UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE BIOCIÊNCIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA VEGETAL

PAULO AECYO FRANCISCO DA SILVA

ELUCIDAÇÃO DAS RELAÇÕES FILOGENÉTICAS E DETECÇÃO DE HÍBRIDOS NO GRUPO CAESALPINIA (Caesalpinioideae, Leguminosae)

Recife
2021

PAULO AECYO FRANCISCO DA SILVA

ELUCIDAÇÃO DAS RELAÇÕES FILOGENÉTICAS E DETECÇÃO DE HÍBRIDOS NO GRUPO CAESALPINIA (Caesalpinioideae, Leguminosae)

Dissertação apresentada ao Programa de Pós-Graduação em Biologia Vegetal da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de mestre em Biologia Vegetal.

Área de Concentração: Sistemática e Evolução.

Orientadora: Profa. Dra. Andrea Pedrosa-Harand

Coorientador: Prof. Dr. Luiz Gustavo Rodrigues Souza

Recife

2021

Catalogação na fonte:
Bibliotecária Claudina Queiroz, CRB4/1752

Silva, Paulo Aecyo Francisco da
Elucidação das relações filogenéticas e detecção de híbridos no grupo Caesalpinia (Caesalpinioideae, Leguminosae) / Paulo Aecyo Francisco da Silva - 2021.
155 folhas: il., fig., tab.

Orientadora: Andrea Pedrosa-Harand
Coorientador: Luiz Gustavo Rodrigues Souza
Dissertação (mestrado) – Universidade Federal de Pernambuco. Centro de Biociências. Programa de Pós-Graduação em Biologia Vegetal. Recife, 2021.

Inclui referências e apêndice.

1. Cenostigma 2. Filogenômica 3.Caatinga
I. Pedrosa-Harand, Andrea (Orientadora) II. Souza, Luiz Gustavo Rodrigues (Coorientador) III.Título

583.749 CDD (22.ed.) UFPE/CB-2022-017

PAULO AECYO FRANCISCO DA SILVA

ELUCIDAÇÃO DAS RELAÇÕES FILOGENÉTICAS E DETECÇÃO DE HÍBRIDOS NO GRUPO CAESALPINIA (Caesalpinioideae, Leguminosae)

Dissertação apresentada ao Programa de Pós-Graduação em Biologia Vegetal da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de mestre em Biologia Vegetal.

Área de Concentração: Sistemática e Evolução.

APROVADA EM: 19/07/2021

BANCA EXAMINADORA

Membros titulares:

Prof. Dra. Andrea Pedrosa Harand - UFPE

Prof. Dra. Clarisse Palma da Silva - UNICAMP

Prof. Dr. Cícero Carlos de Souza Almeida - UFAL

Membros suplentes

Prof. Dr. Rafael Batista Louzada - UFPE

Prof. Dr. Edlley Max Pessoa da Silva – UFMT

AGRADECIMENTOS

Eu quero que fique registrado aqui a gratidão que tenho por todas as histórias lutadas antes de mim para que eu chegasse aqui. A cada preto e preta que permitiram que hoje eu, esse preto, pobre, cotista e favelado, conquistasse o título de Mestre em Biologia Vegetal, meu muito obrigado! Esse título também é de cada um.

Além dos antepassados, também gostaria e devo lembrar dos presentes. À Benedita, minha avó, vinda do agreste de Pernambuco, que viu na educação um caminho para uma possível ascensão pessoal e social e lutou para que seus dois filhos fossem educados a qualquer custo. À Lourdes, minha mãe, que com muita garra lutou contra o machismo, contra o racismo e contra uma estrutura social que fazia de tudo para que ela não avançasse. À Janainna, minha mamãe, que assim como as mulheres de sua família apoiou todo o meu desenvolvimento até aqui e também me deu dois dos maiores presentes de minha vida, Gabi e Bia, minhas primas, as quais, muitas vezes presenciaram mais de perto do que qualquer um o desenvolvimento desta dissertação e me acalantaram por tantas noites antes de dormir. À Paula, de muitos títulos: professora, historiadora, empresária, irmã, amiga, chata e insuportável. Mas também, a primeira graduada da família. Minha maior referência preta e acadêmica. Que me instigou e orientou durante todas as fases da minha vida. E que, juntamente ao meu querido cunhado, amigo e parceiro de cachaça, Felipe (Reis) me abrigou durante este período de pandemia em que passamos, permitindo que boa parte desse trabalho fosse realizado. À todas vocês, meu fortíssimo obrigado. Esse título também é de, por e para cada uma!

No campo ideológico, este título também não é uma aquisição só minha. Por isto, também agradeço aos meus orientadores por todo trabalho e perseverança comigo. A Gustavo, meu co-orientador, por todo o incentivo e conselho. Por todo este tempo, muitas vezes, fui animado e instigado por suas palavras para dar continuidade a este e a outros de seus trabalhos em que colaborei. Além disso, não posso esquecer de todo o aprendizado seja em sala de aula ou diretamente com você, tenha certeza que levarei comigo cada palavra. À Andrea, minha orientadora, a qual eu não tenho palavras pra expressar minha gratidão. Foram cinco anos e meio sob sua orientação. Cada reunião, cada relatório, cada plano de trabalho, cada projeto, cada conversa, cada puxão de orelha, cada plano (Planos A, B, C, D). Você não foi só minha orientadora, mas também amiga e companheira durante todo esse processo e formou tudo que eu sei sobre fazer ciência. Você é uma grande inspiração para o que eu quero ser neste meio acadêmico. Vocês dois são de longe os melhores orientadores que eu podia ter

durante este período. Este meu título de mestre, Doutores, também é de vocês sem a menor sombra de dúvida.

Muitas pessoas se envolveram diretamente com meu trabalho no dia a dia durante os 2 anos e 5 meses em que passei no mestrado. Mas de fato duas delas se envolveram de modo muito especial. À Amália, em especial por também ter me abrigado em sua casa, não apenas nos dias de bebedeira em alguma comemoração ou simplesmente a boa e velha aglomeração, mas também durante o período de pandemia. Você permitiu diretamente que os dados de pelo menos metade do meu segundo capítulo fossem gerados. Você também permitiu que eu não surtasse diversas vezes, me incentivou e me acolheu tantas outras. Por cada conversa no laboratório, na mesa de bar ou em sua casa às três da manhã ao som da mais profunda cultura musical de meu país Pernambuco: o bregafunk. À Mariela, a qual, mesmo da Argentina, sempre manteve contato, ouviu minhas indagações, me orientou como a um aluno, amigo e filho! Mama, você segurou várias das minhas barras e eu preciso reconhecer isso aqui. E como se eu já não aperreasse você o suficiente, ainda revisasse minha fundamentação teórica. Esse título só existe e também é de vocês. Muito obrigado!

Além delas, eu devo o agradecimento ao Laboratório de Citogenética e Evolução Vegetal, o melhor laboratório de Citogenética da América Latina! Este laboratório, sob a orientação de Andrea, Gustavo e Marcelo, com todos os seus alunos atuais e passados, me acolheu e me formou desde março de 2016, quando entrei em meu PIBIC, até o presente momento, me ensinando das mais diversas formas a luta que é para se construir a ciência e permanecer cientista no Brasil. À Bruna, Natália e Rayssa (que sempre vai ter seu pé encostado aqui) eu agradeço a cada momento vivido, cada sorriso, cada desabafo, cada momento de amizade dentro e fora do lab. À Amanda por cada limpeza feita na molecular (seja da bagunça da própria molecular ou da bagunça que eu deixei lá), mas também por cada conselho, cada palavra e gestos trocados (sim, estou falando das suas caras e bocas). A Erton que viu de perto em cada reunião em trio, juntamente com Mariela, os meus desesperos e aperreios, e que sempre esteve disposto e presente para me ajudar. À Brena por cada besteira e surtos trocados no instagram – só Deus sabe o quanto os áudios de Malu me acalmavam e alegravam no dia a dia. A Lucas o qual tirou diversas das minhas dúvidas e contribuiu diretamente no meu segundo capítulo. À Yhanndra que foi parceira de diversas viagens “CDU – Piedade” tornando todo o percurso de 2h entre nossas casas e a universidade mais palatável. À Yennifer a qual me fez rir loucamente durante toda minha formação e compartilhou cada momento comigo, mesmo eu devendo o famoso caldinho de dona Lourdes. À Cláudio, Géssica, Gustavinho, Jéssica e Thiago que estiveram comigo durante todo o processo, desde

cachaças até a escrita do presente trabalho, eu agradeço por cada sorriso, cada fofoca e cada troca de aprendizado. Por fim, mas definitivamente não menos importante, à Mariana *Alejandra* Baez a qual, dentre tantas coisas, me serve de inspiração quanto pessoa e profissional. “Pode vir aqui porque esse *título* também serve pra vocês”.

Algumas pessoas não estavam envolvidas diretamente na produção deste trabalho, mas também merecem os créditos. À Ana Karla, Gabriel e Reginaldo que, apesar de não conhecerem um ao outro, ouviram meus desesperos de mais perto que uma tela de telefone pode permitir (acreditem, a globalização permite uma proximidade gigante). À Fabíola, Fernanda, Natália (e Gabriel), Lucas, Fábio, Marina, Thais e Tomás, meus amigos de graduação que acompanharam todo o processo da seleção para o mestrado e comemoraram como uma vitória deles mesmos (porque foi!). Também agradeço a todos os meus amigos da EREM Augusto Severo e agregados, sendo representados aqui por Acauã, Elice e Nilton (e Ana Karla), só Deus sabe a saudade que sinto dos beijos e abraços compartilhados em nossas aglomerações e o quanto eu preciso daquela praia hoje ao fim deste trabalho. Apesar da distância causada pela pandemia, este título também é fruto de seu apoio e pertence a cada um de vocês.

Por fim, também devo meus agradecimentos à CAPES que, mesmo com os diversos cortes sofridos durante o atual governo, me concedeu e manteve a bolsa durante 2 anos e 3 meses de meu mestrado, tornando possível a obtenção do título de mestre deste preto, pobre, cotista e favelado. À Universidade Federal de Pernambuco, ao Departamento de Botânica e à Pós-graduação em Biologia Vegetal por ter me dado a oportunidade de ingresso e estrutura de trabalho durante todo esse tempo. Também gostaria de agradecer a diversos professores que se envolveram na produção e revisão deste trabalho: os professores Cícero Almeida, Clarisse Palma da Silva, Edlley Pessoa e Rafael Louzada que compõem a minha banca de defesa de mestrado. Desde já agradeço por sua disposição! Mas também gostaria de agradecer aos professores Benoit Loeuille e Ana Christina Brasileiro Vidal que, em algum momento, avaliaram parte deste trabalho ao longo da trajetória de sua produção, no nome dos quais agradeço a todos os professores e profissionais que permitiram que este trabalho tenha sido concluído.

[...] ¡Negra!
¿Y qué? ¡Negra!
Sí ¡Negra!
Soy ¡Negra!
Negra ¡Negra!
Negra soy
[...]
Al fin
Al fin comprendí AL FIN
Ya no retrocedo AL FIN
Y avanzo segura AL FIN
Avanzo y espero AL FIN
Y bendigo al cielo porque quiso Dios
que negro azabache fuese mi color
Y ya comprendí AL FIN
Ya tengo la llave
NEGRO (14x)
¡Negra soy!

Trecho do poema entitulado *Me Gritaron Negra*, de autoria de Victoria Santa Cruz (GAMARRA, 1978)

RESUMO

O grupo Caesalpinia (Leguminosae) apresenta distribuição pantropical, compreendendo 26 gêneros monofiléticos e 225 espécies. O grupo tem sido utilizado como modelo em estudos correlacionando variáveis ambientais e citogenéticas/genômicas. Por outro lado, as relações inter e intragenéricas não estão totalmente elucidadas. Dessa forma, este trabalho visou avançar na compreensão dos processos evolutivos atuantes no grupo Caesalpinia. Para isso, 13 plastomas foram sequenciados, montados, anotados e comparados com outros presentes no GenBank, representando 54% dos gêneros do grupo Caesalpinia. Foi observada uma alta conservação desses plastomas, apesar da idade antiga do grupo (~56 Ma). Uma abordagem filogenômica recuperou dois clados bem suportados, com uma boa congruência com filogenias prévias. Entretanto, foi observado algumas incongruências, como a relação mal-resolvida entre *Cenostigma microphyllum* e *C. pyramidale*. A existência de indivíduos com morfologia intermediária entre essas espécies sugeria a existência de hibridação natural, hipótese que foi testada aqui. Nós analisamos populações naturais destas espécies a partir de marcadores microssatélites e morfometria geométrica. Foi observado que, embora essas espécies mantenham sua integridade genética, há fluxo gênico interespecífico com a formação de híbridos que podem ser identificados pela morfologia foliar intermediaria. Esse fluxo gênico interespecífico é relacionado à ocorrência dessas espécies em simpatria e sobreposição das mesmas em tipos de solo e formações geológicas similares. Dessa forma, a hibridação natural parece ter um papel importante na evolução do gênero *Cenostigma*, contribuindo para relações filogenéticas mal-resolvidas.

Palavras-chave: caatinga; *Cenostigma*; filogenômica; hibridação; plastoma.

ABSTRACT

The Caesalpinia group (Leguminosae) shows a pantropical distribution and comprises 26 monophyletic genera and 225 species. The group has been used as model in studies correlating environmental and cytogenetics/genomic traits. However, the inter- and intrageneric phylogenetic relationships are not fully solved. The aim of this study was to investigate the evolutionary process acting on the Caesalpinia group. Thus, 13 plastomes were sequenced, assembled, annotated and compared with others deposited in the GenBank, representing 54% of the genera from the Caesalpinia group. It was observed a high macrostructural conservation of the plastomes in the group, despite its old age (~56 Ma). A phylogenomic approach recovered two well supported clades, with a high congruence with previous phylogenetic trees. However, some incongruences were observed, such as a non-solved relationship between *Cenostigma microphyllum* and *C. pyramidale*. The existence of individuals with an intermediary morphology among these species suggested that natural hybridization might be occurring. This hypothesis was tested here. We analyzed natural population of both species using microsatellites markers and geometric morphometry. We observed that, although both species maintained its genetic integrity, there is interspecific gene flow among them, with the formation of hybrid individuals, which can be identified by the intermediate foliar morphology. The interspecific gene flow is correlated with the sympatric occurrence and overlapping distribution of these species on the same soil type and geologic formation. Therefore, natural hybridization seems to have an important role in the evolution of *Cenostigma* genus, contributing to the unresolved phylogenetic relationship.

Keywords: caatinga; *Cenostigma*; hybridization; phylogenomics; plastome.

LISTA DE FIGURAS

ARTIGO 1 – DOES NATURAL HYBRIDIZATION CONTRIBUTE TO SYSTEMATIC COMPLEXITY IN *Cenostigma* (LEGUMINOSAE)? INTEGRATIVE EVIDENCE ON THE INTRICACY OF *C. microphyllum* AND *C. pyramidale* SPECIES BOUNDARIES

LISTA DE TABELAS

ARTIGO 1 – DOES NATURAL HYBRIDIZATION CONTRIBUTE TO SYSTEMATIC COMPLEXITY IN *Cenostigma* (LEGUMINOSAE)? INTEGRATIVE EVIDENCE ON THE INTRICACY OF *C. microphyllum* AND *C. pyramidale* SPECIES BOUNDARIES

SUMÁRIO

1 INTRODUÇÃO

O esclarecimento das relações filogenéticas de um clado é fundamental para diversos estudos evolutivos (IBIAPINO et al., 2019; GAGNON et al., 2019; SOUZA et al., 2019; VAN-LUME et al., 2019; KOENEN et al., 2020a; MATA-SUCRE et al., 2020a). Entretanto, a complexa história evolutiva de alguns grupos dificulta a sua inferência filogenética (PIRIE et al., 2016; LPWG, 2017). Com o avanço das técnicas de sequenciamento, tem sido possível realizar a montagem de genomas organelares (mitogenoma e plastoma), assim como o sequenciamento de múltiplos genes, o que permitiu um avanço nas inferências filogenéticas, a resolução de alguns clados e a detecção de incongruências filogenéticas que indicam histórias evolutivas complexas em alguns grupos (DODSWORTH et al., 2019; GONÇALVES et al., 2019; KOENEN et al., 2020b; ZHANG et al., 2020).

Dentre os principais problemas em uma inferência filogenética está a evolução reticulada, que pode ter como causa a separação incompleta de linhagens ou a transferência horizontal de genes como eventos de hibridação interespecífica. A hibridação pode provocar uma baixa resolução na filogenia, com clados apresentando baixos suportes e incongruência entre as árvores nucleares e plastidiais, mesmo em abordagens filogenômicas (STEFANOVIĆ; COSTEA, 2008; STULL et al., 2020).

Por muito tempo a hibridação foi vista como sendo consequência das ações antrópicas e sendo um perigo para a conservação das espécies (ABBOT, 2017). Entretanto, a hibridação natural tem um papel importante na evolução de vários organismos, especialmente em plantas nas quais se estima que cerca de 25% delas possam hibridizar (MALLET, 2005). Ela pode contribuir para a manutenção da diversidade genética dos organismos, propiciar a introgressão de características adaptativas entre as espécies e contribuir para o surgimento de novas linhagens e para a diversificação de clados (SUAREZ-GONZALEZ et al., 2016; MOTA et al., 2019; MA et al., 2020; MEIER et al., 2020).

O grupo Caesalpinia pertence à subfamília Caesalpinioideae (Leguminosae) e apresenta uma grande variação morfológica, podendo apresentar hábito arbustivo, arbóreo e liana (GAGNON et al., 2016, 2019). O grupo tem uma distribuição pantropical, principalmente no Bioma Suculento, podendo estar presente em outros biomas, o que normalmente está associado com mudanças no hábito das espécies (GAGNON et al., 2019). Por causa de suas características morfológicas, ecológicas e genômicas, o grupo tem sido utilizado como modelo em diversos estudos (VAN-LUME et al., 2017, 2019; GAGNON et

al., 2019; SOUZA et al., 2019; MATA-SUCRE et al., 2020a, 2020b). Entretanto, as suas relações inter- e intragenéricas não estão completamente esclarecidas.

Diversos estudos filogenéticos tentaram elucidar as relações do grupo (LEWIS; SCHIRE, 1995; SIMPSON; MIAO, 1997; SIMPSON et al., 2003; NORES et al., 2012; GAGNON et al., 2013; 2016). O estudo mais amplo utilizou cerca de 84% de suas 225 espécies e observou que ele é monofilético e se divide em 26 clados, cada um agora circunscrito como um gênero (GAGNON et al., 2016, 2019). Entretanto, as relações entre alguns gêneros teve baixo suporte, particularmente entre os gêneros *Paubrasilia* Gagnon, H.C.Lima & Lewis, *Caesalpinia* L., *Guilandina* L., *Moullava* Adans, *Mezoneuron* Desf. e *Pterolobium* R.Br. ex Wight & Arn, assim como as relações infragenéricas, como no gênero *Cenostigma* Tul.

O gênero *Cenostigma* apresenta cerca de 14 espécies e 20 taxa. Ele apresenta uma distribuição neotropical especialmente nas Florestas Tropicais Sazonalmente Secas (GAGNON et al., 2016, 2019). No Brasil, o gênero é representado por nove espécies, das quais quatro são endêmicas da Caatinga (FLORA DO BRASIL, 2020). As relações filogenéticas do gênero foram recuperadas pela primeira vez no estudo de Gagnon et al. (2016), utilizando cerca de 74% das espécies, no qual os autores recircunscreveram as espécies pertencentes ao clado B do antigo gênero *Poincianela*, tornando *Cenostigma* monofilético. Entretanto, foram observadas politomias, clados com baixo valores de suportes e espécies não-monofiléticas, como *C. bracteosum* (Tul.) Gagnon & G.P.Lewis, *C. eriostachys* (Benth.) Gagnon & G.P.Lewis, *C. gaumeri* (Greenm.) Gagnon & G.P.Lewis, *C. microphyllum* (Mart. ex G.Don) Gagnon & G.P.Lewis, *C. pluviosum* (DC.) Gagnon & G.P.Lewis e *C. pyramidale* (Tul.) Gagnon & G.P.Lewis. Um estudo morfológico anterior realizado por Lewis (1995) tinha observado indivíduos de morfologia intermediária entre algumas espécies atualmente pertencentes ao gênero, levantando a hipótese de hibridação. Este mesmo padrão foi observado em campo entre as espécies *C. microphyllum* e *C. pyramidale*. Entretanto, esta hipótese nunca foi testada.

Sendo assim, esta dissertação se divide em dois capítulos e tem como objetivo elucidar as relações filogenéticas intergenéricas do grupo Caesalpinia e detectar se a hibridação interespecífica pode estar contribuindo para a complexidade filogenética do grupo. No primeiro capítulo, foram sequenciados por *genome skimming* (sequenciamento de baixa cobertura) um total de 13 espécies de diferentes gêneros do grupo Caesalpinia, seus plastomas foram montados e comparados com outros 13 plastomas presentes no GenBank. As seguintes questões foram levantadas: 1) Quão conservados são os plastomas do grupo Caesalpinia? 2)

Estes plastomas são uteis para inferir as relações inter- e infragenéricas do grupo? 3) Quais relações filogenéticas são inferidas a partir da análise desses plastomas? O segundo capítulo foi focado no gênero *Cenostigma*, visando compreender se a hibridação interespecífica contribui para sua complexidade filogenética. Para isso, foi realizada uma abordagem populacional utilizando marcadores microssatélites e morfometria geométrica para a identificação de fluxo gênico interespecífico entre as espécies *C. microphyllum* e *C. pyramidale*, e comparado com algumas variáveis ambientais, visando responder as seguintes perguntas: 1) Quão definido é o limite entre as espécies *C. microphyllum* e *C. pyramidale*? 2) A hibridação contribui para a complexidade filogenética do gênero? Foi hipotetizado que os plastomas do grupo Caesalpinia apresentam uma estrutura bastante conservada entre as espécies do gênero, mas que a variação de sequência seja útil para análises populacionais e filogenéticas, conseguindo resolver parte das relações complexas do grupo. Além disso, acredita-se que as espécies *C. microphyllum* e *C. pyramidale* hibridizam naturalmente, contribuindo para a complexidade taxonômica do gênero *Cenostigma*.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Genoma de plantas

O genoma de plantas está distribuído em três compartimentos celulares: a mitocôndria, o cloroplasto e o núcleo. Cada genoma apresenta características distintas e é responsável pela tradução de um conjunto de proteínas. O avanço das técnicas de sequenciamento de segunda e terceira geração tem permitido compreender melhor as características de cada conjunto genômico, levando ao avanço em estudos biotecnológicos e servindo também para elucidar aspectos evolutivos de diversos grupos (HESLOP-HARRISON; SCHWARZACHER, 2011; RUHLMAN; JANSEN, 2014; GUALBERTO; NEWTON, 2017).

2.2 O genoma mitocondrial (mitogenoma)

O genoma mitocondrial (ou mitogenoma) está presente em animais e plantas, mas apresenta diferenças entre esses organismos. Enquanto nos mamíferos o mitogenoma varia de 15 a 17 Kpb em tamanho, por exemplo, nas plantas terrestres essa variação pode ser de 200 Kpb a 11 Mpb (GUALBERTO; NEWTON, 2017). Apesar dessa variação em tamanho, não há uma diferença significativa na quantidade de genes presentes nas plantas e nos animais, sendo em sua maioria responsáveis pela produção energética da célula (GUALBERTO; NEWTON, 2017). Sendo assim, o aumento do genoma acontece por meio de perdas e ganhos de introns, assim como transferências horizontais de genes entre os outros compartimentos (MARTINS et al., 2019; CHOI; JANSEN; RUHLMAN, 2020).

O mitogenoma de plantas apresenta uma taxa de mutação bem inferior à de animais e até mesmo em relação aos genomas dos outros compartimentos (2 a 4× menor que o genoma plastidial e 10 a 20× menor que o nuclear). Esse conservacionismo de sequência pode ser explicado pela presença de mecanismos de reparo eficientes, que envolve recombinações homólogas (GUALBERTO; NEWTON, 2017). Esse mecanismo utiliza as sequências repetitivas, que são abundantes no genoma mitocondrial, para o reparo. Por este motivo também, o mitogenoma das plantas, diferentemente de animais, apresenta uma estrutura que é bastante rearranjada mesmo entre espécies próximas (MARTINS et al., 2019) ou até mesmo dentro de uma espécie (ALLEN et al., 2007).

2.3 O genoma plastidial (plastoma)

O cloroplasto é uma organela que tem uma origem a partir do processo de endossimbiose secundária e que participa do metabolismo da fotossíntese, ácidos graxos e outras atividades energéticas (JANSEN; RUHLMAN, 2012). Esta organela apresenta um genoma próprio (plastoma) de herança haploide e, no geral, uniparental nas angiospermas (materna na maioria dos casos). Porém, estima-se que cerca de 32% das plantas com flores possam apresentar uma herança biparental (GONÇALVES et al., 2020). O plastoma apresenta uma macroestrutura que é dividida em quatro regiões: uma região longa de cópia única (*Long single copy* - LSC), uma região curta de cópia única (*Small Single Copy* – SSC) e duas regiões idênticas, porém invertidas (*Inverted Regions* – IR), que separa a região longa da curta (JANSEN; RUHLMAN, 2012). Normalmente, o plastoma é descrito como uma molécula circular, mas estudos recentes apontam que ele pode estar presente no formato linear ou até mesmo ramificado (GONÇALVES et al., 2020).

O avanço das técnicas de sequenciamento e das análises bioinformáticas tem facilitado os estudos comparativos utilizando organismos não modelos. Atualmente, sabe-se que o plastoma apresenta uma variação que vai desde 11 Kpb em *Pilostyles aethiopica* Welw. (Apodanthaceae; BELLOT; RENNER, 2015), uma planta endoparasita, até 217 Kpb em *Pelargonium* × *hortorum* L.H.Bailey (Geraniaceae; CHUMLEY et al., 2006), uma planta ornamental. Esta variação de tamanho pode ser explicada por três diferentes mecanismos: variações pontuais na sequência do plastoma, perda/duplicação gênica ou a expansão/retração do IR.

Perdas gênicas contribuem para a diminuição no tamanho do plastoma. Elas têm sido estudadas em diferentes grupos, e, em alguns casos, estas perdas estão relacionadas com mudanças de hábitos, como no caso do gênero *Cuscuta* L. (Convolvulaceae). Este gênero engloba plantas que são hemi- ou holoparasitas, portanto, é um grupo de plantas que não realiza fotossíntese, ou a faz em níveis muito baixos (VOGEL et al., 2018). Do ponto de vista morfológico, este grupo apresenta a perda ou redução de diversos órgãos (VOGEL et al., 2018). Do ponto de vista molecular, há a perda de diversos genes responsáveis pelos processos de fotossíntese em altas taxas luminosas (BRAUKMAN; KUZMINA; STEFANOVIĆ, 2014; VOGEL et al., 2018). Apesar de haver perdas gênicas em determinado compartimento celular, estas não necessariamente estão vinculadas com a perda de sua função, uma vez que pode haver transferências horizontais entre os diferentes genomas (CAUZ-SANTOS et al., 2020) ou ainda a neofuncionalização de outros genes.

As duplicações gênicas no plastoma podem ocorrer de duas formas: duplicações de genes isolados ou duplicações de genes presentes no IR. Apesar de raras, há casos de duplicações gênicas não relacionadas à expansão do IR (WANG et al., 2017). Em *Pelargonium* × *hortorum* cerca de nove genes do LSC foram parcial ou totalmente duplicados no genoma, chegando a apresentar até seis cópias parciais distribuídos pelo plastoma (CHUMLEY et al., 2006). Entretanto, essas duplicações são frequentemente relacionadas com a expansão do IR.

A região invertida contribui também para o aumento do genoma plastidial, podendo variar bastante em tamanho e composição entre os diferentes grupos. Por exemplo, na tribo Mimosoid (Caesalpinioideae, Leguminosae Juss.), Dugas et al. (2015) observaram um aumento de 13 Kpb no IR dos gêneros *Acacia* Mill. e *Inga* Mill. Esse evento aconteceu junto à incorporação de nove genes, normalmente presentes na SSC. O mesmo é observado em *P.* × *hortorum* que apresenta uma expansão que levou seu IR a apresentar 75 Kpb e também contribuiu para a duplicação de diversos genes na molécula. Da mesma forma que há grandes expansões no IR, já foram relatados alguns casos de retrações no IR, como por exemplo em *Passiflora* L., reduzindo drasticamente o tamanho de alguns genes (RABAH et al., 2018).

Apesar do plastoma ser descrito como uma molécula quadripartite, o IR foi perdido algumas vezes em diferentes grupos de angiospermas: Leguminosae, Geraniaceae Juss. (*Erodium* L'Hér. ex Ait. e *Monsonia* L.), Orobanchaceae Vent. (*Conopholis* Wallr. e *Striga* Lour.), Passifloraceae Juss. ex Roussel (*Passiflora capsularis* L. and *Passiflora costaricensis* Killip), o clado Putranjicoide (que inclui as famílias Lophopyxidaceae (Engl.) H.Pfeiff. e Putranjivaceae Endl.) e em Cactaceae Juss. (*Carnegiea gigantea* Britton & Rose) (JANSEN; RUHLMAN, 2012; SANDERSON et al., 2015; CAUZ-SANTOS et al., 2020; JIN et al., 2020). Na família Leguminosae, a perda do IR é a sinapomorfia do clado IRLC (*Inverted Repeat Lacking Clade* – Papilionoideae, Leguminosae), que compreende quatro gêneros *Cicer* L., *Medicago* L., *Pisum* L. e *Trifolium* L. (RUHLMAN; JANSEN, 2014). Apesar desses grupos serem conhecidos pela perda do IR, nos gêneros *Erodium* e *Medicago* foi observado o reaparecimento do IR em algumas espécies, sendo observado inclusive variações entre populações naturais de *Medicago minima* (L.) L. (BLAZIER et al., 2016; CHOI; JANSEN; RUHLMAN, 2020).

O IR não tem uma origem nem uma função completamente elucidada. Porém, acredita-se que a presença dessa estrutura tenha alguma função adaptativa (BLAZIER et al., 2016). Palmer e Thompson (1982) levantaram a hipótese de que o IR teria um papel importante na estabilidade estrutural do plastoma. Entretanto, estudos recentes com o gênero

Erodium e *Passiflora* não observaram esta relação e novas hipóteses correlacionam a presença de regiões repetidas e aumento de blocos sintênicos (BLAZIER et al., 2016; CAUZ-SANTOS et al., 2020), porém esta relação não é vista em Cactaceae (ALMEIDA et al., submetido).

2.4 O genoma nuclear

Organizado na forma de cromossomos lineares e com uma herança biparental, o genoma nuclear das angiospermas apresenta uma variação no tamanho de 61 Mpb – 150 000 Mpb (FLEISCHMANN et al., 2014; PELLICER; FAY; LEITCH, 2010), havendo uma relação entre o tamanho do genoma e latitude (RICE et al., 2019; SOUZA et al., 2019). Ele é formado tanto por sequências de cópia única – sequências codificadoras de proteínas, íntrons, promotores, sequências reguladoras, etc. -, como também por várias classes de sequências repetidas que compõem até 75% do genoma de algumas espécies de plantas (HESLOP-HARRISON; SCHWARZACHER, 2011; NEUMANN et al., 2019).

Por muito tempo as sequências repetidas foram vistas como “DNA lixo” por não codificar genes. Atualmente, sabe-se que esse grupo de sequências tem papel importante no genoma, podendo estar relacionado com processos tumorais em animais, envolvido na composição do telômero e do centrômero e apresentar papel regulatório (BOURQUE et al., 2018; MARQUES et al., 2015; BRAJKOVIĆ et al., 2018), contribuindo em eventos de diversificação de clados (DOGAN et al, 2021). As sequências de DNA repetitivo são classificadas em dois grupos de acordo com a sua distribuição no genoma: sequências dispersas e as que são repetidas em tandem.

As sequências dispersas podem ser classificadas como transposons e retrotransposons, a depender do mecanismo que usam para se mover no genoma (BOURQUE et al., 2018). Este tipo de sequência é geralmente a mais abundante nos genomas de plantas. Os retroelementos, em especial os do tipo LTR (*Long Terminal Repeats*), apresentam um mecanismo de “copiar e colar” por meio do qual multiplicam suas cópias no genoma a partir da ação de uma transcriptase reversa, contribuindo, assim, para o aumento do genoma das plantas e duplicação de genes (BOURQUE et al., 2018).

O outro grupo de sequências repetidas, as sequências repetidas em tandem, são caracterizadas pela presença de dezenas a milhares de repetições adjacentes na mesma orientação. Este grupo pode ser classificado em satélites, minissatélites e microssatélites. Elas apresentam um papel estrutural, na formação do DNA ribossomal e no processo de recombinação, e pode ainda apresentar um papel regulatório (BRAJKOVIĆ et al., 2018). Os

microssatélites, ou SSR (*Simple Sequence Repeats*), são menores que os satélites e minissatélites, apresentando repetição de 1 - 6 pb (WEISING; GARDNER, 1999). Eles podem ser encontrados não apenas no genoma nuclear, mas também no plastidial e mitocondrial.

Os SSR não apresentam uma distribuição uniforme pelo genoma, sendo encontrados principalmente (mas não exclusivamente) em regiões intergênicas (WEISING; GARDNER, 1999). Eles apresentam uma taxa evolutiva mais rápida, causada por erros da DNA polimerase durante o processo de replicação (ELLEGREN, 2000). Seu alto grau de polimorfismo, a codominância e os padrões de herança de cada genoma (nuclear e plastidial), permite que os microssatélites sejam bastante utilizados como marcadores moleculares em diversos campos de estudos e organismos (WHEELER et al., 2014; ELLEGREN, 2000).

2.5 Inferências Filogenéticas e Filogenômicas e suas dificuldades

Abordagens filogenéticas são amplamente utilizadas para inferir as relações filogenéticas em vários níveis taxonômicos (APG IV, 2016; LPWG, 2017; GAGNON et al., 2016). Para tanto, muitos locos tem sido desenvolvidos, especialmente locos universais que permitam a sua utilização entre diferentes grupos (WHITE et al., 1990; SHAW et al., 2005, 2007). Essas abordagens multiloco tem permitido resolver relações filogenéticas e investigar processos evolutivos e biogeográficos envolvidos na evolução de diferentes grupos (STEFANOVIĆ; COSTEA, 2008; GAGNON et al., 2019).

O sucesso da técnica está ligado diretamente à quantidade de informação que um loco pode oferecer e, portanto, identificar o melhor marcador molecular é um ponto crucial e laborioso no processo (KADLEC et al., 2017). Entretanto, nem sempre é possível encontrar marcadores que sejam informativos para o nível hierárquico e tipo de inferência que se pretende realizar (e.g. Givnish et al. 2007). Com o avanço das técnicas de sequenciamento, investigar as relações filogenéticas a partir de abordagens filogenômicas tem permitido esclarecer relações mal-resolvidas e até mesmo inferir relações distintas, dada a quantidade de dados analisados (GONÇALVES et al., 2019).

Apesar da montagem de genomas completos ainda ser desafiador na biologia, abordagens de sequenciamento de baixa cobertura (*genome skimming*) ou até mesmo que enriqueçam e sequencie regiões específicas do genoma (*target capture*) tem se mostrado promissoras nos estudos filogenéticos (DODSWORTH et al., 2019). Com os dados de *genome skimming*, por exemplo, é possível a montagem do genoma plastidial das plantas e do

cístron do DNA ribossomal (rDNA) nuclear, cujo locos são frequentemente utilizados para inferências filogenéticas em plantas (SUN et al., 1994; SHAW et al., 2005, 2007).

Abordagens filogenômicas com os plastomas tem conseguido resolver relações complexas, obtendo altos valores de suportes em diferentes abordagens e níveis taxonômicos (THODE et al., 2020; ZHANG et al., 2020). Porém, tem-se observado que o sinal filogenético pode variar entre os diferentes locos do plastoma e, portanto, diferentes partições devem ser analisadas para uma melhor compreensão das relações evolutivas de um grupo (GONÇALVES et al., 2019; HE et al., 2019; WALKER et al., 2019; THODE et al., 2020).

Além dos problemas metodológicos, um clado pode apresentar uma história evolutiva complexa e, nesses casos, inferir as relações entre os táxons se torna uma tarefa difícil (KOENEN et al., 2020b). Clados que passem por radiação adaptativa, por exemplo, geralmente apresentam árvores de genes mal resolvidas (PIRIE et al., 2016). Isto acontece por causa da retenção de polimorfismo ancestral causada pela não fixação de alelos antes da subsequente divergência das linhagens, ou seja os táxons não tiveram tempo suficiente para divergir geneticamente, apresentando uma separação incompleta das linhagens (*Incomplete Lineage Sorting*, ILS) (MADDISON; KNOWLES, 2006). Essa falta de informação dificulta a inferência filogenética entre as espécies. Apesar disso, abordagens filogenômicas tem auxiliado a resolução de alguns clados (KOENEN et al., 2020; STULL et al., 2020).

Outro processo que pode dificultar a inferência filogenética de um grupo é a hibridação interespecífica. De fato, a presença de híbridos em uma filogenia tem sido um problema antigo e leva a uma incongruência entre as inferências no genoma nuclear e plastidial (STEFANOVIĆ; COSTEA, 2008; STULL et al., 2020). Porém, este tipo de incongruência tem permitido a identificação de linhagens híbridas que podem ser confirmadas posteriormente com outras abordagens, como em *Cuscuta* (IBIAPINO et al., 2019).

2.6 Hibridação

A hibridação pode ser definida como o cruzamento entre diferentes espécies ou raças dentro de uma mesma espécie (SOLTIS; SOLTIS, 2009). A hibridação interespecífica foi vista por muito tempo como problemática para a evolução, pois teria como causa principal as perturbações humanas (ABBOT, 2017) e como resultado a perda de diversidade e extinção de uma espécie (MALLET, 2005). Essa visão se dava principalmente pela ampla adoção do Conceito Biológico de Espécie, proposto por Mayr (1942). Para ele, um grupo de organismos pode ser definido como da mesma espécie se conseguirem reproduzir e gerar descendentes

férteis, excluindo, portanto, a existência do fluxo gênico interespecífico. Neste sentido, o isolamento reprodutivo é a premissa básica para definir uma espécie.

O estudo da especiação é feito a partir do estudo da evolução das barreiras de isolamento reprodutivo entre pares de populações divergentes (WANG et al., 2020). Estas barreiras são divididas entre barreiras pré- e pós-zigótica e atuam juntas para prevenir o fluxo gênico interespecífico e manter a integridade genética de uma espécie (STANKOWSKI; RAVINET, 2021). Entretanto, apesar do conceito de espécie ser um dos mais importantes dentro da biologia, pois serve de base para diversas áreas (DE QUEIROZ, 2005), definir os limites de uma espécie não é uma tarefa fácil. Existem diferentes conceitos de espécies que utilizam aspectos morfológicos, fisiológicos ou genéticos como base para a sua definição (DE QUEIROZ, 2007). Entretanto, de Queiroz (2007) aponta que todos os conceitos envolvem, em sua premissa básica, a presença de diferentes linhagens dentro de uma metapopulação que podem, ao longo do tempo, se diferenciar e dar origem a novas espécies ao adquirir novas características, sejam elas morfológicas, ecológicas, reprodutivas, entre outras.

Com o avanço da genômica tem sido possível uma maior compreensão do processo de especiação em diferentes grupos, permitindo sair de um conceito “genômico” de espécie para a compreensão do impacto de pequenos genes ou loci na evolução das barreiras reprodutivas, como proposto por Wu e Ting (2004) em seu conceito gênico de espécie. Neste conceito, algumas regiões do genoma contribuem diretamente para a diminuição do fluxo gênico interpopulacional levando ao aumento progressivo do isolamento reprodutivo entre as populações, podendo, ou não, culminar na formação de espécies distintas durante o contínuo de especiação (STANKOWSKI; RAVINET, 2021). Neste sentido, a presença do híbrido tem um papel importante no processo evolutivo, permitindo inclusive um aumento do isolamento reprodutivo entre as populações, a partir de processos como o reforço (HOPKINS, 2013), ou por outro lado, permitindo a homogeneização e manutenção de um conjunto gênico único.

