UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa Interunidades De Pós-Graduação em Bioinformática

Marcele Laux

GENÔMICA COMPARATIVA E DIVERSIDADE GENÉTICA DE CLOROPLASTOS DE PLANTAS VASCULARES DA AMAZÔNIA

Belo Horizonte
Setembro – 2018

Marcele Laux

GENÔMICA COMPARATIVA E DIVERSIDADE GENÉTICA DE CLOROPLASTOS DE PLANTAS VASCULARES DA AMAZÔNIA

Dissertação apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais, como requisito parcial para à obtenção do título de Mestre em Bioinformática.

Orientador: Prof. Dr. Guilherme Oliveira

Belo Horizonte
Setembro – 2018

043 Laux, Marcele.

Genômica comparativa e diversidade genética de cloroplastos de plantas vasculares da amazônia [manuscrito] / Marcele Laux. – 2018.

55 f. : il. ; 29,5 cm.

Orientador: Dr. Guilherme Oliveira.

Dissertação (mestrado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Variação Genética. 3. Genoma de Planta. 4. Cloroplastos. I. Oliveira, Guilherme. II. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. III. Título.

CDU: 573:004

Ficha catalográfica elaborada pela bibliotecária Jéssica Patrícia Silva de Sá – CRB 6 - 3430

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática da UFMG

ATA DA DEFESA DE DISSERTAÇÃO

56/2018
entrada
2°/2016
CPF:
008.299.090-52

Marcele Laux

Às quatorze horas do dia **06 de setembro de 2018**, reuniu-se, no Instituto de Ciências Biológicas da UFMG, a Comissão Examinadora de Dissertação, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"Genômica comparativa e diversidade genética de cloroplastos de plantas vasculares da Amazônia"**, requisito para obtenção do grau de Mestre em **Bioinformática.** Abrindo a sessão, o Presidente da Comissão, **Dr. Guilherme Oliveira**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra à candidata, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa da candidata. Logo após, a Comissão se reuniu, sem a presença da candidata e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Prof./Pesq.	Instituição	CPF	Indicação
Dr. Guilherme Oliveira	Instituto Tecnológico Vale	68655118672	Aprovada
Dr. Thiago Mafra Batista	UFMG	060.041.026-95	APROVADA
Dra. Alessandra Giani	UFMG	378662 496.87	APROVADA
Dr. Tetsu Sakamoto	UFRN	330.246.028-79	APROVADA

Pelas indicações, a candidata foi considerada: Aprovada
O resultado final foi comunicado publicamente à candidata pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.
Belo Horizonte, 06 de setembro de 2018.

Dr. Guilherme Oliveira - Orientador Guilherme Cm de Ol
Dr. Thiago Mafra Batista
Dra. Alessandra Giani Alessandra Giani
Dr. Tetsu Sakamoto

AGRADECIMENTOS

Agradeço ao Programa de Pós-Graduação em Bioinformática da UFMG pela oportunidade de desenvolver esse trabalho e aos docentes do Programa pelos ensinamentos.

Agradeço ao Instituto Tecnológico Vale pela estrutura disponibilizada, à equipe da Genômica Ambiental, ao meu orientador Guilherme Oliveira e a todos os demais colegas do Instituto.

Ter uma equipe e colegas a quem confiar é um presente no nosso caminho.

Muito Obrigada!

RESUMO

A diversidade genética de plantas endêmicas e sob ameaça de extinção no Parque Nacional da Serra dos Carajás foi explorada através da montagem e análise comparativa de seus cloroplastos. Apresentamos neste trabalho a estrutura de nove espécies de *Ipomoea* (Convolvulaceae). Dentre estas, analisamos a diversidade genética de duas espécies consideradas irmãs, *I. cavalcantei* e *I. marabaensis*, e quatro indivíduos potencialmente híbridos, encontrados na zona de contato entre as duas populações. *Ipomoea cavalcantei* é uma espécie endêmica da Serra dos Carajás, apresenta uma distribuição restrita à serra norte e possui corola de cor vermelha, enquanto *I. marabaensis* é encontrada desde a serra norte até a serra sul e apresenta corola de cor lilás. Os indivíduos considerados potencialmente híbridos apresentam um gradiente de coloração, além de outros traços fenotípicos intermediários. O sequenciamento do DNA total foi realizado utilizando a estratégia *shotgun*, plataforma Illumina NextSeq 500, *paired-end* (2x150bp). Os plastomas foram montados utilizando os montadores *de novo* NOVOPlasty, específico para plastomas, e SPAdes. Após a montagem foi realizada a genômica comparativa, na qual avaliamos a estrutura dos plastomas, conteúdo e ordem gênica, similaridade a nível de nucleotídeos e distância genética utilizando distribuição de k-mers. Foi desenvolvido um *pipeline* de busca por regiões hipervariáveis através da dispersão dos k-mers previamente identificados. Os plastomas dos 4 espécimes híbridos apresentam maior similaridade aos plastomas da população de *I. cavalcantei*. Identificamos regiões de maior variabilidade especialmente localizadas nas regiões invertidas do cloroplasto. Destacamos o gene *ycf1*, o qual apresentou o maior valor de dispersão entre todas as regiões analisadas. Este estudo demonstra o potencial de variabilidade e adaptação de *I. cavalcantei* a diferentes condições ambientais e pressões antrópicas. A estrutura genética das espécies e potenciais híbridos é importante para as estratégias de conservação e manejo de populações da Serra dos Carajás.

Palavras-chave: Cloroplasto, Plastoma, *Ipomoea*, Amazônia, Canga

ABSTRACT

The genetic diversity of endemic and endangered plant species from the *Serra dos Carajás* National Park was explored through plastome assembly and comparative analysis. We present the plastome structure of nine *Ipomoea* (Convolvulaceae) from the Carajás National Forest. Among them, we analyzed the genetic diversity of two sister species, *I. cavalcantei* and *I. marabaensis*, and four putative hybrids found in the contact zone between the two populations. *Ipomoea cavalcantei* is endemic from Carajás Mountain Range, presents a remarkable red corolla, and a more restrict spatial distribution on the north range. *I. marabaensis* has a broad distribution, from north to south region and presents a typic lilac corolla. Individuals considered potential hybrids are found in the central region of the study area and display a color gradient, along with other intermediate phenotypic traits. Genomic DNA sequencing was performed using the shotgun strategy, *paired-end* (2x150bp), with the Illumina NextSeq 500 platform. The assembly was carried out using NOVOPlasty *de novo* assembler, designed explicitly for chloroplast genomes, and SPAdes. The comparative analysis encompassed plastome structures, gene content and order, nucleotide similarity, and genetic distance based on k-mer distribution. A pipeline for searching of hypervariable regions was developed based on the dispersion of k-mers. The plastomes of the four putative hybrid individuals showed higher similarity to *I. cavalcantei*. The hypervariable regions were located primarily in the two inverted repeats. We highlight the *ycf*1 gene, which presented the highest dispersion among all regions analyzed. This study shows the variability and adaptation potential of *I. cavalcantei* to distinct environmental conditions and anthropic pressure. The genetic infor*mat*ion will be relevant for conservation and management strategies for this two *Ipomoea* populations of Carajás.

Keywords: Chloroplast, Plastome, *Ipomoea*, Amazon, Canga

LISTA DE ILUSTRAÇÕES

Figura 1 - Localização geográfica e pontos de coleta na Serra de Carajás 15
Figura 2 - Genomas de cloroplastos de *Ipomoea cavalcantei*, *I marabaensis* e os quatro híbridos candidatos 22
Figura 3 - Comparação das quatro junções da estrutura quadripartida do cloroplasto 23
Figura 4 - Árvore filogenética Raxml (método ML) dos plastomas de *Ipomoea* 24
Figura 5 - Dendrograma AAF baseado na distribuição de frequência k-mers entre treze *Ipomoea* e dois *Phylodendron* como grupo externo 25
Figura 6 - Análise de componentes principais explicando 87,6% da variância gerada pelos primeiros 1000 K-mers H-disp 26
Figura 7 - Plastomes de *Isoetes serracarajasensis* sem gene 23S 36
Figura 8 - Plastomes *Isoetes cangae* 37
Figura 9 - Árvore de alinhamento múltiplo MAFFT abrangendo os oito espécimes de *Isoetes* e seis plastomas de referência 38
Figura 10 - Plastomas completos de *Phylodendron wullschlaegelii* (ITV1706) e *Phylodendron carajasense* (ITV3037) 40
Figura Suplementar 1 - Alinhamento malva em modo progressivo para verificar regiões homólogas e similaridade geral 50
Figura Suplementar 2 - A estrutura Pfam, curadoria Artemis e visualização de anotação para o gene Ycf1 51-53
Figura Suplementar 3 - Remapeamento do conjunto de dados original em relação aos plastomas montados 54-55

LISTA DE TABELAS

Tabela 1 – Espécies, código ITV, coletores e localização dos treze indivíduos *Ipomoea* estudados 15

Tabela 2 - Métricas de sequenciamento para cada conjunto de dados. Qual (qualidade), DP (profundidade de cobertura) 20

Tabela 3 - Posição e estrutura do gene Ycf1 nos treze genomas de cloroplastos de *Ipomoea*, mostrando a posição das regiões H-disp encontradas e respectivos escores de Blast para o banco de dados de *Ipomoea* 28

Tabela 4 - Repetição encontrada na primeira e segunda regiões H-disp em onze genomas de cloroplastos de *Ipomoea* (exceto *I. carnea* e *I. setifera*) mais quatro genomas de referência 29

Tabela 5 - Espécies, código ITV, coletores e localização dos oito indivíduos *Isoetes* 32

Tabela 6 - Métricas de sequenciamento para cada conjunto de dados. Qual (qualidade), DP (profundidade de cobertura) 34

Tabela 7 - Métricas de plastoma para cada compartimento dos plastomas *Isoetes* 35

Tabela 8 - Métricas de sequenciamento para cada conjunto de dados. Qual (qualidade), DP (profundidade de cobertura) 39

Tabela 9 - Comprimento e % GC para cada compartimento dos dois plastomas de *Philodendron* montados 40

Tabela Suplementar 1 - Frequência alélica, conteúdo de GC, comprimento, os três aminoácidos mais frequentes (respectivamente), os três códons mais frequentes (além de ATG e TGG) e os genes de cada junção entre os compartimentos do plastoma 47-48

Tabela Suplementar 2 - As primeiras 50 regiões H-disp que apontaram para diferentes genes e regiões intergênicas dos IRs 49

SUMÁRIO

1 INTRODUÇÃO…………………………………………………………………….....10

2 COMPARATIVE GENOMICS OF TWO IPOMOEA (CONVOLVULACEAE) POPULATIONS AND PUTATIVE HYBRIDS UNDER NATURAL AND ANTHROPIC PRESSURES ……………………………………………………………………...12

2.1 Abstract……………………………………………………………………………...12

2.2 Introduction…………………………………………………………………………13

2.3 Material and Methods………………………………………………………………16

2.4 Results and discussion………………………………………………………………19

2.5 Final Considerations……………………………………………………………...29

3 ANÁLISES ADICIONAIS ……………………………………………………………32

3.1 Characterization of the complete chloroplast genome of two Amazonian species of *Isoetes* (Isoetales) showing different habitat preferences……………………………………………32

3.2 Characterization of the complete chloroplast genome of two *Philodendron* species (Araceae)………………………………………………………………………………...39

4 REFERENCES…………………………………………………………………….....41

5 SUPPLEMENTARY MATERIAL 46

1. INTRODUÇÃO

A Floresta Nacional de Carajás, localizada no sudeste do Pará, na Amazônia, é uma unidade de conservação federal, criada em 1998, em partes dos municípios de Parauapebas, Canaã dos Carajás e Água Azul do Norte. Os maciços florestais da Flona Carajás são em geral bem conservados. Nas serras, onde se localizam as formações ferríferas, encontram-se as Cangas, um ecossistema ainda pouco estudado, porém muito explorado, uma vez que abriga grandes reservas de minério. As Cangas caracterizam-se pelo solo raso, rico em ferro e por formações vegetais rasteiras e abertas, com alto grau de especialização e endemismo.

Parte desses geossistemas ferruginosos são diretamente afetados pela mineração, sendo, dessa forma, eminente seu estudo em vista da conservação de espécies nativas e endêmicas e de um desenvolvimento sustentável na região. O Plano de Manejo da Floresta Nacional de Carajás de 2016 determinou que a zona de mineração fosse reduzida a cerca de 30%, uma zona de conservação de outros 30% e o restante como zona de manejo sustentável, sendo a pesquisa em biodiversidade realizada ativamente para subsidiar decisões futuras.

O presente trabalho está inserido em um projeto contínuo do Grupo de Genômica Ambiental do Instituto Tecnológico Vale, que tem por objetivo explorar a diversidade genética de plantas e animais raros e endêmicas e sob ameaça de extinção da Floresta Nacional de Carajás. Para plantas utilizamos abordagens que vão do amplo uso de códigos de barra de DNA, passando pelo sequenciamento de baixa cobertura do genoma, até estudos genômicos completos para espécies de maior interesse. Cada abordagem gera uma perspectiva nova sobre as questões abordadas, com diferentes profundidades de informações sobre a flora. Neste trabalho o foco é a caracterização do genoma completo de cloroplastos de espécies raras ou endêmicas da região das Cangas de Carajás.

O número de genomas de cloroplastos depositados no NCBI vem aumentando nos últimos anos, principalmente em função da redução dos custos de sequenciamento e maior acesso a ferramentas, computadores de alta performance e profissionais da bioinformática. Diversos genes e regiões intergênicas de cloroplastos têm sido usados como marcadores moleculares para reconstrução de filogenias, estudos de polimorfismos e evolução. Ao longo dessa trajetória, a ampliação no número de regiões analisadas passou a ser um requisito para o alcance de reconstruções filogenéticas mais robustas. O primeiro genoma completo de cloroplasto foi o do tabaco (*Nicotiana tabacum*), em 1986 (SHINOZAKI et al., 1986). Hoje temos acesso a mais de 3 mil sequências de genomas completos de cloroplasto (Fonte: https://www.ncbi.nlm.nih.gov/nuccore/?term=%22complete+chloroplast%22, 13/08/2018). Devido ao seu maior nível de conservação e sua herança majoritariamente *mat*erna, o uso de genomas completos de cloroplasto, também chamado de plastoma, tem mostrado alta robustez nas análises inter- e intra-populacionais.

Neste trabalho exploramos a diversidade genética das populações de duas espécies de *Ipomoea* (Convolvulaceae) das Cangas da FLONA Carajás e a estrutura do plastoma de outras sete espécies de *Ipomoea*. De acordo com as análises taxonômicas baseadas em caracteres morfológicos e análises genéticas baseadas em marcadores moleculares, a hipótese inicial é que em um dos pontos de amostragem ocorra um processo de hibridização entre duas espécies, *I. cavalcantei* e *I. marabaensis*. *Ipomoea cavalcantei* apresenta uma distribuição restrita à serra norte e possui corola de cor vermelha, enquanto *I. marabaensis* é encontrada desde a serra norte até a serra sul e apresenta corola de cor lilás. Os indivíduos considerados potencialmente híbridos apresentam um gradiente de coloração, além de outros traços fenotípicos intermediários, que não permitem uma delimitação taxonômica de base morfológica precisa. A estratégia de análise adotada foi a montagem dos plastomas de indivíduos cujas espécies puderam ser determinadas (9 espécies) e de indivíduos identificados apenas a nível de gênero (4 espécimes). Após a montagem foi realizada a genômica comparativa, na qual avaliamos a estrutura dos plastomas, conteúdo e ordem gênica, similaridade a nível de nucleotídeos e distância genética utilizando k-mers. De acordo com a estratégia e os parâmetros de comparação adotados, concluímos que os plastomas dos 4 espécimes considerados potencialmente híbridos apresentam maior similaridade aos plastomas da população de *I. cavalcantei*. Identificamos regiões de maior variabilidade especialmente localizadas nas regiões invertidas do cloroplasto, onde é esperado que exista um maior nível de conservação, uma vez que é a porção onde encontramos os principais genes do aparato de replicação. Destacamos o gene *ycf1*, o qual apresentou o maior valor de dispersão entre todas as regiões analisadas. Este estudo demonstra o potencial de variabilidade e adaptação de *I. cavalcantei* a diferentes condições ambientais e pressões antrópicas. Os resultados obtidos também auxiliam o esforço de conservação e manejo das espécies.

Adicionalmente, apresentamos a montagem de oito plastomas do gênero *Isoetes* (Isoetales), correspondentes a duas espécies endêmicas da Serra dos Carajás, *I. serracarajasense* e *I. cangae*., sendo a última uma espécie recentemente descrita. As duas espécies analisadas apresentam distribuição diferenciada, sendo *I. cangae* encontrada submersa em um único lago natural perene da Serra Sul e *I. serracarajasensis* encontrada em lagos sazonais das Serras Norte e Sul. De acordo com a análise de similaridade entre os plastomas de quatro indivíduos de cada espécie, mais seis plastomas de referência pudemos observar que as espécies de *Isoetes* da Serra dos Carajás diferenciam-se entre si, confirmando sua classificação taxonômica, e também das referências (Fonte: https://www.ncbi.nlm.nih.gov/nuccore/?term=isoetes+complete+chloroplast, 13/08/2014).

Apresentamos também a estrutura e conteúdo gênico de dois plastomas correspondentes a duas espécies do gênero *Philodendron* (Araceae), uma delas endêmica restrita de Carajás. É importante ressaltar que estes são os dois primeiros plastomas completos para o gênero.

Como resultado deste trabalho, foi estabelecido um *pipeline* para a montagem e anotação de plastomas, englobando dois montadores *de novo,* uma etapa de seleção de contigs não nucleares e preenchimento de *gaps* e confirmação das duas regiões invertidas repetidas através de alinhamento *pairwise* e remapeamento de contigs contra o plastoma *draft*, progressivamente. Estabelecemos também um *pipeline* inédito de busca por regiões hipervariáveis através da análise de dispersão de k-mers.

2 COMPARATIVE GENOMICS OF TWO *IPOMOEA* (CONVOLVULACEAE) POPULATIONS AND PUTATIVE HYBRIDS UNDER NATURAL AND ANTHROPIC PRESSURES

2.1 Abstract

The savannah-like ecosystem known as the ferruginous Canga that occurs in the Mountains of the Carajás National Forest, in the eastern Amazonian Forest, is a rich and intriguing vegetation type holding native and endemic plant populations adapted to a shallow soil and phytotoxic levels of metals. Convolvulaceae is one of the main families found in savannah-like ecosystems and presents nine genera and 34 species in the Carajás mountain range, 17 of these occurring in Canga areas, including nine *Ipomoea* species. *Ipomoea cavalcantei* and *I. marabaensis* are sister species from a lineage within the morning glory clade Murucoides and share some chloroplast DNA polymorphisms, but present a distinct pattern of geographic occurrence, the former restricted to the North and the latter spread throughout the region. The two species co-inhabit a region in the northern range where exists a putative zone of hybridization. We explored the *Ipomoea* populations of the Carajás Cangas, focusing on the hypothesis of hybridization between *I. cavalcantei* and *I. marabaensis.* We assembled thirteen *Ipomoea* chloroplast genomes, including four putative hybrids, all previously unpublished in the literature. The total DNA of the thirteen specimens were sequenced at high depth using the paired-end shotgun strategy. After assembly, they presented a quadripartite structure, similar gene content and order and slight differences in the IR-SSC junctions. We could observe a greater similarity among the four putative hybrids and *I. cavalcantei*. Deeper analysis using k-mer frequency pointed to the IRs as the major source of variation. Six regions, including the *ycf*1, *ndh*B, and *rps*15 genes were among the most informative regions defining the groupings among the analyzed species of *Ipomoea*. The results are consistent with the *I. cavalcantei* and *I. marabaensis* hybridization hypothesis. Further analysis would be able to depict the pattern of variation and gene flow among populations of *Ipomoea*.

