arecaceae phylogenomic tree of 40 taxa (37 Arecaceae species and 3 outgroups) based on whole plastomes using bayesian inference. The bayesian posterior probability of all nodes is 1. The branch length is proportional to the inferred divergence level and the scale bar indicates the number of inferred nucleic acids substitutions per site. *Hanguana malayana* (Commelinales) was used to root the tree (omitted from the figure)

noted in the sites of positive selection, including high heterogeneity of amino acid type (e.g. the sites 330, 664, and 686 of *ycf1* gene, Supplementary Fig. S5), unique amino acid type in a particular clade such as subfamily or tribe (e.g. the sites 300, 337, and 338 of *rpoA* gene; Supplementary Fig. S7), and the same amino acid in species or clades that are not closely related (e.g. the sites 89, 142, 219, and 251 of *rbcL* gene; Fig. 8). The latter pattern prevailed in comparison with other types, particularly among photosynthesis-related genes (Fig. 8).

Discussion

The general features of macaw palm plastome are conserved but small expansion and contraction events at the IR boundaries occurred during plastome evolution within Arecoideae

The general structure, the number and type of genes in the plastome of macaw palm are similar to most angiosperms, including other species of Arecaceae (Wicke et al. 2011; Uthaipaisanwong et al. 2012; Huang et al. 2013; Barrett et al. 2016). Nevertheless, at the IR-LSC and IR-SSC junctions we identify sequence variability, indicating the occurrence of small expansion and contraction events at the IR boundaries of Arecoideae plastomes. Similarly, sequence variabilities involving few hundreds of base pairs, especially in the *ycf1* gene at IR-SSC junction and in the *rps19* and *rpl22* genes at IR-LSC junction, are frequently observed as a result of expansion and contractions events by gene conversion (Goulding et al. 1996; Zhu et al. 2016). Despite of these

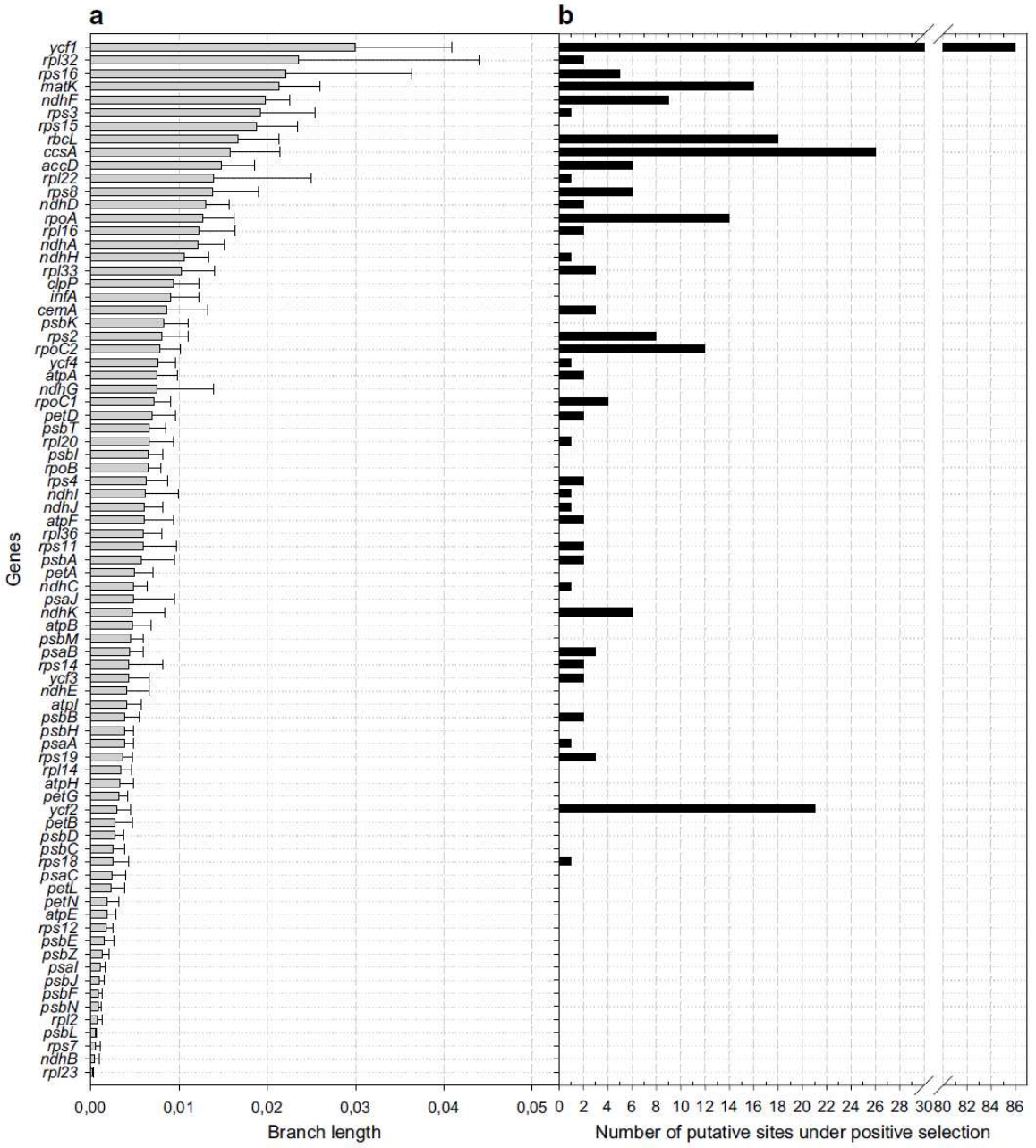


Fig. 7 Molecular evolution analyses of plastid genes within the family Arecaceae. **a** It is shown the divergence of protein-coding genes. The gene divergence was estimated by the sum of total branch lengths in each gene tree inferred (mean ± SD). **b** It is shown the number of putative sites under positive selection

slight differences at IR borders of most Arecoideae plastomes analyzed here, the extremely reduced IRs of *A. vestiaria* and *Tahina spectabilis* plastomes as described by Barrett et al. (2016) are likely the result of large-scale contraction event, which indicates that major changes may have occurred in others species of Arecoideae (Barrett et al. 2016). Large-scale expansion, contraction, and even complete loss of the IR have been reported for some plant lineages belonging to

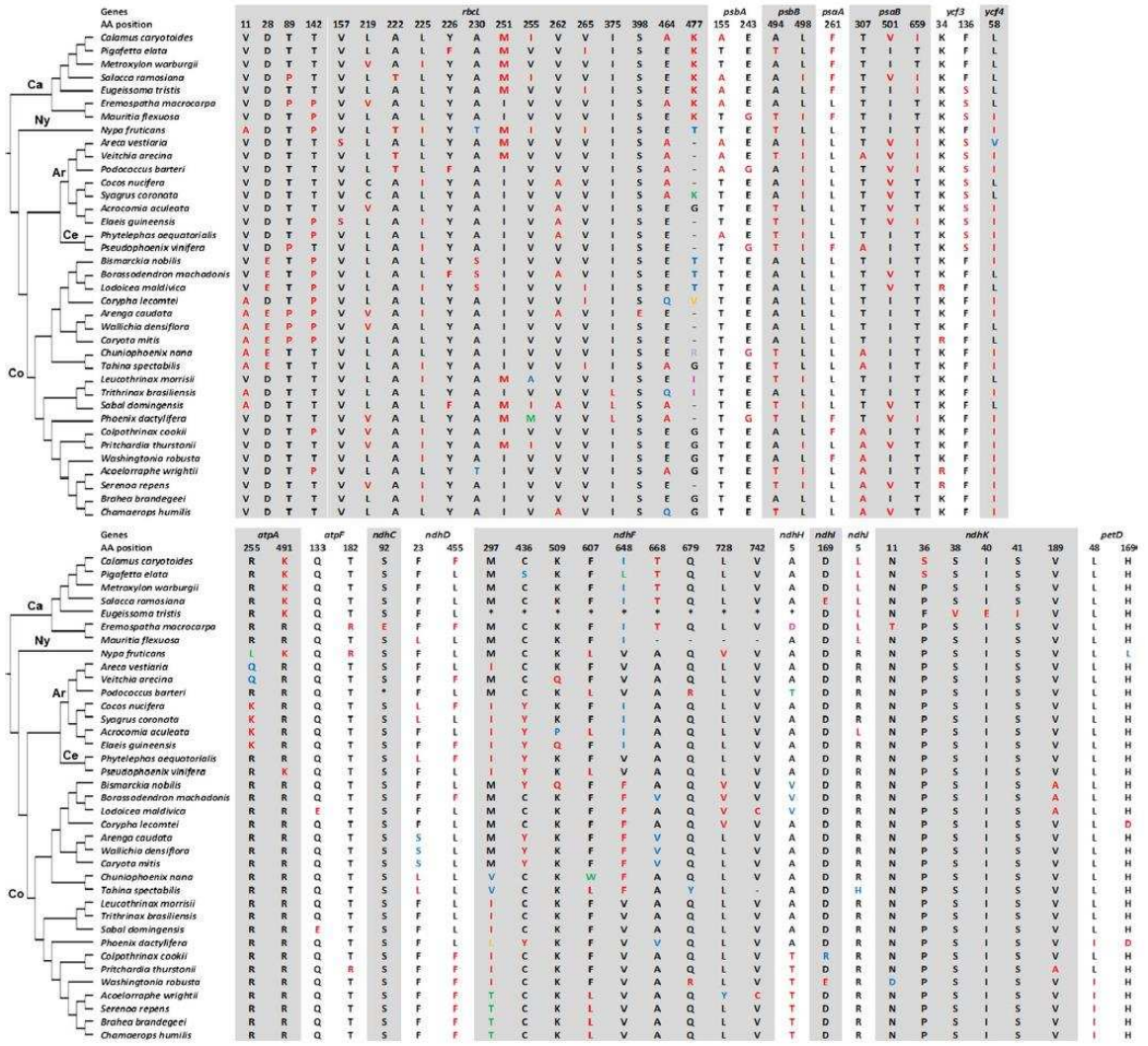


Fig. 8 Sites under positive selection in plastid genes involved in the photosynthetic process. The amino acids are plotted across the palm phylogeny based on whole plastomes. Different amino acid types identified at the same position are highlighted in distinct colors. The amino acid positions are relative to macaw palm plastid genes

the families Campanulaceae (Haberle et al. 2008), Fabaceae (Cai et al. 2008; Guisinger et al. 2011), Geraniaceae (Chumley 2006; Weng et al. 2014), Linaceae (Lopes et al. 2017) and Trochodendraceae (Sun et al. 2013).

The identification of SSR loci and polymorphism hotspots in the plastome of macaw palm represent special sources of molecular markers

Plastid SSRs represent potentially useful markers given that they have demonstrated high levels of intraspecific variability in several studies and due to the nonrecombinant and uniparental inheritance of the plastomes. Such SSRs have been applied in a broad range of researches focusing on studies of genetic diversity within and among natural populations, and characterization of evolutionary history of native and agricultural species (Provan et al. 2001; Ebert and Peakall 2009; Wheeler et al. 2014).

The SSRs located in IGS and introns are more interesting to use as molecular markers since these regions evolve faster than CDSs (Rogalski et al. 2015). In our study we found that 49% of the SSRs loci located in the IGSs and introns of macaw palm plastome are polymorphic when compared with *E. guineesis*. The *ndhC/trnV*-UAC (ten SSRs) and

ndhF/*rpl32* (six SSRs) IGSs, and the introns of *clpP* (eight SSRs) and *trnK*-UUU (seven SSRs) genes contain a high number of SSRs and, therefore, configure important source of molecular markers. The SSRs detailed in the macaw palm plastome, mainly those located in the IGS and introns, represent important sequences to be used together with nuclear molecular markers developed in other groups (Mengistu et al. 2016a, b; Nucci et al. 2008) to study the genetic structure and genetic flow in natural populations of macaw palm (Wheeler et al. 2014). In addition, the SSRs found here are useful tools to characterize germplasm collections, promising genotypes and natural population, in order to search suitable strategies for genetic breeding of this wild species. It is also important to note that strategies using species-specific SSR primers have generated an elevated number of polymorphic loci than those employing universal primers (Wheeler et al. 2014), highlighting the importance of plastome sequencing.

The density of SSR loci was lower in the IRs in Arecoideae species. The lower number of SSR loci in the IRs may be correlated with the duplicative nature of the IRs, which enhances the copy-correction activity by gene conversion, resulting in lower evolutionary rate (Yamane et al. 2006; Zhu et al. 2016). Interestingly, the species *A. vestiaria*, that underwent massive reduction of the IRs, maintains a conserved pattern of SSR distribution in the SSC, LSC, and IR regions as found in other Arecoideae species, suggesting that the contraction of the IR region in *A. vestiaria* may be a recent event.

Lastly, we identify putative fast-evolving plastid sequences in macaw palm plastome based on sliding window analyses. Some IGSs at the level of subfamily and tribe were identified as polymorphism hotspots, especially the *trnC*-GCA/*petN* and *psaC*/*ndhE*. The polymorphism hotspots identified here, mainly in IGSs, represent interesting plastid markers to be used in genetic studies aiming the improvement of characteristics of the interest in macaw palm. Fast-evolving plastid IGSs have been used as efficient tools in several studies involving biogeography, genetic structure, genetic diversity, and germplasm characterization of important commercial species (Tsai et al. 2015; Wambulwa et al. 2016; Roy et al. 2016).

The evolution of putative RNA editing sites within the subfamily Arecoideae

The plastid RNA editing is a posttranscriptional modification (C-to-U and U-to-C conversions) identified in all groups of land plants, except in some species of Marchantiales. The mechanism of RNA editing has a monophyletic origin beginning with the evolution of land plants, when hundreds of editing sites were created in basal lineages. Following the evolution of higher plants several RNA editing sites were lost maintaining a relative conserved number of editing sites in angiosperms (Tillich et al. 2006; Takenaka et al. 2013; He et al. 2016). Plastid RNA editing sites have been identified in mRNA, introns, and untranslated regions, being implicated primordially as error correctors, but also acting in regulatory functions and producing variants of proteins to adapt to different physiological needs (Takenaka et al. 2013; Tseng et al. 2013; He et al. 2016; Chen et al. 2017).

In this study, a total of 100 RNA editing sites were predicted to occur within the subfamily Arecoideae. Like most angiosperms, all editing sites identified are C-to-U conversions, at the first or second position of the codons (Takenaka et al. 2013; He et al. 2016). RNA editing sites at the third position of the codon are less common, generally resulting in synonymous substitutions and having low frequency (He et al. 2016; Chen et al. 2017). From the total sites, 87 are shared among all species sampled in this study and 74 are completely or partially edited in *C. nucifera* (Huang et al., 2013), which suggest high conservation of the RNA editing mechanism within Arecoideae. Interestingly, most RNA editing sites predicted here change the encoded amino acid from polar to apolar, increasing the protein hydrophobicity which can affect structural features such as the creation of new transmembrane regions (He et al. 2016; Chen et al. 2017). Thereby, the general editing process is biased towards to increase the hydrophobicity of plastid proteins, which may be involved in protein–protein interactions and transmembrane domains present in the plastid protein complexes.

Despite the high conservation of RNA editing sites among species of Arecoideae, we identified 13 sites that are not edited in one or more species and six editing sites with codon variants. The phylogenetic tree reconstructed based on these divergent codons may suggest the mode of evolution of the RNA editing sites in the subfamily Arecoideae (Fig. 5a). The reconstructed tree distinguishes the three tribes represented in the analysis (Cocoseae, Podococceae, and Areceae) and the tribe relationships are in accordance with our Arecaceae phylogenomic analysis (discussed below) and Baker et al. (2009). However, the tree based on divergent RNA editing sites does not distinguish the relationships among different genera within the tribe Cocoseae. Therefore, the RNA editing evolution within the subfamily Arecoideae accumulated enough differences (gains, losses, and codon variations) to differentiate tribes but they are not enough to differentiate genera.

The loss of RNA editing sites in higher plants usually occurs when a mutation in DNA sequence changes the C to T, which results in a transcript that codifies the conserved amino acid without need for RNA editing. On the other hand, gains of RNA editing sites arise from the need to correct the mutations that change the T to C in DNA restoring the conserved amino acid (Kahlau et al. 2006; Hein et al. 2016). Our data suggest that at least six gains and five losses

of RNA editing occurred within the subfamily Arecoideae, being species-specific events or events restricted to a closer taxa. Generally, the edited sites conserve the same amino acid among the species sampled here, except three RNA editing sites [*accD* (241), *matK* (219), and *ndhF* (607)] that showed some variability regarding the encoded amino acid. The editing sites that accept some variability is a common feature of plastid transcripts and it can occur in essential and nonessential genes (Fiebig et al. 2004). Indeed, nonessential *ndh* genes (Horváth et al. 2000; Li et al. 2004) have been reported as the most edited genes (Kahlau et al. 2006; He et al. 2016), including in the analysis performed here (37% of the editing sites; Supplementary Tables S4 and S5).

The number of RNA editing sites predicted here for protein-coding genes in palm species is high in comparison with others monocots, as rice (21 sites), maize (26 sites), and orchids (79 sites) (Corneille et al. 2000; Chen et al. 2017). Therefore, future analyses will be important to validate these putative sites identified in Arecoideae species. In addition, comparisons within Arecaceae and other related families may be able to decipher the origin of this unusual high number of RNA editing sites found here.

Arecaceae phylogenomic reconstruction based on whole plastomes

The placement of subfamilies within Arecaceae in our phylogenomic tree was similar to reported by Barrett et al. (2016) based on maximum likelihood analyses using whole plastomes and plastid protein-coding genes. The placement of the tribes within the subfamilies Calamoideae and Coryphoideae is also in accordance with Barrett et al. (2016). However, some incongruences related to the placement of tribes within the subfamily Arecoideae among phylogenies based on nuclear genes (Baker et al. 2011), plastid genes (Comer et al. 2015), and supertree method (Baker et al. 2009) are observed. In our tree, the relationships within Arecoideae are in accordance with Baker et al. (2009). Among Arecoideae species, the macaw palm (*A. aculeata*) is placed within the tribe Cocoseae closed to the genus *Elaeis*, in accordance with Baker et al. (2011).

Molecular evolution of plastid protein-coding genes within Arecaceae

The general structure and gene content of most plastomes of Arecaceae sequenced to date are conserved and a low substitution rate in DNA was observed within Arecaceae in comparison with other commelinids (Barrett et al. 2016). Nevertheless, some exceptions raise questions concerning the plastome evolution of palms. Barrett et al. (2016) reported massive IR loss and a 1944-bp inversion in the LSC in *T. spectabilis* (Coryphoideae), and unusual loss of gene (*ndhF*) and pseudogenization of several genes (*ndhA*, *ndhH*, and *ndhG*) in *Eugeissoma tristis* (Calamoideae). In addition, as previously mentioned, the species *A. vestiaria* (Arecoideae) also lost most part of the IRs (Supplementary Fig. S1). These structural changes are hypothesized to occur idiosyncratically in species or populations instead as synapomorphies in the palm plastid phylogeny (Barrett et al. 2016). Thus, we performed analyses to assess the gene divergence and to investigate the presence of positive selection in each protein-coding gene of palm plastomes to provide insights into the evolution of palm plastid genes. Our gene divergence analysis shows that some genes present wide variation of branch length (e.g. *ycf1*, *rpl32*, and *rps16* genes) because one or two species have notable long branches while the remaining species appear with small branches. In the case of exceptional genes containing nonconserved structure, the particular high gene divergence of some isolated species could be related to gene degeneration process in a species-specific manner rather than a feature shared by a specific clade, subfamily or tribe. Such hypotheses of species-level idiosyncrasies are in accordance with species-level supertree that suggests recent increase and heterogeneity in the diversification rate within genera of family Arecaceae (Faurby et al. 2016).

Moreover, we investigated whether any plastid gene of Arecaceae underwent positive selection. Unexpectedly, we identified a total of 283 putative positive signatures distributed in more than half of plastid genes (40 out of 79 protein-coding genes), including genes related to essential functions such as photosynthesis, plastid gene expression, and fatty acid biosynthesis. In comparison with other monocots, one-third of the plastid genes were reported to evolve under positive selection among species of grasses (Piot et al. 2017) suggesting that positive selection is strongly acting on species of Arecaceae. The occurrence of positive signatures across the palm phylogeny shows frequently the same amino acid change in species or clades that are not closely related. This evolutionary pattern prevailed in comparison with other types among photosynthesis-related genes, which can indicate convergent evolution associated with environmental conditions being shared among unrelated species. Recently, an emblematic case was reported in different lineages of C_4 grasses that show a convergent evolution of several sites in the *rbcL* gene, which underwent positive selection (Piot et al. 2017). Within Arecaceae, most species grow on shaded lower strata of tropical rain forest and their shade adaptation has been correlated with convergent evolution towards high net carbon gain efficiency among different lineages (Ma et al. 2015). It was also recently reported the adaptation of palm species to dry environmental conditions. Bacon et al. (2017) reported recently the shifting of some palm species from tropical rain forest to dry habitats, which is correlated with morphological adaptations. Here, we showed

that several sites in essential genes for photosynthesis (*rbcL*, *psbA*, *psbB*, *psaA*, and *psaB*) are presumably under positive selection. However, the correlation of these sites with the habitats of palm species remains unknown and the existence of specific selective pressures on plastid genes for adaptation to different environmental conditions have to be investigated.

Among photosynthesis-related genes, the *rbcL* gene (which encodes de large subunit of rubisco enzyme) has the highest number of putative positive signatures (Fig. 8). According to Conserved Domains Database (https://www.ncbi.nlm.nih.gov/cdd), the sites 157, 225, 226, and 230 are related to the protein–protein interaction on the heterodimer interface. In addition, the sites 464 and 477 are part of the C-terminal strand that acts on the closing/opening mechanism of the active site and, therefore, they may be important for the catalytic activity (Burisch et al. 2007). Other essential gene for the photosynthesis is the *psbA* gene (encodes the D1 protein, subunit of the reaction center of the PSII). Two positive signatures were identified in this gene at the amino acid position 155, located in the region of chlorophyll biding site, and at the amino acid position 243, situated in the region involved in the protein–protein interaction on the D1-D2 interface (Conserved Domains Database). Positive selection in the amino acid position 155 was also reported among fern species where the evolution of *psbA* gene was related to the competition between ferns and angiosperms for light (Sen et al. 2012). Additionally, we found positive selection in other genes acting on the photochemical machinery, including the *psbB* (encodes the CP47 subunit of PSII), *psaA* (encodes a reaction center subunit of PSI), *psaB* (encodes a reaction center subunit of PSI), *ycf3* (encodes a PSI assembly factor), *ycf4* (encodes a PSI assembly factor), and *petD* (encodes the subunit IV of Cytochrome b_6f) genes. The selective pressures and the consequences of the changes for the photochemical functions are unknown, but positive selections were also identified in *psbB* and *petD* genes of the species of family Brassicaceae (Hu et al. 2015) and grasses (Piot et al. 2017), respectively. Finally, among the eleven *ndh* genes (encode subunits of the chloroplast Ndh complex) seven of them have positive signatures within Arecaceae. In grasses, nine *ndh* genes were found to evolve under positive selection (Piot et al. 2017). The Ndh complex acts in the minor pathway of PSI cyclic electron transport (Shikanai 2016), which is important for plant fitness under various stress conditions (Horváth et al. 2000; Li et al. 2004). The ndh genes have a specific dynamic of evolution, which includes high number of edition sites, pseudogenization, and gene losses or transfer to the nucleus (Martín and Sabater 2010). The selective pressures driving the conservation, positive selection (Fig. 8), or loss (e.g. *E. tristis*) of the Ndh complex within Arecaceae are a matter to be investigated given that it can be an adaptation to different environmental conditions (e.g. light intensities and hydric stress). Similarly, it is also notorious the amount of genes involved in the gene expression machinery containing positive signatures. They include 3 *rpo* genes (*rpoA*, *rpoC1* and *rpoC2*), 5 *rpl* genes (*rpl16*, *rpl20*, *rpl22*, *rpl32* and *rpl33*), 9 *rps* genes (*rps2*, *rps3*, *rps4*, *rps8*, *rps11*, *rps14*, *rps16*, *rps18* and *rps19*), and the *matK* gene (Supplementary Fig. S3, S7, S8, and S9), which totalizes 18 genes. Several studies have reported accelerated evolutionary rates and positive selection on plastid genes involved in gene expression (Krawczyk and Sawicki 2013; Hu et al. 2015; Xu et al. 2015; Weng et al. 2016; Piot et al. 2017). However, the effect of these signatures on the protein function and the adaptive capacity are poorly understood.

The high number of genes containing positive signatures, the wide distribution of these signatures across palm phylogeny, and several cases of putative convergent evolution may be related to the recent increase of diversification rate of Arecaceae species (Faurby et al. 2016). However, we cannot discard the possibility that some putative positive signatures are RNA editing sites involved in the recovery of conserved amino acids. Comparing the RNA editing sites predicted among species of Arecoideae (Supplementary Tables S4 and S5) with the putative positive signatures of Arecaceae (Fig. 8, Supplementary Figs. S3–S9), we identified five positive signatures that could be RNA editing sites. These sites are located at the position 218 and 309 of *matK*, 273 of *ccsA*, 136 of *ycf3*, and 607 of *ndhF* genes, which correspond to the RNA editing sites 219, 310, 274, 138, and 607 in the respective genes in species of Arecoideae. The use of next-generation sequencing (NGS) to investigate RNA editing sites has expanded the number of new RNA editing sites (He et al. 2016; Hein et al. 2016; Chen et al. 2017), including positive pressures acting on the editing sites of some genes (He et al. 2016). Therefore, it will be interesting to carry out RNA editing studies based on NGS technology in Arecaceae to confirm putative positive signatures and new editing sites expanding our knowledge about the selective pressures acting on plastid genes of Arecaceae species.

Conclusions

Here we reported the complete plastome of macaw palm, which was carefully characterized regarding the gene content, structure and evolution. A total of 221 SSR loci were identified and localized along the plastome of macaw palm. Eight regions of high polymorphism (hotspots) at level of subfamily and tribe were determined, being the most divergent the *trnC*-GCA/*petN* and *psaC*/*ndhE* IGSs. In addition, we performed a phylogenomic analysis using whole plastomes of 40 taxa, including 37 species representing all five subfamilies of Arecaceae, which placed the macaw palm within the tribe Cocoseae closed to *Elaeis* genus. Moreover, the analysis of

RNA editing among Arecoideae species, including *A. aculeata*, indentified 100 putative sites within this subfamily and indicated possible events of gain and loss of editing sites within this subfamily. Furthermore, we analyzed extensively the molecular evolution of plastid genes within the family Arecaceae, covering the full set of plastid protein-coding genes. The data revealed the presence of highly divergent genes in a species-specific manner suggesting that gene degeneration processes may be occurring within Arecaceae at the level of genus and/or species. These analyses also revealed unexpectedly that more than half of all plastid protein-coding genes within Arecaceae are under positive selection, which can presumably affect essential plastid functions. The distribution of these positive signatures across the Arecaceae phylogenomic suggests convergent evolution of most sites, including genes involving in photosynthesis. Finally, the SSRs and polymorphism hotspots identified here increase significantly the genetic information available for boost genetic studies in natural populations and germplasm collections of macaw palm aiming domestication and genetic breeding. The findings showed here via molecular analyses have important implications in the areas of genetic, evolution, conservation, breeding, and biotechnology of macaw palm and other species of Arecaceae.

Author contribution statement ASL, TGP, LNV, MPG, EMS, FOP, and MR conceived and designed the research. ASL, TGP, TN, LNV, EMS, FOP, and MR conducted experiments and analyzed the data. MPG, RON, EMS, FOP, and MR contributed with reagents and materials. ASL and MR wrote the manuscript. All authors read and approved the manuscript.