O resultado de eventos de hibridação são os mais variados possíveis. De fato, eventos de hibridação podem ser problemáticos para a estabilidade de uma espécie, podendo leva-la à extinção (MUNIZ et al., 2020). Por outro lado, uma vez que esse híbrido seja fértil, ele pode retrocruzar com as espécies parentais e contribuir para a manutenção da diversidade genética da espécie, introgressão de características adaptativas, formação de novas linhagens e contribuir com a diversificação de clados (MOTA et al., 2019; THE HELICONIUS GENOME CONSORTIUM, 2012; MA et al., 2019; MEIER et al, 2017).

A hibridação, estando ou não associada com eventos de poliploidia, tem contribuído em eventos de diversificação de espécies (SOLTIS; SOLTIS, 2009). De fato, com o avanço

das técnicas de sequenciamento, tem sido possível estudar mais a fundo o papel de eventos de hibridação antigas na evolução de clados (TAYLOR; LARSON, 2019). Atualmente, alguns desses eventos estão associados com radiações adaptativas e tem contribuído com a hipótese de que a hibridação contribua com esses eventos (NACIRI; LINDER, 2020). Além disso, a hibridação interespecífica tem um papel ecológico marcante, uma vez que ela pode interferir na comunidade, como no padrão de herbivoria e de polinizadores (CASTILLO-MENDONZA et al., 2019; WANG et al., 2021).

2.7 Identificação de híbridos

A ocorrência de hibridação pode ser sugerida tanto por características fenotípicas quanto por características genéticas, como as incongruências filogenéticas listadas acima. A capacidade de identificação de um híbrido através de características fenotípicas depende do padrão de herança dessa característica, que pode parecer com um dos parentais, ser intermediário aos dois parentais (efeito aditivo) ou ser uma característica nova (SOLTIS; SOLTIS, 2009). Dentre as características fenotípicas que podem indicar hibridação, as morfológicas são mais utilizadas.

Análises morfológicas que demonstrem características intermediárias entre diferentes espécimes pode ser um indicativo de hibridação (LEWIS, 1995). A morfometria tem sido bastante útil para delimitação de espécies e identificação de híbridos, em especial a morfometria geométrica (VISCOSI; LEPAIS; FORTINI, 2009). Esta técnica utiliza uma abordagem semiautomatizada, o que facilita e acelera o processamento dos dados, além de permitir resultados mais robustos (VISCOSI; LEPAIS; FORTINI, 2009; VISCOSI; CARDINI, 2011).

Apesar das técnicas fenotípicas serem bastante úteis, elas apresentam limitações e não permitem uma total caracterização das zonas híbridas. Abordagens de genética populacional permitem acessar o nível e direção do fluxo gênico entre as populações e espécies. Dentre os marcadores genéticos que podem ser utilizados, os marcadores microssatélites (SSR) estão entre os mais empregados, dada às características listadas acima; mas abordagens mais recentes como RADseq tem se mostrado bastante úteis em estudos populacionais (DODSWORTH et al., 2019; KHAN et al., 2020).

A identificação e caracterização das zonas híbridas a partir de marcadores moleculares tem permitido compreender a complexidade do fluxo gênico interespecífico, que pode acontecer entre mais de uma espécie ao mesmo tempo (KHAN et al., 2020), e o papel das

barreiras reprodutivas nesse processo (MOTA et al., 2019). Também é possível inferir a fertilidade do híbrido, o que pode ser importante para a preservação das espécies puras (MUNIZ et al., 2020), e observar o papel da hibridação na diversificação de linhagens (MA et al., 2019; FU et al., 2020).

2.8 O grupo Caesalpinia

O grupo Caesalpinia pertence à família Leguminosae, a segunda maior família de angiospermas de interesse econômico, e a terceira em número de espécies (LPWG, 2017). Nas últimas décadas, diversos estudos, principalmente guiados pelo *The Legume Phylogeny World Group* (LPWG), têm contribuído para resolver as relações existentes entre seus gêneros, classificando atualmente o grupo em seis subfamílias (LPWG, 2017; KOENEN et al., 2020b; ZHANG et al., 2020). Dentre elas está a subfamília Caesalpinioideae, a segunda maior subfamília em número de espécies.

O grupo Caesalpinia é um grupo monofilético pertencente à subfamília Caesalpinioideae, com cerca de 225 espécies, sendo algumas de interesse econômico e ecológico como *Paubrasilia echinata* (Lam.) Gagnon, H.C.Lima & G.P.Lewis, o pau-brasil (GAGNON et al., 2016). Apesar de não apresentar uma sinapomorfia, o grupo é caracterizado por apresentar hábito em geral arbustivo ou arbóreo, podendo apresentar ou não caules aculeados ou espinhosos, tricomas glandulares e estruturas secretoras como mecanismos de defesa, flores zigomórficas com uma sépala cuculada na base da flor e estames organizados ao redor do pistilo (LEWIS, 2005; GAGNON et al., 2013; 2016).

Por causa de suas características morfológicas, ecológicas e genômicas, o grupo Caesalpinia tem sido utilizado como modelo de estudo em alguns trabalhos correlacionando características genéticas com variáveis ambientais. O grupo apresenta uma distribuição pantropical em ambientes semiáridos, especialmente em Florestas Tropicais Sazonalmente Secas (*Seasonally Tropical Dry Forest*, STDF), com algumas espécies habitando outros habitats (GAGNON et al., 2019). Recentemente, foi observado que há uma associação entre a mudança no hábito da espécie e o ambiente em que ela se encontra (ex.: árvores e arbusto no Bioma Suculento e lianas na Savana e Florestas úmidas; GAGNON et al., 2019). Do ponto de vista citogenético, todas as espécies do grupo apresentam $2n = 24$ cromossomos, com algumas espécies apresentando citótipos polipóides (*Libidibia ferrea* (Mart. ex Tul.) L.P.Queiroz, *Cenostigma bracteosum* (Tul.) E.Gagnon & G.P.Lewis). Apesar da pouca variabilidade numérica, há um certo grau de variação na posição, quantidade e composição da

heterocromatina, assim como no tamanho do genoma, havendo correlação dessas características com características ecológicas e não filogenéticas (VAN-LUME et al., 2017, 2019; SOUZA et al., 2019 MATA-SUCRE et al., 2020a, b).

O grupo apresenta um longo histórico de estudos filogenéticos (LEWIS; SCHIRE, 1995; SIMPSON; MIAO, 1997; SIMPSON et al., 2003; NORES et al., 2012; GAGNON et al., 2013; 2016). O mais recente (GAGNON et al., 2016) reclassificou o grupo baseado em dados de cinco marcadores plastidiais (*rps*16, *trn*D-*trn*T, *ycf*6-*psb*M, *mat*K-3'-*trn*K *intron* e *trn*L-*trn*F) e um marcador nuclear (a região ITS1-5.8S-ITS2 do DNA ribossomal) com uma amostragem ampla (84% das espécies). Eles propuseram a divisão do grupo em 26 gêneros, com a possível existência de um 27º gênero, e fizeram um tratamento taxonômico em diversas espécies, criando novos gêneros, como o monoespecífico *Paubrasilia* Gagnon, H.C.Lima & G.P.Lewis, e extinguindo outros como o gênero *Poincianella*, que teve suas espécies distribuídas entre os gêneros *Erythrostemon* Klotzsch e *Cenostigma* Tul. Apesar disso, algumas relações intergenéricas [ex.: *Paubrasilia* + *Caesalpinia* e (*Guilandina* (*Moullava* (*Mezoneuron* + *Pterolobium*)))] e infragenéricas (ex.: *Cenostigma*) não estão bem resolvidas - apresentando baixos suportes e/ou incongruências - havendo, portanto, a necessidade do emprego de novas ferramentas genômicas para a elucidação de suas relações filogenéticas.

2.9 O gênero *Cenostigma*

O gênero *Cenostigma* Tul. apresenta cerca de 20 taxa distribuídos em 14 espécies (GAGNON et al., 2016) e é caracterizado principalmente pelo seu legume lenhoso com uma margem conspícua e grossa, cavidades secretoras internas na lâmina das pínulas e inflorescência e a presença de um indumento estelado nos caules, folhas e ou inflorescência (GAGNON et al., 2016). O gênero apresenta uma distribuição neotropical, desde o México até o Centro-Oeste do Brasil, ocupando principalmente as Florestas Tropicais Sazonalmente Secas, mas também em área de Savana e Florestas Tropicais (GAGNON et al., 2016, 2019). No Brasil, o gênero é representado por nove espécies que estão distribuídas em diferentes domínios fitogeográficos: o Cerrado, a Mata Atlântica, a Floresta Amazônica, o Pantanal e a Caatinga. Neste, o gênero apresenta quatro espécies endêmicas e bem característica dele, conhecidas popularmente como "catingueiras" (FLORA DO BRASIL, 2020).

A descrição do gênero foi feita em 1843, porém Gagnon et al. (2016) o recircunscreveu para incluir as espécies pertencentes ao clado Poincianella B, tornando o gênero monofilético. Apesar disso, as relações infragenéricas não estão resolvidas,

apresentando politomias, clados com baixo suportes e espécies não-monofiléticas. Em estudos morfológicos com o grupo, Lewis (1995) observou, a partir de exsicatas, que alguns espécimes apresentavam características intermediárias entre *C. laxiflorum* x *C. microphyllum* e *C. laxiflorum* x *C. pluviosum*, indicando possíveis eventos de hibridação dentro do gênero. Entretanto, até então nunca houve uma confirmação da existência desses híbridos naturais.

2.10 *Cenostigma microphyllum* × *Cenostigma pyramidale*

Cenostigma microphyllum (Mart ex G.Don) E.Gagnon G.P.Lewis e *C. pyramidale* (Tul.) E.Gagnon G.P.Lewis são espécies distribuídas na Caatinga, estando *C. pyramidale* também presente na Amazônia (FLORA DO BRASIL 2020). As espécies são muito parecidas morfologicamente, mas se diferenciam principalmente pelo número, posição e tamanho das pinas e pínulas, formato e quantidade de flores por inflorescência, além da maior presença de glândulas em *C. microphyllum* (LEWIS, 1998). Citogeneticamente não há diferença entre elas quanto ao número cromossômico ($2n = 2x = 24$), quantidade de bandas $CMA^{+}/DAPI^{-}$ e DNA ribossomal (1 par de sítios de DNAr 5S e 4 pares de sítios de DNAr 45S nos cromossomos acrocêntricos) (VAN-LUME et al., 2017), nem em relação ao tamanho do genoma (1,88 pg em *C. microphyllum* e 1,80 pg em *C. pyramidale*) (SOUZA et al., 2019).

Com base em aproximadamente 50 espécies estudadas da subfamília Caesalpinioideae (ainda na antiga classificação), 60% são autoincompatíveis. *Cenostigma pyramidale* apresentou autoincompatibilidade de ação tardia e polinização principalmente por abelhas do gênero *Xylocopa* e *Centris* (Apidae) (LEITE & MACHADO, 2009). Apesar de não haver trabalho de polinização aprofundados com *C. microphyllum*, já foi observada a presença de pólen da espécie em abelhas do gênero *Apis* (Apidae), da mesma família de *Xylocopa* (NOVAIS et al., 2010).

Ambas as espécies apresentam importância econômica e ecológica. Estudos recentes utilizando *C. microphyllum* observaram a sua capacidade de estabelecimento em áreas degradadas da Caatinga (SFAIR et al., 2018), servindo como um modelo para observar o impacto humano na diversidade genética de plantas da Caatinga (AECYO et al., in prep). É uma espécie enfermeira, facilitando o estabelecimento de outras plantas nesses ambientes (PATERNO et al., 2016). Já *C. pyramidale* é conhecida por seus compostos secundários, com potencial para a produção de medicamentos (CHAVES et al., 2015). Além disso, estas espécies têm servido de modelo em diversos trabalhos, como os que correlacionam tamanho do genoma e variáveis ambientais (VAN-LUME et al., 2019; SOUZA et al., 2019).

Ambas as espécies estão presentes no Parque Nacional (PARNA) do Catimbau, uma área de proteção ambiental que, com o auxílio do Programa Ecológico de Longa Duração (PELD – Catimbau), tem permitido o desenvolvimento de diversas pesquisas, principalmente sobre a conservação e o processo de desertificação da Caatinga, assim como questões socioambientais relacionadas à conservação deste ambiente (TABARELI et al., 2018). Durante essas pesquisas, alguns pesquisadores encontraram indivíduos, inicialmente identificados como *C. microphyllum*, que apresentaram características morfológicas diferentes do esperado para a espécie, principalmente quanto ao tamanho das pínulas e a quantidade de pinas por folha, levantando a hipótese de possíveis híbridos entre as duas espécies no local.

3 ARTIGO 1 – Does natural hybridization contribute to systematic complexity in *Cenostigma* (Leguminosae)? Integrative evidence on the intricacy of *C. microphyllum* and *C. pyramidale* species boundaries

Paulo Aecyo, Uedson Pereira Jacobina, Lucas Costa, Edeline Gagnon, Inara R. Leal, Gustavo Souza, Andrea Pedrosa-Harand

Artigo a ser submetido ao periódico ***Perspectives in plant ecology, evolution and systematics***

Does natural hybridization contribute to systematic complexity in *Cenostigma* (Leguminosae)? Integrative evidence on the intricacy of *C. microphyllum* and *C. pyramidale* species boundaries

Paulo Aecyo[a], Uedson Pereira Jacobina[b], Lucas Costa[a], Edeline Gagnon[c], Inara R. Leal[d] Gustavo Souza[a], Andrea Pedrosa-Harand[a]*

[a]*Laboratory of Plant Cytogenetics and Evolution, Department of Botany, Federal University of Pernambuco, Recife, Brazil,*

[b]*Laboratório de Sistemática Integrativa Molecular, Campus-Penedo, Universidade Federal de Alagoas, Avenida Beira Rio s/n, Penedo CEP 57200-000, Alagoas, Brazil*

[c]*Royal Botanic Garden of Edinburgh, University of Edinburgh*

[d]*Laboratory of Plant-Animal Interaction, Department of Botany, Federal University of Pernambuco, Recife, Brazil.*

*Corresponding author at: Laboratory of Plant Cytogenetic and Evolution, Department of Botany, Federal University of Pernambuco, Recife, Pernambuco, Brazil.

Contact information: andrea.harand@ufpe.br; telephone +55 81 2126 8846 and Fax +55 81 2126 8348

HIGHLIGHTS

- Molecular data and morphology identified hybridization in the genus *Cenostigma*
- Foliar morphological traits were correlated to genetic clusters and abiotic features
- Soil type may contribute to hybridization between *C. microphyllum* and *C. pyramidale*
- Hybridization contributes to systematic complexity in *Cenostigma*

ABSTRACT

Interspecific hybridization plays an important role in the evolution of plants, allowing introgression that may contribute to ecological adaptation and increase in fitness. However, hybridization also contributes to phylogenetic uncertainty. The genus *Cenostigma* comprises neotropical legume trees which showed morphological intermediary individuals among species and phylogenetic incongruences in previous studies. Thus, it was hypothesized that hybridization contributes to the systematic complexity in the group. Here we tested this hypothesis investigating two endemic species from the Caatinga Domain in Brazil: *C.*

microphyllum and *C. pyramidale*. Eight microsatellites (four nuSSR and four cpSSR) were used to access gene flow among 91 individuals distributed through six sympatric and four allopatric populations. In order to evaluate the morphological differences between the taxa, a geometric morphometric analysis was performed in the same sample of individuals. Species distributions were correlated with Caatinga geomorphologies, soil types, altitude and pluviosity to identify ecological preferences among pure and hybrid individuals. We recovered in both molecular and morphometric analyses a separation between species, but with a high proportion of intermediates among them. The hybrid individuals showed a clear correlation between molecular and foliar traits. Both species are distributed in crystalline and sedimentary areas and co-occurred in different soil types, altitudes and pluviosity regimes. Although they show slightly different preferences in soil types, they occur in similar proportion in *Neossolos*, what may facilitate hybridization. Thus, we demonstrate that hybridization occurs in *Cenostigma*, influencing the systematic complexity of the genus, and it is frequent and ancient between *C. microphyllum* and *C. pyramidale*.

Keywords: Caatinga; Caesalpinia group; geometric morphometrics; gene flow; introgression; Seasonally Dry Tropical Forest

1. Introduction

Speciation may be defined as the evolution of reproductive isolation among pairs of populations (Wang et al. 2020). It occurs as a continuum process during time, which usually culminate in the formation of two isolated species (Stankowski and Ravinet 2021). However, during the continuum of speciation, different lineages may hybridize due to incomplete pre- and/or post-zygotic isolation barriers (Stankowski and Ravinet 2021), leading to some genetic leakage among species caused by a porous genome (Wu and Thing 2004).

Hybridization may be defined not only as crossing between species, but also crossing between populations or races within a species (Soltis and Soltis 2009). Nowadays, it is estimated c. 10% of animals and 25% of plants may hybridize (Mallet 2005). Introgression of adaptive traits has long been seen as significant to the evolution of taxa (Mallet 2005) and its impact has been subject of recent studies (The *Heliconius* genome consortium 2012; Suarez-Gonzallez et al. 2016). Thus, a new hypothesis has been formulated concerning its importance on the diversification of clades, including its role in adaptive radiations (Naciri and Linder 2020).

The acknowledgement of a hybrid lineage is important in several biological fields, given its impact to the evolutive history of a taxa (Lamichhaney et al. 2018; Muniz et al. 2020). In a phylogenetic point of view, the presence of a hybrid individual may produce cytonuclear discordance leading to low supported trees and incorrect phylogenetic inference (Stefanocić and Costea 2008; Stull et al. 2020). Thus, several traits, including molecular markers, have been used to detect interspecific hybridization (Viscosi et al. 2009; Marques et al. 2018; Castillo-Mendonza et al. 2019; Souza et al. 2019).

Regarding phenotypic traits, hybrids may show different patterns of inheritance, such as the same phenotype as one of the parents, an intermediate phenotype or a transgressive trait (Castillo-Mendonza et al. 2019). Thus, phenotype analysis may suggest difficulties to delimit species boundaries, and indicate new ecological features induced by the introgression of adaptive traits. Introgression may, thus, interfere in the biological community (Castillo-Mendonza et al. 2019), lead to habitat shifts (Aizawa et al. 2020; Suarez-Gonzallez et al. 2016), the establishment of a new lineage and further to the formation of a new species (Lamichhaney et al. 2018). Traditionally morphometric studies have been widespread to delimit morphospecies boundaries using both vegetative and reproductive organs (Gagnon et al. 2015). Recently the advance of geometric morphometric analysis allowed to investigate species boundaries and hybridization with a less time-consuming approach (Viscosi et al. 2009).

Although morphological analysis is widely used to detect hybridization, it may fail to classify hybrids, specially second generation (and onward) hybrids and backcrossed individuals (Viscosi et al. 2009). Thus, population genetic studies have been largely chosen to identify the level of gene flow and introgression in a hybrid zone (Hu et al. 2019; Mota et al. 2019; Khan et al. 2020). Using neutral markers, such as microsatellites (SSR), it is possible to access the genetic diversity of species involved in the hybridization and to identify the direction of gene flow, the fertility of the hybrid, genetic swamping, and the level of genomic integrity of the species (Mota et al. 2019).

Cenostigma Tul. is a neotropical genus of the Caesalpinia group (Leguminosae), distributed from Mexico to Brazil through different biomes (Gagnon et al. 2016, 2019). It comprises ca. 14 species and 20 taxa and is characterized by its robust pods with a conspicuous thickened margin, but also by the presence of internal secretory cavities in the leaflet lamina and inflorescences, the presence of stellate indumentum on the stems, leaves and/or inflorescences (Gagnon et al. 2016). In a systematic point of view, the intergeneric relationships of the *Cenostigma* genus has been analysed (Lewis and Schrire 1995; Simpson

and Miao 1997; Nores et al. 2012; Gagnon et al. 2013, 2016), however no study focused on its infrageneric relationship has been so far carried out. The most robust study of the Caesalpinia group was performed by Gagnon et al. (2016), which analysed the phylogenetic relationships of the group based on five plastid loci (*rps16, trnD-trnT, ycf6-psbM, matK-3'-trnK intron,* and *trnL-trnF*) and the ribosomal *ITS1*-5.8S-*ITS2* DNA region. It included 28 accessions from 10 species of what is now considered *Cenostigma*. Although it is well supported, the *Cenostigma* phylogeny is not fully resolved, showing polytomies, low supported clades and non-monophyletic species (Gagnon et al. 2016).

Further, Aecyo et al. (2021) has performed a phylogenomic approach to unravel intergeneric relationships in the Caesalpinia group. They analysed 26 plastomes, representing ~57% of the genus of the group, including three accessions of two *Cenostigma* species from the Brazilian Caatinga: two accessions of *C. pyramidale* (Tul.) E.Gagnon & G.P.Lewis and one of *C. microphyllum* (Mart. Ex. G.Don) E.Gagnon & G.P.Lewis. However, a closer relationship was recovered between one accession of *C. pyramidale* and the *C. microphyllum* accession, not only in the phylogenetic reconstructions, but also in the plastome identity and distribution of SSRs. The authors proposed that this result could be explained by Incomplete Lineage Sorting (ILS) or plastome capture caused by interspecific hybridization.

In a morphological study, Lewis (1995) has proposed the hypothesis of hybridization among other species of the genus. Also, *C. microphyllum* and *C. pyramidale*, which seem to occur at different soil formations in Catimbau, present individuals with intermediate foliar morphology (personal observation). Thus, we addressed the following questions: (i) How clear are species boundaries among *C. microphyllum* and *C. pyramidale* sympatric and allopatric populations? (ii) Does hybridization contribute to the systematic complexity of *Cenostigma*? We hypothesized that (i) both species hybridize, though maintaining species integrity, (ii) contributing to systematic complexity in the genus, and (iii) that the presence of hybrids may be correlated with abiotic factors such as altitude, pluviosity and soil type.

2. Material and Methods

2.1. Study area and plant material

Cenostigma microphyllum and *C. pyramidale* are trees (~2 m high) endemic and well distributed throughout the Caatinga Domain, from the Succulent Biome. This domain is found exclusively in the northeast region of Brazil, ranging from 2º54' to 17º21' and comprising around 912,529 km^2 (da Silva et al. 2017) or around 11% of Brazilian territory. It is characterized by a shrub or arboreal vegetation, a semi-arid climate with pluviosity ranging

from less than 400 up to 1800 mm/year and temperature ranging from 25 to 30ºC (da Silva et al. 2017). The Caatinga is composed by several environments (Mota et al., 2014; Queiroz et al. 2017), but its terrestrial flora can be divided in two major floristic groups associated with the geomorphology: the Crystalline Caatinga (70% of the Caatinga area), which comprise rocky, shallow and medium to high fertile soils in the Sertaneja Depression; and the Sedimentary Caatinga (30% of the Caatinga area), presenting a sandy, profound and unfertile soils (da Silva et al. 2017; Queiroz et al. 2017). Nevertheless, the Caatinga is not a homogeneous region, with 135 geo-environmental areas recognized. Moreover, this domain presents the highest level of biodiversity in the Seasonally Dry Tropical Forest (SDTF) around the world, with high endemism (Queiroz et al. 2017). Furthermore, only 7.4% of its territory is in conservation unities, which contrast with the risk of aridization upon 94% of its region due to anthropogenic disturbance and climate changes (da Silva et al. 2017).

The Catimbau National Park (PARNA Catimbau) and Serra Branca Environmental Protection Area (APA Serra Branca) are two protected areas in the northeast of Brazil where both *Cenostigma* species occur. The PARNA Catimbau (8°24'00" and 8°36'35" S; 37°0′30″ and 37°1′40″ W) is localized in the Pernambuco state, and covers an area of nearly 640 km^2. It shows an annual pluviosity ranging from 480 – 1100 mm per year and a mean annual temperature of 23ºC (Rito et al. 2017). Includes both Crystalline and Sedimentary areas, but the latter is predominant (Rito et al. 2017). The APA Serra Branca (09º53’15.5” and 09º44’34.6” S; 38º49’36.1 and 38º52’20.4” W) covers an area 672.37 km^2 in Bahia state. It is localized in the ecoregion known as Raso da Catarina, which shows a semiarid climate with annual pluviosity around 500 mm per year and mean temperature of 23ºC. The Sedimentary basins predominates, with altitude varying from 350 – 700 m.

Cenostigma microphyllum and *C. pyramidale* share several morphological and ecological traits (Lewis 2005), such as bipinnate leaves, yellow zygomorphic flowers with a red marked standard petal, pollination by bees (Leite & Machado 2009; Novais et al. 2010) and blossoms between March and June during the rainy season. Regarding their distribution, both species are widespread in the Caatinga, being present in both Crystalline and Sedimentary areas (Moro et al. 2014).

The species show foliar morphological differences, such as the length of the petiole, presence of trichomes and number of pinnae (Lewis 2005). *Cenostima microphyllum* has a petiole 1 – 1.5 cm in length, 3 – 10 pairs of alternates to opposite pinnae per leaf, with 11 – 22 glabrescent oblongue-eliptic pinnule. Unlike it, *C. pyramidale* shows larger leaves, with a

petiole length of 1.5 – 2.4 cm, 1 – 3 pairs of opposite pinae per leaf, with pubescent ovate pinnule.

To perform molecular analysis, fresh leaves from a total of 95 individuals distributed in 10 populations were collected and stored in silica (Fig. 1, Table 1). From those populations, four were sympatric populations of *Cenostigma microphyllum* and *C. pyramidale* from PARNA Catimbau, and two were sympatric population from APA Serra Branca. Allopatric populations of each species were also collected, two *C. microphyllum* and *C. pyramidale* populations from Paulo Afonso – BA and two of *C. pyramidale* from Jeremoabo – BA and Belém – Alagoas state (AL).

2.2. SSR detection and primer design

Unassembled reads of *Cenostigma microphyllum* were previously obtained (GenBank accession number: SRX11185454, Aecyo et al. 2021) and used to detect and develop nuclear microsatellites (nuSSR) for population analysis in the species. Microsatellites were identified in the sample using Phobos software (Mayer 2010) as a plugin in Geneious v7.1.9. The minimum number of repeats were: ten for mono-, five for di-, four for three, three for tetra-, penta- or hexanucleotides. Primers designed and polymorphisms tests were performed as described in Aecyo et al. (2021).

2.3. DNA extraction, amplification, cross-amplification and genotyping

Total genomic DNA was extracted from ca. 50 mg leaves stored on silica gel following the cetyltrimethylammonium bromide (CTAB) protocol of Doyle and Doyle (1987), as in Ferreira & Grattapaglia (1995), and stored at -20°C.

To analyse the genetic diversity and structure of the chloroplast genome, a total of four loci previously detect by Aecyo et al. 2021 were used: three developed for *C. microphyllum* (CmCPSSR4, CmCPSSR12 and CmCPSSR15; Aecyo et al. 2021) and one developed for *Nicotiana tabacum* L. (ccmp2; Weising and Gardner 1999) (Table 2).

The nine nuSSR primer pairs designed for *C. microphyllum* were cross-amplified to other five species from the Caesalpinia group (Leguminosae): *Caesalpinia pulcherrima* (L.) Tod, *Cenostigma pyramidale, Guilandina bonduc* L., *Libidibia ferrea* (Jacq.) Schltdl. and *Paubrasilia echinata*. Amplifications were performed following the protocol of Aecyo et al. (2021) for cpSSR loci, testing amplification at 50ºC, 53ºC and 56ºC. The PCR products were

analysed in an 3% agarose gel. A subsampling of eight individuals of *Cenostigma microphyllum* and *C. pyramidale* from our sample were screening for polymorphism in the cross-amplified loci between both species. The four most polymorphic loci were used for detecting gene flow among populations. The loci were amplified using the M13 tail as developed by Schuelke et al (2000). The amplification and genotyping protocol were performed as described by Aecyo et al. (2021).

2.4. Genetic diversity and structure accessed by nuSSR

Null alleles were inferred for each loci using the MSA v. 4.05 software (Dieringer and Schlötterer 2003). The genetic diversity indices were calculated for each locus and population. The Genepop v. 4.7.5 (Raymond and Rousset 1995; Rousset 2008), MSA v. 4.05 and FSTAT v. 2.9.4 (Goudet 1995) software were used to calculate the total number of alleles (*A*), observed heterozygosity (H_O), expected heterozygosity (H_E) and inbreeding coefficient (F_{IS}). Departure from the Hardy-Weinberg equilibrium for each locus within each population was determined using the Genepop software. *F*-statistic were also calculated for each locus and population to estimate genetic diversity of nuclear loci.

To estimate the admixture proportion from *Cenostigma microphyllum* and *C. pyramidale* per individual, we performed a model-based Bayesian clustering analysis on STRUCTURE 2.3.4 (Pritchard et al. 2000). We used the online platform CLUMPAK (http://clumpak.tau.ac.il/) to test the Δ*K* of Evanno (Evanno et al., 2005), and used the genetic cluster (*K*) value = 2, corresponding to the gene pool of two species. We estimated the admixture proportion using a burn-in period of 250,000 and run length of 1,000,000 with the Markov Chain Monte Carlo (MCMC) method. No prior population conditions were used for analysis. We classified each individual as pure or hybrid based on the *q* (admixture proportion) value, as in Mota et al. (2019). Individuals were classified as pure when $0.10 > q > 0.90$. As for hybrid categories, we consider the individuals which showed *q* value between 0.10 and 0.90 ($0.10 < q < 0.90$). To further characterize the genetic structure of the ten population, we performed this analysis varying the *K* value from 2 to 10 clusters, with and without the hybrids identified in the previous analysis.

To access patterns of genomic differentiation between populations, an Analysis of Molecular Variance (AMOVA) was performed among populations and among species within each population, withdrawing the hybrid individuals identified with STRUCTURE. Also, we estimated the genetic differentiation index (F_{ST}) in our sample. Both analyses were performed

with ARLEQUIN v. 3.5 software (Excofier and Lischer 2010) with 9999 permutations and default configurations.

2.5. Genetic diversity and structure accessed by cpSSR

The Haplotype v.1.05 software (Eliades and Eliades 2009) was used to analyse the genetic diversity of the cpSSR loci of the sample. It recovered the total number of haplotypes (A), private haplotypes (P), haplotypic richness (Hr) and the genetic diversity (H). The Network v.5.0.0.3 software (http://www.fluxus-engineering.com) was used to draw the haplotype network. The geographic coordinates for each population were plotted in a map using the QGIS software and the proportion of each haplotype per population was represented in the same map. An analysis of molecular variance (AMOVA) and genetic structure was also performed for plastidial loci as described above.

2.6. Geometric morphometrics

All genotyped specimens analysed were photographed under scale, using a Canon T3i camera attached to a tripod. It was analysed four leaflet per leaf and around ten leaves per individual. The images were analysed using Tpsutil software (Rohlf 2010a) and TpsDig2 software (Rohlf 2010b). In total, four landmarks were chosen (the apex and base of each leaflet, and the second and fourth insertion of the vein on the left side), as well as other four semilandmarks by means of a perpendicular projection, aiming to symmetrically contour the leaflet margin. The landmark and semilandmark configurations for each profile were submitted to Generalized Procrustes Analysis (GPA) (Dryden and Mardia 1998). GPA removes any variations related to specimen position, size, or rotation (Rohlf and Slice 1990). Subsequently, a principal component analysis (PCA) was performed using a covariance matrix, in order to assess the general magnitudes and patterns of variations between the forms of the analysed species and hybrids. In addition, centroid size was used as a measure of shape size in our study for statistical correlation analysis. All analyses were performed with MorphoJ 2.0 software (Klingenberg 2011) and in the Excel 2017 (Microsoft).

2.7. Ecological traits

To investigate if species distribution and ecological traits were correlated with the rate of hybridization between *C. microphyllum* and *C. pyramidale*, we downloaded general occurrence points of both species from the Global Biodiversity Information Facility (GBIF)

website (https://www.gbif.org). To minimize the effect of erroneous distribution data, only individuals with vouchers deposited in herbaria were recorded and the data was cleaned to exclude oceanic points, and locations that were unlikely to be natural occurrence. The coordinates of our collection points were merged with the GBIF coordinates and plotted in a map using the software DIVA-GIS (Hijmans et al. 2004). A shapefile containing the geographic distribution of soil types in Brazil was downloaded from the Brazilian Agricultural Research Corporation (EMBRAPA) geographic database (http://geoinfo.cnps.embrapa.br/) and added to the DIVA-GIS map. Soil types were defined based on the classification of the Brazilian Soil Classification System (SiBCs, Santos et al., 2018). To gather geologic information on the coordinate locations (crystalline basement versus sedimentary cover), a second shapefile containing information of Brazilian geologic provinces was downloaded from the Geological Survey of Brazil database (https://geosgb.cprm.gov.br/geosgb) and was also added to the DIVA-GIS map. Information of soil type and geological province of each collection point was extracted using the "extract value by points" option implemented in DIVA-GIS.

From the collection sites of each species as well as the downloaded GBIF points, we extracted altitude and precipitation variables from the WorldClim 1.4 (5 min) generic grid format (Hijmans et al. 2005) using the "extract values by points" function of DivaGis. Using the R package 'stats' (R Core Team 2019), we employed Pearson's correlation to check if genotypic (*q* value) or phenotypic (leaflet centroid value) values were correlated with altitude and precipitation variables for the collected individuals.

3. Results

3.1. Development of nuSSR loci of *Cenostigma microphyllum* and cross-amplification of nuSSR

A total of nine nuSSR loci were developed for *Cenostigma microphyllum*, since CmSSR10 was monomorphic in the subsample of eight individuals of *C. microphyllum* tested (Table S1). CmSSR1 to 9 showed polymorphism of 2 – 9 alleles per loci in this subsample. All ten primer pairs were tested for cross-amplification in five Caesalpinia group species. The loci CmSSR5, CmSSR7, CmSSR8 and CmSSR9, as well as CmSSR10, cross-amplified in *Cenostigma pyramidale* only, while the locus CmSSR1 was the only one which cross-amplified in all five species (Table S2). The loci CmSSR1, CmSSR2, CmSSR4 and CmSSR6

were the most polymorphic in *C. microphyllum* and thus chose for the population genetic analysis.

3.2. Admixture between *Cenostigma microphyllum* and *C. pyramidale* by STRUCTURE analysis

We successfully genotyped 91 individuals sampled in ten natural populations of *Cenostigma microphyllum* and *C. pyramidale* with the selected nuSSR loci (CmSSR1, CmSSR2, CmSSR4 and CmSSR6) (Table 1, Fig. 1). The Bayesian STRUCTURE analysis performed with $K = 2$ was able to discriminate *Cenostigma microphyllum* and *C. pyramidale* individuals, grouping them into two different clusters. Considering the threshold of $0.10 > q > 0.90$ for hybrid individuals, it was possible to classify a total of 35 individuals as *C. pyramidale*, 20 individuals as *C. microphyllum* and 36 individuals from hybrid origin (Fig. 1, Table 2). Withdrawing the hybrid individuals from Bayesian analysis and screening from $K = 2$ to $K = 10$ to check for population differentiation, it is possible to observe that both species maintained its genomic integrity, and no intraspecific differentiation among populations was observed (Fig. S1). Considering the distribution of hybrid individuals among the six sympatric populations, a higher proportion of hybrid individuals was observed in Alcobaça (64,7%) and Brejo (62,5%), both from the PARNA Catimbau and APA1 (61%) from APA Serra Branca. They were rare in Pedra, APA2 and in three of four allopatric populations, with one hybrid each. Only in Jeremoabo none of the three collected individuals was presumably of hybrid origin (Table 1).

3.3. Genetic diversity in sympatric and allopatric populations accessed by nuSSR and cpSSR markers

The four nuSSR loci were highly polymorphic in the present sample, showing from nine (CmSSR1) to 19 alleles (CmSSR2) per loci (Table 1). The inbreeding coefficient (F_{IS}) was low in almost all loci, except for the loci CmSSR1 ($F_{IS} = 0.245$) that showed a departure from the Hardy-Weinberg equilibrium with an excess of homozygotes (Table 3). Also, the null allele test showed values over 10% for this locus in two populations (30% in Açude and 21.5% in APA1). All loci showed a high genetic diversity, with H_E varying from 0.687 to 0.841 per locus, with an overall diversity of 0.790 (Table 3). A high genetic diversity was observed for all populations, with a mean of 0.783 (Table 4). The lower H_E was found in Pedra ($H_E = 0.611$) and the highest was found in Brejo ($H_E = 0.835$). We also observed

similar levels of genetic diversity between pure and hybrid individuals for each population (Table 1).

Six cpSSR loci were successfully amplified and genotyped for 74 individuals. However, loci ccmp10 and CmCPSSR8 were monomorphic in the present sample and not included in further analyses. The remaining four loci showed from two to six alleles per locus (Table 1), and an overall genetic diversity of H = 0.112 (Table 4). In combination, the thirteen alleles revealed the presence of seven haplotypes (Table S3), showing a haplotypic richness of Rh = 0.190 (Table 4).

Haplotypes H1 and H2 were private and the only found in *Cenostigma pyramidale* populations from Jeremoabo and Belém, respectively. Haplotypes H3, H4 and H5 were shared among APA1 and APA2 populations (two sympatric populations of *C. microphyllum* and *C. pyramidale*), PA1 (one population of *C. pyramidale*) and PA2 (one population of *C. microphyllum*). Haplotypes H6 and H7 were shared among individuals from PARNA Catimbau, with H6 present in Açude and Pedra and H7 in Alcobaça and Brejo (Fig 1).

The haplotype network divided the sample into two groups, which were related to the geographic distance instead of species delimitation (Fig. 1). Thus, the populations from PARNA Catimbau, Pernambuco, were genetically distant from the other six populations from Alagoas and Bahia. Although haplotypes H3 to H7 were found in populations of both species, the haplotypes H3 and H4 were found only in individuals genetically classified as *Cenostigma microphyllum* and hybrids. Haplotype H7 was found in individuals classified as *C. pyramidale* and hybrids. Only H5 and H6 were found in pure individuals of both species (Table S3).

3.4. Analysis of Molecular Variance and genetic structure among species and populations

An overall low degree of genetic differentiation as observed among populations (F_{ST} = 0.0558) and among genetic groups (pure *C. microphyllum*, pure *C. pyramidale* and hybrids) classified by means of the *q* value within each population (Table 1). To investigate the molecular variation between species and among population within each species, we removed the individuals classified as hybrids after STRUCTURE analysis. AMOVA showed that the nuclear genetic variation is present mainly within species (79.23%) rather than between species (20.77%; Table 5). This result is similar within each of the two species, which showed a higher genetic variation within populations than among them (F_{ST} = 0.05 and 0.04 for *C.*

microphyllum and *C. pyramidale*, respectively). Also, both species showed a low genetic differentiation and a high genetic diversity. In the cpSSR analysis, again the genetic variation was mainly observed within species (97.15%), with a low genetic structure among species (F_{ST} = 0.0285). However, most of the variation was observed among, not within populations (F_{ST} = 0.93 and 0.95 for *C. microphyllum* and *C. pyramidale*, respectively), indicating a high genetic structure for the chloroplast loci (Table 5).

3.5. Morphological differentiation revealed by geometric morphometric analysis

Principal component analysis (PCA) revealed two morphologically differentiated groups when considering only genetically pure *C. microphyllum* and *C. pyramidale* individuals. However, overlaps in morphospace were seen between the hybrids and both species, but mostly with *C. microphyllum* (Fig. 2). Indeed, some individuals with intermediate morphology were in fact pure *C. pyramidale* or *C. microphyllum* and some individuals attributed to one or the other species were in fact ybrids (Table 2). Most of the morphological variation found between species were present in the first two axes, PC1 (38%) and PC2 (18%), which explained 56% of the magnitude of the morphological variation detected (Fig 2). Furthermore, the size of the centroid extracted from the landmarks and semilandmarks was significantly correlated with the genetic admixture of the species (R^2 = 0.5155, p = 0.001, DF = 24).