Keywords: Chloroplast, Plastome, *Ipomoea*, Amazon, Canga

2.2 Introduction

The Carajás National Forest (Carajás FLONA) biodiversity has been studied in order to expand our knowledge about the native and endemic populations, hybridization and speciation events of such a unique Amazonian ecosystem (BABIYCHUK *et al.*, 2017; LANES *et al.*, 2018; SIMÃO-BIANCHINI; VASCONCELOS; PASTORE, 2016). The Carajás Mountain range is located in the Southeastern lowland Amazon and presents savannah-like ecosystems known as Canga, that are isolated from each other by matrixes of rainforest (Fig. 1). The Cangas contain shallow soils (0-10 cm), often with potentially phytotoxic levels of metals that are restrictive for the seedling establishment of some species, indicating that the soil properties are the primary drivers of vegetation composition and structure in the Cangas (BABIYCHUK *et al.*, 2017; NUNES *et al.*, 2015).

The Carajás flora is composed of 166 plant families, encompassing 1.066 species (VIANA *et al.*, 2016; VIANA; GIULIETTI-HARLEY, 2018). Convolvulaceae is one of the principal families found in savannah-like ecosystems across the world and present nine genera and 34 species in the Carajás FLONA, 17 of these occurring in Canga areas, including nine *Ipomoea* species (SIMÃO-BIANCHINI; VASCONCELOS; PASTORE, 2016). The genus *Ipomoea* comprises the largest number of species within the Convolvulaceae family, mostly concentrated in the Americas (AUSTIN; HUAMAN, 1996). Morning glories are commercially valuable species, as ornamental plants and food crop, being the sweet potato (*I. batatas*) the most widely planted species (HOSHINO *et al.*, 2016; YAN, L. *et al.*, 2015).

Ipomoea cavalcantei is a scandent liana with a red hypocrateriform corolla, endemic of the Carajás FLONA and considered an endangered species. *Ipomoea marabaensis* has a broader distribution and presents a larger lilac corolla (SIMÃO-BIANCHINI; VASCONCELOS; PASTORE, 2016). According to BABIYCHUK et al. (2017), *I. cavalcantei* and *I. marabaensis* are sister species from a lineage within the morning glory clade Murucoides and share some chloroplast DNA polymorphisms, namely the rpoC1^{Q} variant, which can be an example of convergent evolution. *Ipomoea cavalcantei* is mostly found in the North region of the Carajás FLONA (N1 site), while *I. marabaensis* is found along the entire region (Tab. 1). Several individuals with intermediate phenotypes have been observed at the N4 site and are considered putative hybrids between the two species (BABIYCHUK *et al.*, 2017; SIMÃO-BIANCHINI; VASCONCELOS; PASTORE, 2016). In order to better understand the *Ipomoea* populations of the Carajás mountain range, focusing on the hybridization hypothesis between *I. cavalcantei* and *I. marabaensis*, we assembled and analyzed 13 *Ipomoea* plastomes, including four putative hybrids collected at the N4 site, where *I. cavalcantei* and *I. marabaensis* co-inhabit.

Chloroplasts are intracellular plant organelles which contain the entire machinery for the photosynthesis process and can hold genetic determinants for variegation in higher plants (SUGIURA, 1992). The plastome has been used to reconstruct phylogenies on a broad scale and set inter-population variation, boundaries, and gene flow at the local scale (TONTI-FILIPPINI *et al.*, 2017). The most widely used plant classification systems (GROUP *et al.*, 2016) now rely heavily on molecular data, often from plastid sequences, to determine clade boundaries and relationships. Therefore, complete plastomes may be extremely useful for providing an abundance of additional characters that can be used to resolve polytomies in phylogenetic trees (STRAUB *et al.*, 2012; WILLIAMS *et al.*, 2016) and boost statistical confidence in deeply branching clades (SUN *et al.*, 2016). According to WILLIAMS et al. (2016), the number of characters included have a significant influence in the resolution in phylogenetic approaches and the use of whole genome sequences, as well as the expansion of DNA barcodes, has the potential to increase phylogenetic support.

Nine species were identified according to the International Code of Botanical Nomenclature (SIMÃO-BIANCHINI; VASCONCELOS; PASTORE, 2016) and were deposited in the herbarium of the Museum Emilio Goeldi (Belém, Pará, Brazil). Four specimens were identified as *Ipomoea* putative hybrids, ITV2181, ITV280, ITV2295 and ITV2294 (Tab. 1).

Table 1. Species, ITV code, collectors, and Location from the thirteen *Ipomoea* individuals studied.

Name	Code	Collectors	Location	Sector
I. cavalcantei	3206	Babiytchouk, E.	Parauapebas, Pará, Brazil	N1
I. sp1	2181	Harley, R.M.	Parauapebas, Pará, Brazil	N4
I. sp2	280	Santos, F	Parauapebas, Pará, Brazil	N4
I. sp3	2294	Babiytchouk, E.	Parauapebas, Pará, Brazil	N4
I. sp4	2295	Babiytchouk, E.	Parauapebas, Pará, Brazil	N4
I. goyazensis	4320	Vasconcelos, L.V.	Parauapebas, Pará, Brazil	N4
I. marabaensis	2328	Babiytchouk, E.	Canaã dos Carajás, Pará, Brazil	Serra do Tarzan
I. carnea	2324	Babiytchouk, E.	Salinópolis, Pará, Brazil	Salinas
I. asarifolia	3285	Babiytchouk, E.	Salinópolis, Pará, Brazil	Salinas
I. triloba	4963	Nogueira, M.G.C.	Redenção, Pará, Brazil	-
I. quamoclit	4995	Nogueira, M.G.C.	Redenção, Pará, Brazil	-
I. setifera	2613	Pastore, M.	APA São Geraldo do Araguaia, Pará, Brazil	-
I. maurandioides	4245			

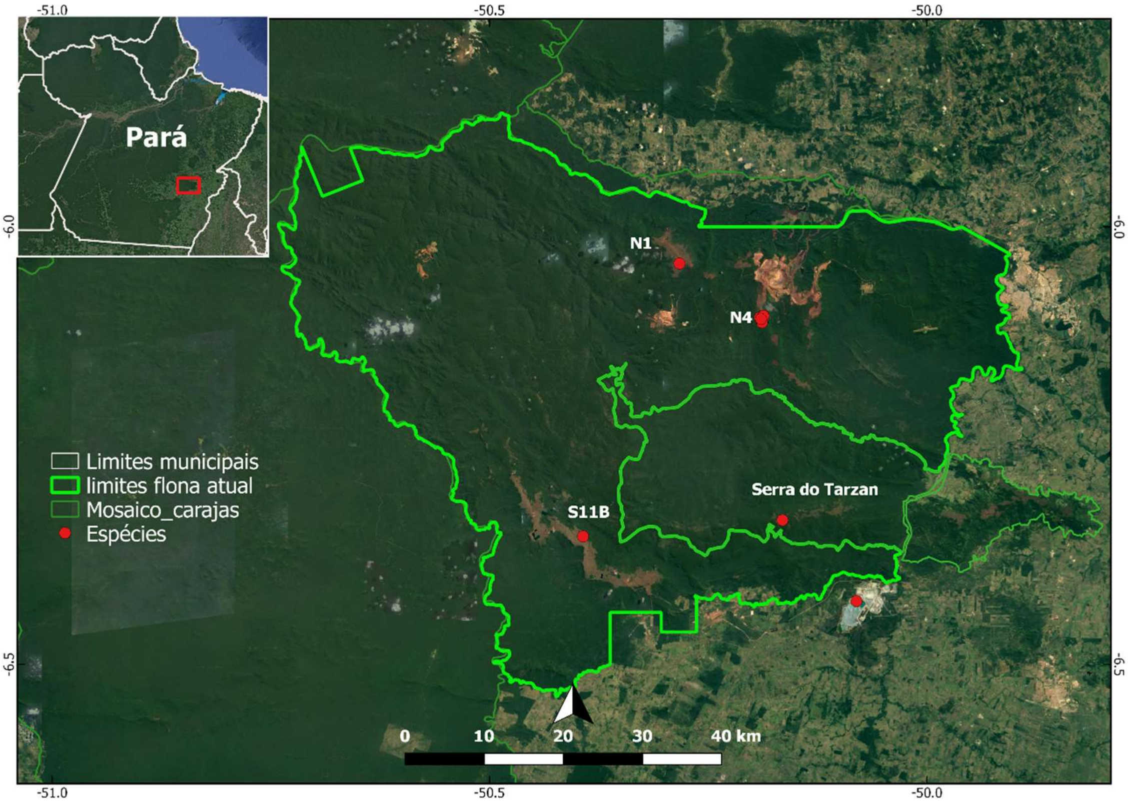


Figure 1. Geographic location and collection points at Carajás Mountain Range.

2.3 Material and Methods

2.3.1 DNA extraction

DNA extraction was carried out using the QIAcube HT robot (Qiagen), with approximately 20 mg of plant material collected in NaCl-saturated CTAB solution. Samples were weighed and transferred to a 96 well-plate containing two 3 mm stainless steel beads (Qiagen). Subsequently, the plates were frozen overnight at -80 °C and the leaf material was pulverized using a TissueLyser II (Qiagen) for 2 min at 30 Hz. After this, 600 µL of extraction buffer [2% CTAB, 0.1 mM Tris-HCl (pH 8.0), 20 mM EDTA (pH 8.0), 1.4 M NaCl] were added to the sample that was kept in a water bath at 60 °C for 40 min under gentle agitation. The plates were then centrifuged for approximately 30 s at 14,000 rpm, and then 200 µL of the supernatant was transferred to a 96 sample S-block. Afterward, DNA extraction was carried out using the QIAamp 96 DNA kit (Qiagen), following the manufacturer's instructions. DNA quantity and quality were checked using the Eon spectrophotometer (Bioteck) and Qubit 3.0 (Invitrogen®).

2.3.2 Genome sequencing

Sequencing runs were performed using the NextSeq 500 Illumina platform. Briefly, paired-end libraries were constructed from 50 ng of DNA. Samples were subjected to a step of enzymatic and random fragmentation in which the DNA was simultaneously fragmented and bound to adapters using the QXT SureSelect (Agilent Technologies®) kit according to the manufacturers' instructions. The fragmented DNA was purified and subjected to an amplification reaction using primers complementary to the adapters. Next, the libraries were quantified using the Qubit® 3.0 Fluorimeter (Life Technologies, USA) and checked for fragments size in the 2100 Bioanalyzer (Agilent Technologies®). Then, the libraries were diluted in a solution of 0.1% Tris-HCl and Tween and pooled. The sequencing run was performed with a NextSeq 500 v2 kit high output (300 cycles).

2.3.3 Chloroplast Genome assembly and annotation

No pre-processing method was applied because initial trials using Prinseq (SCHMIEDER; EDWARDS, 2011), Trimmomatic (BOLGER; LOHSE; USADEL, 2014) and cutadapt (MARTIN, 2011) did not generate contigs better than the original datasets. The quality of the datasets was checked using the FastQC tool (ANDREWS, 2010). The *de novo* assembling was conducted using both NOVOPlasty version 2.6.3 (DIERCKXSENS; MARDULYN; SMITS, 2017) and SPAdes version 3.11 (BANKEVICH *et al.*, 2012) assemblers. NOVOPlasty is an assembler created specifically for assembling organelle genomes from total DNA reads, which can deal with inverted repeats and can produce circular assemblies if both the ends of seed-extend overlap by 200 bp. Its algorithm uses seed-extend based assembler, and hashing tables from the whole genome sequencing runs. The assembly starts with a user-defined seed sequence, which acts as an anchor to the extend the seed bi-directionally. The config file was set as follows: insert size 300, read length 150, type chloro, genome range 120k-200k, k-mer 39, paired-end mode and original dataset with the full content of the DNA extracted as input. The main seeds used for capture and contig extension were (in order of effectiveness) the entire genes *psb*K, *psb*B, *psa*C, *ndh*F, *rbc*L, *psb*C, *rpo*B, *rrn*23, *trn*H, *ycf*1, *rps*15, *mat*K, *rpl*32 and some partial sequences as IRa-SSC junction, IRb-SSC junction, and *ycf*1-*rps*15 intergenic space. We also used DNA barcodes generated by the ITV DNA barcoding efforts (VASCONCELOS et al., unpublished). The resulting contigs were assembled in Geneious R11 (KEARSE *et al.*, 2012) and the consensus sequences were annotated in CpGAVAS (Chloroplast Genome Annotation, Visualization, Analysis and GenBank Submission Tool) web server (LIU *et al.*, 2012).

SPAdes is a de-Bruijn assembler for which a graph is constructed based on k-mers, and the operations are based on graph topology, coverage, and sequence length, but not the sequences themselves. At the last stage, the consensus DNA sequence is restored (BANKEVICH *et al.*, 2012). Spades was run in an iterative short-read genome assembly module, in which the values of K were automatically selected based on the read length and data set type. The parameters were default, and the inputs were the fastq files, paired-end, of the original datasets. Spades was not designed to deal with the chloroplast (cp) genome architecture, especially the inverted repeats, so the larger contigs generated was about 27-89k long. In order to select cp contigs in the SPAdes assembly, the contigs were chosen according to a two-step selection. The cp DNA fragments are expected to be more abundant in the total DNA extracted, since such organelle is found in high numbers within each plant cell (MCKAIN *et al.*, 2018; SAKAMOTO; TAKAMI, 2018), so a coverage cutoff was applied according to the overall contigs depth of coverage (DP) median (select_contigs.pl). The selected high DP contigs were subsequently aligned to the plastid NCBI database (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/plastid/, November 2017) using megablast (MORGULIS *et al.*, 2008) with e-value 1e-5 and minimum percent identity of 85. The cp contigs sequences were then extracted from the original output using extract_seq.pl. The selected SPAdes contigs were used mostly to help gaps filling and sequence discrepancies between inverted repeat regions.

Most plastomes were assembled only with NOVOPlasty, but for manual intervention and gap filling we used the selected cp contigs from SPAdes. The IRs should be identical, so all thirteen genomes needed manual editing for genome finishing. For each sample, after the first draft genome was entirely assembled and annotated in CpGAVAS, the IRs were extracted and pairwise aligned. The NOVOPlasty and SPAdes selected contigs were mapped against the draft genome, and the contigs which fell in such regions were used to guide the sequence edition in the pairwise alignment. The final IR region generated was mapped against the original dataset using Bowtie2 (LANGMEAD; SALZBERG, 2012) and then inserted in the draft genome (forward direction for IRa and reverse for IRb). Further adjustments were performed by the same protocol, using re-mapped contigs and Bowtie2. The final draft genome was then uploaded to CpGAVAS webserver and checked for gene content, order and IRs before considered finished and its GFF file used as input in Artemis tool (RUTHERFORD *et al.*, 2000) for curation. After curation, the edited GFF was re-annotated in CpGAVAS to generate the final GFF file.

CpGAVAS is implemented using Perl Catalyst Web Application Framework and a combination of Perl programs, which provide standard functions to annotate and analyze the cp genome sequences, generate circular genome maps, summary statistics of the annotated genome, and creation of files for GenBank submission. It also included tRNAscan (LOWE; EDDY, 1997) for the prediction of the *trn*A in the chloroplast genomes. None of the annotation tools currently provide submission-ready annotations, requiring subsequent manual curation, primarily because of the extremely short exons of some cp genes that are a feature of some plastid genes, as *pet*B and *pet*D, the intron-containing *trn*A genes, the intron splice sites, identification of start codons non-ATG, fragmented genes and trans-splicing, as for *rps*12 gene (TONTI-FILIPPINI *et al.*, 2017).

The thirteen final plastomes were re-mapped against its original datasets to check for coverage uniformity throughout the entire genome, and the mapped reads were used to calculate the depth of coverage for each one. The metrics as total length, gene content, GC%, amino acid frequency, and codon usage were assessed using the Geneious metrics display.

2.3.4 Comparative analysis

The thirteen *Ipomoea* cp genomes plus four references *I. nil, I. purpurea, I. trifida,* and *I. batatas* were initially aligned using Mauve (DARLING *et al.*, 2004) in progressive mode to check for homologous regions and general similarity. The MAFFT (KATOH *et al.*, 2002) multiple alignment was then performed to check the nucleotide structure since the MAFFT algorithm relies on nucleotide similarity. The dendrogram was generated with the thirteen *Ipomoea* plastomes using maximum parsimony (MP) performed with RAxML 8.2.8 (STAMATAKIS, 2014) using the rapid bootstrapping option with 1000 replicates. The ML analysis was performed as implemented in the CIPRES portal (http://www.phylo.org). The Alignment and Assembly-free (AAF) software was used to check if the same structure would be observed through k-mer frequency-based phylogeny (FAN, H. *et al.*, 2015).

AAF reconstructs phylogeny from a distance matrix based on the proportion of shared k-mers from raw sequencing reads between each sample (FAN, H. *et al.*, 2015). AAF allows the user to choose the k-mer length. We tested k-mers of 17, 25, and 31 nucleotides. The selection of k is a trade-off between avoiding multiple mutations on one k-mer (which favors shorter k) and decreasing the chances of k-mer homoplasy (which favors longer k). We performed a downstream statistical analysis of such k-mer frequency and distribution using R language (R CORE TEAM, 2018). The phylokmer.dat.wc output provides the total k-mer diversity for each sample, and the phylokmer.dat output provides a table with k-mer frequencies from each sample. After checking the histogram (ggplot function) for each sample (exponential distribution), the k-mer dispersion was calculated (sd function) and sorted in decreasing order. The top-1,000 k-mers were clustered using the hclust function and visualized through Principal Component Analysis, PCA function. The contribution of the variance observed was calculated using the fviz_contrib function, factoextra package (KASSAMBARA & MUNDT, 2017) and visualized in a histogram. A cutoff of 0.11 contribution was applied, and a final list of 432 high dispersion (H-disp) k-mers was generated. The list was sorted from the highest contributing k-mers, in decreasing order. The first 100 k-mers were manually searched in the cp genomes to determine the H-disp regions, which were then analyzed and described.

2.4 Results and discussion

2.4.1 Chloroplast genome sequencing and assembly

The total DNA of thirteen specimens were sequenced using the paired-end shotgun strategy in the NextSeq 500 Illumina platform, generating a total of 1,072 Gbp data (Tab. 2). All datasets presented a high average quality (Qual) and depth of coverage (DP) according to the respective plastome.