Acknowledgements This research was supported by the National Council for Scientific and Technological Development, Brazil (CNPq, Grant 459698/2014-1). We are grateful to INCT-FBN and for the scholarships granted by the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES) to TGP and LNV, and those granted by the CNPq to ASL, MPG, RON, EMS and FOP. We would like to thank Dr. Sergio Y. Motoike (Professor—Universidade Federal de Viçosa) and Francisco de Assis Lopes (Technical Assistant—Universidade Federal de Viçosa) for their assistance with the collection of plant material for chloroplast isolation.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Alkatib S, Scharff LB, Rogalski M, Fleischmann TT, Matthes A, Seeger S et al (2012) The conributions of wobbling and superwobbling to the reading of the genetic code. PLoS Genet 8:e1003076. https://doi.org/10.1371/journal.pgen.1003076

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

Bacon CD, Moraes RM, Jaramillo C, Antonelli A (2017) Endemic palm species shed light on habitat shifts and the assembly of the Cerrado and Restinga floras. Mol Phylogenet Evol 110:127–133. https://doi.org/10.1016/j.ympev.2017.03.013

Baker WJ, Savolainen V, Asmussen-Lange CB, Chase MW, Dransfield J, Forest F, Harley MM, Uhl NW, Wilkinson M (2009) Complete generic-level phylogenetic analyses of palms (Arecaceae) with comparisons of supertree and supermatrix approaches. Syst Biol 58:240–256. https://doi.org/10.1093/sysbio/syp021

Baker WJ, Norup MV, Clarkson JJ, Couvreur TLP, Dowe JL, Lewis CE, Pintaud JC, Savolainen V, Wilmot T, Chase MW (2011) Phylogenetic relationships among arecoid palms (Arecaceae: Arecoideae). Ann Bot 108:1417–1432. https://doi.org/10.1093/aob/mcr020

Barfod AS, Hagen M, Borchsenius F (2011) Twenty-five years of progress in understanding pollination mechanisms in palms (Arecaceae). Ann Bot 108:1503–1516. https://doi.org/10.1093/aob/mcr192

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. https://doi.org/10.1111/nph.13617

Berton LHC, de Azevedo Filho JA, Siqueira WJ, Colombo CA (2013) Seed germination and estimates of genetic parameters of promising macaw palm (*Acrocomia aculeata*) progenies for biofuel production. Ind Crops Prod 51:258–266

Bock R (2017) Witnessing genome evolution: experimental reconstruction of endosymbiotic and horizontal gene transfer. Annu Rev Genet. https://doi.org/10.1146/annurev-genet-120215-035329 **(in press)**

Burisch C, Wildner GF, Schlitter J (2007) Bioinformatic tools uncover the C-terminal strand of Rubisco's large subunit as hot-spot for specificity-enhancing mutations. FEBS Lett 581:741–748. https://doi.org/10.1016/j.febslet.2007.01.043

Cai Z, Guisinger M, Kim HG, Ruck E, Blazier JC, McMurtry V, Kuehl JV, Boore J, Jansen RK (2008) Extensive reorganization of the plastid genome of trifolium subterraneum (Fabaceae) is associated with numerous repeated sequences and novel DNA insertions. J Mol Evol 67:696–704. https://doi.org/10.1007/s00239-008-9180-7

Chen TC, Liu YC, Wang X, Wu CH, Huang CH, Chang CC (2017) Whole plastid transcriptomes reveal abundant RNA editing sites and differential editing status in *Phalaenopsis aphrodite* subsp. formosana. Bot Stud 58:38. https://doi.org/10.1186/s40529-017-0193-7

Chumley TW (2006) The complete chloroplast genome sequence of Pelargonium × hortorum: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. Mol Biol Evol 23:2175–2190

Ciconini G, Favaro SP, Roscoe R, Miranda CHB, Tapeti CF, Miyahira MAM, Bearari L, Galvani F, Borsato AV, Colnago LA et al (2013) Biometry and oil contents of *Acrocomia aculeata* fruits from the Cerrados and Pantanal biomes in Mato Grosso do Sul, Brazil. Ind Crops Prod 45:208–214

Comer JR, Zomlefer WB, Barrett CF, Davis JI, Stevenson DW, Heyduk K, Leebens-Mack JH (2015) Resolving relationships within the palm subfamily Arecoideae (Arecaceae) using plastid sequences

derived from next-generation sequencing. Am J Bot 102:888–899. https://doi.org/10.3732/ajb.1500057
Conceição LDHCS, Antoniassi R, Junqueira NTV, Braga MF, de Faria-Machado AF, Rogério JB, Duarte ID, Bizzo HR (2015) Genetic diversity of macauba from natural populations of Brazil. BMC Res Notes 8:406. https://doi.org/10.1186/s13104-015-1335-1
Corneille S, Lutz K, Maliga P (2000) Conservation of RNA editing between rice and maize plastids: are most editing events dispensable? Mol Gen Genet MGG 264:419–424
Coser SM, Motoike SY, Corrêa TR, Pires TP, Resende MDV (2016) Breeding of *Acrocomia aculeata* using genetic diversity parameters and correlations to select accessions based on vegetative, phenological, and reproductive characteristics. Genet Mol Res. https://doi.org/10.4238/gmr15048820
Daniell H, Lin CS, Yu M, Chang WJ (2016) Chloroplast genomes: diversity, evolution, and applications in genetic engineering. Genome Biol 17:134. https://doi.org/10.1186/s13059-016-1004-2
Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. Nat Methods 9:772. https://doi.org/10.1038/nmeth.2109
Dransfield J, Uhl NW, Lange CBA, Baker WJ, Harley MM, Lewis CE (2005) A new phylogenetic classification of the palm family, Arecaceae. Kew Bull 60:559–569
Ebert D, Peakall R (2009) Chloroplast simple sequence repeats (cpSSRs): technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant species. Mol Ecol Resour 9:673–690. https://doi.org/10.1111/j.1755-0998.2008.02319.x
Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340
Faurby S, Eiserhardt WL, Baker WJ, Svenning JC (2016) An all-evidence species-level supertree for the palms (Arecaceae). Mol Phylogenet Evol 100:57–69. https://doi.org/10.1016/j.ympev.2016.03.002
Fiebig A, Stegemann S, Bock R (2004) Rapid evolution of RNA editing sites in a small non-essential plastid gene. Nucleic Acids Res 32:3615–3622. https://doi.org/10.1093/nar/gkh695
Fuentes P, Armarego-Marriott T, Bock R (2017) Plastid transformation and its application in metabolic engineering. Curr Opin Biotechnol 49:10–15. https://doi.org/10.1016/j.copbio.2017.07.004
Gonçalves DB, Batista AF, Rodrigues MQRB, Nogueira KMV, Santos VL (2013) Ethanol production from macaúba (*Acrocomia aculeata*) presscake hemicellulosic hydrolysate by *Candida boidinii* UFMG14. Bioresour Technol 146:261–266. https://doi.org/10.1016/j.biortech.2013.07.075
Goulding SE, Olmstead RG, Morden CW, Wolfe KH (1996) Ebb and flow of the chloroplast inverted repeat. Mol Gen Genet 252:195–206
Guisinger MM, Kuehl JV, Boore JL, Jansen RK (2011) Extreme reconfiguration of plastid genomes in the angiosperm family geraniaceae: rearrangements, repeats, and codon usage. Mol Biol Evol 28:583–600. https://doi.org/10.1093/molbev/msq229
Haberle RC, Fourcade HM, Boore JL, Jansen RK (2008) Extensive rearrangements in the chloroplast genome of *Trachelium caeruleum* are associated with repeats and tRNA genes. J Mol Evol 66:350–361. https://doi.org/10.1007/s00239-008-9086-4
He P, Huang S, Xiao G, Zhang Y, Yu J (2016) Abundant RNA editing sites of chloroplast protein-coding genes in *Ginkgo biloba* and an evolutionary pattern analysis. BMC Plant Biol 16(1):257. https://doi.org/10.1186/s12870-016-0944-8
Hein A, Polsakiewicz M, Knoop V (2016) Frequent chloroplast RNA editing in early-branching flowering plants: pilot studies on angiosperm-wide coexistence of editing sites and their nuclear specificity factors. BMC Evol Biol 16:23. https://doi.org/10.1186/s12862-016-0589-0
Henderson A, Galeano G, Bernal R (1995) Field guide to the palms of the Americas. Princeton University Press, Princeton
Horváth EM, Peter SO, Joët T, Rumeau D, Cournac L, Horváth GV, Kavanagh TA, Schäfer C, Peltier G, Medgyesy P (2000) Targeted inactivation of the plastid ndhB gene in tobacco results in an enhanced sensitivity of photosynthesis to moderate stomatal closure. Plant Physiol 123:1337–1350
Hu S, Sablok G, Wang B, Qu D, Barbaro E, Viola R, Li M, Varotto C (2015) Plastome organization and evolution of chloroplast genes in Cardamine species adapted to contrasting habitats. BMC Genomics 16:306. https://doi.org/10.1186/s12864-015-1498-0
Huang YY, Matzke AJM, Matzke M (2013) Complete sequence and comparative analysis of the chloroplast genome of coconut palm (*Cocos nucifera*). PLoS One 8:e74736. https://doi.org/10.1371/journal.pone.0074736
Kahlau S, Aspinall S, Gray JC, Bock R (2006) Sequence of the tomato chloroplast DNA and evolutionary comparison of solanaceous plastid genomes. J Mol Evol 63:194–207. https://doi.org/10.1007/s00239-005-0254-5
Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010
Krawczyk K, Sawicki J (2013) The uneven rate of the molecular evolution of gene sequences of DNA-dependent RNA polymerase I of the genus *Lamium* L. Int J Mol Sci 14:11376–11391. https://doi.org/10.3390/ijms140611376
Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. Genome Biol 5:R12. https://doi.org/10.1186/gb-2004-5-2-r12
Lanes ÉCM, de Almeida Costa PM, Motoike SY (2014) Alternative fuels: Brazil promotes aviation biofuels. Nature 511:31. https://doi.org/10.1038/511031a
Lanes ÉCM, Motoike SY, Kuki KN, Nick C, Freitas RD (2015) Molecular characterization and population structure of the macaw palm, *Acrocomia aculeata* (Arecaceae), ex situ germplasm collection using microsatellites markers. J Hered 106:102–112. https://doi.org/10.1093/jhered/esu073
Lanes ÉCM, Motoike SY, Kuki KN, Resende MDV, Caixeta ET (2016) Mating system and genetic composition of the macaw palm (*Acrocomia aculeata*): implications for breeding and genetic conservation programs. J Hered 107:527–536. https://doi.org/10.1093/jhered/esw038
Li XG, Duan W, Meng QW, Zou Q, Zhao SJ (2004) The function of chloroplastic NAD(P)H dehydrogenase in tobacco during chilling stress under low irradiance. Plant Cell Physiol 45:103–108
Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451–1452. https://doi.org/10.1093/bioinformatics/btp187
Lohse M, Drechsel O, Kahlau S, Bock R (2013) OrganellarGenomeDRAW—a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res 41:W575–W581. https://doi.org/10.1093/nar/gkt289
Lopes AS, Pacheco TG, Santos KG, Vieira LN, Guerra MP, Nodari RO, Souza EM, Pedrosa FO, Rogalski M (2017) The *Linum usitatissimum* L. plastome reveals atypical structural evolution, new editing sites, and the phylogenetic position of Linaceae within Malpighiales. Plant Cell Rep. https://doi.org/10.1007/s00299-017-2231-z **(in press)**
Lorenzi H, Kahn F, Noblick LR, Ferreira E (2010) Flora Brasileira—Arecaceae (Palmeiras). Instituto Plantarum, Nova Odessa, p 368

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964

Luis ZG, Scherwinski-Pereira EJ (2014) An improved protocol for somatic embryogenesis and plant regeneration in macaw palm (*Acrocomia aculeata*) from mature zygotic embryos. Plant Cell Tissue Organ Cult 118:485–496. https://doi.org/10.1007/s11240-014-0500-x

Ma RY, Zhang JL, Cavaleri MA, Sterck F, Strijk JS, Cao KF (2015) Convergent evolution towards high net carbon gain efficiency contributes to the shade tolerance of palms (Arecaceae). PLoS One 10:e0140384. https://doi.org/10.1371/journal.pone.0140384

Machado W, Figueiredo A, Guimarães MF (2016) Initial development of seedlings of macauba palm (*Acrocomia aculeata*). Ind Crops Prod 87:14–19

Martín M, Sabater B (2010) Plastid ndh genes in plant evolution. Plant Physiol Biochem 48:636–645. https://doi.org/10.1016/j.plaphy.2010.04.009

Mengistu FG, Motoike SY, Caixeta ET, Cruz CD, Kuki KN (2016a) Cross-species amplification and characterization of new microsatellite markers for the macaw palm, *Acrocomia aculeata* (Arecaceae). Plant Genet Resour 14:163–172

Mengistu FG, Motoike SY, Cruz CD (2016b) Molecular characterization and genetic diversity of the macaw palm ex situ germplasm collection revealed by microsatellite markers. Diversity 8:20

Montoya SG, Motoike SY, Kuki KN, Couto AD (2016) Fruit development, growth, and stored reserves in macauba palm (*Acrocomia aculeata*), an alternative bioenergy crop. Planta 244:927–938. https://doi.org/10.1007/s00425-016-2558-7

Motoike SY, Kuki KN (2009) The potential of macaw palm (*Acrocomia aculeata*) as source of biodiesel in Brazil. IRECHE 1:632–635

Moura EF, Motoike SY, Ventrella MC, Sá AQ, Carvalho JM (2009) Somatic embryogenesis in macaw palm (*Acrocomia aculeata*) from zygotic embryos. Sci Hortic 119:447–454. https://doi.org/10.1016/j.scienta.2008.08.033

Mower JP (2009) The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. Nucleic Acids Res 37:W253–W259. https://doi.org/10.1093/nar/gkp337

Nucci SM, Azevedo-Filho JA, Colombo CA, Priolli RHG, Coelho RM, Mata TL, Zucchi MI (2008) Development and characterization of microsatellites markers from the macaw. Mol Ecol Resour 8:224–226. https://doi.org/10.1111/j.1471-8286.2007.01932.x

Padilha JHD, Ribas LLF, Amano É, Quoirin M (2015) Somatic embryogenesis in *Acrocomia aculeata* Jacq. (Lodd.) ex Mart using the thin cell layer technique. Acta Bot Bras 29:516–523. https://doi.org/10.1590/0102-33062015abb0109

Park S, Ruhlman TA, Weng ML, Hajrah NH, Sabir JSM, Jansen RK (2017) Contrasting patterns of nucleotide substitution rates provide insight into dynamic evolution of plastid and mitochondrial genomes of geranium. Genome Biol Evol 9:1766–1780. https://doi.org/10.1093/gbe/evx124

Piot A, Hackel J, Christin PA, Besnard G (2017) One-third of the plastid genes evolved under positive selection in PACMAD grasses. Planta. https://doi.org/10.1007/s00425-017-2781-x

Pires TP, dos Santos Souza E, Kuki KN, Motoike SY (2013) Ecophysiological traits of the macaw palm: a contribution towards the domestication of a novel oil crop. Ind Crops Prod 44:200–210

Provan J, Powell W, Hollingsworth PM (2001) Chloroplast microsatellites: new tools for studies in plant ecology and evolution. Trends Ecol Evol 16:142–147

Rogalski M, Carrer H (2011) Engineering plastid fatty acid biosynthesis to improve food quality and biofuel production in higher plants: plastid fatty acid biosynthesis. Plant Biotechnol J 9:554–564. https://doi.org/10.1111/j.1467-7652.2011.00621.x

Rogalski M, Ruf S, Bock R (2006) Tobacco plastid ribosomal protein S18 is essential for cell survival. Nucleic Acids Res 34:4537–4545. https://doi.org/10.1093/nar/gkl634

Rogalski M, Schoettler MA, Thiele W, Schulze WX, Bock R (2008) Rpl33, a nonessential plastid encoded ribosomal protein in tobacco, is required under cold stress conditions. Plant Cell 20:2221–2237. https://doi.org/10.1105/tpc.108.060392

Rogalski M, do Nascimento Vieira L, Fraga HP, Guerra MP (2015) Plastid genomics in horticultural species: importance and applications for plant population genetics, evolution, and biotechnology. Front Plant Sci 6:586. https://doi.org/10.3389/fpls.2015.00586

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61:539–542. https://doi.org/10.1093/sysbio/sys029

Roy PS, Rao GJN, Jena S, Samal R, Patnaik A, Patnaik SSC, Jambhulkar NN, Sharma S, Mohapatra T (2016) Nuclear and chloroplast DNA Variation provides insights into population structure and multiple origin of native aromatic rices of Odisha, India. PloS One 11:e0162268. https://doi.org/10.1371/journal.pone.0162268

Ruhlman TA, Zhang J, Blazier JC, Sabir JSM, Jansen RK (2017) Recombination-dependent replication and gene conversion homogenize repeat sequences and diversify plastid genome structure. Am J Bot 104:559–572. https://doi.org/10.3732/ajb.1600453

Sen L, Fares M, Su YJ, Wang T (2012) Molecular evolution of psbA gene in ferns: unraveling selective pressure and co-evolutionary pattern. BMC Evol Biol 12:145. https://doi.org/10.1186/1471-2148-12-145

Shikanai T (2016) Chloroplast NDH: a different enzyme with a structure similar to that of respiratory NADH dehydrogenase. Biochim Biophys Acta 1857:1015–1022. https://doi.org/10.1016/j.bbabio.2015.10.013

Stern A, Doron-Faigenboim A, Erez E, Martz E, Bacharach E, Pupko T (2007) Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. Nucleic Acids Res 35:W506–W511. https://doi.org/10.1093/nar/gkm382

Sun Y, Moore MJ, Meng A, Soltis PS, Soltis DE, Li J, Wang H (2013) Complete plastid genome sequencing of Trochodendraceae reveals a significant expansion of the inverted repeat and suggests a Paleogene divergence between the two extant species. PLoS One 8:e60429. https://doi.org/10.1371/journal.pone.0060429

Takenaka M, Zehrmann A, Verbitskiy D, Härtel B, Brennicke A (2013) RNA editing in plants and its evolution. Annu Rev Genet 47:335–352. https://doi.org/10.1146/annurev-genet-111212-133519

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30:2725–2729. https://doi.org/10.1093/molbev/mst197

The Plant List (2013) Version 1.1. Retrieved from http://www.theplantlist.org/

Thiel T, Michalek W, Varshney R, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor Appl Genet 106:411–422. https://doi.org/10.1007/s00122-002-1031-0

Tillich M, Lehwark P, Morton BR, Maier UG (2006) The evolution of chloroplast RNA editing. Mol Biol Evol 23:1912–1921. https://doi.org/10.1093/molbev/msl054

Tsai CC, Chou CH, Wang HV, Ko YZ, Chiang TY, Chiang YC (2015) Biogeography of the *Phalaenopsis amabilis* species complex

inferred from nuclear and plastid DNAs. BMC Plant Biol 15:202. https://doi.org/10.1186/s12870-015-0560-z

Tseng CC, Lee CJ, Chung YT, Sung TY, Hsieh MH (2013) Differential regulation of Arabidopsis plastid gene expression and RNA editing in non-photosynthetic tissues. Plant Mol Biol 82(4–5):375–392

Uthaipaisanwong P, Chanprasert J, Shearman JR, Sangsrakru D, Yoocha T, Jomchai N, Jantasuriyarat C, Tragoonrung S, Tangphatsornruang S (2012) Characterization of the chloroplast genome sequence of oil palm (*Elaeis guineensis* Jacq.). Gene 500:172–180. https://doi.org/10.1016/j.gene.2012.03.061

Vianna SA (2017) A new species of Acrocomia (Arecaceae) from Central Brazil. Phytotaxa 314:45–54. https://doi.org/10.11646/phytotaxa.314.1.2

Vieira LN, Faoro H, Fraga HPF, Rogalski M, de Souza EM, de Oliveira Pedrosa F, Nodari RO, Guerra MP (2014) An improved protocol for intact chloroplasts and cpdna isolation in conifers. PLoS One 9:e84792. https://doi.org/10.1371/journal.pone.0084792

Vieira LN, Dos Anjos KG, Faoro H, Fraga HP, Greco TM, Pedrosa FO, de Souza EM, Rogalski M, de Souza RF, Guerra MP (2016a) Phylogenetic inference and SSR characterization of tropical woody bamboos tribe Bambuseae (Poaceae: Bambusoideae) based on complete plastid genome sequences. Curr Genet 62(2):443–453. https://doi.org/10.1007/s00294-015-0549-z

Vieira LN, Rogalski M, Faoro H, Fraga HP, Anjos KG, Picchi GFA, Nodari RO, Pedrosa FO, Souza EM, Guerra MP (2016b) The plastome sequence of the endemic Amazonian conifer, *Retrophyllum piresii* (Silba) C.N.Page, reveals different recombination events and plastome isoforms. Tree Genet Genomes 12:10. https://doi.org/10.1007/s11295-016-0968-0

Vilas-Boas MA, Carneiro ACO, Vital BR, Carvalho AMML, Martins MA (2010) Efeito da temperatura de carbonização e dos resíduos de macaúba na produção de carvão vegetal. Sci For 38:481–490

Wambulwa MC, Meegahakumbura MK, Kamunya S, Muchugi A, Möller M, Liu J, Xu JC, Ranjitkar S, Li DZ, Gao LM (2016) Insights into the genetic relationships and breeding patterns of the African tea germplasm based on nSSR markers and cpDNA sequences. Front Plant Sci 7:1244. https://doi.org/10.3389/fpls.2016.01244

Weng ML, Blazier JC, Govindu M, Jansen RK (2014) Reconstruction of the ancestral plastid genome in geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. Mol Biol Evol 31:645–659. https://doi.org/10.1093/molbev/mst257

Weng ML, Ruhlman TA, Jansen RK (2016) Plastid-nuclear interaction and accelerated coevolution in plastid ribosomal genes in Geraniaceae. Genome Biol Evol 8:1824–1838. https://doi.org/10.1093/gbe/evw115

Wheeler GL, Dorman HE, Buchanan A, Challagundla L, Wallace LE (2014) A review of the prevalence, utility, and caveats of using chloroplast simple sequence repeats for studies of plant biology. Appl Plant Sci. https://doi.org/10.3732/apps.1400059

Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, Quandt D (2011) The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol 76:273–297. https://doi.org/10.1007/s11103-011-9762-4

Williams AV, Boykin LM, Howell KA, Nevill PG, Small I (2015) The complete sequence of the acacia ligulata chloroplast genome reveals a highly divergent clpP1 gene. PLoS One 10:e0125768. https://doi.org/10.1371/journal.pone.0125768

Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20:3252–3255. https://doi.org/10.1093/bioinformatics/bth352

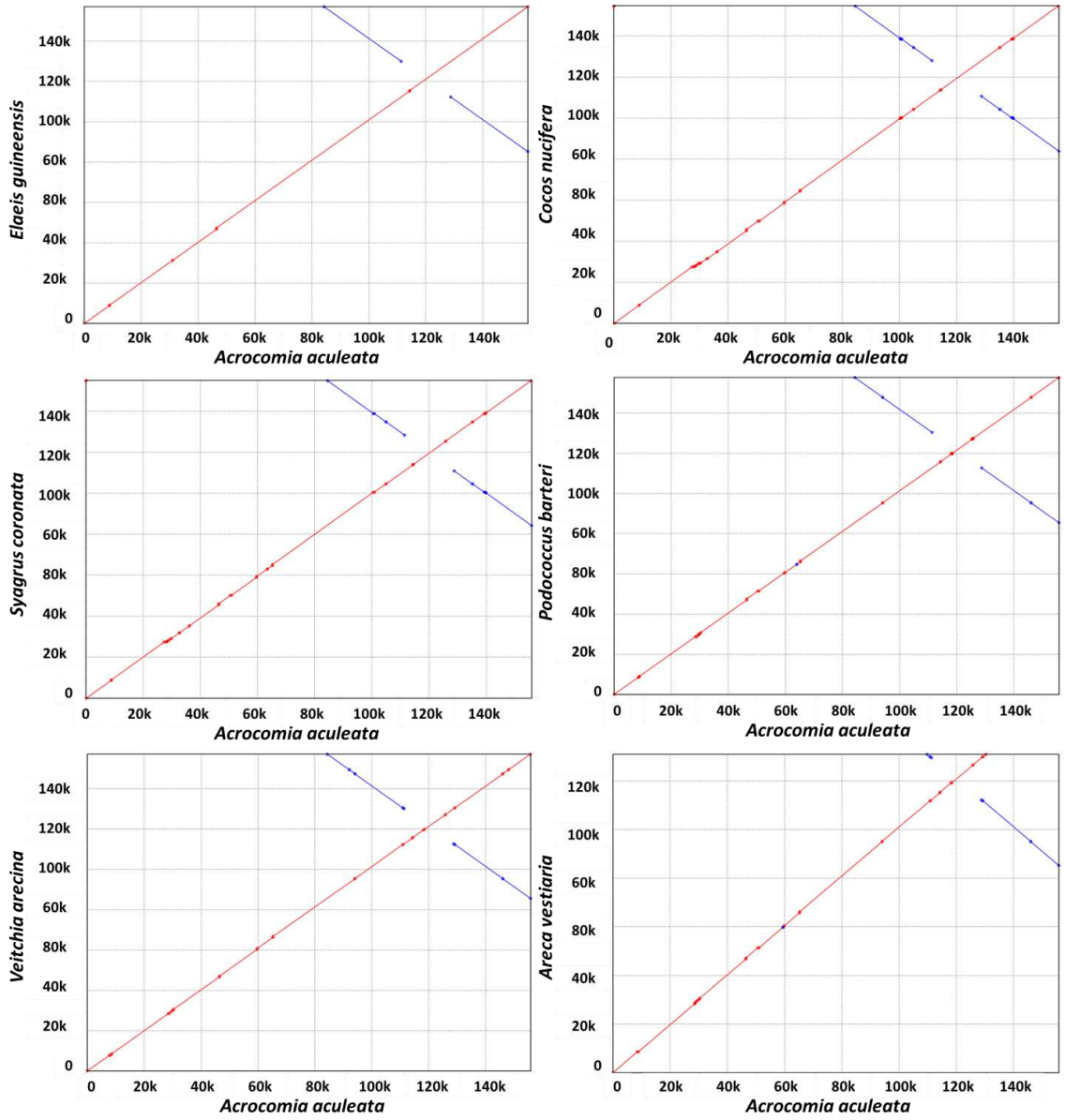
Xu JH, Liu Q, Hu W, Wang T, Xue Q, Messing J (2015) Dynamics of chloroplast genomes in green plants. Genomics 106:221–231. https://doi.org/10.1016/j.ygeno.2015.07.004

Yamane K, Yano K, Kawahara T (2006) Pattern and rate of indel evolution inferred from whole chloroplast intergenic regions in sugarcane, maize and rice. DNA Res Int J Rapid Publ Rep Genes Genomes 13:197–204. https://doi.org/10.1093/dnares/dsl012

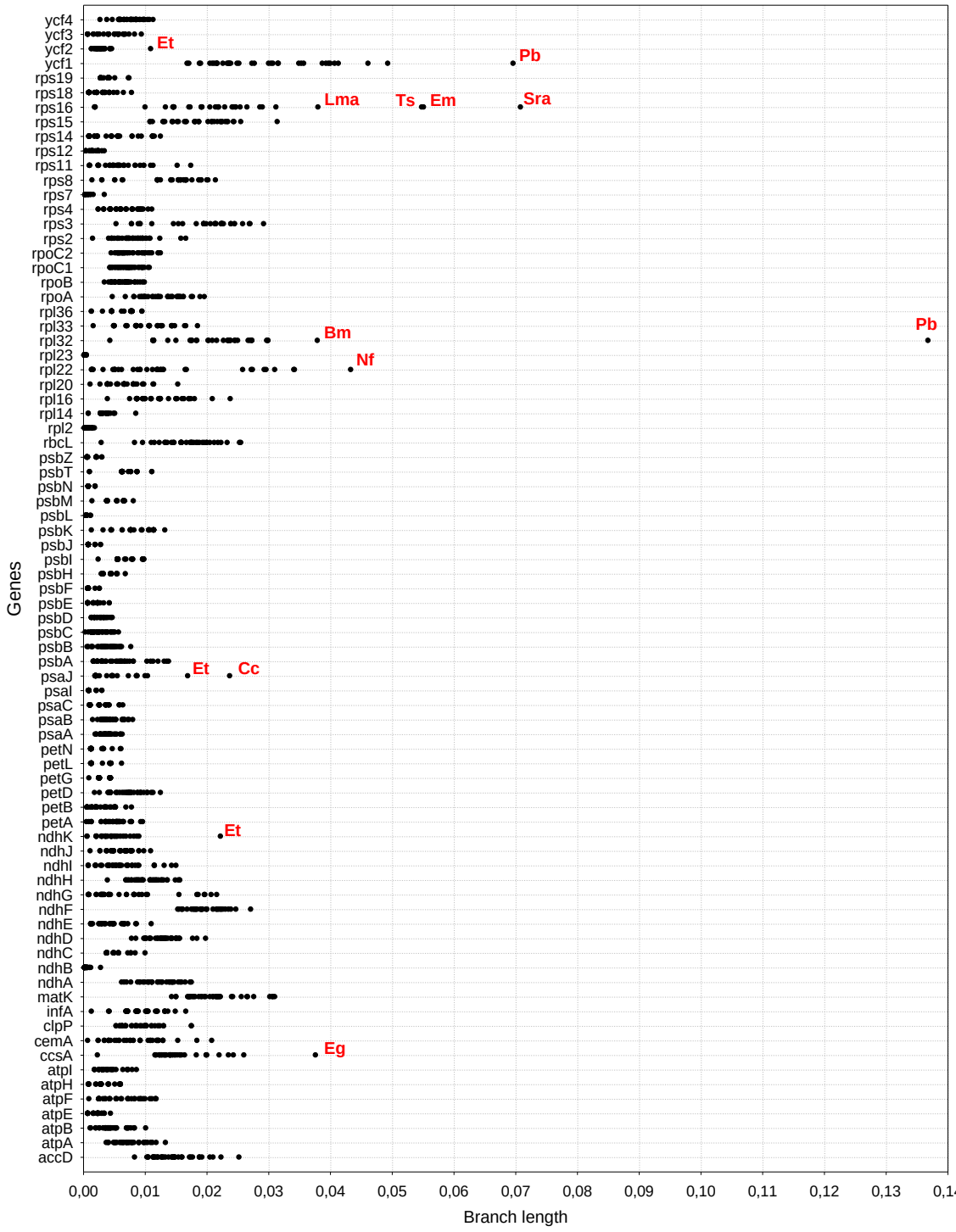
Zhang J, Khan SA, Heckel DG, Bock R (2017) Next-generation insect-resistant plants: RNAi-mediated crop protection. Trends Biotechnol 35:871–882. https://doi.org/10.1016/j.tibtech.2017.04.009