3.6. Ecological traits and correlation among genetic clusters

Despite the high proportion of introgression estimated, the species could be largely differentiated morphologically. Nevertheless, pure parental individuals occasionally showed intermediate morphology, indicating morphological plasticity for both species. Therefore, we investigated if some ecological traits could correlate with morphology, suggesting that phenotypes could be, at least in part, determined by the abiotic environment. For that, we expanded our sampling to the whole geographic distribution of the species. Soil type analysis showed that most distribution points of *Cenostigma microphyllum* and *C. pyramidale* were concentrated in five different soil types: *Planossolo, Luvissolo, Argissolo, Neossolo* and *Latossolo* (Fig. 3A). While *C. pyramidale* preferentially occurred in *Planossolo* and *Luvissolo* soil types, *C. microphyllum* was predominant in *Latossolos*. However, both species had a high frequency of distribution in the *Neossolo* type, with a slightly higher frequency of *C. microphyllum* samples. Five out of the 10 populations sampled in this study were found in *Neossolo* (Fig. 3A), and all were sympatric. The only sympatric population that showed a

different distribution was Alcobaça, which was distributed through *Planossolo* and composed by *C. pyramidale* and hybrid individuals.

Most of the downloaded distribution points of *C. pyramidale* and *C. microphyllum* were present at crystalline formations, but both species showed similar proportions at both geomorphologies (Fig. 3B, B' and B"). However, most of our populations, sympatric and allopatric, were found in sedimentary formations, with the "Açude" population having individuals in both geomorphologies (Fig. 3B). This location was also the most variable in terms of altitude (Fig 3C) and precipitation (Fig 3D), with around half of the individuals being pure *C. pyramidale* and the other half, pure *C. microphyllum*, with two hybrids, the only individuals found in the Crystalline basement in this population. Both populations from Paulo Afonso (1 and 2) occurred in the crystalline basement, but while Paulo Afonso 1 had two *C. pyramidale* individuals and one hybrid, Paulo Afonso 2 showed two *C. microphyllum* individuals and one hybrid.

Cenostigma pyramidale presented a wide range of altitudes, with most points being between 200 to 600 meters. Altitude distribution of *C. microphyllum* was more uniform, with most distribution points being at around 400 meters of altitude (Fig. 3C). The sampled populations were located at higher altitudes (>600 m), with few populations (APA 1 and 2, Paulo Afonso 1 and 2) being at relatively lower altitudes (approx. 200 m). Thus, we could confirm the occurrence of pure individuals of both species between 200 m (APA 1 and 2) and >800 m (Açude and Pedra). We also analysed the variation of annual pluviosity along the distribution of the species. Both species show a similar range of pluviosity along its distribution (Fig. 3D), being the majority of the points present in regions with low annual precipitation (<700 mm). The populations from Bahia and Alagoas states presented a lower annual precipitation than the populations of PARNA Catimbau (~800 mm), except for the population of Belém, which showed the highest annual precipitation among our sample (>1000 mm).

In order to access the influence of abiotic factors to the genetic and morphological variation, we performed correlations between *q* values obtained in the STRUCTURE analysis and the foliar morphology, by the size of the centroid found in our geometry morphometrics analysis, and nine abiotic traits (Table 6). We found no correlation to the genetic assignment of the individuas, but a small, although significant, correlation between leaf morphology and altitude ($R^2 = 0.2$, $p = 0.04$) and pluviosity indices, especially the precipitation of the wettest quarter ($R^2 = 0.35$, $p = 0.0001$). Thus, although the foliar morphology was mainly influenced by genetic inheritance, abiotic traits play a role on the phenotype of the hybrids.

4. Discussion

4.1. Species delimitation and hybridization in *Cenostigma*

To define a group of organisms as belonging to the same or different species is a challenging task since several concepts may be used to define it (De Queiroz 2007). Thus, different markers should be used to access species boundaries among individuals in case complete reproductive isolation and genetic, morphological and ecological differentiation has not been achieved. *Cenostigma microphyllum* and *C. pyramidale* are two endemic species from the Brazilian Caatinga and share several morphological, ecological, cytogenetic and genomics traits (Lewis 2005; Leite & Machado 2009; Novais et al. 2010 Gagnon et al. 2016, Van-lume et al. 2017; Aecyo et al. 2021). Despite the morphological similarity and low reproductive barriers between the species, we recovered that both species are separated into two genetic clusters and by leaf morphological traits (Fig. 1, Fig 2). Nevertheless, it was possible to observe some level of admixture in almost all individuals, with 36 (around 40% of our sample) classified as hybrid by the STRUCTURE analysis, with many showing intermediate morphological traits. Thus, for the first time in the genus *Cenostigma*, we demonstrate that the two arboreous species *C. microphyllum* and *C. pyramidale* do hybridize in nature, possibly frequently and over a long period of time, which is remarkable give the estimated old age of the genus (~14 Mya) and the separation of the species (~6.5 Mya; Gagnon et al. 2019). We also hypothesize that the hybrid individuals may show some fertility due to the continuous degree of admixture among intermediary individuals and the morphometric analysis.

The detection of hybrid individuals is extremely important in phylogenetic studies. Since a hybrid individual is the mixture of two (or more) genetic pools, it may influence the phylogenetic inference of its parental species, resulting in low supported clades and non-monophyletic taxa. Also, since the plastome has mainly a uniparental inheritance, it may lead to cytonuclear discordance, which is observed as an incongruence among plastidial and nuclear phylogenies (Stefanocić and Costea 2008; Zhou et al. 2020), or even as incongruence in the plastome phylogeny itself (Aecyo et al. 2021). All those incongruences were observed in previously phylogenetic studies in the Caesalpinia group (Gagnon et al 2016; Aecyo et al. 2021). Although hybridization may account for these incongruences, there are other natural processes that may have the same effect such as homoplasy observed in the SSR or

Incomplete Lineage Sorting. Regarding homoplasy, we do not believe it is influencing our results given the pattern recovered in the morphometric analysis, which corroborate the STRUCTURE analysis.

Incomplete Lineage Sorting plays an important role in evolution specially in recently formed clades and evolutionary radiations (Maddison 1997; Stull et al. 2020). Although some studies have been focused in discriminating the effects of ILS and introgression (Meyer et al. 2017; Stull et al. 2020), it is not easy to discriminate among them in a phylogenetic inference. Even though the effect of ILS must be analysed in the phylogeny of the *Cenostigma* genus, the effect of hybridization cannot be overseen. The nuSSR markers could clearly demonstrate hybridization between *C. microphyllum* and *C. pyramidale*, which was corroborated by cpSSR and geometric morphometric analysis. In addition, the occurrence of both species in sympatry, in overlapping soil types, altitude and rainfall ranges indicate that extrinsic barriers to hybridization do not exist or are not as impactful.

In fact, other species from *Cenostigma* also seem to hybridize. Lewis (1995) hypothesized that *C. laxiflorum* may hybridize with both *C. microphyllum* and *C. pluviosum* through its distribution. Although hybridization is commonly investigated among two species, the pattern of hybridization in a hybrid zone may be more complex involving more species, as observed in several groups such as *Nothoscordum* Kunth. (Amaryllidaceae, Souza et al. 2012), *Eucalyptus* (Myrtaceae, Robbins et al. 2021), *Melocactus* Link & Otto (Cactaceae, Khan et al. 2020) and *Populus* L. (Salicaceae, Chattre et al. 2018). We believe that a similar pattern could be found in the *Cenostigma* genus and should be further investigated to quantify its impact on the evolution of the group. The SSR developed in this study may be helpful to achieve this purpose because they could be transferred between species from the same genus and eventually closely related genera.

4.2. Effect of introgression in morphology and geographic structure

The role of hybridization in the evolution of a taxon may have different outcomes. If the hybrid is formed, the outcome will be defined by how fertile the hybrid individuals are. If the formed individual is infertile or has less fertility than the parental, it may decrease the genetic diversity of the parental species and could lead it to extinction (Muniz et al. 2020) in a process called genetic swamp. In the other way, if the hybrid is fertile, it may even promote gene flow among species and contributes to maintain their genetic diversity (Ma et al. 2019).

In our study, we neither quantify the level of introgression among the species nor evaluated the hybrid classes among our sampling. However, it is possible to infer the fertility of the hybrids based on the continuous introgression observed with STRUCTURE (Schley et al. 2020), as well as the pattern of the morphometric analysis. Also, nuSSRs identified a high genetic diversity in each species and in the hybrid individuals, and a low genetic structure among all populations (in both nuclear and plastidial loci), which may indicate an ongoing gene flow among them (Khan et al. 2020), what needs to be further tested. Furthermore, genetic differentiation between species was moderate only (Mota et al. 2019; Khan et al. 2020).

Unlike nuSSR, a high genetic structure was observed in cpSSR analysis. A total of seven haplotypes were recovered among the ten populations, with eight populations showing a single haplotype each and a geographical distribution in accordance with the haplotype network. Pollination and seed dispersal work together to keep the genetic diversity with and among populations, however both processes have different impact in the maintenance of the gene flow (Gonçalves-Oliveira et al. 2016; Mota et al. 2019). Notwithstanding, it is known that both *Cenostigma* species are pollinated by bees (Leite & Machado 2009; Novais et al. 2010), although there is a lack of formal studies with *C. microphyllum*, and their seeds are dispersed ballistically (Lewis 2005). Thus, the offspring grow closer to its parental, which explain the geographic structure observed in the plastidial loci. Besides, despite being geographically close, the populations from PARNA Catimbau showed a high genetic structure of cpSSR. This can be explained by the distribution of the populations, since Açude and Pedra populations are in higher altitudes and separated from Alcobaça and Brejo by a geographic barrier, a granitic formation in which the Alcobaça population is localized.

Morphological analysis is useful to identify introgressive traits (Aizawa et al. 2019), however, some vegetative characters such as foliar traits may show some plasticity along ecological clines (Hopkins et al. 2008). Indeed, our study showed plasticity in morphological traits, and overlap between species. Nevertheless, correlation between foliar morphology (centroid size), and genetic composition (*q*-value obtained by the STRUCTURE analysis) was observed, what may be a good indication of the level of introgression among both *Cenostigma* species. But also, some abiotic traits as pluviosity and altitude contributes in some degree to morphological plasticity among the individuals.

4.3. Ecological conditions and hybridization

Once formed, the hybrid may function as a genetic bridge, allowing the introgression of different traits among the species (Wang et al. 2021). The introgressed traits may play a role in the species distribution (Aizawa et al. 2019), resistance to drier and cooler habitats (Ma et al. 2019), photosynthetic capacity (Suarez-Gonzalez et al. 2016) and pollination patterns (Wang et al. 2021).

Cenostigma microphyllum and *C. pyramidale* share some ecological features. Although *C. pyramidale* shows a larger range of distribution, the genetic assignment of individuals to species confirmed that both are found in the same altitude range, more frequently between 400 – 500 m. Furthermore, both species are present in different plant communities and distributed through both crystalline and sedimentary geomorphologies. They occur more frequently in the crystalline basement, which represents around 70% of the Caatinga area (Moro et al. 2014). Nevertheless, although overlapping, each species is more frequent at different soil types, except for *Neossolos*, which showeda similar distribution of both species and where the majority of the sympatric populations were sampled. This may indicate that *Neossolos* could provide ideal conditions for both species to thrive, what could eventually facilitate the establishment of hybrids. However, this hypothesis needs to be tested.

It is known that climate changes influence the distribution of a species and may facilitate secondary contact among almost divergent lineages leading to hybridization (Fu et al. 2020; Ma et al. 2019). However, in our analysis, we did not find any correlation of altitude or rainfall with the distribution of pure or hybrid individuals. Another factor that may facilitate interspecific gene flow is anthropogenic disturbances (Abbot 2017). Although the majority of the Caatinga native vegetation is preserved, around 63% of its area is of human use and its impact is higher in areas with high pluviosity, such as in PARNA Catimbau, than in dryer area as in the APA Serra Branca and the Raso da Catarina itself (Silva and Barbosa 2017; Rito et al. 2017). The anthropogenic disturbance may be influencing the level of hybridization among species by favouring the establishment of hybrid individuals, which should be tested. Indeed, our study find a higher proportion of hybrid individuals in populations at the PARNA Catimbau (45%) than in APA Serra Branca (30%). Future studies should investigate if the human perturbance is influencing the rate of hybridization among *Cenostigma* species or whether introgression may facilitate adaptation in anthropogenic environments.

5. Conclusion

In summary, our study demonstrates that interspecific hybridization occurs between *Cenostigma microphyllum* and *C. pyramidale*, two arboreous species endemic from Caatinga. Also, the hybrid individuals must be fertile and gene flow may occur in both directions given the haplotype share among populations, contributing to maintain the high genetic diversity of both species, as well as reducing genetic differentiation without compromising their genetic integrity. Although plastid haplotypes demonstrated high population structure, species differentiation was very low, which may suggest a long term hybridization. Albeit this high gene flow, morphometric analysis showed correlation between foliar morphology and genetic composition, being able to discriminate among both species with hybrid individuals being plotted between species. It demonstrates that leaf morphology is a good trait to infer hybrid individuals. Both species occurred in similar proportions in both Crystalline and Sedimentary Caatinga, in the same soil types, but with different preferences, demonstrating that extrinsic barriers correlated to soil composition may not exist or be weak among species. Besides, other factors such as human disturbances may be contributing to the frequency of hybridization and must be further analysed.

Author contributions

Paulo Aecyo: Formal analysis, Investigation, Data Curation, Writing - Original draft preparation. **Uedson Pereira Jacobina**: Formal analysis, Investigation, Writing - Original draft preparation. **Lucas Costa**: Formal analysis, Investigation, Writing - Original draft preparation. **Edeline Gagnon**: Conceptualization, Writing – Review & Editing **Inara R. Leal**: Conceptualization, Writing – Review & Editing, Funding acquisition **Gustavo Souza**: Conceptualization, Resources, Writing – Review & Editing, Visualization, Supervision, Funding acquisition **Andrea Pedrosa-Harand**: Conceptualization, Resources, Writing – Review & Editing, Supervision, Project administration, Funding acquisition

Acknowledgments

We thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, PELD process 403770/2012-2, Universal process 426738/2018-7), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES code 0001) and the Fundação de Amparo à Ciência e Tecnologia de Pernambuco (FACEPE, processes BIC-0846-2.02/17, BIC-0624- 2.02/18) for financial support. We also thank Géssica Souza for collecting some material, to Cícero Almeida for a initial genotyping test and the Sequencing Platform of CB-UFPE and Heidi Lacerda from LABBE for their help with the genotyping of our sample.

APH, GS and IRL also thank CNPq for productivity grants (processes 310804/2017-5, 310693/2018-7 and 308300/2018-1, respectively).

Funding: This work was supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, PELD process 403770/2012-2, Universal process 426738/2018-7); the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES code 001) and the Fundação de Amparo à Ciência e Tecnologia de Pernambuco (FACEPE, processes BIC-0846-2.02/17, BIC-0624- 2.02/18).

References

Abbot, R.J., 2017. Plant speciation across environmental gradients and the occurrence and nature of hybrid zones. J. Syst. Evol. 55, 238-258. https://doi.org/10.1111/jse.12267.

Aecyo, P., Marques, A.S., Huettel, B., Silva, A.L.G., Esposito, T., Ribeiro, E., Leal, I.R., Gagnon, E., Souza, G., Pedrosa-Harand, A., 2021. Plastome evolution in the Caesalpinia group (Leguminosae) and its application in phylogenomics and populations genetics. Planta. 254, 27. https://doi.org/10.1007/s00425-021-03655-8

Aizawa, M., Iwaizumi, M.G., 2020. Natural hybridization and introgression of *Abies firma* and *Abies homolepis* along the altitudinal gradient and genetic insights into the origin of *Abies umbellate*. Plant Species Biol. 35, 147-157. https://doi.org/10.1111/1442-1984.12269.

Castillo-Mendonza, E., Salinas-Sánchez, D., Valencia-Cuevas, L., Zamilpa, A., Tovar-Sánchez, E., 2019. Natural hybridisation among *Quercus glabrescens*, *Q. rugosa* and *Q. obtusata* (Fagaceae): Microsatellites and secondary metabolites markers. Plant Biol. 21, 110-121. https://doi.org/10.1111/plb.12899.

Chhatre, V.E., Evans, L.M., DiFazio, S.P., Keller, S.R., 2018. Adaptive introgression and maintenance of a trispecies hybrid complex in range-edge populations of Populus. Mol. Ecol., 27, 4820-4838. https://doi.org/10.1111/mec.14820.

Da Silva, J.M.C., Barbosa, L.C.F., Leal, I.R., Tabarelli, M., 2017. The Caatinga: Understanding the Challenges. In: Silva, J.M.C., Leal, I.R., Tabarelli, M. (Eds), Caatinga. Springer, Cham. https://doi.org/10.1007/978-3-319-68339-3_1.

De Queiroz, K., 2005. Ernst Mayr and the modern concept of species. PNAS. 102, 6600-6607. https://doi.org/10.1073/pnas.0502030102.

De Queiroz, K., 2007. Species Concepts and Species Delimitation. syst. Biol. 56, 879-886. https://doi.org/10.1080/10635150701701083.

Dieringer, D., Schlötterer, C., 2003. microsatellite analyser (MSA): a platform independent analysis tool for large microsatellite data sets. Mol. Ecol. Notes. 3, 167-169. https://doi.org/10.1046/j.1471-8286.2003.00351.x.

Doyle, J.J., Doyle, J.L., 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull. 19: 11-15

Dryden, I.L., Mardia, K.V., 1998. Statistical Analysis of Shape. Wiley, Chichester.

Eliades, N.G.H., Eliades, D.G., 2009. Haplotype Analysis: software for analysis of haplotype data. Distributed by the authors. Forest Genetics and Forest Tree Breeding, Georg-August University Goettingen, Germany. Available at http://www.uni-goettingen.de/en/134935.html

Excoffier, L., Lischer, H.E.L., 2010. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Mol. Ecol. Res.. 10, 564-567. 10.1111/j.1755-0998.2010.02847.x.

Evanno, G., Regnaut, S., Goudet, J., 2005. Detecting the numbers of clusters of individuals using the software structure: a simulation study. Mol Ecol 14, 2611–2620. https://doi.org/10.1111/j.1365-294X.2005. 02553.x.Ferreira, M.E., Grattapaglia, D., 1995. Introdução ao uso de marcadores moleculares em análise genética. EMBRAPA-CENARGEN, Brasília.

FU, P.C., SUN, S.S., KHAN, G., DONG, X.X., TAN, J.Z., FAVRE, A., ZHANG, F.Q., CHEN, S.L., 2020. Population subdivision and hybridization in a species complex of *Gentiana* in the Qinghai-Tibetan Plateau. Ann. Bot., 125, 677-690. https://doi.org/10.1093/aob/mcaa003

Gagnon, E., Lewis, G.P., Sotuyo, J.S., Hughes, C.E., Bruneau, A., 2013. A molecular phylogeny of *Caesalpinia* sensu lato: Increased sampling reveals new insights and more genera than expected. S Afr J Bot 89, 111-127. https://doi.org/10.1016/j.sajb.2013.07.027

Gagnon, E., Hughes, C.E., Lewis, G.P., Bruneau, A., 2015. A new cryptic species in a new cryptic genus in the *Caesalpinia* group (Leguminosae) from the seasonally dry inter-Andean valleys of South America. Taxon. 64, 468-490. http://dx.doi.org/10.12705/643.6.

Gagnon, E., Bruneau, A., Hughes, C.E., De Queiroz, L.P., Lewis, G.P., 2016. A new generic system for the pantropical *Caesalpinia* group (Leguminosae). PhytoKeys 71, 1-160. https://doi.org/10.3897/phytokeys.71.9203.

Gagnon, E., Ringelberg, J.J., Bruneau, A., Lewis, G.P., Hughes, C.E., 2019. Global Succulent Biome phylogenetic conservatism across the pantropical Caesalpinia Group (Leguminosae). New Phytol. 222, 1994-2008. https://doi.org/10.1111/nph.15633.

Gonçalves-Oliveira, R.C., Wöhrmann, T., Benko-Iseppon, A.M., Krapp, F., Alves, M., Wanderley, M.G.L., Weising, K., 2019. Population genetic structure of the rock outcrop species Encholirium spectabile (Bromeliaceae): The role of pollination vs. seed dispersal and evolutionary implications. Am. J. Bot. 104, 868-878. https://doi.org/10.3732/ajb.1600410.

Goudet, J., 1995. FSTAT (version 1.2): a computer program to calculate F-statistics. J Hered. 86:485-486. https://doi.org/10.1093/oxfordjournals.jhered.a111627.

Hijmans, R.J., Guarino, L., Bussink, C., Mathur, P., Cruz, M., Barrentes, I. et al. 2004. DIVA-GIS. Vsn. 5.0. A geographic information system for the analysis of species distribution data. (manual available at: http://www.diva-gis.org)

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25, 1965–1978. https://doi.org/10.1002/joc.1276.

Hopkins, R., Schmitt, J., Stinchcombe, J.R., 2008. A latitudinal cline and response to vernalization in leaf angle and morphology in Arabidopsis thaliana (Brassicaceae). New Phytol. 179, 155-164. https://doi.org/10.1111/j.1469-8137.2008.02447.x

Hu, Y.N., Zhao, L.Z., Buggs, R.J.A., Zhang, X.M., Li, J., Wang, N., 2019. Population structure of *Betula albosinensis* and *Betula platyphylla*: evidence for hybridization and a cryptic lineage. Ann. Bot. 123, 1179-1189. https://doi.org/10.1093/aob/mcz024.

Khan, G., Franco, F.F., Silva, G.A.R., Bombonato, J.R., Machado, M., Alonso, D.P., Ribolla, P.E.M., Albach, D.C., Moraes, E.M., 2020, Maintaining genetic integrity with high promiscuity: Frequent hybridization with low introgression in multiple hybrid zones of *Melocactus* (Cactaceae). Mol. Phylogenetics Evol. 142, 106642. https://doi.org/10.1016/j.ympev.2019.106642

Klingenberg, C.P., 2011. MorphoJ: an integrated software package for geometric morphometrics. Mol Ecol Res 11, 353-357. https://doi.org/doi:10.1111/j.1755-0998.2010.02924.x.

Lamichhaney, S., Han, F., Webster M.T., Andersson, L., 2018. Rapid hybrid speciation in Darwin's finches. Science, 359, 224-228. 10.1126/science.aao4593.

Leite, A.V., Machado, I.C., 2009. Biologia reprodutiva da" catingueira"(Caesalpinia pyramidalis Tul., Leguminosae-Caesalpinioideae), uma espécie endêmica da

Caatinga. Brazilian Journal of Botany. 32, 79-88. https://doi.org/10.1590/S0100-84042009000100008

Lewis, G.P., 1995. Systematic studies in neotropical '*Caesalpinia* L.' (Leguminosae: Caesalpinioideae), including a revision of the 'Poinchianella-Erythrostemon' group. Thesis. University of St Andrews.

Lewis, G.P., 2005. Tribe Caesalpinieae. In: Lewis G, Schrire B, Mackinder B, Lock M (Eds) Legumes of the World. Kew Royal Botanic Gardens, Richmond, pp 127–159.

Lewis, G.P., Schrire, B.D., 1995. A reappraisal of the Caesalpinia group Caesalpinioideae: Caesalpinieae) using phylogenetic analysis. In: Crisp, M.D., Doyle, J.J. (Eds), Advances in Legume Systematics: Part 7, Phylogeny. Kew Royal Botanic Gardens, Richmond, pp 41-52.

Lira, C.F., Cardoso, S.R.S., Ferreira, P.C.G., Cardoso, M.A., Provan, J., 2003. Long-term population isolation in the endangered tropical tree species Caesalpinia echinata Lam. revealed by chloroplast microsatellites. Mol Ecol 12, 3219-3225. https://doi.org/10.1046/j.1365-294X.2003.01991.x.

Ma, Y., Wang, J., Hu, Q., Li, J., Sun, Y., Zhang, L., Abbott, R.J., Liu, J., Mao, K., 2019. Ancient introgression drives adaptation to cooler and drier mountain habitats in a cypress species complex. Commun. Biol. 2, 213 https://doi.org/10.1038/s42003-019-0445-z.

Maddison, W. P., 1997. Gene trees in species trees. Systematic biology. 46, 523-536. https://doi.org/10.1093/sysbio/46.3.523

Mallet, J., 2005. Hybridization as an invasion of the genome. Trends Ecol. Evol. 20, 229-237. https://doi.org/10.1016/j.tree.2005.02.010.

Marques, A., Moraes, L., Santos, M.A., Costa, I., Costa, L., Nunes, T., Melo, N., Simon, M.F., Leitch, A.R., Almeida, C., Souza, G., 2018. Origin and parental genome characterization of the allotetraploid *Stylosanthes scabra* Vogel (Papilionoideae, Leguminosae), an important legume pasture crop. Ann. Bot. 122, 1143-1159. https://doi.org/10.1093/aob/mcy113.

Mayer, C., 2006-2010. Phobos 3.3.11. http://www.rub.de/ecoevo/cm/cm_phobos.htm.

Meyer, B.S., Matschiner, M., Salzburger, W., 2017. Disentangling Incomplete Lineage Sorting and Introgression to Refine Species-Tree Estimates for Lake Tanganyika Cichlid Fishes. Syst. Biol. 66, 531-550. https://doi.org/10.1093/sysbio/syw069.

Moro, M,F., Lughadha, E.N., Filer, D.L., Araújo, F.S., Martins, F.R., 2014. A catalogue of the vascular plants of the Caatinga Phytogeographical Domain: a synthesis of floristic

and phytosociological surveys. Phytotaxa. 160, 19. http://dx.doi.org/10.11646/phytotaxa.160.1.1.

Mota, M.R., Pinheiro, F., Leal, B.S.S., Wendt, T., Palma-Silva, C., 2019. The role of hybridization and introgression in maintaining species integrity and cohesion in naturally isolated inselberg bromeliad populations. Plant Biol. 21, 122-132. https://doi.org/10.1111/plb.12909.

Muniz, A.C., Lemos-Filho, J.P., Souza, H.A., Marinho, R.C., Buzatto, R.S., Heusrtz, M., Lovato, M.B., 2020. The protected tree *Dimorphandra wilsonii* (Fabaceae) is a population of inter-specific hybrids: recommendations for conservation in the Brazilian Cerrado/Atlantic Forest ecotone. Ann. Bot. 126, 191-203. https://doi.org/10.1093/aob/mcaa066.

Naciri, Y., Linder, H.P., 2020. The genetics of evolutionary radiations. Biol. Rev. 95, 1055-1072. https://doi.org/10.1111/brv.12598.

Nores, M.J., Simpson, B.B., Hick, P., Anton, A.M., Fortunato, R.H., 2012. The phylogenetic relationships of four monospecific caesalpinioids (Leguminosae) endemic to southern South America. Taxon. 61, 790-802. https://doi.org/10.1002/tax.614006

Novais, J.S., Lima, L.C.L., Santos, F.A.R., 2010. Bee pollen loads and their use in indicating flowering in the Caatinga region of Brazil. J. Arid Environ. 74, 1355-1358. https://doi.org/10.1016/j.jaridenv.2010.05.005.

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics, 155, 945-959. PMID: 10835412.

Queiroz, L.P., Cardoso, D., Fernandes, M.F., Moro, M.F., 2017. Diversity and Evolution of Flowering Plants of the Caatinga Domain. In: Silva, J.M.C., Leal, I.R., Tabarelli, M. (Eds), Caatinga. Springer, Cham. https://doi.org/10.1007/978-3-319-68339-3_2.

Raymond, M., Rousset, F., 1995. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. J. Heredity. 86, 248-249. https://genepop.curtin.edu.au/.

Rito, K.F., Arroyo-Rodríguez, V., Queiroz, R.T., Leal, I.R., Tabarelli, M., 2017. Precipitation mediates the effect of human disturbance on the Brazilian Caatinga vegetation. J Ecol. 105, 828-838. https://doi.org/10.1111/1365-2745.12712

Robins, T.P., Binks, R.M., Byrne, M., Hopper, S.D., 2021. Landscape and taxon age are associated with differing patterns of hybridization in two Eucalyptus (Myrtaceae) subgenera. Ann. Bot. 127, 49-62. https://doi.org/10.1093/aob/mcaa164.

Rousset, F., 2008. Genepop'007: a complete reimplementation of the Genepop software for Windows and Linux. Mol. Ecol. Resources. 8, 103-106. https://genepop.curtin.edu.au/

Rohlf, F. J., 2010a. tpsUtil, version 1.46, Software. Department of Ecology and Evolution, Stony Brook University of New York.

Rohlf, F.J., 2010b. tps-DIG, Digitize Landmarks and Outlines, Version 2.16. Department of Ecology and Evolution. State University of New York, Stony Brook, New York. http://life.bio.sunysb.edu/morph.

Rohlf, F.J., Slice, D., 1990. Extensions of the Procrustes Method for the Optimal Superimposition of Landmarks. Syst. Biol. 39, 40-59. https://doi.org/10.2307/2992207

Santos, H.G., Jacomine, P.K.T., Dos Anjos, L.H.C., De Oliveira, V.A., Lumbreras, J.F., Coelho, M.R., Almeida, J.A., Araujo-Filho, J.C., Oliveira, J.B., Cunha, T. J. F., 2018. Sistema brasileiro de classificação de solos, fifth ed. Embrapa, Brasília.

Schley, R. J., Pennington, R. T., Pérez-Escobar, O. A., Helmstetter, A. J., de la Estrella, M., Larridon, I., Kikuchi, B., Barraclough, T., Forest, F., Klitgård, B., 2020. Introgression across evolutionary scales suggests reticulation contributes to Amazonian tree diversity. Mol. Ecol., *29*, 4170-4185. https://doi.org/10.1111/mec.15616.Schuelke, M., 2000. An economic method for the fluorescent labeling of PCR fragments: A poor man's approach to genotyping for research and high-throughput diagnostics. Nat Biotechnol, 18, 233-234. https://doi.org/10.1038/72708

Silva, J.M.C., Barbosa, L.C.F., 2017. Impact of Human Activities on the Caatinga. In: Silva, J.M.C., Leal, I.R., Tabarelli, M. (Eds), Caatinga. Springer, Cham. https://doi.org/10.1007/978-3-319-68339-3_13.

Simpson, B.B., Miao, B.M., 1997. The circumscription of Hoffmannseggia (Fabaceae, Caesalpinioideae, Caesalpinieae) and its allies using morphological and cpDNA restriction site data. Plant Syst Evol. 205, 157-178. https://doi.org/10.1007/BF01464402

Soltis, P.S., Soltis, D.E., 2009. The Role of Hybridization in Plant Speciation. Annu. Rev. Plant Biol. 60, 561-588. https://doi.org/10.1146/annurev.arplant.043008.092039.

Souza, L.G.R., Crosa, O., Speranza, P., Guerra, M., 2012. Cytogenetic and molecular evidence suggest multiple origins and geographical parthenogenesis in Nothoscordum gracile (Alliaceae). Ann Bot. 109, 987-999. https://doi.org/10.1093/aob/mcs020.

Souza, G., Marques, A., Ribeiro, T., Dantas, L. G., Speranza, P., Guerra, M., Crosa, O., 2019. Allopolyploidy and extensive rDNA site variation underlie rapid karyotype evolution in *Nothoscordum* section *Nothoscordum* (Amaryllidaceae). Bot. J. Linn. Soc. 190, 215-228. https://doi.org/10.1093/botlinnean/boz008.

Stankowski, S., Ravinet, Mark., 2021. Defining the speciation continuum. **Evolution**. 75, 1256-1273. https://doi.org/10.1111/evo.14215.

Steane, D.A., Jones, R.C., Vaillancourt, R.E., 2005. A set of chloroplast microsatellite primers for Eucalyptus (Myrtaceae). Mol. Ecol. Notes. 5, 538-541. https://doi.org/10.1111/j.1471-8286.2005.00981.x.

Stefanović, S., Costea, M., 2008. Reticulate evolution in the parasitic genus *Cuscuta* (Convolvulaceae): over and over again. Bot. 86, 791-808. https://doi.org/10.1139/B08-033.

Stull, G.W., Soltis, P.S., Soltis D.E., Gitzendanner M.A., Smith S.A., 2020. Nuclear phylogenomic analyses of asterids conflict with plastome trees and support novel relationships among major lineages. Am. J. Bot. 107, 790-805. https://doi.org/10.1002/ajb2.1468.

Suarez-Gonzallez, A., Hefer, C.A., Christe, C., Corea, O., Lexer, C., Cronk, Q.C.B., Douglas, C.J., 2016. Genomic and functional approaches reveal a case of adaptive introgression from *Populus balsamifera* (balsam poplar) in *P. trichocarpa* (black cottonwood). Mol. Ecol. 25, 2427-2442. https://doi.org/10.1111/mec.13539.

The *Heliconius* genome consortium., 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature, 487, 94-98. https://doi.org/10.1038/nature11041.

Viscosi, V., Lepais, O., Gerber, S., Fortini, P., 2009. Leaf morphological analyses in four European oak species (*Quercus*) and their hybrids: A comparison of traditional and geometric morphometric methods. Pl Biosyst, 143, 564-574, https://doi.org/10.1080/11263500902723129.

Van-Lume, B., Esposito, T., Diniz-Filho, J.A.F., Gagnon, E., Lewis, G., Souza, G., 2017. Heterochromatic and cytomolecular diversification in the Caesalpinia group (Leguminosae): Relationships between phylogenetic and cytogeographical data. PPEES. 29, 51-63. https://doi.org/10.1016/j.ppees.2017.11.004.

Wang, G., Zhang, X., Herre, E.A., McKey, D., Machada, C.A., Yu, W.B., Cannon, C.H., Arnold, M.L., Pereira, R.A.S., Ming, R., Liu, Y.F., Wang, Y., Ma, D., Chen, J., 2021. Genomic evidence of prevalent hybridization throughout the evolutionary history of the fig-wasp pollination mutualism. Nat. Commun. 12, 718. https://doi.org/10.1038/s41467-021-20957-3.

Wang, X., He, Z., Shi, S., Wu, C., 2021. Genes and speciation: is it time to abandon the biological species concept?. **Natl. Sci. Rev.**, 7, 1387-1397. https://doi.org/10.1093/nsr/nwz220.

Weising, K., Gardner, R.C., 1999. A set of conserved PCR primers for the analysis of simple sequence repeat polymorphisms in chloroplast genomes of dicotyledonous angiosperms. Genome. 42, 9-19. https://doi.org/10.1139/g98-104.

Wu, C., Ting, C., 2004. Genes and speciation. **Nat. Rev. Gen.**, 5, 114-122, 2004. https://doi.org/10.1038/nrg1269.Zhou, Y., Li, W.W., Zhang, Y.Q., Zhang, J.Q., Ren, Y., 2020. Extensive reticulate evolution within Fargesia (s.l.) (Bambusoideae: Poaceae) and its allies: Evidence from multiple nuclear markers. Mol. Phylogenetics Evol. 149, 106842. https://doi.org/10.1016/j.ympev.2020.106842.