Table 2. Sequencing metrics for each dataset. Qual (quality). DP (depth of coverage).

Name	Code	File size	Qual	reads	Kbp	Mapped reads	Plastome size (bp)	DP
I. cavalcantei	3206	360G	32	1,127,842,550	169,706,864	3,077,468	161,563	2,857x
I. marabaensis	2328	284G	32	820,626,544	123,574,346	2,267,900	161,324	2,108x
I. maurandioides	4245	22G	32	59,864,318	9,008,883	2,827,744	161,242	2,630x
I. carnea	2324	50G	32	142,547,972	21,446,710	14,427,906	160,819	1,3457x
I. asarifolia	3285	32G	32	91,986,912	13,830,802	4,104,928	160,589	3,834x
I. triloba	4963	28G	32	75,654,494	11,377,938	2,661,608	161,835	2,466x
I. goyazensis	4320	40G	32	110,835,364	16,678,648	637,082	160,414	595x
I. quamoclit	4995	28G	32	167,419,792	25,185,853	7,052,366	160,836	6,577x
I. setifera	2613	24G	32	64,363,332	9,678,711	6,015,926	160,082	5,637x
I. sp1	2181	40G	32	110,820,732	16,675,181	3,338,016	161,495	3,100x
I. sp2	280	42G	32	118,798,524	17,867,963	3,900,308	160,765	3,639x
I. sp3	2294	46G	32	130,420,838	19,623,103	2,078,538	160,974	1,936x
I. sp4	2295	46G	32	130,605,640	19,648,319	2,870,856	161,948	2,659x

The *de novo* assembly was performed using a hybrid strategy with the *de novo* NOVOPlasty (NP) assembler in conjunction with selected contigs from SPAdes assembler. The seeds initially used for contig extension in NP assembler were genes from the reference genome of *I. nil* (HOSHINO *et al.*, 2016) and subsequently genes from the very assembled plastomes. A total of 2,693 NP contigs and 11,442 SPAdes contigs were used to perform the assemblies. Besides the higher number of contigs, SPAdes generated only about three consensus sequences for each sample which could be mapped and used in the assembly. The majority of all thirteen genomes were assembled with NP contigs. The inverted repeats and the long, low-complexity AT-rich regions are problematic for almost all assemblers, requiring manual intervention to orientate and attach contigs correctly (TONTI-FILIPPINI *et al.*, 2017).

The average length of the chloroplast genomes was 161 kbp, and they presented 79 coding genes, four ribosomal genes, and 18 *trn*A genes. The allelic frequency, GC content, length, the three more frequent amino acids (respectively), the three most frequent codons (besides ATG and TGG) and the genes of each junction between cp genome compartments are presented in Supplementary Table 1. All chloroplast genomes presented a homogeneous coverage in the remapping analysis (Suppl. Fig. 3). *Ipomoea cavalcantei* and *I. marabaensis* are considered the parental source of the hybridization process, so the circular genome of both and the four putative hybrids are presented in Figure 2.

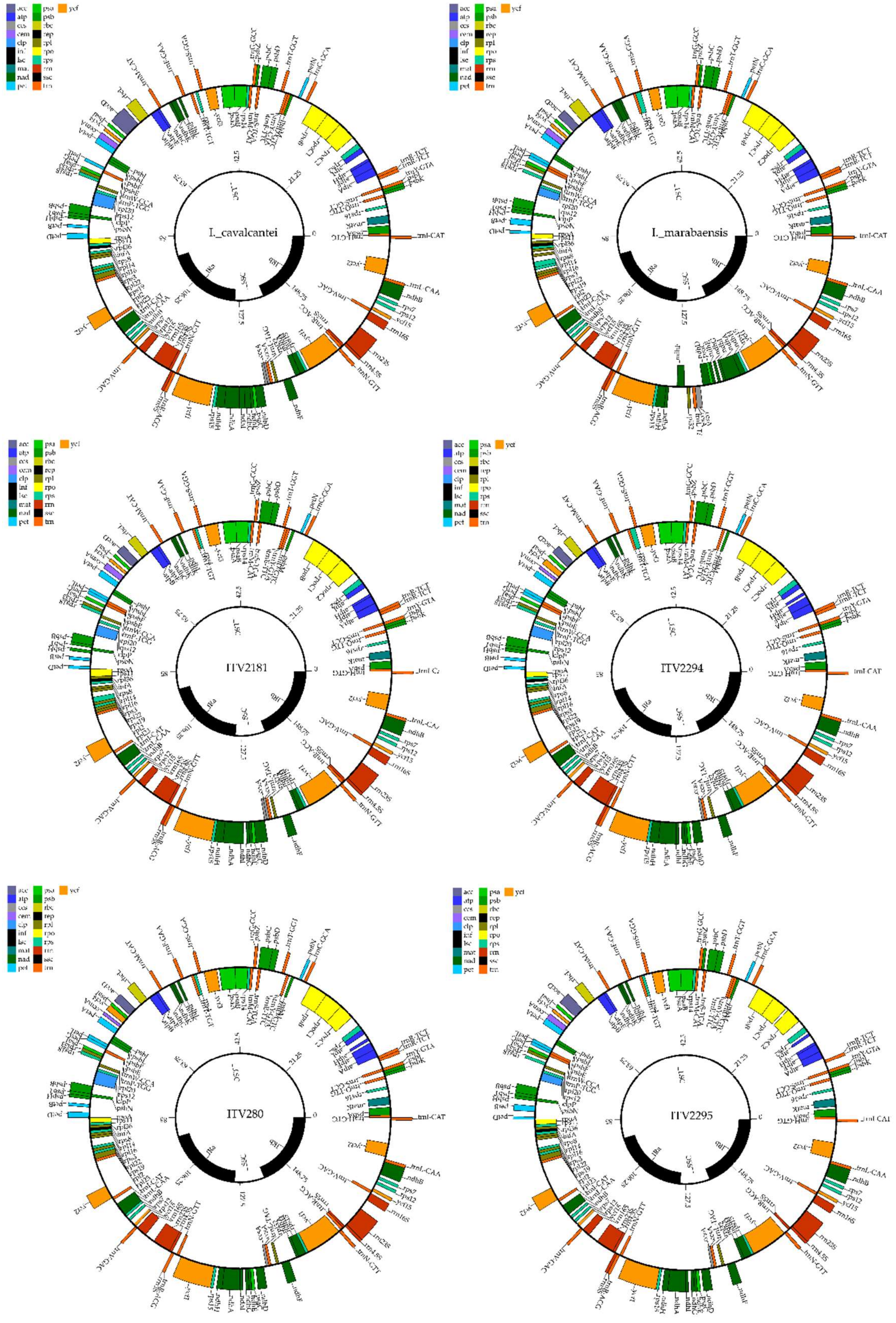


Figure 2. Chloroplast genomes of *I. cavalcantei*, *I marabaensis* and the four putative hybrids.

2.4.2 Comparative Analysis

All thirteen complete *Ipomoea* cp genomes displayed the typical quadripartite structure, including the large single copy (LSC, 87 kbp), the small single copy (SSC, 12 kbp) and the inverted repeats (IRa and IRb, 30 kbp). Most genomes presented a similar gene content, some slight rearrangements in the SSC and some differences in SSC junctions. *Ipomoea setifera* (ITV2613) presented the most discrepant genome, with an expansion of IR towards LSC incorporating the *rpl*2 and *rpl*23 genes and a contraction towards the SSC, showing the *ycf*1 gene, *rps*15 and *ndh*H in the SSC region (Fig. 3). *Ipomoea cavalcantei*, *I. marabaensis,* and the four putative hybrids presented the same junctions between LSC and IRs (*rpl*2, *rpl*23/*trn*I, *ycf*2). *Ipomoea marabaensis* presented an inverted SSC gene order and subsequent difference in the SSC-IRs junctions, where the IRa fully encompassed the *nhd*A gene, and partially encompassed by the IRb, while the opposite was observed for the other five genomes (Fig. 3).

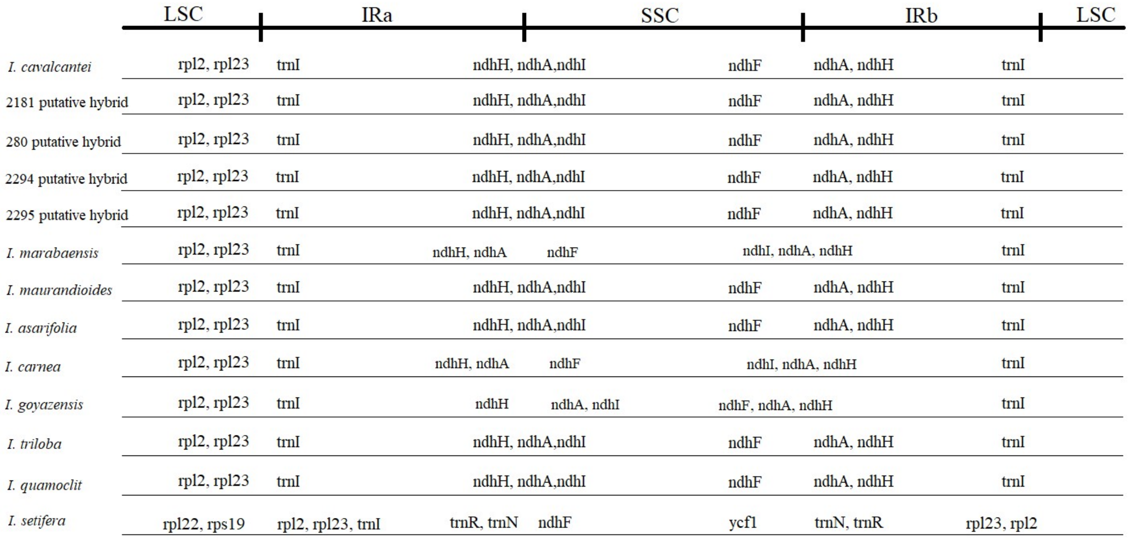


Figure 3. Comparison of the four junctions of the chloroplast quadripartite structure

According to the Mauve alignment (Suppl. Fig. 1), there are two large homologous regions among all 13 genomes, one encompassing the entire LSC and part of the IRa and another encompassing the SSC and respective IR junctions. *Ipomoea setifera* presented no homology related to the IRs, with blocks restricted to LSC and SSC regions. *Ipomoea cavalcantei* and two putative hybrids 2181 and 280 showed a third homologous block at the end of the IRb, from 16S gene to *trn*I. Accordingly, MAFFT multiple alignment and the phylogenetic reconstruction using the thirteen *Ipomoea* plastomes showed higher support values in the ML analysis for *I. triloba*, *I. asarifolia* and *I. maurandioides* branch (100%), *I. marabaensis* and *I. carnea* (100%). Lower support values and unresolved branches were observed concerning *I. cavalcantei* and the four putative hybrids branch, likely due to the higher similarity and few reliable sinapomorphisms among them, corroborating the hypothesis that *I. cavalcantei* is likely the primary source of variation concerning the plastomes (Figure 4).

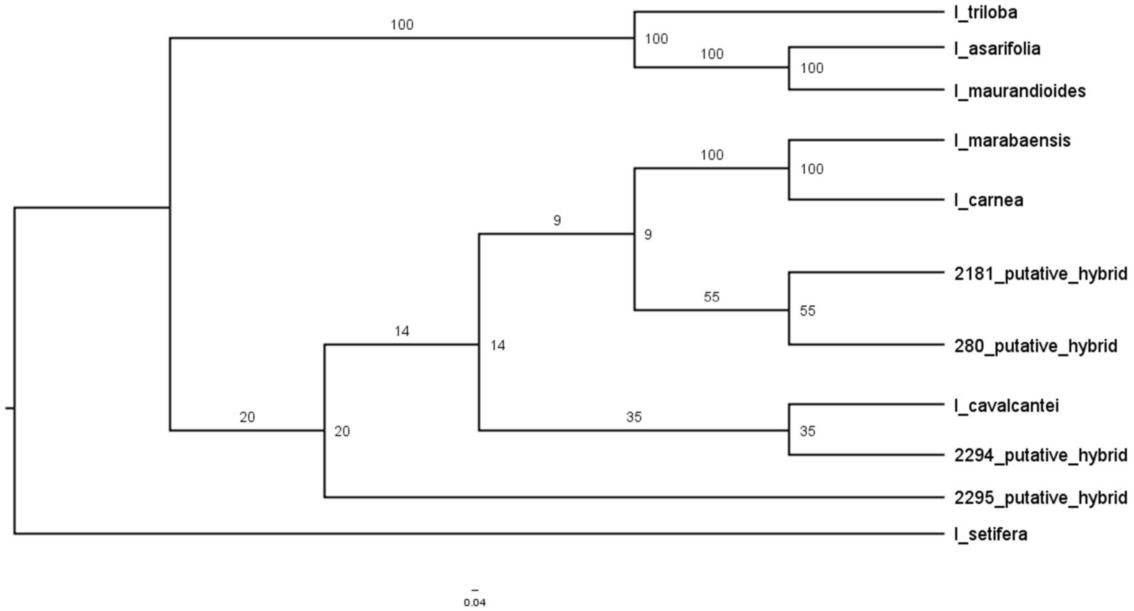


Figure 4. Raxml phylogenetic tree (ML method) of the Carajás *Ipomoea* plastomes.

Although *I. cavalcantei* and *I. marabaensis* are considered sister species, they were found at different branches in phylogenetic analysis. BABIYCHUK et al. (2017) observed significant genetic differentiation in the *I. cavalcantei* species, mostly due to geographic location, grouping the Canga N1 and N4 sites, while *I. marabaensis* was collected at the south (Serra do Tarzan, Tab. 1), what could be an indicative of the combined effects of genetic drift and population bottlenecks occurring in Cangas.

2.4.3 Using K-mer frequency for comparative analysis

The k-mer frequency among the thirteen chloroplast genomes was analyzed in order to find the sequence regions that were more informative in the genetic similarity analyses (Fig. 4). We used the AAF approach (MARÇAIS; KINGSFORD, 2011) to capture the k-mer frequency for each sample. Since the three k-mers 17, 25, and 31 showed similar tree topologies, we choose the 25-mer for subsequent analysis. The 25-

mer AAF tree showed a branch encompassing *I. cavalcantei*, *I. marabaensis,* and the four putative hybrids (Fig. 5). Two *Philodendron* species were included in the analysis as outgroups.

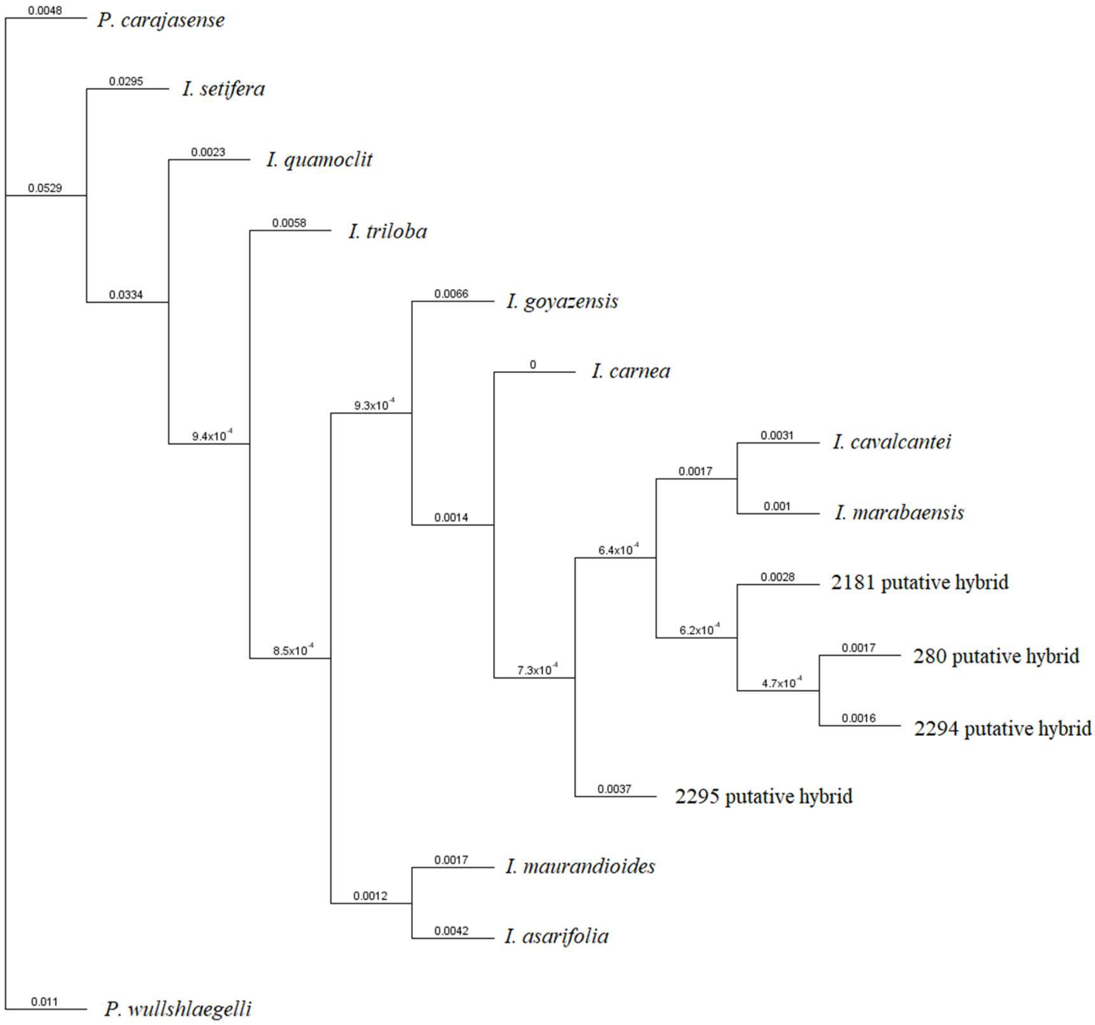


Figure 5. AAF Dendrogram based on k-mers frequency distribution among thirteen *Ipomoea* and two *Phylodendron* as outgroup.

The 25-mer frequency by sample table presented a total of 1,187,671 k-mers, most of which shared among all samples with similar frequency. The samples with higher k-mer diversity were *I. carnea*, *I. quamoclit*, and *I. setifera*. We identified the more infor*mat*ive k-mers, which were mostly the ones not shared among all samples or with highly distinct frequency patterns. Using the dispersion among the frequencies as a contribution value parameter, we were able to progressively reduce the number of k-mers, selecting the ones with the highest contribution, until we reach a dispersion-based clustering still similar to the original dendrogram.

We selected 432 k-mers, several of them as part of the same string, which we call High Dispersion Regions (H-disp). The first 100 H-disp regions pointed to different genes and intergenic regions of the IRs (Suppl. Table 2). *Ipomoea setifera*, which was the most distant species, did not present any of the selected k-mers. The Principal Component analysis explained 87.6% of the variance, and we can see the grouping among *I. cavalcantei* and the four putative hybrids in the second axis (Fig 6).

