Programa de Pós-Graduação em Biodiversidade e Evolução

Universidade Federal da Bahia

Filogenômica e arquitetura de genomas plastidiais em Chrysobalanaceae (Malpighiales)

Rogério Katsuhito Barbosa Maruyama

Salvador

2024

Rogério Katsuhito Barbosa Maruyama

Filogenômica e arquitetura de genomas plastidiais em Chrysobalanaceae (Malpighiales)

Dissertação apresentada ao Instituto de Biologia da Universidade Federal da Bahia para a obtenção do Título de Mestre em Biodiversidade e Evolução pelo Programa de Pós-Graduação em Biodiversidade e Evolução.

Orientadora: Dra. Goia de Mattos Lyra
Coorientadora: Dra. Renata C. Asprino

Salvador

2024

Ficha catalográfica

Dados internacionais de catalogação-na-publicação
(SIBI/UFBA/Biblioteca Universitária Reitor Macedo Costa)

Maruyama, Rogério Katsuhito Barbosa.
Filogenômica e arquitetura de genomas plastidiais em Chrysobalanaceae (Malpighiales) / Rogério Katsuhito Barbosa Maruyama. - 2024.
52 f.: il.

Orientadora: Profa. Dra. Goia de Mattos Lyra.
Coorientadora: Profa. Dra. Renata C. Asprino.
Dissertação (mestrado) - Universidade Federal da Bahia, Instituto de Biologia, Salvador, 2024.

1. Botânica. 2. Genômica. 3. Crisobalanácea - Filogenia. 4. Crisobalanácea - Classificação. 5. Crisobalanácea - Identificação I. Lyra, Goia de Mattos. II. Universidade Federal da Bahia. Instituto de Biologia. III. Título.

CDD - 582.13
CDU - 582.755.1

Programa de Pós-Graduação em
BIODIVERSIDADE E EVOLUÇÃO
UNIVERSIDADE FEDERAL DA BAHIA – INSTITUTO DE BIOLOGIA
Rua Barão de Geremoabo, s/n. Ondina – Salvador – Bahia – CEP. 40.170-000

Título de Dissertação: **Filogenômica e arquitetura de genomas plastidiais em Chrysobalanaceae (Malpighiales)**

Mestrando: **Rogério Katsuhito Barbosa Maruyama**
Orientadora: **Goia de Mattos Lyra**
Co-orientadora: **Renata Camargo Asprino Pereira**

BANCA EXAMINADORA

Documento assinado digitalmente
gov.br
GOIA DE MATTOS LYRA
Data: 12/07/2024 12:23:28-0300
Verifique em https://validar.iti.gov.br

Membro: Dra. Goia de Mattos Lyra
Universidade do Estado do Rio de Janeiro

Documento assinado digitalmente
gov.br
VERONICA AYDOS THODE
Data: 12/07/2024 16:42:36-0300
Verifique em https://validar.iti.gov.br

Universidade Federal do Rio Grande do Sul

Documento assinado digitalmente
gov.br
DANIEL SANTANA DE CARVALHO
Data: 15/07/2024 10:27:39-0300
Verifique em https://validar.iti.gov.br

Universidade Federal da Bahia

PARECER DA BANCA

A Comissão Examinadora considera o trabalho de conclusão de curso:
[X]Aprovado
[ ]Reprovado

À minha tia, Mihoko Maruyama,
pois poderia ter sido fruto jogado ao tempo,
mas a tempo fui recolhido,
e com o tempo frutifico.

Agradecimentos

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa concedida, e ao Programa de Pós-Graduação em Biodiversidade e Evolução pela oportunidade de realização deste trabalho. Ao seu corpo docente, agradeço tanto pela formação quanto pelas bancas de Seminário avaliativo parcial, pelas importantes contribuições que me ajudaram na evolução do meu projeto.

Agradeço à minha orientadora Goia de Mattos Lyra, pela paciência e carinho, desde a minha iniciação científica até o presente momento.

Agradeço à minha coorientadora Renata Asprino, pela generosa apresentação à família Chrysobalanaceae e pela sua disposição em ajudar em todas as etapas deste trabalho.

Agradeço também às minhas colegas de laboratório, Cibele Santos e Edilene Pestana, pelas contribuições tanto nas revisões, quanto pelas longas horas tentando decifrar as metodologias e barreiras tecnológicas.

De forma particular, agradeço aos Orixás pela vida, abundância e caminho até aqui.

Epígrafe

“Kò sí òrìṣà tí kò ní ìgbé.”

Não existe orixá sem sua

própria planta medicinal.

-Provérbio yorubá.

Índice

Lista de figuras

Lista de tabelas

Resumo

Chrysobalanaceae é uma das famílias da ordem Malpighiales cuja taxonomia vem desafiando a ciência há mais de dois séculos. A família compreende 27 gêneros e aproximadamente 550 espécies de distribuição predominantemente tropical, abarcando as Américas, África e sudeste da Ásia. Nos últimos 20 anos, os avanços nos estudos moleculares da família possibilitaram a resolução de problemas na taxonomia do grupo, incluindo a confirmação de seu monofiletismo e grandes mudanças na delimitação genérica. Estudos envolvendo genomas plastidiais completos tiveram como objetivo resolver as relações filogenéticas entre os gêneros da família, mas outras questões permanecem em aberto, como a caracterização de aspectos estruturais dos genomas. O presente estudo explorou a comparação da arquitetura de genomas plastidiais em Chrysobalanaceae, investigando a presença de duplicações gênicas e perdas de genes, além de variações no tamanho dos genes e em regiões de íntrons, com o objetivo de caracterizar os plastomas das espécies e analisar os eventos moleculares em um contexto filogenético. As comparações genômicas revelaram regiões de alta variabilidade, a exemplo do gene *ycf1*b, que pode funcionar como *barcode* para delimitação específica em Chrysobalanaceae. Os resultados permitem concluir que os plastomas já sequenciados de espécies de Chrysobalanaceae são altamente conservados em termos de sintenia gênica, mesmo em clados com história biogeográfica complexa. No entanto, a comparação entre diferentes aspectos da estrutura dos genomas ao longo dos subclados da família mostrou que tais análises podem fornecer subsídios para estudos taxonômicos integrativos e para a melhor compreensão dos processos evolutivos em Chrysobalanaceae.

Palavras-chave: *Barcoding;* Genômica; Plastoma; Sintenia.

Abstract

Chrysobalanaceae is one of the families of the order Malpighiales whose taxonomy has challenged science for over two centuries. The family comprises 27 genera and approximately 550 species with a predominantly tropical distribution, spanning the Americas, Africa, and Southeast Asia. In the last 20 years, advances in molecular studies of the family have made it possible to resolve problems in the group's taxonomy, including the confirmation of its monophyly and major changes in generic delimitation. Studies involving complete plastid genomes aimed to resolve the phylogenetic relationships among the genera of the family, but other questions remain unexplored, such as the characterization of structural aspects of the genomes. The present study explored the comparison of plastid genome architecture of Chrysobalanaceae, investigating the presence of gene duplications and gene losses, in addition to variations in gene size and intron regions, with the aim of characterizing the plastomes of the species and analyzing the molecular events in a phylogenetic context. Genomic comparisons revealed regions of high variability, such as the *ycf1*b gene, which may function as a barcode for species delimitation in Chrysobalanaceae. Our results show that the sequenced plastomes of Chrysobalanaceae species are highly conserved in terms of gene synteny, even in clades with complex biogeographic history. However, the comparison among different aspects of the genome structure across subclades showed that such analyses could provide support for integrative taxonomic studies and for a better understanding of the evolutionary processes in Chrysobalanaceae.

Keywords: *Barcoding; Genomics; Plastome; Synteny.*

1. Introdução geral

Chrysobalanaceae R.Br. (Malpighiales) compreende 27 gêneros e cerca de 550 espécies distribuídas nas regiões tropicais e subtropicais das Américas, África e sudeste da Ásia, chegando ao norte da Austrália e diversas ilhas do Pacífico (Prance & Sothers 2003; Sothers et al. 2016). Espécimes de Chrysobalanaceae são árvores ou arbustos de folhas simples e alternas, com estípulas, flores com disco nectarífero na extremidade do receptáculo, gineceu 1–3-carpelar, estilete ginobásico, óvulos com placentação basal e frutos do tipo drupa (Prance 1972; Prance & White 1988; Prance & Sothers 2003).

Cerca de 80% das espécies de Chrysobalanaceae estão distribuídas na região Neotropical, principalmente na Amazônia, onde se apresenta como uma das famílias com maior riqueza e abundância de espécies arbóreas (Rankin-de-Mérona et al. 1992; Bardon et al. 2016; Cardoso et al. 2017). Chrysobalanaceae inclui espécies com importância econômica, tanto na alimentação humana como no seu potencial farmacológico (Feitosa et al. 2012). Além disso, *Moquilea tomentosa* Benth. é amplamente utilizada na arborização urbana no Brasil, e algumas outras espécies de Chrysobalanaceae estão entre as mais adequadas para a restauração ecológica de áreas degradadas, especialmente na Amazônia (Gastauer et al. 2020). A representatividade florística nos Neotrópicos e a importância econômica das espécies são alguns dos fatores que explicam o investimento em estudos sobre este clado nas últimas décadas, desde aspectos sistemáticos até ecológicos, como na busca pela compreensão dos seus padrões fenológicos (Ruiz & Alencar 1999).

1.1. Contribuições de dados moleculares para a taxonomia da família

A delimitação de categorias supra específicas em Chrysobalanaceae tem sido particularmente difícil na história taxonômica do grupo. Por exemplo, Prance & White (1988) propuseram quatro tribos (*Chrysobalaneae*, *Parinarieae*, *Couepieae* e *Hirtelleae*) largamente baseadas em caracteres morfológicos florais e que não se sustentaram em filogenias posteriores. A partir das duas últimas décadas, com a intensificação das filogenias moleculares (Yakandawala et al. 2010; Bardon et al. 2013; Sothers et al. 2014; Sothers et al. 2016), foi confirmado o monofiletismo da família e os primeiros estudos apontaram sinapomorfias morfológicas para o clado, como a presença de cristais de sílica, o desenvolvimento do hipanto e o estilete ginobásico (Yakandawala et al. 2010; Bardon et al. 2013). Estudos subsequentes baseados em multigenes focaram na delimitação de gêneros neotropicais, e contribuíram para a resolução de alguns complexos taxonômicos, aproximando a

taxonomia da família a um esquema mais estável e que reflita a sua história evolutiva (Sothers et al. 2014; Sothers et al. 2016). Porém, foram mantidos como gêneros aceitos clados sem suporte nas filogenias [e.g. *Moquilea* Aubl. e *Leptobalanus* (Benth.) Sothers & Prance] ou mesmo grupos polifiléticos [e.g. *Hymenopus* (Benth.) Sothers & Prance] (Sothers et al. 2014; Sothers et al. 2016). A delimitação de *Moquilea*, *Leptobalanus*, *Hymenopus* e *Licania* Aubl. é particularmente difícil, baseando-se em caracteres morfológicos que se sobrepõem ou que não refletem a história evolutiva do grupo, ou seja, que não correspondem a sinapomorfias.

Com o avanço e o acesso mais amplo às técnicas de sequenciamento de Nova Geração na última década, foram geradas sequências de genomas plastidiais completos de um grande número de espécies de Chrysobalanaceae, com o principal objetivo de esclarecer suas relações filogenéticas e investigar sua história biogeográfica (Malé et al. 2014, Bardon et al. 2016, Chave et al. 2020). Essas filogenias com amostragens amplas tanto em espécies como em genes contribuíram para uma melhor compreensão da evolução e da história biogeográfica da família. Por outro lado, tais resultados filogenômicos colocaram em dúvida o monofiletismo de gêneros e espécies no clado Neotropical (Bardon et al. 2016; Chave et al. 2020). Alguns gêneros permanecem polifiléticos, como *Hirtella* L., *Leptobalanus* e *Hymenopus*, demandando mais estudos que combinem dados moleculares com aspectos morfológicos das espécies.

A filogenia mais completa até o momento (Chave et al. 2020) resolveu em grande parte a árvore das Chrysobalanaceae, especialmente no que se refere às relações mais profundas, entre gêneros. Principalmente no clado Neotropical, os clados de diversificação mais recente (a exemplo de *Hirtella*) apresentam baixo suporte para as relações interespecíficas. Apesar de contar com um conjunto de dados genômicos relevante, com todos os gêneros representados e ampla representatividade de espécies, em sua maioria, a estrutura e arquitetura dos plastomas na família não foi explorada até então, principalmente em um contexto comparativo entre os clados.

1.2. Sintenia gênica em plastomas ao longo da árvore da vida das plantas

O termo *sintenia* refere-se a loci genéticos no mesmo cromossomo (Passarge et al. 1999) e os estudos sintênicos analisam a posição relativa dos genes (e outros elementos) nos genomas. Os genomas plastidiais são constituídos por um cromossomo circular, o que é uma das evidências da sua origem bacteriana (Stadnichuk & Kusnetsov 2021). Estudos de sintenia dos genomas plastidiais têm se mostrado úteis em revelar eventos moleculares comuns a determinados clados em diversas

regiões da árvore do Reino Plantae. Lee et al. (2016) compararam a arquitetura de plastomas em um conjunto de dados filogeneticamente amplo, incluindo algas vermelhas, algas verdes e plantas vasculares, e observaram uma alta conservação na posição relativa dos genes. Iha et al. (2018) revelaram que eventos de inserções de sequências de plasmídeos nos genomas plastidiais foram relevantes na evolução em Gracilariaceae (Rhodophyta). Também em Rhodophyta, Jesus et al. (2023) observaram uma série de perdas gênicas e inversões em genomas organelares de espécies da família Cystocloniaceae, especialmente nos plastomas.

Diferentes dos genomas de algas vermelhas, os genomas de angiospermas apresentam duas regiões repetidas invertidas (IR1 e IR2), uma região maior de genes com cópias únicas denominada LSC (*Large Single Copy*) e uma região menor também com cópias únicas, a SSC (*Small Single Copy*). Köhler et al. (2020), através de estudos de sintenia, mostraram que o plastoma de *Opuntia quimilo* K.Schum. (Cactaceae, Caryophyllales) difere do padrão de plastomas da maioria das angiospermas, contando com uma expansão da região LSC e redução da região SSC, o que se deveu à translocação de alguns genes comuns do SSC para a região IR e uma translocação massiva do IR para a região LSC. Mais estudos são necessários para compreender se estes eventos são compartilhados por outras espécies de Cactaceae, mas, de forma geral, a estrutura quadripartite do plastoma é conservada nas angiospermas (Lee et al. 2016). Análises de sintenia gênica também têm sido utilizadas com objetivos taxonômicos em outros clados de angiospermas, como em Bignoniaceae (Lamiales), onde foram úteis para resolver complexos de espécies (Firetti et al. 2017; Fonseca & Lohmann 2017; Thode & Lohmann 2019). Além disso, tais análises contribuíram para identificar pontos críticos de mutação e loci sob seleção em um gênero de Urticaceae (Rosales) (Wang et al. 2020), demonstrando a sua utilidade para identificar *hotspots* de diversidade gênica e possíveis regiões de *barcode* úteis para delimitação específica.