Table 1. Nuclear genetic diversity of *Cenostigma microphyllum* and *Cenostigma pyramidale* individuals, including sample size of pure and intermediate individual based on phenotype (Np), sample size of pure and hybrid individuals classified according to *q*-value between 0.10 - 0.90 (Nq), observed heterozygosity (Ho), expected heterozygosity (He) and the fixation index (F_{ST}) among pure species and hybrids assigned by the *q* – value

Species/Form	Np	Nq	Location	Coordinates		Voucher	Ho	He	Fst
Alcobaca									
C. microphyllum	0	0				-	-	-	
C. pyramidale	11	6	PARNA Catimbau - PE	08°31' 47,8" S	37°11'40,7" W	-	0.875	0.76894	0.01643
hybrid	6	11				UFP 88527 / UFP 88528	0.7500	0.75325	
Acude									
C. microphyllum	7	6	PARNA Catimbau - PE	08°27' 22,3" S	37°19'44,9" W	UFP 88534	0.6500	0.600	
C. pyramidale	0	7				UFP 88533	0.71429	0.76374	0.20276
hybrid	8	2				UFP 88529	0.625	0.6667	
Pedra			PARNA						
C. microphyllum	5	4	Catimbau - PE			UFP 88532	0.916667	0.72619	0.25698

C. pyramidale	0	0		08°34' 24,7" S	37°14'49,7" W	-			
hybrid	0	1				UFP 88530	1.000	1.000	
Brejo									
C. microphyllum	7	0	PARNA			-			
C. pyramidale	1	6	Catimbau - PE	08°31' 52" S	37°13'01,2" W	UFP 88535	0.83333	0.82197	0.04395
hybrid	8	10				UFP 88526	0.8000	0.82105	
APA1									
C. microphyllum	2	3	APA Serra			UFP 88540	0.58333	0.58333	
C. pyramidale	4	2	Branca - BA	09°55' 17" S	38°41,8'37" W	UFP 88538	0.7500	0.91667	0.11807
hybrid	7	8				UFP 88531	0.78125	0.81060	
APA2									
C. microphyllum	2	5	APA Serra			UFP 88537 / UFP 88539	0.66667	0.56667	
C. pyramidale	7	7	Branca - BA	09°53' 38" S	38°40'31" W	UFP 88536	0.82143	0.777747	0.17062
hybrid	4	1				-	1.000	1.000	

Belem									
C. microphyllum	0	0	Belém city - AL	9°32'57.9"S	36°31'23.7" W	-			
C. pyramidale	3	2				-	0.625	0.79167	-0.11688
hybrid	0	1				-	1.000	1.000	
Jeremoabo									
C. microphyllum	0	0	Jeremoabo city - BA	09°58'26.3" S	38°52'53.5" W	-			
C. pyramidale	3	3				-	0.58333	0.8333	-
hybrid	0	0				-			
P. Af. 1									
C. microphyllum	0	0	Paulo Afonso city - BA	9°27'23.1"S	38°12'05.3" W	-			
C. pyramidale	3	2				-	0.625	0.8333	-0.19481
hybrid	0	1				-	1.000	1.000	
P. Af. 2			Paulo Afonso						

C. microphyllum	3	2	city - BA			-	0.75000	0.70833	
C. pyramidale	0	0		9°25'05.7"S	38°11'49.6"W	-			0.16364
hybrid	0	1				-	1.000	1.000	
Overall	91	91					0.718739	0.783112	0.05579

Table 2. Nuclear and plastidial SSR loci used in this study for detect gene flow among *Cenostigma microphyllum* and *Cenostigma pyramidale* showing the annealing temperature (Ta) and the number of allele (N_A) of each locus

Genome	Locus	Ta (ºC)	Size (bp)	N_A	Study
Nuclear	CmSSR1	56	159 – 180	9	This study
	CmSSR2	53	158 – 201	19	This study
	CmSSR4	56	159 – 186	12	This study
	CmSSR6	53	123 – 149	14	This study
Plastidial	CmCPSSR4	60	365 – 366	2	Aecyo et al. 2021
	CmCPSSR8[a]	55	426	1	Aecyo et al. 2021
	CmCPSSR12	60	126 - 128	2	Aecyo et al. 2021
	CmCPSSR15	60	225 - 227	3	Aecyo et al. 2021
	ccmp2	55	280 - 354	6	Weising and Gardner 1999

[a]Monomorphic in the present sample, thus excluded from the population analysis

Table 3. Genetic diversity detected by four nuSSR for each locus. Number of genotyped individuals (N), observed heterozygosity (H_O), expected heterozygosity (H_E), inbreeding coefficient (F_{IS})

Locus	N	H_O	H_E	F_{IS}	*p* value
CmSSR1	91	0.527	0.687	0.2451	0.0011
CmSSR2	90	0.761	0.841	-0.0187	0.8348
CmSSR4	91	0.836	0.816	0.0326	0.1310
CmSSR6	90	0.763	0.817	0.0199	0.3490
Mean		0.722	0.790	0.0638	

Table 4. Characterization of the genetic diversity accessed by four nuSSR and four cpSSR loci in ten populations analysed in this study. Including sample size (N), number of alleles per site (N_A), number of private alleles (A_P) observed heterozygosity (H_O), expected heterozygosity (H_E), inbreeding coefficient (F_{IS})number of haplotypes per site (A), private haplotype (P), number of effective haplotype (Ne), haplotypic richness (Rh)

	nuSSR							*cpSSR*					
Population	N	A_P	N_A	H_O	H_E	F_{IS}	*p* value	N	A	P	Ne	Rh	He
Alcobaca	17	2	26	0.79412	0.76337	-0.05004	0.81916	16	1	0	1	0	0
Acude	15	2	31	0.65000	0.79885	0.02382	0.44770	8	1	0	1	0	0
Brejo	16	0	33	0.81250	0.83518	-0.32806	1.000	16	1	0	1	0	0
Pedra	5	0	14	0.70000	0.61111	0.01159	0.47214	5	1	0	1	0	0
APA1	13	2	30	0.73077	0.82736	0.14255	0.11144	5	3	0	2.778	1.4	0.8
APA2	13	3	28	0.75000	0.78692	-0.08717	0.88074	12	3	0	1.412	0.5	0.318
Belém	3	0	16	0.66667	0.80000	0.25581	0.33724	3	1	1	1	0	0
Jeremoabo	3	1	15	0.58333	0.83333	0.34884	0.14663	3	1	1	1	0	0
P. Af. 1	3	0	16	0.66667	0.81667	0.30435	0.30205	3	1	0	1	0	0
P. Af. 2	3	0	13	0.83333	0.75833	-0.21739	1.0000	3	1	0	1	0	0
Mean				0.718739	0.783112	0.06676	0.01075	7.4	1.4	0.2	1.219	0.19	0.112

Table 5. Results of analysis of molecular variance (AMOVA) for *Cenostigma microphyllum* and *Cenostigma pyramidale* for nuclear and plastidial SSR withdrawing hybrid individuals identified by the STRUCTURE analysis

		nuSSR		cpSSR	
		% variation	F_{ST}	% variation	F_{ST}
All populations and species	Among species	20.77	0.20773	2.85	0.0285
	Within species	79.23		97.15	
Cenostigma microphyllum	Among populations	4.75	0.04745	92.85	0.92851
	Within populations	95.25		7.15	
Cenostigma pyramidale	Among populations	3.96	0.03962	95.22	0.95218
	Within populations	96.04		4.78	

Table 6. Summary of correlation analyses between altitude and precipitation values and genotypic (*q* value) and leaf (centroid) characters, showing significance values (*p*), correlation coefficient (*r*) and the degrees of freedom (Df)

Variable	***q* value (Df = 67)**		**Centroid (Df = 106)**	
	p	***r***	***p***	***r***
Altitude	0.06	-0.22	0.04	0.2
Annual Precipitation	0.4	0.1	0.001	0.31
Precipitation of Wettest Month	0.59	0.07	0.001	0.33
Precipitation of Driest Month	0.46	0.09	0.01	0.24
Precipitation Seasonality (Coefficient of Variation)	0.69	-0.05	0.33	-0.1
Precipitation of Wettest Quarter	0.47	0.09	0.0001	0.35
Precipitation of Driest Quarter	0.3	0.13	0.01	0.24
Precipitation of Warmest Quarter	0.38	0.16	0.04	0.19
Precipitation of Coldest Quarter	0.24	0.14	0.003	0.28

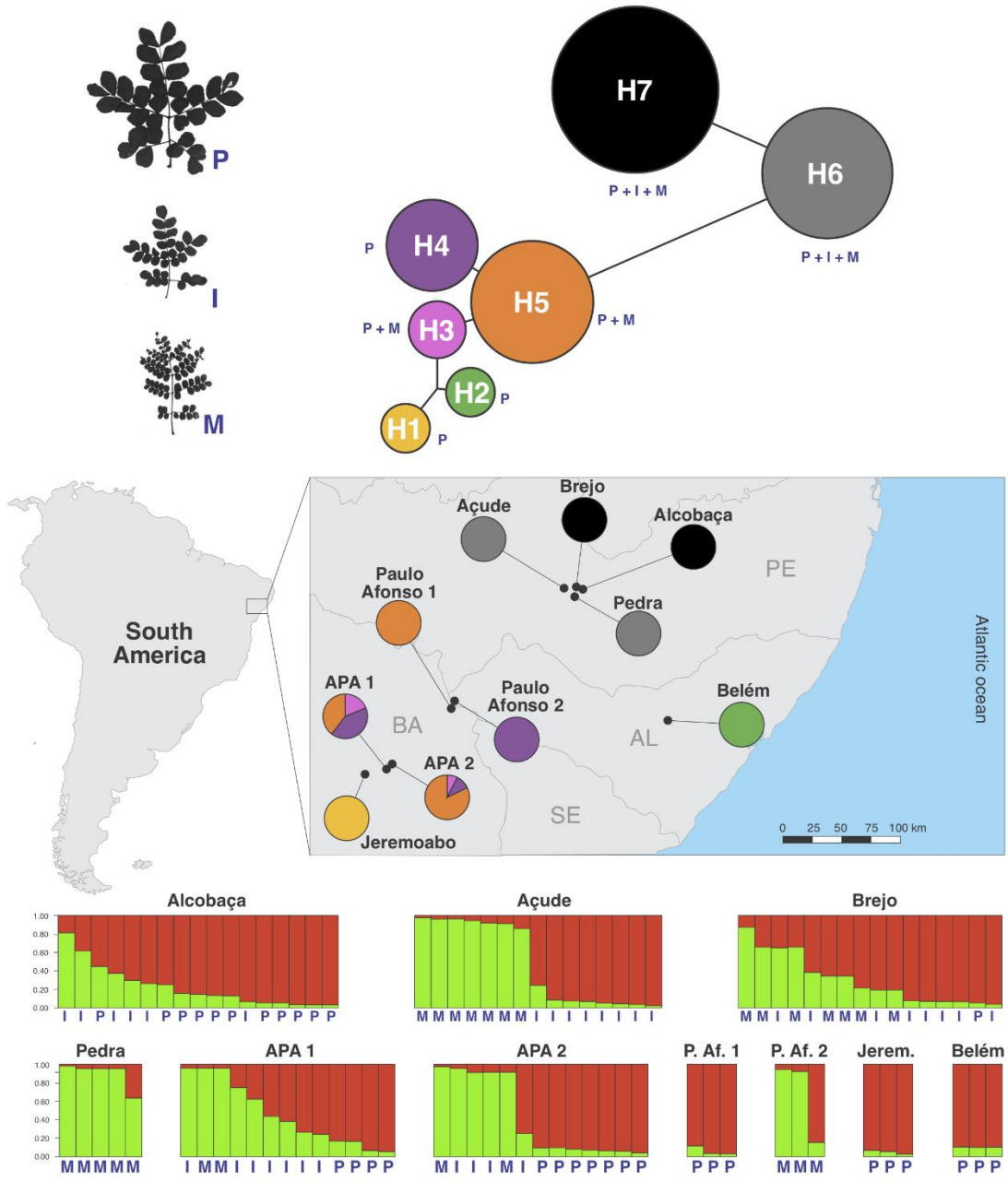

Fig 1. Genetic characterization and geographic distribution of the ten *Cenostigma microphyllum* (M) and *C. pyramidale* (P) individuals, as well as intermediate (I) morphotypes, sampled across the Caatinga domain. H1 to H7 represent the haplotypes obtained with six cpSSR markers. Each haplotype is represented by the same colour in the haplotype network and in the map, where the proportion of each haplotype per population is indicated. Below, genetic clusters (K = 2) assigned to each individual after genotyping with four nuSSR markers. Each bar represents one individual and indicates the proportions of the red (*Cenostigma microphyllum*) and the green cluster (*C. pyramidale*) in its genome

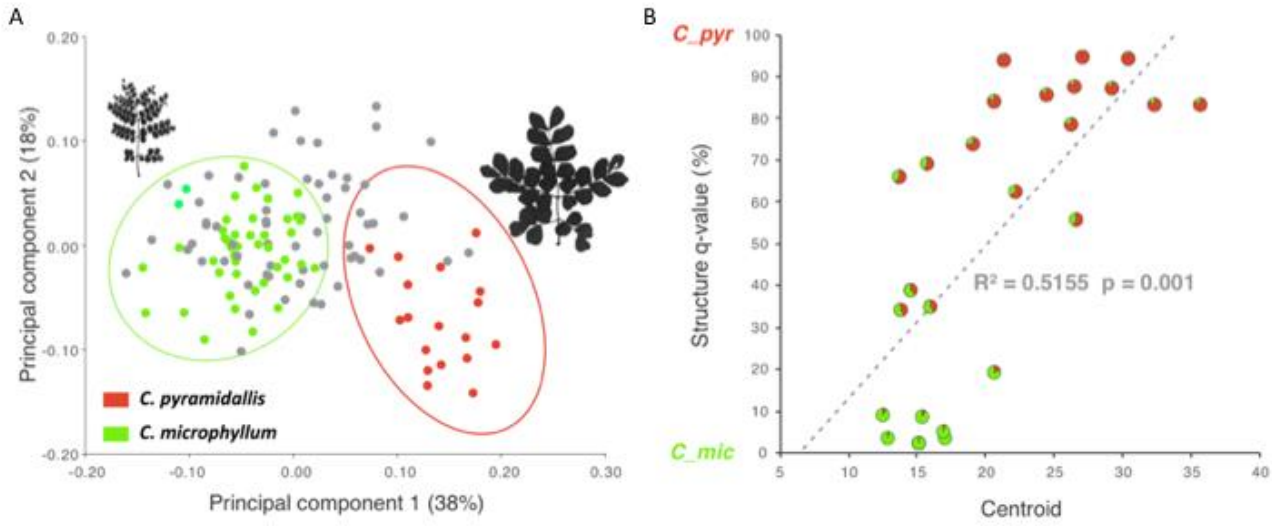

Fig 2. Morphometric analysis compared to the genetic identification of *Cenostigma microphyllum*, *C. pyramidale* and hybrid individuals. A) Principal Component Analysis (PCA) of foliar geometric morphometry, based on four landmarks and four semilandmarks; B) correlation between its centroid with the admixture proportion (*q*-value) of each hybrid individual obtained by the STRUCTURE analysis

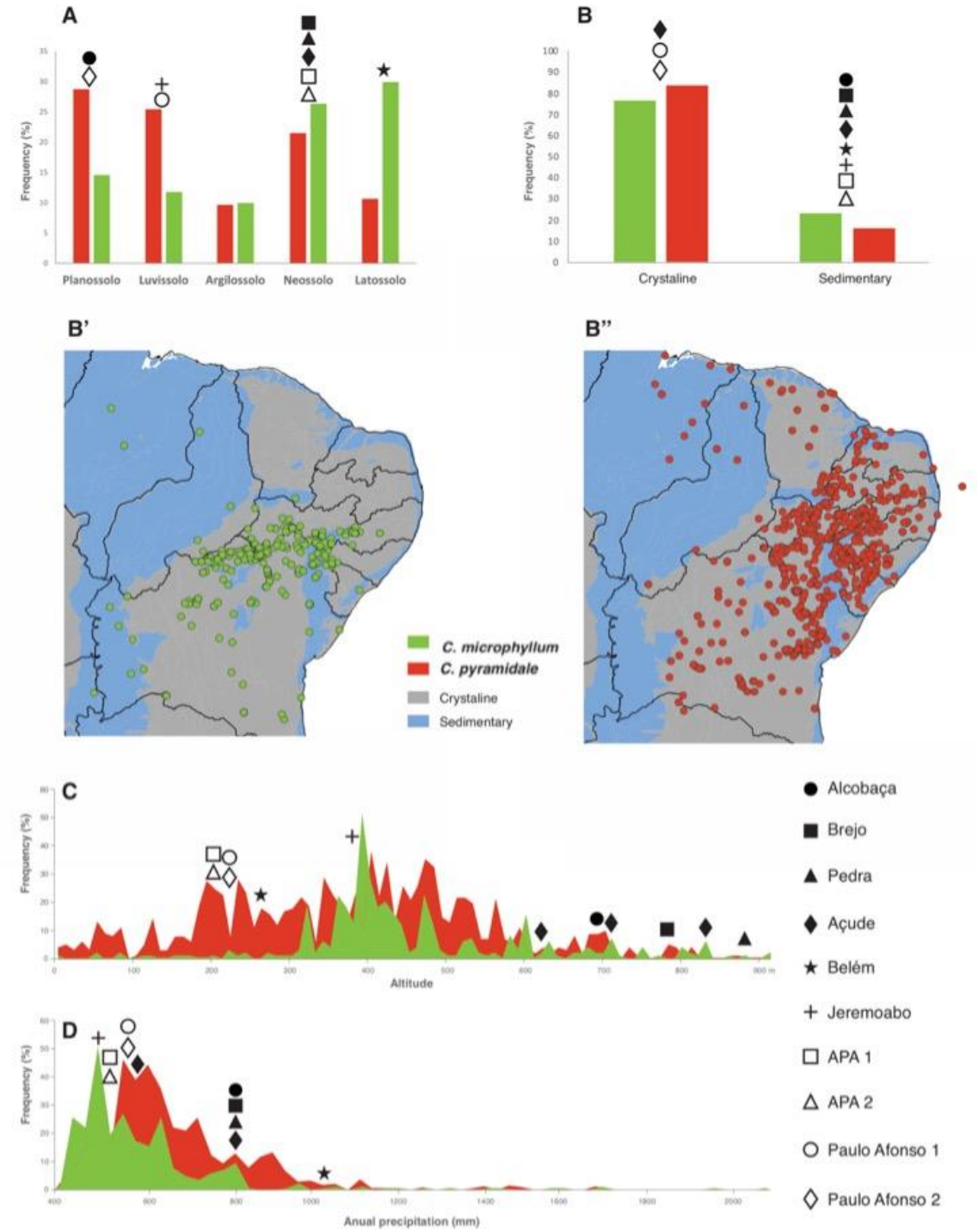

Fig 3. Features of *Cenostigma microphyllum* (green) and *C. pyramidale* (red) distribution throughout the Caatinga domain. (A) Type of soil in which each species occurs, based on SIBICs (Santos et al. 2018). (B) Distribution of *C. microphyllum* (B') and *C. pyramidale* (B") through the both geomorphologies in Northeastern Brazil. (C) Altitude range and (D) annual pluviosity range of these species through Caatinga. The same attributes were analysed for the ten populations sampled in the present study, which are indicated but different symbols in A-D

Table S1. Nuclear microsatellites primers developed to access the genetic diversity of *Cenostigma microphyllum*, including melting temperature (Tm), annealing temperature (Ta) and number of alleles (Na). In bold are the loci used to access the genetic diversity in this study. CmSSR10 turn out to be monomorphic and therefore not included among the nuSSR markers.

Locus	5' – Primer sequence – 3'	Tm (°C)	Ta (ºC)	Expected size (bp)	Observed size (bp)	Na	Motif
***CmSSR1F**	TCAGATATTTCTTAAAGGGCGAAC	60	65	153	153 - 177	6	$(AGG)_{10}$
CmSSR1R	GAGGAGAGTACAGCGATCAGAATC	60,8					
***CmSSR2F**	CAACCTAAAAAGCAGGTTCCTAAC	59,6	62	197	165 - 199	6	$(AT)_{17}$
CmSSR2R	CTGTGTAGTGTTTGCTCTCAGCTT	60,2					
*CmSSR3F	GTTCATTTTATGGGAATCAAGGAC	60	50	176	313 - 350	3	$(AAT)_{16}$
CmSSR3R	GGGCTTCTATTCATAGATTCTTGC	59,7					
***CmSSR4F**	CATTTCTTGTCATGCGTGTTCT	60,2	60	155	165 - 173	4	$(AT)_{12}$
CmSSR4R	GGGCTAAATCCTAAAATCTAACCAG	59,8					
*CmSSR5F	ATCATGGTGAAAGGAATCCAAC	60,1	60	123	135 - 147	9	$(AT)_{10}$
CmSSR5R	TCTACAATCACTCAAGGAGAAGTG	57,7					
***CmSSR6F**	CTCAACCCACAATGGAAGTAAAGT	60,6	60	120	125 - 147	8	$(AT)_{10}$
CmSSR6R	AAGAGGTATTGACTCAAGTGCGTA	59,4					
*CmSSR7F	ATGCTTATCCCAGACTATTCAACG	60,7	55	211	223 - 226	3	$(AAT)_{6}$
CmSSR7R	CCAATCTAACATCAAAACCATGAA	60,6					
*CmSSR8F	TTCCCAAAAGATTTGTGTGTGG	62	55	137	128 - 151	5	$(AAAAT)_{8}$

CmSSR8R	CCCTCGTGTCTCACTCTCTTAAA	59,9					
*CmSSR9F	AGAGAATAAATCGGTCAGCATCTC	60,1	50	112	128, 132	2	$(ATTT)_4$
CmSSR9R	AAGCCAATTCCAGTACACATGAC	60,3					
*CmSSR10F	TCATTTTAGTCCCTCCTCCCTACT	60,7	50	128	237	1	$(AG)_9$
CmSSR10R	AACCAAGAGAAATCCAAGTTGAAG	60					

*A M13 tail sequence (5’ – TGTAAAACGACGGCCAGT – 3’) was added was described by Schuelke et al. (2000)

Table S2. Cross-amplification of ten nuSSR primer pairs developed in this study for *Cenostigma microphyllum* to five species from the Caesalpinia group. "-" symbol means that the loci did not cross-amplify. The optimal annealing temperature of each loci for each species is indicated

Locus	*Cenostigma pyramidale*	*Caesalpinia pulcherrima*	*Guillandina bonduc*	*Libidibia ferrea*	*Paubrasilia echinata*
CmSSR1	56ºC	56ºC	56ºC	56ºC	56ºC
CmSSR2	53ºC	-	-	56ºC	-
CmSSR3	56ºC	56ºC	56ºC	56ºC	-
CmSSR4	56ºC	56ºC	-	56ºC	-
CmSSR5	53ºC	-	-	-	-
CmSSR6	56ºC	-	-	56ºC	-
CmSSR7	56ºC	-	-	-	-
CmSSR8	53ºC	-	-	-	-
CmSSR9	56ºC	-	-	-	-

Table S3. Haplotypes detected by four cpSSR in the ten sympatric and allopatric populations of *Cenostigma microphyllum* and *Cenostigma pyramidale*

Haplotype Code	Haplotype	N	Population	Specie
H1	280 365 126 225	3	Jeremoabo	*Cenostigma pyramidale*
H2	282 366 126 228	3	Belém	*Cenostigma pyramidale*
H3	287 366 126 227	2	APA 1, APA 2	*Cenostigma microphyllum*
H4	288 365 126 227	6	APA 1, APA 2, PA2	*Cenostigma microphyllum*
H5	288 366 126 227	15	APA 1, APA 2, PA1	Both species
H6	338 365 128 228	13	Açude, Pedra	Both species
H7	354 365 128 228	32	Alcobaça, Brejo	*Cenostigma pyramidale*

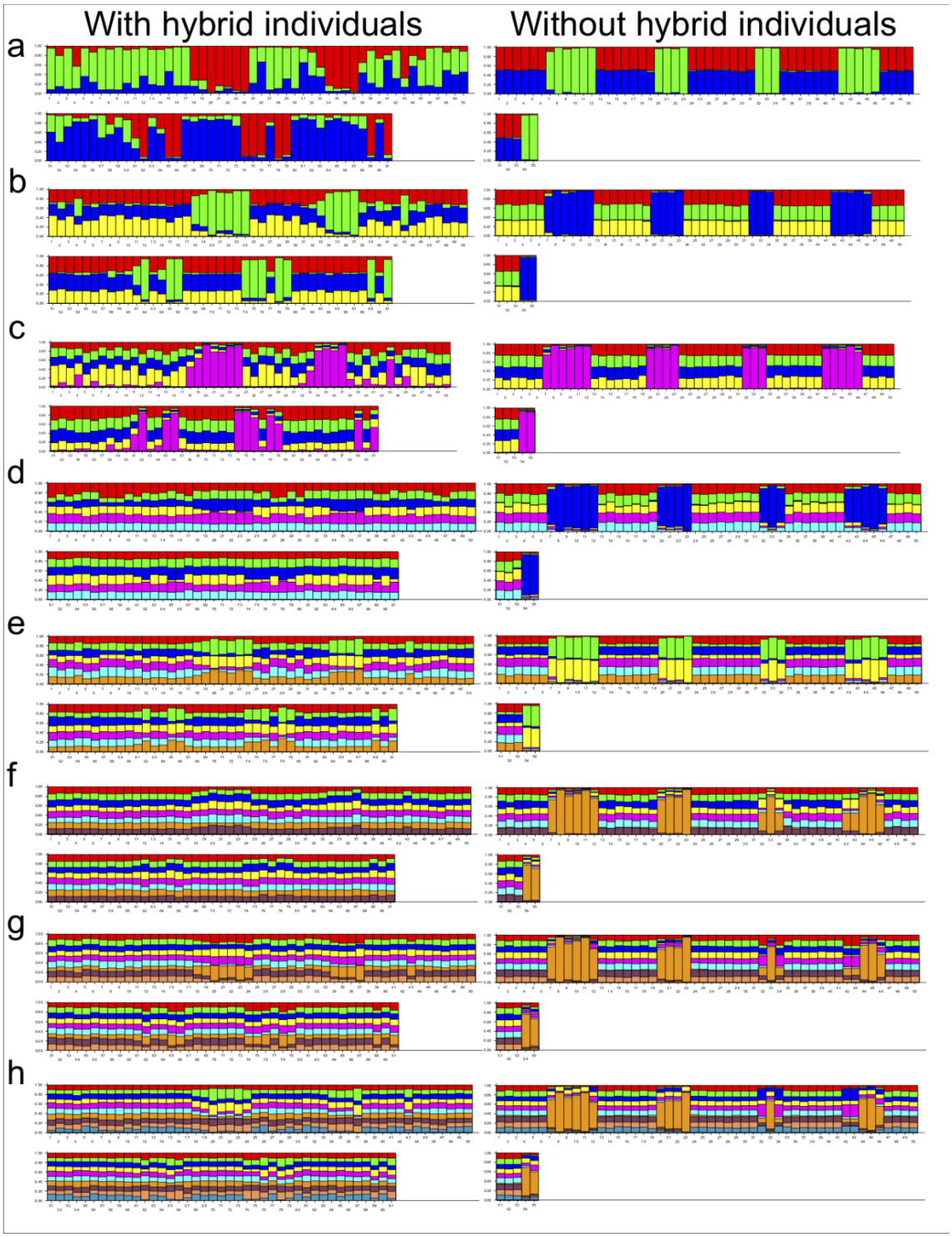

Figure S1. Admixture analysis of *Cenostigma microphyllum* and *C. pyramidale* populations accessed by STRUCTURE with number of clusters (*K*) varying from (a) *K* = 3 to (h) *K* = 10 including or withdrawing the hybrid individuals

4 CONSIDERAÇÕES FINAIS

O grupo Caesalpina apresenta uma história evolutiva complexa necessitando a utilização de diferentes abordagens para completa elucidação das relações evolutivas entre e dentro dos diferentes clados. Esta dissertação permitiu um avanço na compreensão das relações filogenéticas intergenéricas do grupo Caesalpinia a partir de abordagens filogenômicas. Além disso, este trabalho montou pela primeira vez o genoma plastidial de nove espécies, acrescentando novas informações genômicas para nove gêneros, contribuindo com 13 novos plastomas disponibilizados no banco de dados, incluindo o plastoma do Pau-Brasil (*Paubrasilia echinata*). Com o estudo macroevolutivo, foi possível compreender como se deu a evolução da macrosintenia do genoma plastidial no grupo Caesalpinia

O presente trabalho também contribui para a sistemática do gênero *Cenostigma* e para conhecimento do processo de especiação com fluxo gênico a partir da detecção e estudo das zonas híbridas entre *C. microphyllum* e *C. pyrdamidale*. Este estudo permitiu a identificação de indivíduos híbridos, possivelmente férteis, entre as duas espécies arbóreas que são amplamente distribuídas na Caatinga. Apesar disso, possivelmente o fluxo gênico interespecífico deve ser simétrico e não tem afetado a integridade genética das espécies.

REFERÊNCIAS

ABBOTT, Richard J. Plant speciation across environmental gradients and the occurrence and nature of hybrid zones. **Journal of Systematics and Evolution**, v. 55, n. 4, p. 238-258, 2017.

ALLEN, James O.; FAURON, Christiane M.; MINX, Patrick; ROARK, Leah; ODDIRAJU, Swetha; LIN, Guan Ning; MEYER, Louis; SUN, Hui; KIM, Kyung; WANG, Chunyan; DU, Feiyu; XU, Dong; GIBSON, Michael; CIFRESE, Jill; CLIFTON, Sandra W.; NEWTON, Kathleen J. Comparisons among two fertile and three male-sterile mitochondrial genomes of maize. **Genetics**, v. 177, n. 2, p. 1173-1192, 2007.

APG IV. 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, 2016.

BELLOT, Sidonie; RENNER, Susanne S. The plastomes of two species in the endoparasite genus *Pilostyles* (Apodanthaceae) each retain just five or six possibly functional genes. **Genome Biology and Evolution**, v. 8, n. 1, p. 189-201, 2015.

BLAZIER, John C.; JANSEN, Robert K.; MOWER, Jeffrey P.; GOVINDU, Madhu; ZHANG, Jin; WENG, Mao-Lun; RUHLMAN, Tracey A. Variable presence of the inverted repeat and plastome stability in *Erodium*. **Annals of Botany**, v. 117, n. 7, p. 1209-1220, 2016.

BOURQUE, Guillaume; BURNS, Kathleen H.; GEHRING, Mary; GORBUNOVA, Vera; SELUANOV, Andrei; HAMMELL, Molly; IMBEAULT, Michaël; IZSVÁC, Zsuzsanna; LEVIN, Henry L.; MACFARLAN, Todd S.; MAGER, Dixie L.; FESCHOTTE, Cédric. Ten things you should know about transposable elements. **Genome Biology**, v. 19, n. 1, p. 1-12, 2018.

BRAJKOVIĆ, Josip; PEZER, Zeljka; BRUVO-MADARIĆ, Branka; SERMEK, Antonio; FELICIELLO, Isidoro; UGARKOVIĆ, Đurdica. Dispersion profiles and gene associations of repetitive DNAs in the euchromatin of the beetle *Tribolium castaneum*. **G3: Genes, Genomes, Genetics**, v. 8, n. 3, p. 875-886, 2018.

BRAUKMANN, Thomas; KUZMINA, Maria; STEFANOVIĆ, Saša. Plastid genome evolution across the genus *Cuscuta* (Convolvulaceae): two clades within subgenus *Grammica* exhibit extensive gene loss. **Journal of Experimental Botany**, v. 64, n. 4, p. 977-989, 2013.

CASTILLO-MENDONZA, E., SALINAS-SÁNCHEZ, D., VALENCIA-CUEVAS, L., ZAMILPA, A., TOVAR-SÁNCHEZ, E., 2019. Natural hybridisation among *Quercus glabrescens*, *Q. rugosa* and *Q. obtusata* (Fagaceae): Microsatellites and secondary metabolites markers. **Plant Biology**, v. 21, n. 1, p. 110-121, 2019.

CAUZ-SANTOS, Luiz Augusto; DA COSTA, Zirlane Portugal; CALLOT, Caroline; CAUET, Stéphane; ZUCCHI, Maria Imaculada; BERGÈS, Hélène; VAN DEN BERG; Cássio; VIEIRA, Maria Lúcia Carneiro. A repertory of rearrangements and the loss of an inverted repeat region in *Passiflora* chloroplast genomes. **Genome Biology and Evolution**, v. 12, n. 10, p. 1841-1857, 2020.

CHAVES, T.; FERNANDES, F.; SANTANA, C.; SANTOS, J.; MEDEIROS, F.; FELISMINO, D.; SANTOS, V.; CATÃO, R.; COUTINHO, H. Evaluation of the Interaction between the *Poincianella pyramidalis* (Tul.) LP Queiroz Extract and Antimicrobials Using Biological and Analytical Models. **Plos One**, v. 11, n. 5, p.1-23, 2016.

CHOI, In-Su; JANSEN, Robert; RUHLMAN, Tracey. Caught in the Act: Variation in plastid genome inverted repeat expansion within and between populations of *Medicago minima*. **Ecology and Evolution**, v. 10, n. 21, p. 12129-12137, 2020.

CHUMLEY, Timothy W.; PALMER, Jeffrey D.; MOWER, Jeffrey P.; FOURCADE, H. Matthew; CALIE, Patrick J.; BOORE, Jeffrey L.; JANSEN, Robert K. The complete chloroplast genome sequence of *Pelargonium* × *hortorum*: organization and evolution of the largest and most highly rearranged chloroplast genome of land plants. **Molecular Biology and Evolution**, v. 23, n. 11, p. 2175-2190, 2006.

DE QUEIROZ, K. Ernst Mayr and the modern concept of species. **Proceedings Of The National Academy Of Sciences**, v. 102, n. 1, p.6600-6607, 2005.

DE QUEIROZ, K. Species Concepts and Species Delimitation. **Systematic Biology**, v. 56, n. 6, p.879-886, 2007.

DODSWORTH, Steven; POKORNY, Lisa; JOHNSON, Matthew G.; KIM, Jan T.; MAURIN, Olivier; WICKETT, Norman J.; FOREST, Felix; BAKER, William J. Hyb-Seq for flowering plant systematics. **Trends in Plant Science**, v. 24, n. 10, p. 887-891, 2019.

DOGAN, Mert; POUCH, Milan; MANDÁKOVÁ, Terezie; HLOUŠKOVÁ, Petra; GUO, Xinyi; WINTER, Pieter; CHUMOVÁ, Zuzana; VAN NIEKERK, Adriann; MUMMENHOFF, Klaus; AL-SHEHBAZ, Ihsan A.; MUCINA, Ladislav; LYSAK, Martin A. Evolution of tandem repeats is mirroring post-polyploid cladogenesis in *Heliophila* (Brassicaceae). **Frontiers in Plant Science**, v. 11, n. 607893, p. 1-18, 2021.

DUGAS, Diana V.; HERNANDEZ, David; KOENEN, Erik J.M.; SCHWARZ, Erika; STRAUB, Shannon; HUGHES, Colin E.; JANSEN, Robert K.; NAGESWARA-RAO, Madhugiri; STAATS, Martijn; TRUJILLO, Joshua T.; HAJRAH, Nahid H.; ALHARBI, Njud S.; AL-MALKI, Abdulrahman L.; SABIR, Jamal S. M.; BAILEY, Donovan. Mimosoid legume plastome evolution: IR expansion, tandem repeat expansions and accelerated rate of evolution in clpP. **Scientific Reports**, v. 5, n. 1, p. 1-13, 2015.

ELLEGREN, Hans. Microsatellite mutations in the germline: implications for evolutionary inference. **Trends in genetics**, v. 16, n. 12, p. 551-558, 2000.

FLEISCHMANN, Andreas; MICHAEL, Todd P.; RIVADAVIA, Fernando; SOUSA, Aretuza; WANG, Wenqin; TEMSCH, Eva M.; GREILHUBER, Johann; Müller, Kai F.; HEUBL, Günther. Evolution of genome size and chromosome number in the carnivorous plant genus *Genlisea* (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms. **Annals of Botany**, v. 114, n. 8, p. 1651-1663, 2014.

FLORA DO BRASIL 2020. Jardim Botânico do Rio de Janeiro. Disponível em: http://floradobrasil.jbrj.gov.br. Acesso em: 16 Maio 2020

FU, Peng-Cheng; SUN, Shan-Shan; KHAN, Gulzar; DONG, Xiao-Xia; TAN, Jin-Zhou; FAVRE, Adrien; ZHANG, Fa-Qi; CHEN, Shi-Long. Population subdivision and hybridization in a species complex of *Gentiana* in the Qinghai-Tibetan Plateau. **Annals of Botany**, v. 125, n. 4, p. 677-690, 2020.

GAGNON, Edeline; BRUNEAU, Anne; HUGHES, Colin E.; QUEIROZ, Luciano Paganucci; LEWIS, Gwylin P. A new generic system for the pantropical Caesalpinia group (Leguminosae). **Phytokeys**, v. 71, p.1-160, 2016.

GAGNON, Edeline; LEWIS, Gwylin P.; SOTUYO, J. Solange, HUGHES, Colin E.; BRUNEAU, Anne. A molecular phylogeny of *Caesalpinia* sensu lato: Increased sampling reveals new insights and more genera than expected. **South African Journal of Botany**, v. 89, p. 111-127, 2013.

GAGNON, Edeline; RINGELBERG, Jens J.; BRUNEAU, Anne; LEWIS, Gwylin P.; HUGHES, Colin E. Global succulent biome phylogenetic conservatism across the pantropical Caesalpinia Group (Leguminosae). **New Phytologist**, v. 222, n. 4, p. 1994-2008, 2019.

GAMARRA, Victoria Eugênia de Santa Cruz. **Me Gritaron Negra!**, s.l., 1978. Disponível em: https://www.youtube.com/watch?v=cHr8DTNRZdg. Acesso em 27 de junho de 2021.

GIVNISH, T.; MILLAM, K.; BERRY, P.; SYTSMA, K. Phylogeny, adaptive radiation, and historical biogeography of Bromeliaceae inferred from *ndh*F sequence data. **Aliso: A Journal of Systematic and Evolutionary Botany**, v. 23, n. 1, p. 3-26, 2007.

GONÇALVES, Deise J.P.; JANSEN, Robert K.; RUHLMAN, Tracey A.; MANDEL, Jennifer R. Under the rug: abandoning persistent misconceptions that obfuscate organelle evolution. **Molecular Phylogenetics and Evolution**, v. 151, n. 106903, p. 1-4, 2020.

GONÇALVES, Deise J.P.; SIMPSON, Beryl B.; ORTIZ, Edgardo M.; SHIMIZU, Gustavo H.; JANSEN, Robert K. Incongruence between gene trees and species trees and phylogenetic signal variation in plastid genes. **Molecular Phylogenetics and Evolution**, v. 138, [s.n.], p. 219-232, 2019.

GUALBERTO, José M.; NEWTON, Kathleen J. Plant mitochondrial genomes: dynamics and mechanisms of mutation. **Annual Review of Plant Biology**, v. 68, [s.n.], p. 225-252, 2017.

HE, Jian; YAO, Min; LYU, Ru-Dan; LIN, Le-Le; LIU, Hui-Jie; PEI, Lin-Ying; YANG, Shuang-Xi; XIE, Lei; CHENG, Jin. Structural variation of the complete chloroplast genome and plastid phylogenomics of the genus *Asteropyrum* (Ranunculaceae). **Scientific Reports**, v. 9, n. 1, p. 1-13, 2019.

HESLOP-HARRISON, J. S.; SCHWARZACHER, Trude. Organisation of the plant genome in chromosomes. **The Plant Journal**, v. 66, n. 1, p. 18-33, 2011.

HOPKINS, Robin. Reinforcement in plants. **New Phytologist**, v. 197, n. 4, p. 1095-1103, 2013.

IBIAPINO, Amália; GARCIA, Miguel A.; FERRAZ, Maria Eduarda; COSTEA, Mihai; STEFANOVIĆ, Sasa; GUERRA, Marcelo. Allopolyploid origin and genome differentiation of the parasitic species *Cuscuta veatchii* (Convolvulaceae) revealed by genomic in situ hybridization. **Genome**, v. 62, n. 7, p.467-475, 2019.

JANSEN, Robert K.; RUHLMAN, Tracey A. Plastid genomes of seed plants. In: **Genomics of chloroplasts and mitochondria**. Springer, Dordrecht, 2012. p. 103-126.

JIN, Dong-Min; WICKE, Susann; GAN, Lu; YANG, Jun-Bo; JIN, Jian-Jun; YI, Ting-Shuang. The Loss of the Inverted Repeat in the Putranjivoid Clade of Malpighiales. **Frontiers in Plant Science**, v. 11, [s.n.], p. 942, 2020.

KADLEC, Malvina; BELLSTEDT, Dirk U.; LE MAITRE, Nicholas C.; PIRIE, Michael D. Targeted NGS for species level phylogenomics: “made to measure” or “one size fits all”? **PeerJ**, v. 5, p. e3569, 2017.

KHAN, Gulzar; FRANCO, F ernando F.; SILVA, Gislaine A.R.; BOMBONATO, Juliana R.; MACHADO, Marlon; ALONSO, Diego P.; RIBOLLA, Paulo E.M.; ALBACH, Dirk C.; MORAES, Evandro M. Maintaining genetic integrity with high promiscuity: Frequent hybridization with low introgression in multiple hybrid zones of *Melocactus* (Cactaceae). **Molecular Phylogenetics and Evolution**, v. 142, p. 106642, 2020.

KOENEN, Erik J.M.; OJEDA, Dario I.; BAKKER, Freek T.; WIERINGA, Jan J.; KIDNER, Catherine; HARDY, Oliver J.; PENNINGTON, R. Toby; BRUNEAU, Anne; HUGHES, Colin E. The origin of the legumes is a complex paleopolyploid phylogenomic tangle closely associated with the cretaceous–paleogene (K–Pg) mass extinction event. **Systematic Biology**, v. 70, n. 3, p. 508-526, 2020a.

KOENEN, Erik J.M.; OJEDA, Dario I.; STEEVES, Royce; MIGLIORE, Jérémy; BAKKER, Freek T.; WIERINGA, Jan J.; KIDNER, Catherine; HARDY, Oliver J.; PENNINGTON, R. Toby; BRUNEAU, Anne; HUGHES, Colin E. Large-scale genomic sequence data resolve the deepest divergences in the legume phylogeny and support a near-simultaneous evolutionary origin of all six subfamilies. **New Phytologist**, v. 225, n. 3, p. 1355-1369, 2020b.

LEITE, Ana Virgínia; MACHADO, Isabel Cristina. Biologia reprodutiva da "catingueira" (*Caesalpinia pyramidalis* Tul., Leguminosae-Caesalpinioideae), uma espécie endêmica da Caatinga. **Brazilian Journal of Botany**, v. 32, n. 1, p. 79-88, 2009.

LEWIS, G.P.; SCHRIRE, B.D. A reappraisal of the Caesalpinia group Caesalpinioideae: Caesalpinieae) using phylogenetic analysis. In: CRISP, M. D.; DOYLE, J. J. (Eds) **Advances in Legume Systematics**: Part 7, Phylogeny. Kew Royal Botanic Gardens, Richmond, 1995. p. 41–52.

LEWIS, Gwilym P. Tribe Caesalpinia. In: LEWIS, Gwilyn P.; SCHRIRE, B. D.; MACKINDER, B.; LOCK, M. **Legumes of the world**. Kew Royal Botanical Garden: Richmond, 2005. p. 127-159.

LEWIS, Gwilym Peter et al. **Caesalpinia: a revision of the Poincianella-Erythrostemon group**. Royal Botanic Gardens (K-RBG), 1998.

LEWIS, Gwilym Peter. **Systematic studies in neotropical 'Caesalpina L.'(Leguminosae: Caesalpinioideae), including a revision of the 'Poinchianella-Erythrostemon'group**. 1995. Tese de Doutorado. University of St Andrews.

LPWG. A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny – The Legume Phylogeny Working Group (LPWG). **Taxon**, v. 66, n. 1, p.44-77, 2017.

MA, Yazhen; WANG, Ji; HU, Quanjun; LI, Jialiang; SUN, Yongshuai; ZHANG, Lei; ABBOTT, Richard J.; LIU, Jianquan; MAO, Kangshan. Ancient introgression drives adaptation to cooler and drier mountain habitats in a cypress species complex. **Communications Biology**, v. 2, n. 1, p. 1-12, 2019.

MADDISON, Wayne P.; KNOWLES, L. Lacey. Inferring phylogeny despite incomplete lineage sorting. **Systematic Biology**, v. 55, n. 1, p. 21-30, 2006.

MALLET, James. Hybridization as an invasion of the genome. **Trends in Ecology & Evolution**, v. 20, n. 5, p. 229-237, 2005.

MARQUES, André; RIBEIRO, Tiago; NEUMANN, Pavel; MACAS, Jiří; PETR, Novác; SCHUBERT, Veit; PELINO, Marco; FUCHS, Jörg; MA, Wei; KUHLMANN, Markus; BRANDT, Ronny; VANZELA, André L.L.; BESEDA, Tomáš; ŠIMKOVÁ, Hana; PEDROSA-HARAND, Andrea; HOUBEN, Andreas. Holocentromeres in *Rhynchospora* are associated with genome-wide centromere-specific repeat arrays interspersed among euchromatin. **Proceedings of the National Academy of Sciences**, v. 112, n. 44, p. 13633-13638, 2015.