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

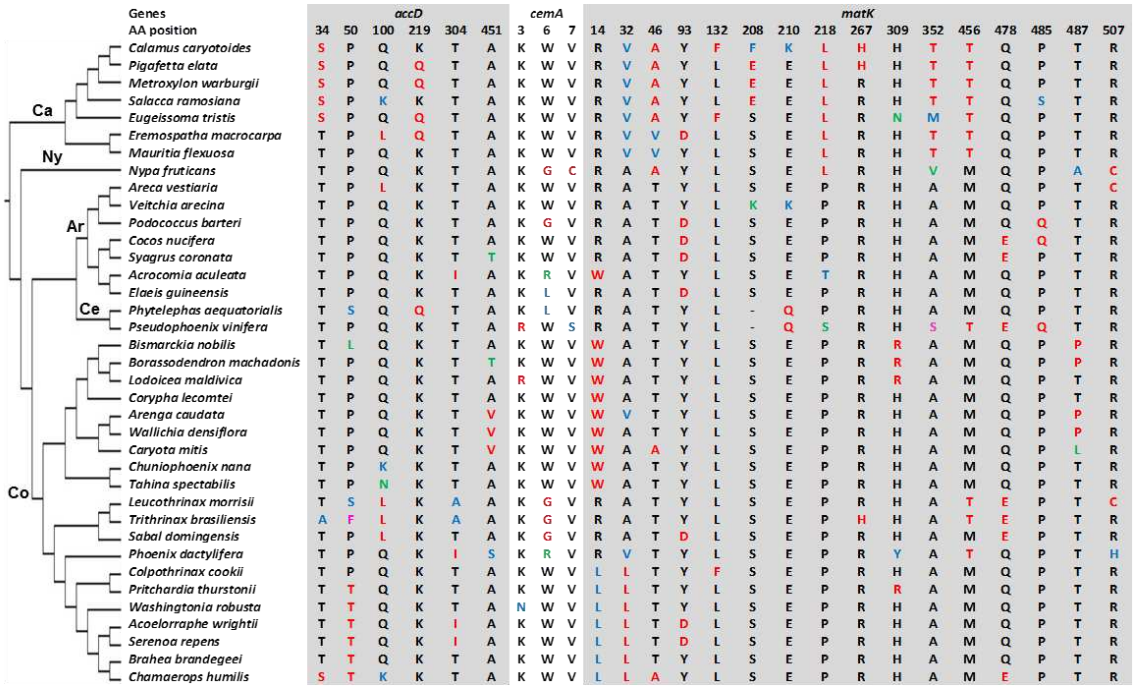
Supplementary Figures



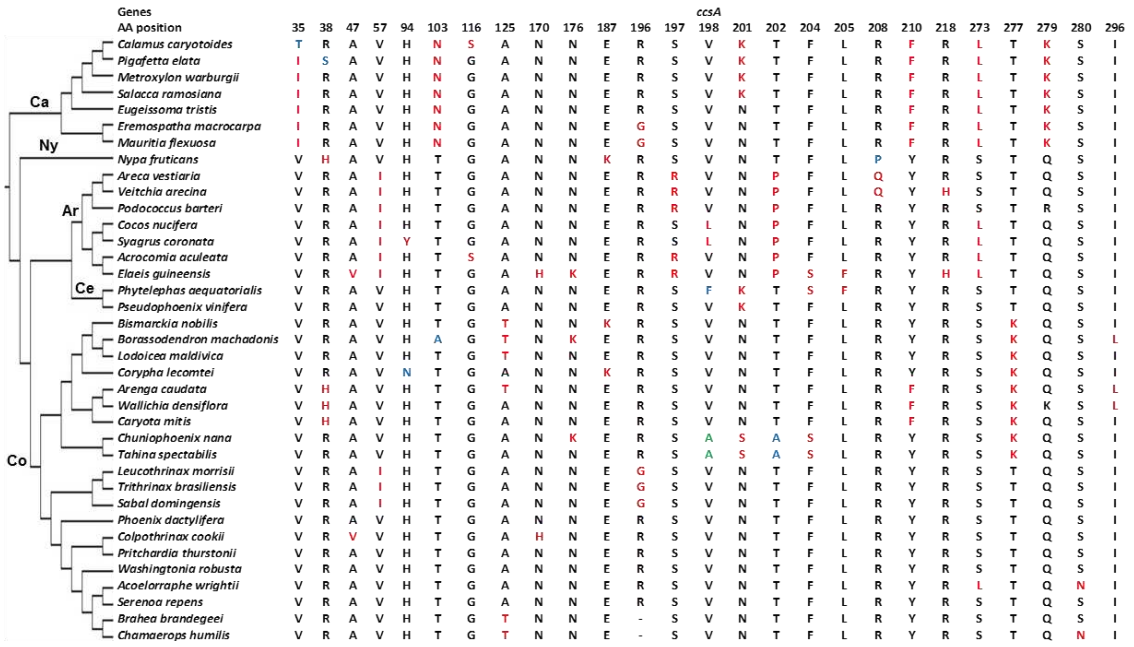
Supplementary Fig. S1. Dot-plot analyses of Acrocomia aculeata (X-axis) plastome against selected species (Y-axis) within the subfamily Arecoideae. A positive slope denotes that the pair of sequences compared is in the same orientation. A negative slope denotes that the pair of sequences compared can be aligned, but their orientation is opposite. Sequences in the same direction are red and inversions are blue



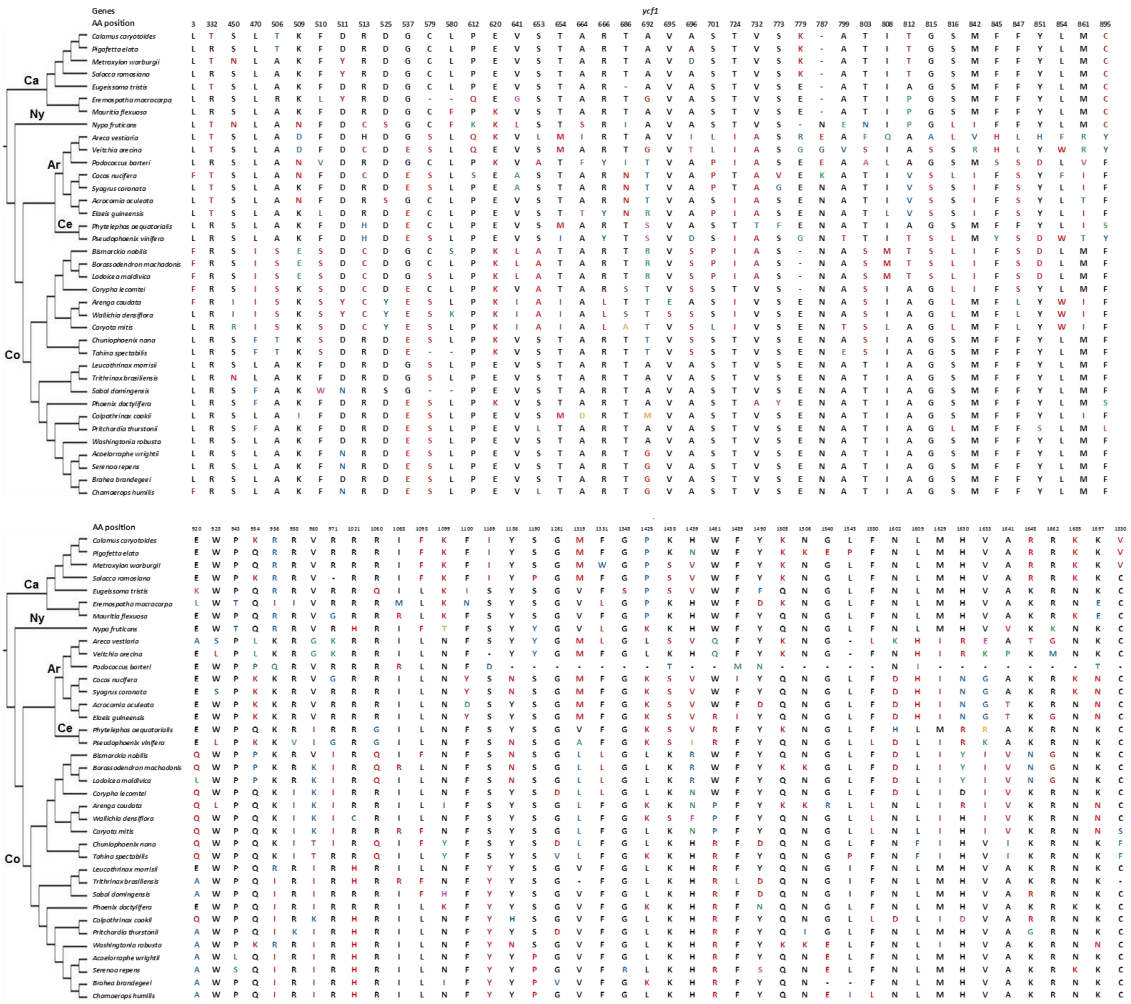
Supplementary Fig. S2. Nucleotide divergence of protein-coding genes in Arecaceae plastomes based on phylogenetic reconstruction. The gene divergence for each species was estimated by the sum of total branch lengths until the common ancestor node. The letters highlighted in red are abbreviations of species: Eugeissoma tristis (Et), Podococcus barteri (Pb), Lodoicea malvidica (Lma), Tahina spectabilis (Ts), Eremospatha macrocarpa (Em), Salacca ramosiana (Sra), Borassodendron machadonis (Bm), Nypa fruticans (Nf), Colpothrinax cookie (Cc), and Elaeis guineensis (Eg)



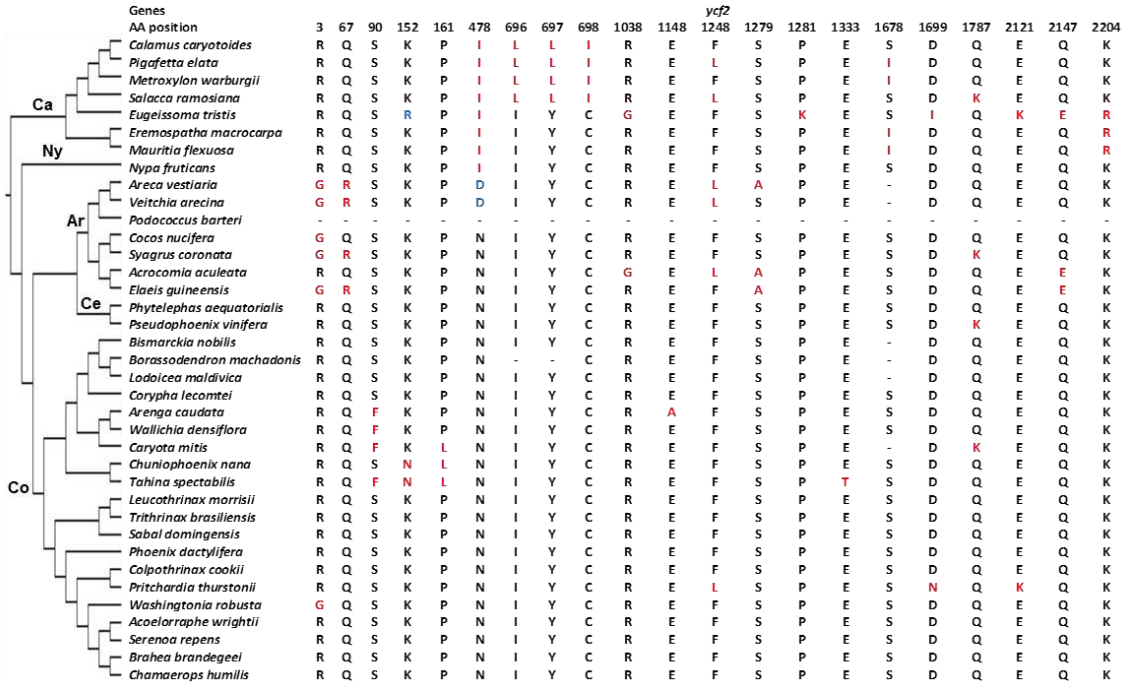
Supplementary Fig. S3. Sites under positive selection across the Arecaceae phylogeny inferred based on whole plastomes. Sites identified in accD, cemA, and matK genes. Different amino acid types identified at the same position are highlighted in distinct colors. The amino acid positions are relative to macaw palm plastid genes



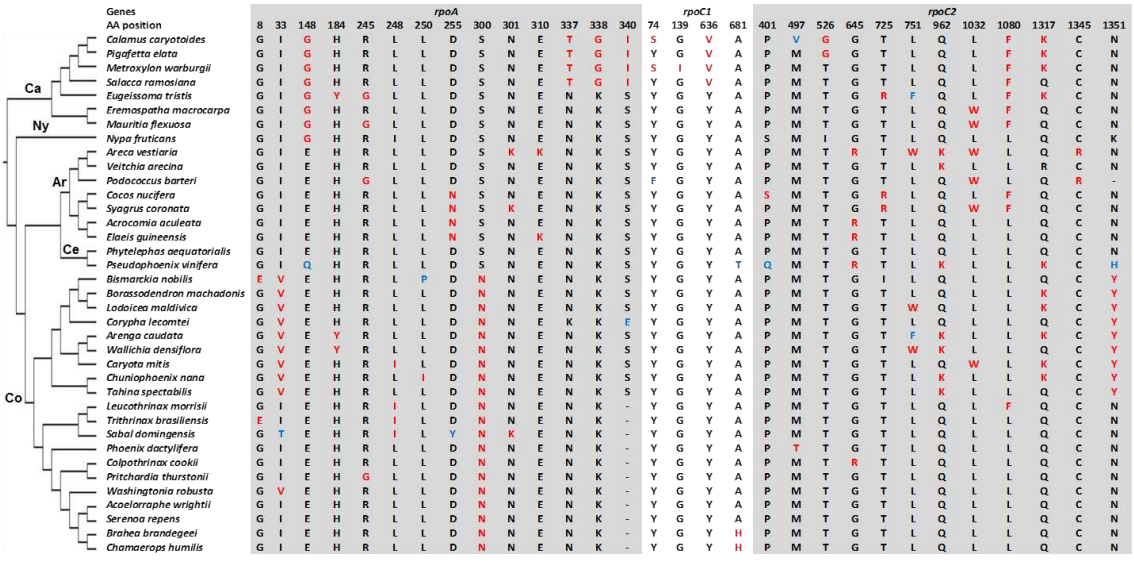
Supplementary Fig. S4 Sites under positive selection across the Arecaceae phylogeny inferred based on whole plastomes. Sites identified in ccsA gene. Different amino acid types identified at the same position are highlighted in distinct colors. The amino acid positions are relative to macaw palm plastid gene



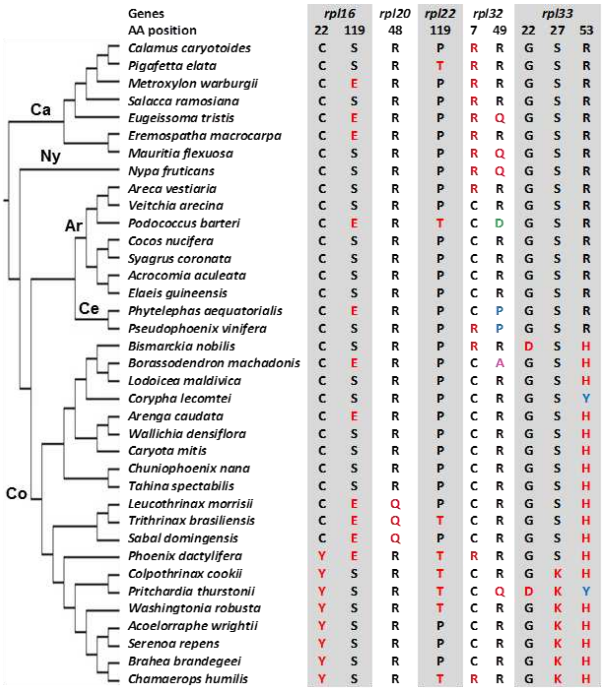
Supplementary Fig. S5. Sites under positive selection across the Arecaceae phylogeny inferred based on whole plastomes. Sites identified in ycf1 gene. Different amino acid types identified at the same position are highlighted in distinct colors. The amino acid positions are relative to macaw palm plastid gene



Supplementary Fig. S6. Sites under positive selection across the Arecaceae phylogeny inferred based on whole plastomes. Sites identified in ycf2 gene. Different amino acid types identified at the same position are highlighted in distinct colors. The amino acid positions are relative to macaw palm plastid gene

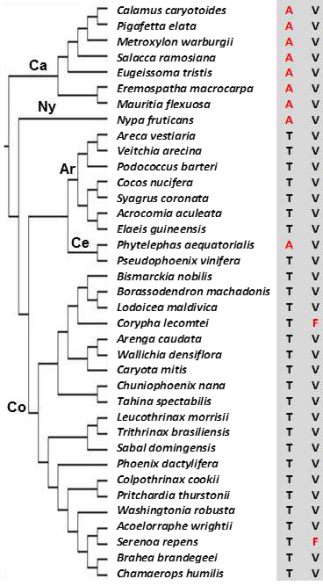


Supplementary Fig. S7. Sites under positive selection across the Arecaceae phylogeny inferred based on whole plastomes. Sites identified in rpoA, rpoC1, and rpoC2 genes. Different amino acid types identified at the same position are highlighted in distinct colors. The amino acid positions are relative to macaw palm plastid genes



Genes	rpl16		rpl20	rpl22	rpl32		rpl33		
AA position	22	119	48	119	7	49	22	27	53
Calamus caryotoides	C	S	R	P	R	R	G	S	R
Pigafetta elata	C	S	R	T	R	R	G	S	R
Metroxylon warburgii	C	E	R	P	R	R	G	S	R
Salacca ramosiana	C	S	R	P	R	R	G	S	R
Eugeissoma tristis	C	E	R	P	R	Q	G	S	R
Eremospatha macrocarpa	C	E	R	P	R	R	G	S	R
Mauritia flexuosa	C	S	R	P	R	Q	G	S	R
Nypa fruticans	C	S	R	P	R	Q	G	S	R
Areca vestiaria	C	S	R	P	R	R	G	S	R
Veitchia arecina	C	S	R	P	C	R	G	S	R
Podococcus barteri	C	E	R	T	C	D	G	S	R
Cocos nucifera	C	S	R	P	C	R	G	S	R
Syagrus coronata	C	S	R	P	C	R	G	S	R
Acrocomia aculeata	C	S	R	P	C	R	G	S	R
Elaeis guineensis	C	S	R	P	C	R	G	S	R
Phytelephas aequatorialis	C	E	R	P	C	P	G	S	R
Pseudophoenix vinifera	C	S	R	P	R	P	G	S	R
Bismarckia nobilis	C	S	R	P	R	R	D	S	H
Borassodendron machadonis	C	E	R	P	C	A	G	S	H
Lodoicea maldivica	C	S	R	P	C	R	G	S	H
Corypha lecomtei	C	S	R	P	C	R	G	S	Y
Arenga caudata	C	E	R	P	C	R	G	S	H
Wallichia densiflora	C	S	R	P	C	R	G	S	H
Caryota mitis	C	S	R	P	C	R	G	S	H
Chuniophoenix nana	C	S	R	P	C	R	G	S	H
Tahina spectabilis	C	S	R	P	C	R	G	S	H
Leucothrinax morrisii	C	E	Q	P	C	R	G	S	H
Trithrinax brasiliensis	C	E	Q	T	C	R	G	S	H
Sabal domingensis	C	E	Q	P	C	R	G	S	H
Phoenix dactylifera	Y	E	R	T	R	R	G	S	H
Colpothrinax cookii	Y	S	R	T	C	R	G	K	H
Pritchardia thurstonii	Y	S	R	T	C	Q	D	K	Y
Washingtonia robusta	Y	S	R	T	C	R	G	K	H
Acoelorraphe wrightii	Y	S	R	P	C	R	G	K	H
Serenoa repens	Y	S	R	P	C	R	G	K	H
Brahea brandegeei	Y	S	R	P	C	R	G	K	H
Chamaerops humilis	Y	S	R	T	R	R	G	K	H

Supplementary Fig. S8. Sites under positive selection across the Arecaceae phylogeny inferred based on whole plastomes. Sites identified in rpl16, rpl20, rpl22, rpl32, and rpl33 genes. Different amino acid types identified at the same position are highlighted in distinct colors. The amino acid positions are relative to macaw palm plastid genes



Genes	rps2							rps3	rps4		rps8						rps11		rps14		rps16					rps18	rps19		
AA position	57	61	91	132	134	181	231	236	218	29	121	23	27	55	64	79	106	75	88	45	89	15	52	82	84	85	94	19	56	58
Calamus caryotoides	A	V	A	C	P	V	G	R	E	R	R	T	A	N	R	H	L	A	L	V	A	A	A	S	R	V	T	I	E	L
Pigafetta elata	A	V	A	C	P	V	G	R	K	G	R	T	A	N	R	H	L	A	L	V	A	A	A	S	R	V	T	I	E	L
Metroxylon warburgii	A	V	A	C	P	V	G	R	K	G	R	T	A	N	R	H	L	A	L	V	A	A	A	S	R	V	T	I	E	L
Salacca ramosiana	A	V	A	C	P	V	G	R	K	G	R	T	A	N	R	H	L	A	L	V	A	A	A	N	D	S	T	M	E	L
Eugeissoma tristis	A	V	A	C	P	V	G	R	K	G	R	T	A	N	R	H	L	A	L	V	A	A	A	S	R	V	T	M	E	L
Eremospatha macrocarpa	A	V	A	C	P	V	G	R	K	G	R	T	A	N	R	H	L	A	L	V	A	V	A	N	D	S	T	I	E	L
Mauritia flexuosa	A	V	A	C	P	V	G	R	K	V	R	T	A	N	R	H	L	A	L	V	A	A	A	S	R	V	T	I	E	L
Nypa fruticans	A	V	A	C	P	V	S	R	K	G	R	T	A	N	R	R	L	T	M	G	A	A	A	S	R	V	T	I	G	L
Areca vestiaria	T	V	A	C	S	V	S	R	F	G	R	T	A	K	R	C	L	A	M	G	A	A	V	S	-	-	T	I	E	L
Veitchia arecina	T	V	A	C	P	V	S	R	F	G	H	T	A	K	R	C	L	A	M	G	T	A	V	S	R	V	T	I	E	L
Podococcus barteri	T	V	A	C	S	V	S	R	F	G	R	T	A	N	R	C	L	T	M	G	A	A	V	-	-	-	T	M	E	L
Cocos nucifera	T	V	A	C	S	V	S	R	F	G	R	T	A	N	R	C	L	A	M	G	A	A	V	S	R	V	T	I	E	L
Syagrus coronata	T	V	A	C	S	V	S	R	K	G	R	T	A	N	R	C	L	A	M	G	A	A	L	S	R	V	T	I	E	L
Acrocomia aculeata	T	V	A	C	S	V	S	R	F	G	R	T	A	N	R	C	L	A	M	G	A	A	V	S	R	V	T	I	E	L
Elaeis guineensis	T	V	A	C	S	V	S	R	F	G	H	T	A	N	R	C	L	A	M	G	A	A	L	S	R	V	T	I	E	L
Phytelephas aequatorialis	A	V	A	C	S	V	S	R	K	G	R	T	A	N	R	C	L	A	M	G	A	A	V	S	R	V	T	I	E	L
Pseudophoenix vinifera	T	V	A	S	S	V	S	R	K	G	R	M	A	N	R	C	L	A	M	G	A	A	V	S	R	V	T	I	E	L
Bismarckia nobilis	T	V	A	C	S	V	S	S	F	G	R	T	A	N	R	C	L	A	L	E	A	A	V	S	R	V	S	I	E	I
Borassodendron machadonis	T	V	A	C	S	V	S	R	F	G	R	T	A	N	R	C	L	A	L	E	A	A	V	S	R	V	S	I	E	L
Lodoicea maldivica	T	V	A	C	S	V	S	R	F	G	R	T	A	N	R	C	L	A	L	E	A	A	V	S	F	V	S	I	E	L
Corypha lecomtei	T	F	A	C	S	V	S	R	F	V	G	T	A	N	R	C	L	A	L	G	A	A	V	S	R	V	S	I	E	L
Arenga caudata	T	V	T	C	S	V	S	-	F	G	R	T	A	N	R	C	L	A	L	G	T	A	V	S	R	G	S	I	E	L
Wallichia densiflora	T	V	T	C	S	V	S	-	F	G	R	T	A	N	L	C	L	A	L	G	T	V	V	S	R	G	S	I	E	L
Caryota mitis	T	V	T	C	S	V	S	R	K	G	R	T	A	N	R	C	L	A	L	G	T	A	V	S	R	G	S	I	E	L
Chuniophoenix nana	T	V	V	S	S	V	S	R	F	G	R	T	A	N	Q	C	I	A	L	G	A	A	V	S	R	A	S	I	V	L
Tahina spectabilis	T	V	A	S	S	V	S	R	F	G	R	T	A	N	Q	C	I	A	L	G	A	A	V	K	K	Q	S	I	E	L
Leucothrinax morrisii	T	V	A	C	S	S	S	H	F	G	R	T	A	K	R	C	L	A	M	G	A	A	V	S	R	V	T	I	E	L
Trithrinax brasiliensis	T	V	A	C	S	A	S	R	F	G	R	T	A	K	R	C	L	A	M	G	A	A	L	S	R	V	T	I	E	L
Sabal domingensis	T	V	A	C	S	V	G	R	F	G	R	T	A	N	R	C	I	A	M	G	A	A	V	S	R	V	T	I	E	L
Phoenix dactylifera	T	V	A	C	S	V	S	R	F	G	R	T	A	N	R	C	L	A	M	G	A	A	L	S	R	V	T	I	E	L
Colpothrinax cookii	T	V	A	C	S	V	S	R	F	G	R	T	T	N	R	C	L	A	M	G	A	A	V	S	R	V	T	I	E	L
Pritchardia thurstonii	T	V	A	C	S	V	T	P	F	G	H	T	K	N	R	C	L	A	M	G	A	V	V	S	R	V	T	I	E	L
Washingtonia robusta	T	V	A	C	S	V	S	R	F	G	R	T	T	N	R	C	L	T	M	G	A	A	L	S	R	V	T	I	E	L
Acoelorraphe wrightii	T	V	A	C	S	V	S	R	F	G	R	R	T	N	R	C	L	A	M	G	A	A	L	S	R	V	T	I	E	F
Serenoa repens	T	F	A	C	S	V	S	R	F	G	R	T	T	N	R	C	L	A	M	G	V	A	L	S	R	V	T	I	E	L
Brahea brandegeei	T	V	A	C	S	V	S	R	F	G	R	T	T	N	R	C	L	A	M	G	A	A	V	S	R	V	T	I	E	L
Chamaerops humilis	T	V	A	C	S	V	S	R	F	G	R	T	T	K	R	C	L	A	M	G	A	A	V	S	R	V	T	I	E	L

Supplementary Fig. S9. Sites under positive selection across the Arecaceae phylogeny inferred based on whole plastomes. Sites identified in rps2, rps3, rps4, rps8, rps11, rps14, rps16, rps18, and rps19 genes. Different amino acid types identified at the same position are highlighted in distinct colors. The amino acid positions are relative to macaw palm plastid genes

Supplementary Tables

Supplementary Table S1. List of species included in the Arecaceae phylogenomic analysis