A ordem Malpighiales tem uma história evolutiva complexa, marcada por separação incompleta de linhagens, introgressões e transferência horizontal de genes, o que Cai et al. (2021) chamaram de "a tempestade perfeita", que dificulta inferências filogenéticas e a compreensão da evolução deste clado. Abordagens genômicas comparativas, incluindo análises de sintenia gênica, podem ser fontes de novos *insights* sobre a evolução desta linhagem. Análises de sintenia dos plastomas em Malpighiales, apesar de ainda raras, têm demonstrado resultados relevantes ou surpreendentes em alguns clados. Revelaram, por exemplo, a perda de uma região repetitiva inteira caracterizando os genomas de duas espécies de um subgênero de *Passiflora* L. (Passifloraceae)

(Cauz-Santos et al. 2020), e foram úteis para a delimitação da família Podostemaceae (Bedoya et al. 2019). Em Calophyllaceae, embora tenham alto grau de conservação, os plastomas apresentaram diversos eventos moleculares, incluindo expansão da LSC, redução de IRs, além de inversões ou perdas de regiões gênicas (Trad et al. 2021), que ajudaram a compreender melhor a evolução deste clado. Em *Byrsonima* Rich. ex Kunth (Malpighiaceae), análises de arquitetura ajudaram a identificar regiões filogeneticamente informativas, além de revelar diferenças no conteúdo gênico entre espécies, bem como a ocorrência de casos de pseudogenização (Menezes et al. 2018). Chrysobalanaceae conta com um conjunto de genomas plastidiais privilegiado em comparação com a maioria das famílias de Malpighiales. Entretanto, análises comparativas desses plastomas, incluindo análises de sintenia, ainda não foram realizadas. O presente estudo realizou a comparação entre genomas plastidiais de uma ampla amostragem de espécies de Chrysobalanaceae, interpretando os resultados no contexto de uma filogenia produzida com base em plastomas completos.

1.3. Objetivo Geral

Caracterizar a estrutura dos genomas plastidiais de 152 espécies de Chrysobalanaceae, e comparar a sintenia gênica entre esses plastomas, analisando os eventos moleculares em um contexto filogenético.

1.4. Objetivos específicos

- Caracterizar os plastomas de 152 espécies de Chrysobalanaceae;
- Comparar a arquitetura dos plastomas de espécies de todos os gêneros de Chrysobalanaceae, bem como de espécies de um grupo externo, identificando eventos como presença e ausência de genes, inversões e translocações;
- Produzir uma filogenia com base nos genes dos genomas plastidiais das espécies estudadas para plotar e interpretar os resultados moleculares em um contexto filogenético;
- Contribuir com o conhecimento da história evolutiva de uma família diversa e ecologicamente importante de Malpighiales.

2. Referências Bibliográficas

Bardon, L., Chamagne, J., Dexter, K.G., Sothers, C.A., Prance, G.T., Chave, J., 2013. Origin and evolution of Chrysobalanaceae: insights into the evolution of plants in the Neotropics. **Botanical Journal of the Linnean Society** 171: 19-37.

Bardon, L., Sothers, C.A., Prance, G.T., Malé, P.G., Xi, Z., Davis, C.C., Murienne, J., García-Villacorta, R., Coissac, E., Lavergne, S., Chave, J., 2016. Unravelling the biogeographical history of Chrysobalanaceae from plastid genomes. **American Journal of Botany** 103: 1-14.

Bedoya, A., Ruhfel, B., Philbrick, T., Madriñán, S., Bove, C., Mesterhazy, A., Olmstead, R., 2019. Plastid Genomes of Five Species of Riverweeds (Podostemaceae): Structural Organization and Comparative Analysis in Malpighiales. **Frontiers in Plant Science**. DOI: 10.3389/fpls.2019.01035.

Cai, L., Xi, Z., Lemmon, E. M., Lemmon, A. R., Mast, A., Buddenhagen, C. E., ... & Davis, C. C., 2021.The perfect storm: gene tree estimation error, incomplete lineage sorting, and ancient gene flow explain the most recalcitrant ancient angiosperm clade, Malpighiales. **Systematic Biology** 70: 491–507 .

Cardoso, D., Särkinen, T., Alexander, S., Amorim, A. M., Bittrich, V., Celis, M., ... & Forzza, R. C., 2017. Amazon plant diversity revealed by a taxonomically verified species list. **Proceedings of the National Academy of Science** 114: 10695–10700.

Cauz-Santos, L.A., da Costa, Z.P., Callot, C., Cauet, S., Zucchi, M.I., Bergès, H., van den Berg, C. and Vieira, M.L.C., 2020. A repertory of rearrangements and the loss of an inverted repeat region in Passiflora chloroplast genomes. **Genome biology and evolution**, *12*(10):1841-1857.

Chave, J., Sothers, C., Iribar, A., Suescun, U., Chase, M. W., & Prance, G. T., 2020. Rapid diversification rates in Amazonian Chrysobalanaceae inferred from plastid genome phylogenetics. **Botanical Journal of the Linnean Society** 194(3): 271–289.

Feitosa, E.A., Xavier, H.S. & Randau, K.P., 2012. Chrysobalanaceae: traditional uses, phytochemistry and pharmacology. **Revista Brasileira de Farmacognosia** 22(5):1181–1186.

Firetti, F., Zuntini, A.R., Gaiarsa, J.W., Oliveira, R.S., Lohmann, L.G., Van Sluys. M.A., 2017. Complete chloroplast genome sequences contribute to plant species delimitation: A case study of

the Anemopaegma species complex. **American Journal of Botany** 104(10): 1493-1509. DOI: 10.3732/ajb.1700302.

Fonseca, L.H.M., Lohmann, L.G., 2017. Plastome Rearrangements in the "Adenocalymma-Neojobertia" Clade (Bignonieae, Bignoniaceae) and Its Phylogenetic Implications. **Frontiers in Plant Science** 1(8): 1875. DOI: 10.3389/fpls.2017.01875.

Gastauer, M., Sarmento, P.S.M., Santos, V.C.A., Caldeira, C.F., Ramos, S.J., Teodoro, G.S. & Siqueira, J.O., 2020. Functional traits guide plant species selection for initial mineland rehabilitation. **Ecological Engineering** 148: 105763.

Iha, C., Grassa, C.J., Lyra, G.M., Davis, C.C., Verbruggen, H., Oliveira, M.C., 2018. Organellar genomics: a useful tool to study evolutionary relationships and molecular evolution in Gracilariaceae (Rhodophyta). **Journal of Phycology** 54: 775-787.

Jesus, P.B., de Mattos Lyra, G., Zhang, H., Fujii, M.T., Nauer, F., de Castro Nunes, J.M., Davis, C.C. and Oliveira, M.C., 2023. Phylogenomics and taxon-rich phylogenies of new and historical specimens shed light on the systematics of Hypnea (Cystocloniaceae, Rhodophyta). **Molecular Phylogenetics and Evolution** 183, p.107752.

Köhler, M., Reginato, M., Souza-Chies, T.T., Majure, L.C., 2020. Insights Into Chloroplast Genome Evolution Across Opuntioideae (Cactaceae) Reveals Robust Yet Sometimes Conflicting Phylogenetic Topologies. **Frontiers in Plant Science** 19(11): 729. DOI: 10.3389/fpls.2020.00729.

Lee, J., Cho, C.H., Park, S.I., Choi, J.W., Song, H.S., West, J.A., Bhattacharya, D. and Yoon, H.S., 2016. Parallel evolution of highly conserved plastid genome architecture in red seaweeds and seed plants. **BMC biology** 14, pp.1-16.

Malé, P.J.G., Bardon, L., Besnard, G., Coissac, E., Delsuc, F., Engel, J., Lhuillier, E., Scotti-Saintagne, C., Tinaut, A. and Chave, J., 2014. Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. **Molecular Ecology Resources**, 14(5), pp.966-975.

Menezes, A.P.A., Resende-Moreira, L.C., Buzatti, R.S.O. *et al.* 2018. Chloroplast genomes of *Byrsonima* species (Malpighiaceae): comparative analysis and screening of high divergence sequences. **Scientific Reports** 8, 2210. https://doi.org/10.1038/s41598-018-20189-4.

Passarge, E., Horsthemke, B. & Farber, R., 1999. Incorrect use of the term synteny. **Nature Genetics**, 23, 387 . https://doi.org/10.1038/70486

Prance, G.T., 1972. Flora Neotropica - Chrysobalanaceae. **Flora Neotropica Monograph 9**. New York: New York Botanical Garden. 410 p.

Prance, G.T. & White, F., 1988. The genera of Chrysobalanaceae: A study of practical and theoretical taxonomy and its relevance to evolutionary biology. **Philosophical Transactions of the Royal Society of London B** 320: 1-184.

Prance, G.T. & Sothers, C.A., 2003. Chrysobalanaceae 1: Chrysobalanus to Parinari. In: **Species plantarum: Flora of the World, Part 9**. Canberra: Australian Biological Resources.

Rankin-de-Mérona, J.M., Prance, G. T., Hutchings, R. W., Silva, M. F. D., Rodrigues, W. A., & Uehling, M. E., 1992. Preliminary results of a large-scale inventory of upland rain forest in the central Amazon. **Acta Amazônica** 22: 493–534.

Ruiz, J.E.A. & Alencar, J.C., 1999. Interpretação fenológica de cinco espécies de Chrysobalanaceae na reserva florestal Adolpho Ducke, Manaus, Amazonas, Brasil. **Acta Amazônica** [Internet], 29, 223–242.

Sothers, C., Prance, G.T., Buerki, S., Kok, R.D., Chase, M.W. 2014. Taxonomic novelties in Neotropical Chrysobalanaceae: towards a monophyletic *Couepia*. **Phytotaxa** 172(2): 176-200.

Sothers, C., Prance, G.T., Chase, M.W. 2016. Towards a monophyletic *Licania*: a new generic classification of the polyphyletic Neotropical genus *Licania* (Chrysobalanaceae). **Kew Bulletin** 71: 58.

Stadnichuk, I.N. & Kusnetsov, V.V., 2021. Endosymbiotic origin of chloroplasts in plant cells' evolution. **Russian Journal of Plant Physiology,** 68, pp.1-16.

Thode, V.A., Lohmann, L.G., 2019. Comparative Chloroplast Genomics at Low Taxonomic Levels: A Case Study Using Amphilophium (Bignonieae, Bignoniaceae). **Front in Plant Science** 19(10): 796. DOI: 10.3389/fpls.2019.00796.

Trad, R.J., Cabral, F.N., Bittrich, V., Silva, S. R. D., & Amaral, M. D. C. E. D., 2021. Calophyllaceae plastomes, their structure and insights in relationships within the clusioids. **Science Reports** 11, 20712. https://doi.org/10.1038/s41598-021-99178-z.

Wang, R.N., Milne, R.I., Du, X.Y., Liu, J. and Wu, Z.Y., 2020. Characteristics and mutational hotspots of plastomes in Debregeasia (Urticaceae). **Frontiers in Genetics**, *11*, p.729.

Yakandawala, D.; Morton, C.M., Prance, G.T., 2010. Phylogenetic Relationships of the Chrysobalanaceae Inferred from Chloroplast, Nuclear, and Morphological Data. **Annals of the Missouri Botanical Garden** 97 (2): 259-281.

Capítulo 1

Título: *Comparative architecture of plastid genomes in Chrysobalanaceae (Malpighiales)**

Autores: Rogério Katsuhito Barbosa Maruyama, Renata C. Asprino, Goia de Mattos Lyra

*A ser submetido à *Molecular Phylogenetics and Evolution* (ISSN: 1055-7903)

Comparative architecture of plastid genomes in Chrysobalanaceae (Malpighiales)

Rogério Katsuhito Barbosa Maruyama[a], Renata C. Asprino[b], Goia de Mattos Lyra[a,c]

[a]Programa de Pós-Graduação em Biodiversidade e Evolução, Instituto de Biologia, Universidade Federal da Bahia, R. Barão de Jeremoabo, s/n, Ondina, Salvador, BA, Brasil, 40170-115.

[b]Programa de Pós-Graduação em Botânica, Universidade Estadual de Feira de Santana, Av. Transnordestina, s/n, Novo Horizonte, Feira de Santana, BA, Brasil, 44036-900.

[c]Departamento de Biologia Vegetal, Instituto de Biologia Roberto Alcântara Gomes, Universidade do Estado do Rio de Janeiro, R. São Francisco Xavier, 524, Maracanã, Rio de Janeiro - RJ, 20550-013, Brasil.