MARTINS, Gleica; BALBINO, Eliane., MARQUES, André; ALMEIDA, Cícero. Complete mitochondrial genomes of the *Spondias tuberosa* Arr. Cam and *Spondias mombin* L. reveal highly repetitive DNA sequences. **Gene**, v. 720, p. 144026, 2019.

MATA-SUCRE, Yennifer; COSTA, Lucas; GAGNON, Edeline; LEWIS, Gwilym P.; LEITCH, Ilia J.; SOUZA, Gustavo. Revisiting the cytomolecular evolution of the Caesalpinia group (Leguminosae): a broad sampling reveals new correlations between cytogenetic and environmental variables. **Plant Systematics and Evolution**, v. 306, n. 2, p. 1-13, 2020b.

MATA-SUCRE, Yennifer; SADER, Mariela; VAN-LUME, Brena; GAGNON, Edeline; PEDROSA-HARAND, Andrea; LEITCH, Ilia J.; LEWIS Gwylym P.; SOUZA, Gustavo. How diverse is heterochromatin in the Caesalpinia group? Cytogenomic characterization of *Erythrostemon hughesii* Gagnon & GP Lewis (Leguminosae: Caesalpinioideae). **Planta**, v. 252, n. 4, p. 1-14, 2020a.

MAYR, Ernst. **Systematics and the Origin of Species**. New York: Columbia Univ. Press, 1942.

MEIER, Joana I.; MARQUES, David A.; MWAIKO, Salome; WAGNER, Catherine E.; EXCOFFIER, Laurent; SEEHAUSEN, Ole. Ancient hybridization fuels rapid cichlid fish adaptive radiations. **Nature Communications**, v. 8, n. 1, p. 1-11, 2017.

MOTA, M.; PINHEIRO, F.; LEAL, B.; WENDT, T; PALMA-SILVA, C. The role of hybridization and introgression in maintaining species integrity and cohesion in naturally isolated inselberg bromeliad populations. **Plant Biology**, v. 21, n. 1, p.122-132, 9 out. 2018.

MUNIZ, André Carneiro; LEMOS-FILHO, José Pires; SOUZA, Helena Augusta; MARINHO, Rafaela Cabral; BUZATTI, Renata Santiago; HEUSRTZ, Myriam; LOVATO, Maria Bernadete. The protected tree *Dimorphandra wilsonii* (Fabaceae) is a population of inter-specific hybrids: recommendations for conservation in the Brazilian Cerrado/Atlantic Forest ecotone. **Annals of Botany**, v. 126, n. 1, p. 191-203, 2020.

NACIRI, Yamama; LINDER, H. Peter. The genetics of evolutionary radiations. **Biological Reviews**, v. 95, n. 4, p. 1055-1072, 2020.

NORES, María J.; SIMPSON, Beryl B.; HICK, Pascale; ANTON, Ana M.; FORTUNATO, René H. The phylogenetic relationships of four monospecific caesalpinioids (Leguminosae) endemic to southern South America. **Taxon**, v. 61, n. 4, p. 790-802, 2012.

NOVAIS, J. S.; LIMA, L.; SANTOS, F. Bee pollen loads and their use in indicating flowering in the Caatinga region of Brazil. **Journal Of Arid Environments**, v. 74, n. 10, p.1355-1358, 2010.

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

PATERNO, G.; SIQUEIRA-FILHO, J.; GANADE, G. Species-specific facilitation, ontogenetic shifts and consequences for plant community succession. **Journal Of Vegetation Science**, v. 27, n. 3, p.606-615, 2016.

PELLICER, Jaume; FAY, Michael F.; LEITCH, Ilia J. The largest eukaryotic genome of them all?. **Botanical Journal of the Linnean Society**, v. 164, n. 1, p. 10-15, 2010.

PIRIE, M. D.; OLIVER, E. G. H.; DE KUPPLER, A. M.; GEHRKE, B.; LE MAITRE, N. C.; KANDZIORA, M.; BELLSTEDT, D. U. The biodiversity hotspot as evolutionary hot-

bed: spectacular radiation of *Erica* in the Cape Floristic Region. **BMC evolutionary biology**, v. 16, n. 1, p. 1-11, 2016.

RABAH, Samar O.; SHRESTHA, Bikash; HAJRAH, Nahid H.; SABIR, Mumdooh J.; ALHARBY, Hesham F.; SABIR, Mernan J.; ALHEBSHI, Alawiah M.; SABIR, Jamal S.M.; GILBERT, Lawrence E.; RUHLMAN, Tracey A.; JANSEN, Robert K. *Passiflora* plastome sequencing reveals widespread genomic rearrangements. **Journal of Systematics and Evolution**, v. 57, n. 1, p. 1-14, 2019.

RICE, Anna; ŠMARDA, Petr; NOVOSOLOV, Maria; DRORI, Michael; GLICK, Lior; SABATH, Niv; MEIRI, Shai; BELMAKER, Jonathan; MAYROSE, Itay. The global biogeography of polyploid plants. **Nature Ecology & Evolution**, v. 3, n. 2, p. 265-273, 2019.

RUHLMAN, Tracey A.; JANSEN, Robert K. The plastid genomes of flowering plants. In: **Chloroplast Biotechnology**. Humana Press, Totowa, NJ, 2014. p. 3-38.

SANDERSON, Michael J.; COPETTI, Dario; BÚRQUEZ, Alberto; BUSTAMANTE, Enriquena; CHARBONEAU, Joseph L. M.; EGUIARTE, Luis E.; KUMAR, Sudhir; LEE, Hyun Oh; MCMAHON, Michelle; STEELE, Kelly; WING, Rod; YANG, Tae-Jin; ZWICKL, Derrick; WOJCIECHOWSKI, Martin F. Exceptional reduction of the plastid genome of saguaro cactus (*Carnegiea gigantea*): Loss of the *ndh* gene suite and inverted repeat. **American Journal of Botany**, v. 102, n. 7, p. 1115-1127, 2015.

SFAIR, J.; BELLO, F.; FRANÇA, T.; BALDAULF, C.; TABARELLI, M. Chronic human disturbance affects plant trait distribution in a seasonally dry tropical forest. **Environmental Research Letters**, v. 13, n. 2, p.1-12, 2018.

SHAW, J.; LICKEY, E.B.; BECK, J.T.; FARMER, S.B.; LIU, W.; MILLER, J.; SIRIPUN, K.C.; WINDER, C.T.; SCHILLING, E.E.; SMALL, R.L. The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. **American Journal of Botany**, v. 92, n. 1, p. 142-166, 2005.

SHAW, J.; LICKEY, E.B.; SCHILLING, E.E.; SMALL, R.L. Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. **American Journal of Botany** v.94, n. 3, p.275–288, 2007.

SIMPSON, B. B.; LEWIS, G. P. New combinations in *Pomaria* (Caesalpinioideae: Leguminosae). **Kew Bulletin**, p. 175-184, 2003.

SIMPSON, Beryl B.; MIAO, Bo-Mao. The circumscription of *Hoffmannseggia* (Fabaceae, Caesalpinioideae, Caesalpinieae) and its allies using morphological and cpDNA restriction site data. **Plant Systematics and Evolution**, v. 205, n. 3, p. 157-178, 1997.

SOLTIS, Pamela S.; SOLTIS, Douglas E. The role of hybridization in plant speciation. **Annual Review of Plant Biology**, v. 60, [s.n.], p. 561-588, 2009.

SOUZA, G; COSTA, L; GUIGNARD, M; VAN-LUME, B; PELLICER, J; GAGNON, E; LEITCH, I; LEWIS, G. Do tropical plants have smaller genomes? Correlation between genome size and climatic variables in the Caesalpinia Group (Caesalpinioideae, Leguminosae). **Perspectives in Plant Ecology, Evolution and Systematics**, v. 38, [s.n.], p.13-23, 2019.

STANKOWSKI, Sean; RAVINET, Mark. Defining the speciation continuum. **Evolution**, v. 75, n. 6, p. 1256-1273, 2021.

STEFANOVIĆ, S.; COSTEA, M. Reticulate evolution in the parasitic genus *Cuscuta* (Convolvulaceae): over and over again. **Botany**, v. 86, n. 8, p. 791-808, 2008.

STULL, Gregory W.; SOLTIS, Pamela S.; SOLTIS Douglas. E.; GITZENDANNER, Matthew A.; SMITH, Stephen A. Nuclear phylogenomic analyses of asterids conflict with plastome trees and support novel relationships among major lineages. **American journal of botany**, v. 107, n. 5, p. 790-805, 2020.

SUAREZ-GONZALLEZ, Adriana; HEFER, Charles A.; CHRISTE, Camille; COREA, Oliver; LEXER, Christian; CRONK, Quentin C. B.; DOUGLAS, Carl J. Genomic and functional approaches reveal a case of adaptive introgression from *Populus balsamifera* (balsam poplar) in *P. trichocarpa* (black cottonwood). **Molecular ecology**, v. 25, n. 11, p. 2427-2442, 2016.

SUN, Y.; SKINNER, D. Z.; LIANG, G. H.; HULBERT, S. H. Phylogenetic analysis of *Sorghum* and related taxa using internal transcribed spacers of nuclear ribosomal DNA. **Theoretical and Applied Genetics**, v. 89, n. 1, p. 26-32, 1994.

TABARELLI, M; LEAL, I; SCARANO, F; SILVA, J. Caatinga: legado, trajetória e desafios rumo à sustentabilidade. **Ciência e Cultura**, v. 70, n. 4, p. 25-29, 2018.

TAYLOR, Scott A.; LARSON, Erica L. Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. **Nature Ecology & Evolution**, v. 3, n. 2, p. 170-177, 2019.

THE HELICONIUS GENOME CONSORTIUM. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. **Nature**, v. 487, n. 7405, p. 94-98, 2012.

THODE, Verônica A.; LOHMANN, Lúcia G.; SANMARTÍN, Isabel. Evaluating character partitioning and molecular models in plastid phylogenomics at low taxonomic levels: A case study using *Amphilophium* (Bignonieae, Bignoniaceae). **Journal of Systematics and Evolution**, v. 58, n. 6, p. 1071-1089, 2020.

VAN-LUME, Brena; ESPOSITO, Tiago; DINIZ-FILHO, J.; GAGNON, Edeline; LEWIS, Gwylin; SOUZA, Gustavo. Heterochromatic and cytomolecular diversification in the Caesalpinia group (Leguminosae): Relationships between phylogenetic and cytogeographical data. **Perspectives in Plant Ecology, Evolution and Systematics**, v. 29, [s.n.], p. 51-63, 2017.

VAN-LUME, Brena; MATA-SUCRE, Yennifer; BÁEZ, Mariana; RIBEIRO, Tiago; HUETTEL, Bruno; GAGNON, Edeline; LEITCH, Ilia J.; PEDROSA-HARAND, Andrea; LEWIS, Gwylin P.; SOUZA, Gustavo. Evolutionary convergence or homology? Comparative cytogenomics of Caesalpinia group species (Leguminosae) reveals diversification in the pericentromeric heterochromatic composition. **Planta**, v. 250, n. 6, p. 2173-2186, 2019.

VISCOSI, V.; LEPAIS, O.; GERBER, S.; FORTINI, P. Leaf morphological analyses in four European oak species (*Quercus*) and their hybrids: A comparison of traditional and geometric morphometric methods. **Plant Biosystems**, v. 143, n. 3, p. 564-574, 2009.

VISCOSI, Vincenzo; CARDINI, Andrea. Leaf morphology, taxonomy and geometric morphometrics: a simplified protocol for beginners. **PloS one**, v. 6, n. 10, p. e25630, 2011.

VOGEL, Alexander; SCHWACKE, Rainer; DENTON, Alisandra K.; USADEL, Björn; HOLLMANN, Julien; FISCHER, Karsten; BOLGER, Anthony; SCHIMIDT, Maximilian H.-W.; BOLGER, Marie E.; GUNDLACH, Heidrun; MAYER, Klaus F.X; WEISS-SCHNEEWEISS, Hanna; TEMSHC, Eva M.; KRAUSE, Kirsten. Footprints of parasitism in the genome of the parasitic flowering plant *Cuscuta campestris*. **Nature Communications**, v. 9, n. 1, p. 1-11, 2018.

WALKER, Joseph F.; WALKER-HALE, Nathanael; VARGAS, Oscar M.; LARSON, Drew A.; STULL, Gregory W. Characterizing gene tree conflict in plastome-inferred phylogenies. **PeerJ**, v. 7, p. e7747, 2019.

WANG, Gang; ZHANG, Xingtan; HERRE, Edward Allen; MCKEY, Doyle; MACHADO, Carlos A.; YU, Wen-Bin; CANNON, Charles H.; ARNOLD, Michael L.; PEREIRA, R odrigo A. S.; MING, Ray; LIU, Yi-Fei; WANG, Yibin; MA, Dongna; CHEN, Jin. Genomic evidence of prevalent hybridization throughout the evolutionary history of the fig-wasp pollination mutualism. **Nature Communications**, v. 12, n. 1, p. 1-14, 2021.

WANG, Xinfeng; HE, Ziwen; SHI, Suhua; WU, Cheng-I Genes and speciation: is it time to abandon the biological species concept?. **National Science Review**, v. 7, n. 8, p. 1387-1397, 2020.

WANG, Yin-Huan; QU, Xiao-Jian; CHEN, Si-Yun; LI, De-Zhu; YI, Ting-Shuang. Plastomes of Mimosoideae: structural and size variation, sequence divergence, and phylogenetic implication. **Tree Genetics & Genomes**, v. 13, n. 2, p. 41-59, 2017.

WEISING, K; GARDNER, R. A set of conserved PCR primers for the analysis of simple sequence repeat polymorphisms in chloroplast genomes of dicotyledonous angiosperms. **Genome**, v. 42, n. 1, p. 9-19, 1999.

WHEELER, Gregory L.; DORMAN, Hanna E.; BUCHANAN, Alenda; CHALLAGUNDLA, Lavanya; WALLACE, Lisa E. A review of the prevalence, utility, and caveats of using chloroplast simple sequence repeats for studies of plant biology. **Applications in Plant Sciences**, v. 2, n. 12, p. 1400059, 2014.

WHITE, T.J.; BRUNS, T.; LEE, S.; TAYLOR, J.W. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: **PCR Protocols: A Guide to**

Methods and Applications. San Diego, CA, USA: Academic Press. ISBN:01-23721-81-4. p.315–22, 1990.

WU, Chung-I.; TING, Chau-Ti. Genes and speciation. **Nature Reviews Genetics**, v. 5, n. 2, p. 114-122, 2004.

ZHANG, Rong; WANG, Yin-Huan; JIN, Jian-Jun; STULL, Gregory W.; BRUNEAU, Anne; CARDOSO, Domingos; DE QUEIROZ, Luciano Paganucci; MOORE, Michael J.; ZHANG, Shu-Dong; CHEN, Si-Yun; WANG, Jian; LI, De-Zhu; YI Ting-Shuang. Exploration of plastid phylogenomic conflict yields new insights into the deep relationships of Leguminosae. **Systematic Biology**, v. 69, n. 4, p. 613-622, 2020.

APÊNDICE A – PLASTOME EVOLUTION IN THE CAESALPINIA GROUP (LEGUMINOSAE) AND ITS APPLICATION IN PHYLOGENOMICS AND POPULATIONS GENETICS

Paulo Aecyo, André Marques, Bruno Huettel, Ana Silva, Tiago Esposito, Elâine Ribeiro, Inara R. Leal, Edeline Gagnon, Gustavo Souza, Andrea Pedrosa-Harand

Artigo publicado no periódico ***PLANTA*.** Esta é uma versão pós-revisão por pares anterior ao processo de revisão de texto. A reprodução deste foi feita com a permissão da Springer Nature. A versão autenticada final está disponível em: https://doi.org/10.1007/s00425-021-03655-8

Plastome evolution in the Caesalpinia group (Leguminosae) and its application in phylogenomics and populations genetics

Paulo Aecyo[1], André Marques[2], Bruno Huettel[2], Ana Silva[1], Tiago Esposito[1], Elâine Ribeiro[3], Inara R. Leal[3], Edeline Gagnon[4], Gustavo Souza[1], Andrea Pedrosa-Harand[1]*

[1]Laboratory of Plant Cytogenetics and Evolution, Department of Botany, Federal University of Pernambuco, Recife, Brazil,

[2]Max Planck Institute for Plant Breeding Research, Cologne, Germany,

[3]Laboratory of Plant-Animal Interaction, Department of Botany, Federal University of Pernambuco, Recife, Brazil.

[4]Royal Botanic Garden of Edinburgh, University of Edinburgh

*andrea.harand@ufpe.br; Phone +55 81 2126 8846 and Fax +55 81 2126 8348

ORCIDs:
Ana Silva 0000-0001-5956-8213
André Marques 0000-0002-9567-2576
Andrea Pedrosa-Harand 0000-0001-5213-4770
Bruno Huettel 0000-0001-7165-1714
Edeline Gagnon 0000-0003-3212-9688
Elâine Ribeiro 0000-0002-3632-1004
Gustavo Souza 0000-0002-5700-6097
Inara R. Leal 0000-0002-8125-2191
Paulo Aecyo 0000-0001-9254-5603
Tiago Esposito 0000-0002-0143-6450

Main conclusion

The chloroplast genomes of Caesalpinia group species are structurally conserved, but sequence level variation is useful for both phylogenomic and population genetic analyses.

Abstract

Variation in chloroplast genomes (plastomes) has been an important source of information in plant biology. The Caesalpinia group has been used as a model in studies correlating ecological and genomic variables, yet its intergeneric and infrageneric relationships are not

fully solved, despite densely sampled phylogenies including nuclear and plastid loci by Sanger sequencing. Here, we present the de novo assembly and characterization of plastomes from 13 species from the Caesalpinia group belonging to eight genera. A comparative analysis was carried out with 13 other plastomes previously available, totalizing 26 plastomes and representing 15 of the 26 known Caesalpinia group genera. All plastomes showed a conserved quadripartite structure and gene repertoire, except for the loss of four *ndh* genes in *Erythrostemon gilliesii*. Thirty polymorphic regions were identified for inter- or intrageneric analyses. The 26 aligned plastomes were used for phylogenetic reconstruction, revealing a well-resolved topology, and dividing the Caesalpinia group into two fully supported clades. Sixteen microsatellite (cpSSR) loci were selected from *Cenostigma microphyllum* for primer development and at least two were cross-amplified in different Leguminosae subfamilies by in vitro or in silico approaches. Four loci were used to assess the genetic diversity of *C. microphyllum* in the Brazilian Caatinga. Our results demonstrate the structural conservation of plastomes in the Caesalpinia group, offering insights into its systematics and evolution, and provides new genomic tools for future phylogenetic, population genetics, and phylogeographic studies.

Keywords: Caesalpinioideae; *Cenostigma*; Chloroplast genome; cpSSR; Genetic diversity; Seasonally Tropical Dry Forest

List of abbreviations

AAF	Assembly and Alignment Free
cpSSR	Chloroplast simple sequence repeat
IR	Inverted Repeat
LSC	Long single copy
Pi	Nucleotide diversity
SSC	Small single copy

Acknowledgements We thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, PELD process 403770/2012-2, Universal process 426738/2018-7), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES code 0001) and the Fundação de Amparo à Ciência e Tecnologia de Pernambuco (FACEPE, processes BIC-0846-2.02/17, BIC-0624-2.02/18) for financial support. We also thank Tatiane Menezes for assistance during field work, Davi Jamelli for drawing the map, Brena Van-Lume and

Yennifer Mata-Sucre for DNA extraction and Benoit Loeuille, Cícero Almeida for critical reading of the manuscript. We also would like to thank the Sequencing Platform of CB-UFPE and Heidi Lacerda from LABBE for their help with genotyping. APH, GS and IRL also thank CNPq for productivity grants (processes 310804/2017-5, 310693/2018-7 and 308300/2018-1, respectively).

Author contribution statement

PA collected the samples, performed the experiments, analysed the data, wrote the manuscript, and prepared the figures and tables. AM, BH, AS, and TE performed experiments and analysed data. ER and IRL designed and helped with the field work. EG discussed methods and results. GS and APH conceived, designed, and guided the experiments. All authors revised and approved the manuscript.

Introduction

Variation in the chloroplast genome (plastome) has been an important source of information in plant biology (Ebert and Peakall 2009). It is usually described as a circular DNA, uniparentally inherited and lacking recombination, but some studies have demonstrated that these character-istics may vary among organisms (Gonçalves et al. 2020). Nevertheless, its small size, haploid nature, high copy num-ber in plant cells, and low paralogy compared to the nuclear genome make plastomes good markers for phylogenetic studies (Shaw et al. 2005; Thode et al. 2020). In general, plastomes have two Inverted Regions (IRs) separated by a Large and a Small Single Copy region (LSC and SSC, respectively) (Jansen and Ruhlman 2012). Although there are exceptions in Leguminosae (Jansen et al. 2008), Berberidaceae (Su et al. 2018), Cactaceae (Solórzano et al. 2019), Convolvulaceae (Banerjee and Stefanović 2019), and Passifloraceae families (Cauz-Santos et al. 2020), the number and order of genes and the structure of the plastidial DNA are usually conserved among angiosperms.

The assembly of plastomes may also facilitate population genetic analyses by the identification of simple sequence repeats (cpSSR), which are largely used as molecular markers (Ebert and Peakall 2009; Wheeler et al. 2014). Simple sequence repeats (SSR) or microsatellites are sequences of 1–6 bp that are repeated in tandem throughout the genome. They are present in coding and non-coding regions in prokaryotes and eukaryotes, both in nuclear (nuSSR) and organellar genomes (Zane et al. 2002). However, unlike nuSSRs, cpSSR present a certain degree of conservatism and are usually uniparentally inherited (mostly maternally in angiosperms), which can complement nuSSR analysis in plants (Provan et al.

2001; Ebert and Peakall 2009; Wheeler et al. 2014; Gonçalves-Oliveira et al. 2017; Mota et al. 2019).

The Caesalpinia group belongs to the family Leguminosae, subfamily Caesalpinioideae, and consists of 26 monophyletic genera with ca. 225 species, including species of economic, ecological, and cultural importance, such as *Paubrasilia echinata* (Lam.) Gagnon, H.C. Lima & G.P.Lewis, the brazilwood (Gagnon et al. 2016). The group has a complex taxonomic history and has undergone recent nomenclatural alterations based on the results of molecular phylogenetic studies (see Gagnon et al. 2016). This group includes species with different growth habits, which are associated with biome shifts (e.g., trees and shrubs in Succulent biomes and lianas in Savanna and Rainforest lineages from Asia; Gagnon et al. 2019). Caesalpinia group species are characterized by the presence of prickles, spines, glandular trichomes, and secretory structures as defence mechanisms, as well as zygomorphic flowers with a lower cucullate sepal on the calyx and stamens crowded around the pistil (Lewis 2005; Gagnon et al. 2013, 2016). The group exhibits a pantropical distribution throughout seasonally dry and semi-arid habitats of the Succulent Biome, especially in Seasonally Tropical Dry Forests in the Neotropics (Gagnon et al. 2016, 2019).

Due to its environmental preferences and correlation between ecological variables and cytogenetic/genomic traits among its genera, the group has been proposed as a model to investigate the relationships between ecological and genomic variables (Van-Lume et al. 2017, 2019; Gagnon et al. 2019; Souza et al. 2019a; Mata-Sucre et al. 2020a, b). However, to further understand the correlation between ecological and genomic variances, it is important to clarify its phylogenetic relationships.

The Caesalpinia group is well studied from a phylogenetic point of view (Lewis and Schrire 1995; Simpson and Miao 1997; Simpson and Lewis 2003; Nores et al. 2012; Gagnon et al. 2013, 2016). The last and most robust study in the group was performed by Gagnon et al. (2016), which was based on five plastid loci (*rps16*, *trnD-trnT*, *ycf6-psbM*, *matK-3'-trnK intron*, and *trnL-trnF*) and the ribosomal *ITS1*-5.8S-*ITS2* DNA region. It included a dense taxonomic sampling with 172 species/taxa (84%), and representatives of all previously described genera in the Caesalpinia group. This led to a new generic classification of the Caesalpinia group, with the recognition of 26 monophyletic genera divided into two subclades (Gagnon et al. 2016). Despite the wide sampling and use of different loci, some relationships remain uncertain, with some clades showing low statistical support [e.g., *Paubrasilia* + *Caesalpinia* and (*Guilandina* (*Moullava* (*Mezoneuron* + *Pterolobium*)))] and a poorly resolved backbone. Besides, several genera show incongruences, such as low supported

clades or no monophyletic species (e.g., *Cenostigma*). Thus, new genomic tools, such as complete plastome sequence and the development of useful molecular markers, might improve the phylogenetic inference in this group.

Complete plastomes have been used to solve phylogenetic relationships across different taxonomic levels (Gonçalves et al. 2019; Duvall et al. 2019; Ji et al. 2019). Plastomes provide information about the whole set of genes, and recent studies have observed that the plastome does not evolve as a single locus, so different gene trees may present different topologies (Gonçalves et al. 2019). In addition, different plastome regions may vary in regard to its phylogenetic signal, being necessary to use plastome partitions (Gonçalves et al. 2019; He et al. 2019; Walker et al. 2019; Thode et al. 2020).

The Caesalpinia group is well-studied from a phylogenetic point of view (Lewis and Schrire 1995; Simpson and Miao 1997; Simpson et al. 2003; Nores et al. 2012; Gagnon et al. 2013, 2016). The last and most robust study in the group was performed by Gagnon et al. (2016), which was based on five plastid loci (*rps16, trnD-trnT, ycf6-psbM, matK-3'-trnK intron,* and *trnL-trnF*) and the ribosomal *ITS1*-5.8S-*ITS2* DNA region. Furthermore, it included a dense taxonomic sampling with 172 species/taxa (84%), and representatives of all previously described genera in the Caesalpinia group. This led to a new generic classification of the Caesalpinia group, with the recognition of 26 monophyletic genera divided into two subclades (Gagnon et al. 2016). Despite the wide sampling and use of different loci, some relationships remain uncertain, with some clades showing low statistical support [e.g. *Paubrasilia* + *Caesalpinia* and (*Guilandina* (*Moullava* (*Mezoneuron* + *Pterolobium*)))] and a poorly resolved backbone. Besides, several genera show incongruences, such as low supported clades or no monophyletic species (e.g. *Cenostigma*). Thus, new genomic tools, such as complete plastome sequence and the development of useful molecular markers, might improve the phylogenetic inference in this group.

Complete plastomes have been used to solve phylogenetic relationships across different taxonomic levels (Gonçalves et al. 2019; Duvall et al. 2019; Ji et al. 2019). Plastomes provide information about the whole set of genes, and recent studies have observed that the plastome do not evolve as a single locus, so different gene trees may present different topologies (Gonçalves et al. 2019). In addition, different plastome regions may vary in regard to its phylogenetic signal, being necessary to use plastome partitions (Gonçalves et al. 2019; He et al. 2019; Walker et al. 2019; Thode et al. 2020).

Currently, 13 plastomes of Caesalpinia group species are available in the NCBI database from previous studies (Zhang et al. 2016; Koenen et al. 2020; Wang et al. 2020;

Zhang et al. 2020). Rearrangements in their structure were so far not evaluated, and neither was their potential for population genetic studies or phylogenomic analysis. Thus, we raised the questions: how conserved are plastomes in the Caesalpinia group? Is plastome variation useful to infer inter- and intrageneric relationships in this group? The aims of this study were (i) to assemble and characterize the plastomes of 13 species from eight genera of the Caesalpinia group and to perform a comparative analysis with 13 other plastomes from the group previously available; (ii) to investigate the usefulness of these plastomes for phylogenomic analysis; (iii) to develop a set of cpSSR for population analyses; (iv) to test this set of markers for population analyses; and (v) to evaluate the cross-amplification of these markers to other Leguminosae species.

Material and methods

Sampling and DNA Extraction

For full plastome NGS sequencing, fresh leaves were collected from 13 species (Table 1) grown in the experimental garden of the Laboratory of Plants Cytogenetics and Evolution, representing eight out of the 26 Caesalpinia group genera. Total genomic DNA (gDNA) was extracted from 50 mg of leaves using the cetyltrimethylammonium bromide (CTAB) protocol (Weising et al. 2005) and purified via isopropanol precipitation.

To assess plastid genetic diversity and structure, we chose *Cenostigma microphyllum* (Mart. ex G.Don) E.Gagnon & G.P.Lewis for population genetic analysis. This species is endemic and widespread in Caatinga Domain - a Seasonally Tropical Dry Forest from Brazil – at the Catimbau National Park (PARNA Catimbau, 8°24'00" and 8°36'35" S; 37°0′30″ and 37°1′40″ W) in Northeast of Brazil (Table S1). The PARNA Catimbau covers an area of nearly 640 km^2, where annual rainfall varies from 480 mm to 1100 mm per year, while the mean annual temperature is 23 ºC (Rito et al. 2017). Quartzite sandy soils are predominant in the Park, supporting a relatively open, low-stature vegetation in which Leguminosae, Euphorbiaceae and Cactaceae are the dominant families (Rito et al. 2017).

Plots were separated by a minimum of 2 km and occurred within an area of 215 km^2 (Rito et al. 2017). Fresh leaves of *Cenostigma microphyllum* were collected from 99 individuals distributed in eight of those plots. Genomic DNA was extracted from leaves stored on silica gel following the protocol of Doyle and Doyle (1987), as in Ferreira & Grattapaglia (1995), and stored at –20 °C.

Plastome sequencing, assembly, and annotation

Ten micrograms of total gDNAs were used for Illumina sequencing (Illumina HiSeq 2000 platform) generating paired-end reads of 250 bp in a genome skimming approach (~0.1× coverage; see genome sizes in Souza et al. 2019a). *De novo* plastome assembly was performed by NOVOPlasty *v.* 3.8.3 (Dierckxsens et al. 2017) using the *rbcL* gene as seed and default parameters. Based on the software instructions, we did not filter or quality trim the reads. Thus, the raw whole Illumina dataset (only removing adapters) was analysed for plastome assembling. The contigs were imported to Geneious v. 9.1.8 (Kearse et al. 2012) and the assembly checked by mapping the raw reads to the contigs using the Geneious mapper with low sensitivity. The plastomes were annotated using two approaches: the Geneious annotation tool, guided by the available *Senna tora* (L.) Roxb plastome (NC_030193), and the GeSeq online software (Tillich et al. 2017). Both annotations were compared, and incongruences were manually checked by BLASTn. Complete annotated plastome maps were generated using OrganellarGenomeDraw (OGDraw v. 1.2; Lohse et al. 2013).

SSR detection and primer design in *Cenostigma microphyllum*

Microsatellites were identified in the plastome of *Cenostigma microphyllum* using Phobos (Mayer 2010) as a plugin in Geneious v. 7.1.9. The minimum numbers of repeats were: ten for mono-, five for di-, four for three, three for tetra-, penta- or hexanucleotides. Primers were designed using 400 bp long sequences including the SSR loci. We used the PRIMER3PLUS online platform (Untergasser et al. 2007) with default parameters, except for the ideal primer size of 24 bp, melting temperature (Tm) = 60 °C and the amplified fragment size of 100 - 400 bp. The primers pairs were tested for the formation of hairpins or dimers at OligoAnalyzer v. 3.1 platform (https://www.idtdna.com/calc/analyzer).

cpSSR amplification and genotyping in *Cenostigma microphyllum*

Initially, eight individuals of *Cenostigma microphyllum* from different plots on PARNA Catimbau were used for polymorphism screening of ten cpSSR developed in this study. PCR amplifications were carried out in a final volume of 10 µl including 20 ng of gDNA, 1× PCR buffer, 2.5 mM $MgCl_2$, 200 µM dNTP, 0.5× TBT pH 8,0 (Samarakoon et al. 2013), 0.25 µM of each primer pair and 0.3 µl of homemade Taq polymerase. The amplification program was: 94 °C for 5 min (1×); 94 °C for 1 min, 50 – 60 °C for 1 min, 72 °C for 3 min (30×); 72 °C for 10 min (1×). The products were checked on a 3% agarose gel. Loci that showed a single band were amplified following the protocol of Schuelke (2000)

with primers indirectly labelled by adding a M13 tail in the 5'– end of the forward primer. PCR was performed in a final volume of 25 µl including 20 ng of gDNA, 1× PCR buffer, 1.50 mM $MgCl_2$, 250 µM dNTP, 1× TBT, 0.125 µM of primer forward with a M13 tail, 0.50 µM of primer reverse, 0.50 µM of a M13 tail attached to a fluorophore (FAM, NED, PET or VIC) and 0.20 µl of homemade Taq polymerase. The amplification program was: 94 °C for 5 min (1×); 94 °C for 30s, 50 – 60 °C for 45s, 72 °C for 45s (30×); 94 °C for 30s, 53 °C for 45s, 72 °C for 45s (8×); 72 °C for 10 min (1×). The products were checked on a 3% agarose gel. The PCR products were genotyped in an ABI 3500 sequencer (Applied Biosystems®) at the Sequencing Platform of the Bioscience Centre at Federal University of Pernambuco. The products were analysed using a multiplex approach with 1µl of the pooled sample, 9.50 µl formamide and 0.50 µl of GeneScan 600 LIZ® size standard as ladder.

Cross-amplification of cpSSR to other legumes

The loci developed in this work were tested for cross-amplification to other legumes *in silico* or *in vitro* by PCR. For *in silico* analyses, Geneious v. 7.1.9 was employed to test the cross-amplification to at least one species of each genus sequenced in this study, as well as other seven Caesalpinia group species from GenBank (*Balsamocarpon brevifolium, Biancaea sappan, Coulteria platyloba, Haematoxylum brasiletto, Mezoneuron cucullatum, Moullava spicata, Pterolobium punctatum*) and five species representing all Leguminosae subfamilies (*Schnella trichosepala* (L.P.Queiroz) Wunderlin., MF135599 - Cercicoideae; *Zenia insignis* Chun, NC_045299 - Dialioideae; *Tamarindus indica* L., KJ468103.1 - Detarioideae; *Duparquetia orchidaceae* Baill., MN709829.1 – Duparquetioideae; and *Vigna unguiculata* (L.) Walp., KJ468104.1 – Papilionoideae). To confirm cross-amplification, the primer annealing regions needed to be conserved (here defined as ≥ 90% of identity) and it was required the presence of the SSR with a length ≥ 10 bp in the region (Weising and Gardner 1999). For further discussion, the presence of SSRs with a length ≥ 6 bp were also analysed. We also evaluated intraspecific polymorphism by *in vitro* analysis for the species which we had plastomes for two accessions (*Cenostigma pyramidale, Erythrostemon gilliesii, Guilandina bonduc, Libidibia coriaria* and *Mezoneuron cucullatum*).

For *in vitro* cross-amplification, gDNA of *Trischidium molle* (Benth.) H.E.Ireland (Papilionoideae subfamily) was extracted from eight individuals from different localities using dried leaves from herbarium specimens (Table S2). DNA extraction and PCR amplification with M13 tail were carried out as described above.

Structure and genetic diversity analysis of *Cenostigma microphyllum*

The four most polymorphic cpSSR loci (CmCPSSR4, CmCPSSR8, CmCPSSR12 and CmCPSSR15) were genotyped for 99 individuals from the eight plots of PARNA Catimbau. The software Haplotype v. 1.05 (Eliades and Eliades 2009) was used for assessing the total number of haplotypes (H), private haplotypes (P) and haplotypic richness (Hr). The haplotype network was obtained with Network v. 5.0.0.3 (http://www.fluxus-engineering.com). The coordinates were plotted on the map using the software ArcGIS. The final version of the map was drawn in CorelDRAW X6.

Detection of high polymorphic regions for phylogenetic analyses of the Caesalpinia group

In order to perform a broader phylogenomic analysis of the Caesalpinia group, 13 additional plastomes from seven genera were downloaded from the GenBank (Table 1). The 26 Caesalpinia group plastomes were aligned by pairwise MAFFT alignment on Geneious v. 7.1.9. Sliding window analysis was conducted to determine DNA polymorphism and nucleotide diversity (Pi) using 200 bp step size and 600 bp window length in DnaSP v. 6.0 (Rozas et al. 2017). Thirty regions with the highest Pi value were identified.

Since at least two plastomes from different species of *Cenostigma*, *Coulteria*, *Erythrostemon* and *Libidibia* are available, the window analysis was also performed to find suitable regions for phylogenetic and barcoding analyses, as described for the whole Caesalpinia group above. The SSC regions of *Erythrostemon gilliesii* and *Guilandina bonduc* (Zhang et al. 2020) were inverted for the alignment. Thirty regions with the highest Pi value were identified for each genus.

Caesalpinia group plastome phylogeny

A total of 28 plastomes were used for phylogenomic analyses, 26 plastomes representing 15 (54%) of the 26 Caesalpinia group genera (Table 1), and two outgroup taxa: *Senna tora* [Cassia clade] and *Leucaena trichandra* Urb. [Mimosoid clade] - KT428297). Alignment of whole plastomes was performed with MAFFT v. 7.222 in an online platform (Katoh et al. 2019).

Five different matrices were used for phylogenetic analysis: (i) total plastomes alignment excluding IRa, (ii) IRa only, (iii) LSC only, (iv) SSC only, (v) and 113 concatenated genes (protein coding genes and transfer RNAs, excluding IRa genes). Phylogenetic relationships were inferred for each matrix by two different phylogenetic

methods: Maximum likelihood (ML) and Bayesian Inference (BI). The ML and BI analyses were performed under a General Time Reversible + Gamma + Invariant sites (GTR + G + I) model (Abadi et al. 2019). The ML analysis was performed using RAxML-HPC2 on CIPRES Science Gateway (Miller et al. 2010). The BI analysis was performed using BEAST v. 1.8.0 (Drummond and Rambaut 2007). Analyses were run using an uncorrelated log normal relaxed clock and a Birth and Death speciation model (Gernhard 2008). A run of 10×10^6 generations as performed, sampling every 1,000 generations. To verify the effective sampling of all parameters and assess convergence of independent chains, we examined their posterior distributions in Tracer v. 1.6 and the MCMC sampling was considered sufficient at effective sampling sizes (ESS) higher than 200. After removing 25% of samples as burn-in, the independent runs were combined, and a maximum clade-credibility (MCC) tree was constructed using TreeAnnotator v. 1.8.2. (Drummond and Rambaut 2007).

Since different plastome partitions may present different phylogenetic signals resulting in different topologies (Gonçalves et al. 2019, He et al. 2019, Walker et al. 2019, Thode et al. 2020), and to avoid the effects of the observed gaps in the tree topology (Duvall et al. 2019), we also performed the ML and BI analysis after excluding all gaps present in the alignment of different partitions with Geneious v. 7.1.9. Furthermore, we performed an Assembly and Alignment-Free (AAF) method based on *k*-mer analysis. The AAF tool (Fan et al. 2015) can infer phylogenetic trees from genome skimming data, without the need of aligning reads. Although this method does not require a sequence alignment, we used the plastome alignment as input option and default parameters. Branch supports were inferred by means of Posterior Probability (PP) in the BI, or Bootstrap support (BS) for the ML and AAF analysis, with 10,000 parametric bootstrap replicates and default parameters. Clades with a PP ≥ 0.95 and BS ≥ 95% were considered highly supported.

Results

Sequencing and comparative analysis of plastome structure

High-throughput sequencing of whole-genomic DNA for 13 Caesalpinia group species was obtained at ~ 0.1 × total genomic coverage (GenBank accession number: PRJNA739461) and assembled de novo to obtain whole plastome sequences. NOVOPlasty was able to produce a single circularized contig for each species (Table 1, Fig. 1). We present, for the first time, plastome information of four Caesalpinia group genera (*Arquita*, *Caesalpinia*, *Paubrasilia*, and *Tara*), among them the monospecific genera Paubra- silia, which comprises the plant symbol of Brazil: *P. echinata*, the brazilwood. We also present five new plastomes

for four other genera (*Cenostigma microphyllum*, *Coulteria mollis*, *Erythrostemon hughesii*, *Erythrostemon pannosus*, and *Libidibia ferrea*) and a second accession for four species (*Cenostigma pyramidale*, *Erythrostemon calycinus*, *Guilandina bonduc*, and *Libidibia coriaria*). The plastomes assembled in this study were compared with 13 Caesalpinia group species plastomes retrieved from GenBank, representing a total of 15 of the 26 genera described for the group.