Species	Subfamily	Family	Order	GenBank
Acrocomia aculeata (Jacq.) Lodd. ex Mart.	Arecoideae	Arecaceae	Arecales	MG020488
Areca vestiaria Giseke	Arecoideae	Arecaceae	Arecales	NC_029972.1
Cocos nucifera L.	Arecoideae	Arecaceae	Arecales	NC_022417.1
Elaeis guineensis Jacq.	Arecoideae	Arecaceae	Arecales	NC_017602.1
Podococcus barteri G.Mann & H.Wendl.	Arecoideae	Arecaceae	Arecales	NC_027276.1
Syagrus coronata (Mart.) Becc.	Arecoideae	Arecaceae	Arecales	NC_029241.1
Veitchia arecina Becc.	Arecoideae	Arecaceae	Arecales	NC_029950.1
Calamus caryotoides A.Cunn. ex Mart.	Calamoideae	Arecaceae	Arecales	NC_020365.1
Eremospatha macrocarpa H.Wendl.	Calamoideae	Arecaceae	Arecales	NC_029964.1
Eugeissona tristes Griff.	Calamoideae	Arecaceae	Arecales	NC_029963.1
Mauritia flexuosa L.f.	Calamoideae	Arecaceae	Arecales	NC_029947.1
Metroxylon warburgii (Heimerl) Becc.	Calamoideae	Arecaceae	Arecales	NC_029959.1
Pigafetta elata (Mart.) H.Wendl.	Calamoideae	Arecaceae	Arecales	NC_029956.1
Salacca ramosiana Mogea	Calamoideae	Arecaceae	Arecales	NC_029954.1
Phytelephas aequatorialis Spruce	Ceroxyloideae	Arecaceae	Arecales	NC_029957.1
Pseudophoenix vinifera (Mart.) Becc.	Ceroxyloideae	Arecaceae	Arecales	NC_020364.1
Acoelorraphe wrightii (Griseb. & H.Wendl.) H.Wendl. ex Becc.	Coryphoideae	Arecaceae	Arecales	NC_029973.1
Arenga caudata (Lour.) H.E.Moore	Coryphoideae	Arecaceae	Arecales	NC_029971.1
Bismarckia nobilis Hildebr. & H.Wendl.	Coryphoideae	Arecaceae	Arecales	NC_020366.1
Borassodendron machadonis (Ridl.) Becc.	Coryphoideae	Arecaceae	Arecales	NC_029969.1
Brahea brandegeei (Purpus) H.E.Moore	Coryphoideae	Arecaceae	Arecales	NC_029968.1
Caryota mitis Lour.	Coryphoideae	Arecaceae	Arecales	NC_029948.1
Chamaerops humilis L.	Coryphoideae	Arecaceae	Arecales	NC_029967.1
Chuniophoenix nana Burret	Coryphoideae	Arecaceae	Arecales	NC_029966.1
Colpothrinax cookii Read	Coryphoideae	Arecaceae	Arecales	NC_028026.1
Corypha lecomtei Becc. ex Lecomte	Coryphoideae	Arecaceae	Arecales	NC_029965.1
Leucothrinax morrisii (H.Wendl.) C.Lewis & Zona	Coryphoideae	Arecaceae	Arecales	NC_029961.1
Lodoicea maldivica (J.F.Gmel.) Pers.	Coryphoideae	Arecaceae	Arecales	NC_029960.1
Phoenix dactylifera L.	Coryphoideae	Arecaceae	Arecales	NC_013991.2
Pritchardia thurstonii (F.Muell.) Drude	Coryphoideae	Arecaceae	Arecales	NC_029955.1
Sabal domingensis Becc.	Coryphoideae	Arecaceae	Arecales	NC_026444.1
Serenoa repens (W.Bartram) Small	Coryphoideae	Arecaceae	Arecales	NC_029953.1
Tahina spectabilis J.Dransf. & Rakotoarin.	Coryphoideae	Arecaceae	Arecales	NC_029952.1
Trithrinax brasiliensis Mart.	Coryphoideae	Arecaceae	Arecales	NC_029951.1
Wallichia densiflora Mart.	Coryphoideae	Arecaceae	Arecales	NC_029949.1
Washingtonia robusta H.Wendl.	Coryphoideae	Arecaceae	Arecales	NC_029974.1
Nypa fruticans Wurmb	Nypoideae	Arecaceae	Arecales	NC_029958.1
Baxteria australis R. Br. ex Hook.*	-	Dasypogonaceae	Arecales	NC_029970.1
Dasypogon bromeliifolius R. Br.*	-	Dasypogonaceae	Arecales	NC_020367.1
Hanguana malayana (Jack) Merr.*	-	Hanguanaceae	Commelinales	NC_029962.1

*outgroup

Supplementary Table S2. List of substitution models for each plastid gene selected using jModelTest v.2.1.7

Gene	Model	Gene	Model	Gene	Model	Gene	Model
accD	GTR+I+G	ndhI	HKY+G	psbH	K80	rpoB	GTR+G
atpA	GTR+G	ndhJ	GTR+G	psbI	K80	rpoC1	GTR+G
tpB	GTR+I	ndhK	HKY+I	psbJ	JC	rpoC2	GTR+I+G
atpE	HKY	petA	HKY+I	psbK	HKY	rps2	HKY+I
atpF	GTR+I	petB	HKY+I	psbL	JC	rps3	HKY+G
atpH	K80+I	petD	HKY+G	psbM	JC	rps4	HKY+I
atpI	GTR+I	petG	JC	psbN	K80	rps7	HKY
ccsA	GTR+G	petL	JC	psbT	JC	rps8	GTR+I
cemA	GTR+G	petN	JC	psbZ	HKY	rps11	HKY+I+G
clpP	HKY+G	psaA	GTR+G	rbcL	GTR+I+G	rps12	HKY
infA	HKY+I	psaB	GTR+I+G	rpl2	HKY	rps14	HKY
matK	GTR+G	psaC	K80+I	rpl14	HKY+G	rps15	HKY+I
ndhA	GTR+I+G	psaI	JC	rpl16	HKY+G	rps16	HKY+G
ndhB	HKY+I	psaJ	K80+I	rpl20	HKY	rps18	HKY+G
ndhC	HKY	psbA	GTR+I+G	rpl22	GTR+I	rps19	F81+G
ndhD	GTR+I+G	psbB	GTR+G	rpl23	F81+I	ycf1	GTR+G
ndhE	HKY+G	psbC	GTR+I+G	rpl32	HKY+G	ycf2	GTR+I+G
ndhF	GTR+I+G	psbD	GTR+I	rpl33	HKY	ycf3	HKY+I+G
ndhG	GTR+G	psbE	HKY+I	rpl36	K80	ycf4	GTR+G
ndhH	GTR+I+G	psbF	JC	rpoA	GTR+G		

Supplementary Table S3. Location of SSRs in the plastome of Acrocomia aculeata (IRB was omitted). The polymorphic loci were identified upon comparison with the SSRs located in the plastome of Elaeis guineensis

SSR type	Sequence	Size	Start	End	Location	Polymorphic locus (P)
di	(AT)4	8	1332	1339	psbA/trnK-UUU (IGS)	
mono	(A)9	9	2649	2657	matK (CDS)	
mono	(T)9	9	2895	2903	matK (CDS)	
di	(AT)4	8	3391	3398	trnK-UUU (intron)	
mono	(A)9	9	3619	3627	trnK-UUU (intron)	P
mono	(T)12	12	3819	3830	trnK-UUU (intron)	P
mono	(A)8	8	3844	3851	trnK-UUU (intron)	
mono	(A)8	8	3931	3938	trnK-UUU (intron)	
mono	(A)10	10	4592	4601	trnK-UUU/rps16 (IGS)	P
mono	(C)11	11	4795	4805	trnK-UUU/rps16 (IGS)	P
di	(GT)4	8	6031	6038	rps16/trnQ-UUG (IGS)	
tetra	(TCTA)5	20	6065	6084	rps16/trnQ-UUG (IGS)	P
mono	(T)10	10	6565	6574	rps16/trnQ-UUG (IGS)	P
mono	(A)11	11	6885	6895	rps16/trnQ-UUG (IGS)	P
mono	(T)9	9	7362	7370	trnQ-UUG/psbK (IGS)	
mono	(A)9	9	7386	7394	trnQ-UUG/psbK (IGS)	
mono	(T)10	10	7415	7424	trnQ-UUG/psbK (IGS)	P
mono	(A)8	8	7432	7439	trnQ-UUG/psbK (IGS)	P
mono	(T)8	8	7622	7629	psbK (CDS)	
mono	(A)8	8	8047	8054	psbK/psbI (IGS)	
mono	(A)8	8	8182	8189	psbI/trnS-GCU (IGS)	
mono	(T)9	9	8221	8229	psbI/trnS-GCU (IGS)	P
di	(GA)4	8	8326	8333	trnS-GCU	
hexa	(TCCCCA)3	18	8409	8426	trnS-GCU/trnG-UCC (IGS)	
di	(TA)8	16	8580	8595	trnS-GCU/trnG-UCC (IGS)	P
di	(AT)4	8	8598	8605	trnS-GCU/trnG-UCC (IGS)	
mono	(T)8	8	9618	9625	trnG-UCC (intron)	P

mono	(T)9	9	11903	11911	atpA/atpF (IGS)	P
mono	(T)10	10	12517	12526	atpF (intron)	
mono	(T)8	8	12727	12734	atpF (intron)	
mono	(T)9	9	12769	12777	atpF (intron)	P
mono	(A)8	8	13205	13212	atpF (CDS)	
mono	(T)11	11	13444	13454	atpF/atpH (IGS)	P
mono	(A)9	9	14130	14138	atpH/atpI (IGS)	
mono	(T)9	9	14192	14200	atpH/atpI (IGS)	P
di	(AT)5	10	14559	14568	atpH/atpI (IGS)	P
di	(AT)9	18	14582	14599	atpH/atpI (IGS)	P
mono	(T)9	9	14784	14792	atpH/atpI (IGS)	
mono	(A)10	10	15874	15883	atpI/rps2 (IGS)	P
mono	(T)9	9	16678	16686	rps2/rpoC2 (IGS)	P
mono	(A)8	8	18041	18048	rpoC2 (CDS)	
mono	(T)10	10	18754	18763	rpoC2 (CDS)	
mono	(T)11	11	18860	18870	rpoC2 (CDS)	
mono	(A)8	8	19003	19010	rpoC2 (CDS)	
di	(AT)4	8	20134	20141	rpoC2 (CDS)	
di	(AT)5	10	20224	20233	rpoC2 (CDS)	
mono	(A)8	8	22676	22683	rpoC1 (CDS)	
di	(TA)4	8	23237	23244	rpoC1 (intron)	
di	(TC)4	8	23264	23271	rpoC1 (intron)	
mono	(T)9	9	23308	23316	rpoC1 (intron)	P
mono	(T)8	8	26559	26566	rpoB (CDS)	
penta	(ATGTA)3	15	27334	27348	rpoB/trnC-GCA (IGS)	P
mono	(A)8	8	27456	27463	rpoB/trnC-GCA (IGS)	
mono	(A)9	9	28358	28366	rpoB/trnC-GCA (IGS)	P
mono	(A)8	8	28785	28792	trnC-GCA/petN (IGS)	
mono	(A)8	8	29013	29020	trnC-GCA/petN (IGS)	P
di	(GT)4	8	29022	29029	trnC-GCA/petN (IGS)	
mono	(A)10	10	29401	29410	petN/psbM (IGS)	
mono	(T)9	9	29436	29444	petN/psbM (IGS)	P
di	(TA)4	8	29451	29458	petN/psbM (IGS)	
mono	(A)9	9	29604	29612	petN/psbM (IGS)	P
di	(TA)4	8	29689	29696	petN/psbM (IGS)	
mono	(T)8	8	30217	30224	psbM/trnD-GUC (IGS)	P
mono	(T)8	8	31036	31043	trnD-GCU/trnY-GUA (IGS)	P
mono	(T)8	8	31076	31083	trnD-GCU/trnY-GUA (IGS)	P
mono	(T)8	8	31884	31891	trnT-GGU/psbD (IGS)	P
mono	(T)8	8	32232	32239	trnT-GGU/psbD (IGS)	
mono	(T)11	11	32520	32530	trnT-GGU/psbD (IGS)	P
di	(TA)4	8	32620	32627	trnT-GGU/psbD (IGS)	
mono	(A)8	8	32688	32695	trnT-GGU/psbD (IGS)	
mono	(G)8	8	34144	34151	psbC (CDS)	
mono	(G)8	8	34415	34422	psbC (CDS)	
di	(GA)4	8	35403	35410	trnS-UGA	
mono	(A)8	8	36116	36123	psbZ/trnG-GCC (IGS)	
mono	(A)8	8	36503	36510	trnG-GCC/trnfM-CAU (IGS)	P
mono	(A)9	9	36672	36680	trnfM-CAU/rps14 (IGS)	
mono	(T)11	11	37022	37032	rps14 (CDS)	
mono	(A)8	8	37247	37254	rps14/psaB (IGS)	P
mono	(C)10	10	40214	40223	psaA (CDS)	
di	(AG)4	8	41116	41123	psaA (CDS)	
penta	(TATTT)3	15	41868	41882	psaA/ycf3 (IGS)	
mono	(A)9	9	42901	42909	ycf3 (intron)	P
mono	(A)9	9	45217	45225	trnS-GGA/rps4 (IGS)	
mono	(G)9	9	45485	45493	rps4 (CDS)	
di	(TA)4	8	46112	46119	rps4/trnT-UGU (IGS)	
mono	(A)8	8	46275	46282	rps4/trnT-UGU (IGS)	
di	(AT)6	12	46935	46946	trnL-UAA (intron)	P
di	(AG)4	8	47080	47087	trnL-UAA (intron)	
di	(AT)4	8	47661	47668	trnF-GAA/ndhJ (IGS)	

mono	(T)8	8	47790	47797	trnF-GAA/ndhJ (IGS)	P
di	(AT)8	16	47865	47880	trnF-GAA/ndhJ (IGS)	P
mono	(A)9	9	47882	47890	trnF-GAA/ndhJ (IGS)	
mono	(T)8	8	50220	50227	ndhC/trnV-UAC (IGS)	
mono	(A)8	8	50356	50363	ndhC/trnV-UAC (IGS)	P
mono	(A)10	10	50397	50406	ndhC/trnV-UAC (IGS)	P
di	(AT)4	8	50565	50572	ndhC/trnV-UAC (IGS)	
di	(AT)4	8	50577	50584	ndhC/trnV-UAC (IGS)	
mono	(A)10	10	50654	50663	ndhC/trnV-UAC (IGS)	P
mono	(T)9	9	50740	50748	ndhC/trnV-UAC (IGS)	
mono	(A)8	8	50944	50951	ndhC/trnV-UAC (IGS)	
mono	(A)9	9	51038	51046	ndhC/trnV-UAC (IGS)	
di	(CA)4	8	51474	51481	ndhC/trnV-UAC (IGS)	
mono	(T)9	9	52828	52836	trnM-CAU/atpE (IGS)	P
mono	(T)8	8	54820	54827	trnM-CAU/atpE (IGS)	
mono	(T)8	8	55155	55162	atpB/rbcL (IGS)	P
di	(AT)4	8	55360	55367	atpB/rbcL (IGS)	
mono	(A)9	9	57103	57111	rbcL/accD (IGS)	
mono	(C)8	8	57632	57639	rbcL/accD (IGS)	
mono	(A)8	16	57640	57647	rbcL/accD (IGS)	P
mono	(T)8	8	58166	58173	accD (CDS)	
mono	(A)10	10	58469	58478	accD (CDS)	
di	(TG)4	8	58898	58905	accD (CDS)	
mono	(A)9	9	59354	59362	accD/psaI (IGS)	P
mono	(A)9	9	59562	59570	accD/psaI (IGS)	
mono	(A)8	8	59605	59612	accD/psaI (IGS)	
di	(TA)4	8	59701	59708	accD/psaI (IGS)	
di	(TA)4	8	59727	59734	accD/psaI (IGS)	P
mono	(T)8	8	60569	60576	ycf4 (CDS)	
mono	(A)8	8	60973	60980	ycf4/cemA (IGS)	P
mono	(A)11	11	61168	61178	cemA (CDS)	P
di	(TC)5	10	61235	61244	cemA (CDS)	
tetra	(AATG)3	12	61845	61856	cemA (CDS)	
mono	(C)8	8	62441	62448	petA (CDS)	
mono	(A)8	8	63571	63578	petA/psbJ (IGS)	
mono	(T)9	9	65219	65227	psbE/petL (IGS)	P
mono	(A)10	10	65326	65335	psbE/petL (IGS)	
mono	(T)8	8	66135	66142	petG/trnW-CCA (IGS)	
mono	(T)8	8	66190	66197	petG/trnW-CCA (IGS)	
mono	(A)8	8	66697	66704	trnP-UGG/psaJ (IGS)	
mono	(A)8	8	66734	66741	trnP-UGG/psaJ (IGS)	P
mono	(A)9	9	66874	66882	trnP-UGG/psaJ (IGS)	P
mono	(T)8	8	67071	67078	psaJ (CDS)	
mono	(T)9	9	67161	67169	psaJ/rpl33 (IGS)	
di	(AT)7	14	67986	67999	rpl33/rps18 (IGS)	
mono	(T)9	9	68552	68560	rps18/rpl20 (IGS)	
mono	(T)8	8	68603	68610	rps18/rpl20 (IGS)	P
mono	(A)9	9	69258	69266	rpl20/rps12 (IGS)	
mono	(T)10	10	69295	69304	rpl20/rps12 (IGS)	
di	(TA)4	8	69884	69891	rps12/clpP (IGS)	
mono	(T)8	8	69954	69961	rps12/clpP (IGS)	P
mono	(A)9	9	70328	70336	clpP (intron)	P
mono	(A)8	8	70511	70518	clpP (intron)	
mono	(T)10	10	70617	70626	clpP (intron)	
mono	(T)10	10	70881	70890	clpP (intron)	P
tetra	(ATAA)3	12	71079	71090	clpP (CDS)	
mono	(A)9	9	71267	71275	clpP (intron)	
mono	(T)9	9	71318	71326	clpP (intron)	P
mono	(A)9	9	71416	71424	clpP (intron)	P
mono	(A)8	8	72102	72109	clpP (intron)	P
mono	(T)8	8	73282	73289	psbB (CDS)	
mono	(T)9	9	74107	74115	psbB/psbT (IGS)	P

mono	(A)9	9	75235	75243	petB (intron)	P
tetra	(AAAT)3	12	75549	75560	petB (intron)	
mono	(T)9	9	78051	78059	rpoA/rps11 (IGS)	P
mono	(T)10	10	79930	79939	rpl36/infA (IGS)	P
mono	(T)13	13	80804	80816	rps8/rpl14 (IGS)	P
di	(TA)4	8	81981	81988	rpl16 (intron)	
mono	(T)10	10	82502	82511	rpl16 (intron)	
mono	(T)8	8	82513	82520	rpl16 (intron)	P
tetra	(TTTA)3	12	82525	82536	rpl16 (intron)	
mono	(T)15	15	82874	82888	rpl16 (intron)	P
penta	(TTTTA)3	15	84297	84311	rpl22/rps19 (IGS)	
mono	(T)8	8	84313	84320	rpl22/rps19 (IGS)	P
mono	(T)9	9	84641	84649	rps19/trnH-GUG (IGS)	
mono	(T)8	8	84677	84684	rps19/trnH-GUG (IGS)	
mono	(A)8	8	84892	84899	trnH-GUG/rpl2 (IGS)	
di	(GA)4	8	87041	87048	ycf2 (CDS)	
di	(GA)4	8	87053	87060	ycf2 (CDS)	
di	(GA)4	8	88055	88062	ycf2 (CDS)	
mono	(A)8	8	88764	88771	ycf2 (CDS)	
mono	(A)8	8	88957	88964	ycf2 (CDS)	
mono	(A)9	9	90225	90233	ycf2 (CDS)	
di	(TA)4	8	93605	93612	ycf2 (CDS)	
di	(TA)4	8	95011	95018	trnL-CAA/ndhB (IGS)	
di	(AG)4	8	95752	95759	ndhB (CDS)	
mono	(T)8	8	96738	96745	ndhB (intron)	P
mono	(T)9	9	99692	99700	rps12/trnV-GAC (IGS)	
mono	(A)8	8	99859	99866	rps12/trnV-GAC (IGS)	
mono	(T)8	8	103725	103732	trnI-GAU (intron)	
di	(CT)4	8	107047	107054	rrn23	
hexa	(CTTTTT)3	18	109158	109175	trnR-ACG/trnN-GUU (IGS)	P
mono	(T)8	8	111190	111197	ycf1 (CDS)	
mono	(A)8	8	111348	111355	ycf1 (CDS)	
mono	(A)8	8	111376	111383	ndhF (CDS)	
mono	(C)10	10	111410	111419	ndhF (CDS)	
mono	(T)9	9	113561	113569	ndhF/rpl32 (IGS)	
mono	(A)8	8	113643	113650	ndhF/rpl32 (IGS)	P
di	(AT)5	10	113661	113670	ndhF/rpl32 (IGS)	P
di	(AT)6	12	113673	113684	ndhF/rpl32 (IGS)	P
mono	(A)10	10	113734	113743	ndhF/rpl32 (IGS)	P
mono	(A)8	8	113778	113785	ndhF/rpl32 (IGS)	
mono	(T)9	9	114275	114283	rpl32/trnL-UAG (IGS)	P
mono	(T)8	8	114334	114341	rpl32/trnL-UAG (IGS)	P
mono	(A)10	10	114346	114355	rpl32/trnL-UAG (IGS)	P
mono	(T)8	8	115301	115308	ccsA (CDS)	
tetra	(AATA)3	12	116178	116189	ndhD (CDS)	
di	(AT)6	12	118164	118175	psaC/ndhE (IGS)	
tri	(AAT)4	12	118239	118250	psaC/ndhE (IGS)	
tetra	(TTTA)3	12	118937	118948	ndhE/ndhG (IGS)	
mono	(T)9	9	119660	119668	ndhG/ndhI (IGS)	P
mono	(T)9	9	120134	120142	ndhI (CDS)	
di	(AT)5	10	121337	121346	ndhA (intron)	
tetra	(ATTC)3	12	121447	121458	ndhA (intron)	
di	(TC)5	10	123399	123408	ndhH (CDS)	
mono	(A)9	9	123897	123905	ndhH/rps15 (IGS)	P
di	(AT)4	8	124610	124617	ycf1 (CDS)	
mono	(T)8	8	125004	125011	ycf1 (CDS)	
mono	(T)11	11	125426	125436	ycf1 (CDS)	P
mono	(T)10	10	125549	125558	ycf1 (CDS)	P
mono	(A)8	8	125755	125762	ycf1 (CDS)	
tri	(ATA)4	12	126526	126537	ycf1 (CDS)	P
mono	(T)8	8	126675	126682	ycf1 (CDS)	
mono	(T)9	9	126954	126962	ycf1 (CDS)	

mono	(T)9	9	127167	127175	ycf1 (CDS)	
mono	(T)10	10	127209	127218	ycf1 (CDS)	
mono	(T)12	12	127330	127341	ycf1 (CDS)	P
mono	(T)9	9	127702	127710	ycf1 (CDS)	
mono	(T)11	11	127718	127728	ycf1 (CDS)	
di	(TC)4	8	127778	127785	ycf1 (CDS)	
mono	(A)10	10	127836	127845	ycf1 (CDS)	
mono	(T)8	8	128683	128690	ycf1 (CDS)	

Supplementary Table S4. List of RNA editing sites predicted by PREP program in Arecoideae plastomes

Gene	AA position	*A. aculeata*	*E. guineensis*	*S. coronata*	*C. nucifera*	*P. barteri*	*V. arecina*	*A. vestiaria*	Huang et al. (2013)*
accD	52	CGG (R) => UGG (W)	CGG (R) => UGG (W)	CGG (R) => UGG (W)	CGG (R) => UGG (W)	CGG (R) => UGG (W)	CGG (R) => UGG (W)	CGG (R) => UGG (W)	-
	241	UUU (F)	UUU (F)	CUU (L) => UUU (F)	UUU (F)	UUU (F)	UGU (C)	UUU (F)	
	267	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	+
	388	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+
	389	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	UAU (Y)	UAU (Y)	UAU (Y)	-
	429	UAC (Y)	UAC (Y)	UAC (Y)	UAC (Y)	UAC (Y)	CAC (H) => UAC (Y)	CAU (H) => UAU (Y)	
	470	CCU (P) => CUU (L)	CCU (P) => CUU (L)	CCU (P) => CUU (L)	CCU (P) => CUU (L)	CCU (P) => CUU (L)	CCU (P) => CUU (L)	CCU (P) => CUU (L)	-
	487	UUG (L)	UCG (S) => UUG (L)	UUG (L)	UUG (L)	UUG (L)	UUG (L)	UUG (L)	
atpA	305	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+
	383	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+/-
atpB	395	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	pseudo*	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+
atpF	31	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	+/-
atpI	143	CCC (P) => CUC (L)	CCC (P) => CUC (L)	CCC (P) => CUC (L)	CCC (P) => CUC (L)	CCC (P) => CUC (L)	CCC (P) => CUC (L)	CCC (P) => CUC (L)	+
	210	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+
ccsA	218	ACU (T) => AUU (I)	ACU (T) => AUU (I)	ACU (T) => AUU (I)	ACU (T) => AUU (I)	ACU (T) => AUU (I)	ACU (T) => AUU (I)	ACU (T) => AUU (I)	-
	274	UUA (L)	UUA (L)	UUA (L)	UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	
clpP	28	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	+
	118	UUC (F)	CUC (L) => UUC (F)	UUC (F)	UUC (F)	UUC (F)	UUC (F)	UUC (F)	
	187	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	+
matK	63	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	-
	219	ACA (T)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	-
	310	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAC (H) => UAC (Y)	CAC (H) => UAC (Y)	CAC (H) => UAC (Y)	-
	312	GCC (A) => GUC (V)	GUC (V)	GUC (V)	GUC (V)	GUC (V)	GUC (V)	GUC (V)	
	426	CAC (H) => UAC (Y)	CAC (H) => UAC (Y)	CAC (H) => UAC (Y)	CAC (H) => UAC (Y)	CAC (H) => UAC (Y)	CAC (H) => UAC (Y)	CAU (H) => UAU (Y)	+

ndhA	17	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	+
	159	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+
	189	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+
	358	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	-
ndhB	50	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+/-
	156	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	+
	181	ACG (T) => AUG (M)	ACG (T) => AUG (M)	ACG (T) => AUG (M)	ACG (T) => AUG (M)	ACG (T) => AUG (M)	ACG (T) => AUG (M)	ACG (T) => AUG (M)	+
	196	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	+
	235	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	+
	246	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	+
	277	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+
	279	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+
	371	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+
	398	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+
	419	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	+
	494	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	+/-
ndhD	1	ACG (T) => AUG (M)	ACG (T) => AUG (M)	ACG (T) => AUG (M)	ACG (T) => AUG (M)	ACG (T) => AUG (M)	ACG (T) => AUG (M)	ACG (T) => AUG (M)	+/-
	20	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+
	128	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+
	225	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCA (S) => UUA (L)	UCG (S) => UUG (L)	+
	316	ACA (T) => AUA (I)	ACA (T) => AUA (I)	ACA (T) => AUA (I)	ACA (T) => AUA (I)	ACA (T) => AUA (I)	ACA (T) => AUA (I)	ACA (T) => AUA (I)	+
	398	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	+
	437	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+
ndhF	21	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+/-
	97	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	CCA (P) => CUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+/-
	131	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UUC (F)	UUC (F)	+
	148	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	UAU (Y)	CAU (H) => UAU (Y)	+
	196	CUU (L) => UUU (F)	CUU (L) => UUU (F)	CUU (L) => UUU (F)	CUU (L) => UUU (F)	CUU (L) => UUU (F)	CUU (L) => UUU (F)	CUU (L) => UUU (F)	-
	465	CAC (H) => UAC (Y)	CAC (H) => UAC (Y)	CAC (H) => UAC (Y)	CAC (H) => UAC (Y)	CAC (H) => UAC (Y)	CAC (H) => UAC (Y)	CAC (H) => UAC (Y)	-
	607	UUG (L)	UUU (F)	UUC (F)	UUC (F)	CUC (L) => UUC (F)	UUC (F)	UUC (F)	
	698	UCU (S) => UUU (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	-
ndhG	105	ACA (T) => AUA (I)	ACA (T) => AUA (I)	ACA (T) => AUA (I)	ACA (T) => AUA (I)	ACA (T) => AUA (I)	ACA (T) => AUA (I)	ACA (T) => AUA (I)	-

petB	140	CGG (R) => UGG (W)	CGG (R) => UGG (W)	CGG (R) => UGG (W)	CGG (R) => UGG (W)	CGG (R) => UGG (W)	CGG (R) => UGG (W)	CGG (R) => UGG (W)	+
	204	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	+
psaI	27	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	+
rpl2	1	ACG (T) => AUG (M)	ACG (T) => AUG (M)	ACG (T) => AUG (M)	ACG (T) => AUG (M)	ACG (T) => AUG (M)	ACG (T) => AUG (M)	ACG (T) => AUG (M)	-
rpl20	9	ACA (T) => AUA (I)	ACA (T) => AUA (I)	ACA (T) => AUA (I)	ACA (T) => AUA (I)	ACA (T) => AUA (I)	ACA (T) => AUA (I)	ACA (T) => AUA (I)	-
	103	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	-
rpoA	123	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+
	277	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+
	296	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	-
rpoB	156	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	+/-
	182	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+
	187	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	+
	206	CCG (P) => CUG (L)	CCG (P) => CUG (L)	CCG (P) => CUG (L)	CCG (P) => CUG (L)	CCG (P) => CUG (L)	CCG (P) => CUG (L)	CCG (P) => CUG (L)	+/-
	665	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	+
	807	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+/-
rpoC1	14	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCG (P) => CUG (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	+
	206	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+
	555	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	-
rpoC2	461	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	-
	759	CGG (R) => UGG (W)	CGG (R) => UGG (W)	CGG (R) => UGG (W)	CGG (R) => UGG (W)	CGG (R) => UGG (W)	CGG (R) => UGG (W)	CGG (R) => UGG (W)	-
	770	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	+
rps2	83	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+
rps8	61	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+
rps14	27	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+
	50	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	+
ycf3	15	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	+
	64	ACG (T) => AUG (M)	ACG (T) => AUG (M)	ACG (T) => AUG (M)	ACG (T) => AUG (M)	ACG (T) => AUG (M)	ACG (T) => AUG (M)	ACG (T) => AUG (M)	+
	66	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	+
	138	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	+

(*) Huang et al (2013) validated by RT-PCR the RNA editing sites in plastid genes of Cocos nucifera
(+) Presence of RNA editing in all transcripts
(+/-) Presence of RNA editing in part of the transcripts
(-) Absence of RNA editing in all transcripts

Supplementary Table S5. List of RNA editing sites predicted in Arecoideae plastomes by comparison with validated sites in Cocos nucifera (Huang et al. 2013) that were not predicted by PREP program

Gene	AA position	*A. aculeata*	*E. guineensis*	*S. coronata*	*C. nucifera*	*P. barteri*	*V. arecina*	*A. vestiaria*	Huang et al. (2013)
ndhA	321	CCU (P) => UCU (S)	CCU (P) => UCU (S)	CCU (P) => UCU (S)	CCU (P) => UCU (S)	CCU (P) => UCU (S)	CCU (P) => UCU (S)	CCU (P) => UCU (S)	+
ndhG	116	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	CCA (P) => CUA (L)	+
ndhH	169	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	+
	182	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UUU (F)	+/-
ndhK	44	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	UCG (S) => UUG (L)	+
psaI	29	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	+
rpl23	24	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	+/-
	30	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+/-
rpoA	67	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	UCU (S) => UUU (F)	+
	176	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	UCC (S) => UUC (F)	+
rpoC1	171	CGG (R) => UGG (W)	CGG (R) => UGG (W)	CGG (R) => UGG (W)	CGG (R) => UGG (W)	CGG (R) => UGG (W)	CGG (R) => UGG (W)	CGG (R) => UGG (W)	+
rps2	45	ACA (T) => AUA (I)	ACA (T) => AUA (I)	ACA (T) => AUA (I)	ACA (T) => AUA (I)	ACA (T) => AUA (I)	ACA (T) => AUA (I)	ACA (T) => AUA (I)	+
rps3	157	ACA (T) => AUA (I)	ACA (T) => AUA (I)	ACA (T) => AUA (I)	ACA (T) => AUA (I)	ACA (T) => AUA (I)	ACA (T) => AUA (I)	ACA (T) => AUA (I)	+/-
	195	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	CAU (H) => UAU (Y)	+
ycf4	85	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	UCA (S) => UUA (L)	+/-

(+) Presence of RNA editing in all transcripts
(+/-) Presence of RNA editing in part of the transcripts

The plastomes of two Amazonian oil palms, Astrocaryum aculeatum G. Mey and A. murumuru Mart., reveal a specific structural rearrangement, events of gain of RNA editing sites, molecular markers and non-synonymous substitutions

Amanda de Santana Lopes[1], Túlio Gomes Pacheco[1], Odyone Nascimento da Silva[1], Gélia Dinah Monteiro Viana[1], Leonardo Magalhães Cruz[2], Eduardo Balsanelli[2], Emanuel Maltempi de Souza[2], Fábio de Oliveira Pedrosa[2], Marcelo Rogalski[1*]

[1] Laboratório de Fisiologia Molecular de Plantas, Departamento de Biologia Vegetal, Universidade Federal de Viçosa, Viçosa-MG, Brazil.
[2] Departamento de Bioquímica e Biologia Molecular, Núcleo de Fixação Biológica de Nitrogênio, Universidade Federal do Paraná, Curitiba-PR, Brazil.