Running title: Architecture of plastomes in Chrysobalanaceae

Corresponding author: katsuhitorogerio@gmail.com

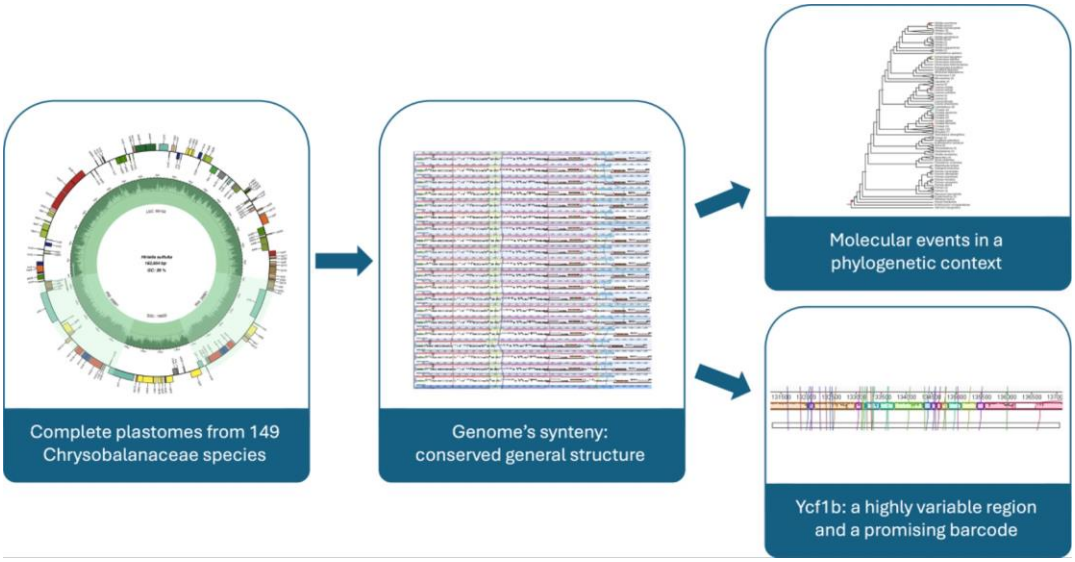
Abstract

Chrysobalanaceae is one of the families of the order Malpighiales whose taxonomy has challenged science for over two centuries. The family comprises 27 genera and approximately 550 species with a predominantly tropical distribution, spanning the Americas, Africa, and Southeast Asia. In the last 20 years, advances in molecular studies of the family have made it possible to resolve problems in the group's taxonomy, including the confirmation of its monophyly and major changes in generic delimitation. Studies involving complete plastid genomes aimed to resolve the phylogenetic relationships among the genera of the family, but other questions remain unexplored such as the characterization of structural aspects of the genomes. The present study explored the comparison of plastid genome architecture, investigating the presence of gene duplications and gene losses, in addition to variations in gene size and intron regions, with the aim of characterizing the plastomes of the species and analyzing the molecular events in a phylogenetic context. Genomic comparisons revealed regions of high variability, such as the *ycf1*b gene, which may function as a barcode for specific delimitation in Chrysobalanaceae. Our results show that the sequenced plastomes of Chrysobalanaceae species are highly conserved in terms of gene synteny, even in clades with complex biogeographic history. However, the comparison among different aspects of

the genome structure across subclades showed that such analyses could provide support for integrative taxonomic studies and for a better understanding of the evolutionary processes in Chrysobalanaceae.

Keywords: Genomics; Plastome; Synteny

Graphical abstract



Highlights

- Plastome architecture is highly conserved in species of Chrysobalanaceae.
- All analyzed Chrysobalanaceae plastomes share the gain of a second copy of *trnM(cau)*.
- Duplications and losses of genes and tRNAs occurred multiple times in the family.
- Highly variable, *ycf1*b is a promising barcode for Chrysobalanaceae.

1. Introduction

Chrysobalanaceae R.Br. (Malpighiales) comprises 27 genera and approximately 550 species distributed in tropical and subtropical regions of the Americas, Africa, and Southeast Asia, reaching northern Australia and several Pacific islands (Prance & Sothers 2003; Sothers et al. 2016). Representatives of Chrysobalanaceae can be recognized for being trees or shrubs with simple and alternate leaves, stipules axillary or sometimes fused to petiole, flowers with a nectariferous disc at the end of the receptacle, 1–3-carpelar gynoecium, gynobasic style, ovules with basal placentation and drupaceous fruits (Prance 1972; Prance & White 1988; Prance & Sothers 2003). The Neotropical region is home to about 80% of the species of Chrysobalanaceae, especially in the Amazon, where it is one of the families with the greatest richness and abundance of tree species (Rankin-de-Mérona et al. 1992; Bardon et al. 2016; Cardoso et al. al. 2017).

In the last decade, complete plastid genome sequences have been generated for a large number of Chrysobalanaceae species with the main objective of clarifying their phylogenetic relationships (Malé et al. 2014, Bardon et al. 2016, Chave et al. 2020). Such phylogenomic results greatly contributed to the understanding of the evolution and biogeographic history of the family. Some genera, however, remain polyphyletic, such as *Hirtella* L., *Leptobalanus* (Benth.) Sothers & Prance and *Hymenopus* (Benth.) Sothers & Prance (Bardon et al. 2016, Chave et al. 2020), requiring further studies that combine molecular data with morphological aspects of the species. The most complete phylogeny to date (Chave et al. 2020) has largely resolved the Chrysobalanaceae tree, especially with regard to the deeper relationships among genera. Mainly in the Neotropical clade, some of the most recently diversified clades (such as *Hirtella*) present low support for interspecific relationships.

Comparative studies of the architecture of plastid genomes have shown to be helpful in revealing molecular events common to certain clades in different branches of the Plant tree of life. Iha et al. (2018) demonstrated that events of insertions of plasmid sequences into plastid genomes were relevant to the evolution of Gracilariaceae (Rhodophyta). Also in Rhodophyta, Jesus et al. (2023) observed gene losses and inversions in organellar genomes of species of the Cystocloniaceae family, especially in plastomes. Unlike the genomes of red algae, typical plastomes of Angiosperms have two inverted repeat regions (IRa and IRb), a larger region of genes with single copies called LSC (Large Single Copy), and a smaller region also with single copies, the SSC (Small Single Copy) (Lee et al. 2016).

Lee et al. (2016) found common events in the land plants' evolution, and characterized patterns of plastome architecture in the seed plants (Gymnosperms and Angiosperms), based on a series of reductions and structural conservations, specifically losses of the IRs, invertions of the SSC, and the conservation of the genes on LSC. According to their characterization, Angiosperms present three main types of plastomes (A1, A2, and A3), the A1-type being the most frequent in the group (present in 45 orders of the 46 analyzed), including the Malpighiales (and, therefore, Chrysobalanaceae, represented in their dataset by *Hirtella physophora* Mart. & Zucc.). The A1-type plastome presents a very conserved LCS, eventual inversion of the whole SSC, and also eventual loss of one of the IRs.

Gene synteny analyses have been performed in plastomes of several clades of Angiosperms, being helpful, for example, in resolving species complexes in Bignoniaceae (Lamiales) (Firetti et al. 2017; Fonseca & Lohmann 2017; Thode & Lohmann 2019), and identifying mutation hotspots and loci under selection in *Debregeasia* Gaudich (Urticaceae, Rosales) (Wang et al. 2020). In Malpighiales, plastomes' architectural analyses were instrumental in: i) revealing the loss of an entire repetitive region characterizing the plastomes of two species of *Passiflora* L. subgenus *Deidamioides* (Passifloraceae) (Cauz-Santos et al. 2020); ii) identifying molecular synapomorphies (an important inversion in the LSC) of the Podostemaceae (Malpighiales), with systematic application (Bedoya et al. 2019); iii) disclosing several molecular events, including expansion of the LSC, reduction of IRs, and inversions or losses of genic regions in Calophyllaceae plastomes (Trad et al. 2021), which helped to better understand the evolution of this clade; iv) identifying phylogenetically informative regions for Malpighiales, in addition to revealing differences in gene content between species of *Byrsonima* Rich. ex Kunth, as well as the occurrence of cases of pseudogenization (Menezes et al. 2018).

The evolutionary history of Malpighiales is marked by incomplete lineage sorting, introgressions, and horizontal gene transfer events, which makes phylogenetic inferences and the understanding of the evolution of this clade difficult (Cai et al. 2021). Comparative analyses of the architecture of plastid genomes have not yet been performed with species of Chrysobalanaceae, even though they may encompass opportunities to shed light on the complex evolution of Malpighiales. Comparative approaches have the potential to contribute to taxonomy, revealing novel characters or mutational hotspots in the genomes, which can constitute new tools for genera and species delimitation. Aiming to be instrumental in this context, the present study has compared

the plastid genomes of a wide range of Chrysobalanaceae species, interpreting the results in the context of a phylogeny based on genes from over 150 complete plastomes.

2. Material and Methods

2.1. Plastome annotations

We obtained sequences for both ingroup and outgroups from the Chave et al. (2020) dataset available on Dryad (https://doi.org/10.5061/dryad.ghx3ffbkp). We found significant differences in gene content in the genomes from Chave's work and decided to reannotate them all using a completely annotated sample. Using *Dactyladenia bellayana* (Baill.) Prance & F.White (GenBank accession code: KX180062) as a reference,complete (149) and incomplete (3) plastomes were annotated in Geneious Prime 2020.2.4 by the annotate from tool, Annotations were manually checked to find if they were biologically viable (i.e. all CDS starts with a start codon, all genes end with a stop codon, all the intron's intervals are respected). The output table of OrthoFinder v. 2.3.1 (Emms & Kelly 2019) was used as a tool to find annotation errors on the dataset. Before concluding that the gene was lost, we analyzed the genome with the Find ORF tool, and most of the times it was found, and annotated.

2.2. Phylogenetic analyses

We extracted genes and intergenic regions using a local Python script. Another local Python script used MAFFT v.7.470 (Katoh & Standley 2013) to align each gene with an automatically selected strategy and trimmed the alignments using trimAl (Capella-Gutierrez et al. 2009). The gene alignments were manually checked and concatenated in Geneious Prime 2020.2.4. Three species from Malpighiales families closely related to Chrysobalanaceae [(*Erythroxylum novogranatens*e (D.Morris) Hieron. (Erythroxylaceae), *Garcinia mangostana* L. (Clusiaceae), and *Hevea brasiliensis* (Willd. ex A.Juss.) Müll.Arg. (Euphorbiaceae)] were used as outgroups for phylogenetic reconstructions. A Maximum Likelihood tree was constructed on CIPRES Science Gateway v.3.3 (https://www.phylo.org/) (Miller et al. 2010) using the RAxML-HPC BlackBox v.8.2.12 (Stamatakis 2014), applying the GTRCAT model, with 1000 bootstrap replications.

2.3 Genomic structure and gene synteny analyses

We analyzed the composition and structure of plastomes from 152 species of Chrysobalanaceae, investigating their total genome size (bp), GC content (%), total number of genes, protein-coding genes (CDS), and tRNA genes. We also analyzed all introns to study their variation in length, losses, gains, and duplications.

Collinear analyses were performed with 149 complete plastomes, including species of all genera in Chrysobalanaceae, representing approximately 27% of the family's biodiversity. Plastids were aligned using a progressive algorithm in Mauve v.2.3.1 (Darling et al. 2010) to identify the putative presence of structural variation between species. We aligned plastids by clade and analyzed each Mauve alignment individually. We selected, from each clade, at least one genome without any difference in architecture from the majority of the genomes in the same clade, and also every genome with any different architectural feature. A Mauve alignment was built with 37 of the selected genomes from each clade. Locally Collinear Blocks (LCBs), identified by Mauve, were color-coded. Another alignment with all 149 Chrysobalanaceae genomes and three plastomes from the outgroup was also performed to assess differences in the formation of the LCBs in a broader sample and in the presence of an external group. To assay the presence and absence of genes, we utilized OrthoFinder v. 2.3.1 (Emms & Kelly 2019), which inferred orthogroups from the dataset of circular and complete organellar genomes, using protein sequences translated from coding sequences (CDS) extracted from the genomes. Gene gain and loss propensity were analyzed by the Dollo method of parsimony, using COUNT (Csűös 2010) and printing the states at all tree nodes.

2.4. Identification of highly variable genes and analyses of genomic divergence

A highly variable gene region (*ycf1*b, varying from 4,524 to 5,790 bp in our genomes dataset) was identified in the Mauve alignment and selected, as well as two historically used genes for barcoding in angiosperms (*rbcL* and *matK*) for the application of divergence analysis methods. We performed analyses using the DNA Polymorphism tool DnaSP ver. 6.12.03 (Rozas et al. 2017) on the genes *ycf1*b, *rbcL,* and *matK* from 152 plastomes to assess the nucleotide diversity metrics throughout each gene dataset: Polymorphic Sites (S), Total Mutations (Eta), Nucleotide Diversity (Pi), and Average Nucleotide Differences (k).

We extracted these three regions (*ycf1*b, *rbcL,* and *matK*) from our plastomes dataset, aligned each of them using Geneious Prime 2020.2.4, and performed the Assemble Species by

Automatic Partitioning (ASAP) delimitation method (Puillandre et al. 2020) to test their performance as barcode genes. ASAP is based on a hierarchical clustering algorithm that uses only paired genetic distances, not using a prior biological view of intraspecific diversity. The number of subsets in each partition represents its number of primary species hypotheses. The analyses were performed using the default configuration of the online server (https://bioinfo.mnhn.fr/abi/public/asap/asapweb.html), with p distances as a model to compute the distances, groups split below the probability of 0.01, and saving the 10 best scores.

3. Results

3.1. Phylogeny of Chrysobalanaceae

Our phylogeny (Fig. 1) was generated from 82 plastid genes and had 152 species of Chrysobalanaceae represented and three from the outgroup. The Chrysobalanaceae had full support, and *Bafodeya* Prance ex F. White was the first lineage to diverge in the family. Six main fully-supported lineages were identified: i) *Kostermanthus* Prance; ii) *Neocarya* (DC.) Prance ex F.White + *Parinari* Aubl.; iii) *Magnistipula* Engl. + *Grangeria* Comm. ex Juss. + *Parastemon* A.DC. + *Atuna* Raf. + *Maranthes* Blume; iv) *Hirtella zanzibarica* Oliv. + *Dactyladenia* Welw.; v) *Chrysobalanus* L. + *Acioa* Aubl. + *Exellodendron* Prance + *Angelesia* Korth. + *Hunga* Prance; and the historically named Neotropical clade, including vi) *Geobalanus* Small + *Moquilea* Aubl. + *Couepia* Aubl. + *Leptobalanus* + *Licania* Aubl. + *Gaulettia* Sothers & Prance + *Microdesmia* (Benth.) Sothers & Prance + *Hymenopus* + *Leptobalanus apetalus* (E.Mey.) Sothers & Prance + *Hirtella*.

Most genera represented by more than one species were monophyletic and well-supported, except for *Hirtella* (as *H. zanzibarica* was fully supported as sister to the *Dactyladenia* clade), *Hymenopus* (with species divided into two non related clades), and *Leptobalanus*, as *Leptobalanus apetalus* was not nested in the main *Leptobalanus* clade. The complete tree, without collapsed genera, is presented in Figure S1.