All plastomes presented the typical quadripartite struc- ture, showing a variation of 11.4 kbp in length, from 149,235 bp in *Erythrostemon gilliesii* to 160,677 bp in *Coulteria platyloba* (Fig. 1; Table 1). This variation was mainly observed in the intergenic regions from the LSC and SSC. However, an increase in length of 957 bp was observed in the *Tara cacalaco* IR when compared to *Arquita mimosifolia*, which showed the shortest IR. This increase at the IR was observed in the intergenic regions with no inclusion or deletion of any genic sequence (Tables 1, 2**)**.

The number of genes was conserved between the species, with 130 genes divided in 84 CDS, 8 rRNA, and 37 tRNA, except for *Erythrostemon gilliesii* which presented a loss of four genes (*ndhJ*, *ndhK*, *ndhC*, and *ndhF*). The order and distribution of the genes were similar in all species. Eighteen genes occurred in the IR, 12 genes occurred in the SSC region (11 in *E. gilliesii*), and 82 genes in the LSC region (79 for *E. gilliesii*) (Fig. 1). A full description of genes found in the plastomes of these species is included in Table 2.

Distribution of simple sequence repeats and primer design for *Cenostigma microphyllum*

A total of 79 cpSSR with a mono-, di-, tri- or tetranucleotide motifs were found in the plastome of *Cenostigma microphyllum*, ranging from 10 bp to 15 bp in size. From those, three were duplicated in the IR. Mononucleotides were the most abundant (62), including only one repetition of guanines (G_{12}). No penta- or hexanucleotides SSR were found. Most SSRs were in intergenic regions (63.29%); however, 36.7% were distributed in 14 genes, with the *ycf1* containing six SSRs (Table S3). A total of 20 loci, which showed the longest SSR sequences, were chosen for primer design, but as four of these loci were already described in the literature (ccmp1, ccmp4, ccmp5 and ccmp10, by Weising and Gardner 1999), they were not further tested here (Table S4).

Cross-amplification of CmCPSSR to other legumes

The sixteen designed primers were transferred in silico to 21 species representing all analysed genera and the six subfamilies of Leguminosae (Table 3). At least two loci were

cross-amplified to each subfamily. The locus with the highest rate of cross-amplification was CmCPSSR6, located in the rpoC2 gene. It was cross-amplified to all tested Leguminosae species, with the exception of *Vigna unguiculata* from the Papilionoideae subfamily (Tables 3; S5). However, this locus had the same length in all species. While most of the annealing region of the primers were conserved among species, it was not possible to find an SSR with a size ≥ 10 bp in the putatively amplified fragment in several cases. Considering these criteria, *Moullava spicata* was the species with the highest degree of cross-amplification, with a total of 13 loci cross-amplified in silico. Nevertheless, it was possible to observe that the great majority of the loci showed SSR ≥ 6 bp and so could potentially present some polymorphism between individuals (Tables 3; S5).

We also performed an in silico analysis to screen for polymorphism in the number of repeats (size of the amplified fragment) in *Cenostigma pyramidale*, *Erythrostemon gilliesii*, *Guilandina bonduc*, *Libidibia coriaria*, and *Mezoneuron cucullatu*m. It was observed polymorphism in two loci in *G. bonduc* (CmCPSSR1 and CmCPSSR5) and *M. cucullatum* (CmCPSSR7 and CmCPSSR12), seven in *L. coriaria* (CmCPSSR5, CmCPSSR8 to 12, and CmCPSSR16), and nine in *C. pyramidale* (CmCPSSR1, CmCPSSR3, CmCPSSR5, CmCPSSR9 to 12, CmCPSSR15, and CmCPSSR16; Tables 4; S5). None of the loci were polymorphic in *E. gilliesii*.

For the ten loci presenting the largest SSRs, we designed primers to test for polymorphism in vitro. Nine of these loci cross-amplified in *Trischidium molle*, a papilionoid legume, except for the loci CmCPSSR16. Five of these loci were polymorphic, showing two-to-four alleles per loci in a subset of eight individuals of *T. molle* from different localities (Table 4).

Genetic diversity and structure in *Cenostigma microphyllum*

A total of six cpSSR, of the ten loci screened, revealed polymorphism of two to three alleles per locus in a subset of *Cenostigma microphyllum* individuals from the eight plots of PARNA Catimbau and four of these loci (CmCPSSR4, CmCPSSR8, CmCPSSR12 and CmCPSSR15) were chosen for a population genetic analysis (Table 5). Ninety-nine individuals of *C. microphyllum* in the Catimbau National Park were successfully genotyped and it was possible to identify a total of nine alleles for the four selected loci in this population. When combined, the four cpSSR markers identified ten haplotypes (H1-10; Fig. 2, Table S6). The most frequent and distributed haplotype was H5, present in 43 individuals from five plots. A total of eight haplotypes were considered private, however they were at low

frequencies, except for H10 and H6 presented in P08 and P23 respectively (Fig. 2, Table S6). One to three haplotypes were observed per plot, showing a mean effective number of 1.272 and a mean haplotypic richness of 0.587. The genetic diversity ranged from zero in plots P12, P15 and P21 to 0.500 in P28, showing a mean of 0.199 (Table 5).

Identification of highly polymorphic regions for phylogenetic analyses

The 30 highest polymorphic regions were identified for the Caesalpinia group, as well as for infrageneric phylogenetic analyses in four genera with plastomes available for more than one species: *Cenostigma, Coulteria, Erythrostemon*, and *Libidibia*. For the Caesalpinia group as a whole, nucleotide diversity indices (Pi value) varied from 0 to 0.04811 in a total of 679 sliding windows. The 30 windows show- ing the highest Pi values were both in intergenic (such as *rps12–clpP* and *trnK-UUU–rps16*) and genic regions (*ycf1, clpP*, and *rpl16*). In some cases, the window included both genic and intergenic regions, and the largest from both regions were considered (Table S7).

The same pattern of polymorphic intergenic and genic regions was observed when performing the window analysis in *Cenostigma, Coulteria, Erythrostemon*, and *Libidibia* genera (Table S7). The most polymorphic regions for each genus differed from each other and from the Caesal- pinia group regions (Table S7). The only region that was shared between the five analyses was the intergenic region *trnK-UUU—rps16* gene. Although the *ycf1* gene showed the highest Pi value in Caesalpinia group (Pi = 0.0481), *Cenostigma* (Pi = 0.078) and *Erythrostemon* (Pi = 0,03267), it was not amongst the most polymorphic loci in *Coulteria* and *Libidibia*.

Phylogenomic relationships between Caesalpinia group species

For the phylogenomic analyses, we generated two groups of matrices, based on alignments with and without the gaps. However, most of the trees recovered from alignments without gaps showed lower support values for some clades (Suppl. Fig. S1). Thus, we used the alignments with gaps for further discussions. Five matrices were generated based on our partition scheme: the total alignment excluding the IRa (150.1 kbp), only the IRa (26.6 kbp), only the SSC (20.7 kbp), only the LSC (102.7 kbp), and 113 concatenated genes (93.8 kbp). The trees performed using the IRa matrix showed a different topology when compared to the other matrices, with low support in several clades, especially in the ML analysis. Therefore, both ML and BI trees for this matrix were not further considered.

For the four remaining matrices and all phylogenetic approaches (ML and BI for four matrices and AAF for total plastome reads only), we observed a completely resolved topology for the Caesalpinia group (Fig. 3, Suppl. Fig. S2). A total of nine phylogenetic trees were analysed. All trees showed the same backbone topology, confirming the monophyly of the Caesalpinia group and the formation of two major subgroups, Clade I and Clade II (Gagnon et al. 2016). Most of the trees (BI and AAF analyses) showed the same topology for all clades and maximum support for most of them (Fig. 3). Within Clade I, *Haematoxylum* was sister to two subclades (*Guilandina* (*Moullava* (*Biancaea* (*Pterolobium* + *Mezoneuron*)))) and (*Caesalpinia* (*Paubrasilia* (*Tara* + *Coulteria*)). *Biancaea*, *Caesalpinia*, and *Pterolobium* did not show the same relationships for all matrices, with branch support varying across plastome partitions and inference methods. The relationship mentioned above was the most frequently recovered and highly supported mainly on BI analysis (Fig. 3, Suppl. Fig. S2).

Clade II showed the same topology for all analyses, with *Cenostigma* as sister to both the *Libidibia* + *Balsamocarpon* and *Arquita* + *Erythrostemon* clades, with maximum support for all clades. However, there was an incongruence in the relationship of *Cenostigma pyramidale* accessions, with one closer to *C. microphyllum* than to the other *C. pyramidale* accession, with maximum support in all phylogenies (Fig. 3, Suppl. Fig. S2). We also observed a 99.9% of similarity of this accession of *C. pyramidale* to *C. microphyllum*, while the other was 98.8% similar. The same pattern was observed comparing the SSRs between the three plastomes.

Discussion

Caesalpinia group plastomes are conserved in structure

In the present paper, we reconstructed the plastomes of 13 Caesalpinia group species, with new plastome information for four genera, and compared to 13 plastomes available for the group. We observed a conservation in their structure, showing a maximum variation of 11.4 kbp in length. Because of its importance to vital functions in plants, plastomes are usually highly conserved, with little or no differences in related species (Jansen and Ruhlman 2012). However, some groups of plants present several rearrangements in the plastome structure (Jansen and Ruhlman 2012), including some legumes (Cai et al. 2008; Dugas et al. 2015; Asaf et al. 2017; Wang et al. 2017b).

The plastome stability reported here is remarkable in view of the ancient age of the Caesalpinia group (~ 54.78 Mya; Gagnon et al. 2019). The Leguminosae family has several studies regarding plastome evolution at different systematic levels (Bai et al. 2020; Koenen et

al. 2020; Zhang et al. 2020) and the macroconservatism in plas- tome structure has been reported for other species from the Caesalpinioideae subfamily (Mimosoid clade; Souza et al. 2019b; Song et al. 2019). This conservatism could probably be explained by both the slow rate of evolution of plastomes and the woody habit of the Caesalpinia group species analysed (Schwarz et al. 2017). The only variation observed between the plastomes assembled and annotated here when compared to previously sequenced species was the orientation of the SSC in *Guilandina bonduc* and *Erythrostemon gilliesii* (Koenen et al. 2020) and a loss of four genes of the ndh complex in *E. gilliesii*. The orientation of the SSC can be explained by a limitation of assembly caused by the similarity of the 250 bp reads, as described by Dierckxsens et al. (2017). Another cause could be a natural state of heteroplasmy of cpDNA, as described by Walker et al. (2015) and the references therein, but more studies would be necessary to confirm this.

The ndh complex encodes the NADH dehydrogenase enzymes, important proteins in the photosynthetic pathway. Losses of ndh genes were observed previously on other groups, usually related to a loss of photosynthesis or change of habits (Jansen and Ruhlman 2012; Braukmann et al. 2013). However, it also occurs in photosynthetic species (Kim et al. 2020), and some hypotheses are that either those genes were transferred to the nucleus or mitochondria genomes or other genes have replaced its function (Kim et al. 2020). Thus, more studies are necessary to test these hypotheses. Although plastomes from the Caesalpinia group were structurally conserved, our data revealed a high degree of sequence polymorphism in genic and intergenic regions. This high level of polymorphism in sequence level is, however, expected for the Caesalpinia group given its ancient origin, with an estimated crown age of ~ 54.78 Mya (Gagnon et al. 2019).

cpSSRs in *Cenostigma microphyllum* and cross-amplification to other legumes

Most of the SSRs found in the *C. microphyllum* plastome were A/T repetitions varying from 10 to 15 bp. The presence of C/G repetitions were rare, as expected for plastomes (Weising and Gardner et al. 1999). The SSRs were mainly distributed in intergenic regions. Among the genic regions, ycf1 showed the largest number of SSRs, in agreement with other legumes plastomes (Asaf et al. 2017; Song et al. 2019; Souza et al. 2019b). *Paubrasilia echinata* is the only species from Caesalpinia group for which three cpSSRs have been developed so far (CECPSSR1–3, Lira et al. 2003), presenting a length of 7, 8, and 13 bp, respectively. However, it was possible to identify larger SSR loci with different motifs in *P.*

echinata plastome, varying from 10 to 15 bp, which may present higher polymorphism and should be further investigated.

In this study, we designed 16 new cpSSR primers for *Cenostigma microphyllum* and most of them were cross- amplified to the species of the Caesalpinia group and to other species from different subfamilies of Leguminosae. *Vigna unguiculata* was the only species with no cross-amplified locus. However, nine loci were cross-amplified in vitro to *Trischidium molle*, a species from an early divergent clade of the Papilionoideae subfamily (Cardoso et al. 2012, 2013). The lack of conserved loci in *V. unguiculata* may be related to its herbaceous habit, because herbaceous plants seem to present a higher rate of substitutions (Schwarz et al. 2017). Besides, 12 loci showed polymorphism in at least one spe- cies on either in silico analyses of the four Caesalpinia group species (*Cenostigma pyramidale, Erythrostemon calycinus, Guilandina bonduc*, and *Mezoneuron cucullatum*) or in the in vitro analysis between different populations of *T. molle*, confirming its potential for phylogeographic analysis for those species.

We considered the loci cross-amplified when the primer annealing region had more than 90% similarity between species and there was an SSR sequence with at least 10 bp in size (Weising and Gardner 1999), given the pattern of SSR evolution (Ellengreen 2000). Here, we observed a high similarity in the annealing region of the primers in all species and the presence of shorter microsatellites in the intervening regions. Indeed, there are some examples of shorter SSR sequences that presents a certain degree of polymorphism (Lira et al. 2003; Lemos et al. 2018). In this sense, the number of cross-amplified loci may be higher than the expected from our conservative analyses.

New set of cpSSR to assess genetic diversity of *Cenostigma microphyllum*

From the 16 new cpSSR primers developed for *Cenostigma microphyllum*, ten were validated through PCR. Six of these loci were polymorphic in a sample of eight individuals from the Catimbau National Park. Four loci were chosen to assess the genetic diversity of these *C. microphyllum* populations. As expected for cpSSRs (Powell et al. 1996), a low degree of genetic diversity was observed in the present sample, with a mean of 0.199. Furthermore, a moderate haplotype richness was observed (Rh = 0.587). Given the geographic proximity of the analysed individuals and the high inter- and intraplot variation, we consider that *C. microphyllum* shows a relatively high plastid genetic diversity when compared with other species (Lira et al. 2003; Wang et al. 2017a; López- Villalobos and Eckbert 2018), what may be further investigated with the developed markers.

Plastomes as a tool for phylogenomic analyses in the Caesalpinia group

The Caesalpinia group was subjected to several taxonomic classifications (Polhill and Vidal 1981; Lewis 2005; Gagnon et al. 2013, 2016) recently supported by molecular phylogenies (Lewis and Schrire 1995; Simpson and Miao 1997; Simpson and Lewis 2003; Nores et al. 2012; Gagnon et al. 2013, 2016). However, multiloci phylogenies with high sampling were not able to generate completely resolved topologies for the Caesalpinia group, particularly in terms of the backbone (Gagnon et al. 2016).

One problem to perform a traditional phylogenetic analysis is to find informative regions for solving the relationships among all hierarchical levels. Babineau et al. (2013) have tested different nuclear and plastidial markers for phylogenetic analysis in the Caesalpinia group and the most informative locus observed was the *rpl16*, which was used for previous phylogenetic analysis in the group (Gagnon et al. 2013, 2016). However, with different plastomes available, it is possible to find, by bioinformatic analysis, more informative regions for phylogenetic analyses in larger samples (Song et al. 2019; Souza et al. 2019b; Henriquez et al. 2020). In this work, we identified several regions with a high nucleotide diversity index (Pi) between the 26 species of Caesalpinia group. These regions varied from intergenic to genic and some of them proved to be useful in previous phylogenetic analysis of other groups, such as *clpP* and *ycf1* (Shawn et al. 2005; Dong et al. 2012, 2015). It seems that these regions are the most appropriate to solve the intergeneric relationships in the Caesalpinia group.

We also performed the nucleotide diversity analysis for *Cenostigma*, *Coulteria*, *Erythrostemon*, and *Libidibia* genera. Although the Caesalpinia group is monophyletic, the infrageneric relationships are not resolved and some species do not appear as monophyletic (Gagnon et al. 2016). Considering the 30 most polymorphic regions, several regions were recovered as good loci for molecular analysis in the group. Although the *trnK-UUU-rps16* region was the only region recovered in all five analysed datasets, alignment of this region presented several gaps and, due to its size, it may be difficult to find a conserved, internal region for primer design. Thus, it seems likely that this region may not be a good marker for phylogenetic and phylogeographic analysis in the Caesalpinia group. However, other loci showed similar or higher polymorphisms, but they vary among each dataset and should be investigated case by case (Table S5).

Plastome phylogenomic analysis has potential to generate robust topologies in both backbones and recent branches, as reported here for the Caesalpinia group and for other groups of plants (Thode et al. 2019, 2020; Zhai et al. 2019). This was confirmed by the

different phylogenetic approaches tested (BI, ML, and AAF). The Maximum Likelihood and the Bayesian Inference approaches are well known, but the Assembly and Alignment-Free (AAF) is a recently developed distance method that infers the phy- logenetic relationships based in k-mer analysis (Fan et al. 2015). AAF constructs phylogenies directly from unassembled genomic sequence data, bypassing both genome assembly and alignment. Using mathematical calculations, models of sequence evolution, and simulated sequencing of published genomes, AAF addresses both evolutionary and sampling issues caused by direct reconstruction, including homoplasy, sequencing errors, and incomplete sequencing coverage. Thus, it calculates the statistical properties of the pairwise distances between genomes, allowing it to optimize parameter selection and perform bootstrapping. Since we detected many gaps in the alignment, the AAF approach was useful to infer the phylogenetic relationship in the group and retrieved the most frequent topology with high support all but two nodes (Suppl. Fig. S2d).

We performed phylogenetic inferences in different plastome data partitions. All partitions generated well-supported trees, except the IRa matrix, which was excluded from the analysis. In fact, the IR genes are duplicated in the plastome, and since our plastome is inferred from 250 bp reads, we may be comparing paralogous gene copies, which might not reflect the evolution of the group (Manzanilla and Bruneau 2012). In general, the relationships observed here (Fig. 3) corroborate the topology proposed by Gagnon et al. (2016) with two main clades in the Caesalpinia group. Nevertheless, we observed a closer relationship for the *Coulteria*, *Tara, Paubrasilia*, and *Caesalpinia* genera, which are present in a monophyletic clade with high support in most analysis. This difference might be caused by our incomplete generic sampling, as we did not include species from *Denisophytum, Gelrebia*, and *Hultholia* genera. However, because the same taxa were placed in low support clades before (Gagnon et al. 2016), Clade I may be a lineage with more complex phylogenetic relationships. The plastome often shows different phylogenetic relationships when compared to nuclear loci (e.g., Koenen et al. 2020; Zhang et al. 2020) and discordance may be high even among gene trees based on whole-genome approaches (Copetti et al. 2017). Deeper phylogenetic analyses will be necessary to further elucidate the relationships within Clade I.

Although phylogenomic analysis increases the number of informative sites and allows the inference of reliable and well-supported trees, it seems that incongruences are frequent (Philippe et al. 2017). Thus, solely relying on traditional statistical support to infer phylogenomic relationships is not sufficient, since incongruences may also present high supports. For plastome phylogenomics, recent studies have indicated that plastid DNA does

not evolve as a single locus and different data partitions may show different topologies (Gonçalves et al. 2019; Walker et al. 2019; Thode et al. 2020). Our results corroborate those previous studies, and it may also explain the observed differences between the phylogenomic and the plastid multiloci phylogenetic approaches.

In all generated trees, we found a closer relationship between one accession of *Cenostigma pyramidale* and *C. microphyllum* than between *C. pyramidale* samples. In fact, both species are endemic to the Brazilian Caatinga, where some of their populations are in sympatry. One possible explanation for this result is the occurrence of interspecific hybridization between *Cenostigma* species. This hypothesis was proposed by Lewis (1995), based on the difficulty of separating out these species on morphology alone. If one accession of *C. pyramidale* is a hybrid individual and inherited the plastome sequence of a *C. microphyllum* progenitor, this could explain the closer relationship found between these accessions than between accessions of the same species and the similarity of their plastomes (99.9% of identity) even in the SSR regions (Table S5). Furthermore, time-calibrated phylogenies of this genus suggest that it is of relative recent origin, with no clear backbone relationships, despite using a sampling that includes 71% of the members of this genus, with multiple individuals per species across their geo- graphic range (Gagnon et al. 2016, 2019). Thus, the use of nuclear markers in a population-wide study is needed to test this hypothesis and fully understand this result. Altogether, our study reinforces the idea that whole plastomes are important tools to study evolution of plants, obtaining more robust phylogenies, and eventually highlighting conflicts that may indicate complex evolutionary trajectories that require deeper investigations (Zhai et al. 2019; Koenen et al 2020; Kim et al. 2020).

Conclusion

In summary, our study is the first that analyses the evolution of the plastomes in the Caesalpinia group. We added plastome information from four previously unreported genera (*Arquita, Caesalpinia, Paubrasilia,* and *Tara*), five new plastomes from species of four other genera (*Cenostigma microphyllum, Coulteria mollis, Erythrostemon hughesii, E. pannosus,* and *Libidibia ferrea*) and a second accession for four species (*Cenostigma pyramidale, E. calycinus, Guilandina bonduc,* and *L. coriaria*) allowing discussions at different taxonomic levels. In a comparative analysis of 26 plastomes, a pattern of strong conservatism was observed in the gene order among genera. A variation in length (11.4 kbp) was caused by increases in the intergenic region especially at LSC and SSC and the loss of four genes from the *ndh* complex in *E. gilliesii*. We presented a set of useful cpSSR markers for Leguminosae

species and demonstrated their use to assess genetic diversity of *C. microphyllum*. The species is endemic to the Caatinga domain and presented relatively high genetic diversity, showing both inter- and intrapopulation variation in the haplotype analysis, comparable to previous studies in larger geographic scales. We also provided a set of polymorphic regions for phylogenetic studies, both for the Caesalpinia group, and for *Cenostigma*, *Coulteria*, *Erythrostemon*, and *Libidibia* genera, to solve infrageneric relationships. The *ycf1* gene showed the highest polymorphic index (Pi) among the 26 Caesalpinia group species, suggesting its usefulness for future phylogenetic analyses. A plastome phylogeny including 15 of the 26 recognized genera corroborate the topology presented by Gagnon et al. (2016), but with better resolution, clarifying the relationship among *Coulteria*, *Tara*, *Caesalpinia*, and *Paubrasilia* genera, which was recovered by different matrices and different analyses with better support. Altogether, our data reinforce the idea that plastomes are useful options for solving complex phylogenies.

Funding Funding for this study was provided by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, PELD process 403770/2012-2, Universal process 426738/2018-7), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES code 001) and the Fundação de Amparo à Ciência e Tecnologia de Pernambuco (FACEPE, processes BIC-0846-2.02/17, BIC-0624- 2.02/18).

Compliance with ethical standards

Conflict of interest All the author declares no conflict of interest.

References

Abadi S, Azouri D, Pupko T, Mayrose I (2019) Model selection may not be a mandatory step for phylogeny reconstruction. Nature Commun 10:1–11. https://doi.org/10.1038/s41467-019-08822-w

Asaf S, Khan AL, Khan MA, Imran QM, Kang S, Al-Hosni K, Jeong EJ, Lee Ke, Lee I (2017) Comparative analysis of complete plastid genomes from wild soybean (*Glycine soja*) and nine other *Glycine* species. PLOS ONE 12:e0182281. https://doi.org/10.1371/journal.pone.0182281

Babineau, M., Gagnon, E., & Bruneau, A. (2013). Phylogenetic utility of 19 low copy nuclear genes in closely related genera and species of caesalpinioid legumes. S Afr J Bot 89:94–105. https://doi.org/10.1016/j.sajb.2013.06.018

Bai HR, Oyebanji O, Zhang R, Yi TS (2020) Plastid phylogenomic insights into the evolution of subfamily Dialioideae (Leguminosae). Plant Divers 42:1-8. https://doi.org/10.1016/j.pld.2020.06.008

Banerjee A, Stefanović S (2019) Caught in action: fine-scale plastome evolution in the parasitic plants of *Cuscuta* section Ceratophorae (Convolvulaceae). Plant Mol Biol 100:621–634. https://doi.org/10.1007/s11103-019-00884-0

Braukmann T, Kuzmina M, Stefanović S (2013) Plastid genome evolution across the genus *Cuscuta* (Convolvulaceae): two clades within subgenus *Grammica* exhibit extensive gene loss. J Exp Bot 64:977–989. https://doi.org/10.1093/jxb/ers391

Cai Z, Guisinger M, Kim H-G, Ruck E, Blazier JC, McMurtry V, Kuehl JV, Boore J, Jansen RK (2008) Extensive Reorganization of the Plastid Genome of *Trifolium subterraneum* (Fabaceae) Is Associated with Numerous Repeated Sequences and Novel DNA Insertions. J Mol Evol 67:696–704. https://doi.org/10.1007/s00239-008-9180-7

Cardoso DBOS, Pennington RT, De Queiroz LP, Boatwright JS, Van Wyk BE, Wojciechowski MF, Lavin M (2013) Reconstructing the deep-branching relationships of the papilionoid legumes. S Afr J Bot, 89:58-75. http://dx.doi.org/10.1016/j.sajb.2013.05.001

Cardoso D, de Queiroz LP, Pennington RT, de Lima HC, Fonty E, Wojciechowski MF, Lavin M (2012) Revisiting the phylogeny of papilionoid legumes: New insights from comprehensively sampled early-branching lineages. Am J Bot 99:1991–2013. https://doi.org/10.3732/ajb.1200380

Cauz-Santos LA, da Costa ZP, Callot C, Cauet S, Zucchi MI, Bergès H, van den Berg C, Vieira MLC (2020) A repertory of rearrangements and the loss of an inverted repeat region in *Passiflora* chloroplast genomes. Genome Biol Evol 12:1841-1857. https://doi.org/10.1093/gbe/evaa155

Copetti D, Búrquez A, Bustamante E, Charboneau JLM, Childs KL, Eguiarte LE, Lee S, Liu TL, McMahon MM, Whiteman NK, Wing RA, Wojciechowski MF, Sanderson MJ (2017) Extensive gene tree discordance and hemiplasy shaped the genomes of North American columnar cacti. Proc Natl Acad Sci U S A 114:12003-12008. https://doi.org/10.1073/pnas.1706367114

Dierckxsens N, Mardulyn P, Smits G, (2017) NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res 45:e18. https://doi.org/10.1093/nar/gkw955

Dong W, Liu J, Yu J, et al (2012) Highly variable chloroplast markers for evaluating plant phylogeny at low taxonomic levels and for DNA barcoding. PLoS ONE 7:e35071. https://doi.org/10.1371/journal.pone.0035071

Dong W, Xu C, Li C, et al (2015) ycf1, the most promising plastid DNA barcode of land plants. Sci Rep 5:8348. https://doi.org/10.1038/srep08348

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull 19:11–15.

Drummond AJ, and Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7:214. https://doi.org/10.1186/1471-2148-7-214

Dugas DV, Hernandez D, Koenen EJM, et al (2015) Mimosoid legume plastome evolution: IR expansion, tandem repeat expansions and accelerated rate of evolution in clpP. Sci Rep 5:16958. https://doi.org/10.1038/srep16958

Duvall MR, Burke SV, Clark DC (2019) Plastome phylogenomics of Poaceae: alternate topologies depend on alignment gaps. Bot J Linn Soc 192:9–20. https://doi.org/10.1093/botlinnean/boz060

Ebert D, Peakall R (2009) Chloroplast simple sequence repeats (cpSSRs): technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant species. Mol Ecol Resour 9:673–690. https://doi.org/10.1111/j.1755-0998.2008.02319.x

Eliades NGH, Eliades DG (2009) Haplotype Analysis: sofware for analysis of haplotype data. Distributed by the authors, Forest Genetics and Forest Tree Breeding, Georg-August University Goettingen, Germany. Available at http://www.uni-goettingen.de/en/134935.html

Ellengreen H (2000) Microsatellite mutations in the germline: implications for evolutionary inference. Trends Genet, 16:551-558. https://doi.org/10.1016/S0168-9525(00)02139-9

Fan H, Ives AR, Surget-Groba Y, Cannon CH (2015) An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. BMC genomics 16:522. https://doi.org/10.1186/s12864-015-1647-5

Ferreira ME, Grattapaglia D (1995) Introdução ao uso de marcadores moleculares em análise genética. EMBRAPA-CENARGEN, Brasília.

Gagnon E, Bruneau A, Hughes CE, de Queiroz LP, Lewis GP (2016) A new generic system for the pantropical Caesalpinia group (Leguminosae). PhytoKeys 71:1–160. https://doi.org/10.3897/phytokeys.71.9203

Gagnon E, Lewis GP, Sotuyo JS, Hughes CE, Bruneau A (2013) A molecular phylogeny of Caesalpinia sensu lato: Increased sampling reveals new insights and more genera than expected. S Afr J Bot 89:111–127. https://doi.org/10.1016/j.sajb.2013.07.027

Gagnon E, Ringelberg JJ, Bruneau A, Lewis GP, Hughes CE (2019) Global Succulent Biome phylogenetic conservatism across the pantropical Caesalpinia Group (Leguminosae). New Phytol 222:1994–2008. https://doi.org/10.1111/nph.15633

Gernhard T (2008) New analytic results for speciation times in neutralmodels. Bull Math Biol 70:1082–1097. https://doi.org/10.1007/s11538-007-9291-0

Gonçalves DJP, Simpson BB, Ortiz EM, Shimizu GH, Jansen RK (2019) Incongruence between gene trees and species trees and phylogenetic signal variation in plastid genes. Mol Phylogenetics Evol 138:219-232. https://doi.org/10.1016/j.ympev.2019.05.022

Gonçalves DJP, Jansen RK, Ruhlman TA, Mandel JR (2020) Under the rug: Abandoning persistent misconceptions that obfuscate organelle evolution. Mol Phylogenetics Evol 151:106903 https://doi.org/10.1016/j.ympev.2020.106903

Gonçalves-Oliveira RC, Wöhrmann T, Benko-Iseppon AM, Krapp F, Alves M, Wanderley MGL, Weising K (2017) Population genetic structure of the rock outcrop species *Encholirium spectabile* (Bromeliaceae): The role of pollination vs. seed dispersal and evolutionary implications. Am J Bot 104:868–878. https://doi.org/10.3732/ajb.1600410

He J, Yao M, Lyu RD, Lin LL, Liu HJ, Pei LY, Yang SX, Xie L, Cheng, J (2019) Structural variation of the complete chloroplast genome and plastid phylogenomics of the genus *Asteropyrum* (Ranunculaceae). Sci Rep 9: 15285. https://doi.org/10.1038/s41598-019-51601-2

Henriquez CL, Ahmed I, Carlsen MM, Zuluaga A, Croat TB, McKain MR (2020) Molecular evolution of chloroplast genomes in Monsteroideae (Araceae). Planta 251:1-16. https://doi.org/10.1007/s00425-020-03365-7

Jansen RK, Ruhlman TA (2012) Plastid Genomes of Seed Plants. In: Bock R, Knoop V (eds) Genomics of Chloroplasts and Mitochondria. Springer Netherlands, Dordrecht, pp 103–126

Jansen RK, Wojciechowski MF, Sanniyasi E, Lee S-B, Daniell Henry (2008) Complete plastid genome sequence of the chickpea (*Cicer arietinum*) and the phylogenetic

distribution of rps12 and clpP intron losses among legumes (Leguminosae). Mol Phylogenetics Evol 48:1204–1217. https://doi.org/10.1016/j.ympev.2008.06.013

Ji Y, Yang L, Chase MW, Liu C, Yang Z, Yang J, Yang JB, Yi TS (2019) Plastome phylogenomics, biogeography, and clade diversification of *Paris* (Melanthiaceae). BMC Plant Biol 19: 15285. https://doi.org/10.1186/s12870-019-2147-6

Katoh K, Rozewicki J, Yamada KD (2019) MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform 20:1160-1166. https://doi.org/10.1093/bib/bbx108

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics, 28:1647–1649. https://doi.org/10.1093/bioinformatics/bts199

Kim Y-K, Jo S, Cheon S-H, Kwak M, Kim Y-D, Kim K-J (2020) Plastome evolution and phylogeny of subtribe Aeridinae (Vandeae, Orchidaceae). Mol Phylogenetics Evol 144:106721. https://doi.org/10.1016/j.ympev.2019.106721

Koenen EJM, Ojeda DI, Steeves R, Migliore J, Bakker FT, Wieringa JJ, Kidner C, Hardy OJ, Pennington RT, Bruneau A, Hughes CE (2020) Large-scale genomic sequence data resolve the deepest divergences in the legume phylogeny and support a near-simultaneous evolutionary origin of all six subfamilies. New Phytol 225:1355–1369. https://doi.org/10.1111/nph.16290

Lemos RPM, Matielo CBD, Beise DC, Da Rosa VG, Sarzi DS, Roesch LFW, Stefenon VM (2018) Characterization of plastidial and nuclear ssr markers for understanding invasion histories and genetic diversity of schinus molle L. Biology 7:43. https://doi.org/10.3390/biology7030043

Lewis GP (1995) Systematic studies in neotropical '*Caesalpinia* L.' (Leguminosae: Caesalpinioideae), including a revision of the 'Poinchianella-Erythrostemon' group. Thesis. University of St Andrews.

Lewis GP (2005) Tribe Caesalpinieae. In: Lewis G, Schrire B, Mackinder B, Lock M (Eds) Legumes of the World. Kew Royal Botanic Gardens, Richmond, pp 127–159.

Lewis GP, Schrire BD (1995) A reappraisal of the Caesalpinia group Caesalpinioideae: Caesalpinieae) using phylogenetic analysis. In: Crisp MD, Doyle JJ (Eds) Advances in Legume Systematics: Part 7, Phylogeny. Kew Royal Botanic Gardens, Richmond, pp 41–52

Lira CF, Cardoso SRS, Ferreira PCG, Cardoso MA, Provan J (2003) Long-term population isolation in the endangered tropical tree species *Caesalpinia echinata* Lam. revealed by chloroplast microsatellites. Mol Ecol 12:3219–3225. https://doi.org/10.1046/j.1365-294X.2003.01991.x

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

López-Villalobos A, Eckert CG (2018) Consequences of multiple mating-system shifts for population and range-wide genetic structure in a coastal dune plant. Mol Ecol 27:675-693. https://doi.org/10.1111/mec.14484

Manzanilla V, Bruneau A (2012) Phylogeny reconstruction in the Caesalpinieae grade (Leguminosae) based on duplicated copies of the sucrose synthase gene and plastid markers. Mol Phylogenetics Evol 65:149-162. http://dx.doi.org/10.1016/j.ympev.2012.05.035

Mata-Sucre Y, Costa L, Gagnon E, et al (2020a) Revisiting the cytomolecular evolution of the Caesalpinia group (Leguminosae): a broad sampling reveals new correlations between cytogenetic and environmental variables. Plant Syst Evol 306:48. https://doi.org/10.1007/s00606-020-01674-8

Mata-Sucre Y, Sader M, Van-Lume B, Gagnon E, Pedrosa-Harand A, Leitch IJ, Lewis GP, Souza G (2020b). How diverse is heterochromatin in the Caesalpinia group? Cytogenomic characterization of Erythrostemon hughesii Gagnon & G.P. Lewis (Leguminosae: Caesalpinioideae). Planta 252:49 https://doi.org/10.1007/s00425-020-03453-8

Mayer C (2006-2010) Phobos 3.3.11. <http://www.rub.de/ecoevo/cm/cm_phobos.htm>.

Miller MA, Pfeiffer W, Schwartz T **(**2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Proceedings of the Gateway Computing Environments Workshop (GCE). http://dx.doi.org/10.1109/GCE.2010.5676129

Mota MR, Pinheiro F, Leal BSS, Wendt T, Palma-Silva C (2019)The role of hybridization and introgression in maintaining species integrity and cohesion in naturally isolated inselberg bromeliad populations. Plant Biol 21:122-132. https://doi.org/10.1111/plb.12909

Nores MJ, Simpson BB, Hick P, Anton AM, Fortunato RH (2012) The phylogenetic relationships of four monospecific caesalpinioids (Leguminosae) endemic to southern South America. Taxon 61:790–802. https://doi.org/10.1002/tax.614006

Philippe H, de Vienne DM, Ranwez V, Roure B, Baurain D, Delsuc F (2017) Pitfalls in supermatrix phylogenomics. Eur J Taxon 283:1–25. http://dx.doi.org/10.5852/ejt.2017.283

Polhill RM, Vidal JE (1981) Caesalpinieae. In: Polhill RM, Raven PH (Eds) Advances in Legume Systematics, Part 1. Kew Royal Botanic Gardens, Richmond, 81–95.