*****Corresponding author
E-mail address: rogalski@ufv.br

Manuscript in preparation to be submitted to **Tree Genetics & Genomes**

Abstract

Astrocaryum murumuru Mart. and A. aculeatum G.Mey. are source of food, oil for industry and row material useful for small Amazonian communities. Genetic studies aiming to establish strategies for conservation and domestication of both species are still in the begining given that the exploitation is mostly by extractivist activity. Therefore, the development of molecular markers is essential to evaluate natural populations, germplasm collection and the selection of elite genotypes. The platomes are great source of molecular markers due to their nonrecombinant and uniparentally inheritance constituting efficient tools for genetic and evolutionary analyses. Here, we present the complete sequence and a full characterization of A. murumuru and A. aculeatum plastomes. Additionally, we carried out structural analysis, RNA editing prediction, and synonymous/non-synonymous substitutions mapping to identify evolutionary traits within the genus Astrocaryum. Moreover, we identify 27 polymorphic SSR loci, 120 SNPs, 35 indels, and six hotspots of nucleotide diversity comparing both species. From the total of 120 SNPs mapped, 54 are located in coding sequences (CDSs), resulting 18 synonymous and 25 non-synonymous substitutions. The non-synonymous substitutions affect the sequence of genes involved in several essential plastid functions such as photosynthesis and gene expression machinery. We also report a gain of RNA editing in the ccsA gene exclusively in A. murumuru. Structural analysis shows that the plastomes of both species present a 4.6-kb inversion between the genes trnT-UGU and trnV-UAC, enwrapping a set of genes involved in chlororespiration and plastid translation. Comparison with other palm species indicates that this 4.6-kb inversion is a lineage-specific structural feature of the genus Astrocaryum originated from a flip-flop recombination. Furthermore, our phylogenetic analysis using whole plastomes of 39 Arecaceae species placed the Astrocaryum species sister to Acrocomia within the Cocoseae tribe. Finally, our results indicate that substantial changes have been occurred in the plastome structure and sequence within Astrocaryum, providing several molecular markers to genetic and evolutionary studies within Arecaceae and their relationship with environmental adaptation.

Keywords: Palm tree, Plastid genome, Flip-flop recombination, Plastome evolution, Plastid SSRs, Plastid SNPs

Introduction

The palm family (Arecaceae) comprises 188 genera and approximately 2,585 species distributed throughout tropical and subtropical ecosystems. This family has great ecological and economical importance in different regions of the planet (Dransfield et al. 2008; Palmweb, http://www.palmweb.org/). The genus Astrocaryum G. Mey. contains 40 species, which are divided into three subgenera, Astrocaryum, Munbaca and Monogynanthus (Kahn 2008). Among the species of Astrocaryum genus, A. murumuru Mart. and A. aculeatum G.Mey produce oil-rich fruits and are two of the most economically important species in the Amazon region as source of food, oil for cosmetic industry, and raw material for several uses (Clement et al. 2005; Bezerra 2012). A. murumuru (subgenus Monogynanthus Burret) is a stemmed palm adapted to seasonal swamp forest of the Amazon region, with occurrence in French Guiana, Guyana, Suriname, and North of Brazil. A. aculeatum (subgenus Astrocaryum), is a large palm adapted to terra firme forest of the Amazon region, with occurrence in Bolivia, Guyana, Suriname, Trinidad, Venezuela and North of Brazil (Kahn 2008).

Since the fruit harvesting of both species is majorly based on extractivist activity, several efforts have been made to access the genetic diversity and structure of natural populations to trace conservation strategies and improve the commercial significance of these palm fruits (Ramos et al. 2011, 2012, 2016; Oliveira et al. 2017). The identification and the characterization of molecular markers has essential importance to assess the genetic diversity of natural populations of both species. Only few nuclear microsatellites loci were developed for A. murumuru and A. aculeatum to date (Ramos et al. 2012; Oliveira et al. 2017). The plastome, a nonrecombinant and uniparentally inherited DNA molecule, is a great source of molecular markers such as single nucleotide polymorphisms (SNPs) and SSRs located majorly in the intergenic spacers (IGSs) and introns, where the mutation rates are higher in comparison with coding sequences (Rogalski et al. 2015; Vieira et al. 2016a; Lopes et al. 2018a). Several genetic studies have been employed plastid sequences, including phylogeographical, population genetic, and germplasm collections analyses (Ebert and Peakall 2009; Wheeler et al. 2014; Tsai et al. 2015; Wambulwa et al. 2016; Roy et al. 2016).

Plastome sequences have been also used to understand evolutionary events and to infer phylogenetic relationships with high efficacy (Lopes et al. 2018b). Due to the conservative nature of plastid genes and plastome structures, the presence of gene degeneration, gene transfer to nucleus, plastome rearrangements, and RNA editing, it is possible to delimit specific lineages (Martin et al. 2014; Vieira et al. 2016b; Bock 2017;

Lopes et al. 2018a, 2018b). The availability of 37 complete plastomes of different species belonging to the palm family, revealed that some species bear uncommon plastome structures, high gene divergence, and gene degeneration (Barret et al. 2016; Lopes et al. 2018a). Additionally, more than half plastid protein-coding genes in Arecaceae show one or more putative positive signatures (Lopes et al. 2018a). Therefore, the sequencing of plastomes belonging to other genera within Arecaceae, as Astrocaryum, may reveal new uncommon features.

Moreover, the genus Astrocaryum belongs to the subtribe Bactridinae (Arecoideae: Cocoseae), which comprises five genera (Astrocaryum, Acrocomia, Aiphanes, Bactris, and Desmoncus) whose relationships have been widely studied, however, they remain unresolved (Hahn 2002; Gunn 2004; Eiserhardt et al. 2011; Ludeña et al. 20011). Similarly, there are incongruences about the relationships between the species within the genus Astrocaryum, and the removal of two taxa (A. mexicanum Liebm. ex Mart. and A. alatum Loomis) and the creation of a new genus representing them has been suggested based on morphological and phylogenetic data (Pintaud et al. 2008; Eiserhardt et al. 2011; Ludeña et al. 2011). Currently, only the plastome of Acrocomia [A. aculeata (Jacq.) Lodd. Ex Mart.] is available within the subtribe Bactridinae, (Lopes et al. 2018a), which makes the sequencing of other genera important to identify evolutionary markers within Bactridinae.

Here, we reported the complete sequencing and characterization in detail of A. murumuru and A. aculeatum plastomes. Structural analysis showed a 4.6 kb inversion in the LSC of both plastomes, which is not present in the other palm species already sequenced, including the related genus Acrocomia. Our analyses demonstrate the occurrence of flip-flop recombination on small inverted repeat sequences which generates this rearrangement in the plastid genome of Astrocaryum. In addition, the rearrangement occurred in a common ancestor plastome. Moreover, we mapped all polymorphic sites by comparison between A. murumuru and A. aculeatum plastomes, which revealed 27 polymorphic SSR loci, 120 SNPs, 35 indels, and six hotspots of nucleotide diversity. Among the 120 SNPs characterized here, 54 are located in 16 coding sequences resulting in 18 synonymous and 25 non-synonymous substitutions, changing the conserved amino acid. Furthermore, we performed a phylogenetic analysis using whole plastomes of 39 Arecaceae species, which placed the Astrocaryum species sister to Acrocomia within the Cocoseae tribe. Finally, our data present a novel lineage-specific rearrangement, several molecular markers, non-synonymous substitutions, gain of RNA editing sites within the

genus Astrocaryum and raise questions about the relationship between plastome evolution and environmental adaptation of close related species within Arecaceae.

Materials and methods

Plant material, chloroplast isolation and DNA extraction

Fresh and young leaves from Astrocaryum murumuru and A. aculeatum plants were collected in the city of Irituia, State of Pará, Brazil, and kept for 1 week at 4 °C to decrease starch level before the chloroplast isolation process. The chloroplast isolation and DNA extraction were carried out according to Vieira et al. (2014).

Plastome sequencing, assembling, and annotation

Approximately 1 ng of plastid DNA was used to prepare sequencing libraries with Nextera XT DNA Sample Prep Kit (Illumina Inc., San Diego, CA, USA) according to the manufacturer's instructions. The obtained library was sequenced using Illumina MiSeq platform (Illumina Inc., San Diego, CA, USA) at the Federal University of Paraná, State of Paraná, Brazil. The reads obtained were trimmed (threshold with probability of error < 0.05) and de novo assembled in contigs using CLC Genomics Workbench 11.0 software (CLC Bio, Aarhus, Denmark). The sequencing of A. murumuru and A. aculeatum resulted, respectively, in 1,344,610 reads of average length 187.0 and 648,656 reads of average length 148.3. The contigs used for assembling the plastomes ranged from 440.57 to 118.38 (A. murumuru) and from 185.06 to 68.87 (A. aculeatum) of average coverage. The program Dual Organellar GenoMe Annotator (DOGMA) (Wyman et al. 2004) and BLAST were used for preliminarily gene annotation. From this initial annotation, putative start codons, stop codons, and intron positions were determined based on comparisons to homologous genes of other plastomes at the GenBank database. All tRNA genes were further verified by using tRNAscan- SE server (Lowe and Eddy 1997). The physical circular map of the plastomes were drawn using Organellar Genome DRAW (OGDRAW) (Lohse et al. 2013).

Comparative analysis of plastome structure

To characterize the general plastome structure of the species A. murmuru and A. aculeatum, nucleotide MUMmer (NUCmer) Perl script in MUMmer 3.0 (Kurtz et al. 2004) was used to visualize and compare the plastome structures within the genus Astrocaryum and among other Arecaceae representatives.

Identification of polymorphic SSRs, SNPs, Indels, nucleotide divergence hotspots and dispersed repeats

Simple sequence repeats (SSRs) loci were detected in the two species of Astrocaryum using the MIcroSAtellite (MISA) Perl script (Thiel et al. 2003). The thresholds were set to eight repeat units for mononucleotide SSRs, four repeat units for di- and trinucleotide SSRs, and three repeat units for tetra-, penta-, and hexanucleotide SSRs. After, we manually located the polymorphic SSRs based on the alignment produced between A. murumuru and A. aculeatum using MAFFT v.7 (Katoh and Standley 2013). To identify single nucleotide polymorphisms (SNPs), small insertions/deletions (indels), and nucleotide divergence hotspots, we used the alignment between A. murumuru and A. aculeatum as input data in the DnaSP v.5 software (Librado and Rozas 2009). The nucleotide divergence hotspots were located by sliding window analysis with window length of 200 bp and step size of 50 bp. Dispersed repeats of length ≥ 30 bp and identity of repeats ≥ 90 % were additionally identified using the program REPuter (Kurtz et al. 2001).

Prediction of RNA-editing sites

Potential RNA editing sites in plastid protein-coding genes of A. murumuru and A. aculeatum were predicted by using the program Predictive RNA Editor for Plants (PREP) (Mower 2009). The program PREP uses 35 reference genes to detect possible RNA editing sites (reference genes: accD, atpA, atpB, atpF, atpI, ccsA, clpP, matK, ndhA, ndhB, ndhD, ndhF, ndhG, petB, petD, petG, petL, psaB, psaI, psbB, psbE, psbF, psbL, rpl2, rpl20, rpl23, rpoA, rpoB, rpoC1, rpoC2, rps2, rps8, rps14, rps16, and ycf3). The cutoff value was set to 0.8. Additional RNA editing sites were predicted by comparison with the RT-PCR and sequencing data from Huang et al. (2013) that identified several RNA editing sites in transcripts of plastid genes of Cocos nucifera L.

Phylogenomic reconstruction

The inference of A. murumuru and A. aculeatum phylogenetic positions within the family Arecaceae was carried out using whole plastomes. The GenBank accession number of each taxon used is shown in the **Supplementary Table S1**, including 39 palm species representing all five subfamilies of Arecaceae. The species Hanguana malayana (Hanguanaceae: Commelinales), Baxteria australis, and Dasypogon bromeliifolius (Dasypogonaceae: Arecales) were used as outgroups. First, whole plastomes were extracted from GenBank and the IRB was withdrawn to prevent overrepresentation of the IR sequences. The plastomes were aligned using MAFFT v.7 (Katoh and Standley 2013) and based on jModelTest v.2.1.7 we select the substitution model GTR+I+G. Last,

Bayesian inference analysis was performed using MrBayes version 3.2 (Ronquist et al. 2012), with one million generations of two runs of four Markov Chains, with three hot and one cold in each run. To check the parameter convergence, we used the software Tracer 1.6 (http://tree.bio.ed.ac.uk/software/tracer/). The software FigTree 1.4.2 (http://tree.bio.ed.ac.uk/software /figtree/) was used to visualize the consensus tree.

Results

Structure and gene content of Astrocaryum plastomes

The plastomes of A. murumuru (**Fig. 1**) and A. aculeatum (**Fig. 2**) are circular molecules with the typical quadripartite structure, including two inverted repeat regions (IRA and IRB) between two single copy regions (LSC and SSC). The size of LSC, SSC, and IR regions of both Astrocaryum plastomes are very similar (**Table 1**), being the total size of A. murumuru plastome only 16 bp larger than A. aculeatum plastome. A more detailed view about the plastome structure, by dot-plot analyses, shows high structure conservation between the two species of Astrocaryum sequenced here (**Fig. 3a**), and highlights an inversion of about 4.6-kb within the LSC region when compared with most common palm plastome structure (**Fig. 3b**), like Acrocomia aculeata (Lopes et al. 2018a). The inverted region comprises of a block of genes, including ndhC, ndhK, ndhJ, trnF-GAA, and trnL-UAA genes.

The gene content of both Astrocaryum plastomes includes 113 unique genes, of which 79 are protein-coding genes, 30 are tRNA genes, and four are rRNA genes (**Table 2**). Within the IR regions there are 20 duplicated genes, being eight protein-coding genes (one of them, the ycf1 gene, is partially duplicated), eight tRNA genes, and four rRNA genes. The clpP and ycf3 genes have two introns, and other 16 genes have only one intron, including six tRNA genes and ten protein-coding genes.

Sequence repeats and polymorphic sites in Astrocaryum plastomes

We identified 220 and 223 SSR loci in the plastomes of A. murumuru and A. aculeatum, respectively. Most of them, about 65-66%, comprises A/T mononucleotide repeats (**Supplementary Fig. S1**). The density of SSR loci (SSR number / kilobase) was similar between the plastomes sequenced here, being 2.2 in the SSC, 1.9 in the LCS, and only 0.7 in the IR regions. The total density was 1.7, considering one IR. These values are in accordance with previously analysis within the subfamily Arecoideae (Lopes et al. 2018a). The number and distribution of SSR loci are well conserved between A. murumuru and A. aculeatum (**Supplementary Table S2**). However, a total of 27 loci have polymorphic sequences (**Table 3**), being 20 of them within intergenic spacers (IGSs)

and only five and two of them located in introns and coding sequences (CDSs), respectively.

The plastomes of Astrocaryum species have also a conserved set of eight dispersed repeats (three inverted and five direct repeats; **Supplementary Table S3**), including an inverted repeat between the trnT-UGU/ndhC and trnL-UAA/trnV-UAC intergenic regions, located in the flanks of the inversion of 4.6 kb abovementioned. Other dispersed repeats identified are in conserved genes (e.g. psaA/psaB and trnS genes), as usually found in plastomes (Raubeson et al. 2007). Additionally, five dispersed repeats are present only in A. murumuru (two repeats) or A. aculeatum (three repeats).

To map all polymorphic sites between the plastomes of A. murmuru and A. aculeatum, we also identified the indels and the SNPs. A total of 35 indels of one to 35 pb are present in the plastomes of Astrocaryum (**Supplementary Table S4**), most of them within IGSs (26 indels), but also located in introns (8 indels) and CDS (one indel in the rps3 gene). Comparing the plastomes of A. murumuru and A. aculeatum we found 120 SNPs, of which 57 are in IGSs, 54 are in CDSs, and nine are in introns (**Supplementary Table S5**). The highest number of SNPs is in the ycf1 gene (28 SNPs). The SNPs located in CDSs changed 44 codons in 16 genes (**Table 4**), resulting in 18 synonymous substitutions and 25 non-synonymous substitutions. The ycf1 gene have the highest number of non-synonymous substitutions (16).

Ultimately, based on sliding window analysis, six regions were identified as hotspots of nucleotide diversity in the plastomes of Astrocaryum (**Fig. 4**). All these IGSs are located within single copy regions, LSC and SSC. The highest nucleotide diversity is in the rps19/psbA IGS and in the ycf1 gene (the part of the gene included within the SSC).

Prediction of RNA editing in plastid genes of Astrocaryum

The prediction of RNA editing sites in plastid genes of A. murumuru and A. aculeatum was carried out based on PREP program and comparison with Huang et al. (2013) that validated several RNA editing sites in plastid genes of Cocos nucifera. All RNA editions predicted are C-to-U conversions, at the first (20.4 %) or second (79.6 %) positions of the codons (**Table 5**). We identified 92 RNA editing sites shared by both species of Astrocaryum sequenced here. Only one RNA editing site, in the ccsA gene (amino acid position 21), seems to be unique for A. murumuru, since in this position there is a T fixed in the A. aculeatum plastome, as well as in the other 37 species of Arecaceae that has a complete plastome published. We found that 91 RNA editions predicted in Astrocaryum are conserved within the subfamily Arecoideae, and one seems to be specific for the tribe Cocoseae (accD gene, amino acid position 387) based on comparison with

our previously analysis (Lopes et al. 2018a). From the 93 RNA editing sites predicted here, according to Huang et al. (2013), 57 sites are completely edited and 16 are partially edited in C. nucifera. Most RNA editions identified here change the encoded amino acid from polar or electrically charged to apolar (60 out of 93), specially changing serine to leucine. Only one predicted RNA edition changes the amino acid from apolar to polar (proline-serine; amino acid position 321 in the ndhA gene), and the last 32 editions do not change the amino acid polarity (15, apolar-apolar; 17, polar-polar).

Phylogenomic inference

We carried out a phylogenomic analysis aiming to infer the position of Astrocaryum within the family Arecaceae based on whole plastomes. This analysis is an update from our Arecaceae phylogenomic previously published in Lopes et al. (2018a), which includes species representing all five subfamilies of Arecaceae, with the addition of A. murumuru and A. aculeatum sequenced here. Bayesian inference (BI) analysis produced a phylogenetic tree with a -lnL = 516,094.965 (**Fig. 5**) and high branch support (BI posterior probability value of 1 for all nodes). The relationships among the subfamilies and within the subfamilies are in accordance to previously published (Lopes et al. 2018a). The two species of Astrocaryum formed a sister-group with Acrocomia aculeate, within the tribe Cocoseae in the subfamily Arecoideae. Within the tribe Cocoseae, two mainly clades are formed: one including the species Elaies guineensis, A. aculeata, Astrocaryum murmuru, and A. aculeatum; and other including the species C. nucifera and Syagrus coronata.

Discussion

The evolution of plastome structure within the genus Astrocaryum included a flip-flop recombination that resulted in a lineage-specific rearrangement

The general features, such as gene content and most part of the gene order, in the plastomes of A. murumuru and A. aculeatum are shared with the plastomes of other palms already known (Huang et al. 2013; Barret et al. 2016; Lopes et al. 2018a). The only exception is the 4.6-kb inversion within the LSC that we identified in both plastomes sequenced. Among the 37 species of palm with a complete plastome sequenced and available at the GenBank, only Tahina spectabilis present a rearrangement event which resulted in a 1.9-kb inversion located between the genes rps16 and trnG-UUC (Barret et al. 2016). Different from Tahina, the 4.6-kb inversion of Astrocaryum plastomes is inserted between the genes trnT-UGU and trnV-UAC, encompassing a block of genes

composed by five genes. Thereby it is very probable that this 4.6-kb inversion is a lineage-specific structural feature of the genus Astrocaryum.

Several studies have been suggested the association between the plastome rearrangements and the presence of dispersed repeats (Milligan et al. 1989; Haberle et al. 2008; Martin et al. 2014; Weng et al. 2014). Rogalski et al. (2006) previously demonstrated the potential of inverted repeats to cause flip-flop recombination, changing the orientation of the segment between the repeated sequences. Curiously, among the dispersed repeats identified in the plastomes of Astrocaryum, we identify a pair of inverted repeats located in the rearrangement endpoints (the rearrangement endpoints were inferred based on alignment with close taxa). This pair of inverted repeats was not found in other species included in the subfamily Arecoideae, although Veitchia arecina, E. guineensis, and A. aculeata conserve a single segment (between the genes ndhC and trnV-UAC) homologue to the pair of repeats in Astrocaryum. Based on these findings, we hypothesized that the ancestor of the genus Astrocaryum evolved a segment between the genes trnT-UGU and trnL-UAA homologue to the conserved sequence present in the ndhc/trnV-UAC IGS, thus acquiring a pair of inverted repeats. After, flip-flop recombination gave rise the current structure of Astrocaryum plastomes (**Fig. 6**).

The presence of active flip-flop recombination between inverted repeats can create isomeric forms of the plastome. In fact, the presence of two isoforms in different individuals, ecotypes or in a single plant have been demonstrated (Guo et al. 2014; Gurdon and Maliga, 2014; Vieira et al. 2016b). However, the assembling of the paired-end reads in contigs do not resulted in any other isoform in both species of Astrocaryum, indicating that the inverted form is the only plastome structure present. Future PCR amplification and sequencing of the rearrangement endpoints using different individuals of A. murumuru and A. aculeatum will be important to clarify this issue.

The evolution of plastid genes in the genus Astrocaryum included more non-synonymous substitutions than synonymous substitutions and gain of RNA editing site

In comparison with other commelinids, the plastomes within Arecaceae has the lower substitution rate (Barret et al. 2016). However, some plastid genes have high nucleotide divergence in a few distinct species/genera and more than half of plastid genes within the palm family has one or more putative positive signatures (Lopes et al. 2018a). Within the genus Astrocaryum, we found high level of conservation among the plastid genes of A. murumuru and A. aculeatum, sharing the same sequence of all tRNA and rRNA genes and in 62 out of 79 protein-coding genes. Only one gene (rps3) diverges due

indel, the other 16 genes are different due presence of SNPs, changing the sequence of 44 codons. Interestingly, most part of the codon changes resulted in non-synonymous substitutions (25 out of 44), affecting genes acting in the photosynthesis (atpF, ndhA, and ndhF genes), gene expression (matK, rpoC2, rps11, and rps15 genes), cytochrome synthesis (ccsA gene), TIC complex (ycf1 gene), and unknown function (ycf2 gene).

Comparing these sites of non-synonymous substitutions in Astrocaryum with the sequence of amino acid coded in other palms (**Supplementary Fig. S2** and **S3**), we observe the tendency of A. aculeatum to keep the most conserved amino acid, while A. murumuru tends to evolve unique amino acids [ccsA (21), ndhA (204), and ycf1 (1072) genes] or to converge in the same amino acid coded by other distantly related species [e.g. matK (230), ndhF (565), rps11 (75), and rps15 (41)]. These non-synonymous substitutions between A. murumuru and A. aculeatum plastid genes may be part of an adaptive response to different environments. While A. murumuru grows in wet or temporarily flooded area and is shade-tolerant (Bezerra 2012; Kahn 2008; Choo et al. 2017), A. aculeatum, in turn, is adapted to non-flooded and non-shaded areas such as deforested areas or areas that underwent anthropic action (Kahn 2008; Ramos et al. 2016).

Among the 25 non-synonymous substitutions identified, the site 21 in the ccsA gene was predicted to be a RNA editing site in A. murumuru, recovering the conserved amino acid [GCG (A) => GUG (V)]. In all Arecales species with a complete plastome sequence available, including two species of the family Dasypogonaceae, is codified a valine in this site without need for edition (**Supplementary Fig. S2**). Therefore, additionally to non-synonymous substitutions, it is probable that the species A. murumuru gained a new RNA editing site. In a previously study about prediction of RNA editing within the subfamily Arecoideae we also found some gain and loss events of RNA editions (Lopes et al. 2018a). Although the number of RNA editing sites underwent a decrease across the evolution of higher plants (Takenaka et al. 2013), the RNA editing sites has a dynamic pattern and is hypothesized that they evolve more readily in genes, or regions of the genes, non-essentials for cell survival (Fiebig et al. 2004). The other 92 RNA editing sites predicted here are shared between both Astrocaryum species and are also present in representatives of the subfamily Arecoideae (Lopes et al. 2018a), being prevalent the editions that increase the protein hydrophobicity, which may improve the hydrophobic interactions of protein complexes and transmembrane domains (He et al. 2016; Chen et al. 2017; Lopes et al. 2018a).