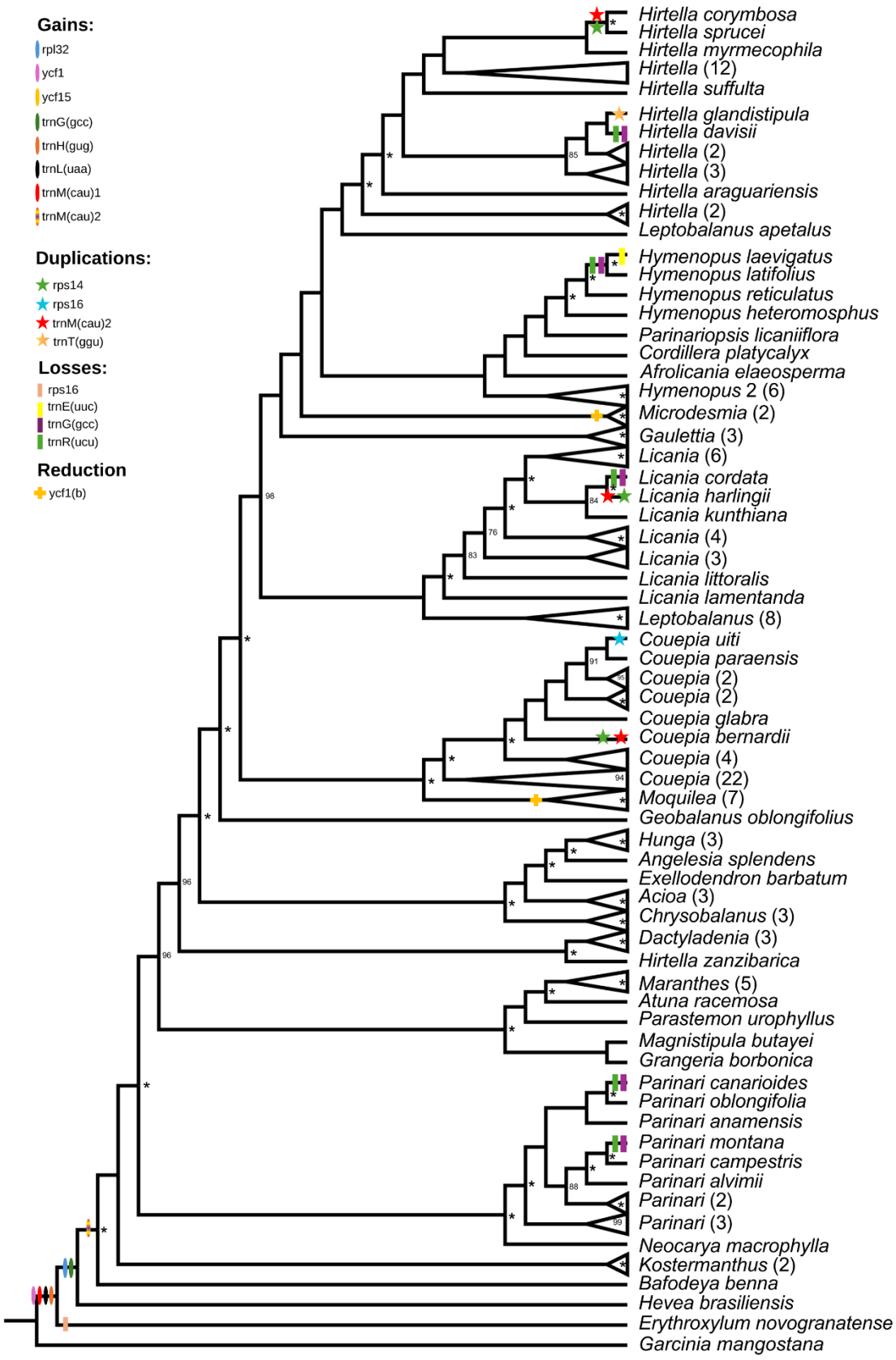

Gains:
rpl32
ycf1
ycf15
trnG(gcc)
trnH(gug)
trnL(uaa)
trnM(cau)1
trnM(cau)2
Duplications:
rps14
rps16
trnM(cau)2
trnT(ggu)
Losses:
rps16
trnE(uuc)
trnG(gcc)
trnR(ucu)
Reduction
ycf1(b)
Hirtella corymbosa
Hirtella sprucei
Hirtella myrmecophila
Hirtella (12)
Hirtella suffulta
Hirtella glandistipula
Hirtella davisii
Hirtella (2)
Hirtella (3)
Hirtella araguariensis
Hirtella (2)
Leptobalanus apetalus
Hymenopus laevigatus
Hymenopus latifolius
Hymenopus reticulatus
Hymenopus heteromosphus
Parinariopsis licaniiflora
Cordillera platycalyx
Afrolicania elaeosperma
Hymenopus 2 (6)
Microdesmia (2)
Gaulettia (3)
Licania (6)
Licania cordata
Licania harlingii
Licania kunthiana
Licania (4)
Licania (3)
Licania littoralis
Licania lamentanda
Leptobalanus (8)
Couepia uiti
Couepia paraensis
Couepia (2)
Couepia (2)
Couepia glabra
Couepia bernardii
Couepia (4)
Couepia (22)
Moquilea (7)
Geobalanus oblongifolius
Hunga (3)
Angelesia splendens
Exellodendron barbatum
Acioa (3)
Chrysobalanus (3)
Dactyladenia (3)
Hirtella zanzibarica
Maranthes (5)
Atuna racemosa
Parastemon urophyllus
Magnistipula butayei
Grangeria borbonica
Parinari canarioides
Parinari oblongifolia
Parinari anamensis
Parinari montana
Parinari campestris
Parinari alvimii
Parinari (2)
Parinari (3)
Neocarya macrophylla
Kostermanthus (2)
Bafodeya benna
Hevea brasiliensis
Erythroxylum novogranatense
Garcinia mangostana

Figure 1. Maximum likelihood (ML) tree based on 82 plastid genes. Bootstrap (BP) values ≥75% are indicated above branches. *=100% BP. Genome's architectural features are plotted on the tree. Clades with no genome architecture feature were collapsed, and numbers in parentheses indicate the number of terminals/species in collapsed clades.

3.2. Genomes' structure

Table 1 summarizes the composition and structure of the complete plastid genomes analyzed to investigate their total genome size (bp), GC content (%), total number of genes, protein-coding genes (CDS), and tRNA genes. In this study, we analyzed the genomes of 152 plant species, of which 149 are from Chrysobalanaceae, representing all 27 genera of the family. We included three species from other families of Malpighiales as outgroups as well. From the total number of genomes, only three were not complete and circularized: *Licania densiflora* Kleinhoonte, *Licania micrantha* Miq., and *Parinari curatellifolia* Planch. ex Benth.

Table 1. Summary of the genomic characteristics of 152 species of Chrysobalanaceae, including the GC content, total genome size, total number of genes, protein-coding genes, and tRNA genes. *Indicate plastomes that are incomplete in the analyzed dataset.

Genome	GC content (%)	Total genome size (bp)	Total number of genes	Protein coding genes	tRNA genes	rRNA genes	N◦ Introns
Acioa edulis	36.2	162845	131	86	37	8	24
Acioa guianensis	36.2	162327	131	86	37	8	24
Acioa longipendula	36.2	162643	131	86	37	8	24
Afrolicania elaeosperma	36.3	162171	131	86	37	8	24
Angelesia splendens	36.3	162196	131	86	37	8	24
Atuna racemosa	36.3	162712	131	86	37	8	24
Bafodeya benna	36.4	161463	131	86	37	8	24
Chrysobalanus cuspidatus	36.2	162887	131	86	37	8	24
Chrysobalanus icaco	36.2	162840	131	86	37	8	24
Chrysobalanus prancei	36.3	162731	131	86	37	8	24
Cordillera platycalyx	36.3	162455	131	86	37	8	24

Genome	GC content (%)	Total genome size (bp)	Total number of genes	Protein coding genes	tRNA genes	rRNA genes	N◦ Introns
Couepia belemii	36.3	162074	131	86	38	8	24
Couepia bernardii	36.3	162947	133	87	37	8	24
Couepia bondarii	36.3	162060	131	86	37	8	24
Couepia bracteosa	36.3	162101	131	86	37	8	24
Couepia caryophylloides	36.3	162009	131	86	37	8	24
Couepia coarctata	36.3	161801	131	86	37	8	24
Couepia excelsa	36.3	162052	131	86	37	8	24
Couepia glabra	36.3	161972	131	86	37	8	24
Couepia grandiflora	36.3	162087	131	86	37	8	24
Couepia guianensis	36.3	162122	131	86	37	8	24
Couepia habrantha	36.3	162143	131	86	37	8	24
Couepia hallwachsiae	36.3	162077	131	86	37	8	24
Couepia impressa	36.3	162014	131	86	37	8	24
Couepia joaquinae	36.3	162084	131	86	37	8	24
Couepia magnoliifolia	36.3	162142	131	86	37	8	24
Couepia monteclarensis	36.3	161885	131	86	37	8	24
Couepia morii	36.3	162126	131	86	37	8	24
Couepia multiflora	36.3	162022	131	86	37	8	24
Couepia obovata	36.3	162025	131	86	37	8	24
Couepia ovalifolia	36.3	161923	131	86	37	8	24
Couepia paraensis	36.3	161925	131	86	37	8	24
Couepia parvifolia	36.3	161897	131	86	37	8	24
Couepia polyandra	36.3	162043	131	86	37	8	24
Couepia rankiniae	36.3	162043	131	86	37	8	24
Couepia robusta	36.3	162080	131	86	37	8	24
Couepia rufa	36.3	161987	131	86	37	8	24
Couepia sandwithii	36.3	162063	131	86	37	8	24

Genome	GC content (%)	Total genome size (bp)	Total number of genes	Protein coding genes	tRNA genes	rRNA genes	N◦ Introns
Couepia schottii	36.3	161845	131	86	37	8	24
Couepia scottmorii	36.3	162161	131	86	37	8	24
Couepia spicata	36.3	161950	131	86	37	8	24
Couepia subcordata	36.3	162076	131	86	37	8	24
Couepia uiti	36.2	162819	132	87	37	8	24
Couepia ulei	36.3	162083	131	86	37	8	24
Couepia williamsii	36.3	161511	131	86	37	8	24
Dactyladenia bellayana	36.2	162281	131	86	37	8	24
Dactyladenia buchneri	36.2	162110	131	86	37	8	24
Dactyladenia floretii	36.2	162288	131	86	37	8	24
Exellodendron barbatum	36.2	162333	131	86	37	8	24
Gaulettia amaraliae	36.3	163533	131	86	37	8	25
Gaulettia elata	36.3	162246	131	86	37	8	24
Gaulettia parillo	36.3	162435	131	86	37	8	24
Geobalanus oblongifolius	36.2	162542	131	86	37	8	24
Grangeria borbonica	36.1	162792	131	86	37	8	24
Hirtella araguariensis	36.2	163072	131	86	37	8	24
Hirtella arenosa	36.3	162143	131	86	37	8	24
Hirtella bicornis	36.3	162909	131	86	37	8	24
Hirtella corymbosa	36.3	164173	133	87	38	8	24
Hirtella davisii	36.4	160638	129	86	35	8	23
Hirtella dorvalii	36.2	162872	131	86	37	8	24
Hirtella duckei	36.3	162921	131	86	37	8	24
Hirtella elongata	36.3	162684	131	86	37	8	24
Hirtella glandistipula	36.2	163798	132	86	38	8	24
Hirtella glandulosa	36.2	162929	131	86	37	8	24
Hirtella guainiae	36.3	162560	131	86	37	8	24

Genome	GC content (%)	Total genome size (bp)	Total number of genes	Protein coding genes	tRNA genes	rRNA genes	N◦ Introns
Hirtella guatemalensis	36.2	162922	131	86	37	8	24
Hirtella lemsii	36.3	164050	131	86	37	8	25
Hirtella macrosepala	36.2	162958	131	86	37	8	24
Hirtella magnifolia	36.3	162385	131	86	37	8	24
Hirtella myrmecophila	36.2	162880	131	86	37	8	24
Hirtella paniculata	36.2	163046	131	86	37	8	24
Hirtella pilosissima	36.3	162380	131	86	37	8	24
Hirtella punctillata	36.2	163079	131	86	37	8	24
Hirtella racemosa	36.2	162889	131	86	37	8	24
Hirtella recurva	36.2	163406	131	86	37	8	24
Hirtella rodriguesii	36.3	162163	131	86	37	8	24
Hirtella sprucei	36.3	164140	133	87	38	8	24
Hirtella suffulta	36.3	162646	131	86	37	8	24
Hirtella tenuifolia	36.3	162714	131	86	37	8	24
Hirtella triandra	36.2	163055	131	86	37	8	24
Hirtella zanzibarica	36.1	162341	131	86	37	8	24
Hunga gerontogea	36.3	162412	131	86	37	8	24
Hunga mackeana	36.3	162422	131	86	37	8	24
Hunga rhamnoides	36.3	162385	131	86	37	8	24
Hymenopus caudatus	36.2	162522	131	86	37	8	24
Hymenopus divaricatus	36.2	162838	131	86	37	8	24
Hymenopus glabriflorus	36.2	163005	131	86	37	8	24
Hymenopus heteromorphus	36.2	162575	131	86	37	8	24
Hymenopus intrapetiolaris	36.2	162833	131	86	37	8	24
Hymenopus laevigatus	36.4	160130	128	86	34	8	23
Hymenopus latifolius	36.4	160450	129	86	35	8	23
Hymenopus latistipulus	36.1	162989	131	86	37	8	24

Genome	GC content (%)	Total genome size (bp)	Total number of genes	Protein coding genes	tRNA genes	rRNA genes	N◦ Introns
Hymenopus macrophyllus	36.2	162672	131	86	37	8	24
Hymenopus reticulatus	36.2	162559	131	86	37	8	24
Kostermanthus heteropetalus	36.4	161909	131	86	37	8	24
Kostermanthus robustus	36.4	161783	131	86	37	8	24
Leptobalanus apetalus	36.2	162466	131	86	37	8	24
Leptobalanus emarginatus	36.3	162191	131	86	37	8	24
Leptobalanus gardneri	36.2	162284	131	86	37	8	24
Leptobalanus granvillei	36.3	162202	131	86	37	8	24
Leptobalanus humilis	36.3	162052	131	86	37	8	24
Leptobalanus latus	36.3	162055	131	86	37	8	24
Leptobalanus octandrus	36.2	162293	131	86	37	8	24
Leptobalanus sclerophyllus	36.2	162372	131	86	37	8	24
Leptobalanus sprucei	36.3	162229	131	86	37	8	24
Licania alba	36.3	162484	131	86	37	8	24
Licania canescens	36.3	162619	131	86	37	8	24
Licania cordata	36.4	160059	129	86	35	8	23
*Licania densiflora**	36.3	151263	121	80	33	8	24
Licania harlingii	36.3	163422	133	87	38	8	24
Licania hypoleuca	36.3	162430	131	86	37	8	24
Licania irwinii	36.2	162681	131	86	37	8	24
Licania kunthiana	36.3	162242	131	86	37	8	24
Licania lamentada	36.2	162678	131	86	37	8	24
Licania littoralis	36.3	162237	131	86	37	8	24
Licania majuscula	36.2	162431	131	86	37	8	24
Licania membranacea	36.3	162199	131	86	37	8	24
*Licania micrantha**	36.2	158093	127	83	36	8	24