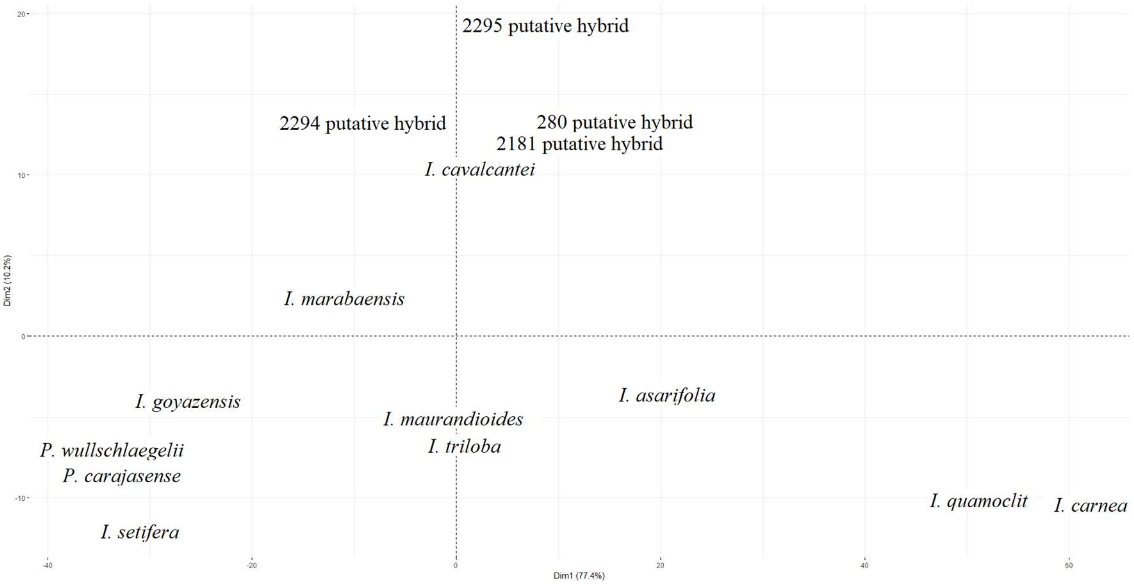


Figure 6. Principal component analysis explaining 87.6% of the variance generated by the first 1000 H-disp K-mers.

The *ycf*1 gene contained the most divergent regions. The Pfam structure, Artemis curation, and annotation view are presented in Supplementary figure 2. According to Pfam database, the domain organization was more similar to the architecture *ycf*1 x 2, from *Glycine max* (soybean), encoding the protein TIC 214, which presents two complete domains, similar to the architecture of *I. setifera*. The remaining specimens presented three domains, very similar among them and with the *I. nil* reference genome. The exception was *I. maurandioides*, which presented a break between the second and third domains, similar to the cp genomes of *I. purpurea* and *I. trifida.* We observed that the gene size is quite similar among them, but the number of exons annotated is variable. All y*cf1* genes required manual curation for the final annotation. The original annotation showed frameshifts, introns and pre*mat*ure stop codons, as previously observed by (MCNEAL *et al.*, 2007) for *Cuscuta* species (Convolvulaceae). After edition, all of them presented the full content of the *ycf*1 gene. H-disp kmers mapped to three different regions of the *ycf*1 gene (Table 3). The first two H-disp regions presented all Blast hits with *Ipomoea* species, but with a low score. The third region presented an even lower Blast scores, and the hits were scattered among several species, probably indicating a particular region from these cp genomes in relation to another *Ipomoea* species.

Table 3. Position and structure of the *Ycf*1 gene in the thirteen *Ipomoea* chloroplast genomes, showing the position of the H-disp regions found and respective Blast scores to the *Ipomoea* database.

Species	*Ycf*1a position	Gene Size	Exons	H-disp region 1	Blast score for *Ipomoea*	H-disp region 2	Blast score for *Ipomoea*	H-disp region 3	Blast score for *Ipomoea*
2295_putative_hybrid	110752-116-715	5964	4	115384-115433	93.5	115489-115538	93.5	116423-116460	71.3*
I. marabaensis	110214-116-108	5895	7	114747-114796	93.5	114852-114901	93.5	115810-115847	71.3*
2181_putative_hybrid	110355-116237	5883	2	114906-114955	93.5	115011-115060	93.5	115945-115982	71.3*
2294_putative_hybrid	110064-115910	5847	7	114597-114646	93.5	114702-114751	93.5	115612-115649	71.3*
280_putative_hybrid	109900-115797	5898	7	114460-114509	93.5	114565-114614	93.5	115499-115536	71.3*
I. cavalcantei	110786-116614	5829	1	115283-115332	93.5	115388-115437	93.5	116322-116359	71.3*
I. carnea	109884-115730	5847	4	114480-114529	93.5	115438-115475	71.3*		
I. triloba	110528-116213	5686	4	115329-115377	91.6	115452-115501	93.5	116410-116447	71.3*
I. quamoclit	110222-115954	5733	4	114581-114630	93.5	114704-114753	93.5	115662-115699	71.3*
I. goyazensis	110240-116086	5847	7	114755-114804	93.5	114878-114927	93.5	115788-115825	71.3*
I. asarifolia	109542-115453	5912	4	114081-114130	93.5	114204-114253	93.5	115161-115198	71.3*
I. maurandioides	110090-116242	6153	4	114893-114942	75	115016-115065	93.5	115950-115987	71.3*

In the first and second H-disp regions, we found a 50 bp repeat that occurred in two copies, except for *I. carnea*, which contained only a single copy. The reference genomes also presented such repeat, in the same position, even for *I. batatas*, which lacks *Ycf*1 gene (YAN, L. *et al.*, 2015) but presents de two copies of the repeat in the same position, between *trn*N and *rps*15 (Table 4). We could not find any reference for such repeat in the repeats databases.

Table 4. Repeat found within the first and second H-disp regions in eleven *Ipomoea* chloroplast genomes (except *I. carnea* and *I. setifera*) plus four reference genomes.

Repeat		
GAAGAGATCAATCCCAGCAGTAATCAAAAGACTCCAATTGGGACTAATAA		
Position		
2295 putative hybrid	**2294 putative hybrid**	**2181 putative hybrid**
4633-4682	4534-4583	4552-4601
4738-4787	4639-4688	4657-4706
280 putative hybrid	**I. marabaensis**	***I. cavalcantei***
4561-4610	4534-4583	4498-4547
4666-4715	4639-4688	4603-4652
I. asarifolia	***I. maurandioides***	***I. triloba***
4540-4589	4804-4853	4802-4851
4663-4712	4927-4976	4925-4974
I. goyazensis	***I. quamoclit***	***I. nil***
4516-4565	4360-4409	4618-4667
4639-4688	4483-4532	4741-4790
I. purpurea	***I. trifida***	***I. batatas***
4708-4757	4708-4757	between *trn*N and *rps*15
4849-4898	4849-4898	

The y*cf*1 gene encodes for a 214 kD protein and is part of the TIC complex which, with the TOC complex, are responsible for the translocation of nuclear-encoded pre-proteins across the double envelope membranes of chloroplasts (KIKUCHI *et al.*, 2013). The y*cf*1 gene is considered a marker gene, involved in species diversification (MENEZES *et al.*, 2018; NEUBIG; ABBOTT, 2010; YAN, M. *et al.*, 2018) & Abbott 2010, Neubig et al. 2009). According to (DONG *et al.*, 2012), within the *ycf*1 gene are the two most variable regions, called *ycf*1a and y*cf*1b. Such regions were verified using 136 genomes belonging to 27 genera, including four Bryophytes, three Monilophytes, four Gymnosperms and 16 angiosperms (DONG *et al.*, 2015).

The third most divergent region was the intergenic space between *rps*7 and *ndh*B genes, except for *I. goyazensis* and *I. setifera*

(“TTCCAATTTCAAAAAAAAATCCCAATTGTGTCGA”). The ferredoxin-dependent plastoquinone reductase (NDH complex) is required for one of the cyclic electron flow pathways around PSI (YAMORI; SHIKANAI, 2016). The chloroplast genome encodes eleven subunits of the NDH complex. Considering the possible physiological role of NDH in photoprotection, the expression of the *ndh* genes may be regulated to cope with changes in environmental conditions and at least two processing sites are found in the intergenic region between *rps*7 and *ndh*B (HASHIMOTO *et al.*, 2003), working as small non-coding RNAs (RUWE; SCHMITZ-LINNEWEBER, 2012). Interestingly, this divergent region is located in the region responsible for the transcription of the *ndh*F gene, in the 5’ *ndh*B (32bp, “TTCATTCTGTACATGCCAGCTCATGAATTAGT”). We propose that the divergent sequence in the transcription regulatory region of the *ndh*F gene may be involved in regulating the appropriate transcription of the gene according to the physiological demands of the species habitat.

The fourth H-disp region was found in the *ndh*H-*rps*15 intergenic spacer, a 25 bp string (“CTTATTTGTTTCGTTTCAATTTTGA”) was observed in all genomes, except for *I. setifera.* Such intergenic spacer was already used as a marker for species discrimination in some studies (SEBASTIANI; CARNEVALE; VENDRAMIN, 2004; SHEPHERD *et al.*, 2016). The fifth H-disp region was found in the *trn*I-*ycf*2 intergenic spacer (“GTTTTCAAGTAATGTTTGATCAATTACGTATTTATACACGTATTCGTATTA ATCAATTTTTGATGAATTATTCTG “, 75 bp long). The sixth H-disp region was located within the *rps*15 gene, next to 3’ end, in all genomes except in *I. setifera* (47 bp, “GTACGTTATAAAGAATTAATTGAGAAATTGGATATTCGAGAGACAAA”).

2.5 Final Considerations

Our current hypothesis is that *I. cavalcantei* and *I. marabaensis* are the source of diversification which originated the four putative hybrids. The first step was the assembly of the plastid genomes, in order to find indications of diversification and variability. To fully explore the hybridization hypothesis, the nuclear genome must be assembled as well.

When analyzing the k-mer frequency of the plastomes we observed that *I. cavalcantei* is closest to the putative hybrids (Fig. 5-6). In order to better understand such structure, we adopted a statistical strategy using the dispersion parameter of the k-mers as a variability metric. Interestingly, the six H-disp regions depicted through such strategy fall in the IR regions. Early studies considered such regions as highly conservative, showing lower levels of rearrangements (KOLODNER; TEWARI, 1979; PALMER; THOMPSON, 1982). IRs are considered essentially identical, so that we could describe the plastid genome structure also as tripartite According to (WICKE *et al.*, 2011), the expansion, contraction, rearrangement, and loss of one of the inverted repeats can be related to rearrangements on the whole cp genome. Additionally, several studies point to the contraction, expansion (ASAF *et al.*, 2017; KIM, 2004; NIE *et al.*, 2012; ZHANG; MA; LI, 2011), rearrangement (GUISINGER *et al.*, 2011) and even the lack of one or both regions, as for the legume tribe Fabeae and related groups (PALMER; THOMPSON, 1982; WU *et al.*, 2011), showing that they could be the primary target of variation within cp genome. The H-disp regions described confirm such observation, since the *ycf*1 (JIANG *et al.*, 2017; YAN, M. *et al.*, 2018), ndhB (MOWER; VICKREY, 2018) and *rps*15 genes (FAN, W.-B. *et al.*, 2018; M. SALIH *et al.*, 2017) are within the IRs and were already mentioned as variable and considered genomic markers.

In general, plastid genomes are uniparentally inherited (GREINER; SOBANSKI; BOCK, 2015), and the entire genome is transferred intact from one species to another following a single hybridization event, what can lead to populations within one species carrying the chloroplast genome of another (related) species (PERCY *et al.*, 2014). We observed that *I. cavalcantei* and *I. marabaensis* putative hybrids carry the cp genome of *I. cavalcantei*. As such putative hybrids present phenotypes that range from *I. cavalcantei* and *I. marabaensis*, our observations remain consistent with the hybridization hypothesis but advance by indicating the origin of the cp genome. Such hybrid individuals are rare, and the analysis of additional specimens will be needed to demonstrate if this is a characteristic of all putative hybrids. There is the possibility that we observed what is known as “chloroplast capture” that occurs when the chloroplast genome *mat*ches the ‘wrong’ species, due to a past hybridization event (FOLK; MANDEL; FREUDENSTEIN, 2016). The broad phenotypic range observed indicates that chloroplast capture did not occur.

Our future research efforts will also target the nuclear genome, as we will be able to relate genes to one or the other species. The AAF approach will also be interesting, as it may reveal an intermediate pattern, as well as the H-disp of the different putative hybrid specimens.

Currently, legal obligations concern the conservation of species, but it will be of interest to understand the hybrids as a source of genomic innovation for both *I. cavalcantei* and *I. marabaensis*.

3 ANÁLISES ADICIONAIS

3.1 Characterization of the complete chloroplast genome of two Amazonian species of *Isoetes* (Isoetales) showing different habitat preferences.

Isoetes is the single living representative genus of the Isoetales order. Despite its ancient origins, its worldwide distribution, and adaptation to diverse habitats, *Isoetes* has a highly conserved morphology, consisting of a lobed subterranean bulb (corm) producing a downward tuft similar to a shoot axis, from which lateral organs named rootles develop, and upwards leaves with four air chambers. In this study, the whole chloroplast genome of eight *Isoetes* specimens from Carajás National Forest were sequenced using the shotgun Illumina technology and *de novo* assembled using NOVOPlasty (DIERCKXSENS; MARDULYN; SMITS, 2017) (Tab. 5). The objective was to better understand the differences between two *Isoetes* species found in savannah-like ecosystems known as Canga, which presents shallow soils (0 –10 cm), often containing potentially phytotoxic levels of metals. Both *Isoetes* species are aquatic, but *I. serracarajensis* can survive in seasonal lakes and ponds, while *I. cangae* occurs submerged in a single natural lake (NUNES *et al.*, 2015; PEREIRA *et al.*, 2016).

Table 5. Species. ITV code, collectors and Location from the eight *Isoetes* individuals studied.

Name	Code	City	Sector
I. serracarajasensis	414	Canaã	Bocaina
I. serracarajasensis	417	Canaã	Bocaina
I. serracarajasensis	2788	Parauapebas	N6
I. serracarajasensis	2781	Parauapebas	N4
I. cangae	1997	Canaã	S11D
I. cangae	2008	Canaã	S11D
I. cangae	2009	Canaã	S11D
I. cangae	2012	Canaã	S11D

The total DNA of eight specimens was sequenced through paired-end shotgun strategy using the NextSeq 500 Illumina platform, generating a total of 155.2 Gb of data (Tab. 6). All datasets presented a high average quality and depth of coverage (DP) according to the respective cp genome. The de novo assembling was conducted using both NOVOPlasty version 2.6.3, an assembler specifically created for assembling organelle genomes from total DNA reads. The config file was set as follows: insert size 300, read length 150, type chloro, genome range 120k-200k, k-mer 39, paired end mode and original dataset with the full content of the DNA extracted as input. The main seeds used for capture and contig extension were *I. flaccida* (KAROL *et al.*, 2010) entire genes *psb*K, *psa*C, *psb*B and *ndh*A. For *I. cangae* specimens, *psa*C and *ndh*A seeds generated circularized assemblies. The resulting contigs were assembled in Geneious R11 (KEARSE *et al.*, 2012) and the consensus sequences were annotated in CpGAVAS (Chloroplast Genome Annotation, Visualization, Analysis and GenBank Submission Tool) webserver (LIU *et al.*, 2012).

Table 6. Sequencing metrics for each dataset. Qual (quality). DP (depth of coverage).

Name	Code	File size	reads	Kbp	Mapped reads	Genome size	DP	% cp reads
I. serracarajasensis	ITV414	13.6G	39299128	2957579	3077468	143.390	3241x	7.83%
I. serracarajasensis	ITV417	32G	91089504	13701142	15811	143.442	16.6x	0.02%
I. serracarajasensis	ITV2788	13.8G	39963064	6013502	270773	143.264	285x	0.7%
I. serracarajasensis	ITV2781	15.2G	43616358	6563663	362914	143.264	382.5x	0.8%
I. cangae	ITV1997	16.8G	48301446	7265852	1936805	143.403	2039x	4%
I. cangae	ITV2008	19.8G	57190814	8603589	2515970	143.401	2649x	4.4%
I. cangae	ITV2009	22G	59814044	8997506	3397356	143.401	3817x	5.7%
I. cangae	ITV2012	22G	62985712	9476027	1496847	143.401	1576x	2.4%

Both species showed very similar genomes (Fig. 7-8), with the typical quadripartite structure. The cp genome metrics are presented in Table 7. The complete chloroplast genome contained a total of 114 unique genes (three *trn* duplicated *trn*N-GTT, *trn*R-ACG and *trn*V-GAC, along with *ndh*K, *rps*12 and the IR double copies), including 87 protein-coding genes, 27 *trn*A genes (2 duplicated) and 4 ribosomal RNA genes (duplicated in IR regions). The most evident difference between the two species was the lack of the 23S rRNA gene in the IRa of the four specimens of *I. serracarajasensis*.

Table 7. Plastome metrics for each compartment of the *Isoetes* plastomes. IC: *Isoetes cangae*, IS: *Isoetes serracarajasensis*.