Comparison between the plastome sequences of A. murumuru and A. aculeatum revels several polymorphic sites

The genetic informations from plastomes are useful tools to several evolutionary and genetic studies, since, overall, they present a nonrecombinant nature and uniparental inheritance (Provan et al. 2001; Wheeler et al. 2014; Rogalski et al. 2015). Here, we mapped all polymorphic sites between the plastomes of A. murumuru and A. aculeatum, resulting in 35 indels, 120 SNPs, and 27 SSRs loci. In addition, based on sliding window analysis, we identified the junction IRB-LSC (IGS between the rps19 and psbA genes) and the ycf1 gene as the major hotspots of nucleotide diversity. The high polymorphism in the junction IRB-LSC may be due expansion and contractions events by gene conversion common in the IR borders (Goulding et al. 1996; Zhu et al. 2016). The other hotspot, the ycf1 gene, has the higher substitution rate among the plastid genes within Arecaceae (Lopes et al. 2018a), and we found a high number of non-synonymous substitutions, suggesting a fast-evolution rate of this gene in the genus Astrocaryum.

Until now, few specific molecular markers have been identified for A. aculeatum and A. murumuru, being limited to nuclear SSRs loci (Ramos et al. 2011, 2012; Oliveira et al. 2017). Scarcelli et al. (2011) developed a set of 100 primer pairs to amplify plastid markers for monocots, including Arecaceae, aiming genetic studies of populations and phylogeny. Nevertheless, studies using specific primers instead universal primers reported a greater number of polymorphic loci (Wheeler et al. 2014). The complete sequencing and mapping of all polymorphisms in the plastomes of A. murumuru and A. aculeatum are important to design specific primers targeting selected markers. The insertion of plastid markers along with the nuclear ones will improve population studies such as access the genetic structure, diversity and gene flows in natural populations (Wheeler et al. 2014), contributing with pioneer studies (Ramos et al. 2011, 2012, 2016; Oliveira et al. 2017).

Further, the plastid markers will contribute to phylogenetic analysis of the subtribe Bactridinae and the genus Astrocaryum, that are still unresolved (Dransfield et al. 2008; Eiserhardt et al. 2011; Ludeña et al. 2011). Our phylogenomic tree placed the genus Astrocaryum sister to Acrocomia, which is in accordance with the accepted classifications that put both genera in the subtribe Bactridinae (Dransfield et al. 2008). However, the absence of data from the other genera precludes us to infer about the relationships at subtribe level. The rearrangement that we found in the plastomes of A. murumuru and A. aculeatum (the 4.6-kb inversion within the LSC) configures in an important evolutionary trait, since plastome rearrangements are rare events in angiosperms (Rogalski et al. 2015).

Thus, major lineages within Bactridinae may be defined based on presence or absence of this 4.6-kb inversion. So far, we know that within Bactridinae, Astrocaryum bears this rearrangement while Acrocomia does not, but remains unknow if the other three genera of the subtribe bear this trait.

Conclusions

In this study we reported the complete plastomes of two species of the genus Astrocaryum, A. murumuru and A. aculeatum. From detailed comparison between both plastomes we mapped all polymorphic sites, including 27 polymorphic SSR loci, 120 SNPs, 35 indels, and six hotspots of nucleotide diversity, providing several molecular markers for genetic studies of these important natural resources of Amazon region. We also found unique features between A. murumuru and A. aculeatum plastid genes, such as non-synonymous substitutions, some of them in codons that codify amino acid very conserved within Arecaceae, and a putative gain of RNA editing. These molecular divergences associated with the different environmental adaptations between A. murumuru and A. aculeatum may indicate that substantial molecular evolution occurs in plastids even between very close related species. Finally, the plastomes of Astrocaryum species bear a 4.6-kb inversion originated from a flip-flop recombination, and comparison with other Arecaceae genera indicates that this rearrangement is a lineage-specific structural feature. The plastomes of the family Arecaceae are, in general, conserved and has a low substitution rate, therefore, the rearrangement found here is an important evolutionary trait within the family and can be a helpful tool to resolve phylogeny conflicts within the subtribe Bactridinae.

Acknowledgements

This research was support by the National Council for Scientific and Technological Development, Brazil (CNPq, Grant 459698/2014-1). We are grateful to INCT-FBN and for the scholarships granted by the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES) to TGP, ONS, GDMV, and LNV, and those granted by the CNPq to ASL, LMS, EB, EMS and FOP. We are also greatful to the Núcleo de Análise de Biomoléculas (NuBiomol) of the Universidade Federal de Viçosa for providing the software CLC Genomics.

References

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. doi: 10.1111/nph.13617

Bezerra VS (2012) Considerações sobre a palmeira murumuruzeiro (Astrocaryum murumuru Mart.). Embrapa, Comunicado Técnico 130

Bock R (2017) Witnessing Genome Evolution: Experimental Reconstruction of Endosymbiotic and Horizontal Gene Transfer. Annu Rev Genet (In press). doi: 10.1146/annurev-genet-120215-035329

Chen TC, Liu YC, Wang X, Wu CH, Huang CH, Chang CC (2017) Whole plastid transcriptomes reveal abundant RNA editing sites and differential editing status in Phalaenopsis aphrodite subsp. formosana. Bot Stud 58:38. doi: 10.1186/s40529-017-0193-7

Choo J, Carasco C, Alvarez-Loayza P, Simpson BB, Economo EP (2017) Life history traits influence the strength of distance- and density-dependence at different life stages of two Amazonian palms. Ann Bot 120: 147–158. doi: 10.1093/aob/mcx051

Clement CR, Lleras Pérez E, Van Leeuwen J (2005) O potencial das palmeiras tropicais no Brasil: acertos e fracassos das últimas décadas. Agrociências 9: 67-71

Dransfield J, Uhl NW, Asmussen CB, Baker WJ, Harley M, Lewis C (2008) Genera Palmarum: the evolution and classification of palms. Kew Publishing, Royal Botanical Garden, Londres, 732

Ebert D, Peakall R (2009) Chloroplast simple sequence repeats (cpSSRs): technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant species. Mol Ecol Resour 9:673-690. doi: 10.1111/j.1755-0998.2008.02319.x

Eiserhardt WL, Pintaud JC, Asmussen-Lange C, et al. 2011. Phylogeny and divergence times of Bactridinae (Arecaceae, Palmae) based on plastid and nuclear DNA sequences. Taxon 60(2): 485–498

Fiebig A, Stegemann S, Bock R (2004) Rapid evolution of RNA editing sites in a small non-essential plastid gene. Nucleic Acids Res 32:3615-3622. doi: 10.1093/nar/gkh695

Goulding SE, Olmstead RG, Morden CW, Wolfe KH (1996) Ebb and flow of the chloroplast inverted repeat. Mol Gen Genet 252:195-206

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

Guo W, Grewe F, Cobo-Clark A, Fan W, Duan Z, Adams RP, Schwarzbach AE, Mower JP (2014) Predominant and substoichiometric isomers of the plastid genome coexist within Juniperus plants and have shifted multiple times during cupressophyte evolution. Genome Biol Evol 6: 580–590. doi: 10.1093/gbe/evu046

Gurdon C, Maliga P (2014) Two distinct plastid genome configurations and unprecedented intraspecies length variation in the accD coding region in Medicago truncatula. DNA Res Int J Rapid Publ Rep Genes Genomes 21: 417–427. doi: org/10.1093/dnares/dsu007

Haberle RC, Fourcade HM, Boore JL, Jansen RK (2008) Extensive Rearrangements in the Chloroplast Genome of Trachelium caeruleum Are Associated with Repeats and tRNA Genes. J Mol Evol 66:350-361. doi: 10.1007/s00239-008-9086-4

Hahn WJ (2002) A phylogenetic analysis of the Arecoid Line of palms based on plastid DNA sequence data. Mol Phylogenet Evol 23: 189–204. doi: 10.1016/S1055-7903(02)00022-2

He P, Huang S, Xiao G, Zhang Y, Yu J (2016) Abundant RNA editing sites of chloroplast protein-coding genes in Ginkgo biloba and an evolutionary pattern analysis. BMC Plant Biol 16(1):257. doi: 10.1186/s12870-016-0944-8

Huang YY, Matzke AJM, Matzke M (2013) Complete Sequence and Comparative Analysis of the Chloroplast Genome of Coconut Palm (Cocos nucifera). PLoS ONE 8:e74736. doi: 10.1371/journal.pone.0074736

Kahn F (2008) El género Astrocaryum (Arecaceae). Rev Peru Biol 15: 31–48

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772-780. doi: 10.1093/molbev/mst010

Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29:4633–4642. doi:10.1093/nar/29.22.4633

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. Genome Biol 5:R12. doi: 10.1186/gb-2004-5-2-r12

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451-1452. doi: 10.1093/bioinformatics/btp187

Lohse M, Drechsel O, Kahlau S, Bock R (2013) OrganellarGenomeDRAW – a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. Nucleic Acids Res 41:W575-W581. doi: 10.1093/nar/gkt289

Lopes AS, Pacheco TG, Nimz T, Vieira LN, Guerra MP, Nodari RO, de Souza EM, Pedrosa FO, Rogalski M (2018a) 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. doi: 10.1007/s00425-018-2841-x

Lopes AS, Pacheco TG, Santos KGD, Vieira LN, Guerra MP, Nodari RO, de Souza EM, Pedrosa FO, Rogalski M (2018b) The Linum usitatissimum L. plastome reveals atypical structural evolution, new editing sites, and the phylogenetic position of Linaceae within Malpighiales. Plant Cell Rep 37: 307–328. doi: 10.1007/s00299-017-2231-z

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955-964

Ludeña B, Chabrillange N, Aberlenc-Bertossi F, Adam H, Tregear JW, Pintaud J-C (2011) Phylogenetic utility of the nuclear genes AGAMOUS 1 and PHYTOCHROME B in palms (Arecaceae): an example within Bactridinae. Ann Bot 108: 1433–1444. doi: 10.1093/aob/mcr191

Martin GE, Rousseau-Gueutin M, Cordonnier S, Lima O, Michon-Coudouel S, Naquin D, de Carvalho JF, Aïnouche M, Salmon A, Aïnouche A (2014) The first complete chloroplast genome of the Genistoid legume Lupinus luteus: evidence for a novel major lineage-specific rearrangement and new insights regarding plastome evolution in the legume family. Ann Bot 113: 1197–1210. doi: 10.1093/aob/mcu050

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

Mower JP (2009) The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. Nucleic Acids Res 37:W253-W259. doi: 10.1093/nar/gkp337

Oliveira NP, Oliveira MSP, Davide LC, Kalisz S (2017) Population genetic structure of three species in the genus Astrocaryum G. Mey. (Arecaceae). Genet Mol Res GMR 16. doi: 10.4238/gmr16039676

Pintaud J-C, Millán B, Kahn F (2008) The genus Hexopetion Burret (Arecaceae). Rev Peru Biol 15: 49–54

Provan J, Powell W, Hollingsworth PM (2001) Chloroplast microsatellites: new tools for studies in plant ecology and evolution. Trends Ecol Evol 16:142-147

Ramos SLF, de Macêdo JLV, Lopes MTG, Batista JS, Formiga KM, da Silva PP, Saulo-Machado AC, Veasey EA (2012) Microsatellite loci for tucumã of Amazonas (Astrocaryum aculeatum) and amplification in other Arecaceae. Am J Bot 99: e508–e510. doi: 10.3732/ajb.1100607

Ramos SLF, Dequigiovanni G, Sebbenn AM, Lopes MTG, Kageyama PY, de Macêdo JLV, Kirst M, Veasey EA (2016) Spatial genetic structure, genetic diversity and pollen dispersal in a harvested population of Astrocaryum aculeatum in the Brazilian Amazon. BMC Genet 17(63). doi: 10.1186/s12863-016-0371-8

Ramos SLF, Lopes MTG, Lopes R, Cunha RNV da, Macêdo JLV de, Contim LAS, Clement CR, Rodrigues DP, Bernardes LG (2011) Determination of the mating system of Tucumã palm using microsatellite markers. Crop Breed Appl Biotechnol 11: 181–185. doi: 10.1590/S1984-70332011000200011

Raubeson LA, Peery R, Chumley TW, Dziubek C, Fourcade HM, Boore JL, Jansen RK (2007) Comparative chloroplast genomics: analyses including new sequences from the angiosperms Nuphar advena and Ranunculus macranthus. BMC Genom 8:174. doi:10.1186/1471-2164-8-174

Rogalski M, do Nascimento Vieira L, Fraga HP, Guerra MP (2015) Plastid genomics in horticultural species: importance and applications for plant population genetics, evolution, and biotechnology. Front Plant Sci 6:586. doi: 10.3389/fpls.2015.00586

Rogalski M, Ruf S, Bock R (2006) Tobacco plastid ribosomal protein S18 is essential for cell survival. Nucleic Acids Res 34:4537–4545. doi: 10.1093/nar/gkl634

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP (2012) MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Syst Biol 61:539-542. doi: 10.1093/sysbio/sys029

Roy PS, Rao GJN, Jena S, Samal R, Patnaik A, Patnaik SSC, Jambhulkar NN, Sharma S, Mohapatra T (2016) Nuclear and Chloroplast DNA Variation Provides Insights into Population Structure and Multiple Origin of Native Aromatic Rices of Odisha, India. PloS One 11:e0162268. doi: 10.1371/journal.pone.0162268

Scarcelli N, Barnaud A, Eiserhardt W, Treier UA, Seveno M, d'Anfray A, Vigouroux Y, Pintaud J-C (2011) A Set of 100 Chloroplast DNA Primer Pairs to Study Population Genetics and Phylogeny in Monocotyledons. PLoS ONE 6. doi: 10.1371/journal.pone.0019954

Takenaka M, Zehrmann A, Verbitskiy D, Härtel B, Brennicke A (2013) RNA Editing in Plants and Its Evolution. Annu Rev Genet 47:335-352. doi: 10.1146/annurev-genet-111212-133519

Thiel T, Michalek W, Varshney R, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). Theor Appl Genet 106:411-422. doi: 10.1007/s00122-002-1031-0

Tsai CC, Chou CH, Wang HV, Ko YZ, Chiang TY, Chiang YC (2015) Biogeography of the Phalaenopsis amabilis species complex inferred from nuclear and plastid DNAs. BMC Plant Biol 15:202. doi: 10.1186/s12870-015-0560-z

Vieira LN, Dos Anjos KG, Faoro H, Fraga HP, Greco TM, Pedrosa FO, de Souza EM, Rogalski M, de Souza RF, Guerra MP (2016a) Phylogenetic inference and SSR characterization of tropical woody bamboos tribe Bambuseae (Poaceae: Bambusoideae) based on complete plastid genome sequences. Curr Genet 62(2):443-453. doi: 10.1007/s00294-015-0549-z

Vieira LN, Faoro H, Fraga HPF, Rogalski M, de Souza EM, de Oliveira Pedrosa F, Nodari RO, Guerra MP (2014) An Improved Protocol for Intact Chloroplasts and cpDNA Isolation in Conifers. PLoS ONE 9:e84792. doi: 10.1371/journal.pone.0084792

Vieira LN, Rogalski M, Faoro H, Fraga HP, Anjos KG, Picchi GFA, Nodari RO, Pedrosa FO, Souza EM, Guerra MP (2016b) The plastome sequence of the endemic Amazonian conifer, Retrophyllum piresii (Silba) C.N.Page, reveals different recombination events and plastome isoforms. Tree Genet Genomes 12:10. doi: 10.1007/s11295-016-0968-0

Wambulwa MC, Meegahakumbura MK, Kamunya S, Muchugi A, Möller M, Liu J, Xu JC, Ranjitkar S, Li DZ, Gao LM (2016) Insights into the Genetic Relationships and Breeding Patterns of the African Tea Germplasm Based on nSSR Markers and cpDNA Sequences. Front Plant Sci 7:1244. doi: 10.3389/fpls.2016.01244

Weng ML, Blazier JC, Govindu M, Jansen RK (2014) Reconstruction of the Ancestral Plastid Genome in Geraniaceae Reveals a Correlation between Genome Rearrangements, Repeats, and Nucleotide Substitution Rates. Mol Biol Evol 31:645-659. doi: 10.1093/molbev/mst257

Wheeler GL, Dorman HE, Buchanan A, Challagundla L, Wallace LE (2014) A review of the prevalence, utility, and caveats of using chloroplast simple sequence repeats for studies of plant biology. Appl Plant Sci 2(12). doi: 10.3732/apps.1400059

Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20:3252-3255. doi: 10.1093/bioinformatics/bth352

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

FIGURES

Fig. 1 Gene map of Astrocaryum murumuru plastome. Genes drawn inside the circle are transcribed in the clockwise direction, and genes drawn outside are transcribed in the counterclockwise direction. Different functional groups of genes are color-coded. The darker gray in the inner circle corresponds to GC content, and the lighter gray corresponds to AT content. LSC, Large Single Copy; SSC, Small Single Copy; IRA/B, Inverted Repeat A/B

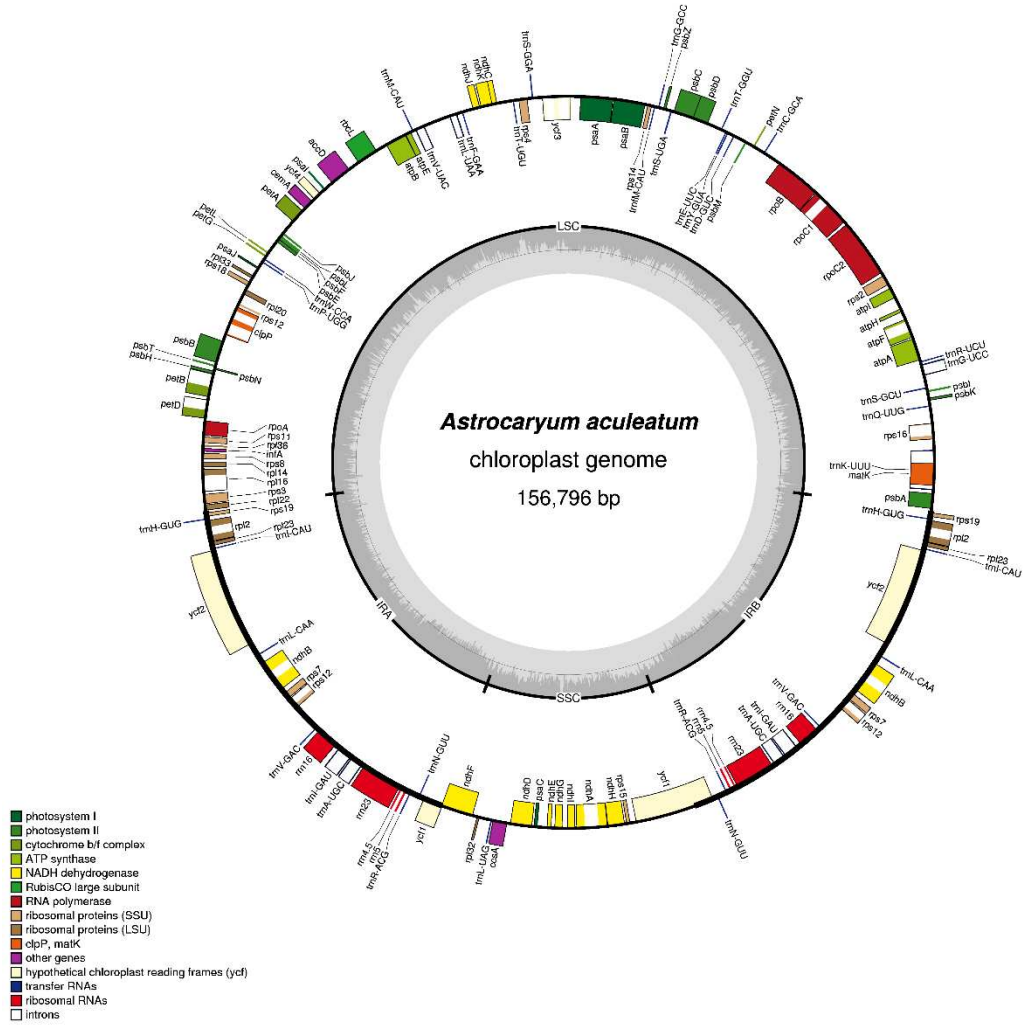


Fig. 2 Gene map of Astrocaryum aculeatum plastome. Genes drawn inside the circle are transcribed in the clockwise direction, and genes drawn outside are transcribed in the counterclockwise direction. Different functional groups of genes are color-coded. The darker gray in the inner circle corresponds to GC content, and the lighter gray corresponds to AT content. LSC, Large Single Copy; SSC, Small Single Copy; IRA/B, Inverted Repeat A/B

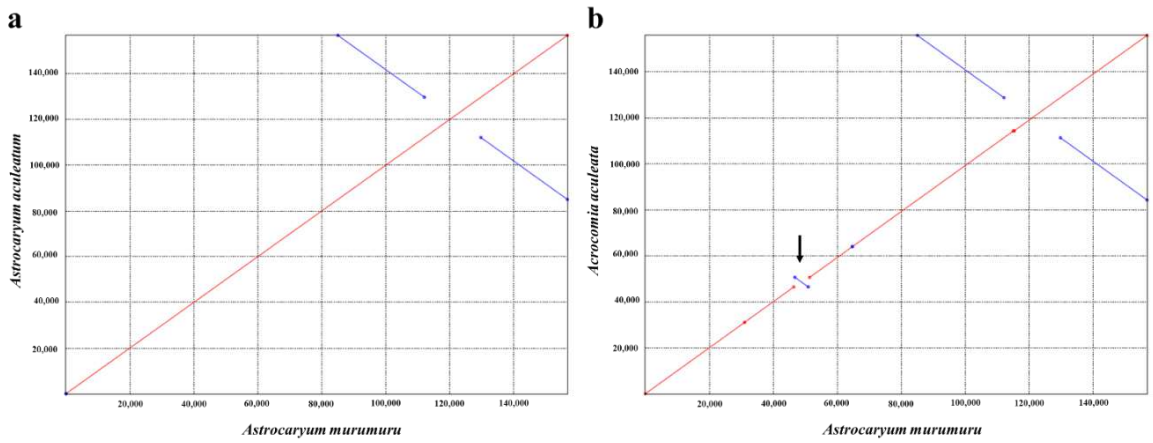


Fig. 3 Dot-plot analyses comparing the plastomes of (**a**) Astrocaryum murumuru (X-axis) against A. aculeatum (Y-axis) and (**b**) A. murumuru (X-axis) against Acrocomia aculeata (Y-axis). A positive slope denotes that the pair of sequences compared is in the same orientation. A negative slope denotes that the pair of sequences compared can be aligned, but their orientation is opposite. Sequences in the same direction are red and inversions are blue. The arrow highlights an inversion of approximately 4.6 kb in the LSC region

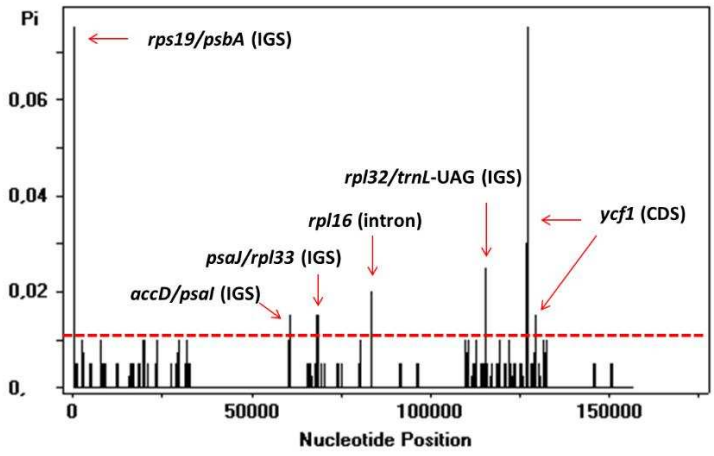


Fig. 4 Sliding window analysis of aligned whole plastomes of Astrocaryum murumuru and A. aculeatum. The regions with high nucleotide variability (Pi > 0.01) are indicated. Pi, nucleotide diversity of each window. Window length, 200 pb. Step size, 50 pb

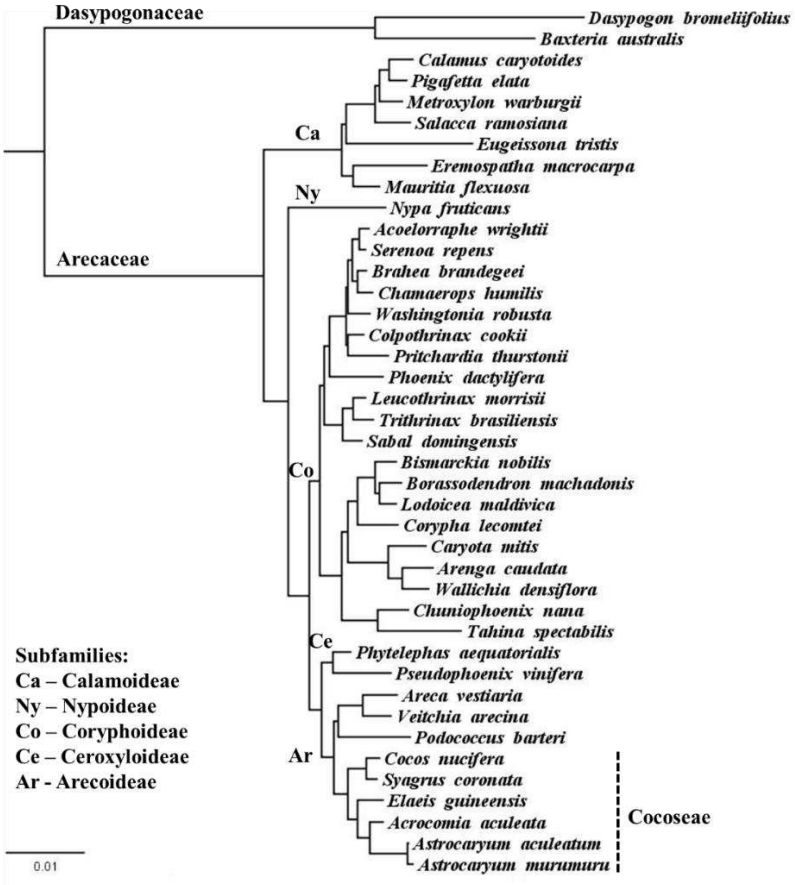


Fig. 5 Arecaceae phylogenomic tree of 42 taxa (39 Arecaceae species and 3 outgroups) based on whole plastomes using bayesian inference. The bayesian posterior probability of all nodes is 1. The branch length is proportional to the inferred divergence level and the scale bar indicates the number of inferred nucleic acids substitutions per site. Hanguana malayana (Commelinales) was used to root the tree (omitted from the figure).