Genome	GC content (%)	Total genome size (bp)	Total number of genes	Protein coding genes	tRNA genes	rRNA genes	N◦ Introns
Licania niloi	36.3	162513	131	86	37	8	24
Licania orbicularis	36.3	162254	131	86	37	8	24
Licania ovalifolia	36.3	162315	131	86	37	8	24
Licania pallida	36.3	162401	131	86	37	8	24
Licania rodriguesii	36.3	162309	131	86	37	8	24
Magnistipula butayei	36.1	162541	131	86	37	8	24
Maranthes gabunensis	36.2	162769	131	86	37	8	24
Maranthes glabra	36.2	162285	131	86	37	8	24
Maranthes kerstingii	36.2	162395	131	86	37	8	24
Maranthes panamensis	36.3	162457	131	86	37	8	24
Maranthes robusta	36.2	162412	131	86	37	8	24
Microdesmia arborea	36.3	161411	131	86	37	8	24
Microdesmia rigida	36.3	161507	131	86	37	8	24
Moquilea brittoniana	36.3	161337	131	86	37	8	24
Moquilea corniculata	36.3	161332	131	86	37	8	24
Moquilea guianensis	36.4	160722	131	86	37	8	24
Moquilea macrocarpa	36.3	161283	131	86	37	8	24
Moquilea minutiflora	36.3	161332	131	86	37	8	24
Moquilea pyrifolia	36.3	161387	131	86	37	8	24
Moquilea tomentosa	36.3	161449	131	86	37	8	24
Neocarya macrophylla	36.2	162405	131	86	37	8	24
Parastemon urophyllus	36.2	162676	131	86	37	8	24
Parinari alvimii	36.3	162315	131	86	37	8	24
Parinari anamensis	36.3	161805	131	86	37	8	24
Parinari campestris	36.2	162633	131	86	37	8	24
Parinari canarioides	36.5	159641	129	86	35	8	23
Parinari capensis	36.2	162524	131	86	37	8	24

Genome	GC content (%)	Total genome size (bp)	Total number of genes	Protein coding genes	tRNA genes	rRNA genes	N◦ Introns
*Parinari curatellifolia**	36.3	157967	125	83	34	8	23
Parinari excelsa	36.2	162407	131	86	37	8	24
Parinari montana	36.4	160278	129	86	35	8	23
Parinari nonda	36.3	162363	131	86	37	8	24
Parinari oblongifolia	36.2	162510	131	86	37	8	24
Parinari obtusifolia	36.2	162513	131	86	37	8	24
Parinariopsis licaniiflora	36.5	160388	131	86	37	8	24

The total genome size (bp) varied from 161,463,00 to 162,947,00. The genome with the largest genome size was *Couepia bernardii* Prance, while the genome with the smallest genome size was *Bafodeya benna* (Scott Elliot) Prance ex F.White. The GC content (%) of the genomes ranged from 36.20% to 36.40%. The genome with the highest GC content was *Bafodeya benna*, while the genome with the lowest GC content was *Acioa edulis* Prance. The total number of genes in the genomes varies from 128 to 133, although it was consistent across almost all species, with a pattern of around 131 genes. Similarly, the number of protein-coding genes was also largely consistent across all species, with each genome having 86 protein-coding genes, with the exception of five genomes that had duplication events and 87 CDS each. The majority of genomes had 37 tRNA genes, with the exception of *Couepia coarctata* Prance, which had 38 tRNA genes. In summary, the genomes analyzed in this study exhibited a similar total number of genes and protein-coding genes, and slightly varied in terms of GC content, total genome size, and tRNA gene count.

Figure 2 represents the complete annotated genome of *Hirtella suffulta* Prance, which represents the majority of the family's plastid genomic structure. The annotated genes included the following: 15 small ribosomal proteins (*rps*), 11 large ribosomal proteins (*rpl*), 4 DNA-dependent RNA polymerases (*rpo*), 8 different rRNA coding genes (*rrn*), 37 different tRNA coding genes (*trn*), 5 photosystem I proteins (*psa*), 15 photosystem II proteins (*psb*), 12 NADH dehydrogenase proteins (*ndh*), 6 cytochrome b6/f complex proteins (*pet*), 6 ATP synthase complex proteins (*atp*), the major subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase (*rbcL*), maturase K (*matK*), the proteolytic subunit of ATP-dependent Clp protease (*clpP*), membrane envelope protein (*cemA*), beta

subunit of acetyl-CoA carboxylase (*accD*), cytochrome C biogenesis protein (*ccsA*), and 6 hypothetical proteins of unknown function (*ycf*). This makes a total of 131 different genes (counting the two copies of each IR), 18 of which present introns (16 have one intron and 2 have two introns) (Table 2).

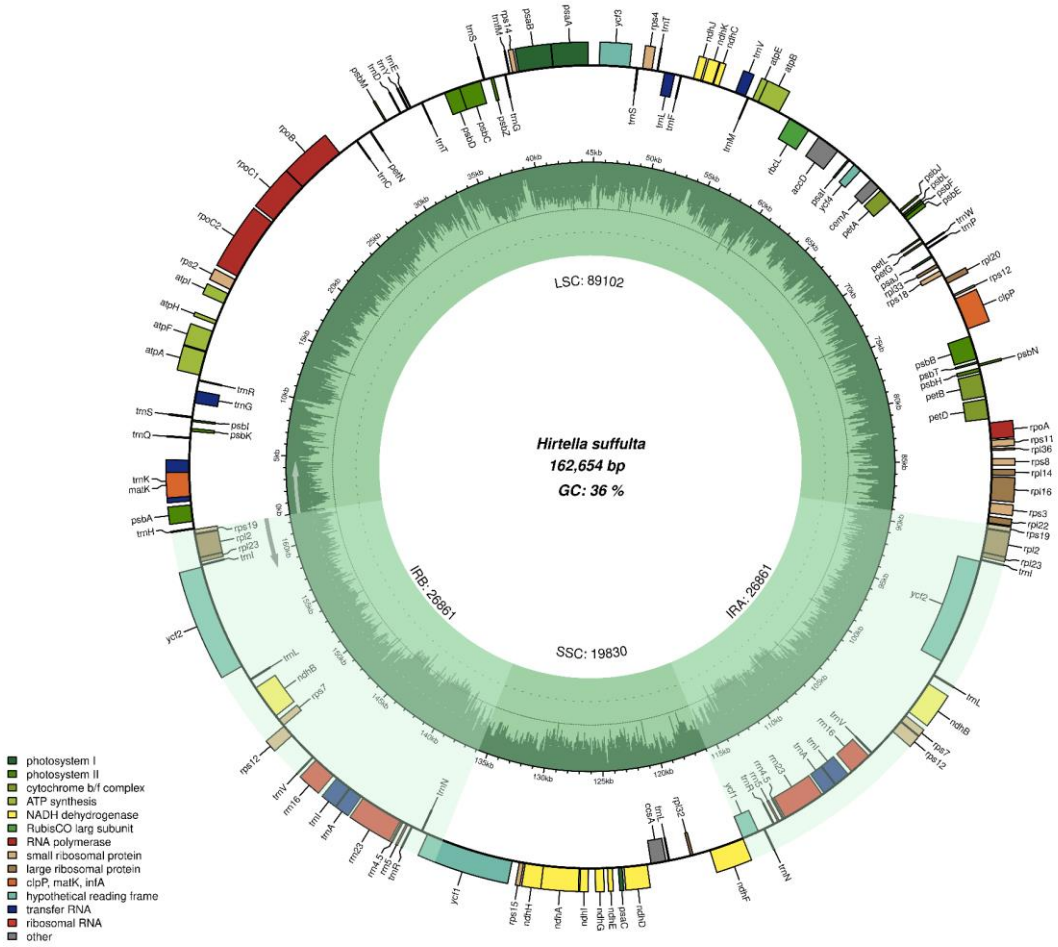


Figure 2. Chloroplast genome map of *Hirtella suffulta* representing the organization of the chloroplast genome, with genes located outside of the circular diagram transcribed in a clockwise direction, and genes inside the circle transcribed counterclockwise. The figure also labels the regions known as small single copy (SSC), large single copy (LSC), and inverted repeats (IRa, IRb).

Table 2. Information on gene's introns. Some genes have more than one intron; in this case, they are followed by their number. Copies located in IRa: (a); Copies located in IRb: (b).

Gene	Mean Length (bp)	Std. Dev. (bp)	Size range	Gene Function	Total genomes	Location
atpF intron	793,05	4,74	784-823	atp	152	LSC
clpP intron 1	865,56	11,66	835-896	clpP	152	LSC
clpP intron 2	624,56	5,204	614-653	clpP	152	LSC
ndhA intron	1188,89	9,753	1124-1221	ndh	152	SSC
ndhB intron(a)	688	0	688-688	ndh	152	IRa
ndhB intron(b)	688	0	688-688	ndh	152	IRb
petB intron	836,47	6,245	830-866	pet	152	LSC
petD intron	794,61	9,113	769-844	pet	152	LSC
rpl16 intron	1077,77	17,829	1028-1157	rpl	151	LSC
rpl2 intron(a)	675,22	1,111	675-681	rpl	152	IRa
rpl2 intron(b)	675,22	1,111	675-681	rpl	152	IRb
rpoC1 intron	787,28	4,243	781-797	rpo	152	LSC
rps12 intron(a)	536,01	0,43	535-541	rps	152	IRa
rps12 intron(b)	535,94	0,84	528-541	rps	152	IRb
trnA(ugc) intron(a)	807,88	1,192	800-813	trn	152	IRa
trnA(ugc) intron(b)	807,88	1,192	800-813	trn	152	IRb
trnG(gcc) intron	722,41	4,852	701-742	trn	146	LSC
trnI(gau) intron(a)	951,81	0,442	950-952	trn	152	IRa
trnI(gau) intron(b)	951,81	0,442	950-952	trn	152	IRb
trnK(uuu) intron	2553,57	11,427	2528-2583	trn	152	LSC
trnL(uaa) intron	579,62	9,323	559-638	trn	152	LSC
trnV(uac) intron	609,9	4,129	608-645	trn	151	LSC
ycf3 intron 1	732,78	3,605	723-757	ycf	152	LSC
ycf3 intron 2	708,89	20,91	634-887	ycf	152	LSC
ycf3 intron 3	868	19	723-757	ycf	2	LSC

3.4. Introns' variation

We analyzed 18 introns in all 152 Chrysobalanaceae genomes, with the exception of *trnV(uac)* intron, which was absent in the incomplete genome of *Parinari curatellifolia* (the summarized results are presented in Table 2). Most of the introns were located in the LSC region (13), followed by 10 in the IRs (five for each IR), and just one intron was found in the SSC (*ndhA* intron). The largest intron was in the *trnK(uuu)* intron (mean length 2553,57 bp), and the smallest one was in the *rps12* (mean length 535,94).

In terms of variation in length within the dataset, the rate was higher on the introns belonging to the LSC region, and very low on the introns belonging to the IRs. We only found differences in length when comparing both copies of the same intron from the two IRs in the *rps12* introns. Most of the genomes presented both copies of *rps*12 with 536bp in length, with the exception of *Hymenopus caudatus* (Prance) Sothers & Prance (IRa:536 and IRb:532bp) and *H. reticulatus* (Prance) Sothers & Prance (IRa:535 and IRb:528bp). Furthermore, *Hymenopus laevigatus* (Prance) Sothers & Prance and *Hymenopus latifolius* (Benth. ex Hook.f.) Sothers & Prance presented both copies with 535bp, and *Parastemon urophyllus* (Wall. ex A.DC.) A.DC. wich presented both copies of the intron measuring 541 bp.

For all other introns there were no differences in length between IRa and IRb, only slight differences in length from one genome to another. Those differences were consistent with taxonomic groupings, such as: i) the intron of *trnA(ugc),* which showed the same length for all the analyzed species (808bp), except for those in *Hunga* and *Parinariopsis* (813bp); ii) *trnI(gau)* introns (a and b), which presented the same length (952bp) for most of the genomes in our dataset, with small variations in *Hirtella* [950bp in *H. corymbosa* Cham. & Schltdl., *H. myrmecophila* Prance and *H. sprucei* Benth. ex Hook.f., and 951bp in the remaining 23 *Hirtella* representatives]. The *ndhB* intron (688 bp) presented no variation in all the analyzed genomes.

We also found two separate events of expansion of the *ycf3*, caused by the duplication of the region of the gene involving its intron 2 and exon 2 ([in *Hirtella lemsii* L.O.Williams & Prance and in *Gaulettia amaraliae* (Prance) Sothers & Prance)]. Differences in intron's contents were also observed in the outgroup. For instance, the intron 2 and exon 2 of *ycf3* are absent in *Garcinia mangostana*, and the intron of *rpl16* was also absent in *Erythroxylum novogranatense*.

3.3. Highly conserved gene synteny

Figure 1 presents the phylogenetic tree plotted with the gains, losses, and duplications of genes, introns, and tRNAs, observed in the MAUVE alignment (Fig. 3).