	IC ITV1997		**IC ITV2008**		**IC ITV2009**		**IC ITV2012**	
	Length	%GC	Length	%GC	Length	%GC	Length	%GC
Total	143.403	37.7	143.401	37.7	143.401	37.7	143.401	37.7
LSC	90.835	36.2	90.832	36.1	90.832	36.1	90.832	36.1
IR	12.744	48.1	12.744	48.1	12.744	48.1	12.744	48.1
SSC	27.081	33.1	27.082	33.1	27.082	33.1	27.082	33.1
	IS ITV414		**IS ITV417**		**IS ITV2781**		**IS ITV2788**	
	Length	%GC	Length	%GC	Length	%GC	Length	%GC
Total	143.390	37.7	143.442	37.7	143.264	37.7	143.264	37.7
LSC	90.632	36.2	90.683	36.2	90.707	36.2	90.707	36.2
IR	12.816	47.9	12.816	47.9	12.744	48.1	12.744	48.1
SSC	27.070	33.2	27.071	33.2	27.069	33.2	27.069	33.2

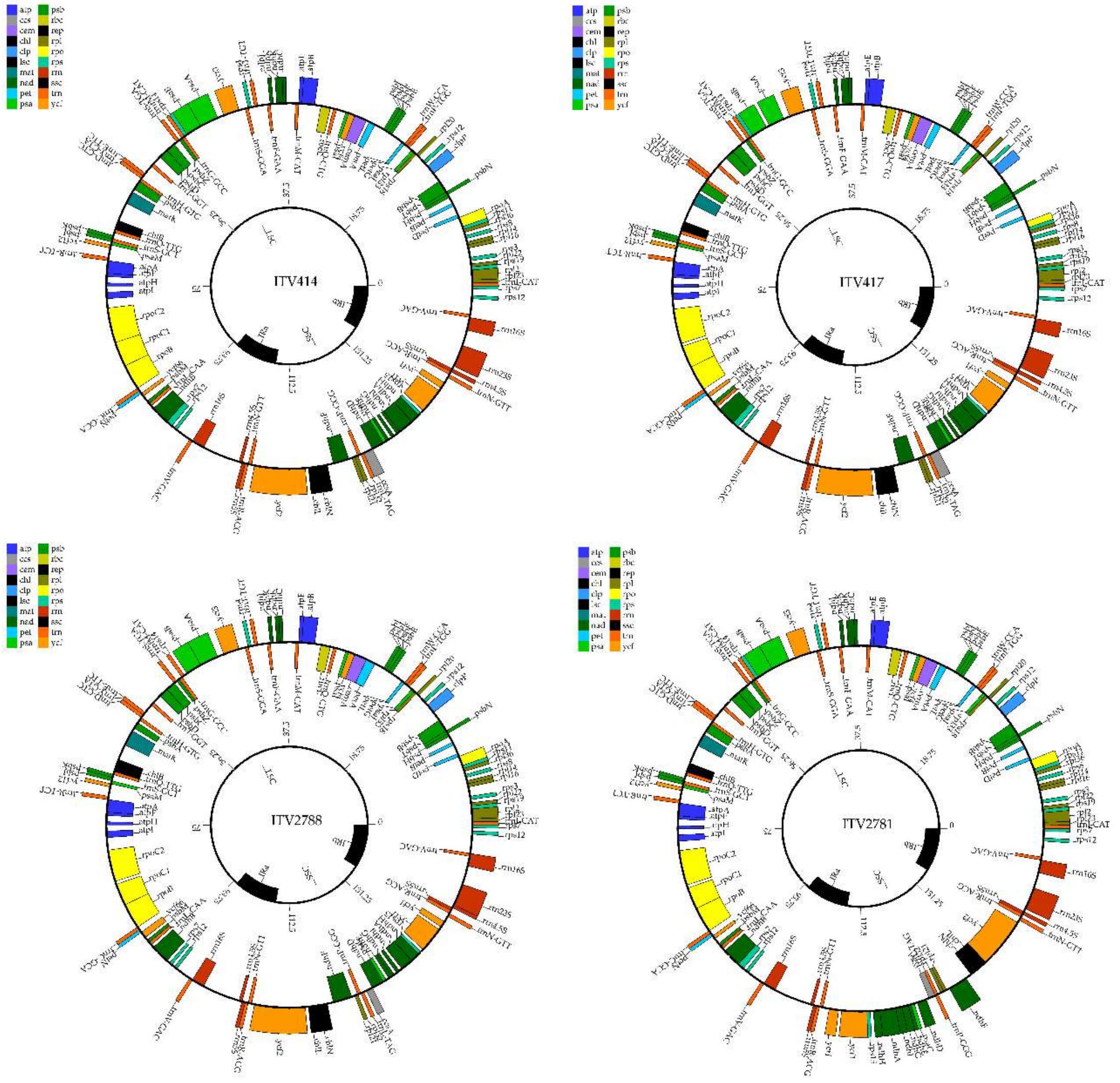


Figure 7. *Isoetes serracarajasensis* plastomes with the lack of 23S gene.

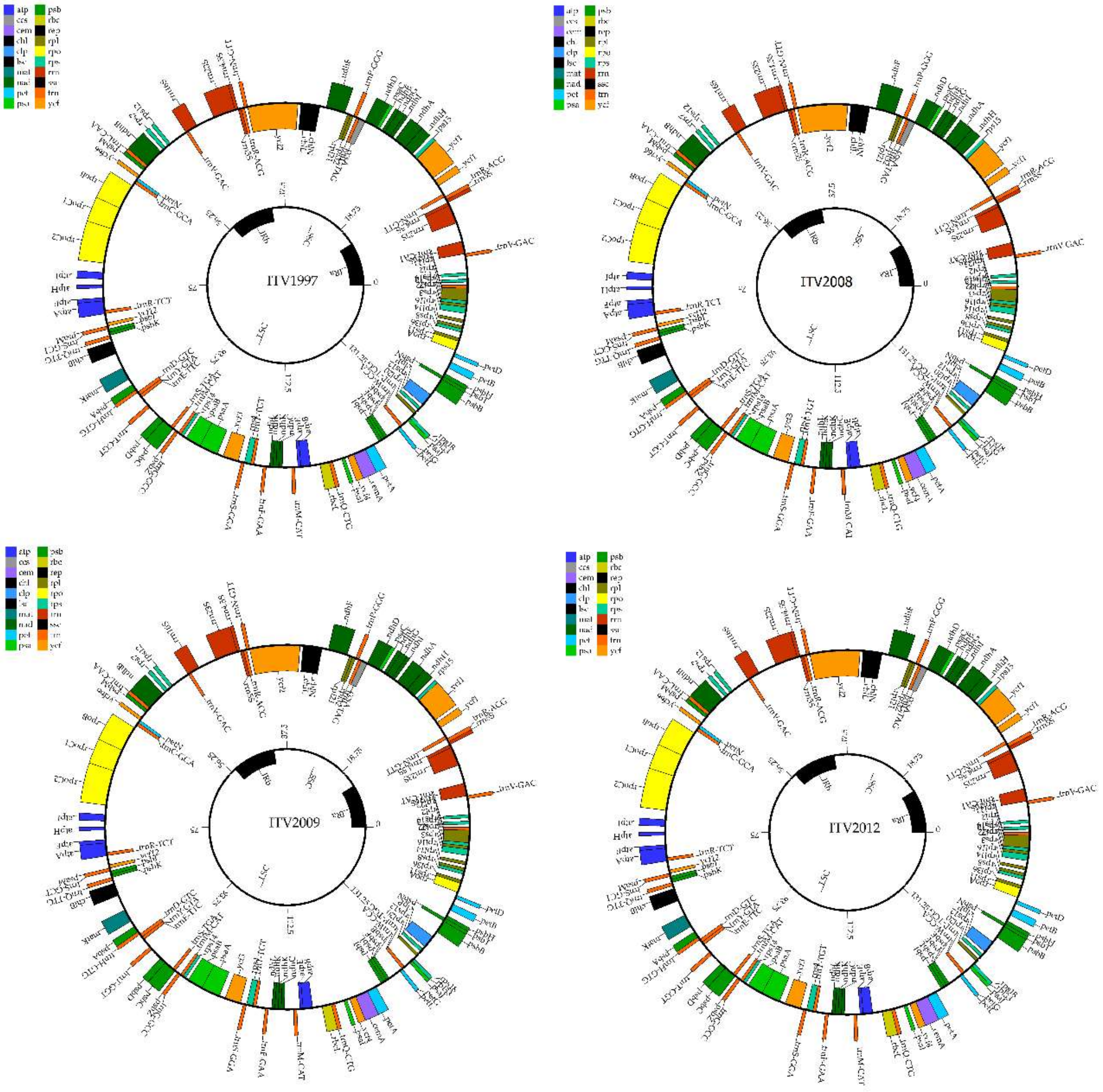


Figure 8. *Isoetes cangae* plastomes.

We could observe a clear separation between *Isoetes* species from Carajás and the six reference genomes (*I. flaccida, I valida, I. butleri, I. melanospora, I. engelmannii* and *I. nuttallii*), while also presenting separated clades for *I. serracarajasensis* and *I. cangae* specimens (Fig. 9).

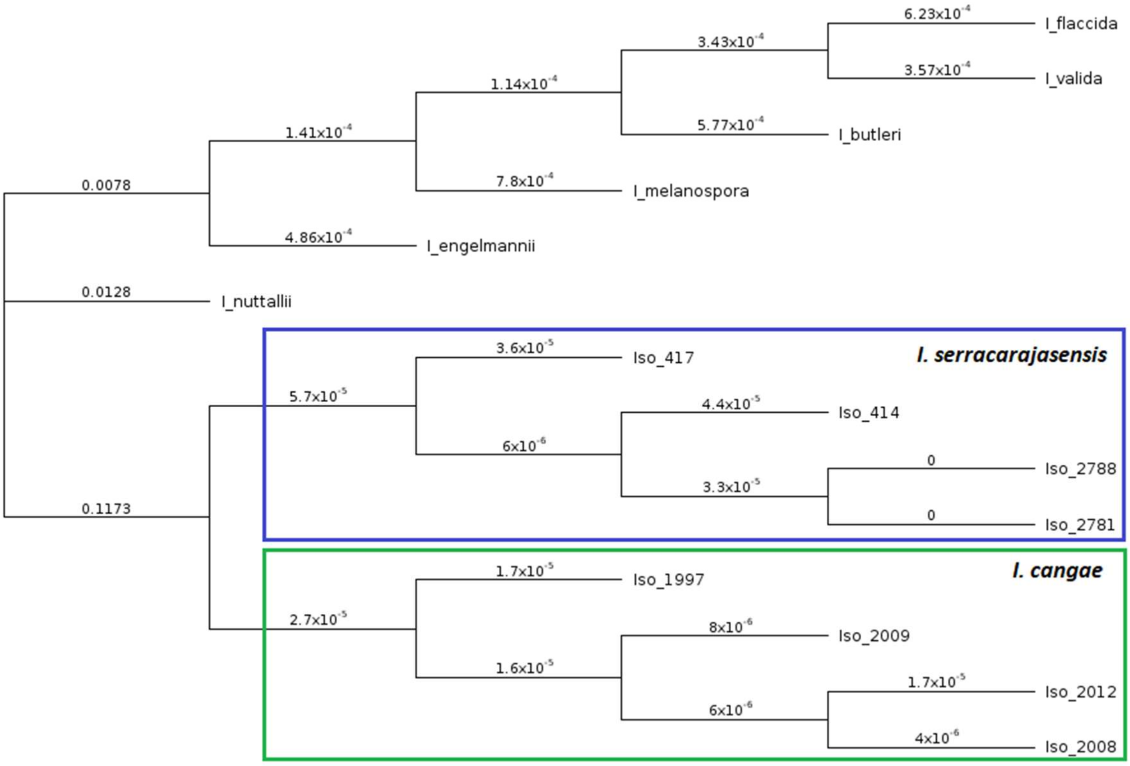


Figure 9. MAFFT multiple alignment tree encompassing the eight *Isoetes* specimens and six reference plastomes.

The whole chloroplast genome is useful for plant evolutionary and population genomic studies, providing important insights into the conservation and maintenance of genetic resource, specially concerning a global biodiversity "hotspot" region as the Carajás Mountain range, within the Amazonian Rain Forest. This report provides essential data for further study on the accurate identification and phylogenetic resolution of *Isoetes* species.

3.2 Characterization of the complete chloroplast genome of two *Philodendron* species (Araceae)

Philodendron Schott is one of the largest Neotropical plant taxa, being the second most species-rich genus within Araceae. The genus shows considerable morphological and ecological diversity, occurring mainly in the humid forests of tropical America and regarded as one of the most important epiphytic components of the Neotropical flora (IRUME *et al.*, 2013; SAKURAGUI; MAYO, 1997). According to (VASCONCELOS *et al.*, 2018) *Philodendron* was recovered as strongly monophyletic with all the three main morphological subdivisions (*Thaumatophyllum*, *P.* subg. *Pteromischum* and *P.* subg. *Philodendron*) as independent lineages within the Homalomena clade. Two species, *P. wullschlaegelii* and *P. carajasense* were sequenced through Illumina platform and *de novo* assembled using NOVOPlasty. We report the first two complete chloroplast genomes of Philodendron.

The total DNA was sequenced using the paired-end shotgun strategy in the NextSeq 500 Illumina platform, generating a total of 155.2G data (Tab. 8). All datasets presented a high average quality and depth of coverage (DP) according to the respective cp genome. The *de novo* assembling was conducted using both NOVOPlasty version 2.6.3, an assembler specifically created for assembling organelle genomes from total DNA reads. The config file was set as follows: insert size 300, read length 150, type chloro, genome range 120k-200k, k-mer 39, paired end mode and original dataset with the full content of the DNA extracted as input. The resulting contigs were assembled in Geneious R11 and the consensus sequences were annotated in CpGAVAS (Chloroplast Genome Annotation, Visualization, Analysis and GenBank Submission Tool) webserver.

Table 8. Sequencing metrics for each dataset. Qual (quality). DP (depth of coverage).

Name	Code	File size	reads	Kbp	Mapped reads	Genome size	DP
P. wullschlaegelii	ITV1706	22G	91671928	138424611	2085047	163.105	1930x
P. carajasense	ITV3037	28G	87813424	13259827	1397412	162.119	2603x

Both species showed very similar genomes (Fig. 10), with the typical quadripartite structure. The cp genome metrics are presented in Table 9.

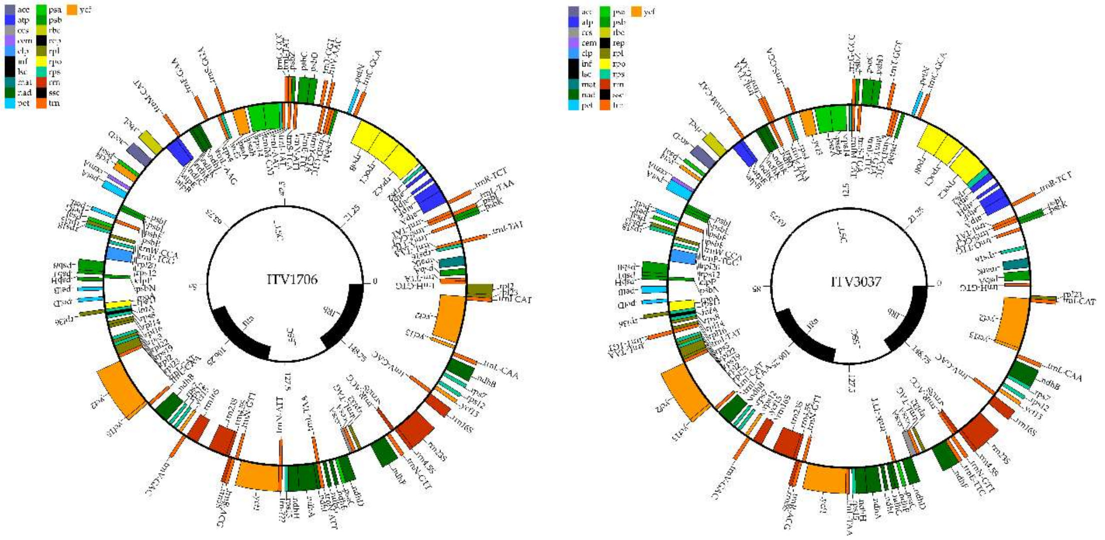


Figure 10. Complete plastomes from *P. wullschlaegelii* (ITV1706) and *P. carajasense* (ITV3037).

Table 9. Length and %GC for each cp compartment of the two *Philodendron* plastomes assembled.

	1706 *P. wullschlaegelii*	
	length	%GC
Total	163.105	36.0
LSC	90.316	34.0
IR	26.056	42.1
SSC	20.677	29.7
	3037 *P. carajasense*	
	length	%GC
Total	162.119	36.3
LSC	89.991	34.2
IR	25.966	42.3
SSC	20.196	30.5

P. wullschlaegelii presented 121 genes, 31 from it transfer RNA's and 11 duplicated (up to 4 copies) genes (besides IR copies). *P. carajasense* presented 120 genes, 27 transfer RNA's and 11 duplicated (up to 4 copies) genes (besides IR copies).

4 REFERENCES

ANDREWS, S. *Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data*. . [S.l: s.n.]. Disponível em: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. Acesso em: 9 jun. 2021. , 2010

ASAF, S. *et al.* The complete chloroplast genome of wild rice (Oryza minuta) and its comparison to related species. *Frontiers in Plant Science*, v. 8, n. March, p. 1–15, 2017.

AUSTIN, D. F.; HUAMAN, Z. A Synopsis of Ipomoea (Convolvulaceae) in the Americas. *Taxon*, v. 45, n. 1, p. 3, fev. 1996. Disponível em: <https://www.jstor.org/stable/1222581?origin=crossref>. Acesso em: 31 maio 2018.

BABIYCHUK, E. *et al.* Natural history of the narrow endemics Ipomoea cavalcantei and I. marabaensis from Amazon Canga savannahs. *Scientific Reports 2017 7:1*, v. 7, n. 1, p. 1–15, 8 ago. 2017. Disponível em: <https://www.nature.com/articles/s41598-017-07398-z>. Acesso em: 7 maio 2022.

BANKEVICH, A. *et al.* SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, v. 19, n. 5, p. 455–477, 1 maio 2012. Disponível em: </pmc/articles/PMC3342519/>. Acesso em: 9 jun. 2021.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, v. 30, n. 15, p. 2114–2120, 1 ago. 2014. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/24695404/>. Acesso em: 9 jun. 2021.

DARLING, A. C. E. *et al.* Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research*, v. 14, n. 7, p. 1394–403, 1 jul. 2004. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/15231754>. Acesso em: 31 maio 2018.

DIERCKXSENS, N.; MARDULYN, P.; SMITS, G. NOVOPlasty: De novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research*, v. 45, n. 4, 2017.

DONG, W. *et al.* Highly Variable Chloroplast Markers for Evaluating Plant Phylogeny at Low Taxonomic Levels and for DNA Barcoding. *PLoS ONE*, v. 7, n. 4, p. e35071, 12 abr. 2012. Disponível em: <http://dx.plos.org/10.1371/journal.pone.0035071>. Acesso em: 31 maio 2018.

DONG, W. *et al.* ycf1, the most promising plastid DNA barcode of land plants. *Scientific Reports*, v. 5, n. 1, p. 8348, 12 fev. 2015.

FAN, H. *et al.* An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. *BMC Genomics*, v. 16, n. 1, p. 1–18, 14 jul. 2015. Disponível em: <https://sourceforge.net/projects/aaf->. Acesso em: 9 jun. 2021.

FAN, W.-B. *et al.* Comparative Chloroplast Genomics of Dipsacales Species: Insights Into Sequence Variation, Adaptive Evolution, and Phylogenetic Relationships. *Frontiers in Plant Science*, v. 9, p. 689, 23 maio 2018. Disponível em: <https://www.frontiersin.org/article/10.3389/fpls.2018.00689/full>. Acesso em: 21 mar. 2019.

FOLK, R. A.; MANDEL, J. R.; FREUDENSTEIN, J. V. Ancestral Gene Flow and Parallel Organellar Genome Capture Result in Extreme Phylogenomic Discord in a Lineage of Angiosperms.

Systematic Biology, v. 66, n. 3, p. syw083, 16 set. 2016. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/27637567>. Acesso em: 31 maio 2018.

GREINER, S.; SOBANSKI, J.; BOCK, R. Why are most organelle genomes transmitted maternally? *BioEssays*, v. 37, n. 1, p. 80–94, 10 jan. 2015.

GROUP, T. A. P. *et al.* An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society*, v. 181, n. 1, p. 1–20, 1 maio 2016. Disponível em: <https://academic.oup.com/botlinnean/article/181/1/1/2416499>. Acesso em: 16 set. 2021.

GUISINGER, M. M. *et al.* Extreme Reconfiguration of Plastid Genomes in the Angiosperm Family Geraniaceae: Rearrangements, Repeats, and Codon Usage. *Molecular Biology and Evolution*, v. 28, n. 1, p. 583–600, 1 jan. 2011. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/20805190>. Acesso em: 31 maio 2018.

HASHIMOTO, M. *et al.* A nucleus-encoded factor, CRR2, is essential for the expression of chloroplast *ndhB* in *Arabidopsis*. *The Plant Journal*, v. 36, n. 4, p. 541–549, 17 nov. 2003.