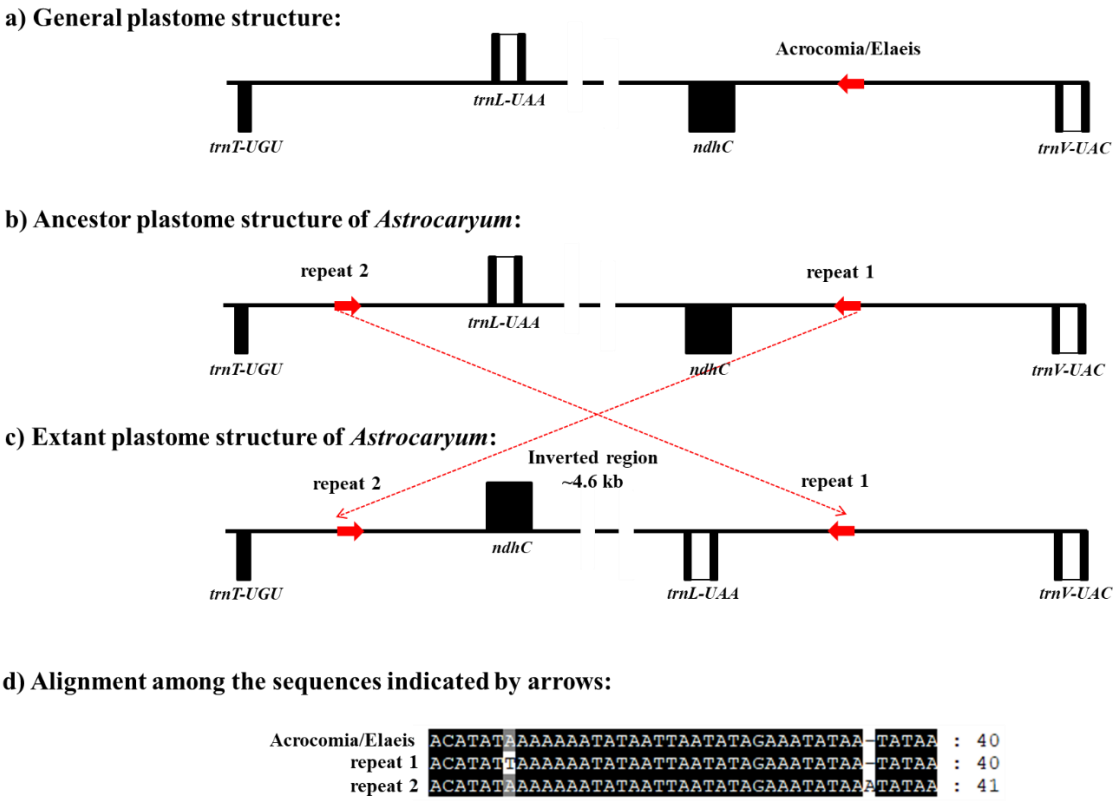


Fig. 6 Hypothesis about the mechanism that generated the inversion in the plastomes of Astrocaryum. **a)** General gene order of most plastomes, including Acrocomia aculeata and Elaeis guineensis that have a conserved sequence (highlighted by red arrow) in the ndhC/trnV-UAC intergenic region. **b)** The ancestor of Astrocaryum plastomes gained inverted repeats with the arising of an inverted sequence in the trnT-UGU/trnL-UAA intergenic region. **c)** A flip-flop rearrangement between the inverted repeats gave rise to the extant plastome structure of Astrocaryum. **d)** The repeats 1 and 2 identified in the plastomes of Astrocaryum species have high identity with a stretch of 40 nucleotides identified in the plastomes of A. aculeata and E. guineensis

TABLES

Table 1. General features of Astrocaryum murumuru and A. aculeatum

	Astrocaryum murumuru	**Astrocaryum aculeatum**
Total size	156,812	156,796
Length of LSC region	85,028	85,046
Length of IR region	27,081	27,081
Length of SSC region	17,622	17,588
GC content (%)	37.42	37.43

Table 2. List of genes identified in the plastomes of Astrocaryum murumuru and A. aculeatum

Group of gene	**Name of gene**
Gene expression machinery	
Ribosomal RNA genes	rrn16[b]; rrn23[b]; rrn5[b]; rrn4.5[b]
Transfer RNA genes	trnA –UGC[ab]; trnC –GCA; trnD –GUC; trnE –UUC; trnF –GAA; trnfM –CAU; trnG –UCC[a]; trnG –GCC; trnH –GUG[b]; trnI –CAU[b]; trnI –GAU[ab]; 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
Small subunit of ribosome	rps2; rps3; rps4; rps7[b]; rps8; rps11; rps12[ab]; rps14; rps15; rps16[a]; rps18; rps19[b]
Large subunit of ribosome	rpl2[ab]; rpl14; rpl16[a]; rpl20; rpl22; rpl23[b]; rpl32; rpl33; rpl36
DNA-dependent RNA polymerase	rpoA; rpoB; rpoC1[a]; rpoC2
Translational initiation factor	infA
Intron maturase	matK
Genes for photosynthesis	
Subunits of photosystem I (PSI)	psaA; psaB; psaC; psaI; psaJ; ycf3[a]; ycf4
Subunits of photosystem II (PSII)	psbA; psbB; psbC; psbD; psbE; psbF; psbH; psbI; psbJ; psbK; psbL; psbM; psbN; psbT; psbZ
Subunits of cytochrome b_6f	petA; petB[a]; petD[a]; petG; petL; petN
Subunits of ATP synthase	atpA; atpB; atpE; atpF[a]; atpH; atpI
Subunits of NADH dehydrogenase	ndhA[a]; ndhB[ab]; ndhC; ndhD; ndhE; ndhF; ndhG; ndhH; ndhI; ndhJ; ndhK
Large subunit of Rubisco	rbcL
Other functions	
Envelope membrane protein	cemA
Subunit of acetyl-CoA carboxylase	accD
C-type cytochrome synthesis	ccsA
Subunit of protease Clp	clpP[a]
Component of TIC complex	ycf1[c]
Unknown function	ycf2[b]

[a]Genes containing introns; [b]Duplicated gene; [c]Partially duplicated genes

Table 3. Location of polymorphic SSR loci in the plastomes of Astrocaryum murumuru and A. aculeatum

	Astrocaryum murumuru					Astrocaryum aculeatum				
Location	**Type**	**Sequence**	**Size**	**Start**	**End**	**Type**	**Sequence**	**Size**	**Start**	**End**
psbA (CDS)	tri	(CAG)4	12	710	721	-	-	-	-	-
trnK-UUU (intron)	mono	(T)12	12	3794	3805	mono	(T)10	10	3818	3827
trnK-UUU/rps16 (IGS)	mono	(T)9	9	4843	4851	mono	(T)11	11	4863	4873
rps16/trnQ-UUG (IGS)	mono	(T)9	9	6610	6618	mono	(T)10	10	6632	6641
rps16/trnQ-UUG (IGS)	mono	(A)12	12	6930	6941	mono	(A)11	11	6952	6962
trnQ-UUG/psbK (IGS)	mono	(T)10	10	7459	7468	mono	(T)9	9	7480	7488
trnS-GCU/trnG-UCC (IGS)	-	-	-	-	-	hexa	(TCCCCA)3	18	8475	8492
atpF/atpH (IGS)	mono	(T)9	9	13444	13452	mono	(T)10	10	13470	13479
atpH/atpI (IGS)	mono	(T)10	10	14777	14786	mono	(T)11	11	14804	14814
rps2/rpoC2 (IGS)	-	-	-	-	-	mono	(T)12	12	16699	16710
rpoC1 (intron)	mono	(T)14	14	22970	22983	mono	(T)12	12	23008	23019
rpoC1 (intron)	mono	(T)10	10	23294	23303	mono	(T)11	11	23324	23334
rpoB/trnC-GCA (IGS)	-	-	-	-	-	mono	(A)8	8	27474	27481
petN/psbM (IGS)	mono	(T)15	15	29413	29427	mono	(T)13	13	29434	29446
petN/psbM (IGS)	di	(AT)4	8	29453	29460	-	-	-	-	-
trnT-GGU/psbD (IGS)	-	-	-	-	-	mono	(T)8	8	32206	32213
trnfM-CAU/rps14 (IGS)	mono	(A)9	9	36608	36616	mono	(A)10	10	36644	36653
psbE/petL (IGS)	mono	(T)9	9	65929	65937	mono	(T)10	10	65966	65975
psbE/petL (IGS)	mono	(A)11	11	66037	66047	mono	(A)12	12	66075	66086
clpP (intron)	mono	(T)10	10	71631	71640	mono	(T)11	11	71635	71645
psbB/psbT (IGS)	mono	(T)11	11	74856	74866	mono	(T)10	10	74861	74870
rpl36/infA (IGS)	mono	(T)10	10	80700	80709	mono	(T)9	9	80704	80712
rpl16 (intron)	mono	(T)9	9	83279	83287	mono	(T)10	10	83284	83293
ndhF/rpl32 (IGS)	mono	(A)12	12	114475	114486	mono	(A)9	9	114493	114501
rpl32/trnL-UAG (IGS)	-	-	-	-	-	mono	(T)8	8	115215	115222
rpl32/trnL-UAG (IGS)	-	-	-	-	-	mono	(A)10	10	115227	115236
ycf1 (CDS)	mono	(T)11	11	126531	126541	mono	(T)10	10	126515	126524

Table 4. List of synonymous (S) and non-synonymous (N) substitutions in plastid genes of Astrocaryum murumuru and A. aculeatum

Gene	S/N	Nt position	AA position	A. murumuru – A. aculeatum	
				Codon change	AA change
atpF	N	256	86	CGG-TGG	R-W
ccsA	N	62	21	GCG-GTG	A-V
matK	S	612	204	TCG-TCT	S-S
	N	690	230	TTC-TTA	F-L
ndhA	N	280	94	GAT-TAT	D-Y
	N	611	204	TTG-TGG	L-W
	S	753	251	ATA-ATC	I-I
ndhF	S	54	18	GTG-GTT	V-V
	S	246	82	TCC-TCT	S-S
	N	1695	565	ATA-ATG	I-M
	S	1842	614	ATA-ATT	I-I
petN	S	60	20	TCG-TCA	S-S
psbA	S	459	153	GCT-CGA	A-A
psbB	S	561	187	GCA-GCG	A-A
psbK	S	93	317	GCT-GCC	A-A
psbT	S	45	15	GGA-GGG	G-G
rpl32	S	39	13	AAA-AAG	K-K
rpoC2	S	216	72	GTT-GTC	V-V
	S	1254	418	GTA-GTG	V-V
	N	1336	446	AGA-AAA	E-K
	S	2652	884	TCT-TCC	S-S
rps11	N	223	75	ACA-GCA	T-A
	S	249	83	GTA-GTC	V-V
rps15	N	121	41	AAA-GAA	K-E
ycf2	N	3625	1209	GGG-CGG	G-R
ycf1	S	828	276	GAG-GAA	E-E
	N	1882	628	AAA-GAA	K-E
	N	1927	643	CTT-GTT	L-V
	N	1964	655	TTG-TCG	L-S
	S	2280	760	CTT-CTC	L-L
	S	2658	886	GAT-GAC	D-D
	N	3215	1072	CGA-CTA	R-L
	N	4005	1335	TTA-TTT	L-F
	N	4049	1350	AGT-AAT	S-N
	N	4051/4052	1351	CCT-TAT	P-Y
	N	4054/4055/4056	1352	TCC-GGG	S-G
	N	4057	1353	TAT-AAT	Y-N
	N	4060/4061/4062	1354	TCC-AGG	S-R
	N	4063/4065	1355	CAT-AAG	H-K
	N	4066/4068	1356	AAT-GAC	N-D
	N	4325/4326	1442	AAA-AGT	K-S
	N	4327/4328/4329	1443	CAC-GTT	H-V
	N	4474	1492	CAT-GAT	H-D
	N	4531	1511	AAA-CAA	K-Q

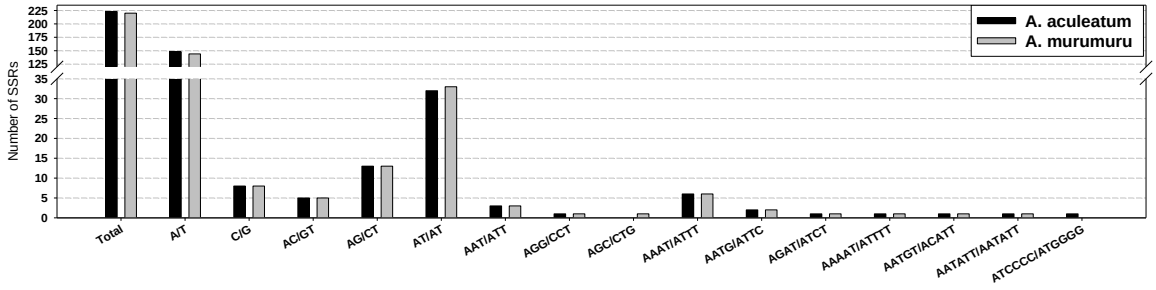
Table 5. List of RNA editing sites predicted in plastid genes of Astrocaryum murumuru and A. aculeatum

Gene	Nt Pos	AA Pos	Effect	Observations based on Lopes et al. (2018a)	Huang et al. (2013)*
accD	154	52	CGG (R) => UGG (W)	conserved in the subfamily Arecoideae	-
	794	265	UCG (S) => UUG (L)	conserved in the subfamily Arecoideae	+
	1157	386	UCA (S) => UUA (L)	conserved in the subfamily Arecoideae	+
	1159	387	CAU (H) => UAU (Y)	conserved in the tribe Cocoseae	-
	1403	468	CCU (P) => CUU (L)	conserved in the subfamily Arecoideae	-
atpA	914	305	UCA (S) => UUA (L)	conserved in the subfamily Arecoideae	+
	1148	383	UCA (S) => UUA (L)	conserved in the subfamily Arecoideae	+/-
atpB	1184	395	UCA (S) => UUA (L)	conserved in the subfamily Arecoideae	+
atpF	92	31	CCA (P) => CUA (L)	conserved in the subfamily Arecoideae	+/-
atpI	428	143	CCC (P) => CUC (L)	conserved in the subfamily Arecoideae	+
	629	210	UCA (S) => UUA (L)	conserved in the subfamily Arecoideae	+
ccsA	62	21	GCG (A) => GUG (V)	present only in A. murumuru; the other species of the subfamily Arecoideae has a T fixed	T
	647	216	ACU (T) => AUU (I)	conserved in the subfamily Arecoideae	-
clpP	559	187	CAU (H) => UAU (Y)	conserved in the subfamily Arecoideae	+
matK	188	63	UCA (S) => UUA (L)	conserved in the subfamily Arecoideae	-
	653	218	CCA (P) => CUA (L)	conserved in the subfamily Arecoideae; exception: lost in A. aculeata	-
	919	307	CAU (H) => UAU (Y)	conserved in the subfamily Arecoideae	-
	1267	423	CAC (H) => UAC (Y)	conserved in the subfamily Arecoideae	+
ndhA	50	17	UCG (S) => UUG (L)	conserved in the subfamily Arecoideae	+
	476	159	UCA (S) => UUA (L)	conserved in the subfamily Arecoideae	+
	566	189	UCA (S) => UUA (L)	conserved in the subfamily Arecoideae	+
	961	321	CCU (P) => UCU (S)	conserved in the subfamily Arecoideae	+
	1073	358	UCC (S) => UUC (F)	conserved in the subfamily Arecoideae	-
ndhB	149	50	UCA (S) => UUA (L)	conserved in the subfamily Arecoideae	+/-
	467	156	CCA (P) => CUA (L)	conserved in the subfamily Arecoideae	+
	542	181	ACG (T) => AUG (M)	conserved in the subfamily Arecoideae	+
	586	196	CAU (H) => UAU (Y)	conserved in the subfamily Arecoideae	+
	704	235	UCC (S) => UUC (F)	conserved in the subfamily Arecoideae	+
	737	246	CCA (P) => CUA (L)	conserved in the subfamily Arecoideae	+
	830	277	UCA (S) => UUA (L)	conserved in the subfamily Arecoideae	+
	836	279	UCA (S) => UUA (L)	conserved in the subfamily Arecoideae	+
	1112	371	UCA (S) => UUA (L)	conserved in the subfamily Arecoideae	+
	1193	398	UCA (S) => UUA (L)	conserved in the subfamily Arecoideae	+
	1255	419	CAU (H) => UAU (Y)	conserved in the subfamily Arecoideae	+
	1481	494	CCA (P) => CUA (L)	conserved in the subfamily Arecoideae	+/-
ndhD	2	1	ACG (T) => AUG (M)	conserved in the subfamily Arecoideae	+/-
	59	20	UCA (S) => UUA (L)	conserved in the subfamily Arecoideae	+
	383	128	UCA (S) => UUA (L)	conserved in the subfamily Arecoideae	+
	674	225	UCG (S) => UUG (L)	conserved in the subfamily Arecoideae	+
	947	316	ACA (T) => AUA (I)	conserved in the subfamily Arecoideae	+
	1193	398	UCA (S) => UUA (L)	conserved in the subfamily Arecoideae	+
	1310	437	UCA (S) => UUA (L)	conserved in the subfamily Arecoideae	+
ndhF	62	21	UCA (S) => UUA (L)	conserved in the subfamily Arecoideae	+/-
	290	97	UCA (S) => UUA (L)	conserved in the subfamily Arecoideae	+/-
	392	131	UCC (S) => UUC (F)	conserved in the subfamily Arecoideae; exception: lost in V. arecina and A. vestiaria	+
	442	148	CAU (H) => UAU (Y)	conserved in the subfamily Arecoideae; exception: lost in V. arecina	+
	586	196	CUU (L) => UUU (F)	conserved in the subfamily Arecoideae	-
	1393	465	CAC (H) => UAC (Y)	conserved in the subfamily Arecoideae	-
	2093	698	UCC (S) => UUC (F)	conserved in the subfamily Arecoideae	-
ndhG	314	105	ACA (T) => AUA (I)	conserved in the subfamily Arecoideae	-
	347	116	CCA (P) => CUA (L)	conserved in the subfamily Arecoideae	+
ndhH	505	169	CAU (H) => UAU (Y)	conserved in the subfamily Arecoideae	+
	545	182	UCU (S) => UUU (F)	conserved in the subfamily Arecoideae; exception: lost in A. vestiaria	+/-
ndhK	131	44	UCG (S) => UUG (L)	conserved in the subfamily Arecoideae	+
petB	418	140	CGG (R) => UGG (W)	conserved in the subfamily Arecoideae	+
	611	204	CCA (P) => CUA (L)	conserved in the subfamily Arecoideae	+
psaI	80	27	UCU (S) => UUU (F)	conserved in the subfamily Arecoideae	+
	85	29	CAU (H) => UAU (Y)	conserved in the subfamily Arecoideae	+
rpl2	2	1	ACG (T) => AUG (M)	conserved in the subfamily Arecoideae	-
rpl20	26	9	ACA (T) => AUA (I)	conserved in the subfamily Arecoideae	-
	308	103	UCA (S) => UUA (L)	conserved in the subfamily Arecoideae	-
rpl23	71	24	UCU (S) => UUU (F)	conserved in the subfamily Arecoideae	+/-
	89	30	UCA (S) => UUA (L)	conserved in the subfamily Arecoideae	+/-
rpoA	200	67	UCU (S) => UUU (F)	conserved in the subfamily Arecoideae	+
	368	123	UCA (S) => UUA (L)	conserved in the subfamily Arecoideae	+
	527	176	UCC (S) => UUC (F)	conserved in the subfamily Arecoideae	+
	830	277	UCA (S) => UUA (L)	conserved in the subfamily Arecoideae	+

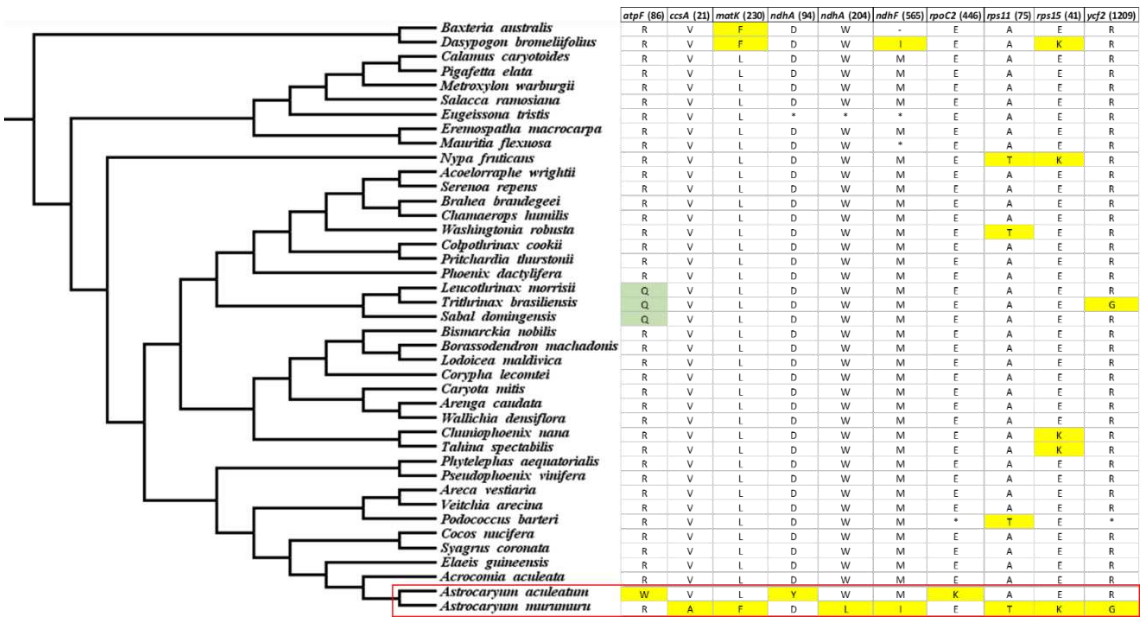
(*) Huang et al. (2013) validated by RT-PCR and sequencing the RNA editing sites in plastid genes of C. nucifera
(+) Presence of RNA editing in all transcripts
(+/-) Presence or RNA editing in part of the transcripts
(-) Absence of RNA editing in all transcripts

SUPPLEMENTARY MATERIAL

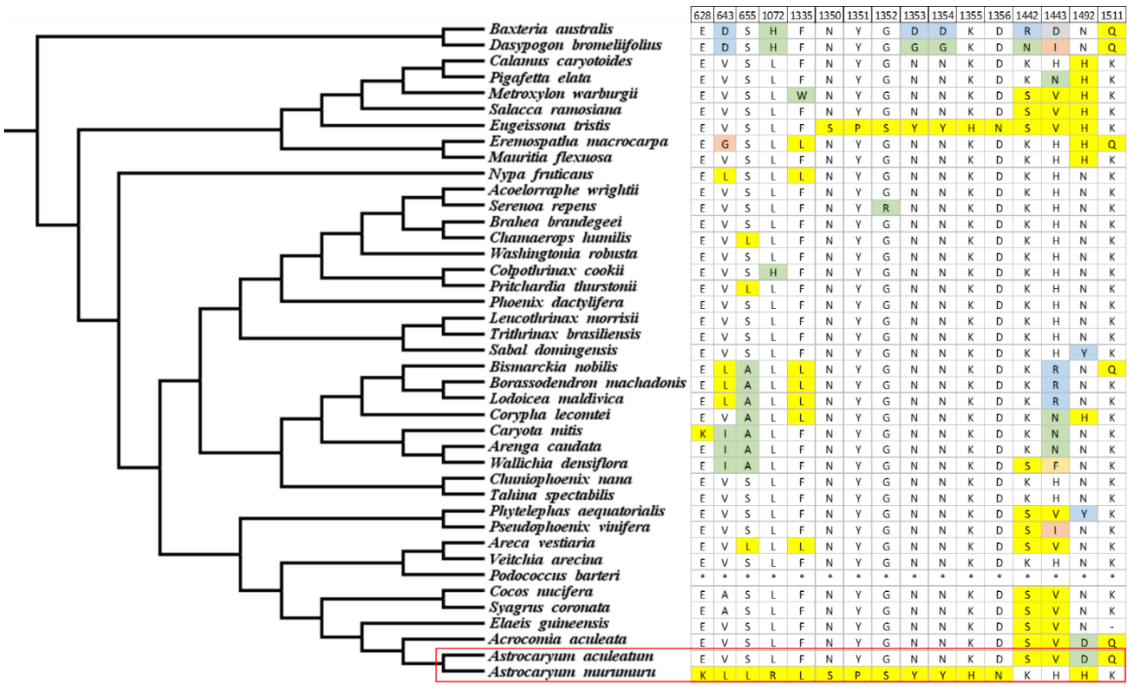
Supplementary Figures



Supplementary Fig. S1. Number of SSR loci in the plastomes of Astrocaryum murumuru and A. aculeatum (IRB omitted)



Supplementary Fig. S2. Non-synonymous substitutions in conserved sites. The amino acids are plotted across the palm phylogeny based on whole plastomes. Different amino acid types identified at the same position are showed in distinct colors. The two species of Astrocaryum are highlighted by red square. The amino acid positions are relative to Astrocaryum



Supplementary Fig. S3. Non-synonymous substitutions in the ycf1 gene. The amino acids are plotted across the palm phylogeny based on whole plastomes. Different amino acid types identified at the same position are showed in distinct colors. The two species of Astrocaryum are highlighted by red square. The amino acid positions are relative to Astrocaryum

Supplementary Tables

Supplementary Table S1. List of species used in the phylogenomic analysis

Species	Subfamily	Family	Order	GenBank
Acrocomia aculeata (Jacq.) Lodd. ex Mart.	Arecoideae	Arecaceae	Arecales	MG020488
Areca vestiaria Giseke	Arecoideae	Arecaceae	Arecales	NC_029972.1
Cocos nucifera L.	Arecoideae	Arecaceae	Arecales	NC_022417.1
Elaeis guineensis Jacq.	Arecoideae	Arecaceae	Arecales	NC_017602.1
Podococcus barteri G.Mann & H.Wendl.	Arecoideae	Arecaceae	Arecales	NC_027276.1
Syagrus coronata (Mart.) Becc.	Arecoideae	Arecaceae	Arecales	NC_029241.1
Veitchia arecina Becc.	Arecoideae	Arecaceae	Arecales	NC_029950.1
Calamus caryotoides A.Cunn. ex Mart.	Calamoideae	Arecaceae	Arecales	NC_020365.1
Eremospatha macrocarpa H.Wendl.	Calamoideae	Arecaceae	Arecales	NC_029964.1
Eugeissona tristes Griff.	Calamoideae	Arecaceae	Arecales	NC_029963.1
Mauritia flexuosa L.f.	Calamoideae	Arecaceae	Arecales	NC_029947.1
Metroxylon warburgii (Heimerl) Becc.	Calamoideae	Arecaceae	Arecales	NC_029959.1
Pigafetta elata (Mart.) H.Wendl.	Calamoideae	Arecaceae	Arecales	NC_029956.1
Salacca ramosiana Mogea	Calamoideae	Arecaceae	Arecales	NC_029954.1
Phytelephas aequatorialis Spruce	Ceroxyloideae	Arecaceae	Arecales	NC_029957.1
Pseudophoenix vinifera (Mart.) Becc.	Ceroxyloideae	Arecaceae	Arecales	NC_020364.1
Acoelorraphe wrightii (Griseb. & H.Wendl.) H.Wendl. ex Becc.	Coryphoideae	Arecaceae	Arecales	NC_029973.1
Arenga caudata (Lour.) H.E.Moore	Coryphoideae	Arecaceae	Arecales	NC_029971.1
Bismarckia nobilis Hildebr. & H.Wendl.	Coryphoideae	Arecaceae	Arecales	NC_020366.1
Borassodendron machadonis (Ridl.) Becc.	Coryphoideae	Arecaceae	Arecales	NC_029969.1
Brahea brandegeei (Purpus) H.E.Moore	Coryphoideae	Arecaceae	Arecales	NC_029968.1
Caryota mitis Lour.	Coryphoideae	Arecaceae	Arecales	NC_029948.1
Chamaerops humilis L.	Coryphoideae	Arecaceae	Arecales	NC_029967.1
Chuniophoenix nana Burret	Coryphoideae	Arecaceae	Arecales	NC_029966.1
Colpothrinax cookii Read	Coryphoideae	Arecaceae	Arecales	NC_028026.1
Corypha lecomtei Becc. ex Lecomte	Coryphoideae	Arecaceae	Arecales	NC_029965.1
Leucothrinax morrisii (H.Wendl.) C.Lewis & Zona	Coryphoideae	Arecaceae	Arecales	NC_029961.1
Lodoicea maldivica (J.F.Gmel.) Pers.	Coryphoideae	Arecaceae	Arecales	NC_029960.1
Phoenix dactylifera L.	Coryphoideae	Arecaceae	Arecales	NC_013991.2
Pritchardia thurstonii (F.Muell.) Drude	Coryphoideae	Arecaceae	Arecales	NC_029955.1
Sabal domingensis Becc.	Coryphoideae	Arecaceae	Arecales	NC_026444.1
Serenoa repens (W.Bartram) Small	Coryphoideae	Arecaceae	Arecales	NC_029953.1
Tahina spectabilis J.Dransf. & Rakotoarin.	Coryphoideae	Arecaceae	Arecales	NC_029952.1
Trithrinax brasiliensis Mart.	Coryphoideae	Arecaceae	Arecales	NC_029951.1
Wallichia densiflora Mart.	Coryphoideae	Arecaceae	Arecales	NC_029949.1
Washingtonia robusta H.Wendl.	Coryphoideae	Arecaceae	Arecales	NC_029974.1
Nypa fruticans Wurmb	Nypoideae	Arecaceae	Arecales	NC_029958.1
Baxteria australis R. Br. ex Hook.	-	Dasypogonaceae	Arecales	NC_029970.1
Dasypogon bromeliifolius R. Br.	-	Dasypogonaceae	Arecales	NC_020367.1
Hanguana malayana (Jack) Merr.	-	Hanguanaceae	Commelinales	NC_029962.1

Supplementary Table S2. List of conservative SSR loci identified in the plastomes of Astrocaryum murumuru and A. aculeatum