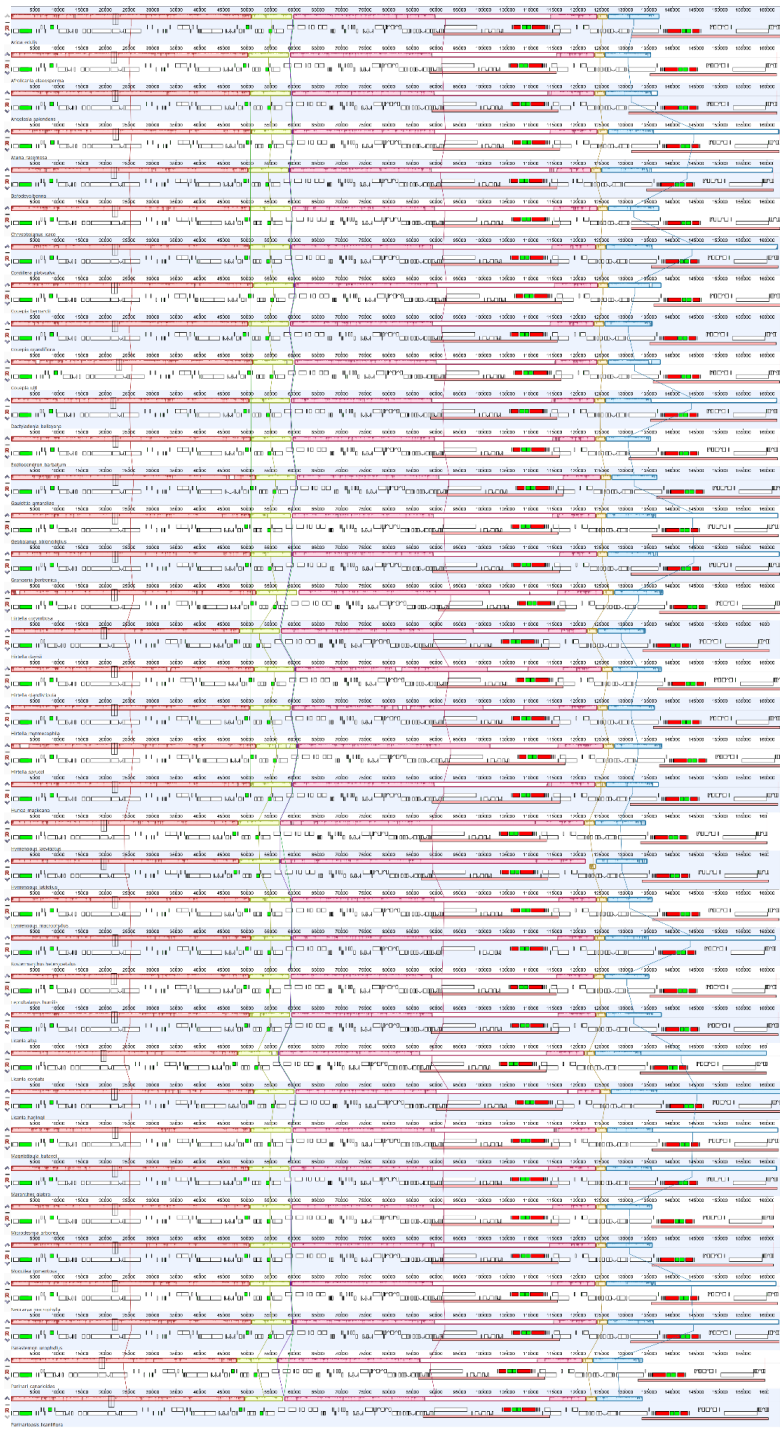


Figure 3. Progressive MAUVE alignment of plastomes from 37 species of Chrysobalanaceae. The horizontal axis above refers to genome length in base pairs (bp). Synteny between genomes is

represented by locally collinear blocks (LCB). LCB above the line is the forward strand; below the line is the reverse strand. Color lines of LCB indicate similar regions among the genomes. Gene maps below the LCBs, show protein-coding genes in white, rRNA genes in red, tRNA genes in dark green, and long bars of light red under the genes represent the repeat regions. The position of boxes above (forward) and below (reverse) line also refers to gene orientation.

We found three protein-coding genes (*rpl32*, *ycf1* and *ycf15*) and five tRNAs [*trnG(gcc)*, *trnH(gug)*, *trnL(uaa)*, *trnM(cau)*1, and *trnM(cau)*2], which were gained in different lineages of our phylogeny. All the gains occurred early in the phylogeny, mostly in the outgroups, except for the *trnM(cau)2*, whose gain is common to the whole Chrysobalanaceae. Two protein-coding genes (*rps14* and *rps16*) and two tRNAs [(*trnM(cau)*2 and *trnT(ggu)*] were involved in events of duplication across the phylogeny of Chrysobalanaceae. Three tRNAs were lost in some clades [*trnE(ucc)*, *trnG(gcc)*, and *trnR(ucu)*], and the gene *rps16* was absent in one of the species in our outgroup. Most of the events of duplications and losses happened in pairs, which might be due to the closeness of such regions in the chromosome; for instance, the simultaneous loss of *trnG(gcc)* and *trnR(ucu)* happened five times in the analyzed genomes. Together with the gain of *trnM(cau)2*, the reduction of the *ycf1*b gene (more than 1000bp at the beginning of the gene) was one of the few molecular events that were recorded for whole clades, occurring in both *Microdesmia* (two species) and *Moquilea* (seven species).

The Mauve alignment with all 149 Chrysobalanaceae + three outside groups formed 104 LCBs (Fig. S3), but no variation in the gene sinteny of all 149 Chrysobalanaceae species was found. From the three species used as outside groups, *Hevea brasiliensis* (Willd. ex A.Juss.) presented a large block (23530 bp) composed of 15 genes is inverted when compared to the rest of the genomes A final and smaller alignment using 37 genomes from all six lineages (Fig. 3) formed eight LCBs. These LCBs are identified in the Mauve alignment process as conserved segments that appear to be internally free from genome rearrangements. No rearrangements of LCBs were present among the Chrysobalanaceae genomes analyzed, which can be appreciated in Figure 3 by the lines connecting the center of each LCB with their corresponding block on the other genomes. Such lines are considerably straight, only bending in those genomes that are smaller (e.g. *Hirtella davisii* Sandwith, *Licania cordata* Prance, *Parinari canarioides* Kosterm), which is caused by a slight shift to the left of the whole genome. The last LCB (LCB8, colored light blue) is an exception, shifting abruptly along the genomes due to the lack of homogeneity in the size of this LCB. Nonetheless, vertical lines

never cross each other, indicating that there is no translocation of LCBs or genes in LCBs. No gene inversions were observed. Furthermore, by analyzing the consensus alignment to each LCB provided by the MAUVE tool, we verified that the most variable region of the analyzed plastomes is the LSC, which also presented most of the observed gains, losses, duplications, and reductions, with the exception of the events regarding the genes *rpl32* (SSC) (Fig. 1) and *ycf1*b (IR). Our Mauve alignments did not form LCBs in the IRs. In cases when an IR is included in a large LCB, this region's identity is very low.

3.5. *Ycf1*b: a greatly variable gene in Chrysobalanaceae

Analyzing the Mauve alignment for all 152 genomes (Fig. S2), we noticed the presence of multiple LCBs inside the sequence of the *ycf1*b, indicating that there might be several homologous sequences inside this gene. The alternation of the number of LCBs and their individual lengths from one genome to another visually depicts a significant amount of internal variation in this gene. Results from the DNA polymorphism tool from DNAsp software are summarized in Table 3.

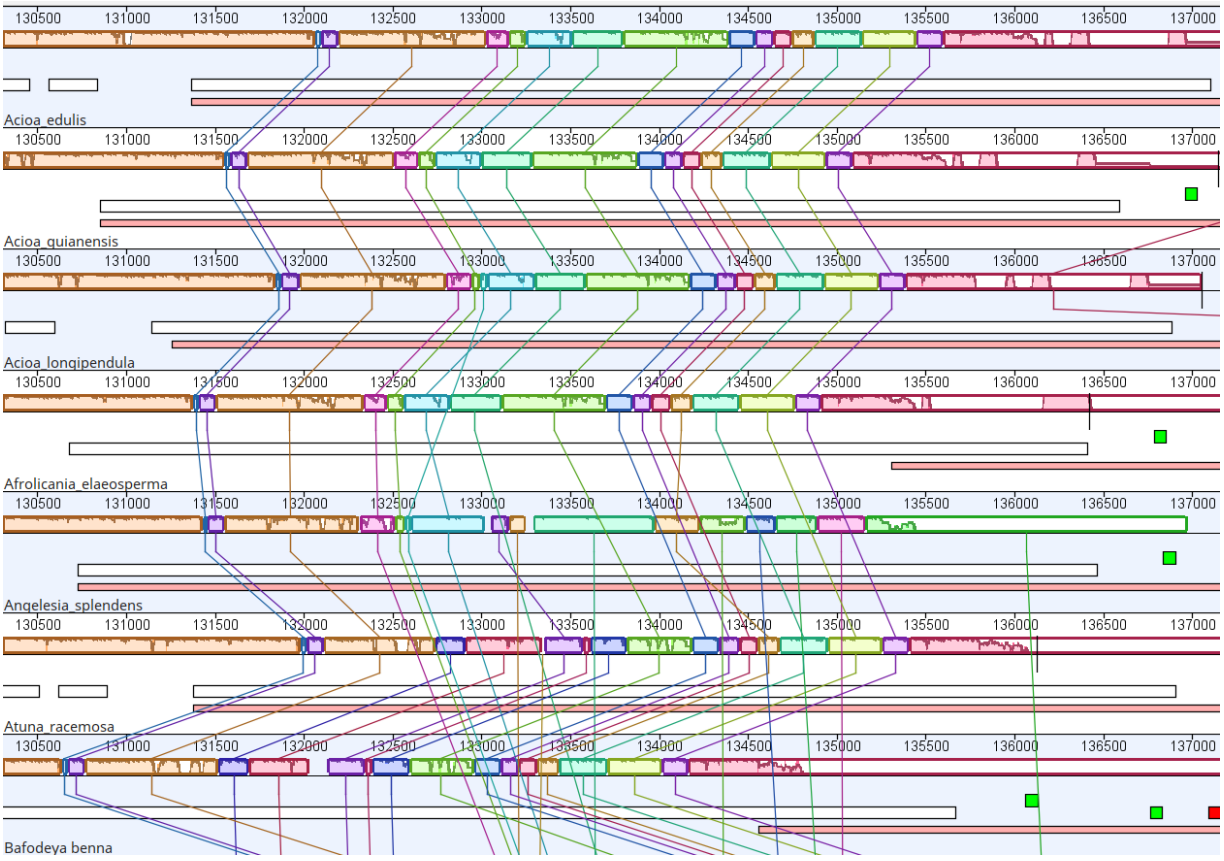


Figure 4. Zoom in the *ycf1*b area from progressive MAUVE alignment from 152 genomes. Six genomes are depicted and shows the formation of multiple LCBs (colored blocks) inside the gene (white bar under the LCBs).

Table 3. Genetic Diversity Metrics for *rbcl*, *matK*, and *ycf1b*.

Gene	Polymorphic Sites (S)	Total Mutations (Eta)	Nucleotide Diversity (Pi)	Average Nucleotide Differences (k)
rbcl	89	95	588	8.275
matK	198	214	673	10.336
*ycf1*b	833	923	1.178	46.656

*Ycf1*b has the highest scores in all metrics regarding genetic variability. The analysis recovered a total of 89 polymorphic sites(S) for *rbcl* gene, 198 sites for *matK*, and 833 sites for *ycf1*b. The total mutation (ETA) found in each gene dataset was 95 for *rbcl*, 214 for *matK*, and 923 for *ycf1*b. . The nucleotide diversity (Pi) is 588 for *rbcl*, 673 for *matK*, and 1,178 for *ycf1*b; and the average nucleotide differences (k) is 8,275 for *rbcl*, 10,336 for *matK*, and 46,656 for *ycf1*b.

Results from ASAP for *ycf1*b showed that the first partition, with a lower ASAP score (=2.00) and p-value (=0,214), was mostly useful for delimiting genera, presenting 31 subsets, and separating species only in *Microdesmia* and *Kostermanthus* (among the genera represented by more than one species). However, the second partition, with a higher ASAP score (=10, p value=0,687), formed 142 subsets and delimited most of the species, with a few exceptions of multispecies subsets: i) *Parinari oblongifolia* Hook.f. + *P. canarioides*; ii) *Parinari nonda* + *P. capensis* Harv.; iii) *Hymenopus intrapetiolaris* (Spreng. ex Hook. f.) Sothers & Prance + *H. divaricatus* (Benth.) Sothers & Prance; iv) *Hirtella racemosa* Ruiz & Pav. + *H. paniculata* Sw.; v) *Dactyladenia floretii* Breteler + *D. bellayana*; vi) *Couepia sandwithii* Prance + *C. rankiniae* Prance; vii) *Couepia rufa* Ducke + *C. impressa* Prance; viii) *Couepia parvifolia* Prance + *C. monteclarensis* Prance; ix) *Couepia magnoliifolia* Benth. ex Hook.f. + *C. habrantha* Standl.; x) *Licania membranacea* Sagot ex Laness. + *L. micrantha*.

ASAP applied to the *rbcL* dataset generated a maximum of 81 subsets (ASAP score=4.5, p-value lower than 0.001). Results for *matK* generated up to 152 subsets, distinguishing all the species from the dataset, with ASAP score of 9.0 and p value=1. The *matK* partition (ASAP score

4.5, p value=0,589) that presented a p-value closer to the one of the second partition from the *ycf1* alignment (0,651) generated only 46 subsets. When applied to concatenated datasets, ASAP results from *matK+ycf1*b generated up to 143 subsets (ASAP score 6, p value=0,413), and results from *rbcL+matK+ycf1*b also formed a maximum of 143 subsets, with lower confidence (ASAP score 9, p value=0,667).

4. Discussion

4.1. Phylogenetic relationships in Chrysobalanaceae

Our phylogeny (Fig. 1) was largely consistent with the results from Chave et al. (2020), with the exception that our analyses did not recover a sister relationship between *Kostermanthus* and *Bafodeya*, the latter being the first lineage to diverge in Chrysobalanaceae. Other differences involved weakly supported relations, such as the placement of *Parastemon* as closely related to the clade comprising *Atuna* and *Maranthes* (67% BP, Fig. S1), while Chave et al. (2020) found the clade including *Grangeria* and *Magnistipula* to be closer to the *Atuna* + *Maranthes* clade (also with low support). A third combination was found by Bardon et al. (2016), who recovered a well-supported sister relationship between *Parastemon* and *Magnistipula*, this clade being sister to the clade comprising *Atuna* and *Maranthes*, demonstrating that the currently available sampling (both in genes and taxa) may be insufficient for understanding the evolution of this lineage.

The paraphyly of *Hirtella* has been well documented (Bardon et al. 2016, Chave et al. 2020), and morphological analyses are needed to evaluate if *H. zanzibarica* is consistent with the circumscription of *Dactyladenia*, making the appropriate taxonomic changes. Further analyses of the sampled specimen of *Leptobalanus apetalus* are also needed to determine if its identification is accurate and to make proper taxonomic decisions.