Powell W, Machray GC, Provan J (1996) Polymorphism revealed by simple sequence repeats. Trends Plant Sci 1:215–222. https://doi.org/10.1016/1360-1385(96)86898-1

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

Rito K F, Arroyo-Rodríguez V, Queiroz RT, Leal IR, Tabarelli M (2017) Precipitation mediates the effect of human disturbance on the Brazilian Caatinga vegetation. J Ecol, 105:828-838. https://doi.org/10.1111/1365-2745.12712

Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins S, Sánchez-Gracia A (2017) DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. Mol Biol Evol, 34:3299-3302. https://doi.org/10.1093/molbev/msx248

Samarakoon T, Wang SY, Alford MH (2013) Enhancing PCR amplification of DNA from recalcitrant plant specimens using a trehalose-based additive. Appl Plant Sci 1: 1200236. https://doi.org/10.3732/apps.1200236

Schuelke M, (2000) An economic method for the fluorescent labeling of PCR fragments: A poor man's approach to genotyping for research and high-throughput diagnostics. Nat Biotechnol, 18:233-234. https://doi.org/10.1038/72708

Schwarz EN, Ruhlman TA, Weng M-L, Khiyami MA, Sabir JSM, Hajarah NH, Alharbi NS, Rabah SO, Jansen RK (2017) Plastome-Wide Nucleotide Substitution Rates Reveal Accelerated Rates in Papilionoideae and Correlations with Genome Features Across Legume Subfamilies. J Mol Evol 84:187–203. https://doi.org/10.1007/s00239-017-9792-x

Shaw J, Lickey EB, Beck JT, Farmer SB, Liu W, Miller J, Siripun KC, Winder CT, Schilling EE, Small RL (2005) The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. Am J Bot 92:142–166. https://doi.org/10.3732/ajb.92.1.142

Simpson BB, Miao BM (1997) The circumscription of *Hoffmannseggia* (Fabaceae, Caesalpinioideae, Caesalpinieae) and its allies using morphological and cpDNA restriction site data. Plant Syst Evol 205:157–178. https://doi.org/10.1007/BF01464402

Simpson BB, Lewis GP (2003) New combinations in *Pomaria* (Caesalpinioideae: Leguminosae). Kew Bulletin 58:175–184. https://doi.org/10.2307/4119360

Solórzano S, Chincoya DA, Sanchez-Flores A, Estrada K, Díaz-Velásquez CE, González-Rodríguez A, Vaca-Paniagua F, Dávila P, Arias S (2019) De Novo Assembly Discovered Novel Structures in Genome of Plastids and Revealed Divergent Inverted Repeats in *Mammillaria* (Cactaceae, Caryophyllales). Plants 8:392. https://doi.org/10.3390/plants8100392

Song Y, Zhang Y, Xu J, Li W, Li MF (2019) Characterization of the complete chloroplast genome sequence of *Dalbergia* species and its phylogenetic implications. Sci Rep 9:1-10. https://doi.org/10.1038/s41598-019-56727-x

Souza UJB de, Nunes R, Targueta CP, Diniz-Filho JAF, Telles MPC (2019b) The complete chloroplast genome of *Stryphnodendron adstringens* (Leguminosae - Caesalpinioideae): comparative analysis with related Mimosoid species. Sci Rep 9: 14206. https://doi.org/10.1038/s41598-019-50620-3

Souza G, Costa L, Guignard MS, Van-Lume B, Pellicer J, Gagnon E, Leitch IJ, Lewis JP (2019a) Do tropical plants have smaller genomes? Correlation between genome size and climatic variables in the Caesalpinia Group (Caesalpinioideae, Leguminosae). Perspect Plant Ecol Evol Syst 38:13-23. https://doi.org/10.1016/j.ppees.2019.03.002

Su Y, Huang L, Wang Z, Wang T (2018) Comparative chloroplast genomics between the invasive weed *Mikania micrantha* and its indigenous congener *Mikania cordata*: Structure variation, identification of highly divergent regions, divergence time estimation, and phylogenetic analysis. Mol Phylogenetics Evol 126:181–195. https://doi.org/10.1016/j.ympev.2018.04.015

Thode VA, Sanmartín I, Lohmann LG (2019) Contrasting patterns of diversification between Amazonian and Atlantic forest clades of Neotropical lianas (*Amphilophium*, Bignonieae) inferred from plastid genomic data. Mol Phylogenet Evol 133:92–106. https://doi.org/10.1016/j.ympev.2018.12.021

Thode VA, Lohmann LG, Sanmartín I (2020) Evaluating character partitioning and molecular models in plastid phylogenomics at low taxonomic levels: A case study using *Amphilophium* (Bignonieae, Bignoniaceae). J Syst Evol 58:1071-1089. https://doi.org/10.1111/jse.12579

Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R and Greiner S (2017) GeSeq – versatile and accurate annotation of organelle genomes. Nucleic Acids Res 45:W6-W11. https://doi.org/10.1093/nar/gkx391

Untergasser A, Nijveen N, Rao X, Bisseling T, Geurts R, Leunissen JAM (2007) Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res, 35:W71-W74. https://doi.org/10.1093/nar/gkm306

Van-Lume B, Esposito T, Diniz-Filho JAF, Gagnon E, Lewis G P, Souza G (2017). Heterochromatic and cytomolecular diversification in the Caesalpinia group (Leguminosae): Relationships between phylogenetic and cytogeographical data. Perspect Plant Ecol Evol Syst 29:51-63. https://doi.org/10.1016/j.ppees.2017.11.004

Van-Lume B, Mata-Sucre Y, Báez M, Ribeiro T, Huettel B, Gagnon E, Leitch IJ, Pedrosa-Harand A, Lewis GP, Souza G (2019) Evolutionary convergence or homology? Comparative cytogenomics of Caesalpinia group species (Leguminosae) reveals diversification in the pericentromeric heterochromatic composition. Planta 250:2173–2186. https://doi.org/10.1007/s00425-019-03287-z

Walker JF, Jansen RK, Zanis MJ, Emery NC (2015) Sources of inversion variation in the small single copy (SSC) region of chloroplast genomes. Am J Bot 102:1751–1752. https://doi.org/10.3732/ajb.1500299

Walker JF, Walker-Hale N, Vargas OM, Larson DA, Stull GW (2019) Characterizing gene tree conflict in plastome-inferred phylogenies. PeerJ 7:e7747. http://doi.org/10.7717/peerj.7747

Wang J, Jiang Y, Qian J, Xu L, Duan B (2020) Characterization of the complete chloroplast genome of *Caesalpinia sappan* L. (Leguminosae). Mitochondrial DNA Part B 5:1642–1643. https://doi.org/10.1080/23802359.2020.1745714

Wang Z, Zeng Y, Zhang Z, Sheng S, Tian J, Wu R, Pang X (2017a) Phylogeography Study of the Siberian Apricot (*Prunus sibirica* L.) in Northern China Assessed by Chloroplast Microsatellite and DNA Makers. Front Plant Sci 8:1989. https://doi.org/10.3389/fpls.2017.01989

Wang Y-H, Qu X-J, Chen S-Y, Li D-Z, Yi T-S (2017b) Plastomes of Mimosoideae: structural and size variation, sequence divergence, and phylogenetic implication. Tree Genet Genomes 13:41. https://doi.org/10.1007/s11295-017-1124-1

Weising K, Gardner RC (1999) A set of conserved PCR primers for the analysis of simple sequence repeat polymorphisms in chloroplast genomes of dicotyledonous angiosperms. Genome 42:9-19. https://doi.org/10.1139/g98-104

Weising K, Nybom H, Pfenninger M, Wolf K, Kahl G (2005) DNA fingerprinting in plants: principles, methods, and applications. CRC Press, Boca Raton.

Wheeler GL, Dorman HE, Buchanan A, Challagundla L, Wallace LE (2014) A Review of the Prevalence, Utility, and Caveats of Using Chloroplast Simple Sequence Repeats for Studies of Plant Biology. Appl Plant Sci 2:1400059. https://doi.org/10.3732/apps.1400059

Zane L, Bargelloni L, Patarnello T (2002) Strategies for microsatellite isolation: a review. Mol Eco 11:1–16. https://doi.org/10.1046/j. 0962-1083.2001.01418.x

Zhai W, Duan X, Zhang R, Guo C, Li L, Xu G, Shan H, Kong H, Ren Y (2019) Chloroplast genomic data provide new and robust insights into the phylogeny and evolution of the Ranunculaceae. Mol Phylogenetics Evol 135:12–21. https://doi.org/10.1016/j.ympev.2019.02.024

Zhang R, Wang Y-H, Jin J-J, Stull GW, Bruneau A, Cardoso D, De Queiroz LP, Moore MJ, Zhang S-D, Chen S-Y, Wang J, Li D-Z, Yi T-S (2020) Exploration of Plastid Phylogenomic Conflict Yields New Insights into the Deep Relationships of Leguminosae. Syst Biol 69:613–622. https://doi.org/10.1093/sysbio/syaa013

Zhang T, Zeng C X, Yang J B, Li H T, Li D Z (2016) Fifteen novel universal primer pairs for sequencing whole chloroplast genomes and a primer pair for nuclear ribosomal DNAs. J Syst Evol, 54:219-227. https://doi.org/10.1111/jse.12197

Table 1 Summary of the chloroplast genome characterization within the Caesalpinia group species. Coverage represents the proportion of reads mapped to the plastome assembly. The geographical distribution indicated is the genus distribution. Columns: Genome size (bp), GC content (%), LSC (large single copy region, bp), SSC (small single copy region, bp), IR (inverted repeat, bp)

Species	Geographical distribution	Clade	Number of reads (coverage)	Size (bp)	GC (%)	LSC (bp)	SSC (bp)	IR – each (bp)	Accession number	Source
Arquita mimosifolia (Griseb.) E.Gagnon, G.P.Lewis & C.E.Hughes	Andes	II	53,698 (86×)	155,884	36.7	85,664	18,184	26,018	MZ441390	This study
Balsamocarpon brevifolium Clos	Andes	II	NA	156,953	36.6	86,592	18,197	26,082	MH252075.1	Jimenez et al. unpublished
Biancaea sappan (L.) Tod.	Asia	I	NA	160,192	36.0	89,726	18,358	26,054	MN933929	Wang et al. 2020
Caesalpinia pulcherrima (L.)	Mesoamerica (cultivated)	I	87,510 (139×)	157,480	36.2	87,050	17,192	26,619	MZ441391	This study

SW.

Cenostigma microphyllum (Mart. ex G.Don) E.Gagnon & G.P.Lewis	Northeastern Brazil	II	49,192 (77×)	159,456	36.4	88,991	18,357	26,054	MZ441392	This study
Cenostigma pyramidale (Tul) E.Gagnon & G.P.Lewis	Northeastern Brazil	II	14,469,232 (331.8×)	159,795	36.4	89,351	18,354	26,045	MN709843.1	Zhang et al. 2020
Cenostigma pyramidale (Tul) E.Gagnon & G.P.Lewis 2	Northeastern Brazil	II	212,544 (217×)	159,424	36.4	88,999	18,357	26,034	MZ441393	This study
Coulteria mollis Kunth	Mesoamerica	I	747,176 (760×)	160,553	36.3	89,892	18,511	26,075	MZ441394	This study

Coulteria platyloba(S.Watson) N.Zamora	Mesoamerica	I	14,985,744 (424.2×)	160,677	36.3	90,034	18,493	26,075	MN709822	Zhang et al. 2020
Erytrhostemon calycinus (Benth.) L.P.Queiroz	Northeastern Brazil	II	13,775,276 (1098.9×)	156,190	36.6	86,955	18,077	26,079	MN709844.1	Zhang et al. 2020
Erytrhostemon calycinus (Benth.) L.P.Queiroz 2	Northeastern Brazil	II	452,358 (492×)	156,079	36.6	85,864	18,057	26,079	MZ441395	This study
Erythrostemon gilliesii (Hook.) Link et al.	Andes	II	NA	149,235	36.8	81,683	15,424	26,064	NC_045296.1	Koenen et al. 2020
Erytrhostemon hughesii (G.P.Lewis) Gagnon & G.P.Lewis	Mesoamerica	II	81,210 (130×)	156,366	36.8	86,071	18,207	26,044	MZ441396	This study

Erytrhostemon pannosus (Brandegee) E.Gagnon & G.P.Lewis	Mesoamerica	II	358,606 (389×)	154,726	37.0	84,581	18,105	26,020	MZ441397	This study
Guilandina bonduc L.	Cosmopolitan	I	111,388 (178×)	156,843	36.7	86,809	17,852	26,091	MZ441398	This study
Guilandina bonduc L. 2	Cosmopolitan	I	15,047,772 (377.2×)	156,847	36.7	86,807	17,858	26,091	MN709864	Zhang et al. 2020
Haematoxylum brasiletto H.Karst.	Mesoamerica	I	14,384,874 (449.3×)	157,616	36.7	87,334	18,194	26,044	KJ468097.1	Zhang et al. 2020
Libidibia coriaria (Jacq.) Schltdl.	North of South America	II	NA	158,045	36.5	87,602	18,147	26,148	NC_026677	Unpublished
Libidibia coriaria (Jacq.) Schltdl. 2	North of South America	II	526,068 (542×)	158,122	36.4	87,751	18,129	26,121	MZ441399	This study

Libidibia ferrea (Mart. ex Tul.) L.P.Queiroz	Northeastern Brazil	II	260,840 (413×)	158,488	36.4	88,117	18,077	26,147	MZ441400	This study
Mezoneuron cucullatum (Roxb.) Wight & Arn.	Asia	I	15,217,436 (685.5×)	158,487	36.3	87,862	18,075	26,275	MN709870.1	Zhang et al. 2020
Mezoneuron cucullatum (Roxb.) Wight & Arn. 2	Asia	I	1×[a]	158,357	36.4	87,663	18,106	26,294	KU569489	Zhang et al. 2015
Moullava spicata (Dalzell) Nicolson	Asia	I	15,107,462 (486.0×)	159,260	36.0	88,346	17,856	26,529	MN709867.1	Zhang et al. 2020
Paubrasilia echinata (Lam.) Gagnon, H.C.Lima & G.P.Lewis	Atlantic Forest (cultivated)	I	20,716 (32×)	158,713	36.5	88,293	18,354	26,033	MZ441401	This study
Pterolobium	Asia	I	15,079,290	158,801	36.2	88,832	17,761	26,104	MN709875.1	Zhang et al.

punctatum Hemsl. ex Forb. & Hemsl.			(474.9×)							2020
Tara cacalaco (Humb. & Bonpl.) Molinari & Sánchez	Mesoamerica	I	705,592 (721×)	159,603	36.3	89,147	18,506	25,975	MZ441402	This study

Table 2 List of genes identified in the Caesalpinia group plastomes and described functions. Genes in bold are present in the IR and duplicated in the genome

Gene function	Gene group	Gene name
Self-replication	Ribossomal RNA	rrn4.5, rrn5, rrn16, rrn23
	Transfer RNA	**trnA-UGC**, trnC-GCA, trnD-GUC, trnE-UUC, trnF-GAA, trnfM-CAU, trnG-GCC, trnG-UCC, trnH-GUG, **trnI-CAU**, **trnI-GAU**, trnK-UUU, **trnL-CAA**, trnL-UAA, trnL-UAG, trnM-CAU, **trnN-GUU**, trnP-UGG, trnQ-UUG, **trnR-ACG**, trnR-UCU, trnS-GCU, trnS-GGA, trnS-UGA**,** trnT-GGU, trnT-UGU, **trnV-GAC**, trnV-UAC, trnW-CCA, trnY-GUA
	Small subunit of ribosome	rps2, rps3, rps4, **rps7**, rps8, rps11, **rps12**, rps14, rps15, rps16, rps18, rps19[a]
	Large subunit of ribosome	**rpl2**, rpl14, rpl16, rpl20, **rpl23**, rpl32, rpl33, rpl36
	RNA polymerase subunits	rpoA, rpoB, rpoC1, rpoC2
Photosynthesis	Subunits of NADH dehydrogenase	ndhA, **ndhB**, ndhC[b], ndhD, ndhE, ndhF[b], ndhG, ndhH, ndhI, ndhJ[b], ndhK[b]
	Subunits of photosystem I	psaA, psaB, psaC, psaI, psaJ
	Subunits of photosystem II	psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ
	Subunits of cytochrome b/f complex	petA, petB, petD, petG, petL, petN
	Subunit of ATP synthase	atpA, atpB, atpE, atpF, atpH, atpI
Others	Large subunit of RUBISCO	RbcL
	Maturase	MatK
	Envelop membrane	CemA

protein	
C-type cytochrome synthesis	ccsA
ORFs	ycf1[a], ycf2, ycf3, ycf4
Acetyl-CoA carboxylase	accD
ATP-dependent protease	clpP

[a]Genes presenting one full and one truncated copies in the borders of IR

[b]Genes are not present in *Erythrostemon gilliesii* plastome

Table 3 *In silico* transferability of CmCPSSR loci to twenty-one Leguminosae species of different subfamilies

Subfamily	Species	Loci with similarity in annealing region >90%	SSR motif ≥ 10		SSR motif ≥ 6	
			Loci with a SSR	Loci transferred	Loci with a SSR	Loci transferred
Caesalpinioideae	*Arquita mimosifolia*	15	9	9	16	15
	Balsamocarpon brevifolium	15	8	7	16	15
	Biancaea sappan	16	8	8	16	16
	Caesalpinia pulcherrima	14	12	11	16	14
	Coulteria platyloba	16	9	9	16	16
	Erythrostemon huguesii	16	8	8	16	16
	Guilandina bonduc	15	9	8	16	15
	Haematoxylum brasiletto	15	5	5	15	15
	Libidibia ferrea	16	9	9	16	16
	Mezoneuron cucullatum	15	11	10	16	15
	Moullava spicata	16	13	13	16	16

	Paubrasilia echinata	16	7	7	15	15
	Pterolobium punctatum	16	11	11	16	16
	Tara cacalaco	14	7	5	16	14
Cercidoideae	*Schnella tricosepala*	6	7	2	15	6
Detarioideae	*Tamarindus indica*	11	3	2	15	11
Dialioideae	*Zenia insignis*	10	11	7	16	10
Duparquetioideae	*Duparquetia orchidaceae*	12	5	3	15	11
Papilionoidae	*Vigna unguiculata*	5	2	0	14	5

Table 4 Intraspecific polymorphism of CmCPSSR loci by *in vitro* (*Cenostigma microphyllum* and *Trischidium molle*) and *in silico* (remaining species) screening. Ta – Annealing temperature for loci analysed *in vitro*

	Cenostigma microphyllum		*Trischidium molle*		*Cenostigma pyramidale*	*Erythrostemon gilliesii*	*Guilandina bonduc*	*Libidibia coriaria*	*Mezoneuron cucullatum*
Loci	Ta (ºC)	Allele sizes (bp)	Ta (ºC)	Allele sizes (pb)	Allele sizes (bp)	Allele sizes (bp)	Allele sizes (bp)	Allele sizes (bp)	Allele sizes (bp)
CmCPSSR1	-	276	-	-	275, 276	256	296, 297	269	287
CmCPSSR2	-	220	-	-	220	221	218	215	226
CmCPSSR3	-	296	-	-	296, 297	309	306	302	305
CmCPSSR4	60	365, 366	52	366, 378, 379, 384	352	349	323	344	358
CmCPSSR5	-	423	-	-	420, 423	419	422, 424	419, 449	420
CmCPSSR6	60	390	56	390	373	373	373	373	373
CmCPSSR7	60	189, 190	60	189, 190, 191	177	173	175	177	178,179
CmCPSSR8	60	428, 429	52	414, 415	414	414	415	394, 417	416
CmCPSSR9	-	430	-	-	429, 430	424	423	426, 428	423

CmCPSSR10	-	246	-	-	246, 250	249	267	245, 251	256
CmCPSSR11	60	158	50	162, 390, 391	144, 145	145	149	140, 417	148
CmCPSSR12	58	125, 127, 128	50	128, 129	111, 113	114	112	113, 417	117, 118
CmCPSSR13	60	186	65	180	172	172	172	172	178
CmCPSSR14	60	209	60	164	196	189	195	195	195
CmCPSSR15	60	229, 230	55	259	210, 214	221	205	218	197
CmCPSSR16	60	203, 204	-	-	189, 190	191	194	199,200	192

Table 5 Haplotypic diversity of *Cenostigma microphyllum* accessed with four cpSSR loci in the eight plots in Catimbau National Park (*n* = 99). N, Sample size; A, no. of haplotypes; P, private haplotypes; Ne, effective haplotypes; Rh, haplotypic richness; He, genetic diversity

Population	N	A	P	Ne	Rh	He
P08	6	2	1	1.385	1.000	0.333
R12	9	1	0	1.000	0.000	0.000
R15	13	1	0	1.000	0.000	0.000
P21	11	1	0	1.000	0.000	0.000
P22	19	3	2	1.241	0.632	0.205
R23	15	3	2	1.316	0.800	0.257
P27	13	3	2	1.374	0.923	0.295
P28	13	3	1	1.857	1.339	0.500
Mean	**12.375**	**2.125**	**1.000**	**1.272**	**0.587**	**0.199**

Fig. 1 Circular map of the Caesalpinia Group plastomes generate by OGDrawn. Color of each gene correspond to its functional group. The *ndh*C, *ndh*F, *ndh*J and *ndh*K are not present in *Erythrostemon gilliesii*

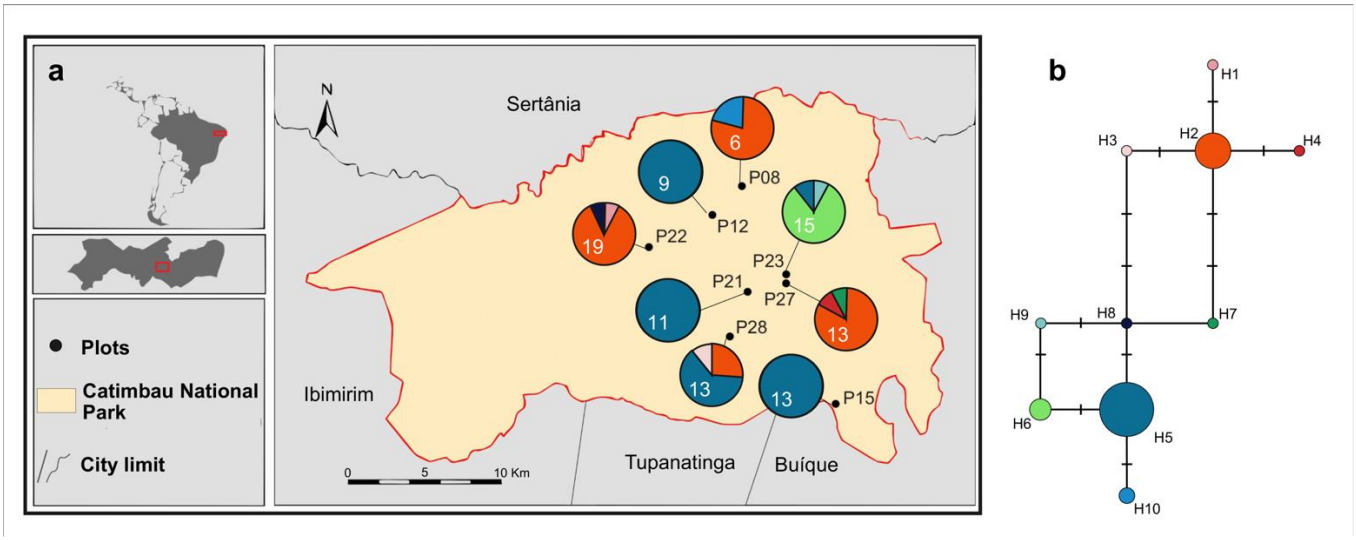

Fig. 2 Distribution of haplotypes generated with four cpSSR loci from *Cenostigma microphyllum* in eight plots of Catimbau National Park. **a** Haplotype frequencies in each of the sample plots. **b** Haplotype network. Numbers inside the charts are the number of individuals sampled in each population

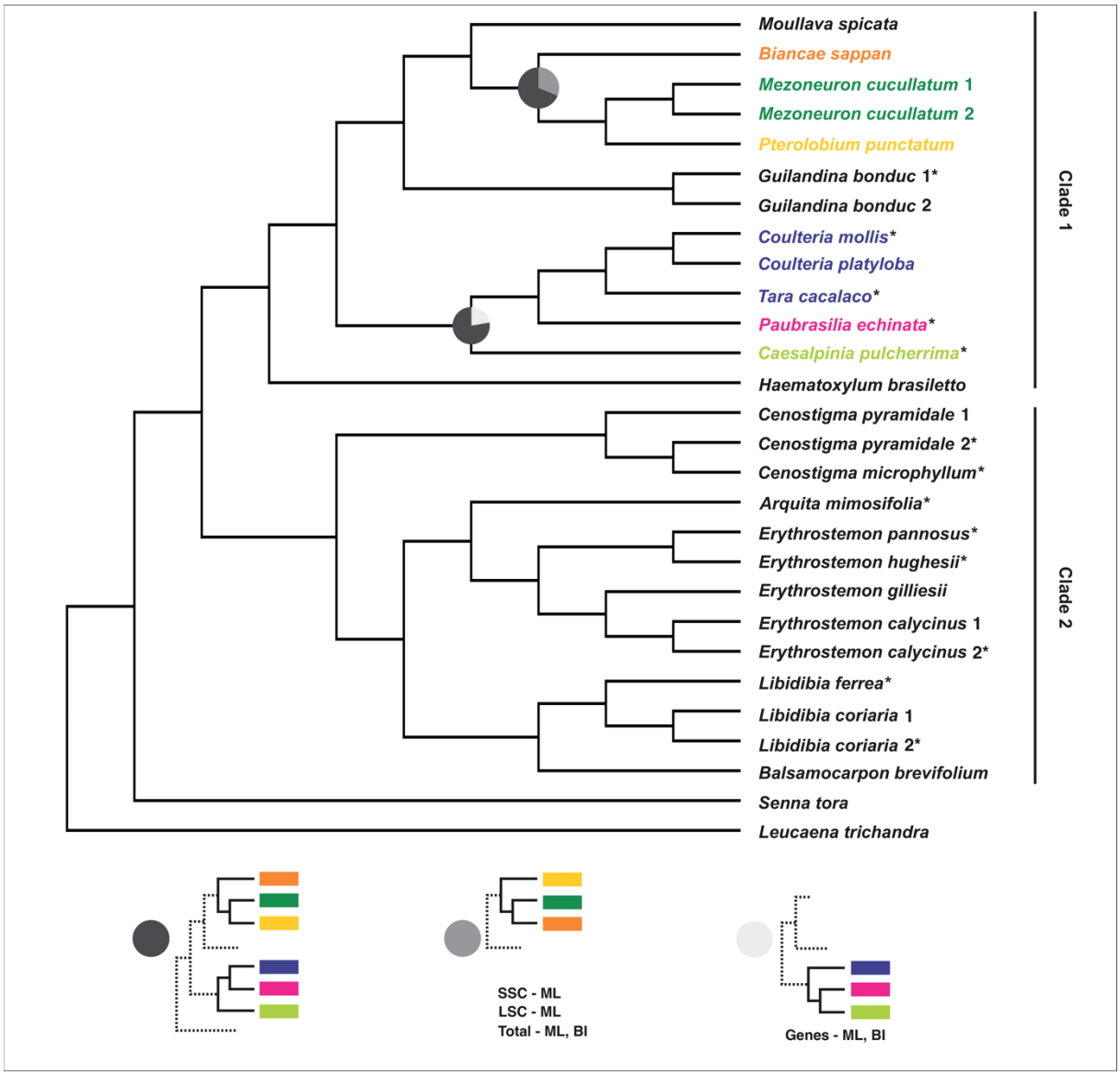

Fig. 3 The most frequently recovered topology of the nine phylogenetic trees analysed in this study for the Caesalpinia group with the total plastome alignment of 28 taxa with outgroup. The chart graphics indicate in how many trees the clade was recovered. Each colour is a different topology observed in different plastome data partition. Clades without charts were recovered in all analysis. Support values varied between the different matrix mainly on the conflicting topologies (60% ≤ BS ≤ 100% and 0.92 ≤ PP ≤ 1.00). Leucaena trichandara and Senna tora were used as outgroup. To further information see Figure S1

*Species which plastome was assembled in this study

Table S1 Geographic coordinates of the *Cenostigma microphyllum* plots analysed in the Catimbau National Park in Pernambuco, with number of individuals collected at each site

Cenostigma microphyllum			
Plots	Latitude	Longitude	Number of individuals
P08	8° 27.064'S	37° 18.086' W	6
P12	8° 27.663'S	37° 18.347'W	9
P15	8° 34.672’S	37° 14.833’W	13
P21	8° 31.241'S	37° 17.779'W	11
P22	8° 29.167'S	37° 20.418'W	19
P23	8° 31.020' S	37° 18.712'W	15
P27	8° 30.634'S	37° 16.520'W	13
P28	8° 32.163'S	37° 18.580'W	13

Table S2 Geographic coordinates from *Trischidium molle* individuals from different locations from Northeastern Brazil obtained from herbarium vouchers

Trischidium molle				
State	City	Herbarium	Latitude	Longitude
Bahia	Canudos	PEUFR	10°1’58”S	39°9’0”W
Bahia	Casa Nova	PEUFR	9°23’8”S	41º49’4”W
Bahia	Itaparica	UFP	9º04’11”S[a]	44°24113”W[a]
Bahia	Jeremoabo	UHP	10°10’00”S	38º47’00W
Pernambuco	Ibimirim	PEUFR	-	-
Pernambuco	Ibimirim	PEUFR	-	-
Pernambuco	Mirandiba	UFP	08º07’.849S	38°41.45W
Piauí	São Raimundo Nonato	UFP	08º44’01”S	042º29’2”W

[a]Data converted from UTM to degrees. (-) Data not available

Table S3 Characterization of cpSSR motifs from *Cenostigma microphyllum* plastome

Total SSR	79
Mononucleotides	62
Dinucleotides	7
Trinucleotides	2
Tetranucleotides	8
Intergenics	50
Genics	29
*atp*B	1
*atp*F	1
*ccs*A	1
*ndh*D	1
*ndh*F	1
*rpo*B	1
*rps*18	1
*trn*G-UCC	1
*trn*K-UUU	1
*rpo*C1	2
*ycf*3	3
*clp*P	4
*rpo*C2	5
*ycf*1	6

Table S4 Chloroplastidial microsatellites (cpSSR) primers developed in this work for *Cenostigma microphyllum*. F – Primer Forward; R – Primer Reverse; Tm – Melting temperature. All forward primers contain a 5'-M13 tail as described by Schuelke (2000)

Primer	5' – Primer Sequence – 3'	Tm (°C)	Expected product size (pb)	Locus	Motif
CmCPSSR1	F: CGGGGTACTTTATTGCTTAGTCTG R: AAGAAGGCGAGTCTGTATGCTAAT	60,4 59,9	276	*atp*H - *atp*F	$(T)_{12}$
CmCPSSR2	F: AATCTTTCTCGATCAATCCTTTTG R: AGAAAATGGGTCAGATTCTACAGG	60,0 59,9	220	*trn*V-GAC	$(T)_{10}$
CmCPSSR3	F: CAAGCGGTATTCAAGCTCTTATTT R: GATCTTTTAGATGGACCTCTTTCC	60,2 58,7	296	*atp*I - *atp*H	$(T)_{10}$
CmCPSSR4	F: TGGAGCTTTTGAATAACAGTCAAG R: ACCTATTGAGAAATCCCTAACTTT	59,8 55,9	352	*rpo*A - *pet*D[a]	$(A)_{12}$
CmCPSSR5	F: GCTATGGTGAAATCGGTAGACAC R: AACCCCATGAAAGAAAGATTACTG	59,9 59,8	423	*trn*L-UAG - *ccs*A[a]	$(T)_{11}$
CmCPSSR6	F: ATCAGATGCTGGATATCTTACACG R: AATACCTACAGCTTCTCCCAGTTC	59,5 59,3	373	*rpo*C2	$(AT)_5$
CmCPSSR7	F: TGGGTTAGGTCAACTCTACTCCAC R: AAAATACGAACGAGATGGATGTTT	60,8 60,1	177	*ycf*3	$(A)_{11}$
CmCPSSR8	F: GGCTAGGTAAGCGTCCTGTAGTAA R: CCAATCATTGTGGGTATAATGGTA	60,2 59,7	414	*rpl*2 – *rps*19[a]	$(A)_{10}$
CmCPSSR9	F: ATATTGGGTAGCTGTCGTTAAACC R: TTTGGTACTGCTTCCTTGATTACA	59,7 60,1	430	*rps*16 – *rpl*14[a]	$(T)_{15}$
CmCPSSR10	F: TCCCGAAAATCCTATTTCTTACAA R: GAATTGAGAAAATTCTGTCCCTGT	60,2 59,9	246	*rps18* – rpl*33*	$(AT)_5$
CmCPSSR11	F: TTCTAGGGAAGGAACTGAGGTTTA R: GGAACTAGTCGGATGGAGTAGATAA	59,7 59,1	144	*trn*K - UUU	$(T)_{10}$
CmCPSSR12	F: TATGCCTCTCCTTCTAGCATTTTT R: GAAACATTCCCTTATTTTCATTCG	59,8 60,1	113	*trn*L-UAA – *trn*T-UGU	$(T)_{11}$
CmCPSSR13	F: ACAAGGATCAAGATCAAATGAACA R: ATGATTCGATTAAGCCCTGAATAA	59,9 60,2	172	*rpo*C2	$(A)_{10}$
CmCPSSR14	F: GTGACGCATCATCCTCATTTTAAG	62,0	196	*trn*F-GAA –	$(A)_{10}$

	R: TATCTTTGAGCAAGGAATCCTCAT	60,4		*trn*L-UAA	
CmCPSSR15	F: AGCATTACACAATCTCCAAGATCA	60,0	213	*psb*I – *psb*K	$(A)_{10}$
	R: ATCCTATTATGAGTCCTCCACCAA	60,1			
CmCPSSR16	F: TTAACAAATGGGAGACGTAAACAA	59,8	190	*trn*G-UCC	$(T)_{10}$
	R: AAAATGCTTTTCCTAAGGATTCTC	58,5			

[a]Primers anchored in genic regions

Table S5 *In silico* cross-amplification of the sixteen primers (F and R) developed in this work, showing percentage of identity, size (in bp) of amplified product and SSR content in the locus, for *Cenostigma microphyllum* to 27 plastomes of 22 species representing the six Leguminosae subfamily

Loci	*Cenostigma microphyllum*			*Cenostigma pyramidale*				*Cenostigma pyramidale* 2				*Arquita mimosifolia*			
	Region	Size	SSR	F	R	Size	SSR	F	R	Size	SSR	F	R	Size	SSR
CmCPSSR1	*atpH - atpF*	276	T12	100%	100%	275	A11GA6	100%	100%	276	A12GA6	95,83%	100%	273	A10...T10
CmCPSSR2	*trnV-GAC - rps12*	220	T10	100%	100%	220	T10	100%	100%	220	T10	100%	100%	218	A8
CmCPSSR3	*atpI - atpH*	296	T10	100%	100%	297	T11...A9	100%	100%	296	T9...A10	100%	100%	285	T10
CmCPSSR4*	*rpoA - petD*	352	A12	100%	100%	352	T8... A12GA8	100%	100%	352	T12... A8GA8	100%	100%	354	(A10)G(A9)
CmCPSSR5*	*trnL-UAG - ccsA*	423	T11	100%	100%	420	T8...T8	100%	100%	423	T8...T11	100%	91,67%	437	T7
CmCPSSR6	*rpoC2*	373	(AT)5	100%	100%	373	(AT)5	100%	100%	373	(AT)5	100%	100%	373	(AT)5
CmCPSSR7	*ycf3*	177	A11	100%	100%	177	T11	100%	100%	177	T11	100%	100%	175	T9
CmCPSSR8*	*rpl2 - rps19*	414	A10	100%	100%	414	T9...T10...T8... T7	100%	100%	414	T9...T10...T8... T7	100%	100%	414	T8...T9
CmCPSSR9*	*rps16 - rpl14*	430	T15	100%	100%	429	T8...A8...A14	100%	100%	430	T8...A8...A15	100%	100%	421	T13...A11
CmCPSSR10	*rps18 -rpl33*	246	(AT)5	100%	100%	250	(TA)5T9	100%	100%	246	(TA)5T5	100%	100%	253	(TA)4
CmCPSSR11	*trnK-UUU*	144	T10	100%	100%	144	A10	100%	100%	145	A11	95,83%	95,83%	145	A7
CmCPSSR12	*trnL-UAA - trnT-UGU*	113	T11	100%	100%	111	A9	100%	100%	113	A11	91,67%	95,83%	115	A13
CmCPSSR13	*rpoC2*	172	A10	100%	100%	172	T10	100%	100%	172	T10	91,67%	100%	172	T10
CmCPSSR14	*trnF-GAA - trnL-UAA*	196	A10	100%	100%	196	T10	100%	100%	196	T10	75%	95,83%	193	T7
CmCPSSR15	*psbI - psbK*	213	A10	100%	100%	210	T7	100%	100%	214	T11	100%	100%	219	T11
CmCPSSR16	*trnG-UCC*	190	T10	70,84%	100%	189	A9...T8	100%	100%	190	A10...T8	91,67%	95,83%	196	A10
Primers > 90% identity						15				16				15	
SSR >10 bp						11				16				9	
SSR > 6 bp						16				16				16	
Cross-amplified > 10 bp						11				16				9	
Cross-amplified > 6 bp						15				16				15	

*Primers anchored in genic regions

Table S5 Continued

Loci	*Balsamocarpon brevifolium*				*Biancaea sappan*				*Caesalpinia pulcherrima*			
	F	R	Size	SSR	F	R	Size	SSR	F	R	Size	SSR
CmCPSSR1	95,83%	100%	279	A9	100%	95,83%	281	A9GA8	100%	100%	282	T9CT14
CmCPSSR2	100%	100%	219	A9	100%	100%	214	T8	79,16%	100%	213	T8
CmCPSSR3	100%	70,84%	206	A10	100%	100%	309	A7	100%	100%	306	T10...A10
CmCPSSR4*	100%	95,83%	345	(T8)AG(A)13	100%	100%	351	T8 AG A11 G A8	100%	100%	346	T12...A8
CmCPSSR5*	100%	100%	422	T8...T9	100%	100%	420	T7...T8	100%	100%	415	T10
CmCPSSR6	100%	100%	373	(AT)5	100%	100%	373	(AT)5	100%	100%	373	(AT)5
CmCPSSR7	100%	100%	175	T9	100%	100%	177	T11	100%	100%	174	A9
CmCPSSR8*	100%	100%	393	T9	100%	100%	392	T9...T8	100%	100%	414	T9
CmCPSSR9*	100%	100%	423	A13...T9	100%	100%	424	T10...A8... A11...T11	100%	100%	430	T11...A14
CmCPSSR10	100%	100%	238	(TA)5	95,83%	100%	267	(TA)3	100%	100%	251	(AT)3
CmCPSSR11	100%	100%	147	A8	95,83%	100%	148	A7	100%	100%	137	T8 ... A10
CmCPSSR12	95,83%	100%	119	A14	100%	95,83%	112	A10	95,83%	91,67%	112	T10
CmCPSSR13	100%	100%	172	T10	100%	100%	178	T10	95,83%	100%	172	A10
CmCPSSR14	100%	100%	194	T8	100%	100%	194	T7...T8	87,5%	95,83%	188	A8 G A11
CmCPSSR15	100%	100%	217	T9	100%	100%	196	T10	95,83%	100%	242	T10
CmCPSSR16	95,83%	100%	201	A14	95,83%	100%	194	A12...T9	95,83%	100%	195	T13
Primers > 90% identity		15				16				14		
SSR >10 bp		8				8				12		
SSR > 6 bp		16				16				16		
Cross-amplified > 10 bp		7				8				11		
Cross-amplified > 6 bp		15				16				14		

*Primers anchored in genic regions

Table S5 Continued

Loci	*Coulteria platyloba*				*Erythrostemon huguesii*				*Erythrostemon gilliesi*			
	F	R	Size	SSR	F	R	Size	SSR	F	R	Size	SSR
CmCPSSR1	100%	100%	281	A10	100%	95,83%	265	A7	100%	100%	256	A8
CmCPSSR2	100%	100%	224	A9	100%	100%	211	A7	100%	100%	221	T11
CmCPSSR3	100%	100%	305	T10A9	95,83%	100%	304	T9...A10	100%	95,83%	309	T10...T7...T9...A10
CmCPSSR4*	100%	100%	355	T10AGA12GA9	100%	100%	364	T11AGA12GA7	95,83%	100%	349	T7... A13GA6
CmCPSSR5*	100%	100%	420	T8...T7	100%	100%	419	T7...T7	100%	100%	419	T7...T7
CmCPSSR6	100%	95,83%	373	(AT)5	100%	100%	373	(AT)5	100%	100%	373	(AT)5
CmCPSSR7	100%	100%	175	T9	100%	100%	174	T9	100%	100%	173	T8
CmCPSSR8*	100%	100%	415	T9...T9...T8...T7	100%	100%	416	T9...T10	100%	100%	414	T9...T9...T7...T7
CmCPSSR9*	100%	100%	420	T8...A8...A12...T10	100%	100%	424	A10...A8...T10	100%	100%	424	T8...A11...A7...T11
CmCPSSR10	100%	100%	263	(TA)4	100%	100%	249	(TA)4	100%	100%	249	(AT)3...(TA)4...A6
CmCPSSR11	100%	100%	147	A8	91,67%	95,83%	146	A6	95,83%	95,83%	145	A6
CmCPSSR12	100%	95,83%	113	A11	100%	95,83%	113	A11	100%	91,67%	114	A12
CmCPSSR13	100%	91,67%	172	T10	91,67%	100%	172	T10	91,67%	100%	172	T10
CmCPSSR14	95,83%	100%	195	T9	100%	100%	175	T8	95,83%	95,83%	189	T9
CmCPSSR15	100%	100%	219	T11	91,67%	100%	221	T11	100%	95,83%	221	T12
CmCPSSR16	95,83%	100%	193	A11...T8	91,67%	91,67%	164	T6	91,67%	95,83%	191	A10...T9
Primers > 90% identity	16				16				16			
SSR >10 bp	9				8				9			
SSR > 6 bp	16				16				16			
Cross-amplified > 10 bp	9				8				9			
Cross-amplified > 6 bp	16				16				16			