HOSHINO, A. *et al.* Genome sequence and analysis of the Japanese morning glory Ipomoea nil. *Nature Communications*, v. 7, n. 1, p. 13295, 8 dez. 2016. Disponível em: <http://www.nature.com/articles/ncomms13295>. Acesso em: 21 mar. 2019.

IRUME, M. V. *et al.* Floristic composition and community structure of epiphytic angiosperms in a terra firme forest in central Amazonia. *Acta Botanica Brasilica*, v. 27, n. 2, p. 378–393, jun. 2013.

JIANG, D. *et al.* The Chloroplast Genome Sequence of Scutellaria baicalensis Provides Insight into Intraspecific and Interspecific Chloroplast Genome Diversity in Scutellaria. *Genes 2017, Vol. 8, Page 227*, v. 8, n. 9, p. 227, 13 set. 2017. Disponível em: <https://www.mdpi.com/2073-4425/8/9/227/htm>. Acesso em: 16 set. 2021.

KAROL, K. G. *et al.* Complete plastome sequences of Equisetum arvense and Isoetes flaccida: implications for phylogeny and plastid genome evolution of early land plant lineages. *BMC Evolutionary Biology*, v. 10, n. 1, p. 321, 2010.

KATOH, K. *et al.* MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, v. 30, n. 14, p. 3059–3066, 15 jul. 2002. Disponível em: <https://academic.oup.com/nar/article/30/14/3059/2904316>. Acesso em: 9 jun. 2021.

KEARSE, M. *et al.* Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, v. 28, n. 12, p. 1647–1649, 15 jun. 2012. Disponível em: <https://academic.oup.com/bioinformatics/article/28/12/1647/267326>. Acesso em: 16 set. 2021.

KIKUCHI, S. *et al.* Uncovering the Protein Translocon at the Chloroplast Inner Envelope Membrane. *Science*, v. 339, n. 6119, p. 571–574, fev. 2013.

KIM, K.-J. Complete Chloroplast Genome Sequences from Korean Ginseng (Panax schinseng Nees) and Comparative Analysis of Sequence Evolution among 17 Vascular Plants. *DNA Research*, v. 11, n. 4, p. 247–261, 1 jan. 2004.

KOLODNER, R.; TEWARI, K. K. Inverted repeats in chloroplast DNA from higher plants. *Proceedings of the National Academy of Sciences*, v. 76, n. 1, p. 41–45, jan. 1979.

LANES, É. C. *et al.* Landscape genomic conservation assessment of a narrow-endemic and a widespread morning glory from amazonian savannas. *Frontiers in Plant Science*, v. 9, p. 532, 7 maio 2018. Disponível em: <www.frontiersin.org>. Acesso em: 6 jun. 2021.

LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, v. 9, n. 4, p. 357–359, 4 abr. 2012. Disponível em: <https://www.nature.com/articles/nmeth.1923>. Acesso em: 9 jun. 2021.

LIU, C. *et al.* CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. *BMC Genomics*, v. 13, n. 1, p. 715, 20 dez. 2012. Disponível em: <http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-13-715>. Acesso em: 21 mar. 2019.

LOWE, T. M.; EDDY, S. R. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Research*, v. 25, n. 5, p. 955–964, 1 mar. 1997.

M. SALIH, R. H. *et al.* Complete chloroplast genomes from apomictic Taraxacum (Asteraceae): Identity and variation between three microspecies. *PLOS ONE*, v. 12, n. 2, p. e0168008, 9 fev. 2017.

MARÇAIS, G.; KINGSFORD, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, v. 27, n. 6, p. 764–770, 2011.

MARTIN, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, v. 17, n. 1, p. 10, 2 maio 2011. Disponível em: <http://www-huber.embl.de/users/an->. Acesso em: 15 dez. 2020.

MCKAIN, M. R. *et al.* Practical considerations for plant phylogenomics. *Applications in Plant Sciences*, v. 6, n. 3, p. 1–15, 2018.

MCNEAL, J. R. *et al.* Complete plastid genome sequences suggest strong selection for retention of photosynthetic genes in the parasitic plant genus Cuscuta. *BMC Plant Biology*, v. 7, n. 1, p. 57, 24 out. 2007. Disponível em: <http://bmcplantbiol.biomedcentral.com/articles/10.1186/1471-2229-7-57>. Acesso em: 29 jan. 2019.

MENEZES, A. P. A. *et al.* Chloroplast genomes of Byrsonima species (Malpighiaceae): comparative analysis and screening of high divergence sequences. *Scientific Reports*, v. 8, n. 1, p. 2210, 2 fev. 2018.

MORGULIS, A. *et al.* Database indexing for production MegaBLAST searches. 15 ago. 2008, [S.l.]: Oxford Academic, 15 ago. 2008. p. 1757–1764. Disponível em: <https://academic.oup.com/bioinformatics/article/24/16/1757/202524>. Acesso em: 9 jun. 2021.

MOWER, J. P.; VICKREY, T. L. Structural Diversity Among Plastid Genomes of Land Plants. *Advances in Botanical Research*. [S.l.]: Academic Press Inc., 2018. v. 85. p. 263–292. . Acesso em: 9 jun. 2021.

NEUBIG, K. M.; ABBOTT, J. R. Primer development for the plastid region *ycf1* in Annonaceae and other magnoliids. *American Journal of Botany*, v. 97, n. 6, jun. 2010.

NIE, X. *et al.* Complete chloroplast genome sequence of a major invasive species, crofton weed (Ageratina adenophora). *PLoS ONE*, v. 7, n. 5, 2012.

NUNES, J. A. *et al.* Soil-vegetation relationships on a banded ironstone 'island', Carajás Plateau, Brazilian Eastern Amazonia. *Anais da Academia Brasileira de Ciencias*, v. 87, n. 4, p. 2097–2110, 1 out. 2015. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/26648541/>. Acesso em: 8 jun. 2021.

PALMER, J. D.; THOMPSON, W. F. Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. *Cell*, v. 29, n. 2, p. 537–550, jun. 1982.

PERCY, D. M. *et al.* Understanding the spectacular failure of <scp>DNA</scp> barcoding in willows (*Salix*): Does this result from a trans-specific selective sweep? *Molecular Ecology*, v. 23, n. 19, p. 4737–4756, 15 out. 2014.

PEREIRA, J. B. D. S. *et al.* Two New Species of Isoetes (Isoetaceae) from northern Brazil. *Phytotaxa*, v. 272, n. 2, p. 141, 29 ago. 2016.

RUTHERFORD, K. *et al.* Artemis: Sequence visualization and annotation. *Bioinformatics*, v. 16, n. 10, p. 944–945, 1 out. 2000. Disponível em: <http://www.acedb.org/>. Acesso em: 9 jun. 2021.

RUWE, H.; SCHMITZ-LINNEWEBER, C. Short non-coding RNA fragments accumulating in chloroplasts: footprints of RNA binding proteins? *Nucleic Acids Research*, v. 40, n. 7, p. 3106–3116, abr. 2012.

SAKAMOTO, W.; TAKAMI, T. Chloroplast DNA Dynamics: Copy Number, Quality Control and Degradation. *Plant and Cell Physiology*, v. 59, n. 6, p. 1120–1127, 1 jun. 2018. Disponível em: <https://academic.oup.com/pcp/article/59/6/1120/4980302>. Acesso em: 21 mar. 2019.

SAKURAGUI, C. M.; MAYO, S. J. Three New Species of Philodendron (Araceae) from South-Eastern Brazil. *Kew Bulletin*, v. 52, n. 3, p. 673, 1997.

SCHMIEDER, R.; EDWARDS, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, v. 27, n. 6, 2011.

SEBASTIANI, F.; CARNEVALE, S.; VENDRAMIN, G. G. A new set of mono- and dinucleotide chloroplast microsatellites in Fagaceae. *Molecular Ecology Notes*, v. 4, n. 2, p. 259–261, jun. 2004. Disponível em: <http://doi.wiley.com/10.1111/j.1471-8286.2004.00635.x>. Acesso em: 31 maio 2018.

SHEPHERD, L. D. *et al.* Evidence of a Strong Domestication Bottleneck in the Recently Cultivated New Zealand Endemic Root Crop, Arthropodium cirratum (Asparagaceae). *PLOS ONE*, v. 11, n. 3, p. e0152455, 1 mar. 2016. Disponível em: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0152455>. Acesso em: 16 set. 2021.

SHINOZAKI, K. *et al.* The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *The EMBO Journal*, v. 5, n. 9, p. 2043–2049, [S.d.]. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1460-2075.1986.tb04464.x>. Acesso em: 10 jun. 2018.

SIMÃO-BIANCHINI, R.; VASCONCELOS, L. V.; PASTORE, M. Flora das cangas da Serra dos Carajás, Pará, Brasil: Convolvulaceae. *Rodriguesia*, v. 67, n. 5, p. 1301–1318, 1 dez. 2016. Disponível em: <http://rodriguesia.jbrj.gov.br>. Acesso em: 8 jun. 2021.

STAMATAKIS, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, v. 30, n. 9, p. 1312–1313, 1 maio 2014. Disponível em: <https://academic.oup.com/bioinformatics/article/30/9/1312/238053>. Acesso em: 9 jun. 2021.

STRAUB, S. C. K. *et al.* Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *American Journal of Botany*, v. 99, n. 2, p. 349–364, 1 fev. 2012. Disponível em: <http://www.amjbot.org/>. Acesso em: 9 jun. 2021.

SUGIURA, M. The chloroplast genome. *Plant Molecular Biologylecular*, v. 19, p. 149–168, 1992.

SUN, Y. *et al.* Phylogenomic and structural analyses of 18 complete plastomes across nearly all families of early-diverging eudicots, including an angiosperm-wide analysis of IR gene content evolution. *Molecular Phylogenetics and Evolution*, v. 96, p. 93–101, mar. 2016. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/26724406>. Acesso em: 31 maio 2018.

TONTI-FILIPPINI, J. *et al.* What can we do with 1000 plastid genomes? *Plant Journal*, v. 90, n. 4, p. 808–818, 1 maio 2017. Disponível em: <http://www.bioplatforms.com/dna-barcoding/>. Acesso em: 9 jun. 2021.

VASCONCELOS, S. *et al.* New insights on the phylogenetic relationships among the traditional Philodendron subgenera and the other groups of the Homalomena clade (Araceae). *Molecular Phylogenetics and Evolution*, v. 127, p. 168–178, out. 2018.

VIANA, P. L. *et al. Flora das cangas da Serra dos Carajás, Pará, Brasil: História, área de estudos e metodologia*. *Rodriguesia*. [S.l.]: Instituto de Pesquisas Jardim Botanico do Rio de Janeiro. Disponível em: <http://rodriguesia.jbrj.gov.br>. Acesso em: 8 jun. 2021. , 1 dez. 2016

VIANA, P. L.; GIULIETTI-HARLEY, A. M. *Flora das cangas de carajás: Taxonomia preparando novos caminhos*. *Rodriguesia*. [S.l.]: Instituto de Pesquisas Jardim Botanico do Rio de Janeiro. Disponível em: <http://rodriguesia.jbrj.gov.br>. Acesso em: 8 jun. 2021. , 1 jul. 2018

WICKE, S. *et al.* The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Molecular Biology*, v. 76, n. 3–5, p. 273–297, 22 jul. 2011.

WILLIAMS, A. V. *et al.* Integration of complete chloroplast genome sequences with small amplicon datasets improves phylogenetic resolution in Acacia. *Molecular Phylogenetics and Evolution*, v. 96, p. 1–8, 2016. Disponível em: <http://dx.doi.org/10.1016/j.ympev.2015.11.021>.

WU, C.-S. *et al.* Loss of Different Inverted Repeat Copies from the Chloroplast Genomes of Pinaceae and Cupressophytes and Influence of Heterotachy on the Evaluation of Gymnosperm Phylogeny. *Genome Biology and Evolution*, v. 3, p. 1284–1295, 1 jan. 2011.

YAMORI, W.; SHIKANAI, T. Physiological Functions of Cyclic Electron Transport Around Photosystem I in Sustaining Photosynthesis and Plant Growth. *Annual Review of Plant Biology*, v. 67, n. 1, p. 81–106, 29 abr. 2016.

YAN, L. *et al.* Analyses of the Complete Genome and Gene Expression of Chloroplast of Sweet Potato [Ipomoea batata]. *PLOS ONE*, v. 10, n. 4, p. e0124083, 15 abr. 2015. Disponível em:

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0124083>. Acesso em: 9 set. 2021.

YAN, M. *et al.* The Application and Limitation of Universal Chloroplast Markers in Discriminating East Asian Evergreen Oaks. *Frontiers in Plant Science*, v. 9, 8 maio 2018.

ZHANG, Y.-J.; MA, P.-F.; LI, D.-Z. High-Throughput Sequencing of Six Bamboo Chloroplast Genomes: Phylogenetic Implications for Temperate Woody Bamboos (Poaceae: Bambusoideae). *PLoS ONE*, v. 6, n. 5, p. e20596, 31 maio 2011.

5 SUPPLEMENTARY MATERIAL

S.Table 1. Allelic frequency, GC content, length, the three more frequent amino acids (respectively), the three most frequent codons (besides ATG and TGG) and the genes of each junction between cp genome compartments

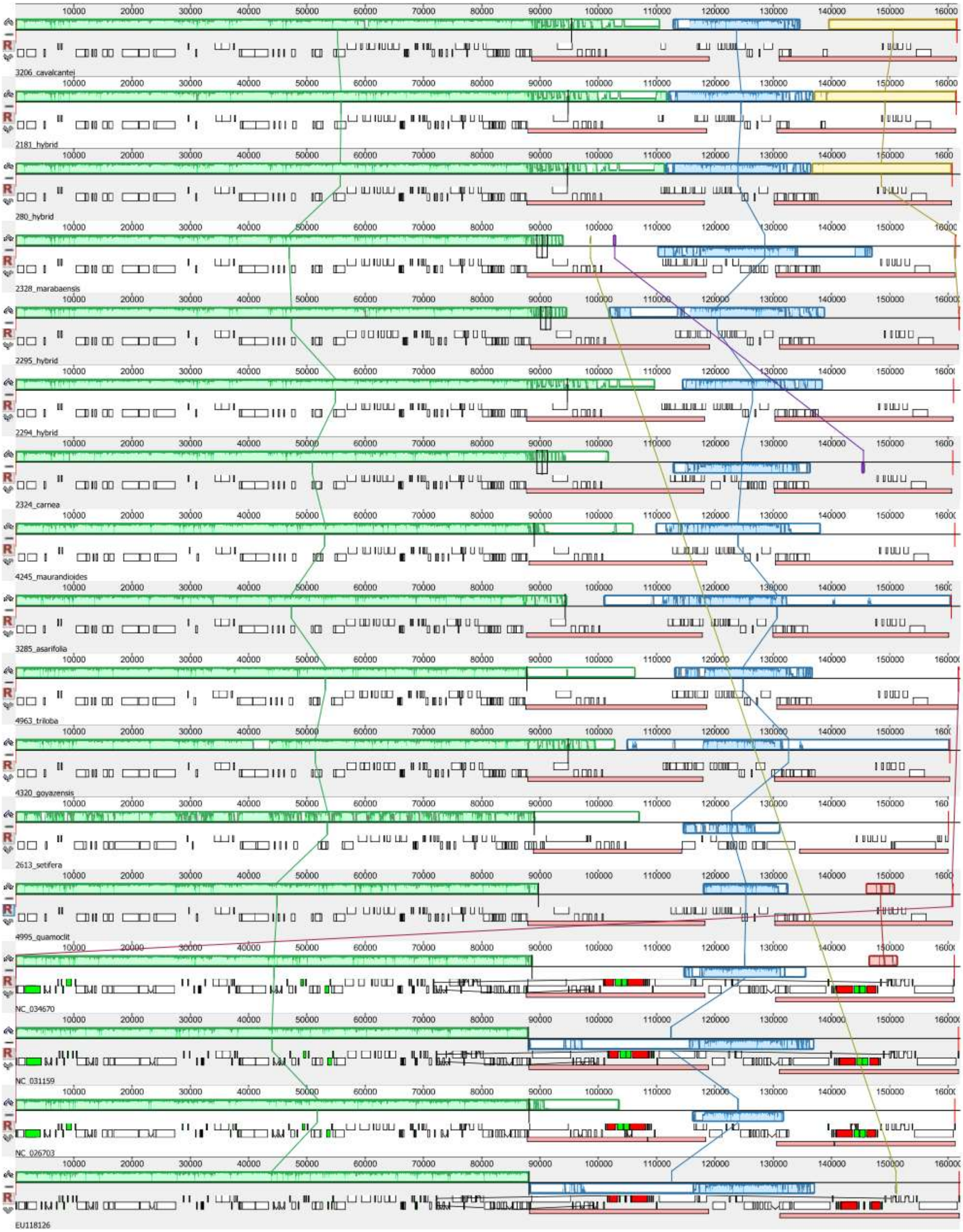
	I. marabaensis	***I. cavalcantei***	**2181 put. hybrid**	**280 put. hybrid**	**2294 put. hybrid**	**2295 put. hybrid**	***I. maurandioides***	***I. asarifolia***
Total lenght	161.324	161.563	161.495	160.765	160.974	161.948	161.242	160.589
LSC lenght	87.817	88.602	87.914	87.752	87.684	88.417	88.144	87.686
N. genes	63	63	63	63	63	63	63	63
trn	14	14	14	14	14	14	14	15
LSC GC%	36.1	36	36.1	36.1	36.1	36	36.1	36
LSC A%	31.3	31.4	31.3	31.3	31.3	31.5	31.3	31.4
LSC C%	18.4	18.3	18.4	18.4	18.4	18.3	18.4	18.4
LSC G%	17.6	17.7	17.6	17.6	17.7	17.7	17.6	17.6
LSC T%	32.6	32.5	32.6	32.6	32.6	32.5	32.6	32.6
LSC AA%	L/I/S	L/S/I	L/I/S	L/S/I	L/S/I	L/S/I	L/I/S	L/I/S
codon usage	GAT/GAA/AAA	GAT/CAA/GAA	GAT/CAA/GAA	GAT/CAA/GAA	CAA/GAT/CAT	GAT/CAA/GAA	GAT/GAA/CAA	CAA/GAA/AAT
Junction LSC-IRa	*rpl*2,*rpl*23/*trn*I,*Ycf*2	*rpl*2,*rpl*23/*trn*I,*Ycf*2	*rpl*2,*rpl*23/*trn*I,*Ycf*2	*rpl*2,*rpl*23/*trn*I,*Ycf*2	*rpl*2,*rpl*23/*trn*I,*Ycf*2	*rpl*2,*rpl*23/*trn*I,*Ycf*2	*rpl*2,*rpl*23,*trn*I/*Ycf*2	*rpl*23,*trn*I/*Ycf*2
IRa lenght	30.717	30.444	30.755	30.471	30.651	30.730	30.529	30.194
N. genes	13	13	13	13	13	13	13	13
trn	5	5	5	5	5	5	5	4
IRa GC%	40.6	40.8	40.6	40.8	40.7	40.7	40.7	40.9
IRa A%	31.1	31.1	31.1	31.1	31.1	31.2	31.3	31.1
IRa C%	19.2	19.2	19.2	19.2	19.2	19.2	19.2	19.2
IRa G%	21.5	21.6	21.5	21.6	21.4	21.5	21.5	21.6
IRa T%	28.2	28.1	28.2	28.1	28.3	28.2	28	28
IRa AA%	S/L/R	S/L/R	L/S/R	S/L/R	S/L/R	L/S/R	L/S/R	L/S/I
codon usage	GAT/GAA/TAT	GAT/AAT/CAT	GAT/AAT/CAA	GAT/AAT/CAT	GAT/GAA/TAT	GAT/TAT/AAT	GAT/AAT/CAA	AAT/CAA/GAT
Junction Ira-SSC	ndhH,ndhA/ndhF,*rpl*32	ndhH,ndhA-ndhI,ndhG	ndhH,ndhA-ndhI,ndhG	ndhH,ndhA-ndhI,ndhG	ndhH,ndhA-ndhI,ndhG	ndhH,ndhA-ndhI,ndhG	ndhH,ndhA-ndhI,ndhG	ndhH,ndhA-mdhI,ndhG
SSC lenght	11.979	11.979	11.977	11.977	11.988	11.977	11.716	12.105
N. genes	9	9	9	9	9	9	9	9
trn	1	1	1	1	1	1	1	1
SSC GC%	32.3	32.3	32.3	32.3	32.3	32.3	32.2	32.3
SSC A%	36.5	31.2	31.2	31.2	31.2	31.2	31.2	31.1
SSC C%	16.6	15.7	15.7	15.7	15.7	15.7	15.7	15.8
SSC G%	15.7	16.6	16.6	16.6	16.6	16.6	16.5	16.5
SSC T%	31.2	36.5	36.5	36.5	36.5	36.5	36.5	36.6
SSC AA%	I/L/K	L/F/I	L/I/F	L/I/F	L/F/I	L/F/I	L/F/I	L/F/I
codon usage	CAA/GAT/AAA	GAT/CAA/AAA	CAT/AAT/CAA	CAT/AAT/CAA	GAT/CAA/TAT	AAA/AAT/GAT	GAT/AAA/GAA	GAA/TAT/CAA
Junction SSC-IRb	ndhI,ndhA-ndhH,*rps*15	ndhF/ndhA	ndhF/ndhA	ndhF/ndhA,ndhH	ndhF/ndhA,ndhH	ndhF/ndhA,ndhH	ndhF,ndhA-ndhH,*rps*15	ndhF/ndhA,ndhH
IRb lenght	30717	30.444	30.755	30.471	30.651	30.730	30.529	30.194
N. genes	13	13	13	13	13	13	13	13
trn	5	5	5	5	5	5	5	5
IRb GC%	40.6	40.8	40.6	40.8	40.7	40.7	40.7	40.9
IRb A%	28.2	28.1	28.2	28.1	28.3	28.2	28	28
IRb C%	21.5	21.6	21.5	21.6	21.4	21.5	21.5	21.6
IRb G%	19.2	19.2	19.2	19.2	19.2	19.2	19.2	19.3
IRb T%	31.1	31.1	31.1	31.1	31.1	31.2	31.3	31.1
IRb AA%	S/L/F	L/S/F	S/L/I	L/S/F	S/L/I	S/L/F	S/L/R	S/L/F
IRb codon usage	GAT/CAA/GAA	GAT/GAA/AAT	GAT/CAA/CAT	GAT/AAT/GAA	CAA/CAT/GAA	GAT/AAT/CAT	GAT/AAT/AAA	GAT/GAA/AAT
Junction IRb-LSC	*Ycf*2,*trn*I/*trn*H	*Ycf*2,*trn*I/*trn*H	*Ycf*2,*trn*I/*trn*H	*Ycf*2,*trn*I/*trn*H	*Ycf*2,*trn*I/*trn*H	*Ycf*2,*trn*I/*trn*H	*Ycf*2/*trn*I,*trn*H	*Ycf*2/*trn*I,*trn*H