4.2. Genome's structure and intron's variation

The number of introns found in Chrysobalanaceae (18) is consistent with the majority of the Angiosperms (Jansen et al. 2007). Most of the differences in the length of the introns studied were found in the LSC region, in opposition to the highly conserved lengths found in the IRs, including the lengths of both copies of the same introns. These findings align with broader observations in plant genomics (Jansen et al. 2007; Caycho 2023; Long 2023). The IR regions are known for their stability and reduced variability compared to LSC and SSC regions (Lee et al. 2016). This stability is likely due to the duplication of the IR regions, providing a mechanism for error correction during

DNA replication, thus maintaining sequence conservation. The studies of Wang et al. (2022) on *Nicotina* L. (Solanaceae) support this observation, noting significantly lower SNP (Single Nucleotide Polymorphism) density and indel (insertion-deletion) events in IR regions, in comparison with LSC and SSC regions. Interestingly, the discrete differences found in the lengths of the introns in the IR regions were correlated to some clades on their phylogeny.

4.2. Highly conserved gene synteny in plastomes of Chrysobalanaceae species

Our synteny analysis, which utilized a broad sampling including all the 27 currently accepted genera, found that the family's clade has a very conserved plastidial architecture, confirming the A1 plastome type (as Lee et al. 2016 suggested) for all the analyzed species. There were no losses of any IR, an event that is present in several plant clades (Lee et al. 2016), and also documented in Malpighiales [e.g., Passifloraceae (Cauz-Santos et al. 2020)]. This is particularly informative since the conservation of both IRs has been associated with the maintenance of plastomes' structure (Turudić et al., 2022). Moreover, Geraniaceae (Geraniales) exhibits extreme events regarding the IRs such as: IR losses, massive expansions into the SSC region, and extensive reductions of the IR region. Their plastomes are recognized as the most rearranged ones among angiosperms (Blazier et al. 2011). Similar patterns are also found in Passifloraceae (Cauz-Santos et al. 2020) and Cactaceae (Sanderson et al. 2015), in which alterations and losses of IRs might be correlated to high variability in genome structure. This highlights the unique evolutionary trajectory of Chrysobalanaceae, and the conserved nature of its plastomes underscores the significance of IR regions in plastid genome evolution and stability.

In our final synteny analysis (Fig. 3) using representatives of all genera, we identified eight Locally Collinear Blocks (LCBs), they presented conserved genomic segments, which are free from translocations or inversions. The absence of such rearrangements suggests that these regions have remained stable across the genomes analyzed. This stability is crucial for comparative genomic studies, as it indicates regions where orthologous sequences can be reliably compared without the confounding effects of recombination-induced rearrangements (Darling et al. 2010). An example that might be a implication of this characteristic of the Mauve alignment is the absence of LCBs in the IRs, since sequences belonging to IRs are under homologous recombination rearrangements (Maréchal & Brisson 2010).

Since the synteny of the analyzed genomes was highly homogeneous, the formation of multiple LCBs is mostly due to differences in the composition of the nucleotides in the genomes (i.e.,

substitutions, deletions, insertions, minor inversions, and translocations), and those happened on the level of genes and intergenic regions. In the LCB6 (dark pink) and LCB8 (light blue), the identity level, represented by the white bars inside the LCB, was nearly zero in the IR regions. It is especially visible in the LCB8, which had a shortage in almost all the genomes, and the IRb was excluded from this block and did not form any other LCB. Understanding the conservation and stability of LCBs provides insights into the evolutionary relationships and genomic integrity among the studied species.

.

4.3. *Ycf1*b: a good barcode candidate in Chrysobalanaceae

The comparison of *rbcl*, *matK*, and *ycf1*b reveals varying degrees of genetic diversity and evolutionary dynamics within this dataset. While all three genes demonstrate substantial genetic variation, *ycf1*b emerges as the most genetically diverse among the three, exhibiting higher values in polymorphic sites, mutations, haplotype diversity, and nucleotide diversity (Table 3). These findings suggest that *ycf1*b harbors a richer reservoir of genetic variation, making it a promising candidate for barcoding and phylogenetic studies within the Chrysobalanaceae. This finding is particularly relevant to this family, as all analyzed Chrysobalanaceae genomes maintain IR regions, where *ycf1* is present, which means that it can be useful for species delimitation in virtually all genera. The use of *ycf1* as a barcode is not as promising for several other families of plants, as it is lost in various lineages (Mohanta et al. 2020). Chrysobalanaceae encompasses species complexes in multiple genera [e.g., *Hirtella*, *Leptobalanus*, and *Hymenopus* (Prance 2020)], and species delimitation based on morphological characters is often challenging. In light of molecular data, historically used characters have become obsolete as they did not reflect the evolutionary history of clades (Prance 2020). In this scenario, it is not possible to be completely confident that the failure in delimiting the ten mentioned pairs of species was related to the lack of effectiveness of *ycf1*b, or if we could be referring to possible cases of hiper estimation of species diversity. The taxonomy of Chrysobalanaceae species could benefit from studies focused on intraspecific morphological variation, investigating species plasticity.

Recent studies also highlight *ycf*1 as one of the most variable regions in chloroplast genomes, and potentially helpful as DNA barcoding of closely related plant species (Dong et al. 2012, 2015; Sun et al. 2022; Long et al. 2023). The higher nucleotide diversity (Pi values) and number of polymorphic sites (S) in *ycf1* compared to other chloroplast markers such as *rbcL* and

matK indicate its superior potential for genetic studies (Dong et al. 2012). Additionally, *ycf1*b's significant variation aligns with its high mutation rate and sequence divergence observed in various plant genera, further establishing it as a key marker for species delimitation analyses (Sun et al. 2022).

Our results from ASAP also indicate that *ycf1*b may be a suitable barcode candidate for Chrysobalanaceae, performing better than *rbcL* and similarly to *matK* in species delimitation. It's important to note that *matK* and *rbcL* still offer valuable genetic information and could complement *ycf1*b in studies focused on species delimitation. Our ASAP results indicate that a combination of *matK* and *ycf1*b constitutes a more powerful tool to be used in integrative taxonomy than each of these genes alone, or in combination with *rbcL*.

5. Conclusions

The plastid genomes of Chrysobalanaceae species are highly conserved in terms of gene synteny, which makes the family a particularly interesting case of study on the maintenance of plastome structure, given its complex biogeographic history. Identified instances of gene duplications and losses, as well as variations in gene sizes and regions of introns, may be an additional step to understanding the evolutionary mechanisms shaping plastid genomes in Chrysobalanaceae. Mauve Alignments were helpful in identifying the presence of highly variable regions, such as the *ycf1b* gene, a potential barcode for species delimitation in the family. Our phylogenomic results indicate the need for more comprehensive sampling and integrative approaches combining molecular and morphological data in Chrysobalanaceae. Genomic comparisons of a broader dataset in terms of number of species, especially in unresolved clades, have the potential to reveal synapomorphies, which is crucial for resolving taxonomic ambiguities and achieving a stable classification. In this sense, the selection of a set of candidate morphological characters for species and genera delimitation is crucial for analyses of ancestral character evolution, which can provide the basis for a taxonomy that reflects the evolutionary history of the group. The identification of molecular characters with implications for taxonomy also echoes the call for integrative approaches, which have been successfully applied in other plant groups (Firetti et al. 2017; Bedoya et al. 2019).

6. Acknowledgments

The authors thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for a scholarship to RKBM.

7. References

Bardon, L., Sothers, C.A., Prance, G.T., Malé, P.G., Xi, Z., Davis, C.C., Murienne, J., García-Villacorta, R., Coissac, E., Lavergne, S., Chave, J., 2016. Unravelling the biogeographical history of Chrysobalanaceae from plastid genomes. **American Journal of Botany** 103: 1-14.

Bedoya, A., Ruhfel, B., Philbrick, T., Madriñán, S., Bove, C., Mesterhazy, A., Olmstead, R.., 2019. Plastid Genomes of Five Species of Riverweeds (Podostemaceae): Structural Organization and Comparative Analysis in Malpighiales. **Frontiers in Plant Science**, 10, p.1035. DOI: 10.3389/fpls.2019.01035.

Blazier C.J., Guisinger, M.M., Jansen, R.K., 2011. Recent loss of plastid-encoded ndh genes within Erodium (Geraniaceae). **Plant molecular biology**, 76(3-5): 263-72. doi: 10.1007/s11103-011-9753-5.

Cai, L. *et al.* The perfect storm: gene tree estimation error, incomplete lineage sorting, and ancient gene flow explain the most recalcitrant ancient angiosperm clade, Malpighiales. **Syst. Biol.** 70, 491–507 (2021).

Capella-Gutiérrez, S., Silla-Martínez, J.M., Gabaldón, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses, **Bioinformatics** 25(15): 1972–1973, https://doi.org/10.1093/bioinformatics/btp348.

Cardoso, D.B.O.S. *et al.*, 2017. Amazon plant diversity revealed by a taxonomically verified species list. **Proceedings of the National Academy of Science** 114: 10695–10700.

Cauz-Santos, L.A., da Costa, Z.P., Callot, C., Cauet, S., Zucchi, M.I., Bergès, H., van den Berg, C. and Vieira, M.L.C., 2020. A repertory of rearrangements and the loss of an inverted repeat region in Passiflora chloroplast genomes. **Genome biology and evolution**, *12*(10), pp.1841-1857.

Chave, J. *et al.*, 2020. Rapid diversification rates in Amazonian Chrysobalanaceae inferred from plastid genome phylogenetics. **Botanical Journal of the Linnean Society** 194(3): 271–289.

Csűös, M. 2010. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. **Bioinformatics** 26: 1910–1912.

Caycho, E., La Torre R, O.G., 2023. Assembly, annotation and analysis of the chloroplast genome of the Algarrobo tree Neltuma pallida (subfamily: Caesalpinioideae). **BMC Plant Biol**. 16;23(1):570. doi: 10.1186/s12870-023-04581-5.

Darling, A. E., Mau, B. & Perna, N. T. 2010. progressive Mauve: multiple genome alignment with gene gain, loss and rearrangement. **PLoS ONE** 5: e11147.

Dong W, Liu J, Yu J, Wang L, Zhou S., 2012. Highly Variable Chloroplast Markers for Evaluating Plant Phylogeny at Low Taxonomic Levels and for DNA Barcoding. **PLOS ONE** 7(4): 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

Emms, D. M., & Kelly, S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. **Genome biology** 20: 1–14.

Feitosa, E.A., Xavier, H.S. & Randau, K.P., 2012. Chrysobalanaceae: traditional uses, phytochemistry and pharmacology. **Rev. Bras. Farmacogn.** 22(5):1181–1186.

Firetti, F., Zuntini, A.R., Gaiarsa, J.W., Oliveira, R.S., Lohmann, L.G., Van Sluys. M.A., 2017. Complete chloroplast genome sequences contribute to plant species delimitation: A case study of the Anemopaegma species complex. **Am J Bot**. 104(10): 1493-1509. DOI: 10.3732/ajb.1700302.

Fonseca, L.H.M., Lohmann, L.G., 2017. Plastome Rearrangements in the "Adenocalymma-Neojobertia" Clade (Bignonieae, Bignoniaceae) and Its Phylogenetic Implications. **Front Plant Sci.** 1(8): 1875. DOI: 10.3389/fpls.2017.01875.

Gastauer, M., Sarmento, P.S.M., Santos, V.C.A., Caldeira, C.F., Ramos, S.J., Teodoro, G.S. & Siqueira, J.O., 2020. Functional traits guide plant species selection for initial mineland rehabilitation. **Ecological Engineering** 148: 105763.

Iha, C., Grassa, C.J., Lyra, G.M., Davis, C.C., Verbruggen, H., Oliveira, M.C., 2018. Organellar genomics: a useful tool to study evolutionary relationships and molecular evolution in Gracilariaceae (Rhodophyta). **J. Phycol.** 54: 775-787.

Jansen, R.K., Cai, Z., Raubeson, L.A., Daniell, H., Depamphilis, C.W., Leebens-Mack, J., Müller, K.F., Guisinger-Bellian, M., Haberle, R.C., Hansen, A.K., Chumley, T.W., Lee, S.B., Peery, R., McNeal, J.R., Kuehl, J.V., Boore, J.L., 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. **Proc Natl Acad Sci USA**, 104(49):19369-74.

Jesus, P.B., de Mattos Lyra, G., Zhang, H., Fujii, M.T., Nauer, F., de Castro Nunes, J.M., Davis, C.C. and Oliveira, M.C., 2023. Phylogenomics and taxon-rich phylogenies of new and historical specimens shed light on the systematics of Hypnea (Cystocloniaceae, Rhodophyta). **Molecular Phylogenetics and Evolution** 183, p.107752.

Rozas, J., Ferrer-Mata, A., Sánchez-DelBarrio, J.C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S.E., Sánchez-Gracia, A., 2017. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets, **Molecular Biology and Evolution** 34(2): 3299–3302, https://doi.org/10.1093/molbev/msx248.

Katoh, K. and Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. **Molecular biology and evolution** 30(4), pp.772-780.

Lee, J., Cho, C.H., Park, S.I., Choi, J.W., Song, H.S., West, J.A., Bhattacharya, D. and Yoon, H.S., 2016. Parallel evolution of highly conserved plastid genome architecture in red seaweeds and seed plants. **BMC biology**, 14, pp.1-16.

Long, L., Li, Y., Wang, S., Liu, Z., Wang, J. and Yang, M., 2023. Complete chloroplast genomes and comparative analysis of Ligustrum species. **Scientific Reports** *13*(1), p.212.

Malé, P.J.G., Bardon, L., Besnard, G., Coissac, E., Delsuc, F., Engel, J., Lhuillier, E., Scotti-Saintagne, C., Tinaut, A. and Chave, J., 2014. Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. **Molecular ecology resources**, 14(5), pp.966-975.

Maréchal, A. and Brisson, N. (2010), Recombination and the maintenance of plant organelle genome stability. **New Phytologist**, 186: 299-317. https://doi.org/10.1111/j.1469-8137.2010.03195.x

Menezes, A.P.A., Resende-Moreira, L.C., Buzatti, R.S.O. *et al.* 2018. Chloroplast genomes of *Byrsonima* species (Malpighiaceae): comparative analysis and screening of high divergence sequences. **Sci Rep** 8, 2210. https://doi.org/10.1038/s41598-018-20189-4

Miller, M.A., Pfeiffer, W., Schwartz, T., 2010. Creating the CIPRES Science Gateway for Inference of Large Phylogenetic Trees. Gateway Computing Environments Workshop (GCE), 14 November, New Orleans, Louisiana. https://www.phylo.org/.