*Primers anchored in genic regions

Table S5 Continued

Loci	*Erythrostemon gilliesii* 2				*Guillandina bounduc*				*Guilandina bonduc* 2			
	F	R	Size	SSR	F	R	Size	SSR	F	R	Size	SSR
CmCPSSR1	100%	100%	256	A8	100%	95,83%	296	A12GA9	100%	100%	297	A13GA7
CmCPSSR2	100%	100%	221	T11	100%	100%	218	T8	100%	100%	218	T8
CmCPSSR3	100%	95,83%	309	T10...T7... T9...A10	100%	62,5%	306	T10...A10	100%	62,5%	306	T10...A10
CmCPSSR4*	95,83%	100%	349	T7... A13GA6	100%	100%	323	T8	100%	100%	323	T8AGA7
CmCPSSR5*	100%	100%	419	T7...T7	100%	100%	422	T11	100%	100%	424	T13
CmCPSSR6	100%	100%	373	(AT)5	95,83%	100%	373	(AT)5	95,83%	100%	373	(AT)5
CmCPSSR7	100%	100%	173	T8	100%	100%	175	T10	100%	100%	175	T10
CmCPSSR8*	100%	100%	414	T9...T9... T7...T7	100%	100%	415	T9...T9	100%	100%	415	T9...T9...T8...T7
CmCPSSR9*	100%	100%	424	T8...A11...A7...T11	100%	100%	423	A9...A11..T9	100%	100%	423	T10...A9...A11...T9
CmCPSSR10	100%	100%	249	(AT)3...(TA)4... A6	100%	100%	267	(AT)3	100%	100%	267	(AT)3...A6
CmCPSSR11	95,83%	95,83%	145	A6	95,83%	100%	149	A4 T A9	95,83%	100%	149	A4 T A9
CmCPSSR12	100%	91,67%	114	A12	91,67%	95,83%	112	A10	91,67%	95,83%	112	A10
CmCPSSR13	91,67%	100%	172	T10	100%	100%	172	T10	100%	100%	172	T10
CmCPSSR14	95,83%	95,83%	189	T9	100%	100%	195	T9	100%	100%	195	T9
CmCPSSR15	100%	95,83%	221	T12	100%	100%	205	T10CT8	100%	100%	205	T10 C T8
CmCPSSR16	91,67%	95,83%	191	A10...T9	95,83%	95,83%	194	A8...T9	95,83%	95,83%	194	A8...T9
Primers > 90% identity		16				15				15		
SSR >10 bp		9				9				9		
SSR > 6 bp		16				16				16		
Cross-amplified > 10 bp		9				8				8		
Cross-amplified > 6 bp		16				15				15		

*Primers anchored in genic regions

Table S5 Continued

Loci	*Haematoxylum brasiletto*				*Libidibia ferrea*				*Libidibia coriaria*			
	F	R	Size	SSR	F	R	Size	SSR	F	R	Size	SSR
CmCPSSR1	100%	100%	278	A7GA6	100%	100%	274	A9GA6	100%	100%	269	A8GA6
CmCPSSR2	100%	100%	220	A10	100%	100%	216	A14	100%	100%	215	T7...T14
CmCPSSR3	100%	100%	303	T8...A9	100%	100%	305	T11...A8	100%	100%	302	T8...A8
CmCPSSR4*	100%	100%	342	T10AGA9	100%	100%	343	T8...A12	100%	100%	344	T8 AG A13
CmCPSSR5*	100%	100%	416	T6	100%	100%	419	T7...T7	100%	100%	449	T7...T8
CmCPSSR6	100%	100%	373	(AT)5	100%	100%	373	(AT)5	100%	100%	373	(AT)5
CmCPSSR7	95,83%	100%	175	T9	100%	100%	176	T10	100%	100%	177	T11
CmCPSSR8*	100%	100%	413	T7...T8...T8...T7...T7	100%	100%	415	T9...T9	100%	100%	417	T9...T9...T10...T7
CmCPSSR9*	100%	100%	419	T8...A8...A11...T8	100%	100%	425	T10...A8...A13...T9	100%	100%	428	T11...A8...A13...T11
CmCPSSR10	100%	100%	256	(TA)4	100%	100%	267	(TA)4	100%	100%	251	(TA)3
CmCPSSR11	100%	100%	141	A6	100%	100%	147	A4 T A8	100%	100%	141	A10
CmCPSSR12	95,83%	91,67%	115	A8	100%	95,83%	113	A12	100%	95,83%	115	A13
CmCPSSR13	100%	100%	172	T10	95,83%	100%	167	T10	100%	100%	172	T10
CmCPSSR14	100%	95,83%	118	-	100%	100%	195	T9	100%	100%	195	T9
CmCPSSR15	100%	100%	217	T6...T9	95,83%	100%	223	T9	100%	95,83%	218	T9
CmCPSSR16	79,17%	100%	184	A7...T8	95,83%	100%	193	T11	95,83%	100%	200	A13...T8
Primers > 90% identity	15				16				16			
SSR >10 bp	5				9				10			
SSR > 6 bp	15				16				16			
Cross-amplified > 10 bp	5				9				10			
Cross-amplified > 6 bp	15				16				16			

*Primers anchored in genic regions

Table S5 Continued

Loci	*Libidibia coriaria* 2				*Mezoenuron cucullatum*				*Mezoneuron cucullatum* 2			
	F	R	Size	SSR	F	R	Size	SSR	F	R	Size	SSR
CmCPSSR1	100%	100%	269	A8GA6	100%	100%	287	(A)11T(A)10	100%	100%	287	(A)11T(A)10
CmCPSSR2	100%	100%	215	T7...T14	100%	100%	226	A11	100%	100%	226	T11
CmCPSSR3	100%	100%	302	T8...A8	100%	100%	305	T12...A12	100%	100%	305	T11...T7...A12
CmCPSSR4*	100%	100%	344	T8 AG A13	100%	100%	358	T12CTA13	100%	100%	358	T12 CT A13 GAT A7
CmCPSSR5*	100%	100%	419	T7...T8	100%	100%	420	T7...T7	100%	100%	420	T7...T7
CmCPSSR6	100%	100%	373	(AT)5	100%	100%	373	(AT)5	100%	100%	373	(AT)5
CmCPSSR7	100%	100%	177	T11	100%	100%	178	T7	100%	100%	179	T7
CmCPSSR8*	100%	100%	394	T10...T7	100%	100%	416	T9...T10	100%	100%	416	T9...T8...T7
CmCPSSR9*	100%	100%	426	T10...A8...A13... T10	100%	95,83%	423	A12...T11	100%	95,83%	423	T8...A8...A12... T11
CmCPSSR10	100%	100%	245	(TA)3	95,83%	100%	256	(TA)4	95,83%	100%	256	(TA)4
CmCPSSR11	100%	100%	140	A9	100%	100%	148	A7	100%	100%	148	A7
CmCPSSR12	100%	95,83%	113	A11	95,83%	33,34%	117	A15	95,83%	33,34%	118	A15
CmCPSSR13	100%	100%	172	T10	100%	100%	178	T10	100%	100%	178	T10
CmCPSSR14	100%	100%	195	T9	100%	100%	195	T9	100%	100%	195	T9
CmCPSSR15	100%	95,83%	218	T9	100%	100%	197	T11	100%	100%	197	T11
CmCPSSR16	95,83%	100%	199	A12...T8	95,83%	100%	192	A10	95,83%	100%	192	A10...T9
Primers > 90% identity			16				15				15	
SSR >10 bp			9				11				10	
SSR > 6 bp			16				16				16	
Cross-amplified > 10 bp			9				10				9	
Cross-amplified > 6 bp			16				15				15	

*Primers anchored in genic regions

Table S5 Continued

Loci	*Moullava spicata*				*Paubrasilia echinata*				*Pterolobium punctatum*			
	F	R	Size	SSR	F	R	Size	SSR	F	R	Size	SSR
CmCPSSR1	100%	100%	292	A14	100%	100%	276	T12	100%	100%	278	A13
CmCPSSR2	100%	100%	225	A7...A10	100%	100%	198	A8	100%	100%	220	A10
CmCPSSR3	100%	100%	303	T7...A10	100%	100%	304	T9	100%	100%	310	T12...A12
CmCPSSR4*	100%	95,83%	353	T11...A9GA9	100%	100%	342	A11	100%	100%	358	(T11)AG(A14)G(A9)
CmCPSSR5*	100%	100%	422	T9...T11	100%	100%	414	T8...T7	100%	100%	425	T7...T7
CmCPSSR6	100%	100%	373	(AT)5	100%	100%	373	(AT)5	100%	100%	373	(AT)5
CmCPSSR7	100%	100%	178	T10	100%	100%	175	A9	95,83%	100%	176	A10
CmCPSSR8*	100%	100%	417	T9...T10...T9	100%	100%	415	T9...T9	100%	100%	415	T9...T9
CmCPSSR9*	100%	100%	427	T13...A8...A12...T10	100%	100%	426	T10	100%	95,83%	428	A10...A16...T10
CmCPSSR10	100%	100%	261	(TA)4	100%	100%	249	(AT)3	100%	100%	250	(TA)4
CmCPSSR11	100%	100%	147	A8	95,83%	100%	144	T5	100%	100%	148	A7
CmCPSSR12	100%	91,67%	113	A10	91,67%	95,83%	108	T13	100%	95,83%	114	A12
CmCPSSR13	100%	100%	178	T10	95,83%	100%	172	A10	100%	100%	172	T10
CmCPSSR14	95,83%	100%	202	T10	100%	100%	195	A9	100%	100%	200	T9
CmCPSSR15	95,83%	100%	194	T8	100%	100%	217	T9	100%	91,67%	200	T15
CmCPSSR16	91,67%	100%	193	A10... T10	95,83%	100%	194	T11	91,67%	100%	199	A16...T10
Primers > 90% identity			16				16				16	
SSR >10 bp			13				7				11	
SSR > 6 bp			16				15				16	
Cross-amplified > 10 bp			13				7				11	
Cross-amplified > 6 bp			16				15				16	

*Primers anchored in genic regions

Table S5 Continued

Loci	*Tara cacalaco*				*Schnella tricosepala*				*Tamarindus indica*			
	F	R	Size	SSR	F	R	Size	SSR	F	R	Size	SSR
CmCPSSR1	100%	100%	274	A9	79,16%	100%	267	(AT)5	91,67%	95,83%	273	A9
CmCPSSR2	100%	100%	181	T9	33,33%	100%	206	T9	100%	100%	215	T6 ... T7
CmCPSSR3	100%	100%	304	T9... A9	100%	16,67%	301	A10	91,67%	100%	319	T9...T8
CmCPSSR4*	100%	100%	354	T9...A12GA9	100%	83,33%	386	T12	91,67%	91,67%	363	T9
CmCPSSR5*	100%	100%	420	T8	100%	95,83%	437	T6...T6	100%	91,67%	413	T6
CmCPSSR6	100%	100%	373	(AT)5	100%	100%	373	(AT)5	100%	95,83%	373	(AT)5
CmCPSSR7	100%	100%	176	T10	100%	91,67%	175	C5	100%	91,67%	171	T5
CmCPSSR8*	100%	100%	415	T9...T9...T8	100%	95,83%	424	T9	100%	100%	395	T11
CmCPSSR9*	100%	100%	417	A12...T11	91,67%	100%	431	A8	95,83%	95,83%	430	A8
CmCPSSR10	100%	100%	251	(AT)3	100%	87,5%	250	(TA)4	91,67%	91,67%	274	(TA)3
CmCPSSR11	100%	100%	143	A9	100%	91,67%	154	A13	100%	91,67%	158	A9
CmCPSSR12	95,83%	95,83%	111	A9	70,83%	79,16%	118	A10	75%	79,16%	109	A6CA5
CmCPSSR13	100%	100%	172	T10	87,5%	95,83%	166	T6	83,33%	95,83%	166	T10
CmCPSSR14	100%	100%	195	T7...T9	58,33%	91,67%	192%	T8	79,16%	100%	214	T8
CmCPSSR15	100%	83,34%	222	T14	91,67%	37%	254	T7	100%	83,34%	219	T9
CmCPSSR16	91,67%	100%	196	A14...T9	95,83%	41,67%	185	A10	91,67%	41,67%	196	A7
Primers > 90% identity			14				6				11	
SSR >10 bp			7				7				3	
SSR > 6 bp			16				15				15	
Cross-amplified > 10 bp			5				2				2	
Cross-amplified > 6 bp			14				6				11	

*Primers anchored in genic regions

Table S5 Continued

Loci	*Vigna unguiculata*				*Zenia insignis*				*Duparquetia orchidaceae*			
	F	R	Size	SSR	F	R	Size	SSR	F	R	Size	SSR
CmCPSSR1	95,83%	41,67%	248	A5	91,66%	91,66%	279	A10	100%	91,67%	260	A8
CmCPSSR2	95,83%	91,67%	189	A7	100%	100%	221	A8	100%	100%	259	A7
CmCPSSR3	33,34%	95,83%	295	A8	100%	87,5%	346	T11...T10...AT7	100%	95,83%	326	T7...A9
CmCPSSR4*	95,83%	91,67%	386	A8	95,83%	100%	346	T8...A8	95,83%	100%	357	T13...A8...A9
CmCPSSR5*	95,83%	83,33%	393	A8...T6	95,83%	100%	445	T7...T7	100%	95,83%	423	A8G9
CmCPSSR6	100%	91,67%	373	(AT)3	100%	100%	373	TA5	100%	95,83%	373	(TA)5
CmCPSSR7	70%	91,67%	171	-	87,5%	100%	174	T8	95,83%	95,83%	177	T5...C5
CmCPSSR8*	95,83%	91,67%	385	A9	100%	95,83%	416	T10	100%	95,83%	417	T9...T8...T9
CmCPSSR9*	95,83%	91,67%	406	T9	100%	100%	422	A11	100%	100%	437	A12...A11
CmCPSSR10	100%	87,5%	111	A8	100%	95,83%	300	TG3...TA4...TA3...A13...A11	25%	100%	249	(TA)6
CmCPSSR11	95,83%	79,16%	153	A10	70,83%	100%	146	A7	100%	100%	155	A8
CmCPSSR12	54,17%	62,5%	110	T13	87,5%	91,67%	112	A10	83,34%	91,67%	111	A6
CmCPSSR13	70,84%	83,34%	166	A8	95,83%	91,67%	166	T6	100%	100%	166	T6
CmCPSSR14	83,34%	-	-	A9	95,83%	95,83%	182	T10	91,67%	91,67%	181	T9
CmCPSSR15	83,4%	12,5%	246	T7	100%	79,16%	227	T16	95,83%	79,16%	223	T11
CmCPSSR16	58,33%	79,16%	193	T8	75%	100%	200	A13	87,5%	100%	196	A9...T7
Primers > 90% identity			5				10				12	
SSR >10 bp			2				11				5	
SSR > 6 bp			14				16				15	
Cross-amplified > 10 bp			0				7				3	
Cross-amplified > 6 bp			5				10				11	

*Primers anchored in genic regions

Table S6 Characterization of the ten haplotypes of *Cenostigma microphyllum* plots at PARNA Catimbau generated by four cpSSR markers

Haplotype number	Alellic composition	Number of individuals	Percentage of individuals per population							
			P08	R12	R15	P21	P22	R23	P27	P28
H1	125 366 426 227	1	-	-	-	-	5%	-	-	-
H2	125 367 426 227	32	17%	-	-	-	89%	-	85%	23%
H3	125 367 426 228	1	-	-	-	-	-	-	-	8%
H4	125 367 427 227	1	-	-	-	-	-	-	8%	-
H5	127 366 426 228	43	-	100%	100%	100%	-	7%	-	69%
H6	127 366 427 228	13	-	-	-	-	-	87%	-	-
H7	127 367 426 227	1	-	-	-	-	-	-	8%	-
H8	127 367 426 228	1	-	-	-	-	5%	-	-	-
H9	127 367 427 228	1	-	-	-	-	-	7%	-	-
H10	128 366 426 228	5	83%	-	-	-	-	-	-	-

Table S7 Thirty plastome intervals grouped by loci (genes or intergenic regions) showing the highest SNP frequencies (Pi) identified by comparing all Caesalpinia group species plastomes, or plastomes within the four genera with two or more species available. Loci that are polymorphic in all five samples are indicated in bold

Caesalpinia group	Loci	*ycf1*	*rps12 – clpP*	*petA – psbJ*	*psbI – trnS-GCU - trnG-GCC*	*clpP*	*psbZ - trnG-GCC - trnfM-CAU – rps14*
	Number of Windows	13	1	1	3	4	1
	Highest PI value	0,04811	0,04129	0,04024	0,03998	0,03892	0,03804
Cenostigma	Loci	*trnG-UCC - trnR-UCU – atpA*	*ndhC - trnV-UAC*	*rps11 - rpl36 – rps8*	*rps16 - trnQ-UUG – psbK - psbI*	*petD - rpoA*	*trnH-GUG – psbA*
	Number of Windows	4	3	3	4	3	2
	Highest PI value	0,02	0,01444	0,01444	0,01111	0,01111	0,01
Coulteria	Loci	*rps16 - trnQ-UUG*	*rpoC2 - rpoC1*	*trnT-GGU – psbD*	*ycf3*	*accD - psaI - ycf4*	*ccsA*
	Number of Windows	4	2	1	2	2	1
	Highest PI value	0,00333	0,00333	0,00333	0,00333	0,00333	0,00333
Erythrostemon	Loci	*psbZ - trnG-GCC - trnfM-CAU*	*rps3 - rps19 - rpl2*	***trnK-UUU - rps16***	*rps16 - trnQ-UUG - psbK*	*psbI - trnS-GCU - trnG-UCC*	*ycf1*
	Number of Windows	2	3	3	3	3	4
	Highest PI value	0,06483	0,05317	0,04667	0,044	0,04033	0,03983
Libidibia	Loci	*rpoB - trnC-GCA – petN*	*psbI - trnS-GCU - trnG-UCC*	*rps8 - rpl14*	*rps11 - rpl36 - rps8*	*psbE – petL*	*rpl33 - rps18*
	Number of Windows	5	3	3	3	3	1
	Highest PI value	0,04778	0.03222	0,03222	0,02944	0,02333	0,02222

Table S7 Continued

rpl16	*rps16 - trnQ-UUG – psbK*	*rpl16 - rps3 – rps19*	*trnT-UGU - trnL-UAA*	*rpl20 - rps12 – clpP*	***trnK-UUU - rps16***	*ndhF - rpl32 – trnL-UAG*	
1	1	1	1	1	1	1	
0,03782	0,03766	0,03606	0,03605	0,03545	0,03538	0,03462	
trnK-UUU - rps16	*rpoC1*	*trnC-GCA – petN*	*psbZ - trnG-GCC – trnfM-CAU*	*psbE – petL*	*ndhF – rpl32*	*ndhA*	*ycf1*
3	2	1	1	1	1	1	1
0,00889	0,00778	0,00778	0,00778	0,00778	0,00778	0,00778	0,00778
ndhG – ndhI	*psbA - trnK-UUU*	*trnK-UUU – matK*	***trnK-UUU - rps16***	*rps16*	*trnQ-UUG – psbK*	*trnG-UCC - trnR-UCU – atpA*	*atpH – atpI*
2	4	2	2	1	3	3	1
0,00333	0,00167	0,00167	0,00167	0,00167	0,00167	0,00167	0,00167
trnL-UAA - trnF-GAA	*psbA - matK*	*petA - psbJ*	*rps11 - rpl36 - rps8*	*rpl32 - trnL-UAG - ccsA*	*trnG-UCC - trnR-UCU - atpA*	*petN - psbM*	*rps18 - rpl20*
1	1	1	2	2	3	1	1
0,038	0,03417	0,0335	0,03333	0,033	0,03283	0,03267	0,03133
psbZ - trnG-GCC - trnfM - rps14	*petG - trnW-CCA - trnP-UGG - psaJ*	*ndhF – rpl32*	***trnK-UUU - rps16***	*ndhC - trnV-UAC*	*rpl16 - rps3*		
3	2	2	2	2	1		
0,02167	0,02111	0,02111	0,01889	0,01889	0,01778		

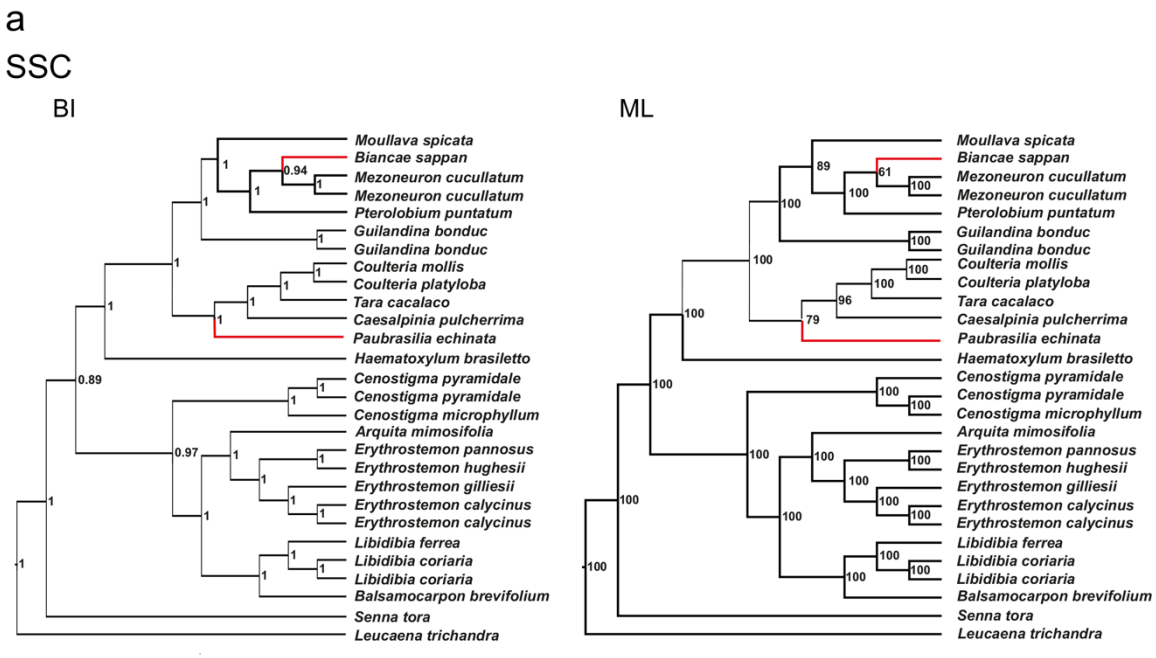

Fig. S1 Phylogenetic tree of twenty-six species from the Caesalpinia group obtained by different partitions of the plastome withdrawing the gaps in the alignment. Maximum Likelihood (ML) and Bayesian Inference (BI) analyses were performed for each partition. Bootstraps and Posterior probabilities are indicated. Conflict positions are marked in red. *Leucaena trichandra* and *Senna tora* are outgroups. a, Small Single Copy region (SSC, 13,959 bp). b, Long Single Copy region (LSC, 70,992 bp). c, Alignment of 113 genes (83,728 bp). d, Total alignment without the IRa (110,495 bp)

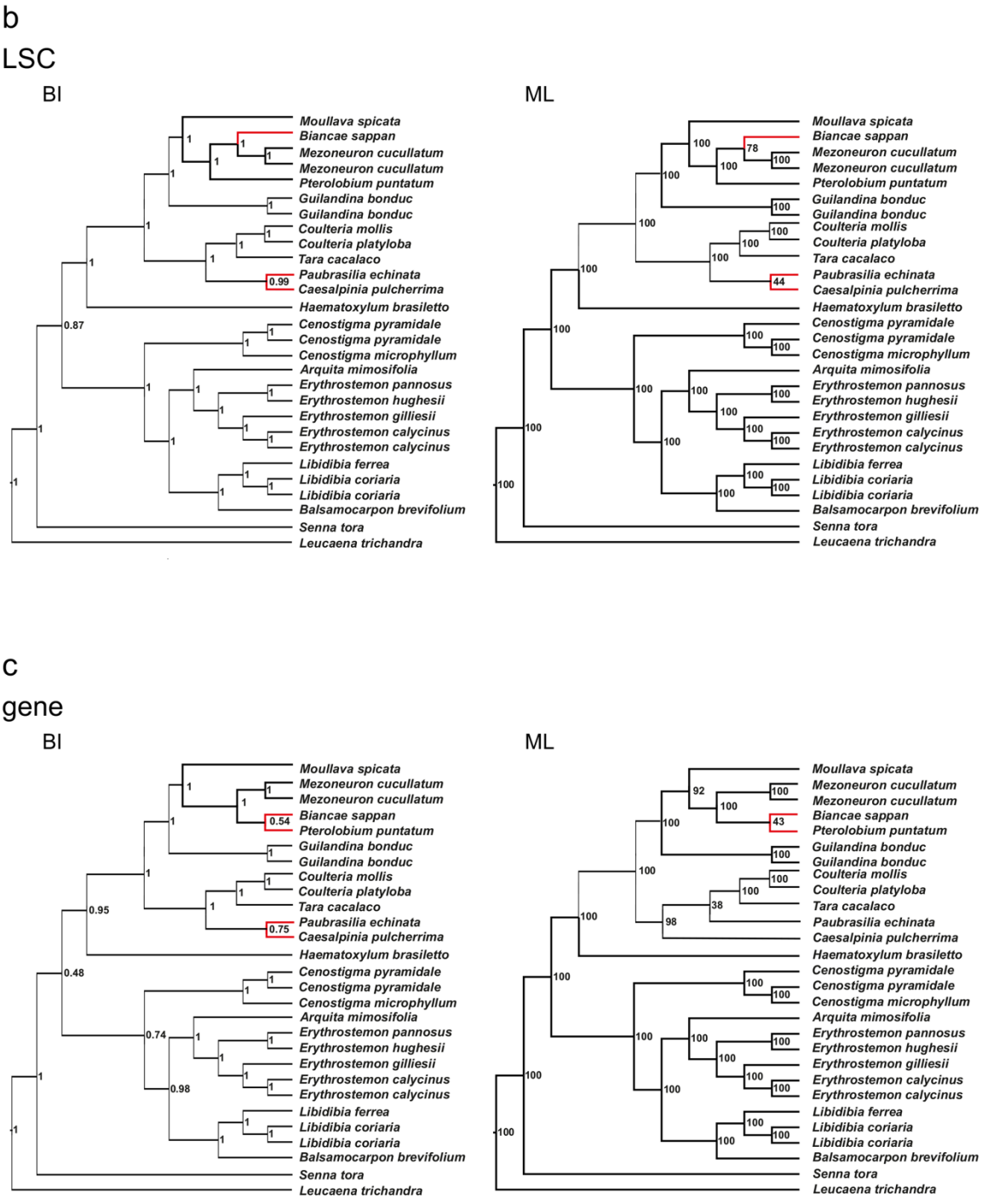

Fig. S1 Continued

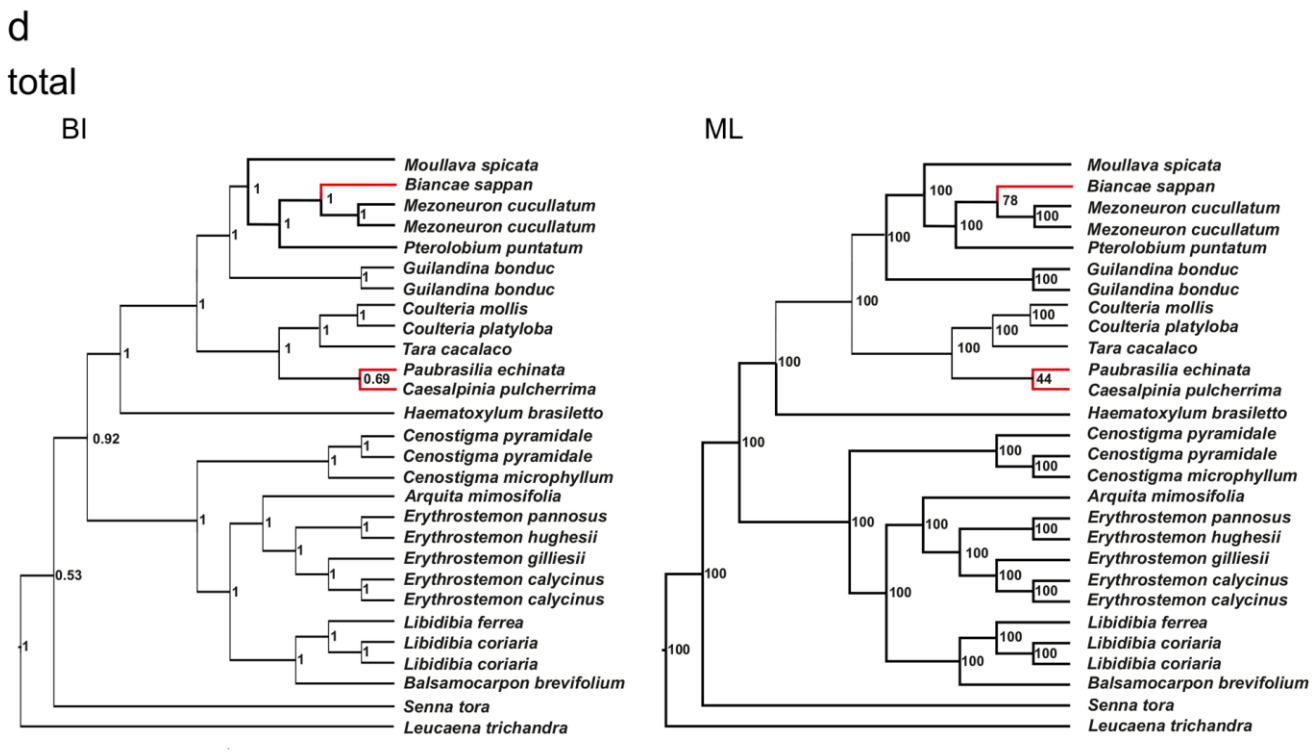

Fig. S1 Continued

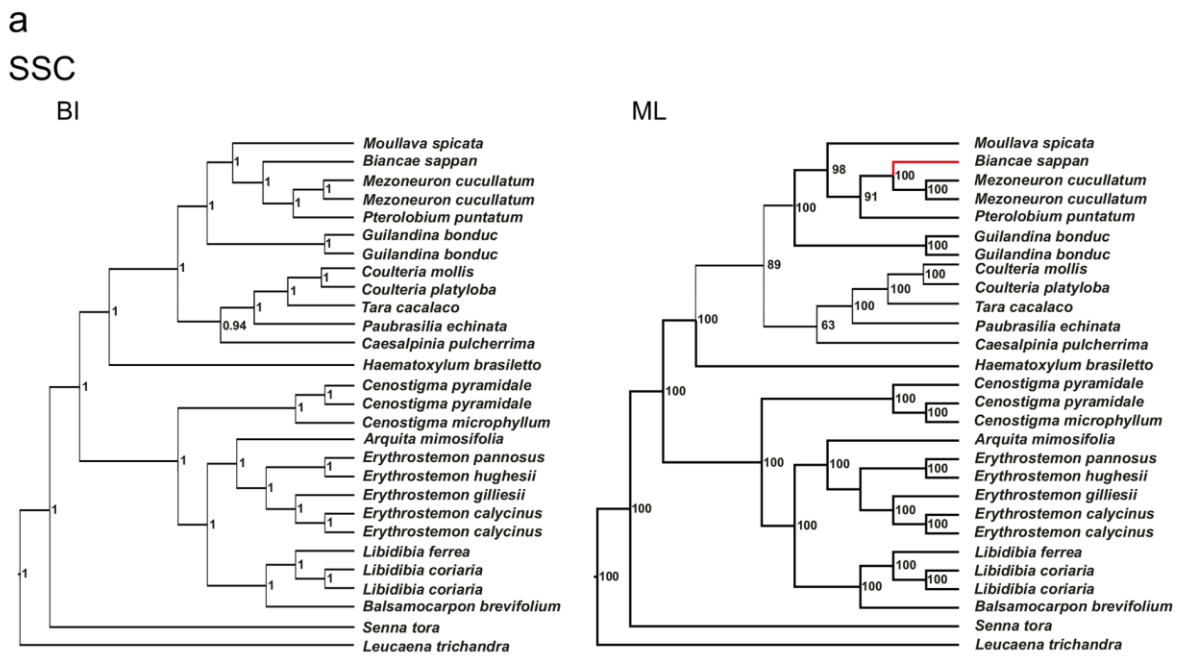

Fig. S2 Phylogenetic tree of twenty-six species from the Caesalpinia group obtained by different partitions of the plastome and including the gaps. Maximum Likelihood (ML) and Bayesian Inference (BI) analyses were performed for each partition. The Assembly and Alignment-Free (AAF) analysis was performed with total plastome data. Bootstraps and Posterior probabilities are indicated. Conflict positions are marked in red. *Leucaena trichandra* and *Senna tora* are outgroups. a, Small Single Copy region (SSC). b, Long Single Copy region (LSC). c, Alignment of 113 genes. d, Total alignment without IRa, including the Assembly and Alignment Free (AAF) phylogenetic tree

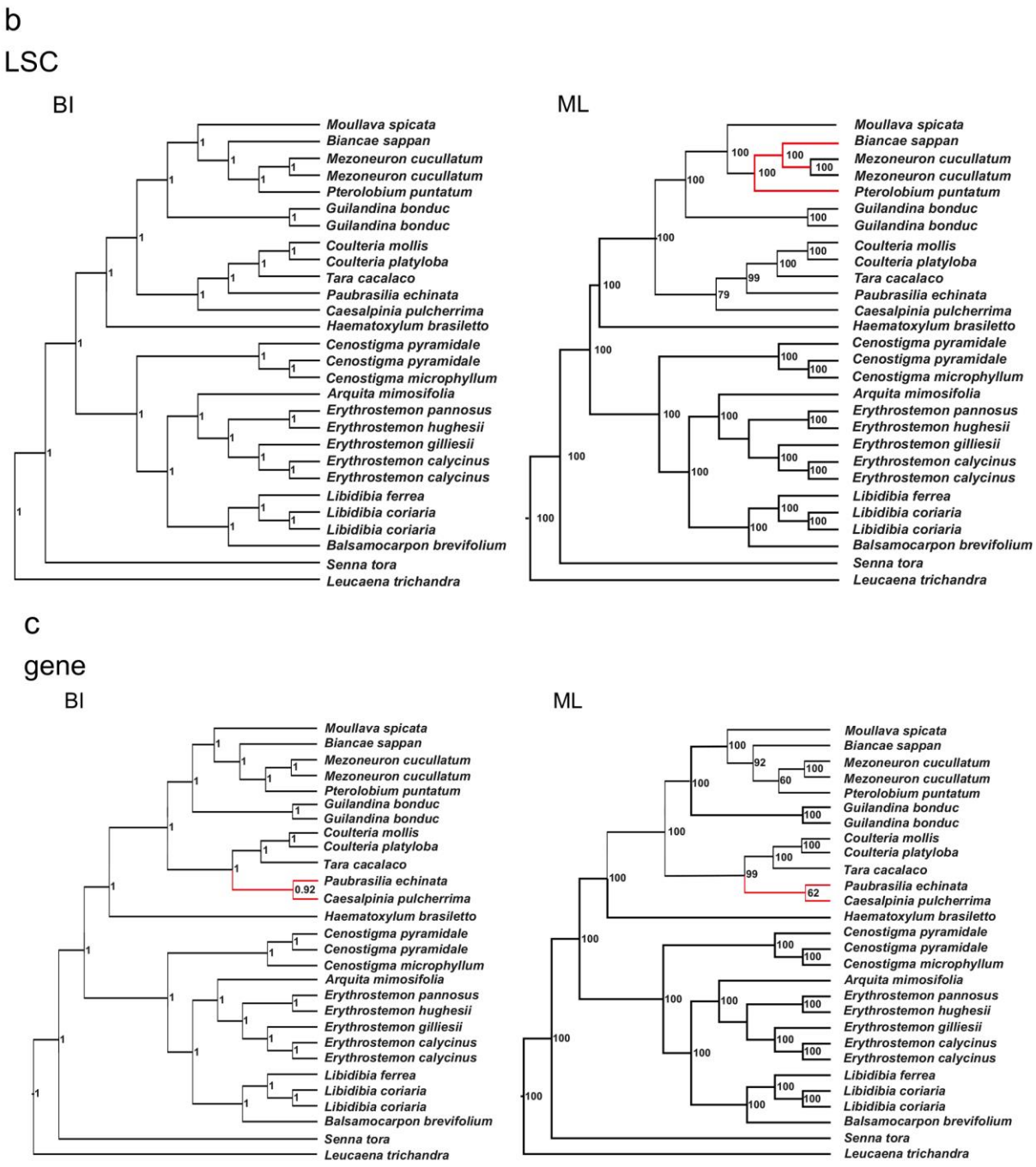

Fig. S2 Continued

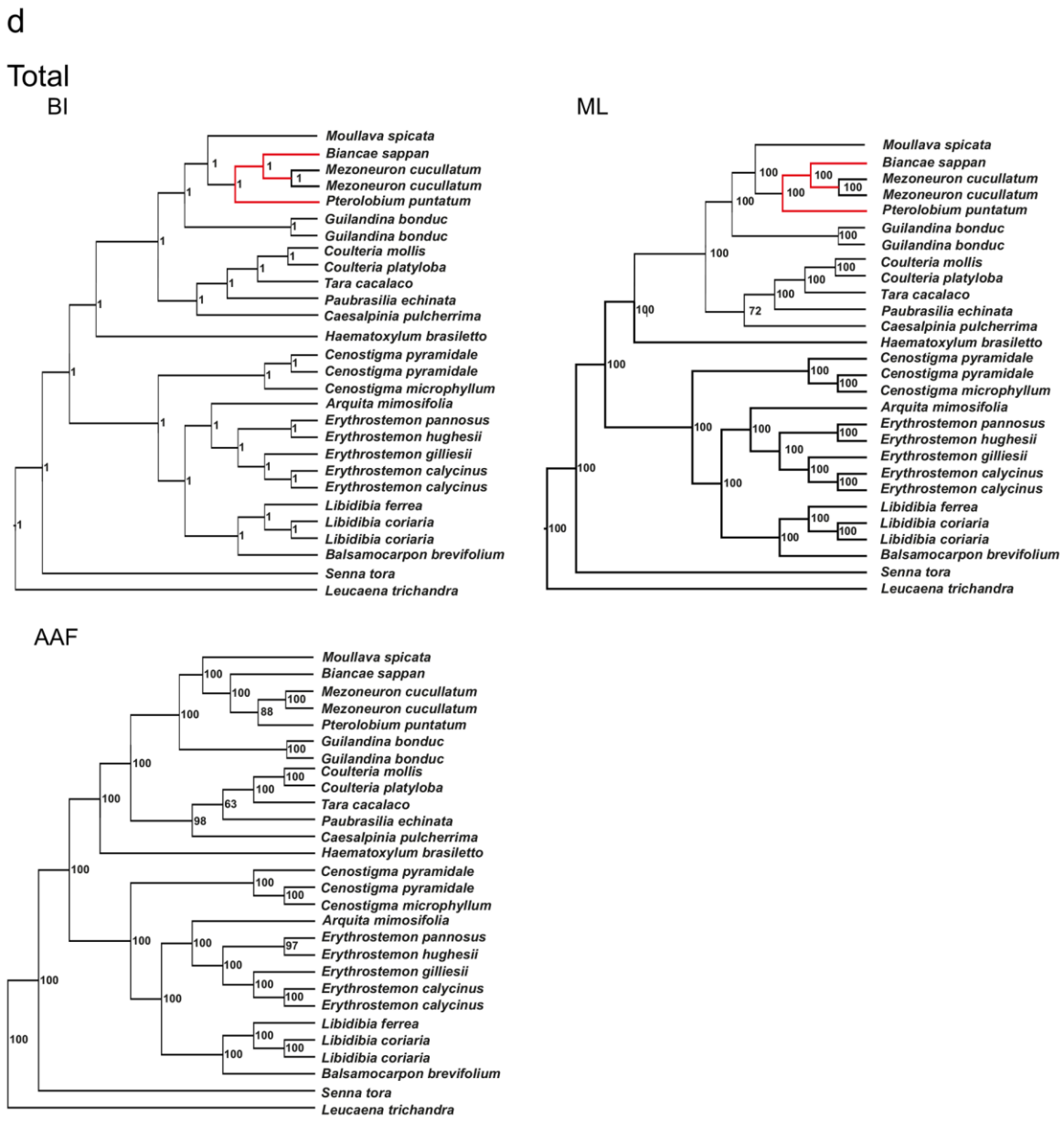

Fig. S2 Continued