	I. carnea	***I. triloba***	***I. goyazensis***	***I. quamoclit***	***I. setifera***	**I. trifida**	**I.nil**	**I. batatas**	**I. purpurea**
Total lenght	160.819	161.835	160.414	160.836	160.082	161.133	161.897	161.303	162.046
LSC lenght	87.660	87.538	87.930	87.911	88.879	87.649	88.117	87.823	88.172
N. genes	63	63	63	63	61				
trn	14	14	14	14	14				
LSC GC%	36.1	36.1	36.1	36.1	35.2	36.1	36.1	36.1	36
LSC A%	31.3	31.3	31.3	31.3	31.6	31.3	31.3	31.3	31.3
LSC C%	18.5	18.5	18.4	18.4	18.1	18.5	18.4	18.5	18.4
LSC G%	17.6	17.7	17.7	17.6	17.1	17.7	17.6	17.7	17.6
LSC T%	32.6	32.6	32.6	32.6	33.1	32.6	32.7	32.6	32.7
LSC AA%	S/L/I	L/S/I	S/L/I	L/S/I	L/I/S	L/I/S	L/I/S	L/S/I	L/I/S
LSC codon	GAT/CAA/GAA	GAT/CAA/GAA	CAA/GAA/GAT	GAA/CAA/GAT	GAT/CAA/AAT	GAT/CAA/GAA	CAA/GAA/GAT	CAA/GAA/GAT	GAT/CAA/GAA
Junction	*rpl*2,*rpl*23/*trn*I,*Ycf*	*rpl*2,*rpl*23/*trn*I,*Yc*	*rpl*2,*rpl*23/*trn*I,*Ycf*	*rpl*2,*rpl*23/*trn*I,*Yc*	*rpl*22,*rps*19/*rpl*2,	*rpl*2,*rpl*23/*trn*I,*Ycf*2	*rpl*2,*rpl*23/*trn*I,*Ycf*2	*rpl*2,*rpl*23/*trn*I,*Ycf*2	*rpl*2,*rpl*23/*trn*I,*Ycf*2
IRa lenght	30.502	31.106	30.098	30.474	25.543	30.716	30.847	30.690	30.882
N. genes	13	13	12	13	12				
trn	5	5	5	5	5				
IRa GC%	40.7	40.1	40.7	40.7	43.1	40.8	40.5	40.7	40.6
IRa A%	31.2	31.9	31.2	31.2	28.5	31.2	31.5	31.3	31.5
IRa C%	19.2	19	19.3	19.2	20.7	19.2	19.1	19.2	19.1
IRa G%	21.5	21.1	21.5	21.5	22.4	21.6	21.4	21.5	21.4
IRa T%	28.1	28	28.1	28.2	28.4	28	28	28	28
IRa AA%	S/L/K	S/R/L	S/R/L	L/S/R	S/L/R	S/L/R	L/S/K	S/L/R	S/R/L
IRa codon	GAA/CAT/GAT	AAA/GAA/GAT	GAA/CAA/GAT	GAT/TAT/AAT	CAA/GAT/CAT	GAT/CAT/CAA	(TGG/ATG)/GAT/CA	GAT/GAA/AAA	(TGG/ATG)/CAA/GA
Junction Ira-	ndhH,ndhA/ndhF,	ndhH,ndhA-	ndhH/ndhA,ndhI,n	ndhH,ndhA-	*trn*R,*trn*N/ndhF,*r*	*rps*15,ndhH/ndhF,*r*	*rps*15,ndhH/ndhF,*rpl*3	ndhH,orf188-	*rps*15,ndhH/ndhF,*rpl*3
SSC lenght	12.061	12.017	12.229	11.992	20.117	12.052	12.086	12.077	12.110
N. genes	9	9	10	9	12				
trn	1	1	1	1	1				
SSC GC%	32.4	32.2	32.6	32.3	31.7	32.2	32.2	32.2	32.2
SSC A%	36.5	31.3	31.3	31.2	33.7	36.6	36.6	36.5	36.6
SSC C%	16.6	15.7	15.9	15.8	16.4	16.5	16.4	16.5	16.5
SSC G%	15.8	16.5	16.7	16.5	15.3	15.7	15.8	15.7	15.7
SSC T%	31.1	36.5	36.1	36.5	34.7	31.2	32.2	31.3	31.2
SSC AA%	L/K/I	L/S/F	L/I/F	L/I/F	L/I/F	I/K/L	L/I/K	L/I/K	L/K/I
SSC codon	CAA/GAT/AAA	GAT/CAA/AAA	GAT/CAA/CAT	TGT/CAA/GAT	CAA/CAT/GAA	CAA/GAA/AAA	GAT/CAA/CAT	CAA/GAA/AAA	CAA/AAA/GAA
Junction SSC-	ndhI,ndhA-	ndhF/ndhA,ndhH	ndhF,ndhA/ndhH,	ndhF/ndhA,ndhH	*rps*15,*Ycf*1/*trn*N,*t*	ndhI,ndhA-	ndhI,ndhA-	ndhI,ndhA-	ndhI,ndhA-
IRb lenght	30.502	31.106	30.098	30.474	25.543	30.716	30.847	30.703	30.882
N. genes	13	13	12	13	12				
trn	5	5	5	5	5				
IRb GC%	40.7	40.1	40.7	40.6	43.1	40.8	40.5	40.7	40.6
IRb A%	28.1	28	28.1	28.2	28.4	28	28	28	28
IRb C%	21.5	21.1	21.5	21.5	22.4	21.6	21.4	21.5	21.4
IRb G%	19.2	19	19.3	19.2	20.7	19.2	19.1	19.2	19.1
IRb T%	31.2	31.9	31.2	31.2	28.5	31.2	31.5	31.3	31.5
IRb AA%	S/L/F	S/L/F	S/L/F	S/L/I	L/S/R	S/L/R	L/S/K	S/L/R	S/R/L
IRb codon	GAT/CAT/GAA	AAT/CAT/CAA	AAT/GAT/AAA	GAT/GAA/CAT	GAT/CAT/GAA	GAT/CAT/CAA	(TGG/ATG)/GAT/CA	GAT/GAA/AAA	(TGG/ATG)/CAA/GA
Junction IRb-	*Ycf*2,*trn*I/*trn*H	*Ycf*2,*trn*I/*trn*H	*Ycf*2,*trn*I/*trn*H	*Ycf*2,*trn*I/*trn*H	*rpl*23,*rpl*2/*trn*H,p	*Ycf*2,*trn*I/*trn*H	*Ycf*2,*trn*I/*trn*H	*Ycf*2,*trn*I/*trn*H	*Ycf*2,*trn*I/*trn*H

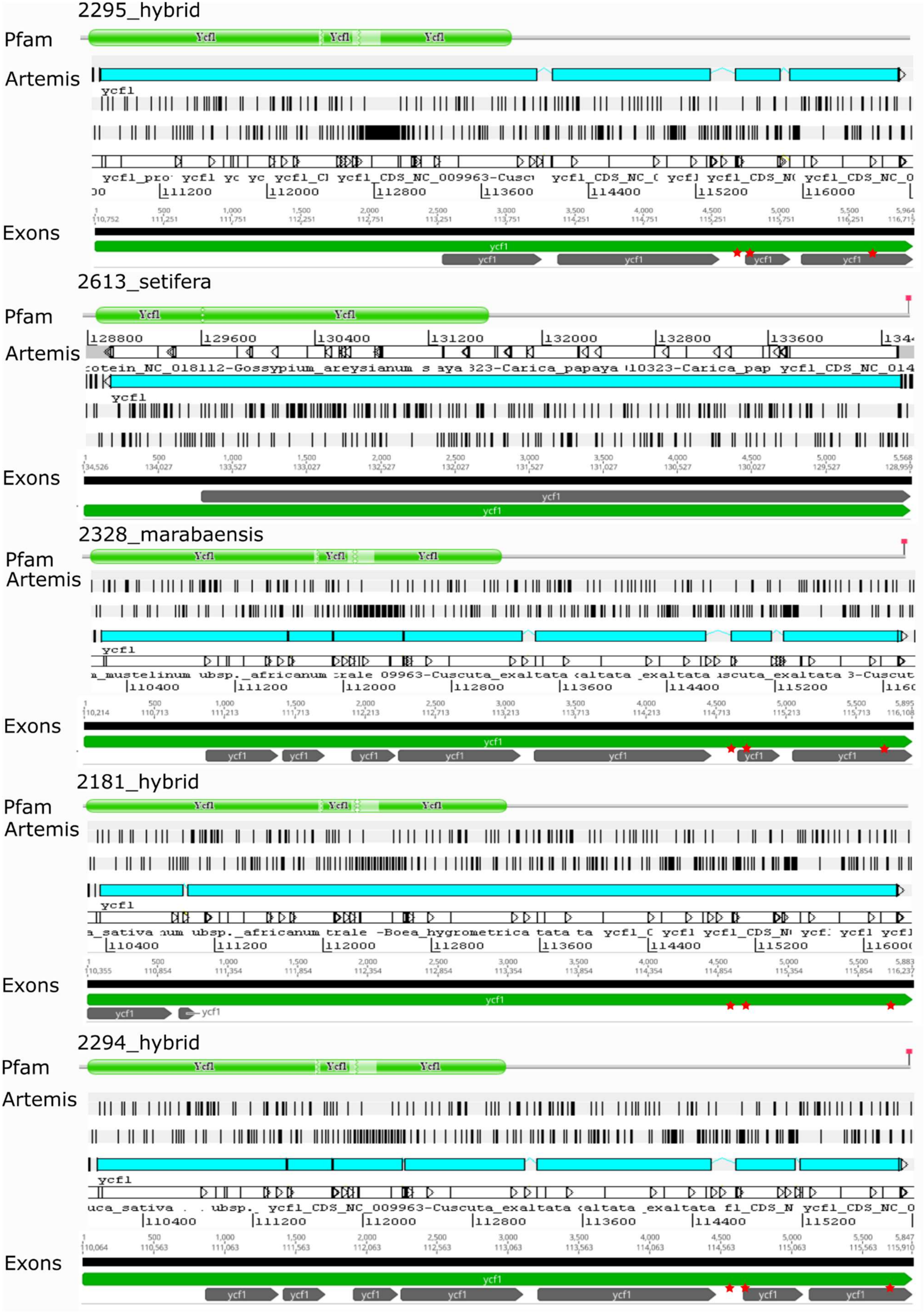
S.Table 2. The first 50 H-disp regions pointed to different genes and intergenic regions of the IRs

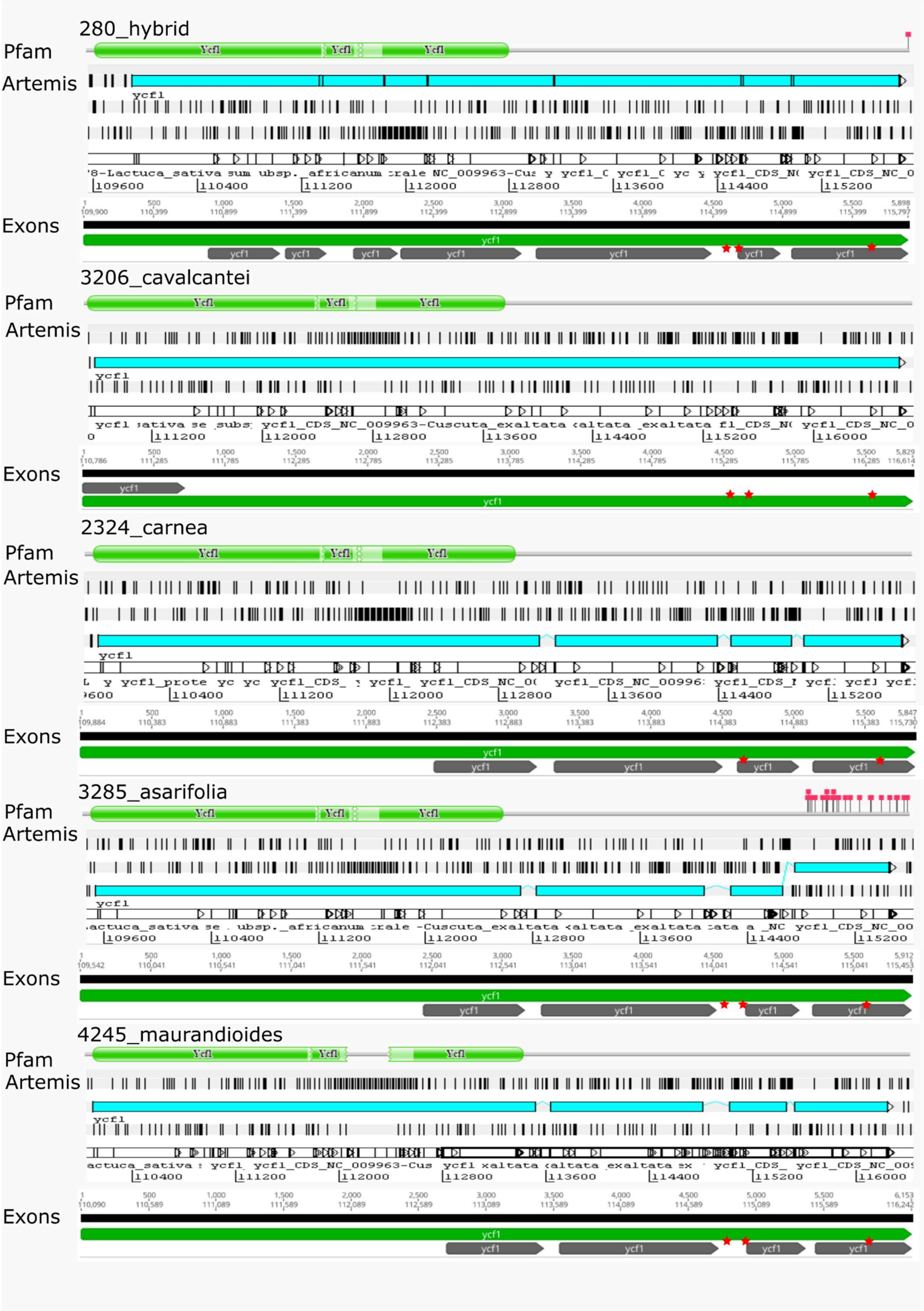
	2328	3206	2181	280	2294	2295	2324	4245	3285	4320	4963	4995	2613
CCAATTGGAGTCTTTTGATTACTGC	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	x
CAGTAATCAAAAGACTCCAATTGGG	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	x
ATTGGGATTTTTTTTTGAAATTGGA	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	x	*rps*7-ndhB	*rps*7-ndhB	x
TTCCAATTTCAAAAAAAAATCCCAA	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	x	*rps*7-ndhB	*rps*7-ndhB	x
CAAAAGACTCCAATTGGGACTAATA	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	x
ATCAAAAGACTCCAATTGGGACTAA	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	x
ATTAGTCCCAATTGGAGTCTTTTGA	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	x
AGTAATCAAAAGACTCCAATTGGGA	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	x
ATTGGAGTCTTTTGATTACTGCTGG	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	x
CAATTTCTCAATTAATTCTTTATAA	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	x
ATCCAATTTCTCAATTAATTCTTTA	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	x
CCAATTTCTCAATTAATTCTTTATA	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	x
TCCCAGCAGTAATCAAAAGACTCCA	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	x
AAAGAATTAATTGAGAAATTGGATA	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	x
AATTGGGATTTTTTTTTGAAATTGG	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	x	*rps*7-ndhB	*rps*7-ndhB	x
CAATTGGGATTTTTTTTTGAAATTG	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	x	*rps*7-ndhB	*rps*7-ndhB	x
AATTTCAAAAAAAAATCCCAATTGT	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	x	*rps*7-ndhB	*rps*7-ndhB	x
AAGAATTAATTGAGAAATTGGATAT	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	x
AATATCCAATTTCTCAATTAATTCT	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	x
ATCCCAGCAGTAATCAAAAGACTCC	*Ycf*1	*rps*15-*Ycf*1	*rps*15-*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	x
ATTTCAAAAAAAAATCCCAATTGTG	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*Ycf*1	*rps*7-ndhB	*rps*7-ndhB	x
ATTTCTCAATTAATTCTTTATAACG	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	x
AATTTCTCAATTAATTCTTTATAAC	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	x
GTCTTTTGATTACTGCTGGGATTGA	*Ycf*1	*rps*15-*Ycf*1	*rps*15-*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	x
ATAAAGAATTAATTGAGAAATTGGA	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	x
ACACAATTGGGATTTTTTTTTGAAA	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	x	*rps*7-ndhB	*rps*7-ndhB	x
CTTTTGATTACTGCTGGGATTGATC	*Ycf*1	*rps*15-*Ycf*1	*rps*15-*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	x
GACACAATTGGGATTTTTTTTTGAA	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	x	*rps*7-ndhB	*rps*7-ndhB	x
ATCAATCCCAGCAGTAATCAAAAGA	*Ycf*1	*rps*15-*Ycf*1	*rps*15-*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	x
CGACACAATTGGGATTTTTTTTTGA	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	*rps*7-ndhB	x	*rps*7-ndhB	*rps*7-ndhB	x
AATTGATTAATACGAATACGTGTAT	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	x
ATTGATTAATACGAATACGTGTATA	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	x
TTATACACGTATTCGTATTAATCAA	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	x
TGATTAATACGAATACGTGTATAAA	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	x
ATTTATACACGTATTCGTATTAATC	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	x
AATTAATTGAGAAATTGGATATTCG	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	x
GAATAATTCATCAAAAATTGATTAA	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	x
AATAATTCATCAAAAATTGATTAAT	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	x
AGTCTTTTGATTACTGCTGGGATTG	*Ycf*1	*rps*15-*Ycf*1	*rps*15-*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	*Ycf*1	x
ATCAAAATTGAAACGAAACAAATAA	ndhH-*rps*15	ndhH-*rps*15	ndhH-*rps*15	ndhH-*rps*15	ndhH-*rps*15	ndhH-*rps*15	ndhH-*rps*15	ndhH-*rps*15	ndhH-*rps*15	ndhH-*rps*15	ndhH-*rps*15	ndhH-*rps*15	x
AGAATAATTCATCAAAAATTGATTA	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	*Ycf*2-*trn*I	x
GAATATCCAATTTCTCAATTAATTC	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	*rps*15	x