Mohanta, T.K., Mishra, A.K., Khan, A., Hashem, A., Abd_Allah, E.F. and Al-Harrasi, A., 2020. Gene loss and evolution of the plastome. **Genes**, 11(10), p.1133.

Prance, G.T., 1972. Flora Neotropica - Chrysobalanaceae. **Flora Neotropica Monograph 9**. New York: New York Botanical Garden. 410 p.

Prance, G.T. & White, F., 1988. The genera of Chrysobalanaceae: A study of practical and theoretical taxonomy and its relevance to evolutionary biology. **Philosophical Transactions of the Royal Society of London B** 320: 1-184.

Prance, G.T. & Sothers, C.A., 2003. Chrysobalanaceae 1: Chrysobalanus to Parinari. In: **Species plantarum: Flora of the World, Part 9**. Canberra: Australian Biological Resources.

Prance, G.T., 2021. Sixty years with the Chrysobalanaceae. **The Botanical Review**, *87*, pp.197-232.

Puillandre, N., Brouillet, S., Achaz, G. 2020. ASAP: Assemble Species by Automatic Partitioning. **Molecular Ecology Resources.** https://doi.org/10.1111/1755-0998.13281.

Rankin-de-Mérona, J.M. *et al.*, 1992. Preliminary results of a large-scale inventory of upland rain forest in the central Amazon. **Acta Amazônica** 22: 493–534.

Sanderson, M.J., Copetti, D., Burquez, A., Bustamante, E., Charboneau, J.L.M., Eguiarte, L.E., Kumar, S., Lee, H.O., Lee, J., McMahon, M., et al, 2015. Exceptional Reduction of the Plastid Genome of Saguaro Cactus (Carnegiea Gigantea): Loss of the Ndh Gene Suite and Inverted Repeat. **Am. J. Bot**., 102: 1115–1127. https://doi.org/10.3732/ajb.1500184.

Sothers, C., Prance, G.T., Buerki, S., Kok, R.D., Chase, M.W., 2014. Taxonomic novelties in Neotropical Chrysobalanaceae: towards a monophyletic *Couepia*. **Phytotaxa** 172(2): 176-200.

Sothers, C., Prance, G.T., Chase, M.W., 2016. Towards a monophyletic *Licania*: a new generic classification of the polyphyletic Neotropical genus *Licania* (Chrysobalanaceae). **Kew Bulletin** 71: 58.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. **Bioinformatics** 30(9), pp.1312-1313.

Sun, J., Wang, S., Wang, Y., Wang, R., Liu, K., Li, E., Qiao, P., Shi, L., Dong, W., Huang, L., & Guo, L., 2022. Phylogenomics and Genetic Diversity of Arnebiae Radix and Its Allies (Arnebia, Boraginaceae) in China. **Frontiers in Plant Science**, 13.

Thode, V.A., Lohmann, L.G., 2019. Comparative Chloroplast Genomics at Low Taxonomic Levels: A Case Study Using Amphilophium (Bignonieae, Bignoniaceae). **Front Plant Sci**. 19(10): 796. DOI: 10.3389/fpls.2019.00796.

Trad, R.J., Cabral, F.N., Bittrich, V. *et al.*, 2021. Calophyllaceae plastomes, their structure and insights in relationships within the clusioids. **Sci Rep** 11, 20712. https://doi.org/10.1038/s41598-021-99178-z.

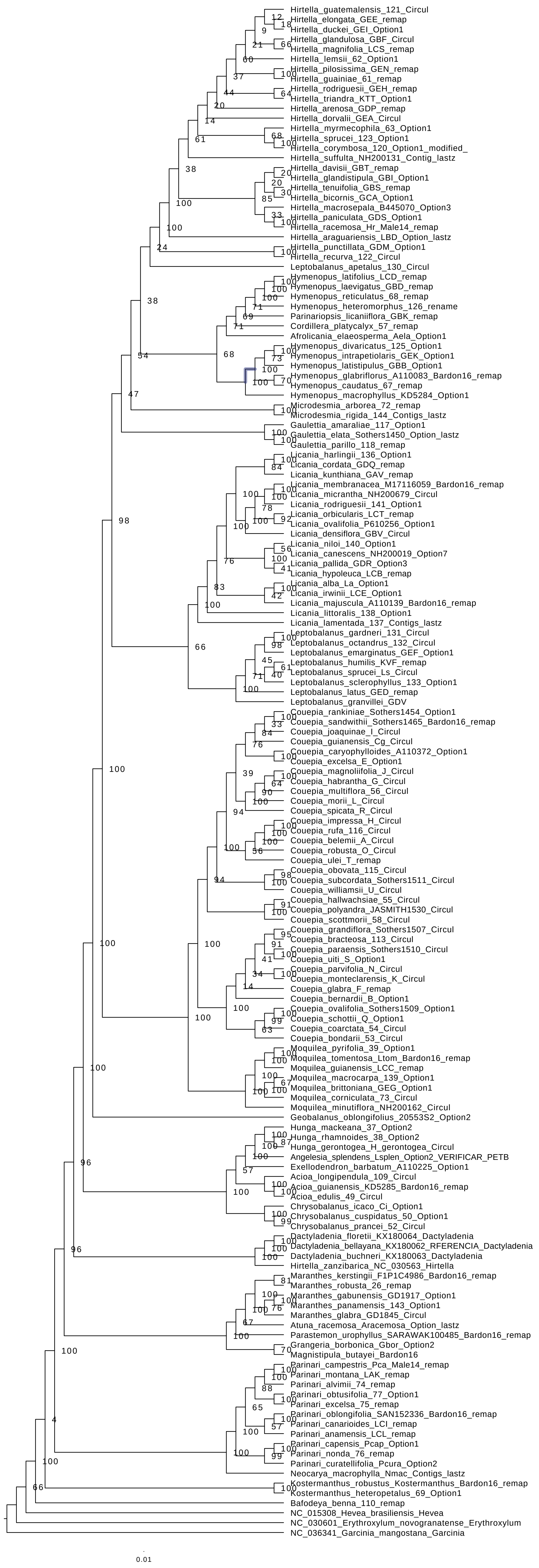
Turudić, A., Liber, Z., Grdiša, M., Jakše, J., Varga, F., Šatović, Z., 2022. Chloroplast Genome Annotation Tools: Prolegomena to the Identification of Inverted Repeats. **Int J Mol Sci**. 16;23(18):10804. doi: 10.3390/ijms231810804.

Wang, R.N., Milne, R.I., Du, X.Y., Liu, J. and Wu, Z.Y., 2020. Characteristics and mutational hotspots of plastomes in Debregeasia (Urticaceae). **Frontiers in genetics**, *11*, p.729.

Wang, S., Gao, J., Chao, H., Li, Z., Pu, W., Wang, Y. and Chen, M., 2022. Comparative chloroplast genomes of Nicotiana Species (Solanaceae): Insights into the genetic variation, phylogenetic relationship, and polyploid speciation. **Frontiers in plant science** 13, p.899252.

Yakandawala, D.; Morton, C.M., Prance, G.T., 2010. Phylogenetic Relationships of the Chrysobalanaceae Inferred from Chloroplast, Nuclear, and Morphological Data. **Annals of the Missouri Botanical Garden** 97 (2): 259-281.

Anexos


Hirtella_guatemalensis_121_Circul
Hirtella_elongata_GEE_remap
Hirtella_duckei_GEI_Option1
Hirtella_glandulosa_GBF_Circul
Hirtella_magnifolia_LCS_remap
Hirtella_lemsii_62_Option1
Hirtella_pilosissima_GEN_remap
Hirtella_guainiae_61_remap
Hirtella_rodriguesii_GEH_remap
Hirtella_triandra_KTT_Option1
Hirtella_arenosa_GDP_remap
Hirtella_dorvalii_GEA_Circul
Hirtella_myrmecophila_63_Option1
Hirtella_sprucei_123_Option1
Hirtella_corymbosa_120_Option1_modified_
Hirtella_suffulta_NH200131_Contig_lastz
Hirtella_davisii_GBT_remap
Hirtella_glandistipula_GBI_Option1
Hirtella_tenuifolia_GBS_remap
Hirtella_bicornis_GCA_Option1
Hirtella_macrosepala_B445070_Option3
Hirtella_paniculata_GDS_Option1
Hirtella_racemosa_Hr_Male14_remap
Hirtella_araguariensis_LBD_Option_lastz
Hirtella_punctillata_GDM_Option1
Hirtella_recurva_122_Circul
Leptobalanus_apetalus_130_Circul
Hymenopus_latifolius_LCD_remap
Hymenopus_laevigatus_GBD_remap
Hymenopus_reticulatus_68_remap
Hymenopus_heteromorphus_126_rename
Parinariopsis_licaniiflora_GBK_remap
Cordillera_platycalyx_57_remap
Afrolicania_elaeosperma_Aela_Option1
Hymenopus_divaricatus_125_Option1
Hymenopus_intrapetiolaris_GEK_Option1
Hymenopus_latistipulus_GBB_Option1
Hymenopus_glabriflorus_A110083_Bardon16_remap
Hymenopus_caudatus_67_remap
Hymenopus_macrophyllus_KD5284_Option1
Microdesmia_arborea_72_remap
Microdesmia_rigida_144_Contigs_lastz
Gaulettia_amaraliae_117_Option1
Gaulettia_elata_Sothers1450_Option_lastz
Gaulettia_parillo_118_remap
Licania_harlingii_136_Option1
Licania_cordata_GDQ_remap
Licania_kunthiana_GAV_remap
Licania_membranacea_M17116059_Bardon16_remap
Licania_micrantha_NH200679_Circul
Licania_rodriguesii_141_Option1
Licania_orbicularis_LCT_remap
Licania_ovalifolia_P610256_Option1
Licania_densiflora_GBV_Circul
Licania_niloi_140_Option1
Licania_canescens_NH200019_Option7
Licania_pallida_GDR_Option3
Licania_hypoleuca_LCB_remap
Licania_alba_La_Option1
Licania_irwinii_LCE_Option1
Licania_majuscula_A110139_Bardon16_remap
Licania_littoralis_138_Option1
Licania_lamentada_137_Contigs_lastz
Leptobalanus_gardneri_131_Circul
Leptobalanus_octandrus_132_Circul
Leptobalanus_emarginatus_GEF_Option1
Leptobalanus_humilis_KVF_remap
Leptobalanus_sprucei_Ls_Circul
Leptobalanus_sclerophyllus_133_Option1
Leptobalanus_latus_GED_remap
Leptobalanus_granvillei_GDV
Couepia_rankiniae_Sothers1454_Option1
Couepia_sandwithii_Sothers1465_Bardon16_remap
Couepia_joaquinae_I_Circul
Couepia_guianensis_Cg_Circul
Couepia_caryophylloides_A110372_Option1
Couepia_excelsa_E_Option1
Couepia_magnoliifolia_J_Circul
Couepia_habrantha_G_Circul
Couepia_multiflora_56_Circul
Couepia_morii_L_Circul
Couepia_spicata_R_Circul
Couepia_impressa_H_Circul
Couepia_rufa_116_Circul
Couepia_belemii_A_Circul
Couepia_robusta_O_Circul
Couepia_ulei_T_remap
Couepia_obovata_115_Circul
Couepia_subcordata_Sothers1511_Circul
Couepia_williamsii_U_Circul
Couepia_hallwachsiae_55_Circul
Couepia_polyandra_JASMITH1530_Circul
Couepia_scottmorii_58_Circul
Couepia_grandiflora_Sothers1507_Circul
Couepia_bracteosa_113_Circul
Couepia_paraensis_Sothers1510_Circul
Couepia_uiti_S_Option1
Couepia_parvifolia_N_Circul
Couepia_monteclarensis_K_Circul
Couepia_glabra_F_remap
Couepia_bernardii_B_Option1
Couepia_ovalifolia_Sothers1509_Option1
Couepia_schottii_Q_Option1
Couepia_coarctata_54_Circul
Couepia_bondarii_53_Circul
Moquilea_pyrifolia_39_Option1
Moquilea_tomentosa_Ltom_Bardon16_remap
Moquilea_guianensis_LCC_remap
Moquilea_macrocarpa_139_Option1
Moquilea_brittoniana_GEG_Option1
Moquilea_corniculata_73_Circul
Moquilea_minutiflora_NH200162_Circul
Geobalanus_oblongifolius_20553S2_Option2
Hunga_mackeana_37_Option2
Hunga_rhamnoides_38_Option2
Hunga_gerontogea_H_gerontogea_Circul
Angelesia_splendens_Lsplen_Option2_VERIFICAR_PETB
Exellodendron_barbatum_A110225_Option1
Acioa_longipendula_109_Circul
Acioa_guianensis_KD5285_Bardon16_remap
Acioa_edulis_49_Circul
Chrysobalanus_icaco_Ci_Option1
Chrysobalanus_cuspidatus_50_Option1
Chrysobalanus_prancei_52_Circul
Dactyladenia_floretii_KX180064_Dactyladenia
Dactyladenia_bellayana_KX180062_RFERENCIA_Dactyladenia
Dactyladenia_buchneri_KX180063_Dactyladenia
Hirtella_zanzibarica_NC_030563_Hirtella
Maranthes_kerstingii_F1P1C4986_Bardon16_remap
Maranthes_robusta_26_remap
Maranthes_gabunensis_GD1917_Option1
Maranthes_panamensis_143_Option1
Maranthes_glabra_GD1845_Circul
Atuna_racemosa_Aracemosa_Option_lastz
Parastemon_urophyllus_SARAWAK100485_Bardon16_remap
Grangeria_borbonica_Gbor_Option2
Magnistipula_butayei_Bardon16
Parinari_campestris_Pca_Male14_remap
Parinari_montana_LAK_remap
Parinari_alvimii_74_remap
Parinari_obtusifolia_77_Option1
Parinari_excelsa_75_remap
Parinari_oblongifolia_SAN152336_Bardon16_remap
Parinari_canarioides_LCI_remap
Parinari_anamensis_LCL_remap
Parinari_capensis_Pcap_Option1
Parinari_nonda_76_remap
Parinari_curatellifolia_Pcura_Option2
Neocarya_macrophylla_Nmac_Contigs_lastz
Kostermanthus_robustus_Kostermanthus_Bardon16_remap
Kostermanthus_heteropetalus_69_Option1
Bafodeya_benna_110_remap
NC_015308_Hevea_brasiliensis_Hevea
NC_030601_Erythroxylum_novogranatense_Erythroxylum
NC_036341_Garcinia_mangostana_Garcinia
0.01