			A. murumuru		A. aculeatum		
Type	**Sequence**	**Size**	**Start**	**End**	**Start**	**End**	**Location**
di	(AT)4	8	1313	1320	1337	1344	psbA/trnK-UUU (IGS)
mono	(A)9	9	2624	2632	2648	2656	matK (CDS)
mono	(T)9	9	2870	2878	2894	2902	matK (CDS)
mono	(A)9	9	3594	3602	3618	3626	trnK-UUU (intron)
mono	(A)8	8	3819	3826	3841	3848	trnK-UUU (intron)
mono	(A)8	8	3972	3979	3992	3999	trnK-UUU (intron)
mono	(A)8	8	4633	4640	4653	4660	trnK-UUU/rps16 (IGS)
mono	(C)9	9	4834	4842	4854	4862	trnK-UUU/rps16 (IGS)
di	(GT)4	8	6070	6077	6092	6099	rps16/trnQ-UUG (IGS)
tretra	(TCTA)5	20	6104	6123	6126	6145	rps16/trnQ-UUG (IGS)
hexa	(TTAATA)3	18	6339	6356	6361	6378	rps16/trnQ-UUG (IGS)
mono	(T)9	9	7408	7416	7429	7437	trnQ-UUG/psbK (IGS)
mono	(A)9	9	7476	7484	7496	7504	trnQ-UUG/psbK (IGS)
mono	(T)8	8	7667	7674	7687	7694	psbK (CDS)
mono	(A)8	8	8092	8099	8112	8119	psbK/psbI (IGS)
mono	(A)9	9	8227	8235	8247	8255	psbI/trnS-GCU (IGS)
mono	(T)9	9	8267	8275	8287	8295	psbI/trnS-GCU (IGS)
di	(GA)4	8	8372	8379	8392	8399	trnS-GCU
di	(AT)4	8	8537	8544	8563	8570	trnS-GCU/trnG-UCC (IGS)
di	(AT)7	14	8558	8571	8584	8597	trnS-GCU/trnG-UCC (IGS)

mono	(T)9	9	9619	9627	9645	9653	trnG-UCC (intron)
mono	(T)8	8	11907	11914	11933	11940	atpA/atpF (IGS)
mono	(T)9	9	12520	12528	12546	12554	atpF (intron)
mono	(T)8	8	12770	12777	12796	12803	atpF (intron)
mono	(A)8	8	13205	13212	13231	13238	atpF (CDS)
mono	(A)9	9	14128	14136	14155	14163	atpH/atpI (IGS)
mono	(T)8	8	14190	14197	14217	14224	atpH/atpI (IGS)
di	(AT)5	10	14556	14565	14583	14592	atpH/atpI (IGS)
di	(AT)7	14	14579	14592	14606	14619	atpH/atpI (IGS)
mono	(A)9	9	15868	15876	15896	15904	atpI/rps2 (IGS)
mono	(A)8	8	18033	18040	18065	18072	rpoC2 (CDS)
mono	(T)10	10	18746	18755	18778	18787	rpoC2 (CDS)
mono	(T)11	11	18852	18862	18884	18894	rpoC2 (CDS)
mono	(A)8	8	18995	19002	19027	19034	rpoC2 (CDS)
di	(AT)4	8	20126	20133	20158	20165	rpoC2 (CDS)
di	(AT)5	10	20216	20225	20248	20257	rpoC2 (CDS)
mono	(A)8	8	22662	22669	22700	22707	rpoC1 (CDS)
di	(TA)4	8	23223	23230	23253	23260	rpoC1 (intron)
di	(TC)4	8	23250	23257	23280	23287	rpoC1 (intron)
mono	(T)8	8	26546	26553	26577	26584	rpoB (CDS)
di	(CA)4	8	26976	26983	27007	27014	rpoB (CDS)
penta	(ATGTA)3	15	27321	27335	27352	27366	rpoB/trnC-GCA (IGS)
mono	(A)9	9	28319	28327	28346	28354	rpoB/trnC-GCA (IGS)
mono	(A)8	8	28752	28759	28773	28780	trnC-GCA/petN (IGS)
mono	(A)9	9	28984	28992	29005	29013	trnC-GCA/petN (IGS)
di	(GT)4	8	28994	29001	29015	29022	trnC-GCA/petN (IGS)
mono	(A)15	15	29373	29387	29394	29408	petN/psbM (IGS)
di	(TA)4	8	29444	29451	29463	29470	petN/psbM (IGS)
mono	(A)9	9	29600	29608	29609	29617	petN/psbM (IGS)
di	(TA)4	8	29690	29697	29699	29706	petN/psbM (IGS)
mono	(T)8	8	30213	30220	30229	30236	psbM/trnD-GUC (IGS)
mono	(T)8	8	31033	31040	31049	31056	trnD-GUC/trnY-GUA (IGS)
mono	(T)9	9	31841	31849	31857	31865	trnT-GGU/psbD (IGS)
mono	(T)9	9	32478	32486	32494	32502	trnT-GGU/psbD (IGS)
di	(TA)4	8	32556	32563	32592	32599	trnT-GGU/psbD (IGS)
mono	(A)8	8	32624	32631	32660	32667	trnT-GGU/psbD (IGS)
mono	(G)8	8	34080	34087	34116	34123	psbC (CDS)
mono	(G)8	8	34351	34358	34387	34394	psbC (CDS)
di	(GA)4	8	35339	35346	35375	35382	trnS-UGA
mono	(T)8	8	35683	35690	35719	35726	trnS-UGA/psbZ (IGS)
mono	(A)8	8	36052	36059	36088	36095	psbZ/trnG-GCC (IGS)
mono	(A)8	8	36439	36446	36475	36482	trnG-GCC/trnfM-CAU (IGS)
mono	(T)11	11	36958	36968	36995	37005	rps14 (CDS)
mono	(A)8	8	37183	37190	37220	37227	rps14/psaB (IGS)
mono	(C)10	10	40150	40159	40187	40196	psaA (CDS)
di	(AG)4	8	41052	41059	41089	41096	psaA (CDS)
penta	(TATTT)3	15	41804	41818	41841	41855	psaA/ycf3 (IGS)
mono	(A)9	9	42837	42845	42874	42882	ycf3 (intron)
mono	(A)9	9	45169	45177	45206	45214	trnS-GGA/rps4 (IGS)
mono	(G)9	9	45437	45445	45474	45482	rps4 (CDS)
di	(TA)4	8	46064	46071	46101	46108	rps4/trnT-UGU (IGS)
mono	(A)8	8	46226	46233	46263	46270	rps4/trnT-UGU (IGS)
di	(TA)5	10	46456	46465	46493	46502	trnT-UGU/ndhC (IGS)
di	(TA)4	8	46808	46815	46845	46852	trnT-UGU/ndhC (IGS)
mono	(T)9	9	46996	47004	47033	47041	trnT-UGU/ndhC (IGS)
mono	(T)8	8	47038	47045	47075	47082	trnT-UGU/ndhC (IGS)
mono	(A)8	8	47174	47181	47211	47218	trnT-UGU/ndhC (IGS)
mono	(T)9	9	49511	49519	49548	49556	ndhJ/trnF-GAA (IGS)
di	(AT)9	18	49521	49538	49558	49575	ndhJ/trnF-GAA (IGS)
di	(AT)4	8	49734	49741	49771	49778	ndhJ/trnF-GAA (IGS)
di	(CT)4	8	50315	50322	50352	50359	trnL-UAA (intron)
di	(AT)8	16	50456	50471	50493	50508	trnL-UAA (intron)

di	(AT)4	8	50992	50999	51029	51036	trnL-UAA/trnV-UAC (IGS)
mono	(T)10	10	51066	51075	51103	51112	trnL-UAA/trnV-UAC (IGS)
tretra	(AAAT)3	12	51220	51231	51257	51268	trnL-UAA/trnV-UAC (IGS)
di	(AT)4	8	51293	51300	51330	51337	trnL-UAA/trnV-UAC (IGS)
di	(TA)4	8	51301	51308	51338	51345	trnL-UAA/trnV-UAC (IGS)
di	(AT)4	8	51322	51329	51359	51366	trnL-UAA/trnV-UAC (IGS)
mono	(A)11	11	51368	51378	51405	51415	trnL-UAA/trnV-UAC (IGS)
mono	(T)9	9	51455	51463	51492	51500	trnL-UAA/trnV-UAC (IGS)
mono	(A)9	9	51752	51760	51789	51797	trnL-UAA/trnV-UAC (IGS)
di	(CA)4	8	52188	52195	52225	52232	trnL-UAA/trnV-UAC (IGS)
mono	(T)9	9	53546	53554	53583	53591	trnM-CAU/atpE (IGS)
mono	(T)8	8	55538	55545	55575	55582	atpB/rbcL (IGS)
mono	(T)8	8	55868	55875	55905	55912	atpB/rbcL (IGS)
di	(AT)4	8	56073	56080	56110	56117	atpB/rbcL (IGS)
mono	(A)9	9	57816	57824	57853	57861	rbcL/accD (IGS)
mono	(C)9	9	58347	58355	58384	58392	rbcL/accD (IGS)
mono	(T)8	8	58879	58886	58916	58923	accD (CDS)
mono	(A)10	10	59182	59191	59219	59228	accD (CDS)
di	(TG)4	8	59611	59618	59648	59655	accD (CDS)
mono	(A)9	9	60062	60070	60099	60107	accD/psaI (IGS)
mono	(A)10	10	60270	60279	60307	60316	accD/psaI (IGS)
mono	(A)8	8	60314	60321	60351	60358	accD/psaI (IGS)
di	(TA)4	8	60441	60448	60478	60485	accD/psaI (IGS)
mono	(T)8	8	61277	61284	61314	61321	ycf4 (CDS)
mono	(A)10	10	61681	61690	61718	61727	ycf4/cemA (IGS)
mono	(A)11	11	61878	61888	61915	61925	cemA (CDS)
di	(TC)5	10	61945	61954	61982	61991	cemA (CDS)
tretra	(AATG)3	12	62555	62566	62592	62603	cemA (CDS)
mono	(C)8	8	63151	63158	63188	63195	petA (CDS)
mono	(A)8	8	64281	64288	64318	64325	petA/psbJ (IGS)
mono	(T)8	8	66847	66854	66886	66893	petG/trnW-CCA (IGS)
mono	(T)8	8	66902	66909	66941	66948	petG/trnW-CCA (IGS)
mono	(A)8	8	67415	67422	67454	67461	trnP-UGG/psaJ (IGS)
mono	(A)8	8	67452	67459	67491	67498	trnP-UGG/psaJ (IGS)
mono	(A)10	10	67592	67601	67631	67640	trnP-UGG/psaJ (IGS)
mono	(T)8	8	67790	67797	67829	67836	psaJ (CDS)
mono	(T)9	9	67880	67888	67919	67927	psaJ/rpl33 (IGS)
di	(AT)7	14	68737	68750	68741	68754	rpl33/rps18 (IGS)
mono	(T)9	9	69303	69311	69307	69315	rps18/rpl20 (IGS)
mono	(T)8	8	69354	69361	69358	69365	rps18/rpl20 (IGS)
mono	(A)9	9	70009	70017	70013	70021	rpl20/rps12 (IGS)
mono	(T)10	10	70046	70055	70050	70059	rpl20/rps12 (IGS)
di	(TA)4	8	70635	70642	70639	70646	rps12/clpP (IGS)
mono	(T)8	8	70705	70712	70709	70716	rps12/clpP (IGS)
mono	(A)8	8	71079	71086	71083	71090	clpP (intron)
mono	(A)8	8	71261	71268	71265	71272	clpP (intron)
mono	(T)10	10	71367	71376	71371	71380	clpP (intron)
tretra	(ATAA)3	12	71829	71840	71834	71845	clpP (CDS)
mono	(A)9	9	72017	72025	72022	72030	clpP (intron)
mono	(T)9	9	72068	72076	72073	72081	clpP (intron)
mono	(A)9	9	72166	72174	72171	72179	clpP (intron)
mono	(A)8	8	72851	72858	72856	72863	clpP/psbB (IGS)
mono	(T)8	8	74031	74038	74036	74043	psbB (CDS)
mono	(A)8	8	75986	75993	75990	75997	petB (intron)
tretra	(AAAT)3	12	76299	76310	76303	76314	petB (intron)
mono	(T)9	9	78821	78829	78825	78833	petD/rpoA (IGS)
tri	(GGA)4	12	81092	81103	81095	81106	infA/rps8 (IGS)
mono	(T)10	10	81574	81583	81577	81586	rps8/rpl14 (IGS)
di	(TA)4	8	82760	82767	82763	82770	rpl16 (intron)
mono	(A)8	8	82896	82903	82894	82901	rpl16 (intron)
tretra	(TTTA)3	12	83300	83311	83298	83309	rpl16 (intron)
mono	(T)18	18	83670	83687	83668	83685	rpl16 (intron)

mono	(T)9	9	85399	85407	85417	85425	rps19 (CDS)
mono	(T)8	8	85435	85442	85453	85460	rps19/trnH-GUG (IGS)
mono	(A)8	8	85650	85657	85668	85675	trnH-GUG/rpl2 (IGS)
di	(GA)4	8	87799	87806	87817	87824	ycf2 (CDS)
di	(GA)4	8	87811	87818	87829	87836	ycf2 (CDS)
di	(GA)4	8	88813	88820	88831	88838	ycf2 (CDS)
mono	(A)8	8	89522	89529	89540	89547	ycf2 (CDS)
mono	(A)8	8	89715	89722	89733	89740	ycf2 (CDS)
mono	(A)9	9	90983	90991	91001	91009	ycf2 (CDS)
di	(TA)4	8	94363	94370	94381	94388	ycf2 (CDS)
di	(TA)4	8	95769	95776	95787	95794	trnL-CAA/ndhB (IGS)
di	(AG)4	8	96510	96517	96528	96535	ndhB (CDS)
mono	(T)8	8	97496	97503	97514	97521	ndhB (intron)
mono	(T)9	9	100450	100458	100468	100476	rps12/trnV-GAC (IGS)
mono	(A)8	8	100617	100624	100635	100642	rps12/trnV-GAC (IGS)
mono	(T)8	8	104483	104490	104501	104508	trnI-GAU (intron)
di	(CT)4	8	107805	107812	107823	107830	rrn23
mono	(T)8	8	111942	111949	111960	111967	ycf1 (CDS)
mono	(A)8	8	112100	112107	112118	112125	ycf1 (CDS)
mono	(A)8	8	112128	112135	112146	112153	ndhF (CDS)
mono	(C)10	10	112162	112171	112180	112189	ndhF (CDS)
mono	(T)10	10	114313	114322	114331	114340	ndhF/rpl32 (IGS)
mono	(A)8	8	114396	114403	114414	114421	ndhF/rpl32 (IGS)
di	(AT)6	12	114414	114425	114432	114443	ndhF/rpl32 (IGS)
mono	(A)8	8	114521	114528	114536	114543	ndhF/rpl32 (IGS)
mono	(A)9	9	115287	115295	115295	115303	rpl32/trnL-UAG (IGS)
mono	(T)8	8	116241	116248	116249	116256	ccsA (CDS)
tretra	(AATA)3	12	117118	117129	117126	117137	ndhD (CDS)
di	(AT)6	12	119104	119115	119095	119106	psaC/ndhE (IGS)
mono	(A)8	8	119182	119189	119173	119180	psaC/ndhE (IGS)
tri	(AAT)4	12	119188	119199	119179	119190	psaC/ndhE (IGS)
tri	(AAT)4	12	119202	119213	119193	119204	psaC/ndhE (IGS)
tretra	(TTTA)3	12	119913	119924	119904	119915	ndhE/ndhG (IGS)
mono	(T)9	9	121108	121116	121099	121107	ndhI (CDS)
di	(AT)5	10	122311	122320	122302	122311	ndhA (intron)
tretra	(ATTC)3	12	122421	122432	122412	122423	ndhA (intron)
mono	(A)8	8	122973	122980	122957	122964	ndhA (intron)
di	(TC)5	10	124381	124390	124365	124374	ndhH (CDS)
mono	(A)9	9	124879	124887	124863	124871	ndhH/rps15 (IGS)
di	(AT)4	8	125592	125599	125576	125583	ycf1 (CDS)
mono	(T)8	8	125986	125993	125970	125977	ycf1 (CDS)
mono	(T)11	11	126408	126418	126392	126402	ycf1 (CDS)
mono	(T)8	8	126746	126753	126721	126728	ycf1 (CDS)
tri	(ATA)4	12	127508	127519	127492	127503	ycf1 (CDS)
mono	(T)8	8	127663	127670	127647	127654	ycf1 (CDS)
mono	(T)9	9	127942	127950	127926	127934	ycf1 (CDS)
mono	(T)9	9	128155	128163	128139	128147	ycf1 (CDS)
mono	(T)10	10	128197	128206	128181	128190	ycf1 (CDS)
mono	(T)12	12	128318	128329	128302	128313	ycf1 (CDS)
mono	(T)9	9	128690	128698	128674	128682	ycf1 (CDS)
mono	(T)11	11	128706	128716	128690	128700	ycf1 (CDS)
di	(TC)4	8	128766	128773	128750	128757	ycf1 (CDS)
mono	(A)10	10	128824	128833	128808	128817	ycf1 (CDS)
mono	(T)8	8	129677	129684	129661	129668	ycf1 (CDS)

Supplementary Table S3. List of dispersed repeats in the plastomes of Astrocaryum murumuru and A. aculeatum

Astrocaryum murumuru

		Repeat 1		Repeat 2		
Type	**Size**	**Nt position**	**Location**	**Nt position**	**Location**	**E-value**
P	35	46742	trnT-UGU/ndhC (IGS)	51329	trnL-UAA/trnV-UAC (IGS)	4.21e-10
F	37	43483	ycf3 (intron)	100079	rps12/trnV-GAC (IGS)	5.26e-08
P	34	8364	trnS-GCU	44964	trnS-GGA	8.10e-08
F	32	38558	psaB (CDS)	40782	psaA (CDS)	2.59e-06
F	32	8366	trnS-GCU	35333	trnS-UGA	3.44e-05
P	31	35333	trnS-UGA	44964	trnS-GGA	3.44e-05
F	31	3956	trnK-UUU (intron)	119169	psaC/ndhE (IGS)	1.25e-04
P	30	46736	trnT-UGU/ndhC (IGS)	60450	accD/psaI (IGS)	1.25e-04
F	30	3957	trnK-UUU (intron)	119167	psaC/ndhE (IGS)	4.50e-04
F	30	46992	trnT-UGU/ndhC (IGS)	83275	rpl16 (intron)	4.50e-04

Astrocaryum aculeatum

		Repeat 1		Repeat 2		
Type	**Size**	**Nt position**	**Location**	**Nt position**	**Location**	**E-value**
P	35	46779	trnT-UGU/ndhC (IGS)	51366	trnL-UAA/trnV-UAC (IGS)	4.21e-10
F	37	43520	ycf3 (intron)	100097	rps12/trnV-GAC (IGS)	5.26e-08
P	34	8384	trnS-GCU	45001	trnS-GGA	8.10e-08
F	34	38595	psaB (CDS)	40819	psaA (CDS)	2.59e-06
F	32	8386	trnS-GCU	35369	trnS-UGA	3.44e-05
P	32	35369	trnS-UGA	45001	trnS-GGA	3.44e-05
F	31	3976	trnK-UUU	119160	psaC/ndhE (IGS)	1.25e-04
F	31	29428	petN/psbM (IGS)	47026	trnT-UGU/ndhC (IGS)	1.25e-04
P	31	60476	accD/psaI (IGS)	82707	rpl16 (intron)	1.25e-04
P	30	3969	trnK-UUU	83256	rpl16 (intron)	4.50e-04
F	30	3977	trnK-UUU	119158	psaC/ndhE (IGS)	4.50e-04

Supplementary Table S4. List of indels identified in the plastomes of Astrocaryum murumuru and A. aculeatum. The nucleotide positions showed are based on the positions in the alignment site

Indel position	Length	Location
1	24	rps19/psbA (IGS)
3818	2	trnK-UUU (intron)
3920	8	trnK-UUU (intron)
3977	10	trnK-UUU (intron)
4875	2	trnK-UUU/rps16 (IGS)
6644	1	rps16/trnG-UUG (IGS)
6964	2	rps16/trnG-UUG (IGS)
7494	1	trnQ-UUG/psbK (IGS)
8495	6	trnS-GCU/trnG-UCC (IGS)
13485	1	atpF/atpH (IGS)
14819	1	atpH/atpI (IGS)
16714	4	rps2/rpoC2 (IGS)
21094	6	rpoC2/rpoC1 (IGS)
23035	8	rpoC1 (intron)
23356	1	rpoC1 (intron)
27712	4	rpoB/trnC-GCA (IGS)
28792	6	trnC-GCA/petN (IGS)
29467	2	petN/psbM (IGS)
29498	10	petN/psbM (IGS)
30223	7	psbM/trnD-GUC (IGS)
32563	20	trnT-GGU/psbD (IGS)
36689	1	trnfM-CAU/rps14 (IGS)
66011	1	psbE/petL (IGS)
66120	1	psbE/petL (IGS)
68238	35	psaJ/rpl33 (IGS)
71715	1	clpP (intron)
74941	1	psbB/psbT (IGS)
80785	1	rpl36/infA (IGS)
82903	5	rpl16 (intron)
83982	20	rps3 (CDS)
114580	3	ndhF/rpl32 (IGS)
115251	10	rpl32/trnL-UAG (IGS)
115324	3	rpl32/trnL-UAG (IGS)
119091	17	psaC/ndhE (IGS)
122600	7	ndhA (intron)

Supplementary Table S5. List of SNPs identified in the plastomes of Astrocaryum murumuru and A. aculeatum. The nucleotide positions showed are based on the positions in the alignment site

SNP position	Location	Change	SNP position	Location	Change
26	rps19/psbA (IGS)	A-T	83370	rpl16 (intron)	T-A
28	rps19/psbA (IGS)	T-A	83373	rpl16 (intron)	C-T
73	rps19/psbA (IGS)	T-C	91490	ycf2 (CDS)	G-C
74	rps19/psbA (IGS)	A-T	96262	trnL-CAA/ndhB (IGS)	A-C
75	rps19/psbA (IGS)	T-A	109607	rrn5/trnR-ACG (IGS)	T-A
76	rps19/psbA (IGS)	C-A	109641	rrn5/trnR-ACG (IGS)	A-C
77	rps19/psbA (IGS)	T-A	110334	trnR-ACG/trnN-GUU (IGS)	A-G
78	rps19/psbA (IGS)	T-A	110410	trnR-ACG/trnN-GUU (IGS)	G-A
79	rps19/psbA (IGS)	G-C	111702	ycf1 (CDS)	G-A
80	rps19/psbA (IGS)	T-A	112555	ndhF (CDS)	T-A
81	rps19/psbA (IGS)	T-A	112702	ndhF (CDS)	T-C
82	rps19/psbA (IGS)	T-G	114151	ndhF (CDS)	G-A
83	rps19/psbA (IGS)	T-A	114343	ndhF (CDS)	C-A
84	rps19/psbA (IGS)	A-T	114740	rpl32 (CDS)	A-G
85	rps19/psbA (IGS)	G-A	115322	rpl32/trnL-UAG (IGS)	C-T
744	psbA (CDS)	A-T	115328	rpl32/trnL-UAG (IGS)	T-A
1233	psbA/trnK-UUU (IGS)	A-G	115363	rpl32/trnL-UAG (IGS)	G-T
2580	matK (CDS)	G-T	115364	rpl32/trnL-UAG (IGS)	A-T
2658	matK (CDS)	C-A	115365	rpl32/trnL-UAG (IGS)	A-C
4823	trnK-UUU/rps16 (IGS)	A-G	115901	ccsA (CDS)	C-T
7658	psbK (CDS)	T-C	116979	ccsA/ndhD (IGS)	T-A
7790	psbK/psbI (IGS)	A-C	118630	ndhD/psaC (IGS)	A-G
8280	psbI/trnS-GCU (IGS)	C-A	119257	psaC/ndhE (IGS)	C-A
8511	trnS-GCU/trnG-UCC (IGS)	T-A	119273	psaC/ndhE (IGS)	C-A
8766	trnS-GCU/trnG-UCC (IGS)	C-A	120840	ndhG/ndhI (IGS)	C-A
12321	atpF (CDS)	G-A	121919	ndhA (CDS)	T-G
15921	atpI/rps2 (IGS)	A-G	122061	ndhA (CDS)	A-C
16722	rps2/rpoC2 (IGS)	C-T	122243	ndhA (intron)	C-T
18372	rpoC2 (CDS)	A-G	123132	ndhA (intron)	G-A
19688	rpoC2 (CDS)	C-T	123442	ndhA (CDS)	C-A
19770	rpoC2 (CDS)	T-C	125162	rps15 (CDS)	T-C
20808	rpoC2 (CDS)	A-G	125582	rps15/ycf1 (IGS)	T-C
23044	rpoC1 (intron)	A-C	126649	ycf1 (CDS)	T-G
23300	rpoC1 (intron)	G-A	126706	ycf1 (CDS)	G-C
23475	rpoC1 (intron)	C-A	126851	ycf1 (CDS)	G-A
27503	rpoB/trnC (IGS)	G-A	126852	ycf1 (CDS)	T-A
29015	trnC-GCA/petN (IGS)	G-A	126853	ycf1 (CDS)	G-C
29323	petN (CDS)	G-A	126854	ycf1 (CDS)	T-A
29483	petN/psbM (IGS)	G-T	126855	ycf1 (CDS)	T-C
31689	trnE-UUC/trnT-GGU (IGS)	T-C	127112	ycf1 (CDS)	A-G
31694	trnE-UUC/trnT-GGU (IGS)	G-A	127114	ycf1 (CDS)	T-C
32258	trnT-GGU/psbD (IGS)	G-T	127115	ycf1 (CDS)	A-C
32687	trnT-GGU/psbD (IGS)	T-A	127117	ycf1 (CDS)	G-T
60468	accD/psaI (IGS)	T-G	127118	ycf1 (CDS)	G-C
60548	accD/psaI (IGS)	A-T	127119	ycf1 (CDS)	G-C
60550	accD/psaI (IGS)	A-T	127120	ycf1 (CDS)	A-T
65664	psbE/petL (IGS)	A-G	127123	ycf1 (CDS)	A-T
66473	psbE/petL (IGS)	G-T	127124	ycf1 (CDS)	G-C
67947	psaJ/rpl33 (IGS)	A-C	127125	ycf1 (CDS)	G-C
68277	psaJ/rpl33 (IGS)	A-T	127126	ycf1 (CDS)	A-C
68280	psaJ/rpl33 (IGS)	T-C	127128	ycf1 (CDS)	G-T
68283	psaJ/rpl33 (IGS)	A-T	127129	ycf1 (CDS)	G-A
69364	rps18/rpl20 (IGS)	A-C	127131	ycf1 (CDS)	C-T
70238	rpl20/rps12 (IGS)	T-C	127175	ycf1 (CDS)	T-A
73943	psbB (CDS)	A-G	127965	ycf1 (CDS)	C-A
75149	psbT (CDS)	A-G	128522	ycf1 (CDS)	A-G
80224	rps11 (CDS)	T-G	128900	ycf1 (CDS)	A-G
80250	rps11 (CDS)	T-C	129216	ycf1 (CDS)	A-G
83362	rpl16 (intron)	A-T	129253	ycf1 (CDS)	G-C
83367	rpl16 (intron)	T-A	129298	ycf1 (CDS)	T-C

Here, it is reported the complete plastomes of the oilseed and oil palm species Linum ustitatissimum (Linaceae), Crambe abyssinica (Brassicaceae), Acrocomia aculeata, Astrocaryum murumuru, and A. aculeatum (Arecaceae). In the plastomes of all families represented here it was identified unique evolutionary traits. Structurally, the plastome of L. usitatissimum underwent large expansion and contraction events in the borders of the IRs, which changed gene order and gene copy number. Similarly, within the family Arecaceae, the species of genus Astrocaryum bear a 4.6-kb inversion, which is a result from a flip-flop recombination event between short inverted repeat sequences. Since the plastome rearrangements are rare events, these unique structural modifications can be useful synapomorphies within the families Linaceae and Arecaceae.

Concerning the gene content, the plastome of L. usitatissimum also presents uncommon features such as gene loss (rsp16), pseudogenes (ndhF, rpl23), loss of introns (clpP), and highly divergent genes (ycf1, ycf2, and clpP). ycf1 and rps16 genes are also among the most divergent genes within the families Brassicaceae and Arecaceae. In Brassicaceae the rps16 gene was lost in many species, but in others, including C. abyssinica, although theoretically still functional, it shows evidences for degeneration. Additionally, more than half of the plastid genes in Arecaceae bear signatures of positive selection, including several photosynthesis-related genes, indicating that subtle changes are selected as adaptive response to different environmental cues. The features of plastid genes of the plastomes sequenced here also included events of gain and loss of RNA editing sites, inclusive specifically at the genera level, which indicates a relatively high evolutionary rate in the RNA editing machinery.

Taken together, the exceptional features concerning structure and gene content of L. usitatissimum indicates the family Linaceae as an interesting lineage to study evolutionary traits in plastids, since this family has a high morphological and ecological diversity. Within the family Arecaceae, the high incidence of positive selection raise question about the role of plastid genes in the adaptation to specific contrasting environmental conditions, an intriguing topic for future studies. In Brassicaceae, the plastome of C. abyssinica confirms the high conservation of plastome sequence and structure within this family, although minor changes concerning the gene content and RNA editing have been also identified.

Moreover, the phylogenies presented here, based on concatenated plastid genes and specially based on whole plastomes, resulted in well-supported trees and were able

to resolve deep and close relationships. Furthermore, a plenty of plastid molecular markers were mapped, providing genetic information useful for several genetic studies aiming to stablish strategies for genetic breeding, domestication, and conservation of natural genetic resources. Finally, the complete plastome sequencing and its characterization in detail provide tools to several biotechnology applications based on plastid transformation for these species or related ones.