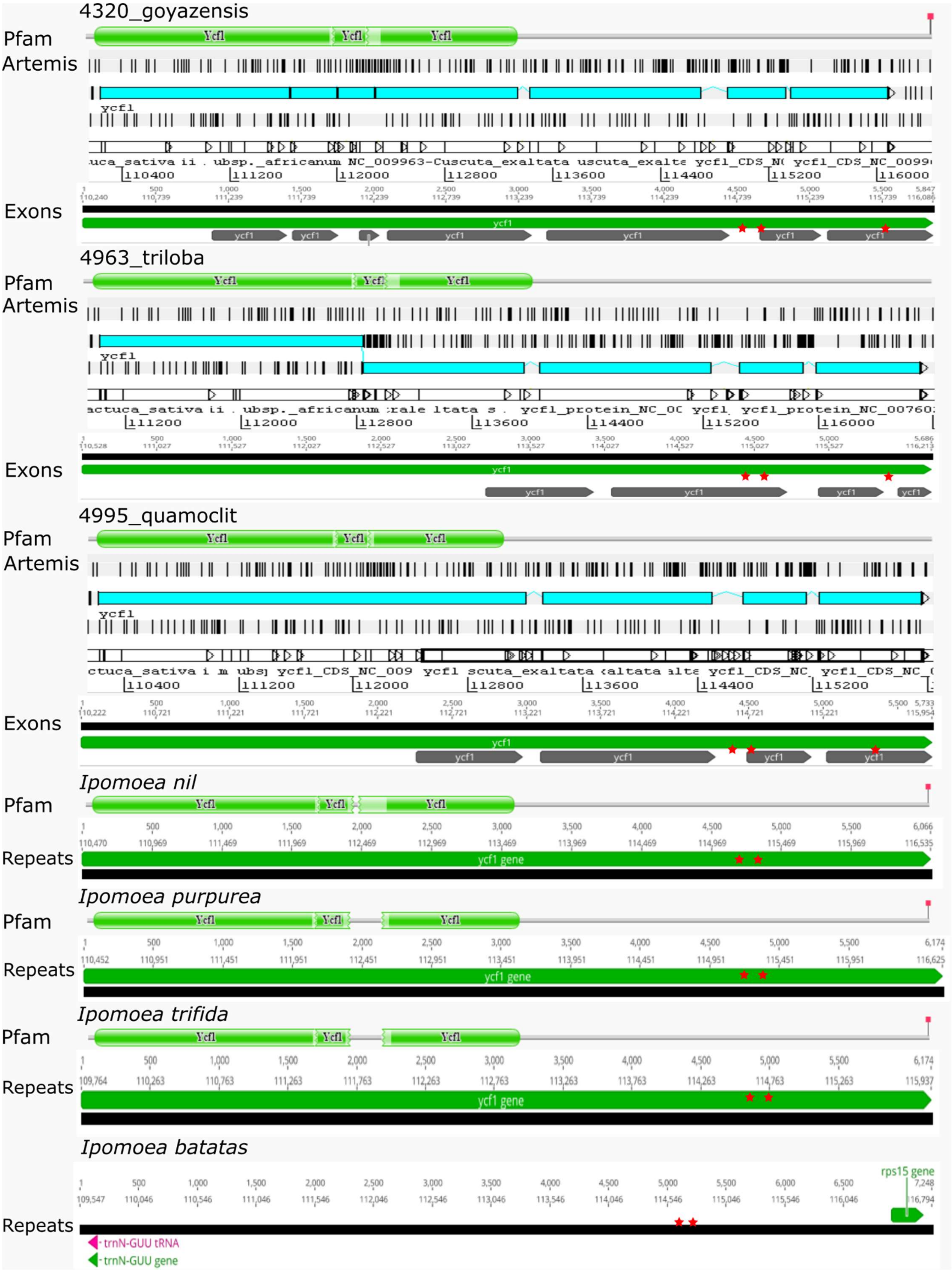
S.Figure 1. Mauve alignment in progressive mode to check for homologous regions and general similarity.



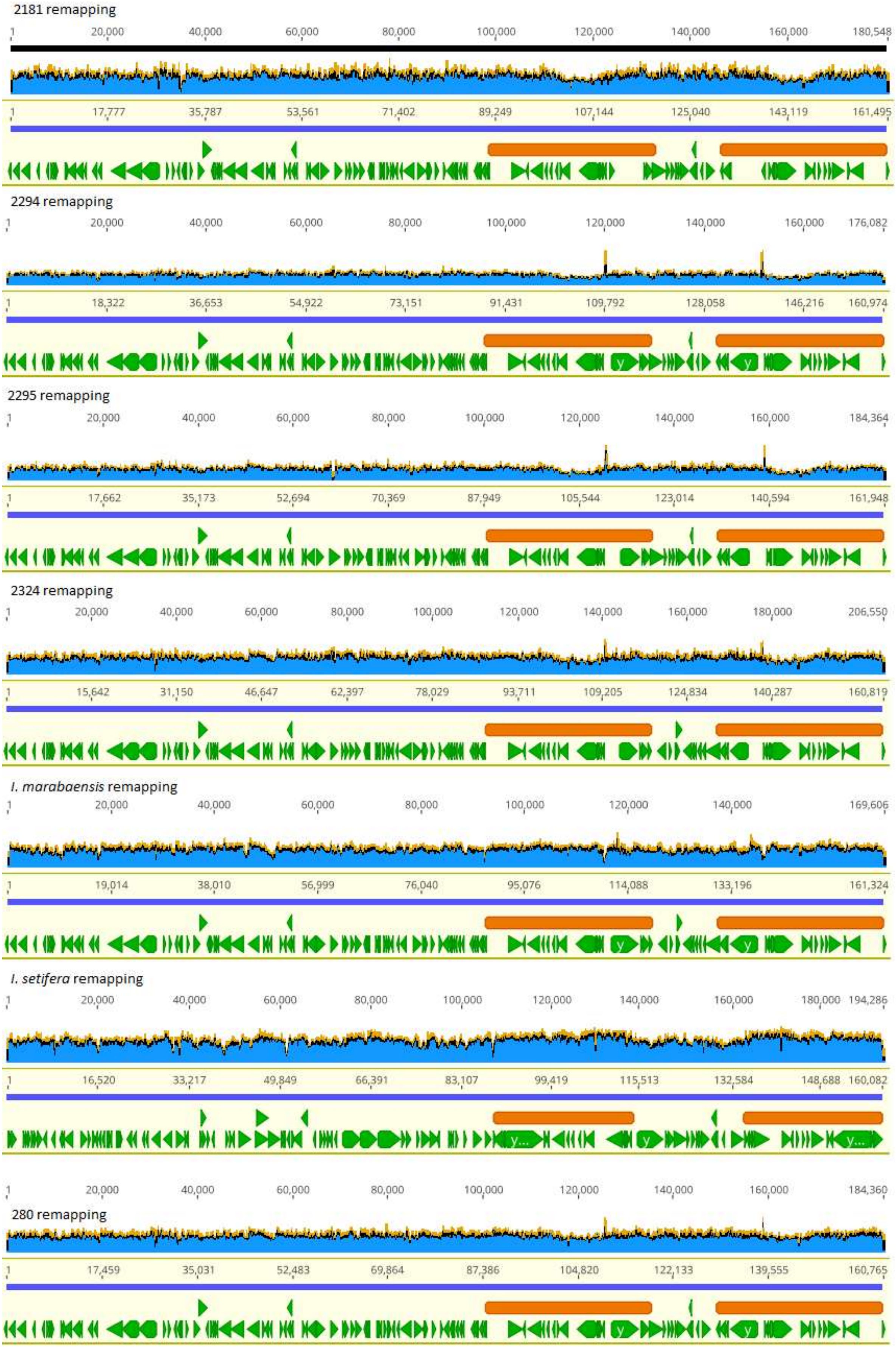
S. Figure 2. The Pfam structure, Artemis curation, and annotation view for *Ycf*1 gene.



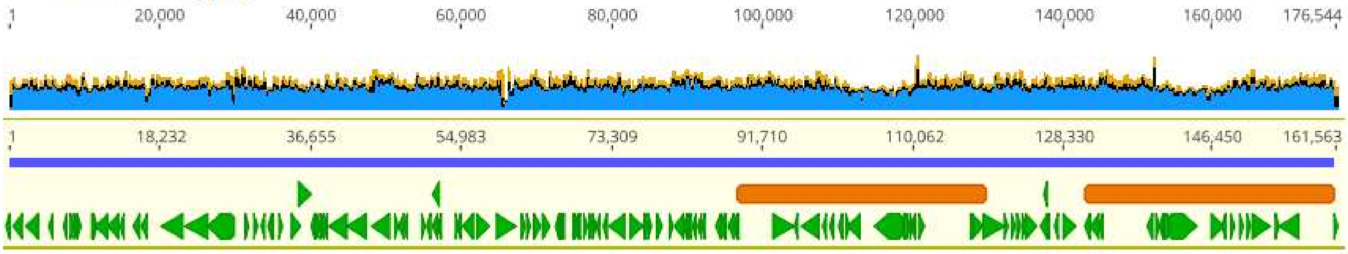

280_hybrid
Pfam
Artemis
Exons
Ycf1
ycf1
3206_cavalcantei
Pfam
Artemis
Exons
Ycf1
ycf1
2324_carnea
Pfam
Artemis
Exons
Ycf1
ycf1
3285_asarifolia
Pfam
Artemis
Exons
Ycf1
ycf1
4245_maurandioides
Pfam
Artemis
Exons
Ycf1
ycf1


4320_goyazensis
Pfam
Ycf1
Ycf1
Ycf1
Artemis
ycf1
Exons
ycf1
4963_triloba
Pfam
Ycf1
Ycf1
Ycf1
Artemis
ycf1
Exons
ycf1
4995_quamoclit
Pfam
Ycf1
Ycf1
Ycf1
Artemis
ycf1
Exons
ycf1
Ipomoea nil
Pfam
Ycf1
Ycf1
Ycf1
Repeats
ycf1 gene
Ipomoea purpurea
Pfam
Ycf1
Ycf1
Ycf1
Repeats
ycf1 gene
Ipomoea trifida
Pfam
Ycf1
Ycf1
Ycf1
Repeats
ycf1 gene
Ipomoea batatas
Repeats
rps15 gene
trnN-GUU tRNA
trnN-GUU gene

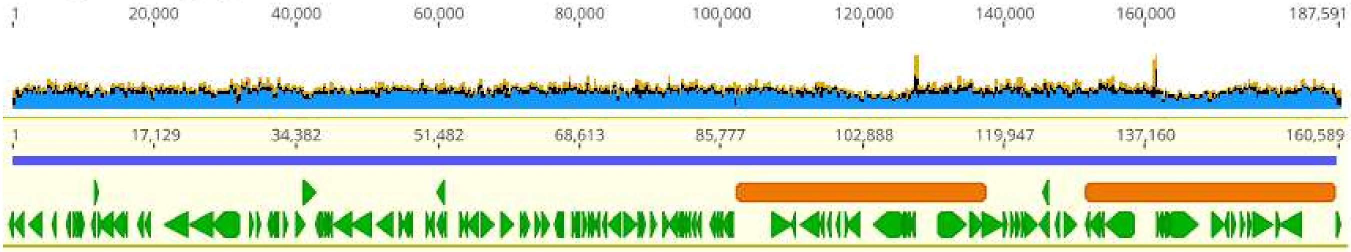
S.Figure 3. remapping of the original dataset against the assembled plastomes



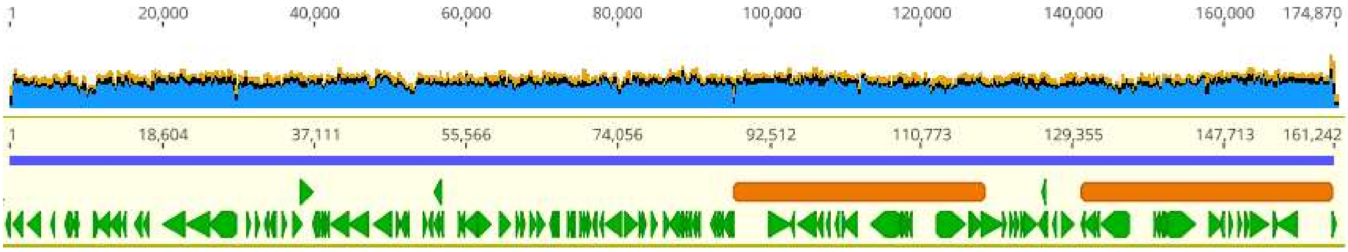
I. cavalcantei remapping



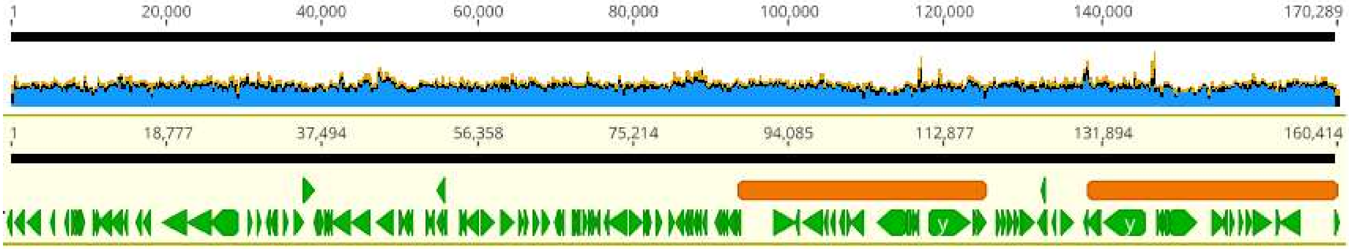
I. asarifolia remapping



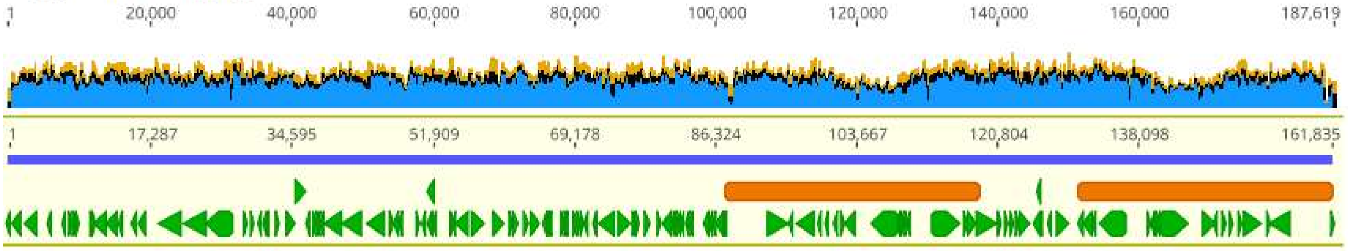
I. maurandioides remapping



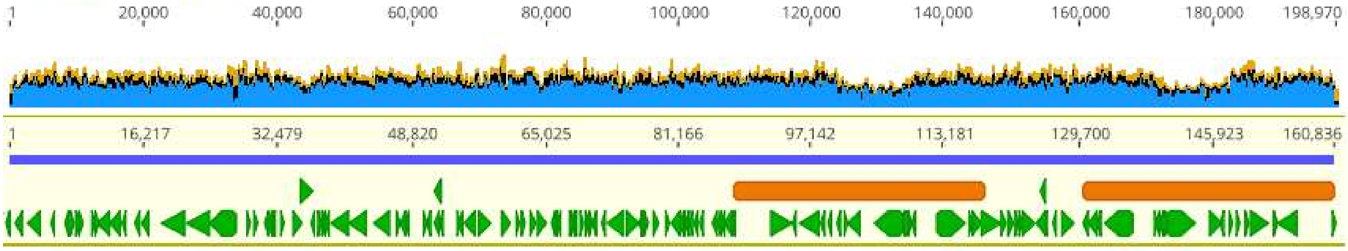
I. triloba remapping



I. goyazensis remapping



I. quamoclit remapping


1
20,000
40,000
60,000
80,000
100,000
120,000
140,000
160,000
180,000
198,970
1
16,217
32,479
48,820
65,025
81,166
97,142
113,181
129,700
145,923
160,836